© Copyright 2019

Joshua A Mathias

Contextual Scripture Recommendation for Writers

Joshua A Mathias

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2019

Reading Committee:

Ryan Georgi, Chair

Deryle Lonsdale

Program Authorized to Offer Degree:

Linguistics

University of Washington

Abstract

Contextual Scripture Recommendation for Writers

Joshua A Mathias

Chair of the Supervisory Committee:
Dr. Ryan Georgi
Department of Linguistics

Recommendation of book passages, quotes, or citations based on a given text can aid writing, research, literary analysis, and the incorporation of legal references (e.g. laws, previous cases). Each of these applications requires domain-specific data or knowledge. This thesis focuses on recommending verses of scripture, including the highly cited King James Bible, given a paragraph as input. To represent verses and queries, I compare the TF-IDF Bag-of-Words representation with the Universal Sentence Encoder. I experiment with a novel approach to weight the context of citations in the training data. I propose the use of a transformation matrix to recommend uncited verses using the contexts of cited verses. I discuss the need to evaluate both utility and diversity of recommendations, and compare automatic metrics with a small human evaluation.

TABLE OF CONTENTS

List of Figures

List of Tables

ACKNOWLEDGEMENTS

I thank Ryan Georgi for taking the time to be my advisor and give me sincere feedback and ideas, providing a fresh perspective. I thank Fei Xia for the initial discussions, ensuring my research task can be properly tested, and her Topics in Computational Linguistics class where this research began. I thank Deryle Lonsdale for miraculously becoming my reader in short notice and for helping me become a computational linguist in multiple ways these past few years.

DEDICATION

I dedicate this thesis to my supportive wife who provided qualitative analysis, my daughter for being a sounding board for my research the first few months of her life, and all students of scripture who may benefit from this work.

Chapter 1. INTRODUCTION

"The art of writing is the art of discovering what you believe." — Gustave Flaubert

1.1 Motivation

Millions of people of various religions spend years or their whole lives studying specific books of scripture (including the Bible). For many, scripture study involves writing. Here are a few examples of writing tasks that incorporate or cite scripture:

- A minister or pastor prepares a speech for a congregation. They seek scriptures that relate to the topic of discourse and add persuasive power.
- A teacher or missionary prepares a lesson for a class or individual. They seek scriptures that apply to the circumstances and questions of those they teach.
- A blogger writes about a life lesson or experience. They seek related scriptures that connect with their readers.
- An individual writes notes in a study journal about personal questions or a topic, which may not appear directly in scripture. They seek scriptures that they could apply to their question or topic.

The author can be aided in each of these purposes by a system that automatically recommends scriptures based on an input text containing the issues the author wishes to address. This thesis provides a framework for this task and a detailed analysis of possible challenges and solutions.

Scripture recommendation for writers is also broadly applicable to other domains, where recommendations come from other sources of text (especially semi-structured or commonly cited texts) instead of scriptures. In the legal domain, a lawyer may wish to incorporate passages from previous legal proceedings (legal precedent) or laws in a document they're writing for a current case. In journalism, an author may wish to incorporate quotes or statements in relation to an event or topic. Researchers and students may wish to incorporate quotes or information from previous research (Bradshaw 2007; Duma 2014). While this thesis doesn't investigate these applications, I hope it is clear that this work applies to other citation or quote recommendation tasks.

1.2 Description of Task

The proposed task is to recommend passages of a specific book or document given a text, to aid writers. A possible use case is to suggest passages or quotes to a writer as they're composing a document or speech. Hence, the given input text could be a sentence, a paragraph, or a whole document. First, I will define a few terms for the scripture recommendation task:

- **Verse:** A small numbered section within a chapter of a book of scripture. A verse may contain a sentence fragment or multiple sentences.
- **Volume:** A collection of books of scripture (e.g. the Old Testament or the New Testament in the Bible).
- **Content:** In recommendation systems, the "content" is characteristics of the items to be recommended, in this case the text of verses.
- **Document:** A General Conference talk or chapter in Teachings of Presidents of the Church (see 3.1), usually speaking on a specific topic. In a single document, there are many contexts (paragraphs or headers).
- **Document structure:** The relationships among parts of a document. For contextual data, within a document there are sections which consist of headers and paragraphs. For scripture, the document structure refers to the hierarchical sections within all books of scripture considered (see 3.2): verses, chapters, books, and volumes.
- **Citation:** In a document, a reference to one or more verses of scripture. Citations are generally incorporated within the document or writing, but I also make use of "related scriptures" or references that are listed at the end of documents.
- **Context:** In recommendation systems, "context" is information to be taken into account when recommending items. For this task, context for a citation and its verses is any text in the document containing a citation. A single context within a document is a paragraph or header, so a context may be a sentence fragment or multiple sentences. At test time, contexts are used as queries.
- **Query:** A query is the input that the recommendation system uses to find and recommend relevant verses. In this thesis, queries are generally taken from contextual data in the form

of a paragraph. However, in 5.6 I discuss the results for a query taken from elsewhere (the United States Pledge of Allegiance).

- **Serendipity:** A recommendation system is serendipitous if it recommends items that are both useful to the user and unexpected, or not a recommendation they would come up with on their own.

With recommender systems, a good list of recommendations is determined not only by relevance to the given text (Ge 2010), but the overall helpfulness of the recommendations provided and customer satisfaction (see 2.2). Therefore, part of this task involves ensuring that recommendations are novel, useful, and diverse. In this work I evaluate utility (relevance or accuracy) and diversity.

The King James Bible is a common source of passage quotations by authors and speakers, and is the main data source of recommendations for this thesis. However, the proposed methods should apply to other books as well, and I include other similar scriptural texts as recommendation sources (see Chapter 3). The King James Bible was written in early modern English (published in 1611), but translations of the Bible are found in many languages, including modern English (Christodoulopoulos 2015). The contextual data used for this thesis is also available in many languages (see 3.1). Consequently, scripture recommendation and the methods in this thesis can be applied to other languages. However, the pretrained English model (Universal Sentence Encoder; see 4.6.2) is not available in other languages.

Also, since we're focusing on specific texts to be recommended, representations of the text of verses (content) can be specialized (without concerns for overfitting the model), but representations of contexts or queries (and verses in relation to these queries) needs to be robust. A few different options for the length of recommendations are a single verse, multiple verses, or a chapter. I'm choosing to recommend a single verse, the idea being that the user could then look at the surrounding context. As with the length of input, deciding whether to recommend part of a verse or multiple verses could be a project of future work.

Blackbox description of system:

Input: Paragraph of context for which to make recommendations (user-generated).

Output: An ordered list of useful verses of scripture relevant to the input text.

With this thesis I hope to encourage and advance the field of text-to-text recommendation, especially book passage recommendation for writers. Consequently, much of Chapter 2 and the entire Chapter 6 are dedicated to exploring future work.

1.3 Research Questions and Hypotheses

1. Main research question: Does using document structure improve document modeling for book passage recommendation?

Sub-questions (applied to scripture verse recommendation):

a. Does weighting context according to document structure and proximity improve context representation?
    i. Hypothesis: Yes, weighting contexts (using WC-QRM, defined in 4.4) will show improved scores for scripture recommendation compared to not using weightings.
b. Does using hierarchical sections of a document to train word deviations improve content representation? (left for future work: see 6.1)

2. Secondary research question: Can training contexts of cited passages be extended effectively to other (uncited) passages according to content similarity among passages?

Hypothesis: Yes, transformed verse representations (see 4.5) based on the training contexts of similar verses will result in scripture recommendation for uncited verses better than the following baselines that don't use contexts:

a. Random (see 4.7.1)
b. Fixed Popularity (see 4.7.2)
c. Direct Query-to-Verse Similarity (see 4.7.3)

1.4 Outline

Chapter 2 provides a survey of previous work that relates to scripture recommendation for writers and explores possible ways to apply previous methods. I also describe each evaluation metric to be used in this study.

Chapter 3 describes the contextual data for citations and the text to be recommended, with statistics on how citations are distributed.

Chapter 4 describes the chosen methods for the experiments of this thesis, beginning with data processing and then the recommendation model.

Chapter 5 presents experiments and results, with discussion, including some qualitative analysis and human evaluation.

Chapter 6 outlines my proposed model, potential additions to the methods of this work, and other methods to compare. This chapter also discusses options for enhancing context selection, diversity, and serendipity. Finally, I propose a method for detailed human evaluation.

Chapter 7 summarizes contributions of this work and how it applies to other fields or tasks.

Chapter 2. Previous Work

Recommender systems have gained popularity in recent years for online services such as Amazon products, Netflix movies, Google Play Music, etc. There are also book and research paper recommender systems. Most or many recommendation applications and research use personalized recommendations (by modeling and comparing users) with methods such as collaborative filtering. My proposed research is user-agnostic: the information I have of the user only consists of a given input text. I was unable to find previous research on suggesting passages from the Bible to writers. However, I found two systems that automatically recommend Bible verses: Havas Cognitive (using IBM's API) recommends Bible verses on social media to advertise the television show Young Pope (Havas n.d.), and openbible.info (Smith n.d.) recommends Bible verses based on topical queries or keywords. Google's Talk to Books (Semantic Experiences n.d.) is also a recent example of book passage recommendation, in this case using a question as input. In this chapter, I seek to provide a comprehensive analysis of research relating to different aspects of scripture recommendation for writers.

2.1 Methods

Many methods apply to scripture recommendation for writers, which I review, but I do not experiment with most methods. At the end of each section I briefly explain how the methods in question apply to this thesis. See 2.1.15 for a list of methods I experiment with directly.

2.1.1 Topic Modeling

Topic modeling is commonly used for recommendation. A popular method for topic modeling is Latent Dirichlet allocation (LDA), which models a document as a "mixture over an underlying set of topics" (Blei 2003). LDA is usually trained or estimated using Collapsed Gibbs Sampling (Canini 2009). Bean (2016) applies a recommendation system to the religious domain, in this case recommending religious speeches (from General Conference) by The Church of Jesus Christ of Latter-day Saints, given another religious speech as input. I also use speeches

from General Conference in this thesis (see 3.1.1), but as contextual data instead of as a source for recommendations. Bean compares TF-IDF and LDA using k-NN to rank the recommended speeches, where cosine similarity is used for TF-IDF and Hellinger distance is used for LDA. He finds the LDA topic model to be more effective, in terms of coverage, than TF-IDF, as LDA creates "a more descriptive model of underlying structure" (Bean 2016). To train LDA, Bean uses Collapsed Gibbs Sampling and the Mallet implementation. LDA is a common baseline, and many variants of LDA have been proposed. In the case of scripture recommendation for writers, however, I suspect that the language of the user's writing and the language of scriptures are so divergent that a word-based topic comparison would be ineffective in most cases. I experiment with TF-IDF in this thesis because it is the original method used in the C-QRM approach for modeling quotes by their contexts (see 2.1.3). Applying this approach to LDA is left for future work.

2.1.2 Modeling Segments and Hierarchy

The Bible consists of hierarchical sections, namely verses, chapters, books, and volumes (the Old and New Testaments). A chapter consists of verses (containing a phrase or a few sentences), a book consists of chapters, etc. Coeckelbergs and van Hooland (2016) experiment using LDA to model the Hebrew Bible and recommend researching "the similarity between similar topics from different distributions, and several smaller levels of scope to perform Topic Modeling on, such as individual Bible books, chapters, or other meaningful units." They emphasize that the books of the Bible should be modeled separately, because they "diverge enormously regarding length, content, and style." While some style differences depend on Biblical translation or the collection in question, research has shown (L. Du 2012; H. Zhang 2016) that leveraging sectional information is effective in multiple domains. Other examples of meaningful units in the Bible are story lines spanning multiple chapters (Bible Study Guide 2018), the five sections of Psalms, the twenty-two stanzas within Psalms 119, sequential books of the same name (e.g. 1 and 2 Corinthians) and books of the same author (e.g. the thirteen Epistles of Paul). Such sectional diversity within the Bible leaves much room for future research, including automatic segmentation at multiple levels. In addition, trends in using ranges among

cited scripture references (e.g. Gen. 1:1–2:7) and scriptures that are commonly cited together may provide additional information for segmentation.

L. Du et al. (2012) propose SeqLDA, a probabilistic model based on LDA that leverages the structure of chapters or sections in books for improved topic modeling, and for a given segment takes into account the previous segment. L. Du et al. (2013) build on this idea and propose a hierarchical Bayesian topic model to segment text linearly. Eisenstein (2009) proposes a Bayesian model for hierarchical text segmentation. J. Chen et al. (2016) use hierarchical LDA to create (and visually present) a topic hierarchy tree, given a predefined number of layers. Their model automatically merges paragraphs of a similar topic into a larger semantic unit or topic based on a threshold. These and similar methods could be used in passage (or verse) recommendation to encode the intuition that the salient topics in a chapter are likely the most salient topics in passages within that chapter; this can be further expanded to books, volumes, etc.

More recently, Amoualian et al. (2017) propose a probabilistic model based on LDA to automatically segment text into "topically coherent segments of words" (2017). This type of automatic segmentation could be used in future work to determine how large of a book passage to recommend (e.g. multiple verses, parts of a verse or quote) as well as to identify portions of a document to which recommendations should be assigned (proactively aiding the writer). As Amoualian et al. suggest, this type of topic segmentation can also help identify the local topics of a segment that would otherwise be missed or overshadowed by the topics of the whole document.

TF-IDF bag-of-words vectors have also been used for hierarchical topic modeling (Nistér 2006). However, for recommending book passages, I wish to model hierarchy that already exists in the text (e.g. verses, chapters, books), whereas most hierarchical methods in previous research create hierarchical topic models independent of the structure (though this may also prove useful) or find hierarchical segments. Nonetheless, the methods mentioned and others provide insight into how hierarchical information can be modeled.

H. Zhang et al. (2016) do model existing textual hierarchy; they propose using a multilayer self-organizing map (MLSOM) for recommending authors and books by modeling the

hierarchical topics of authors, books, pages, and paragraphs, noting that a flat representation such as term frequency is not reliable with large texts that contain large vocabularies but with varying spatial distributions of terms. For recommending authors, they use authors' biographies as well as their respective books. Their MLSOM method outperforms LDA, but they don't compare against hierarchical LDA or other hierarchical methods. In 2018, H. Zhang et al. propose a Tree2Vector representation of books and locality reconstruction (2018a, 2018b), which uses information from higher level nodes to represent lower levels and vice versa. The model is trained to weight the importance of each child node to its parent node.

Overall, previous research has shown that there are many ways to improve text representation using hierarchy. Zheng et al. (2015) specifically apply hierarchical CNN to represent a paragraph as a question in order to suggest an answer; during training they use two separate objective functions for sentence and paragraph representation which are then summed for the final objective function. Tai et al. (2015) find that tree long short-term memory (LSTM) consistently outperforms sequential LSTM for shorter and longer sentences, demonstrating the applicability of hierarchical models even for short texts. For tag recommendation of scientific papers, Hassan et al. (2018) apply a hierarchical attention network to model the importance of words and sentences in a document. For other examples of modeling hierarchy, see Appendix D.

This thesis does not apply hierarchical methods, but I see it as an important step for future improvements (see Chapter 6 for recommended methods).

2.1.3 Context-based Citation Recommendation

As the Bible is often quoted or cited in writings, frequently by numeric reference (e.g. Gen. 1:1), the contexts of such citations may be used to model what types of contexts are associated with verses in the Bible, and which information could be leveraged further with cross-references among verses. He et al. (2010) define a probabilistic model for context-based recommendation and represent contexts as TF-IDF scores normalized into unit vectors; similarity to the query context is then computed using the Frobenius norm. He et al. compare their context-based method with other non-context methods such as finding documents that share authors with the given document and finding the similarity of the title and abstract of the other

documents to the given document. As the context-aware methods of He et al. have best results on citation recommendation, and combining them with other methods further improves results, they make a strong case for the use of context for recommendation.

W. Huang et al. (2015) define a neural probabilistic model that learns to represent context words (using Skip-Grams) and their relationship to documents (using noise contrastive estimation) with stochastic gradient descent. By comparing with baselines such as Cite PLSA LDA (Kataria 2010), Citation Translation Model (using GIZA++; E. H. Huang 2012) and word2vec (Mikolov 2013) word and document representations, W. Huang et al. (2015) show that directly modeling context words in relation to cited documents achieves state-of-the-art results.

Ahn et al. (2016) compare context clustering and context lumping. Context clustering uses the cosine similarity of the query context and the candidate quotes' most similar context cluster, using TF-IDF vector representations. In context lumping, all contexts of a candidate quote are concatenated and similarly compared with the query context using TF-IDF and cosine similarity. Context lumping performed significantly better, likely because of the sparsity problem of context clustering, and with context lumping no context information is lost. Tan et al. (2018) show similar success with an analogous method called C-QRM, where the TF-IDF vectors of a quotes' contexts are summed and then normalized. In both methods, quotes are represented directly by their contexts in the training data. C-QRM is more versatile than context lumping as it doesn't require concatenating words and can be used with representations of contexts other than TF-IDF bag-of-words vectors. In this thesis I experiment with C-QRM.

2.1.4 Context Selection

To use context effectively for recommendation, selecting the right context before or after the location of a potential citation can significantly affect results. In addition, the optimal amount of context to use depends on the algorithm. Ritchie (2008) investigates different lengths of context and finds that using a fixed window (such as 100 words) is simple and effective. Indeed, a fixed window of words is a common approach in recent research, with alternatives employing a fixed number of characters, sentences or a paragraph. He et al. (2010), based on Ritchie's research, opt to use 50 words before and after a citation. Duma and Klein (2014) experiment with

using 10, 20, and 30 words before and after a citation as a query, finding 20 words to work best. They similarly experiment with different windows of context words to represent the documents to be recommended and find that 10 words before and 5 after has best results when only using context to suggest recommendations. It may be worth investigating whether using more context before a citation than after, or weighting pre-context higher than post-context, would improve results for other methods. Furthermore, pre-context and post-context could be represented or trained separately to model their different relationships to the cited texts.

Ahn et al. (2016), for quote recommendation, use different amounts of context for the algorithms they compare and for the various methods they combine using rank aggregation, based on empirical preference: 150 characters before and after for context clustering and lumping, 50 words for Random Forest, and 30 words for CNN (convolutional neural network) and RNN (recurrent neural network). They only use pre-context for Twitter dialogs to reflect the real-world application of quote recommendation in dialogs (where post-context would be unavailable).

Tan et al. (2018), also for quote recommendation, use 300 words before and after a quote for English and 100 words for Chinese (presumably because Chinese uses less written words than English to express the same meaning). How much context to use, or which context to use, may vary greatly between languages and domains. Also, the context before and after a quote is encoded separately using LSTM because LSTM performs poorly on longer texts, and then the resulting context vectors are summed (Tan 2018). For idiom recommendation, Y. Liu et al. (2018) distinguish between global context as all the context available for an idiom (the document), and local context as the sentence in which an idiom is found (or to be recommended). Embedding both types of context together performs better than either type of context alone. Duma et al. (2016) compare context selection methods, including human-aided context selection, where a human annotator indicated which sentences were relevant or useful context for the citation, within a window of two sentences (before and after). Using these human-chosen sentences performed best, and the second best selection strategy used a relatively large range of 500 words before and after. These results suggest that a context-based recommendation system can likely be improved by an effective context selection approach that selects context based on

some measure of utility, such as similarity to the cited text. Indeed, Duma and Klein (2014) recommend a “more targeted extraction of text” as future work. As will be discussed in the section entitled Document Representation, previous research has shown that representing a larger document by its best matching passages is more effective for retrieval than using the whole document. As a note, research on content selection for summarization may also apply to context selection for recommendation.

Context selection varies greatly among previous research and merits further experimentation. For serendipitous scripture recommendation, context that doesn’t directly relate to the citation (such as paragraphs far from the citation but in the same document) may help train the system to recommend verses that are indirectly related. In the ideal scenario, given any paragraph in a religious writing, the recommendation system could recommend verses of scripture that would be relevant or useful to cite anywhere else in the document, with a preference to verses that should be cited closer to the given paragraph. As this particular task is novel, I experiment with a new weighting scheme (see 4.2) that uses document structure to weight every text segment (paragraph) of context in a document in relation to every verse cited in that document.

2.1.5 Reference Identification

Depending on the corpus, identifying quotations of or references to book passages is a challenging task that may or may not require machine learning, depending on whether one is seeking only exact matches or how comprehensively one wishes to find such references or even allusions to the text. For recommending quotes to writers, even indirect references or allusions are relevant. The following sections discuss two types of reference identification: finding quotations or allusions in the text and parsing citations by name and number (e.g. Genesis 1:20).

2.1.5.1 Quotation Identification

Tan et al. (2018) use quotes from libraryofquotes.com to train their quote recommendation system. They remove duplicates and have an initial 380,000 quotes. They search Project Gutenberg for these quotes and find 102,407 total context quote pairs. They do not

specify how matches are found, so I assume they search for exact matches (in both English and Chinese), after which they perform preprocessing (in the case of English: lowercasing, stemming, removing stop words). I suspect that quote matching in plain text would be more successful in many languages (at least European languages) after preprocessing. After removing quotes that are quoted less than five times they have 3,158 quotes with 64,323 pairs. For Chinese, they use Chinese poems and don't include poem sentences that appear only once, which brings the total number of contexts from 96,927 to 56,949. Y. Liu et al. (2018) go through a similar process to obtain data for idiom recommendation, using a Chinese microblogging site and a news search engine. For scripture recommendation, the set of possible quotes (verses) to recommend is fixed and not dependent on whether they are quoted in the data.

In the case of book passage recommendation, we may want to include all passages as possible recommendations. As with many books, some Bible verses are quoted much more often than others, and some may not be quoted at all. For cases where verses have no contexts or citations in the training data, a context based model may need to be adapted, such as using only the content of the verses. For verses with relatively few contexts, context information (as opposed to verse content) could be weighted by how many training contexts were available. Also, the Bible often quotes itself (especially the New Testament quoting the Old Testament), so one issue is merging such quotation (such as labeling the quotation by its earliest occurrence in the Bible).

Mullen (2016) researches quotations of Bible verses in American newspapers from the Library of Congress, focusing on the most frequently quoted passages and their relative frequency over the years. He also creates a visual network of chapters in the Bible that were quoted on the same page at least 350 times. Mullen uses machine learning to capture not only direct quotes but also paraphrases and, to some degree, allusions to Bible verses (one example being "Jug not lest ye be Jugged" in reference to Matthew 7:1). Mullen's research uses the King James translation of the Bible and trains a neural network classifier using four features: frequency of matching n-gram tokens, TF-IDF values of words, the proportion of the verse that is quoted, and Wolfowitz Runs Test to measure the proximity of parts of a verse. A manually created dataset of 1,700 classified possible matches is used to optimize the receiver operating

characteristic. Mullen's code and data is open source and is a potential data source for Bible verse recommendation.

JSTOR Labs' "Understanding Shakespeare" system searches journal articles for quotes from Shakespeare's plays. First, they filter articles to include only articles that mention "Shakespeare" and the name of one of the players; then, only block quotes and passages within quotation marks are considered. A similarity score using the Levenshtein distance and quote length determines the best match for each quote. Matches are then filtered based on similarity score as well as the length of the original match in the play (which is related to Mullen's use of the ratio of the Bible verse that is quoted). However, JSTOR Labs opts to not include quotes of less than 15 characters, to avoid false positives, and suggest taking into account "how common a phrase is in modern usage." Mullen's strategy of using TF-IDF would likely work for this purpose.

For the contextual data used in this thesis, there is no need to identify verses of scripture by word matching, as all scriptures are found in the data as an HTML link (the link is parsed to identify the verses). Furthermore, my recommendation system considers all verses for recommendation regardless of the number of times they are quoted or cited, but methods that only use previous contexts for citations will prefer verses that have been cited previously (if there are no contexts for a verse, its relevance score is 0).

2.1.5.2 Citation Parsing

The Bible is one of the most (perhaps the most) frequently cited sources, and its structured nature (book, chapter, verse) makes it convenient to cite, but while there are common acronyms and formats for citing the Bible, they have many variants and in some cases coincide with words or acronyms not referring to the Bible (such as "The" for Thessalonians). Smith (n.d.) created an open source Bible Chapter Verse (BCV) parser to recognize and parse Bible citations in Twitter posts, including ranges (e.g. Matt. 1:5 - Matt. 3:7), misspellings, different punctuation, and different versions of the Bible. Shafer (n.d.) also released a customizable Bible reference parser.

How strict requirements should be when finding Bible references (such as requiring the first letter of references to a book to be uppercase) depends on the corpus. When searching only books or published works (as opposed to Twitter) one can expect the proper use of case and few misspellings. In the data used for this thesis, scripture verses are cited explicitly using a standard HTML and URL format (all scripture citations are links), so a Bible reference parser (or citation parser) is not needed, except to parse scripture URLs according to the format used by lds.org (now ChurchofJesusChrist.org).

2.1.6 Cross-Domain and Multilingual Recommendation

The English language or word use of the King James Bible is often very different from the paragraph input or query (context) due to diachronic mismatches (see 3.2.1). This applies to quote recommendation as well (Tan 2018). There has been some research in cross-language topic modeling, such as with Bi-LSTM (F. Li 2017), bilingual context-citation embedding (X. Tang 2014), BiLDA (Mimno 2009), and BiSTM (Bilingual Segmented Topic Model), which makes use of aligned segments within documents (Tamura 2016). These methods could be applied to translating and comparing the topic of a book passage to the topic of a given context. Lu et al. (2011) directly apply a translation model to context-based recommendation as the probability of words in a citation's context corresponding with words in the cited document or abstract. This model may be a useful feature combined with other methods.

The QuoteRec system of Tan et al. (2018) models this language difference more directly by using a neural network that trains a deviation matrix for the words of a quote based on the quote's topic and author, while also sharing a word embedding among all quotes. A separate word embedding is trained for contexts, which is used in the loss function for training quote embeddings. Tan et al. achieve state-of-the-art results in recommending historic Chinese poetry as well as English quotations from Project Gutenberg; hence, this method is likely well-suited to Bible verse recommendation, or recommending passages from historic texts.

Paul et al. (2009) propose cross-collection LDA (ccLDA) to jointly model the topics of all collections and the topics of specific collections (subtopics), and then compare the subtopics. They also propose a method to identify "problematic topics," where a specific collection doesn't

truly contain the topic in question (or to a lesser extent). They further extend ccLDA to a hierarchical model which automatically finds subtopics (or no subtopic). ccLDA could be used to compare topics of different sections in the Bible or different collections of citation context. Bao et al. (2013) extend ccLDA with a partially supervised model (PSCCLDA) and C. Chen et al. (2015) propose differential topic modeling using the full Pitman-Yor process, also creating topic hierarchies.

In this thesis, I do not apply these previous methods, but in section 6.1 I propose using the QuoteRec model (Tan 2018) to model word use in scripture. In order to relate scripture verses with modern language queries, I represent verses using the context-based C-QRM approach, which forms part of the QuoteRec model (Tan 2018). To apply this context-based approach to verses with no citations, I represent words in verses by words in contexts to provide a contextual representation of uncited verses (see 4.5). I also experiment with comparing the words of verses and queries directly (see 4.7.3).

2.1.7 External Information

The QuoteRec model (Tan 2018) also provides a framework for leveraging external or meta information about passages. For example, which book in the Bible a verse corresponds to could be used instead of the quote author. Hierarchical segment information could be used by treating each chapter, book, and Old and New Testaments as distinct authors, and then summing each corresponding vector similar to how multiple topics are summed in the QuoteRec model.

A recent survey of deep learning based recommender systems (S. Zhang 2017) reviews the use of different neural network paradigms for incorporating heterogeneous information and suggests continuing in this direction for future work, as neural networks can optimize the use of different features and automatize feature engineering. S. Zhang et al. recommend DeepFM (Deep Factorization Machine) as one potential method for this as it balances memorizing and generalizing (with a multilayer perceptron) which is similar to the Wide and Deep Learning approach, which Google successfully applied to recommendation (Cheng 2016). FM (Rendle 2010) is especially useful for modeling low level information such as movie frames (Deldjoo 2016).

Learning to rank is another method to effectively combine heterogeneous features or information, focused on optimizing ranking (H. Li 2011). Tan et al. (2015) use 16 features for quote recommendation (such as LDA, TF-IDF, user votes, popularity, and context similarity) and optimize NDCG@*5* (comparing the ideal ranking, where only the first five ranked items are considered). While their later neural QuoteRec system performed better, the learning to rank approach is a strong baseline and could be improved upon through the use of different features. Ahn et al. (2016) propose a similar approach called rank aggregation, where the final ranking is weighted by the rankings of multiple algorithms, where a higher original ranking has a greater weight. They further improve this method by using only the top-k ranked items for each algorithm (setting *k* to 3).

There are various sources of external information on scripture verses (such as verses organized by topic) that could be used to improve recommendations, but this is left for future work (see 6.4).

2.1.8 Collaborative Filtering

Collaborative filtering, which leverages the preferences, ratings, or history of other similar users, is a common approach to recommender systems. For sports news recommendation, Lenhart et al. (2016) use an adjusted cosine between news articles based on users who rated both articles, taking into account the average rating of each user. This method combined with content-based similarity (based on keywords automatically generated from an article's text) had highest results. This type of collaborative filtering approach may be used with Bible verse recommendation to leverage user-generated ratings of verses based on a queried topic (topic scores), where each topic is analogous to a user in collaborative filtering.

Consequently, past research strongly encourages a hybrid approach that combines external information and various methods to make a recommendation. This type of information is often available in production settings; in the case of the Bible, there are user topic scores, verses categorized by topic, cross-references between verses, chapter and book information, etc.

Collaborative filtering isn't used in this work but should be considered for a production system with users or as a way to incorporate other data sources.

2.1.9 Word Representation

A common approach to word representation in text-based recommendation is TF-IDF, as it takes topical importance into account. He et al. (2010) normalize TF-IDF scores into a unit vector to represent each citation context. Nguyen et al. (2015) propose combining a latent feature word representation with topic modeling such as LDA, and they experiment with Google's word2vec and Stanford's glove word representations, which perform about the same. There are two common approaches for neurally trained word vectors: Continuous Bag-of-Words (CBOW) and Skip-gram (Mikolov 2013), and Skip-gram usually uses negative sampling. Rosati et al. (2016) compare both methods using different word vector sizes and find that Skip-gram performs better for movie and book recommendation. W. Huang et al. (2015) use Skip-gram for citation context word embeddings using a sliding window, and then "noise contrastive estimation to model the relationship between words and documents"; Y. Liu et al. (2018) also use Skip-gram for idiom recommendation. Tan et al. (2016) use CBOW for quote recommendation, but as discussed earlier, in their 2018 paper they find that training word embeddings using author, topic, and context information together has better results.

In addition to Tan et al. (2018), other research has shown promising results with multiple representations for each word. E. H. Huang et al. (2012) use both local context (nearby words) and global context (the whole document as a bag-of-words vector), where context words are weighted by IDF. Then spherical k-means (k is set to 10) is used to cluster context representations for a word. At test time the new context is compared to each cluster for the given word. J. Li et al. (2015a) also use global context and clustering for multi-sense embeddings, but their model automatically determines the number of representations for each word based on the probability of a given context being in an existing word sense. They found improved results for some but not all natural language understanding tasks.

C. Li et al. (2014) propose Combination with Complementary Languages (CCL) to improve word embeddings by choosing the center word embedding for each word among multiple related languages, which significantly improved results. As there are many translations

of the Bible (Christodoulopoulos 2015), even within the same language (e.g. modern vs. old English), parallel translations could be used to improve word embeddings.

For this thesis, TF-IDF weightings are used to represent words in a bag-of-words vector (where the vocabulary size is the vector length or dimension) as in He et al. (2010). Instead of distinguishing between global context and local context as in E. H. Huang et al. (2012), I weight contexts based on whether they appear near the citation or are in the same section and use this weighting when summing contexts in the C-QRM approach of Tan et al. (2018). As an avenue for future work, the neural word embeddings discussed could be used under this approach instead of TF-IDF weights, allowing for a more rich representation of each word.

2.1.10 Preprocessing

As text-based recommendation generally focuses on topic information, some common preprocessing methods are case normalization (Bean 2016; W. Huang 2015; Xiong 2016), rare word removal (W. Huang 2015, who removes words that appear less than 5 times), stemming (Xiong 2016; Suzuki 2009), lemmatization (Suzuki 2009), and stop word removal (in most or all of the text recommendation studies mentioned). Bean (2016) researched recommendation of religious speeches and also removed frequent domain-specific stop words such as (case-normalized) “christ” and “god,” but without proving that this aided the recommendation system. Schofield et al. (2017) evaluates the effect of removing stop words when training LDA for document classification and find that removing stop words helps, but removing stop words after training performs about the same as removing stop words before training. They also suggest that a curated domain-specific stop words list is unlikely to be beneficial, or at least the domain specific words need not be removed as an initial preprocessing step. This is a potential area of future work, especially in the religious domain. Recommendations for users interested in a very specific but infrequent topic or word requires retaining infrequent words, and if I were to remove common yet meaningful words (e.g. Christ, God) I would limit the recommendation of these common topics. In many cases a user may seek a specific or less frequent topic in relation to a common topic (e.g. Job in relation to prayer or birth in relation to Christ).

H. Zhang et al. (2016, 2018) model paragraphs, pages, and books from Project Gutenberg. Since Project Gutenberg texts contain many errors (such as from OCR), they use an English dictionary to form a vocabulary of 15,601 words after removing stop words and stemming. Paragraphs are separated by new lines. Pages consisting of complete paragraphs are segmented after reaching at least 1000 words; then, paragraphs are merged to reach at least 50 words. Depending on the data, this type of process could be used in preparation for context selection.

In this thesis, I remove no domain specific (or religious) stop words, but I do remove a few technical terms (see 4.1) in addition to typical stop words. I apply stemming and case normalization (see 4.1). Also, the data is already segmented into paragraphs using HTML formatting.

2.1.11 Document and Paragraph Representation

Sarwar et al. (2017) find using smaller representative passages for document representation to be effective. This type of document representation could be applied to context selection or representing Bible chapters, where the verse to be recommended could be chosen from the best matching chapter, based on multiple matching verses within the chapter. Similarly, for recommending scientific articles based on context, Duma and Klein (2014) propose an "internal" method which, for each document in the bibliography, finds the passage with the greatest cosine similarity to the query (the context of the new citation). They found that focusing on a larger matching passage (400 words) was more effective than a smaller passage when this is the only representation of the document, but using the best matching passage was more effective than using the whole document. Duma and Klein found improved results by combining this method with an "external" method, which compares the contexts where the proposed document has been cited previously with the context of the query (new citation). They also find that using just a cited document's contexts by itself performs better than using the best passage of the document, indicating that for citation recommendation, context may be a more useful or important feature than content.

Y. Liu et al. (2018), for idiom recommendation, model the global context (all context available) using a CNN with a sliding window "to capture the semantics of unigrams, bigrams and trigrams." Within local context (the sentence surrounding the idiom) they find "hint words" for the idiom by calculating the mutual information of each word in the idiom to each word in the local context, as the joint probability of the words being in the same sentence, normalized by the probability of each individual word. Combining both global and local context (for both the query and the recommended idiom) achieves best results. S. Bai et al. (2018) apply CNN to various NLP or language modeling tasks and show that CNN can outperform RNN in many cases, including with longer documents.

Cer et al. (2018) at Google released a pretrained model for sentence and paragraph embedding called Universal Sentence Encoder (USE). They show that this model can effectively be fine-tuned with in-domain data. There are two types of USE: a Deep Averaging Network (DAN) model and a Transformer Model. The Transformer model performs better but at a significant cost of memory and time, which increases with the length of input ($O(n^2)$), as opposed to $O(n)$ for the DAN model.

In this thesis I apply the pretrained DAN Universal Sentence Encoder to model paragraphs of context as well as verses to be recommended. Instead of selecting particular representative paragraphs from a document to use as context, I weight each paragraph and header in the document in relation to cited verses in the document (see 4.2).

2.1.12 Language Modeling

As we're seeking to recommend passages to writers, it may be helpful to use language modeling to recommend the best subsequent text based on the current text. Parapar et al. (2013) and Valcarge et al. (2015) specifically explore statistical language modeling in recommender systems with promising results. Recent research uses neural topic models for language modeling. A method introduced by Wang et al. (2017) reportedly outperforms Microsoft's TopicRNN (Dieng 2016) and produces a set of RNNs representing each topic, which can then be used as a language model, even to produce new sentences. Notably, Wang et al. (2017) report that a smaller language model performs better than a larger language model, which they conclude is

because the larger language model has too much impact on the topic model (they are trained together). Lau et al. (2017) create document vectors using a convolutional network. The document vector is used to create a weighted mean of topic vectors, which is then used with a standard LSTM language model. Lau et al. compare their model with LDA and outperform on topic coherence for two out of three datasets. These two papers suggest that combining topic modeling with language modeling improves performance on both tasks, especially for longer documents or contexts.

Another potential method is using hierarchical attentive memory as input to LSTM, which Andrychowicz and Kurach (2016) use to significantly increase the length of input that LSTM can effectively use, though this has not yet been applied to language modeling.

I do not apply language modeling by RNNs to scripture recommendation in this work. In Chapter 6 I propose experimenting with LSTM as future work.

2.1.13 Dialog Recommendation

As with language modeling, recommending book passages based on the current writing or text could be considered a predictive dialog task, and indeed Ahn et al. (2016) apply quote recommendation to dialogs in Twitter conversations. Serban et al. (2016) review the use of neural networks in dialog, a popular method being sequence to sequence (seq2seq) using RNN (Sutskever 2014), and they distinguish between a generative model, which creates new dialog, and a discriminative model, which chooses among existing candidate responses (similar to quote recommendation). Serban et al. (2016) propose a hierarchical model for generating dialog. Kannan et al. (2016) use seq2seq for Google's SmartReply system to suggest email responses, and Google's Talk to Books (Semantic Experiences) system uses a related dialog-based approach.

Henderson et al. (2017) suggest that seq2seq is an inadequate solution for SmartReply because seq2seq is a generative model whereas SmartReply is discriminative. They propose a hierarchical feed forward neural network instead, using n-gram bag-of-words embeddings and nearest neighbor search to find the best response. Ahn et al. (2016) address this weakness of seq2seq for quote recommendation by using CNN in addition to RNN.

Methods in dialog recommendation or generation are not applied in this thesis, but these methods provide another perspective on a very similar task (recommending specific text passages for writers).

2.1.14 Diversity

As some topics may be closely related, merging similar topics for topic compression can aid diversity by enabling recommending items from diverse topics in the recommendation list. L. Liu et al. (2015) perform topic compression for learning resource recommendation based on a user-selected passage, by removing low probability topics and topics with high cosine similarity or low KL divergence with other topics in the passage; the final ranking is weighted to take into account the importance of each topic to the passage. This type of topic compression may work well for recommending diverse Bible verses. L. Liu et al. (2015) also rank suggested resources based on similarity of reading level to the selected passage (using New Dale-Chall Score or Flesch Reading Ease).

Determining the optimal number of topics during training time and at test time given a query (or context) will likely improve diversity. L. Liu et al. (2015) suggest limiting a query's topics to two or three relatively diverse topics. As LDA requires a predefined number of topics, there has been some work in automatically finding the right number of topics, such as using KL divergence to compare word-topic and document-topic distributions (Arun 2010).

External information from users can also improve the diversity of recommendations. Lenhart et al. (2016) note that their use of collaborative filtering increased diversity (according to human evaluation). Niemann (2013) discusses the use of collaborative filtering for diversity in more detail.

In general, books are sorted into larger or hierarchical sections, and this is valuable information to leverage for diversity (to avoid recommending multiple passages from the same section), as well as whether passages or verses cite each other (indicating that they discuss a similar topic).

Conroy and O'Leary (2001) propose using QR matrix decomposition (QRD) to produce diverse summaries. When a sentence is added to the summary, the values of its terms are

subtracted from all other sentences to be considered. This method can be combined with various weighting schemes or representations of words as well as other features in order to produce a diverse list of recommendations. QRD is an intuitive and simple way to reward relevance to the query only when the relevant aspect is different from other items ranked higher in the recommendation list, yet previous research has not yet applied QRD to increasing diversity in recommender systems.

Maximizing entropy maximizes the equal distribution of probability among the different classes or categories given the available information; this is also an intuitive approach to optimize diversity while maintaining accuracy or relevance. Kawamura et al. (2018) use a thesaurus to identify hypernyms, hyponyms, and synonyms among terms and increase semantic diversity by maximizing the entropy of concepts (consisting of words). Then, paragraph vectors are created using cluster vectors instead of word vectors, where a cluster is represented by the concept with the highest entropy. Because terms that are more broad (hypernyms) have a higher entropy, diversity is increased by encouraging diverse representation from different broad concepts or clusters instead of only different specific concepts within one broad concept.

Gogna et al. (2017) propose a convex matrix completion method to balance accuracy and diversity during training in a unified model, as opposed to weighting the final ranking for diversity. To optimize diversity, the system encourages equal probability among the different categories (genres); they introduce a variance regularization term which is the variance of each genre, taking into account the average rating of each movie genre. While Gogna et al. just use this for movie genre, they describe how this regularization can be used with other features or domains as well. Further research is needed to determine if this method can be effectively used with latent topics or with terms (similar to the QRD approach).

For this thesis, diversity is evaluated (see 2.2.2.3), but I do not yet apply methods to diversify the recommendation list. However, as I explain in Chapter 4, I use the Transformation approach (see 4.5) and the Direct approach (see 4.7.3) to enable recommending verses that were not cited during training.

2.1.15 Summary (Methods)

NLP and machine learning research has provided a strong foundation of methods and insights which can be applied to book passage or scripture recommendation. The following strategies have shown promising results:

1. Combining methods, features, and metadata.
2. Leveraging hierarchy, in document structure and in models.
3. Representing and leveraging both context and content for recommendation.
4. Focusing on the most important parts of a text or document.

Furthermore, the following methods from previous work are directly applied in this thesis:

- TF-IDF bag-of-words vectors (Tan 2018)
- Universal Sentence Encoder (Cer 2018)
- C-QRM (Context-based Quote Recommendation Model; Tan 2018)

2.2 Evaluation

There is a wide variety of automatic evaluation metrics for recommendation. They focus on different goals of a recommendation system, such as finding relevant content, providing a representative (or diverse) summary of relevant content, or introducing a user to new items they might want to consume. First, I will review previous work on automatic metrics for recommendation; then, I will describe which metrics are used for this thesis and why. Finally, I will discuss the use of human evaluation in previous research on recommendation.

2.2.1 Automatic Evaluation

To evaluate recommendation of religious writings, Bean (2016) uses catalog coverage, which is a measure of the percentage of documents recommended out of the whole catalog after a large number of trial recommendations. Bean suggests evaluating serendipity, which jointly measures whether recommendations are useful and unexpected, as future work.

Ge (2010) describes coverage and serendipity in more detail, noting that a successful recommendation system isn't simply one with high accuracy or results similar to the user query, but rather results that are novel and useful to the user (this is a subtle point that may depend on the application). In other words, serendipity invokes positive surprise in users by giving them what they didn't know they wanted (and which they wouldn't likely find on their own; Ge 2010; Adamopoulos 2015). Coverage, on the other hand, is helpful in ensuring that the recommendation system applies to all types of users, and doesn't focus on only a small set of highly relevant (or popular) items. Coverage, and variants of it, is often used as a measure of diversity. The distinctions among unexpectedness, novelty, serendipity, coverage, and accuracy merit further analysis.

2.2.1.1 Unexpectedness, Novelty, and Serendipity

Adamopoulos and Tuzhilin (2015) formalize unexpectedness and distinguish it from novelty, serendipity, and diversity. According to their descriptions, novel items are specifically unknown to the user and not necessarily relevant or positive. Serendipitous items are both novel and positive, but accidental and, hence, more risky. Unexpected items differ from the set of expected items of a user, and Adamopoulos and Tuzhilin define the utility of unexpectedness as the most preferred distance from the set of expected items, as too large a distance from a user's expectations may be "perceived as irrelevant by the user and, hence, it is not truly unexpected" (2015).

Silveira et al. (2017) compare five unexpectedness metrics from the literature. Each metric makes use of a primitive model (baseline) or an alternative such as user history or popularity. The first metric, based on Ge et al. (2010), describes the general formula for serendipity:

```
L = recommendation list (system being evaluated)
PM = recommendations by the primitive model
UNEXP = L / PM
Serendipity = |useful items in UNEXP| / |UNEXP|
```

Bean (2016) decided not to use serendipity since his task was novel and there was no baseline to compare for unexpectedness. However, the baseline or primitive model can be achieved with a

measure of popularity or basic similarity (such as the number of words that the query and recommendation have in common). As an appropriate way to measure serendipity in scripture recommendation for writers is not clear or well established, I leave this for future work (see 6.2.7).

Novelty is often measured using a task-dependent measure of popularity. Gogna and Majumbar (2017) evaluate novelty by summing the log of the number of users over the number of ratings for each recommended item, to reward recommending items with fewer ratings. For citation or quote recommendation, popularity can be determined by the number of citations for each item. In this work I evaluate novelty only qualitatively as the number of citations of recommended verses, and I quantitatively analyze diversity (discussed in the next section). For scripture recommendation, my goal is to recommend diverse verses, without knowledge or preference as to whether a verse is novel to the user (this preference may vary by user).

2.2.1.2 Diversity

In general, diversity measures the variety or dissimilarity of recommendations. Gogna and Majumbar (2017) distinguish between two types of diversity: individual diversity, which considers each user individually, and aggregate diversity, which measures the diversity of recommendation among all users. They measure individual diversity by summing the cosine similarities between the items recommended to each user and averaging over all users. For aggregate diversity they use the total number of items ever recommended (catalog coverage). B. Bai et al. (2017) measure diversity using Diversity@N, which is catalog coverage, but explicit about how many (N) recommended items in each list are included in the metric. Niemann et al. (2013) when calculating aggregate diversity only reward for relevant or useful diverse items by using the intersection of all highly rated items and all recommended items. In section 2.2.2.3.1 I give this intersection of relevant items and aggregate diversity (or catalog coverage) the name Precision Coverage.

Gogna and Majumbar (2017) and Rosati et al. (2016) also measure aggregate diversity with the Gini coefficient. The Gini coefficient evaluates the uniformity of the probability distribution of the chance of each item being recommended. Note that if such an equally diverse

probability distribution is what one is optimizing, this evaluation metric may be biased. Furthermore, it may be hard to define and evaluate the right type of diversity for a given task. Where the data doesn't make this distinction clear, human evaluation may be required. For example, one may perform topic compression for diversity but need human evaluation to ensure that the topic compression is effective for the task at hand, or human evaluators may realize that another type of diversity is more or less important (e.g. time period, formality of writing, style, or book section).

2.2.1.3 Accuracy

One of the most popular metrics for recommender systems is Normalized Cumulative Discounted Gain (NDCG), which compares the generated recommendation list with the ideal or gold standard recommendation list, taking rank into account by discounting logarithmically the gain of each recommended item (Ricci 2011). Rosati et al. (2016) use NDCG@N (NDCG including only the first N recommendations), Recall@N (the number of gold standard recommendations included in the top N), and Precision@N (the fraction of relevant items within the top N). F-score is a popular formula for combining recall and precision (either equally as F1 Score or weighting one more than the other); He et al. 2010 compare precision, recall, and F-score for translation recommendation and suggest that there are tradeoffs between precision and recall, where some users may prefer a different balance between the two. MRR is another common metric similar to NDCG, but more quickly discounts the reward by rank and only considers the first relevant recommendation. Tan et al. (2018) compare thirteen methods for quote recommendation using mean reciprocal rank (MRR), NDCG@*5*, and Recall@*{1, 3, 5, 10, 20, 30}*, and these metrics usually but not always correlate. In this thesis, I refer to accuracy metrics as utility metrics, which is a more broad designation that also includes measures such as click-through rate (see 2.2.3 and 2.2.4).

2.2.2 Selected Metrics for Scripture Recommendation

For scripture recommendation, I consider two main aspects of recommendations to evaluate: utility and diversity. In this section, I explain my motivations in selecting each metric and define

explicitly how they are calculated. As the concept of "top k" applies to both aspects, I will discuss this first.

2.2.2.1 Top *k* Recommendations

Throughout this work, *k* or *@k* refers to the number of top recommendations considered for a given metric. Some previous research use the letter "N" (Rosati 2016) and others "k" (Tan 2018). In previous work I've considered, 5 is the most common value of *k*. However, comparing results at various values of *k* can provide insight into tradeoffs among methods or whether recommendations are still effective after the top few.

For serendipitous scripture recommendations, many verses could be relevant for a given query yet were not cited in the test document. This encourages using a larger value of *k*, even 100 or 1000, to take more information into account for each method's recommendations (especially for NDCG, which discounts by rank within the top *k*). However, in production, users mostly consider the top few recommendations, and perhaps the top 20 for a more in-depth search.

Consequently, I evaluate metrics *@5* to follow the standard and *@20* to compromise between encouraging other valid recommendations not in the test set and ensuring that recommendations are effective in a production system. Furthermore, NDCG lends itself to a comparison of greater values of *k*.

2.2.2.2 Utility Metrics

Utility refers to whether a scripture verse is relevant enough to be useful to the user (in relation to the query), and can be measured automatically by whether the system accurately predicts the verses that were cited near the query in the test set. Utility metric equations in this section refer to the score for a specific query or sample, which is then averaged over all test samples.

2.2.2.2.1 MRR (Mean Reciprocal Rank)

Of the metrics used in this study, only MRR is not dependent on *k*, or a certain number of recommendations. Instead, MRR uses the rank of just the first relevant item recommended.

MRR = 1 / rank of first correct recommendation

2.2.2.2.2 NDCG (Normalized Discounted Cumulative Gain)

NDCG is the only metric in this study that considers the relevance scores (both predicted and ideal), and it does so while also taking into account (discounting) the rank of each relevant recommendation. Hence, I consider NDCG to be the most reliable utility metric used in this study. Also, it is expected that leveraging sample weights (ideal relevance scores) during training particularly improves NDCG scores.

$$\text{DCG} = \sum_{rank=1}^{k} \frac{2^{relevance-1}}{\log_2(rank+1)}$$

```
NDCG = predicted DCG / ideal DCG
```

2.2.2.2.3 Precision, Recall, F-score

Precision, Recall, and F-score allow measuring more directly (or simply) whether recommendations are relevant at all, regardless of weight. Precision as used here is also known as mean average precision (MAP). These three metrics answer the following questions:

- Precision: Are recommendations usually relevant?
    - `precision@k = number correct@k / k`
- Recall: Are the relevant items for each sample recommended? (takes into account the number of correct recommendations for each sample)
    - `recall@k = number correct@k / total possible correct`
- F-score: How effective are recommendations, considering both precision and recall?
    - `F@k = precision * recall / (precision + recall)`

For this study, F-score refers to F1 Score, as precision and recall are weighted equally.

2.2.2.3 Diversity Metrics

Diversity was described in 2.2.1.2 and for this task refers to the variety of verses that are recommended. In this work, I consider all verses to be equally diverse (see 6.2.6 for possible distinctions). Unless recommendations are random, diversity is a strong indicator that the recommendation system appropriately considers many options instead of fixating on a small subset of items with certain qualities (like many training citations, many words, or few words).

Diversity can be measured based on certain features of items or verses (topic, book), but for now I simply measure the number of unique verses.

For diversity metrics, "#" means that the raw number of items is used for the metric, instead of a percentage.

2.2.2.3.1 Coverage

I measure coverage as catalog coverage, or the number of unique verses in the top k recommendations for all samples. This requires keeping track of the set of all top recommended verses.

```
coverage@k = |unique verses@k| / |unique verses|
#coverage@k = |unique verses@k|
```

2.2.2.3.1 Precision Coverage

I propose the name Precision Coverage (PCoverage) to describe the combination of precision (the percentage of correct recommendations), with coverage, resulting in the number of unique verses that were accurately recommended in the test set. PCoverage is a more reliable metric because diverse recommendations are only desirable if they are also useful.

```
pcoverage@k = |unique correct verses@k| / |unique verses|
#pcoverage@k = |unique correct verses@k|
```

2.2.3 Human Evaluation

As Ge et al. (2010) and Bean (2016) discuss, “high-quality recommendations need to be fitting their intended purpose and the actors behind this purpose are the ultimate judges of the quality of recommendations.” Examples of metrics based on human input are click-through rate, link-through rate, and cite-through rate (Bean 2016), which seek to answer the basic question, “Did users use the recommendations?” A user study can evaluate whether recommendations are used for their intended purpose, and such a study provides an opportunity for users to give direct feedback or comments on the recommender system.

Lenhart et al. (2016) use Google Analytics to analyze user click behavior on a large scale in addition to a questionnaire. In their system, users could click to view an item or to remove it

from the recommendation list. The questionnaire asked two questions (exact wording not provided):

1. How well did the recommendation system fit their interests?
2. How diversified were the recommendations?

Note that the entire recommendation display or system was evaluated, not individual recommendations, and the user may have used the recommendation system multiple times before answering the questions. Users rate their answers from 1 to 5. Lenhart et al. also include in the questionnaire the System Usability Scale (SUS) by Brooke et al. (1996), containing 10 questions with answers ranging from strongly agree to strongly disagree (Likert scale).

Vagliano et al. (2016) use a website to collect human evaluations that compare four algorithms, and they include the practical details involved in such a comparison. First, the user chooses a movie from a randomly generated list of movies; hence, the recommendations are based on the user's choice. For the chosen movie, 20 recommended movies were displayed in random order, such that the 20 movies represented the top 10 recommendations from each of the four algorithms after removing duplicates and keeping only the top 20 from the merged list. Vagliano et al. sought to answer two research questions:

1. Which of the considered algorithms provides the highest number of novel recommendations?
    - Associated question: Did you already know this recommendation?
    - Statistical metric: Total number of movies not known / Total number of user evaluations
2. Which of the considered algorithms is more accurate?
    - Associated question: Is it related to the movie you have chosen?
    - Statistical metric: Root Mean Squared Error (RMSE)

Participant responses could range from strongly agree to strongly disagree. The users could perform the evaluation as many times as they wanted, choosing a different movie each time, and the two questions must be answered for each of the 20 recommended movies. Note that, for rating novelty, it's important to ask about each individual recommendation, and Vagliano et al. suggest that recommending some items that are already known and liked by the user along with

novel recommendations may improve trust in the recommendation system, making the user more likely to try the novel recommendations.

2.2.4 Summary (Evaluation)

While studies on recommender systems have differed in the details of metrics and names for these metrics, a few key principles emerge. In general, or preferably, a recommendation system should be evaluated on the following:

1. The utility of top recommendations (e.g. click through rate, NDCG)
2. The diversity of top recommendations (e.g. coverage, novelty)

Evaluation strategies such as serendipity and unexpectedness can help to provide insight into these two aspects of a recommendation system and how they interact. Also, note that utility focuses on individual recommendations, whereas diversity directly measures the recommendation list as a whole.

Furthermore, in cases where recommendations are for human users (most cases), human evaluation may prove vital to understand why certain recommendations are more effective.

2.3 Summary

This survey of previous research led to the choice of methods described in Chapter 4 and to the ideas outlined in Chapter 6. In addition, it is evident that relevant research in NLP, machine learning, and recommendation systems continues to evolve rapidly, as shown by the last couple years of research.

Chapter 3. DATA

In this chapter I describe and list statistics, with commentary, on the data of contexts surrounding citations and the content to be recommended (scripture verses) along with how citations or references are distributed among the data. For information on how the data was processed or prepared, see 4.1. I present data statements in accordance with some of the recommendations by Bender et al. (2018) to promote understanding of the advantages, biases, and limits inherent in each data source. These data statements are intended as an overview from a linguistics standpoint and most information comes from my own working knowledge of the data.

3.1 Contexts

Contexts are text segments (e.g. paragraphs, headings) within documents that contain citations or references to verses of scripture. Contexts that are near the citation are weighted higher (see 4.2).

3.1.1 Data Statement - Contextual Data

Two data sources are used to obtain citations in context: General Conference talks and a book series entitled Teachings of Presidents of the Church. Both were provided by The Church of Jesus Christ of Latter-day Saints. These two sources are similar with notable differences.

3.1.1.1 Curation Rationale

My goal in collecting contextual data was to find a large number of scriptural citations that could be identified in a consistent way in context. For both data sources, scripture citations are identified in a standard HTML form. Furthermore, each document is structured in a similar way in HTML (though for General Conference talks, only some of the more recent talks are divided into sections). I also chose these data sources because I am personally familiar with them. As these data sources are used with permission, the Church requested I include the following statement:

> The Product offered by Joshua Mathias is neither made, provided, approved, nor endorsed by, Intellectual Reserve, Inc. or The Church of Jesus Christ of Latter-day Saints. Any content or opinions expressed, implied, or included in or with the Product offered by Joshua Mathias are solely those of Joshua Mathias and not those of Intellectual Reserve, Inc. or The Church of Jesus Christ of Latter-day Saints.

3.1.1.2 General Conference

Example from corpus:

> `Act now, so that a thousand years from now, when you look back at this moment, you can say this was a moment that mattered—this was a day of determination.`

General Conference talks are religious speeches (also known as talks), about 5 to 20 minutes in length, spoken live at a broadcasted worldwide meeting held every six months by The Church of Jesus Christ of Latter-day Saints in Salt Lake City, Utah (Church Newsroom 2011). The talks used for this thesis (those available online) include General Conference sessions from 1971 to 2018. Speeches are written and spoken in English (en-US) originally, except for a few rare cases, and then translated into 93 other languages (Church Newsroom 2011). Transcripts for each speech are later edited and published, usually with only minor changes, and this is the form used for this thesis.

The authors or speakers of each talk are native of many countries and languages, though most authors have lived in the United States and spoke English natively. Each General Conference (consisting of five two-hours sessions) includes about 30 talks; at the last conference, there were 34 talks and 28 unique speakers. 15 speakers (called Apostles) are the same each conference until they pass away and are replaced; the other speakers change each conference and are selected from among hundreds of other general church leaders. Most speakers are male; 4 speakers were women at the last conference.

Talks discuss religious topics (relating to God or the Church). In discussing these topics, speakers often include personal narratives; consequently, the language includes many terms unrelated to religion or the scriptures. As all speakers speak on behalf of the same Church, only

one religion is represented, which means interpretations of scripture may be particular to this religion.

3.1.1.3 Teachings of Presidents

Example from corpus:

> We learn from these scriptures from which I have read and from other revelations from the Lord, that man is the most important of all our Father's creations. In the same vision given to Moses, the Father said: "And as one earth shall pass away, and the heavens thereof, even so shall another come; and there is no end to my works, neither to my words. For behold, this is my work and my glory—to bring to pass the immortality and eternal life of man." [Moses 1:38–39]

The book series Teachings of Presidents of the Church was created for religious instruction by the Church. There are 15 books in the series, one for each of the first 15 presidents of the Church, from 1830 to 2008. Most of the content of these books consists of selections from speeches authored by the president, but it also includes stories from the president's life, which are written by the books' editors.

The content of Teachings of Presidents is similar to that of General Conference and many of the quotations in these books come from General Conference. However, while the General Conference corpus covers the years 1871 to 2018, the Teachings of Presidents book series includes language from an earlier time period (starting at about 1830). Each of the presidents or speakers in this book series spoke English as their native language and were born in the United States (except for one, who was born in England). Translations exist for Teachings of Presidents, but not for as many languages as for General Conference.

3.1.2 Statistics

- Number of General Conference (GC) talks from 1971 to 2018: 3,771
    - Number of talks that contain at least one Bible citation: 2,843 or 75.36%
- Chapters in Teachings of Presidents (ToP): 435
    - Number of chapters that contain at least one Bible citation: 406 or 93.33%
- Documents with more than section

- GC: 343 talks
- ToP: 405 chapters

Table 1: Contextual Data Counts

Corpus	Contexts	Words	Vocabulary	Characters
General Conference (GC)	132,278	8,320,430	178,682	31,987,386
Teachings of Presidents (ToP)	27,793	1,669,566	49,514	6,335,143
Totals	160,071	9,989,996	180,799	38,322,529

Number of context segments in each data source. For these stats, words are tokenized but not stemmed, and stop words are included.

3.2 Content

The content is the texts to be recommended, in this case individual verses from one of five volumes of scripture. All scriptures are cited in the same format in the contextual data.

3.2.1 Data Statement - Scripture

I will provide an overview of the linguistic origin of each volume of scripture, based on my own knowledge and introductory information found in each volume. In describing the origin of each volume of scripture, I seek to follow the claims of the original publishers of the work. Any years or facts about the origin of these historic works are intentionally broad and are not authoritative.

3.2.1.1 Curation Rationale

My goal in collecting scriptural texts was to have a source of text passages to recommend that have already been recommended in many contexts and that could be relevant in new contexts or to new queries. Almost by definition, scriptural or sacred texts are used repeatedly in many contexts and are applied in different ways.

Other books of scripture satisfy these requirements, but I chose to use scriptures I'm familiar with and for which I could easily identify citations in context (see 3.1).

The texts for all scriptures were obtained from scriptures.nephi.org and are in the public domain (LDS Scriptures 2018). The volumes of scripture included are known as the Standard Works and are considered canon by The Church of Jesus Christ of Latter-day Saints. They are

also published by this Church, though the King James Bible has had many publishers. The content of each volume of scripture is that which was included in the versions published by this Church in 1979.

3.2.1.2 The Bible

The Bible consists of the Old and New Testaments, which I consider as separate volumes. Different versions or translations of the Bible may have differences in what content or books are included, but I use the King James version of the Bible, which was published in 1611 (this is the version used by The Church of Jesus Christ of Latter-day Saints). There are likely minor or typographical edits in the copy I obtained online.

The King James Bible was written mostly in Hebrew and Greek and translated into early modern English (early 1600s) by multiple translators in England, by command of King James. The Bible has been translated into over one hundred languages (Christodoulopoulos 2015).

While the original works of the Bible were written as described in the following sections, the Old and New Testaments were compiled centuries after the original works were written.

3.2.1.2.1 The Old Testament

Example from corpus:

> *Joshua 1:8*—This book of the law shall not depart out of thy mouth; but thou shalt meditate therein day and night, that thou mayest observe to do according to all that is written therein: for then thou shalt make thy way prosperous, and then thou shalt have good success.

The Old Testament, also known as the Hebrew Bible (as in Coeckelbergs et al. 2018), was originally written in Hebrew and Aramaic (Wikipedia 2019a). The Old Testament was written mostly by leaders of the Israelites (many known as prophets, such as Moses) near the Mediterranean Sea or what is now known as the Middle East (particularly Israel or Palestine). It contains stories, genealogies, commandments or laws, and counsel. The Old Testament was written within the years 2000 BC and 1 BC (approximately).

3.2.1.2.2 The New Testament

Example from corpus:

> *1 John 3:18*—My little children, let us not love in word, neither in tongue; but in deed and in truth.

The New Testament was originally written mostly in Greek (Wikipedia 2019a) by various authors in the same area as the Old Testament and elsewhere within the Roman Empire. The books of the New Testament were written between 0 AD and 200 AD (approximately) and mostly consist of four accounts of the life and teachings of Jesus Christ plus letters (epistles) from Church leaders (e.g. Paul).

3.2.1.3 The Book of Mormon

Example from corpus:

> *1 Nephi 19:23*—And I did read many things unto them which were written in the books of Moses; but that I might more fully persuade them to believe in the Lord their Redeemer I did read unto them that which was written by the prophet Isaiah; for I did liken all scriptures unto us, that it might be for our profit and learning.

Like the Old Testament, the Book of Mormon was written originally by multiple authors over a long period of time. The first author came from the Jerusalem spoken of in the Old Testament around 600 BC, moved to the American continent, and his descendants continued to author the volume until before 500 AD. The second to last author (Mormon) compiled most of the Book of Mormon. According to this author, it was written in a reformed Egyptian, though the authors spoke a form of Hebrew. As the authors were separated from other Egyptian and Hebrew speakers, over time their language became their own.

The version of the Book of Mormon in this project was translated into English by Joseph Smith, with the help of scribes, and published in 1830 with minor edits since that time. The English of the Book of Mormon was written in a similar style as the King James Bible and also influenced by the English of Joseph Smith's time in the United States (New York). The Book of Mormon contains portions of the Bible with some variations in wording.

3.2.1.3 The Doctrine and Covenants

> *D&C 88:118*—And as all have not faith, seek ye diligently and teach one another words of wisdom; yea, seek ye out of the best books words of wisdom; seek learning, even by study and also by faith.

The Doctrine and Covenants is a collection of revelations, mostly dictated or written by Joseph Smith in English in the 1830s and 1840s. Unlike the Bible, the Doctrine and Covenants is not separated into books but contains 138 sections (and 2 declarations), which are similar to chapters. A few sections and the two declarations were added by later Church leaders.

3.2.1.4 The Pearl of Great Price

Example from corpus:

> *Joseph Smith–History 1:34*—He said there was a book deposited, written upon gold plates, giving an account of the former inhabitants of this continent, and the source from whence they sprang. He also said that the fulness of the everlasting Gospel was contained in it, as delivered by the Savior to the ancient inhabitants;

The Pearl of Great Price is a relatively small volume of scripture that consists of four types of content:

1. Translations of the books Abraham and Moses by Joseph Smith into English. These are written in a similar style as the Old Testament.
2. An alternative translation of Matthew 23:39 and Matthew 24 of the New Testament, translated by Joseph Smith into English.
3. A religious history written by Joseph Smith in English.
4. The Articles of Faith, a statement by Joseph Smith published in 1842.

3.2.2 Statistics

Table 1: Volume Counts

Volume Counts	Verses	Chapters	Books	Words	Vocabulary	Characters
Old Testament	23,145	929	39	706,730	10,553	2,579,790
New Testament	7,957	260	27	210,396	5,975	769,101
Book of Mormon	6,604	239	15	303,053	5,563	1,146,400
Doctrine and Covenants	3,654	138	1	129,125	4,633	487,478
Pearl of Great Price	635	16	5	30,661	2,414	112,099
Totals	41,995	1582	87	1,379,965	15,001	5,094,868

Number of segments and distinct words in each volume of scripture. Here, words are tokenized but not stemmed, and stop words are included.

The Old Testament alone contains over half of all verses and words in this set of scripture.

3.3 Citations

Citations include footnotes cited in the text as well as the list of related scriptures at the end of every chapter of Teachings of Presidents of the Church. A given citation may be for multiple verses. Where a citation spans multiple chapters (only a few cases), currently only verses in the first chapter are included (this could be made more robust in the future).

3.3.1 Citation Statistics

Here I show statistics on citations for each volume. The following list describes each column in Table 3:

- Citations: The number of citations for verses that are included in the volume.
- Verse Citations: The number of citations x the number of verses in each citation
- Context-Verse Pairs: The number of verse citations x the number of contexts in the citing document.
- Verses per Citation: The average number of verses cited in a scripture citation.

- Verses cited: The number of unique verses that are cited at least once.
- Chapters cited: The number of unique chapters that are cited at least once.
- All books are cited at least once, so this column is not shown.

Table 2: Citations by Volume

Volume	Citations	Verse Citations	Context Verse Pairs	Verses per Citation	Verses Cited	Chapters Cited
Old Testament	4,052	12,282	402,046	2.81	0.18	0.60
New Testament	11,712	28,645	968,057	2.34	0.64	0.98
Book of Mormon	10,233	25,556	802,947	2.57	0.64	0.95
Doctrine and Covenants	9,630	28,672	1,077,898	3.08	0.89	0.94
Pearl of Great Price	1,940	4,768	188,193	2.64	0.92	1
Totals	37,567	99,923	3,439,141	2.66	0.41	0.75

The first three columns are counts within each volume, and the last three columns are percentages. Totals are listed in the last row.

Note that while only about 26% of Bible verses are in the New Testament, about 70% of Biblical citations are of verses in the New Testament. The most cited verse in the Bible is Matthew 22:37 with 123 citations. See Appendix C for the most cited verses by volume.

3.3.2 Training, Validation, and Test Sets

As experiments with weighting paragraphs will require keeping paragraphs of documents together, I separate the data by documents instead of by citations, using a .7/.15/.15 split. However, this is done such that no more documents are added to a set once it has reached its portion of context-verse pairs. For the training, validation, and test sets I include citations to all 5 volumes of scripture in the data. Here is the resulting split for the number of context-verse pairs in each set (since some documents have more or fewer citations):

Total number of context citation pairs: 1,464,194 (each may cite multiple verses)

Table 3: Context-Verse Pairs by Dataset

	Training	Validation	Test	Total
Context-verse pairs	3,439,141	847,431	735,450	5,022,022

Context-verse pairs correspond to samples. In each pair, the context is considered the query (input) and the verse is considered the recommendation.

Only documents with citations that are placed in the training set are used to train each method. There are 105,496 contexts in the training set, which is 65.93% of 160,021 total.

3.4 Summary

General Conference talks and Teachings of Presidents of the Church provide a reasonably large set of scripture citations, multiplied by the number of text segments in each document and verses in each citation. However, this data source, while convenient and structured, does not cover all relevant domains or interpretations of verses. In addition, only 37% of potential recommendations (verses) are cited even once.

Chapter 4. METHODS

This chapter covers how data was prepared for use, context selection and weighting, and the implemented methods and baselines.

4.1 Preprocessing

As in Tan et al. (2018), before encoding text into TF-IDF Bag-of-Words vectors, both contexts and scripture verses are preprocessed by removing stopwords (using NLTK's stopwords module) and stemming with the Porter Stemmer (Porter 1980; also using NLTK). The following strings are added to the stop words list: 'com', 'edu', 'org', 'net', ':'. Tokenization is done via nltk.sent_tokenize and nltk.word_tokenize, with the addition that all periods and different types of dashes ('—','–','-') are separated from words as distinct tokens. For sentence tokenization (sent_tokenize), NLTK uses its PunktSentenceTokenizer module, which uses "an unsupervised algorithm to build a model for abbreviation words, collocations, and words that start sentences" (Bird 2009). Future work for scripture recommendation could experiment with training the sentence tokenization on the scriptural texts to be recommended, as well as queries or contexts for a particular domain. For word tokenization (word_tokenize), NLTK uses its TreebankWordTokenizer module, which uses regular expressions to separate punctuation that doesn't form part of a word. I remove punctuation for the bag-of-words methods.

All words are lowercased. Lowercasing allows for a more condense vocabulary and sharing of data between instances with different casing ("olive" and "Olive"). Especially in the case of scriptural texts, the in-domain text available (scriptures) is relatively small, making a more condensed vocabulary more important; lowercasing achieves this while still allowing queries for rare terms. However, some significant distinctions are lost by lowercasing, such as "Job," a person and book in the Bible, and "job," a task to complete (which is a term found in queries but not scripture included in this study). As discussed in 2.1.10, both infrequent and frequent words are retained, because I seek to fit the recommendation model to a specific set of

books (terms rare in the Bible may be common in queries). For this reason, numbers and references to verses such as 23:2 are also retained.

No preprocessing is performed before embedding texts using Google's Universal Sentence Encoder module, because this module handles preprocessing (Google n.d.). The authors do not specify what preprocessing is performed.

4.2 Context Selection

Each document (General Conference talk or chapter in Teachings of Presidents of the Church) was obtained from lds.org (now ChurchofJesusChrist.org) in HTML form. To take full advantage of the citations in context contained in this data, each paragraph, header, and title in the document are used as distinct contexts (or queries during testing) for a citation. Knowledge of the structure of the document (such as section, headers, and distance from the citation) is used to weight each context (sample weight or context-verse weight) when training the recommendation system. The weighting is normalized such that each citation has in total the same weight among its contexts. The different weighting methods are discussed in the following section.

4.2.1 Calculation of sample weights by distance

Here, a "section" refers to a portion of a General Conference talk or chapter in Teachings of Presidents of the Church. Sections represent subtopics or whatever the author desired to assign to a section. Each section begins with a header. Formally, document parts are weighted as follows:

```
For each citation:
      If context is within the same section as the citation:
            Calculate distance (number of paragraphs) from the citation (the
                  same paragraph has distance 0)
      total_section_weight = .5 (assign half of all weight for the document to
            the same section as the citation)
      total_other_weight = .5 (assign other half to all other contexts in the
            document, or 1 if there are no sections)
      total_section_distance = sum of (context distance + 1) for each section
            context
      For each context in the same section:
            Context's sample_weight = context's distance /
                  total_section_distance * total_section_weight
```

```
For each other context:
      Context's sample_weight = 1 / |other contexts| *
            total_other_weight (|_|denotes number of items)
```

In addition, since many citations are of multiple verses (as a range), the sample weight is divided by the number of verses in the citation, following the intuition that a citation for a single verse is more representative of that verse than a citation of multiple verses or the whole chapter.

In summary, the sample weights of all contexts for a particular citation sum to 1, greater weight is assigned to contexts in the same section, and weighting is biased within that section by distance (the number of contexts (paragraphs) between the citation and the context).

4.2.2 Special cases

Each section begins with a header, which is included in its section. The document title and the opening quote (labeled as a "kicker" or "intro") are a separate section.

Each chapter in Teachings of Presidents contains a list of related scriptures. For each Bible verse listed, each context in that chapter is given a sample weight of `1 / |contexts in chapter |`.

Note that contexts will be included as a sample once for each citation in its containing document, biasing the data toward contexts near many citations. This is a desirable effect.

4.2.3 Motivation

This particular weighting scheme follows the principle of maximizing uncertainty in the sense that if we have no other reason to do otherwise, we divide weight evenly among all possibilities. For this reason, the weight for a section (`total_section_weight`) and the weight for the rest of the document (`total_other_weight`) is divided evenly (.5), as I know that contexts within a section are more likely relevant to a citation, but I do not know to what degree (in the future these numbers could be optimized). As there are many more contexts in other sections than in one particular section, contexts in the same section are given a greater weight than other contexts. Then, within all contexts not in the same section, the weight is divided evenly, without further information as to whether they are relevant to the citation. For contexts within the same section, greater weight is given to contexts that are closer to the citation, as

within the same section proximity is a more trustworthy indicator that a context is related to the citation (as a section represents a specific subtopic or train of thought). However, I do not know to what degree proximity is a factor, so the weight difference between consecutive distances is the same; this is acheived by summing all context distances from the citation within the section (`total_section_distance`) and then dividing by this sum (and multiplying by `total_section_weight`). The same principle applies to dividing weight by the number of verses in the citation. As I do not know which verses are more relevant to the citation's context, all verses receive equal weight.

Furthermore, weight for a citation is divided carefully such that the weights for all context-verse pairs sum to 1. As the goal of the recommendation system is to recommend citations, the total weight should be based on the citation. For example, I wouldn't want to bias the recommendation system toward citations that recommend many verses or that are found within a document containing many contexts. A convenient result of this strategy is that for a particular verse, I can sum its associated weight for all of its contexts to obtain the number of times a verse was cited, taking into account whether the citation included other verses (resulting in a citation count that may not be an integer).

This weighting scheme is novel and customized to this particular task, but my hope is that it provides an generalizable example of how document structure can be used to weight contexts in relation to citations, for citation recommendation (or even to represent a paragraph by other paragraphs in the document). Document structure such as chapters, sections, and headers are common in literature.

4.2.4 Summary

These measures provide a much larger number of samples per citation while minimizing irrelevant bias (such as favoring verses in longer documents or with more citations). In the following formula describing the total number of data samples (to be split into train, validation, test sets), "x" means scalar multiplication:

```
Total number of samples =
      the number of citations x
      the number of verses in each citation x
      the number of contexts in the containing document
```

4.3 Calculating Relevance

As in Tan et al. (2018), the focus of each method I experiment with is to represent each context and verse as vectors such that I can compute the relevance of each verse to a given context as the dot product. Since each context and verse is first normalized using L2 normalization, taking the dot product results in the cosine similarity.

Using the same syntax as Tan et al. (2018), variables are defined as follows:

- c_* is a query context.
- q is a verse (quote).
- Q is the set of all verses to be recommended.
- v_{c*} v_{c*} is a context vector query.
- v_q is a verse vector.
- V_Q is a matrix of size $|Q|x|v_q|$, where each row represents a verse.
- $\hat{v}$ is an L2 normalized vector.
- $\cdot$ (small dot) signifies multiplication where at least one operand is a scalar.
- $\odot$ signifies element-wise multiplication (dot/inner product).
- r_{c*} is a vector containing the relevance scores between c_* and each verse.
- $r_{c*} = \hat{v}_{c*} \odot \hat{V}_Q^T$
- `Best verse = argmax(`r_{c*}`)`
- In practice, r_{c*} and $\hat{v}_{c*}$ are matrices with batch size rows.

Hence, relevances are calculated for a given set of contexts C_* as a single matrix multiplication of $\hat{V}_{c*}$ with the transpose of $\hat{V}_Q$, resulting in a relevance score (or similarity) between 0 and 1. Finally, verses are ranked according to their relevance scores.

4.4 Weighted Context Lumping (WC-QRM)

As described in 2.1.3, the context lumping method (Ahn 2016) concatenates the contexts of a particular verse (as words) before representing the lumped context as a TF-IDF vector. This is analogous to the Context-based Quote Recommendation Model (C-QRM), as the verse "is represented by all the words of its contexts in the training data" (Tan 2018). In C-QRM, however, each context for a verse is converted into a TF-IDF vector individually, normalized (L2), and summed together. The sum is then L2 normalized.

Formally,

- C_q is the set of training contexts for a verse.
- $v_C q$ is a verse vector in the context space.

$$V_C q = \sum_{c \in C_q} \hat{v}_{Cc}$$

- For each q in Q,
- $r_{c*} = \hat{v}_{Cc_*} \odot \hat{V}_{CQ}$

Each context for a verse has an associated weight representing how relevant the context and verse are to each other (see 4.2). To incorporate this information, I multiply the weight for each context by its respective context vector during the summation. Using L1 normalization on all weights for a verse showed no effect on results on the validation set, so I left the weights as is (normalized by citation). Normalizing the weights by verse would remove information as to whether a particular verse was well represented by the training data (whether there are many contexts with high weights).

I define weighted context lumping or WC-QRM as follows:

- w_c is the weight for a particular context
- `For each q in Q,` $v_{Cq} = \sum c \in C_q \hat{v}_c w_c$
- $r_{c*} = \hat{v}_{Cc_*} \odot \hat{V}_{CQ}$

4.5 Verse to Context Transformation (TWC-QRM)

The context-based approach described in 4.4 only provides a representation for verses that have been cited in the training data (and thus have contexts). In this section I extend this representation to all verses by comparing their content. Intuitively, I can represent an uncited verse by cited verses which are textually similar. Formally,

- Q_T is the set of all cited (trained) verses.
- Q_N is the set of all uncited (novel) verses.
- v_{Qq} is a verse vector in the verse content space (as opposed to v_{Cq}).
- $v_{Cq_N} = \sum_{q_T \in Q_T} ((\hat{v}_{Qq_N} \odot \hat{v}_{Qq_T}) \cdot \hat{v}_{Cq_N})$
    - $\hat{v}_{Qq_N} \odot \hat{v}_{Qq_T}$ is a scalar representing the similarity between a novel verse v_{Qq_N} and a trained verse v_{Qq_T}.

However, I can achieve this same effect by creating a single transformation matrix to map features or words in verses to features in contexts. I do this by multiplying the verse vectors in content space (V_{QQ}) by each of the verse vectors in context space (V_{QQ}). Formally,

- W_Q is the set of all features (e.g. vocabulary) for the content of a verse or quote.
- W_C is the set of all features for a context.
- $T_{QxC} = \hat{V}_Q^T \odot \hat{V}_C$.
- T_{QxC} is a verse to context transformation matrix of size $|W_Q| x |W_C|$.
- $V_{CQ} = \hat{V}_{QQ} \odot \hat{T}_{QxC}$
- $r_{c_*} = \hat{v}_{Cc_*} \odot \hat{V}_{CQ}$

In this paper, I apply this transformation to all verses, instead of just uncited verses (but evaluate recommendation of uncited verses separately). A content-to-context transformation matrix enables recommendation of novel passages and even to reverse the recommendation (using the transpose of T_{QxC}), in this case recommending paragraphs given a scripture verse.

4.6 Text Representations

The methods described in previous sections can be applied to a variety of word and paragraph representation methods. In the following section I describe two text representation methods, to which I apply C-QRM, WC-QRM, Verse to Context Transformation, and Direct Query-to-Verse Similarity (described later in 4.7.3).

4.6.1 TF-IDF Bag-of-Words

The TF-IDF bag-of-words representation is well researched and has successfully been used for quote recommendation (Ahn 2016; Tan 2018). The C-QRM method was originally used with TF-IDF bag-of-words vectors (Tan 2018). In the present study, I use Scikit-learn's TfidfVectorizer module (Scikit-learn 2011) to train TF-IDF vectors (the vector dimension is the vocabulary size) for verses and contexts separately, with distinct vocabularies, except in the case of the Direct baseline method (see 4.7.3). Stop words and punctuation are not included, and words are lowercased (see 4.1).

Note that while TF-IDF vectors are originally sparse after WC-QRM, the Verse to Context Transformation method produces dense vectors. This allows a verse to be recommended whose contexts do not directly contain words of the query context. However, this dense representation becomes unwieldy (high memory) with a large context vocabulary.

4.6.2 Universal Sentence Encoder (USE)

As with bag-of-words vectors, the same similarity calculations can be applied to paragraph-level embeddings of contexts and verses. I apply the various methods in this thesis to the output vectors of Google's pretrained Universal Sentence Encoder (Cer 2018). In the case of transforming verses to contexts, each embedding dimension is transformed instead of words. The embedding size is 512 and the Deep Averaging Network (DAN) version of the Universal Sentence Encoder is used (as a Tensorflow module) as the Transformer version requires a large amount of memory and time for long inputs.

USE was trained by Google on general text from the web; in particular, “Wikipedia, web news, web question-answer pages and discussion forums” (Cer 2018) as well as the Stanford Natural Language Inference corpus (Bowman 2015). I expect these data sources to represent well the modern language to be used as queries to the scripture recommendation system. Wikipedia (and some discussion pages) include religious concepts and terminology, but may not include much text from scripture or the Bible apart from commonly quoted verses. Consequently, USE may not be as prepared to model the early modern English used in the King James Bible, compared to modern English.

4.7 Baselines

The first two baselines, Random and Fixed Popularity, do not use any information about the query to recommend verses. The third baseline, Direct, does use the query but does not use training citations.

4.7.1 Random Recommendation

The random recommendation model generates a random score between 0 and 1 for each verse (or recommendation candidate). As with other methods, the recommendations are ordered according to their score. For the results presented in this thesis, the following random seeds are used: 1, 2, 3, 4, 5. Five random seeds are required because the test set is split and evaluated in five batches. For the qualitative examples, just the first random seed of (1) is used.

A random representation system likely represents the highest possible coverage or diversity without explicitly recommending items that weren't recommended in previous queries. However, I would expect accuracy or the usefulness of recommendations to be relatively low.

4.7.2 Fixed Popularity

The Fixed Popularity recommender always recommends the same verses, ordered by the number of citations in the training data. The number of citations is calculated by summing the weights of each verse’s contexts, since weights are normalized (sum to 1) by citation. Note that a verse’s contexts only receive partial weight for a citation of multiple verses (1 / number of verses in

citation). Relevance scores are calculated for each verse by sorting verses in descending order by number of citations and then assigning a score of 1 / index, where index is the verse's ranking in the sorted list.

Popularity-based recommendation stands in contrast to random recommendation in that coverage is as low as possible (always recommending the same items), but a higher accuracy is expected over all samples because the most popular items are the most likely to be relevant in a any given situation (without using any information about the query). If a popularity-based recommendation system performs relatively well, this implies that certain verses are recommended more than others (perhaps much more). For the Bible (Old and New Testaments), the following analysis shows that this is true (citation distribution is uneven among verses).

Average number of citations for each verse: 1.368 (including those that have no citations)
Unique verses cited: 9527 or 30.63% of all Bible verses
Verses with only one citation: 4105 or 43.09% of all cited verses in the Bible
Highest number of citations for a verse: 123 (Matthew 22:37)

4.7.3 Direct Query-to-Verse Similarity

The "Direct" baseline compares the representations of the verse's content (instead of contexts) with the query representation using cosine similarity. The Direct method makes no use of training citations.

4.8 Summary

This chapter described details for implementing a context-based recommendation system for scripture verses. Contexts for each citation are weighted according to section and proximity in relation to the citation, and all contexts in the document are included. C-QRM or context lumping is an intuitive way to represent a verse by its contexts, and C-QRM allows weighting a verse's contexts (WC-QRM). Furthermore, applying C-QRM to both TF-IDF and neural paragraph embeddings illustrates the versatility of this approach for a context-based representation. Throughout this thesis, the word “approach” refers to the method used to represent verses (Direct, C-QRM, WC-QRM, TWC-QRM), regardless of the method used to embed or encode text (TF-IDF BOW or USE). For each approach used to represent verses and

queries (or contexts), similarity can be calculated by simple matrix multiplication (assuming the verse and query representations are of the same dimension).

The random and popularity-based recommendation baselines represent recommendation systems focused solely on diversity (random) or solely on high accuracy over all samples (popularity) in a simple manner.

Chapter 5. RESULTS

In this chapter I present and discuss results on the test set, using both automated metrics and human evaluation.

5.1 Test Set

Citations for all 5 volumes of scripture are split by document, as described in 3.3.2. Each method described in Methods is evaluated on the test set. Tables 5 and 6 summarize the test set.

Table 4: Query Overview

Dataset	Queries	Recommended Verses
Total	160,021	5,022,022
Training	105,496 (65.93%)	3,439,141 (68.48%)
Test	20,446 (12.77%)	735,450 (14.64%)

Context segments are used as queries. Context-verse pairs are used as recommended verses.

Each query-verse pair (recommended verse) comes with a relevance weight, calculated according to the method defined in 4.2. This relevance weight is used in Normalized Discounted Cumulative Gain (NDCG) to calculate Ideal DCG.

Table 5: Verse Overview

Dataset	All	Cited	Uncited
Total	41,995	17,209	24,786
In Training	41,995	15,478 (36.86%)	26,517
In Testing	41,995	7,291 (17.36%)	34,704
Test Verses in Training	7,291	6,101 (83.66%)	1,191

The number of unique verses cited in each set and the number of unique verses that are cited in the test set but in not the training set.

The number of unique verses cited should be considered when looking at #Precision Coverage, or the number of unique verses that are correctly recommended. When all verses are considered,

the maximum #Precision Coverage is 7,291. When only uncited verses are considered, the maximum is 1,191. The number of cited verses does not affect Coverage, because it rewards unique verse recommendations regardless of whether they were recommended in the test set.

5.2 Metrics

See 2.2.2 for a discussion on each metric for scripture recommendation. Utility metrics and diversity metrics are considered separately. Each metric (and value of *k* for each metric) forms a column in the tables of scores.

5.3 Experiments

Two text representation methods are evaluated: TF-IDF and Universal Sentence Encoder (USE). For each text representation method, I compare the Direct baseline, C-QRM (without weighting), WC-QRM (with weighting), and TWC-QRM (with weighting and transformed). These three options are indicated as D, C, WC, and TWC respectively, and are separate rows in each table.

5.4 Cited and Uncited

Cited verses refer to verses that are cited in the training set and therefore have training contexts. The C-QRM and WC-QRM approaches provide no representation for verses with no contexts, but the Transformation approach (TWC-QRM) and the Direct approach do. Three separate categories of metric scores are provided for all verses, cited verses only, and uncited verses only. These scores are calculated by including or not including the relevant verses when scoring the test set.

5.5 Scores and Discussion

5.5.1 Overview

In each scores table, the best scores for each metric are bolded, and the best score that isn't one of the Random (Random) or Fixed Popularity (Popular) baselines is bolded and also

highlighted yellow (where one of these baselines has the highest score, there are two bolded scores).

Table 6: Method Acronyms

TF-IDF	Bag-of-words vector with Term Frequency - Inverse Document Frequency weightings
USE	Universal Sentence Encoder embedding
D	Direct Query-to-Verse Similarity
C	Context-Based Quote Representation Model (C-QRM)
WC	Weighted Context-Based Quote Representation Model (WC-QRM)
TWC	Transformed Weighted Context-Based Quote Representation Model (WC-QRM)

These acronyms for methods are used in tables of scores, and some are used in discussion.

Table 7: Metric Acronyms

@5, *@20*	Only the top 5 or 20 recommendations are considered
MRR	Mean Reciprocal Rank
NDCG	Normalized Discounted Cumulative Gain
Prec	Precision (Mean Average Precision)
Rec	Recall
F	F1 Score
Cov	Coverage
PCov	Precision Coverage
#Cov	Number for unique verses used to calculate Coverage
#PCov	Number for unique verses used to calculate Precision Coverage

These acronyms for metrics are used in tables of scores, and some are used in discussion.

All metrics with *k @5* and *@20* are shown in Appendix A, where scores are shown separately for all verses, cited verses (cited in the training data), and uncited verses. To provide an overview, the following table shows results for representative metrics, evaluated on all verses.

Table 8: Results on All Methods for All Test Citations

Method	MRR	NDCG@*20*	F@*20*	#PCov@*20*	#Cov@*20*
Random	0.005019	0.000240	0.000430	310	**41994**
Popular	**0.132500**	**0.032659**	**0.023706**	20	20
TF-IDF-D	0.004050	0.000236	0.000316	154	22957
TF-IDF-C	0.027770	0.006453	0.007000	1113	14844
TF-IDF-WC	0.048617	0.016486	**0.007746**	1349	15109
TF-IDF-TWC	0.013903	0.002060	0.001512	329	11887
USE-D	**0.068195**	**0.018188**	0.005061	**1468**	**34682**
USE-C	0.009324	0.001368	0.001736	418	12140
USE-WC	0.014136	0.002426	0.002018	589	13859
USE-TWC	0.002881	0.000162	0.000264	49	2278

MRR always assigns a score to each sample regardless of *@k*, NDCG represents a weighted utility metric, F combines both recall and precision, Precision Coverage represents the diversity of relevant recommendations, and Coverage represents raw diversity regardless of relevance.

The Fixed Popularity baseline outperforms other methods for all utility metrics on previously cited verses (see 5.5.3). This table shows that metrics do not always agree on which method performs best overall, but in general two methods outperform the others (not including Fixed Popularity): the Direct approach with USE embeddings and WC-QRM with TF-IDF Bag-of-Words vectors. As Precision Coverage combines both relevance and diversity, I consider it to be the most trustworthy metric among those included, but others provide additional insight. As Precision and Recall scores are similar, I only show F1 Score in this chapter, but scores for all metrics are provided in Appendix A.

5.5.2 Statistical Significance of Hypotheses

I present statistical significance of each of my hypotheses (see 1.3). I calculated the t-statistic and p-value using scipy's stats.ttest_ind module (Jones 2001). Both text representation methods are considered, TF-IDF BOW and USE (Universal Sentence Encoder). Coverage and Precision Coverage only have a single score over all test samples, so they are not considered for statistical significance.

Hypothesis 1: Weighting

Statement: Weighting contexts (using WC-QRM) by document structure and proximity will show improved scores for scripture recommendation compared to not using weightings.

The following table shows that this hypothesis is true for TF-IDF BOW and USE text representations. Weighted refers to the WC-QRM approach, and unweighted refers to the C-QRM approach. As NDCG uses weights for test samples as ideal relevance scores, and F-score does not, I show both NDCG*@20* and F*@20* (see 2.2.2.1 for a discussion on *@20*, or the number of top recommendations to consider). Results using *@5* are similar for these comparisons.

Table 9: Weighted vs. Unweighted - Statistical Significance with NDCG and F-score

	NDCG*@20*		**F*@20***	
Method	*p-value*	*t-statistic*	*p-value*	*t-statistic*
TF-IDF Weighted	3.41E-62	16.67	0.00677	2.71
USE Weighted	1.59E-07	5.24	0.03700	2.09

This table shows (for each text representation method) statistical significance of WC-QRM compared to C-QRM, where a positive t-statistic indicates a higher score for WC-QRM (weighted).

Weighting greatly improves scores for TF-IDF BOW, and to a lesser degree for USE. This may be because USE creates a more abstract representation of the text and has been trained on a large body of text which is independent of the weighting scheme, whereas TF-IDF is a simple weighting of contexts that lends itself to be combined with another weighting method.

Weighted compared to unweighted has a much greater t-statistic for NDCG*@20* than for F*@20*, which I suspect is because the test set weights used in NDCG to calculate ideal relevances were calculated the same way as the training set weights used for WC-QRM (see 4.4). As a verse with a higher weight for a context is more likely relevant (as the citation was closer), it's important to take these weights into account at test time. The F-score comparisons show that even without taking weights into account, the weighted approach recommends more relevant verses in the top 20 results.

Hypothesis 2: Transformation

Statement: Transformed verse representations (see 4.5) based on the training contexts of similar verses will result in scripture recommendation for uncited verses better than the following baselines that don't use contexts:

a. Random (see 4.7.1)
b. Fixed Popularity (see 4.7.2)
c. Direct Query-to-Verse Similarity (see 4.7.3)

As Table 10 shows (see also tables 15 and 16 in Appendix A), this hypothesis was not proven true for all baseline comparisons on uncited verses, but the Transformation approach using the TF-IDF BOW text representation method (as opposed to USE) outperformed the three baselines. However, this improvement was not statistically significant compared to the Random baseline (and borderline compared to the Direct approach). Here, statistical significance is evaluated using MRR, as some methods do not have any relevant recommendations of uncited verses within the top *k* recommendations. It should be noted that Random outperformed TF-IDF Transformed on Precision Coverage.

Table 10: Transformation vs. Baselines - Statistical Significance with MRR on Uncited Verses

	Random		**Popular**		**Direct**	
Method	*p-value*	*t-statistic*	*p-value*	*t-statistic*	*p-value*	*t-statistic*
TF-IDF Transformed	0.39451	0.85148	5.82E-12	6.88597	0.01860	2.35349
USE Transformed	0.00001	-4.35854	7.52E-109	22.23163	2.10E-20	-9.26211

Statistical significance with MRR scores. A positive t-statistic indicates a higher score for the Transformation approach, whereas a negative t-statistic indicates that the baseline method outperformed.

Note that relatively few verses cited in the test set were uncited in the training set (16%), and these comparisons don’t reward recommending cited verses.

It’s hard to make conclusions based on these results, but the failure of the Transformation approach is likely because there is too much noise in the verse to context transformation matrix. To mitigate this issue, one could include only the top words (e.g. top 5) for each context when creating the transformation matrix. Theoretically, this would represent verses by only their most

important topics according to contextual data, enabling salient topics to emerge. Including only the top context words could improve results for the WC-QRM approach as well.

To better visualize the success of each approach on uncited verses, Figure 1 presents Precision Coverage*@20* on uncited verses.

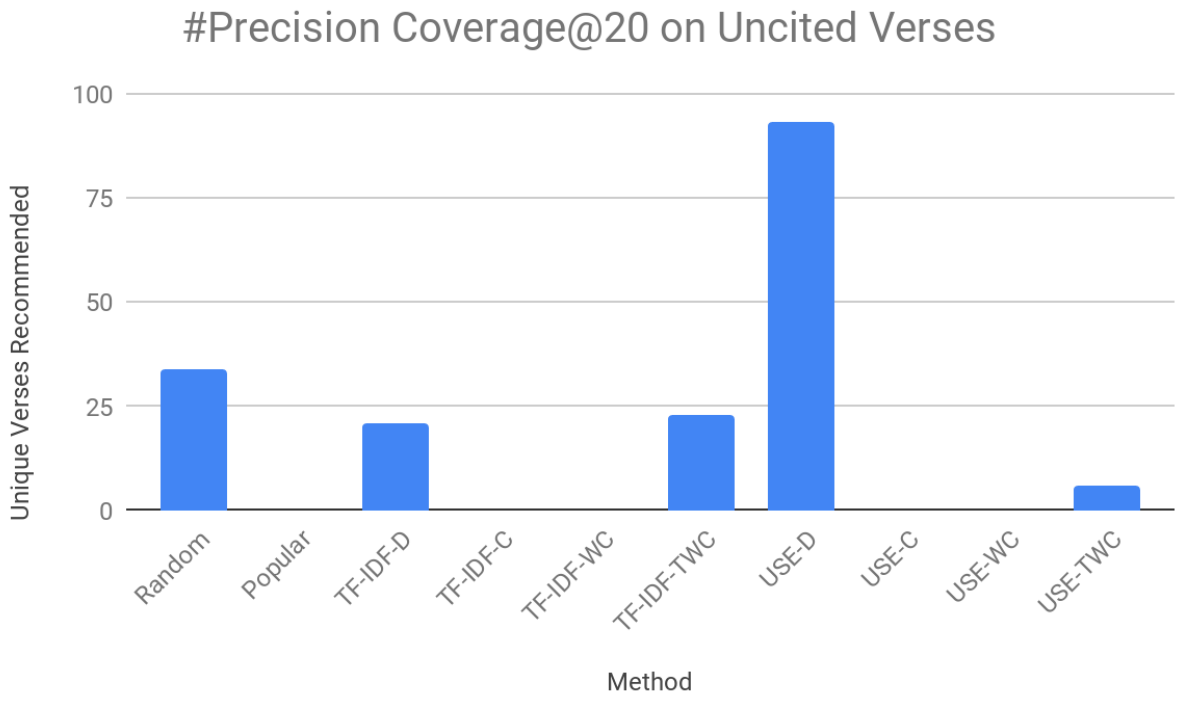

Figure 1: #Precision Coverage*@20* for all methods on verses cited in the test set that were not cited in the training set.

USE performed worse than TF-IDF BOW for the Transformation approach, which correlates with USE's worse results for WC-QRM as well (the WC-QRM vectors are used to create the Transformation matrix). For Precision Coverage on all verses, TF-IDF Transformed performs 4.10 times worse than TF-IDF WC-QRM, whereas USE Transformed performs 12.02 times worse than USE WC-QRM. This demonstrates that compared to a BOW approach, USE embeddings don't lend themselves to cross-domain transformations, as they were trained for direct comparisons between the original embeddings produced by the neural model. Fine tuning USE embeddings for scripture recommendation given text queries would likely improve results (as in Tan et al. (2018) using LSTM).

5.5.3 Random and Popular Baselines

As expected, Random Recommendation performed well (and the best overall) for Coverage, though not for Precision Coverage, nor any metric that uses utility or accuracy. Figure 2 illustrates the success of Random Recommendation for Coverage:

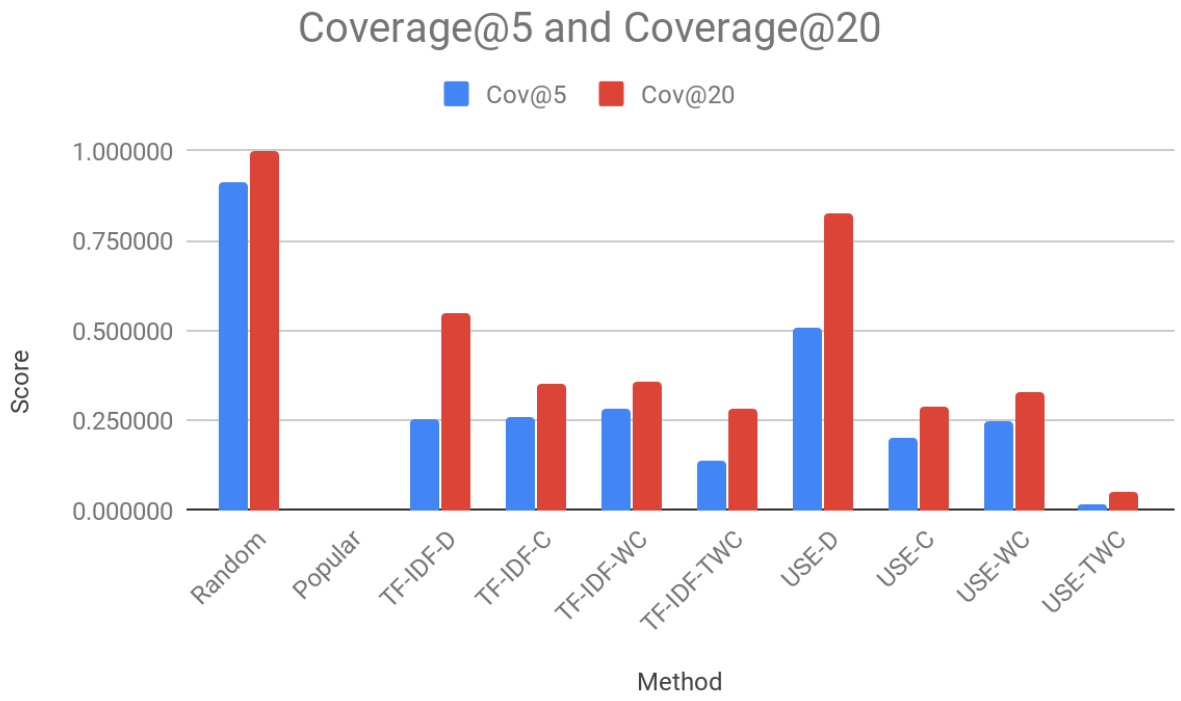

Figure 2: Coverage@*5* and Coverage@*20* for all methods on the test set.

Coverage is a useful and simple metric to ensure diversity in recommendation systems, and doesn't require a dataset of gold standard recommendations. In this study, the relative scores for each metric were similar for coverage as for NDCG except for the Popular and Random baselines. However, if possible, a utility metric should be compared (or a combination such as Precision Coverage) to take into account potential noise or elements of randomness in a recommendation system. Comparing utility also helps ensure that a recommendation system performs better than random.

The Fixed Popularity baseline performed best for all utility metrics except for uncited verses (the Fixed Popularity method never recommends uncited verses). Figure 3 shows NDCG@*5* and NDCG@*20* alongside Precision Coverage@*20* (a diversity metric) to compare:

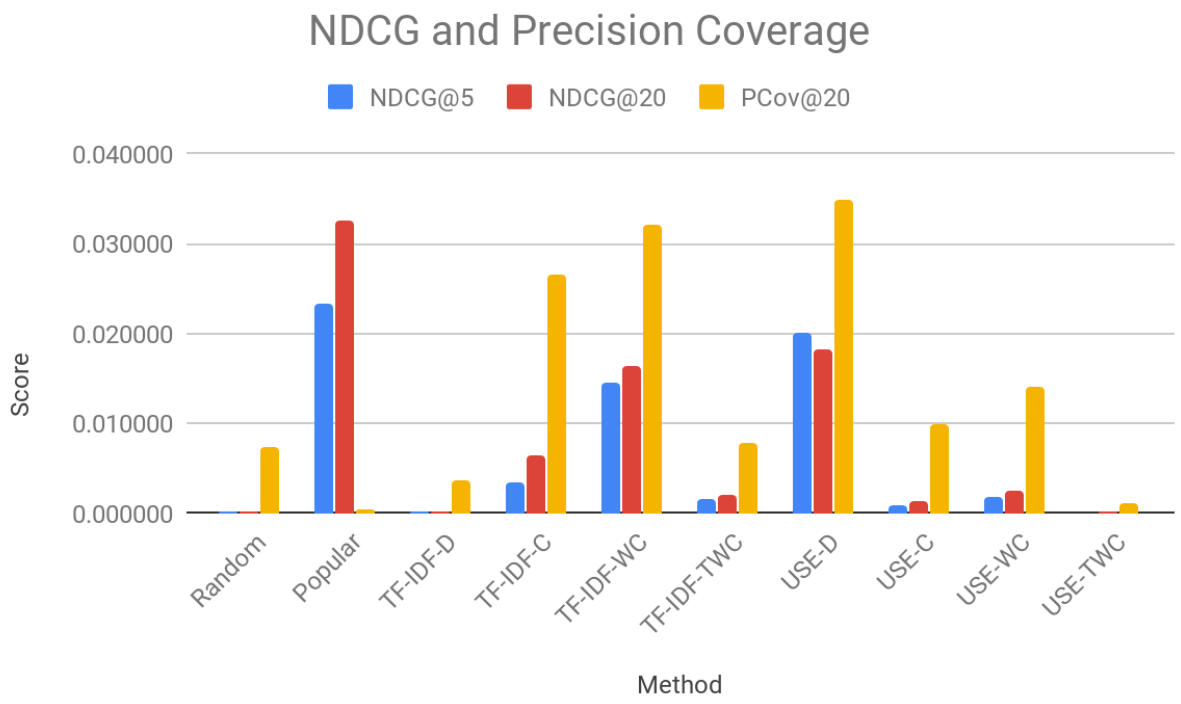

Figure 3: NDCG@*5*, NDCG@*20*, and Precision Coverage@*20* for all methods on all verse citations in the test set.

Precision Coverage is not fooled by either of these baseline methods because it uses both diversity (Coverage) and utility (Precision). This emphasizes the need to measure both aspects of a recommendation system, and combining both may result in a more trustworthy metric (see 6.5 for a proposed metric, NDCG Coverage). It's also noteworthy that Fixed Popularity had a high jump in accuracy between NDCG@*5* and NDCG@*20*, since allowing the top 20 recommendations gives fixed recommendations more opportunity to be relevant to the given query.

The success of the Fixed Popularity baseline suggests that popularity is a critical factor to take into account in many recommendation systems, which is intuitive in terms of Bayes Rule (popularity can be used as a prior). If the goal of the recommendation system is only to show results that are most likely to be preferred by users, then popularity based recommendation may

be effective. However, if recommendations are not specific to the user's interests and do not change upon receiving additional information, the user is likely to have less trust in the recommendation system. With less trust in the recommendation system, a user may be less inclined to adopt or use recommendations.

My goal for this task is to recommend serendipitous scripture passages based on a given text. For recommendations to be serendipitous, recommendations should be diverse, yet still useful. For the Fixed Popularity and Random Recommendations systems, diversity and utility seem to be contradictory. My goal, and the goal of a serendipitous recommendation system, is to suggest recommendations within the intersection of diversity and utility.

5.6.2 Content vs. Context

The Direct Query-to-Verse Similarity (content) approach with USE embeddings outperformed context-based approaches on most metrics, for both TF-IDF Bag-of-Words (BOW) and Universal Sentence Encoder (USE). Where context-based approaches outperformed (C-QRM and WC-QRM) was for TF-IDF BOW on the utility metrics Prec*@20*, Rec*@20*, and F*@20*. Figure 4 compares Precision Coverage for each metric.

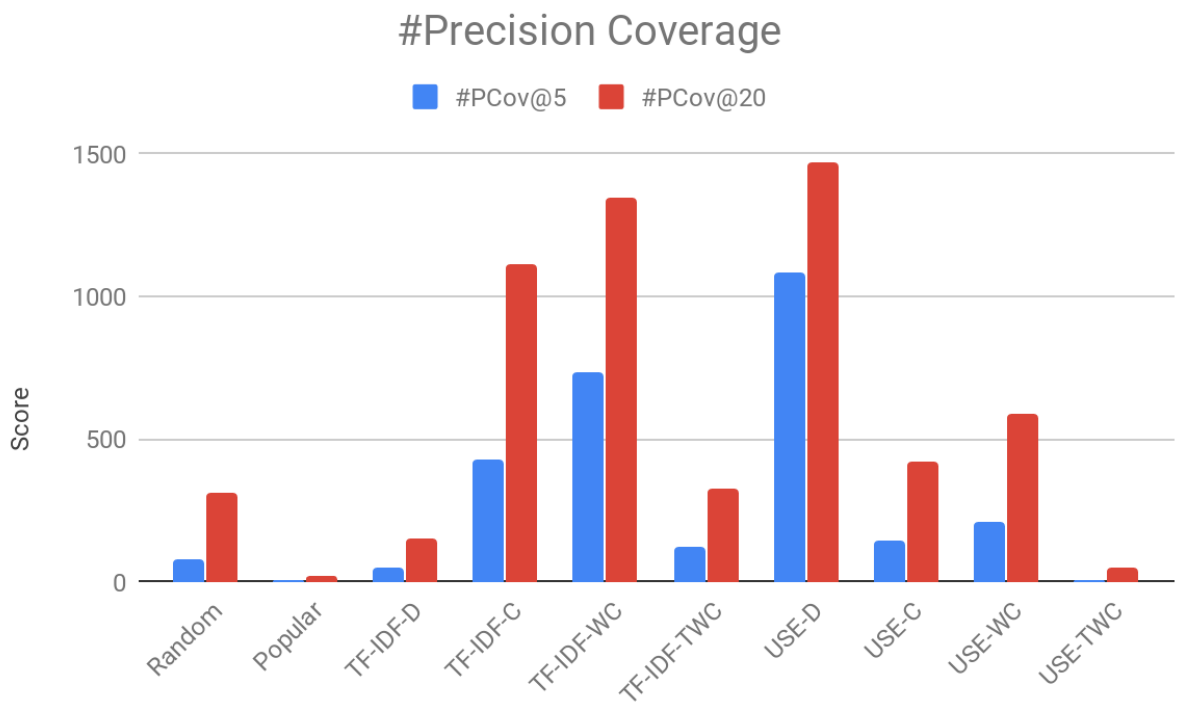

Figure 4: #Precision Coverage*@5* and #Precision Coverage*@20* for all methods on all verse citations in the test set.

While Direct approaches had higher jumps in Coverage from *@5* to *@20* (see Figure 2), context-based approaches had higher jumps in Precision Coverage (see Figure 4). The content-based (Direct) approach considers many more options because it includes uncited verses (most verses are uncited), but these options aren't necessarily relevant or commonly cited. That a context-based approach performed better with a longer recommendation list may imply that context information provides a recommendation system more relevant options or topics to consider in relation to the given query. However, with USE embeddings, the Direct approach still outperforms the C-QRM approaches *@20*.

With TF-IDF Bag-of-Words (BOW), WC-QRM outperforms the Direct approach on all metrics except Coverage, yet with USE the Direct approach outperforms context-based approaches. Furthermore, on diversity metrics the USE Direct approach outperforms TF-IDF Direct (and all other approaches other than Random). I believe this is because the Universal Sentence Encoder creates a more robust and abstract representation of content (including relationships among contiguous words) compared to TF-IDF BOW. For one, USE was trained on

a much larger dataset (likely including some scripture) and is more prepared to cross the domains (using cosine similarity) between modern English queries and scriptural or early-modern English verses, compared to bag-of-words vectors trained only on the contextual training data and scriptures.

Despite USE being more robust, all TF-IDF BOW context-based approaches outperform USE context-based approaches. This suggests that, at least for context-based recommendation, a BOW vector representation more effectively lends itself to a method that sums multiple vectors (C-QRM), compared to a neural embedding. An alternative explanation is that USE was not effective at comparing contexts, or that with USE contexts didn't provide sufficient information for effective recommendation.

Overall, content-based and context-based methods have distinct strengths, or at least distinct information, which can likely be combined to create a recommendation system that is more relevant and also diverse.

5.6.4 Summary (Scores)

While USE embeddings work well for the Direct approach, TF-IDF has more success with the WC-QRM approach and the Verse to Context Transformation approach. Results and baselines show that utility and diversity should be considered together. Finally, each method appears to have distinct advantages and disadvantages in regards to utility, diversity, and ability to recommend verses not cited in the training set. These differences I discuss in detail in 5.6.

5.6 Qualitative Discussion

In this section I provide qualitative analysis on the recommendations from each method. To enable this I use a common query, the United States (US) Pledge of Allegiance (Wikipedia 2019b), and analyze the top 5 recommendations. The recommendations by all methods are listed in Appendix B.

The following query was used for this analysis:

> I pledge allegiance to the Flag of the United States of America, and to the Republic for which it stands, one Nation under God, indivisible, with liberty and justice for all.

In this discussion, a relevant or useful recommendation is a verse of scripture that one might cite when writing about the US Pledge of Allegiance.

5.6.1 Random Recommendation

The Random baseline recommends four verses from the Old Testament and one verse from the New Testament. This demonstrates how random recommendations treat all verses as equally likely, so more verses will be recommended from the volume of scripture that has the most verses (the Old Testament). Furthermore, the recommendations do not seem relevant to the query. The number of citations for each verse does not affect random recommendation.

5.6.2 Fixed Popularity

For Fixed Popularity, the number of citations for each verse directly correlates with the verse's score. I recognize all top 5 recommendations as commonly used verses of scripture (see Appendix C for the most popular verses from each volume of scripture). The last three recommendations do not appear relevant to the query, but I can make an argument for the first two, which are broadly applicable:

> *Moses 1:39*—For behold, this is my work and my glory--to bring to pass the immortality and eternal life of man
>
> *Mosiah 18:9*—Yea, and are willing to mourn with those that mourn; yea, and comfort those that stand in need of comfort, and to stand as witnesses of God at all times and in all things, and in all places that ye may be in, even until death, that ye may be redeemed of God, and be numbered with those of the first resurrection, that ye may have eternal life--

The phrase "bring to pass the immortality and eternal life of man" in Moses 1:39 relates to the phrase "with liberty and justice for all" in the US Pledge of Allegiance, as both phrases promote general privileges for all people. In Mosiah 18:9, the passage "stand as witnesses of God at all times and in all things, and in all places that ye may be in, even until death" is essentially a pledge of allegiance to God, and the word "stand" is used similarly in both Mosiah 18:9 and the US Pledge of Allegiance ("the Republic for which it stands"). This illustrates how some broadly

applicable (and popular) scriptures can apply to a new query, though other popular scriptures do not make sense as recommendations.

5.6.3 TF-IDF BOW Direct

All five recommendations have no citations in the training data, yet all verses relate in some way to the US Pledge of Allegiance, with words such as “justice,” “pledge,” “nation,” or “liberty.” However, some verses use these terms in unrelated ways and the verse as a whole is not very relevant or useful to the query. For example, the first recommendation speaks of the “justice” of God but not about the liberty of a nation. The fact that the BOW Direct approach often recommends uncited verses means that the utility metrics used in this study (including Precision Coverage) may not be a fair metric for this method (when verses cited in the training data are included) because verses that are likely to be cited in the test set are those that were cited in the training set, yet there may be other uncited verses that are more relevant. For example, this method’s second recommendation had no training citations:

> *Galatians 5:13*—For, brethren, ye have been called unto liberty; only use not liberty for an occasion to the flesh, but by love serve one another.

Being “called unto liberty” and serving one another relates to the US Pledge of Allegiance’s call for “liberty and justice for all.”

5.6.4 TF-IDF BOW Unweighted C-QRM

The last four recommendations at least indirectly refer to a nation’s liberty, with phrases such as “I am a leader of the people” and “withdraw your armies into your own lands,” though the content or words of the verses don’t directly relate to the query. It is evident that these scriptures were recommended because they were cited in a speech that referred to the liberty of a nation. In fact, all four verses are from the same chapter and were only cited once, almost certainly in the same speech. For all four verses the top words (which are stemmed) for their bag-of-words representation are the following (many of which are also found in the US Pledge of Allegiance):

> land, yea, behold, countri, america, man, peopl, liberti, freedom, nation

This repetition from the same chapter (and the same talk) illustrates that additional measures to improve diversity could improve the recommendation system, such as only recommending a single verse from the same chapter and comparing vector representations of recommendations to not recommend verses that are very similar to each other (or even have the exact same contextual representation in this case). Furthermore, combining a content-based representation would allow choosing the best verse (most similar to the query) among each of the verses that were cited together in the same context.

The first verse recommended isn't very relevant except that it relates to a "nation under God" by the phrase "do not risk one more offense against your God." For some reason, this verse was cited in a speech that mentions the words "republ" and "pledg," as these words are found among the top weighted words for this verse. As this verse has only .5 citations (meaning it was cited with another verse), this illustrates how C-QRM requires more data to make up for noise in contexts for a verse that may not relate to the verse. However, the weighting of WC-QRM helps account for this noise by taking section and proximity into account.

5.6.5 TF-IDF BOW Weighted C-QRM

The recommendations of WC-QRM are more diversified than those of C-QRM, though there are two that are in the same chapter. The weighting of WC-QRM appears to have resulted in verse recommendations that more directly relate to the query. Recommendations 1 and 4 don't speak of a nation, but they refer to being all inclusive (as does "liberty and justice for all"). The other three recommendations refer at least indirectly to the liberty of a nation, or an ideal nation.

5.6.6 TF-IDF BOW Transformed Weighted C-QRM

In some ways, recommendations by the Transformation approach show mixed characteristics of the Direct and WC-QRM approaches. Like with Direct, some recommendations (3) had no citations. Like with WC-QRM, recommendations did not directly mention words found in the query. While some recommendations are not useful, all are in some way related to the concept of a nation. For example "Dedan, and Tema, and Buz, and all that are in the utmost corners" refers

to nations, but the verse itself is not useful in the context of the US Pledge of Allegiance. The following recommendation does relate to the concept of a "nation under God":

> *Deuteronomy 33:13*—And of Joseph he said, Blessed of the LORD be his land, for the precious things of heaven, for the dew, and for the deep that coucheth beneath,

As the Transformation approach uses a more indirect method to represent verses by both their contexts and their content, it enables serendipity in a way that the other methods do not, but it also risks recommending verses that are not useful, or relevance to the query may be unclear to users.

5.6.7 Universal Sentence Encoder Direct

The Direct approach with USE embeddings recommends verses that all have to do with a nation. One recommendation is particularly relevant, as it mentions the "Constitution" of a nation:

> *Doctrine and Covenants 109:54*—Have mercy, O Lord, upon all the nations of the earth; have mercy upon the rulers of our land; may those principles, which were evidence so honorably and nobly defended, namely, the Constitution of our land, by our fathers, be established forever.

Three of the recommendations are from the Doctrine and Covenants. This preference for the Doctrine and Covenants could be because the Doctrine and Covenants uses relatively modern English to speak about nations, and the Universal Sentence Encoder was trained on modern English.

5.6.8 Universal Sentence Encoder Unweighted C-QRM

Results for this method are similar to that of the BOW Unweighted C-QRM, with the addition that the first recommendation (from the Doctrine and Covenants) is particularly relevant. That each recommendation is at least somewhat relevant shows that the C-QRM approach can work on neural embeddings.

5.6.9 Universal Sentence Encoder Weighted C-QRM

A few recommendations for this method refer to hunger in relation to a people or nation, without directly mentioning concepts like “nation” or “liberty” (from the US Pledge of Allegiance). These verses were likely cited in a speech that referred to these concepts.

5.6.10 Universal Sentence Encoder Transformed Weighted C-QRM

Most recommendations for this method are related, though in a relatively indirect way. For example, “Burning for burning, wound for wound, stripe for stripe” refers to justice (and this verse has no citations). While this verse isn’t particularly related to the Pledge of Allegiance, it demonstrates that the Universal Sentence Encoder can encode relatively abstract concepts (however, I noticed that this verse is recommended first for many other queries as well). The following recommendation is also abstractly related to a nation:

> *Ecclesiastes 4:9*—Two are better than one; because they have a good reward for their labour.

The following recommendation appears to have been chosen due to the presence of the word “flag,” but is not relevant:

> *Job 8:11*—Can the rush grow up without mire? can the flag grow without water?

Contrary to what the metric scores imply, these results show that Transformation approach does work with USE embeddings, but with limited success.

5.6.11 Summary (Qualitative Discussion)

Other than random recommendation, each method suggested recommendations that had at least some relevance to the query, some more relevant than others. This qualitative analysis emphasizes that there are many valid methods or features to use to recommend scripture verses, and which method is more effective may depend on the user or the task. Furthermore, this analysis demonstrates that automatic metrics other than Coverage do not reward relevance for many valid recommendations, particularly verses that were never cited in the test set.

5.7 Human Evaluation Results

As a final evaluation, I requested the help of human judges to evaluate each of the 10 recommendation methods compared in this study. This evaluation is not intended to show statistical significance or to represent an unbiased population, but provides an additional perspective that represents the intended use of a scripture recommendation system for writers.

5.7.1 Details of the Study

There were 4 judges in the study, all female, between the ages 20 and 63. All judges were members of the same family, of the same religious background (The Church of Jesus Christ of Latter-day Saints), and familiar with all of the volumes of scripture used in this study.

To create the dataset, I selected random queries from the test set under the following criteria:

- Contains at least 50 characters
- Contains less than 501 characters
- Does not contain citations to scripture
- Does not contain quotation marks
- Does not directly quote scripture (manual curation)
- Is sensible as a query without further context (manual curation)

Due to the curated nature of the queries for this study, the queries were of a higher quality than those used by the automated evaluation metrics. For each query, I created a set of the first recommendation (*@1*) predicted by each method. Duplicates were removed and the order was randomized before presenting recommendations to the judges.

I instructed each judge to rate each recommended verse by usefulness in relation to the query, under the mindset that a useful verse is one that the author of the query would use or cite in the same speech. Ratings were from 1 to 5. Each of the 4 judges evaluated 10 queries, making a total of 40 queries.

5.7.2 Human Ratings and Discussion

In Table 11 and Figure 5, I present the average ratings (mean rating) for each method. Because the raters did not rate the same queries, I do not provide inter-rater agreement; to take into account differences among raters, I provide the standard deviation among raters for each method. The standard deviation is calculated using the average scores of each rater (four total) for the given method.

Table 11: Human Evaluation Ratings

Method	Mean Rating	Standard Deviation
Random	1.85	0.33
Popular	3.60	0.69
TF-IDF-D	3.53	0.63
TF-IDF-C	2.75	0.39
TF-IDF-WC	2.70	0.61
TF-IDF-TWC	1.78	0.34
USE-D	2.83	0.56
USE-C	2.33	0.51
USE-WC	2.55	0.42
USE-TWC	1.35	0.25

Average ratings for each method, from 1 to 5. Also, the standard deviation among raters.

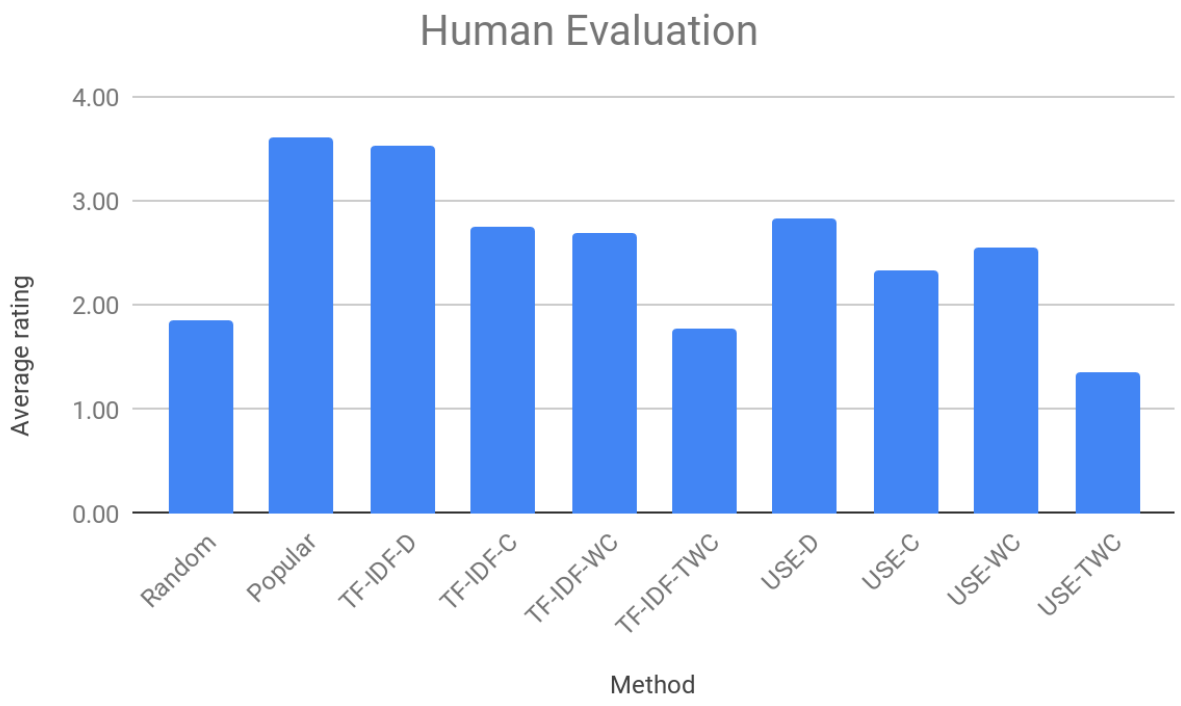

Figure 5: Average rating by human judges on a curated set of queries.

The Fixed Popularity method always recommended the same verse:

> *Moses 1:39*—For behold, this is my work and my glory--to bring to pass the immortality and eternal life of man.

However, two of the judges commented independently that they found Moses 1:39 to be relevant for many queries. One said she liked the scripture so much that she often gave it a 5 (which she did 6 times out of 10). Without including the results from this judge, Fixed Popularity is still second highest (3.27 for Popular and 3.63 for TF-IDF-D). This illustrates that a popularity-based quote recommendation system is a hard baseline to beat. However, the judges noticed the repetition and I assume that if a recommendation system always recommended similar verses they would stop using the recommendation system or wouldn't take the recommendations seriously (trust). As with automatic metrics, human evaluation has limits and requires asking the right questions (e.g. "How surprised are you that this verse was recommended?"). A study that directly monitored the use of recommended verses in writing may show different results, though I suspect that users would tend to use the verses they are most familiar with.

I was surprised by the high ratings for the TF-IDF Direct method, even compared to USE Direct (which was third overall), since TF-IDF Direct had low results on all automatic metrics other than Coverage. This provides evidence for my theory in section 5.6.3 that utility metrics do not fairly evaluate TF-IDF Direct since it often recommends uncommon verses. Also, TF-IDF BOW methods outperform USE methods for all other approaches (C-QRM, WC-QRM, TWC-QRM), so it's possible that TF-IDF Direct also outperforms USE Direct in general (except this doesn't explain why USE Direct has much higher Coverage). It is informative that the two methods with the highest score (Fixed Popularity and TF-IDF Direct) are very different from one another: TF-IDF Direct is very diverse and recommends uncommon verses, while Fixed Popularity is not diverse at all and by definition recommends the most common verses. This implies that both methods provide value to the user in different ways.

That TF-IDF Direct has higher ratings than USE Direct, and other methods, could be due to implicit bias by the judges. TF-IDF Direct directly compares the weights of important words in the query with those of the verse. A human user who seeks to know whether a verse is appropriate for the query will be glad to see the same exact words in a verse as in the query, as this requires less mental effort than understanding how a verse is relevant in an abstract way. USE embeddings create a relatively abstract representation of texts. In addition, even when abstract relevance is understood, human judges may consider use of the same words to be a significant factor in judging the relevance of a verse, whether consciously or not. However, the relative success of USE Direct over TF-IDF Direct for all automated metrics implies that in a real use case (writing a speech) an abstract representation may be more useful for recommending verses.

TF-IDF C-QRM had slightly higher ratings than its WC-QRM (weighted) counterpart (2.75 vs. 2.70), but their ratings are too similar to make a conclusion. For USE, WC-QRM had higher ratings than unweighted (2.55 vs. 2.33), though the difference is not significant (with standard deviations .42 and .51).

As with the automated metrics, TF-IDF BOW outperforms USE for context-based approaches. Recommendations by the Transformation approach were rated worse than random

overall, which I believe is because the Transformation approach often recommends verses that aren't useful, even if the words relate to the query (see 5.6.6 and 5.6.10).

5.7.3 Summary (Human Evaluation)

Popularity is a significant factor in determining whether a user will positively rate a recommendation, but does not necessarily mean the recommendation is useful. Human users are better than automated metrics at recognizing novel relevance (verses that have never been cited), but may tend to overemphasize relatively simple types of relevance (the same words). Human evaluation helps to recognize significant human elements of a recommendation system, such as how they are presented and for what purpose a user is adopting (or consuming) recommendations (e.g. Are they using them or just rating them?).

5.8 Summary

I presented results for automatic metrics for each recommendation method, demonstrating the importance of measuring both utility and diversity. I analyzed individual recommendations for each method with error analysis. Finally, I presented the results of human evaluation and discussed the implications and limitations of human evaluation.

Chapter 6. FUTURE WORK

This chapter presents candidate methods for successful scripture recommendation to writers, which will be reserved for future work.

6.1 Proposed Model: Neural Word Deviations

My proposed model makes use of the QuoteRec (Tan 2018) model to model word deviations among verses of scripture. While QuoteRec models deviations using a weight matrix for the author and for the topic of each quote, my proposed model has individual weight matrices for the chapter, book, and volume (e.g. New Testament vs. Old Testament) of the verse, which is a simple way to model word deviations based on hierarchical sections.

Each of these types of sections can have a significant effect on the word use and topic of a verse. For example, the meaning of a word is often particular to a specific volume. “Temple” in the Old Testament usually refers to the Temple of Solomon or someone’s forehead, whereas “temple” in the Doctrine and Covenants often refers to other temples. Furthermore, the word “temple” is associated with different stories in different volumes and books, so the word temple has different connotations (and topics) depending on where it is used. As another example, if many verses of a chapter discuss “faith,” then a verse recommendation system should consider “faith” as a possible topic of other verses in the chapter that don’t directly mention faith. The motivation for this is that someone may seek scriptures about both “faith” and another more specific topic that is found in a verse that doesn’t mention faith, but is in the same chapter.

The C-QRM method as described in section 4.4 and Tan et al. (2018) is used as part of the QuoteRec model (in the loss function for content modeling) to train verse representations based on their training contexts. The following sections present possible additions and alternatives to the version of the QuoteRec model that I’m proposing.

6.2 Enhancements

Enhancements describe methods that could be added to my proposed model without opposing its core methods.

6.2.1 Sample weights

For neural models, the weights assigned to each context-verse pair in the training set can be used as a sample weight, meaning that pairs with higher weights will be sampled more frequently in training. Alternatively, weights could be multiplied directly with the gradient update, essentially weighting the learning rate by sample.

6.2.2 Topical Language Modeling

As my purpose in modeling the language of scriptures and contexts is to compare their topics, a topical language model, such as in Wang et al. (2017) or Lau et al. (2017), would likely be more appropriate for this task than a standard LSTM encoder. However, a topical language model by itself provides no solution to recommend verses by their contexts. This can be used in place of the standard LSTM encoder in the QuoteRec model.

6.2.3 Hierarchical RNN

A hierarchical RNN such as TreeLSTM (Tai 2015) or HRNN (Hwang 2017) could also be used to improve the LSTM encoder in the QuoteRec model. TreeLSTM is particularly suited to modeling the hierarchical sections of scripture (and books in general).

6.2.4 Hierarchical TF-IDF

Níster has shown success with hierarchical TF-IDF, which could be adapted to modeling topical hierarchy in book sections in addition to the hierarchy of the topics themselves. However, the TF-DF-IDF (Riggin 2018) approach for multi-document summarization could be applied directly to modeling the words of a verse based on the topic of its chapter. That is, words of a verse would be considered more important if they are also frequent in the chapter. This can be further

extended to the book and volume of a verse, though I suspect that the chapter should be considered more than the book, etc.

6.2.5 Context Selection

There is much room for exploration of context selection for recommendation. As with document representation (see 2.1.11), the document context around a citation could be better represented by smaller yet more significant portions, particularly in relation to the verse. For example, training context (e.g. a paragraph) for a cited verse could be weighted higher if it's similar to contexts for that verse in other documents, though this may bias the context representation for a verse to a single topic. Consequently, a topic modeling approach (e.g. LDA) for the context of a particular verse may be effective in modeling multiple topics of a verse. The context of a citation could also be weighted by similarity to the content of the verse, but this requires a cross-domain comparison and focusing on the content directly would likely decrease serendipity.

6.2.6 Diversity

Manual analysis of verse recommendations showed that verses of the same chapter are often recommended together, which is likely because these verses were cited together in the training data. Diversity could be improved by only recommending one verse for a single chapter or book; alternatively, one could recommend as a group verses in the same chapter that are commonly cited together or that are nearby in the recommendation list.

In addition, the recommendation list could penalize verse recommendations that are very similar to verses higher in the list, or the recommendation list could maximize including topics that are representative of the query's topics. Such diversity merits further analysis by measuring individual diversity (see the first paragraph of 2.2.1.2).

6.2.7 Serendipity

Serendipity can be improved or measured by taking popularity of verses (number of citations, number of search results) into account to not bias recommendations to the most popular verses. The Fixed Popularity baseline method could be used for this purpose. The level of serendipity

could be applied at the end and be optional, as some users would want more serendipitous verses (providing information they wouldn't have thought of) and others would want more popular verses or verses that relate more directly to their writing.

6.2.8 Limiting to Top Words

As discussed in 5.5.2, results for C-QRM, WC-QRM, and the Transformation approaches may improve by only including the words with the highest TF-IDF weightings when representing each context in the training data, to remove noise and focus on key topics.

6.3 Alternatives

Alternatives describe methods that could be used instead of parts of my proposed model, or the whole recommendation model itself.

6.3.1 Transformer

The recent transformer model, the Universal Sentence Encoder or BERT (Bidirectional Encoder Representations from Transformers; Devlin 2018), could be used to model the language of contexts and verses (starting with the pretrained model), and then be tuned for sentence pair classification to predict a verse text given an input text or context. I started experimenting with training a BERT model for sentence pair classification, but found the memory requirements prohibitive for even a small embedding and batch size (and the larger the size, the more effective the model).

6.3.2 LDA

LDA would likely show improvements over TF-IDF and can still be used for the C-QRM heuristic that forms part of the QuoteRec model. Compared to TF-IDF, LDA has already shown improvements in coverage for recommendation of General Conference talks (Bean 2016). Sequential LDA (L. Du 2012) particularly applies to writing recommendation, and, as with TF-IDF, hierarchical LDA (J. Chen 2016) could be adapted to model hierarchical sections.

6.3.3 Learning to Rank

A Learning to Rank approach is an important baseline (as in QuoteRec 2018) and is an intuitive method to take into account a wide variety of information sources, including language modeling and human knowledge sources on scripture.

6.3.4 CNN

S. Bai et al. (2018) has shown that CNN may be more effective than RNN for language modeling, especially for long texts (such as the paragraphs of contexts). TCN could be applied directly to scripture recommendation as a classifier, where the inputs are contexts and the classes are verse IDs.

6.4 Data Sources

There are many data sources that can be leveraged to improve scripture recommendation. Here I discuss how additional data can be incorporated into the methods discussed earlier.

6.4.1 Human labeled topics

As with the QuoteRec (2018) model, there are human labeled topics for each verse, allowing the use of a topic weight matrix for word deviations, where each verse may have multiple topics. The following data could be used for this purpose:

- ChurchofJesusChrist.org Topical Guide
    - Large and fairly comprehensive indexing of scripture verses by topic.
- ChurchofJesusChrist.org Bible Dictionary
    - Biblical topics, with commentary and references (not as comprehensive).
- OpenBible.info topic scores
    - 166,561 topic-reference pairs (with human votes)
    - Normalized scores based on user input (positive or negative votes)
- Treasure of Scripture Knowledge
    - Close to 500,000 references by topic

In the case of topic scores, when applying the topic's weights (a row in the weight matrix) to a verse, the weights can be multiplied with the topic score for that verse. This same method can be used with references of verses in a range, to represent a verse more by topics that single out that verse when referencing it (see 4.2.1).

6.4.2 Cross-references

Cross-references (as well as topic scores) could be used to check how related different verses are when providing recommendations, to provide diversity in the recommendation list. In addition to the above mentioned data sources, there are the following:

- ChurchofJesusChrist.org Scripture Footnotes
    - References directly between verses of scripture, associated with a specific location or word in the verse.
- OpenBible.info Cross-references (Smith n.d.)
    - Votes for relatedness from user input
    - 344,789 cross references (with human votes)

6.4.3 In-context Citations

While often less structured than the data used in this thesis, there are many other texts containing citations of Bible verses and other scriptures.

- The Gutenberg Project (Project Gutenberg n.d.)
    - 54264 books, ~50 GB
    - Includes the Bible, Bible commentary, and other religious texts.
    - I did an analysis on books mentioning religion, and found hundreds of books citing the Bible.
    - A list of religious works related to The Church of Jesus Christ of Latter-day Saints is found at the following link: https://mormontextsproject.org/available-texts/

- There are many other sources of scripture citations on ChurchofJesusChrist.org that are in similar formatting as General Conference talks, such as the Ensign and manuals (requires permission to use).
- America's Public Bible (Mullen 2016)
    - A dataset of Bible citations extracted from historical newspapers in the United States, as well as code to extract the citations.

6.5 Automatic Evaluation

NDCG Coverage

NDCG provides discounting by rank and takes into account the predicted relevance score. Coverage ensures diversity. I propose a single metric NDCG Coverage (NDCGC), or Normalized Discounted Cumulative Gain and Coverage, which also discounts catalog coverage according to the total number of times an item has been recommended in the test set (this requires a second pass over all predicted recommendations). The only difference from NDCG is the addition of log2(frequency) in the denominator when calculating DCG, where frequency is the number of times an item has been recommended in all trials.

$$\text{DCG Coverage} = \sum_{rank=1}^{k} \frac{2^{relevance-1}}{\log_2(rank+1) + \log_2(frequency+1)}$$

NDCG Coverage = predicted DCG Coverage / ideal DCG Coverage

Since catalog coverage measures diversity over all queries and recommendations, item frequencies for ideal DCG are calculated over all gold standard recommendations for all test samples. The result of NDCG Coverage is that the system is rewarded for relevant items that are not only highly ranked but that haven't been recommended frequently for other queries. In practice, one could experiment with giving more importance or weight to rank or item frequency. In the current version of the formula, frequency would likely affect scores more than rank, as frequency can vary greatly among items and could be very large (whereas rank is limited to *k*, which is usually 5).

6.5 Human Evaluation

This section suggests possible methods for human evaluation of scripture recommendation. See 2.2.3 for a discussion on human evaluation in recommendation systems.

There are two possibilities for selecting the input text for human evaluation.

1. Provide a text (paragraph) for which to receive recommendations.
2. Have human testers write a sentence or paragraph of their own.

After generating scripture references for the given text, participants may rate the following:

- How helpful is each recommended scripture? (not helpful, neutral, helpful, very helpful)
    - General utility for the recommended verse in this context.
- How helpful is the entire list of recommendations?
    - General utility of the recommendation list.
    - Takes into account interaction among verses (including diversity)
- How varied is the list of recommended scriptures? (not varied, neutral, varied)
    - Measures diversity directly.
- How familiar are you with each of the recommended scriptures? (not familiar, familiar)
    - Measures novelty.
- How surprised are you that each scripture was recommended? (not surprised, surprised)
    - Measures serendipity (distinguishing serendipity from novelty)

Consciously rating utility is different than demonstrating that recommendations are used, however. For example, some recommendation systems measure click-through rate (see 2.2.3). I can measure the effectiveness of scripture (or quote) recommendations for writers by having users write an additional paragraph using any number of recommended scriptures (from 0 to 3 choices, or a specific number), and measuring the number of recommended scriptures used and whether the scriptures used were the first to be recommended. Allowing the user to choose any number of recommendations better measures overall success of the recommendation list/system (the more chosen, the better). The user should indicate which scriptures they used (preferably with a citation), and the use of a scripture as a citation doesn't require a direct quote or paraphrase.

6.6 Summary

I can expect a machine learning model that directly optimizes scripture recommendation given an input text to have better results, especially if the model makes use of available information to model word deviations and topics, such as hierarchical sections and human-labeled topics. Diversity and serendipity for this task can be enhanced by discouraging recommendations from the same section, using cross-references and textual similarity among verses, and by controlling for the popularity of recommended verses. Finally, human evaluation can more reliably measure utility, diversity, novelty, and serendipity of a recommendation system.

Chapter 7. CONTRIBUTIONS

I will explain how this thesis and the associated research provides a reference, includes modularized code, and can guide future research. See 1.1 for examples of how my research may apply to other tasks.

7.1 As a Reference

I wrote the Previous Work section (Chapter 2) such that it is a reasonably comprehensive guide or summary on research or methods that relate to scripture recommendation for writers. It is evident from this work that many subfields of NLP, such as dialog, quote recommendation, and domain adaptation, closely interact and can learn from one another. Furthermore, some subfields, like text representation (and I would suggest context selection), are broadly applicable.

In Chapter 6, I recommended future methods to pursue based on both previous work and my own research.

7.2 As Code

I implemented this research as organized code modules and Python notebooks, including the following:

7.2.1 Repositories and modules

- scripture_rec: The repository for the scripture recommendation models and data processing.
  https://github.com/JoshuaMathias/scripture_rec
- recmetrics_sweep: Convenience functions and class for running many recommendation system evaluation metrics efficiently at once, with multiple values of *k*, and over multiple batches, as well as providing a flexible way to automatically generate tables comparing methods and experiments. I hope to add more metrics to this library in the future.
  https://github.com/JoshuaMathias/recmetrics_sweep

- splitpy: Faced with memory limitations, splitpy splits matrices into files for batch processing of pipelined operations on matrices. I use this for all recommendation methods (including the baselines), and it was particularly essential for the TF-IDF transformation method, which uses a dense TF-IDF vocabulary matrix (sharing word scores among all verses) instead of a sparse matrix.
  https://github.com/JoshuaMathias/splitpy
- easyinfo: easyinfo isn't directly related to recommendation, but it contains general utilities that I created as part of my thesis research. Of this list it's the library I used the most by far.
  https://github.com/JoshuaMathias/easyinfo

7.2.2 Python Notebooks

- tfidf_scripture_rec: Turning scriptures and contexts into TF-IDF BOW vectors.
  https://colab.research.google.com/drive/1wtFtoGAuHM5WgNIwXH4tAoxQsHhhflJ9
- use_scripture_rec: Turning scriptures and contexts into USE embeddings.
  https://colab.research.google.com/drive/1oVrqW07K7kzIz2vVVMICpgzhxI3gfRjS
- contextualized_verse_representations: C-QRM, WC-QRM, and Verse to Context Transformation
  https://colab.research.google.com/drive/18jxVx_1ZsB1Ren_QcWuuYiE3XSFVASaL
- evaluate_scripture_rec: Scripture Rec systems, auto evaluation, demo, and human evaluation
  https://colab.research.google.com/drive/1XnKUufCXbzimixMhKU6iz1t_IFwlpFle
- results_scripture_rec: Compiling scores into tables
  https://colab.research.google.com/drive/11tVzBmPqNx8lzYYfUEwk8zSIoU3nWGy5
- Contact me if you are interested in the code for context selection or other data processing.

7.3 Error Analysis

In addition to providing error analysis on the specific methods implemented for this thesis, my results and discussion demonstrate potential pitfalls when evaluating utility and diversity. I also

show how common recommendation metrics compare to human evaluation (with a small sample).

7.4 Summary

The research and implementations done as part of this thesis are a framework to build upon, allowing future research to be done with much less effort (experimenting with an additional method is relatively easy to do). In addition, Chapters 2 and 6 provide clear directions to pursue to improve scripture or book passage recommendation and related tasks.

Chapter 8. Conclusion

Scripture recommendation for writers is a matter of representing a given query and each passage or verse of scripture in such a way that they can be effectively compared. I experimented with content-based and context-based representations of verses; overall, content-based approaches had higher results. However, both types of verse representations are useful and provide different types of recommendations. I showed that verses can be represented by all contexts in its citing document and that weighting these contexts by their proximity to the citation significantly improves results. I also demonstrated that a context-based representation for uncited verses can be generated by creating a verse to context transformation matrix; however, this method had relatively poor results.

TF-IDF performs better than USE embeddings for combining representations of contexts, but USE has higher results on automated metrics for comparing the similarity of queries and verses directly. Human evaluation can reward valid recommendations that were not cited in the test set. However, humans may be biased towards rewarding a method that compares words directly. Both the diversity and utility or relevance of recommendations should be evaluated in some way, for both automatic and human evaluation.

Bibliography

Adamopoulos, P., & Tuzhilin, A. (2015). On unexpectedness in recommender systems: Or how to better expect the unexpected. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(4), 54.

Ahn, Y., Lee, H., Jeon, H., Ha, S., & Lee, S. G. (2016, August). Quote Recommendation for Dialogs and Writings. In *CBRecSys@ RecSys* (pp. 39-42).

Amoualian, H., Lu, W., Gaussier, E., Balikas, G., Amini, M. R., & Clausel, M. (2017). Topical Coherence in LDA-based Models through Induced Segmentation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vol. 1, pp. 1799-1809).

Andrychowicz, M., & Kurach, K. (2016). Learning efficient algorithms with hierarchical attentive memory. *arXiv preprint arXiv:1602.03218*.

Arun, R., Suresh, V., Madhavan, C. V., & Murthy, M. N. (2010, June). On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Pacific-Asia conference on knowledge discovery and data mining* (pp. 391-402). Springer, Berlin, Heidelberg.

Bai, B., Fan, Y., Tan, W., & Zhang, J. (2017). DLTSR: a deep learning framework for recommendation of long-tail web services. *IEEE Transactions on Services Computing*.

Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*.

Bao, Y., Collier, N., & Datta, A. (2013, October). A partially supervised cross-collection topic model for cross-domain text classification. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management* (pp. 239-248). ACM.

Bean, M. G. (2016). A Framework for Evaluating Recommender Systems. *MA Thesis, Brigham Young University*.

Bender, E. M., & Friedman, B. (2018). Data statements for NLP: Toward mitigating system bias and enabling better science.

Bible Study Guide: Joshua | Resources | American Bible Society. (2018). Retrieved from http://bibleresources.americanbible.org/resource/bible-study-guide-joshua

Bird, S., Loper, E., & Klein, E. (2009), *Natural Language Processing with Python*. O'Reilly Media Inc.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Bradshaw, S., & Light, M. (2007, September). Annotation consensus: implications for passage recommendation in scientific literature. In *Proceedings of the eighteenth conference on Hypertext and hypermedia* (pp. 209-216). ACM.

Brooke, J. (1996). SUS-A quick and dirty usability scale. *Usability evaluation in industry*, *189*(194), 4-7.

Canini, K., Shi, L., & Griffiths, T. (2009, April). Online inference of topics with latent Dirichlet allocation. In *Artificial Intelligence and Statistics* (pp. 65-72).

Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., ... & Sung, Y. H. (2018). Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Chen, C., Buntine, W., Ding, N., Xie, L., & Du, L. (2015). Differential topic models. *IEEE transactions on pattern analysis and machine intelligence*, *37*(2), 230-242.

Chen, J., Wang, T. T., & Lu, Q. (2016). THC-DAT: a document analysis tool based on topic hierarchy and context information. *Library Hi Tech*, *34*(1), 64-86.

Cheng, H. T., Koc, L., Harmsen, J., Shaked, T., Chandra, T., Aradhye, H., ... & Anil, R. (2016, September). Wide & deep learning for recommender systems. In *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems* (pp. 7-10). ACM.

Christodouloupoulos, C., & Steedman, M. (2015). A massively parallel corpus: the bible in 100 languages. *Language resources and evaluation*, *49*(2), 375-395.

Chung, J., Ahn, S., & Bengio, Y. (2016). Hierarchical multiscale recurrent neural networks. *arXiv preprint arXiv:1609.01704*.

Church Newsroom. (2011). Latter-day Saints Gather Twice a Year for General Conference. Retrieved from https://www.mormonnewsroom.org/article/general-conference

Coeckelbergs, M., & Van Hooland, S. (2016). Modeling the Hebrew Bible: Potential of Topic Modeling Techniques for Semantic Annotation and Historical Analysis. In *SWASH@ ESWC* (pp. 47-52).

Conroy, J. M., & O'leary, D. P. (2001, September). Text summarization via hidden markov models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 406-407). ACM.

Deldjoo, Y., Elahi, M., Cremonesi, P., Garzotto, F., Piazzolla, P., & Quadrana, M. (2016). Content-based video recommendation system based on stylistic visual features. *Journal on Data Semantics*, *5*(2), 99-113.

Dieng, A. B., Wang, C., Gao, J., & Paisley, J. (2016). Topicrnn: A recurrent neural network with long-range semantic dependency. *arXiv preprint arXiv:1611.01702*.

Du, L., Buntine, W., Jin, H., & Chen, C. (2012). Sequential latent Dirichlet allocation. *Knowledge and information systems*, *31*(3), 475-503.

Du, L., Buntine, W., & Johnson, M. (2013). Topic segmentation with a structured topic model. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 190-200).

Du, Y., Wang, W., & Wang, L. (2015). Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1110-1118).

Duma, D., & Klein, E. (2014). Citation resolution: A method for evaluating context-based citation recommendation systems. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (Vol. 2, pp. 358-363).

Duma, D., Liakata, M., Clare, A., Ravenscroft, J., & Klein, E. (2016). Applying Core Scientific Concepts to Context-Based Citation Recommendation. In *LREC*.

Eisenstein, J. (2009, May). Hierarchical text segmentation from multi-scale lexical cohesion. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 353-361). Association for Computational Linguistics.

El Hihi, S., & Bengio, Y. (1996). Hierarchical recurrent neural networks for long-term dependencies. In *Advances in neural information processing systems* (pp. 493-499).

Gan, Z., Pu, Y., Henao, R., Li, C., He, X., & Carin, L. (2016). Unsupervised learning of sentence representations using convolutional neural networks. *arXiv preprint arXiv:1611.07897*.

Ge, M., Delgado-Battenfeld, C., & Jannach, D. (2010, September). Beyond accuracy: evaluating recommender systems by coverage and serendipity. In *Proceedings of the fourth ACM conference on Recommender systems* (pp. 257-260). ACM.

Ghosh, S., Vinyals, O., Strope, B., Roy, S., Dean, T., & Heck, L. (2016). Contextual lstm (clstm) models for large scale nlp tasks. *arXiv preprint arXiv:1602.06291*.

Gogna, A., & Majumdar, A. (2015). Matrix completion incorporating auxiliary information for recommender system design. *Expert Systems with Applications*, *42*(14), 5789-5799.

Google (n.d.). Tensorflow Hub universal-sentence-encoder. Retrieved April 3, 2019, from https://tfhub.dev/google/universal-sentence-encoder/2

Hassan, H. A. M., Sansonetti, G., Gasparetti, F., & Micarelli, A. (2018, September). Semantic-based tag recommendation in scientific bookmarking systems. In *Proceedings of the 12th ACM Conference on Recommender Systems* (pp. 465-469). ACM.

Havas Cognitive. Young Pope. Retrieved February 13, 2018, from http://cognitive.havas.com/case-studies/young-pope

He, Y., Ma, Y., van Genabith, J., & Way, A. (2010, July). Bridging SMT and TM with translation recommendation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 622-630). Association for Computational Linguistics.

Henderson, M., Al-Rfou, R., Strope, B., Sung, Y. H., Lukacs, L., Guo, R., ... & Kurzweil, R. (2017). Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652*.

Hu, L., Li, J., Nie, L., Li, X. L., & Shao, C. (2017, February). What happens next? Future subevent prediction using contextual hierarchical LSTM. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Huang, E. H., Socher, R., Manning, C. D., & Ng, A. Y. (2012, July). Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1* (pp. 873-882). Association for Computational Linguistics.

Huang, W., Kataria, S., Caragea, C., Mitra, P., Giles, C. L., & Rokach, L. (2012, October). Recommending citations: translating papers into references. In *Proceedings of the 21st ACM international conference on Information and knowledge management* (pp. 1910-1914). ACM.

Huang, W., Wu, Z., Chen, L., Mitra, P., & Giles, C. L. (2015, January). A Neural Probabilistic Model for Context Based Citation Recommendation. In *AAAI* (pp. 2404-2410).

Hwang, K., & Sung, W. (2017, March). Character-level language modeling with hierarchical recurrent neural networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5720-5724). IEEE.

Jones, E., Oliphant, E., Peterson, P., et al. (2001-). *SciPy: Open Source Scientific Tools for Python*, http://www.scipy.org/ [Online; accessed 2019-04-05].

Kannan, A., Kurach, K., Ravi, S., Kaufmann, T., Tomkins, A., Miklos, B., ... & Ramavajjala, V. (2016, August). Smart reply: Automated response suggestion for email. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 955-964). ACM.

Kataria, S., Mitra, P., & Bhatia, S. (2010, July). Utilizing Context in Generative Bayesian Models for Linked Corpus. In *AAAI* (Vol. 10, p. 1).

Kawamura, T., Watanabe, K., Matsumoto, N., Egami, S., & Jibu, M. (2018). Funding map using paragraph embedding based on semantic diversity. *Scientometrics*, *116*(2), 941-958.

Krause, J., Johnson, J., Krishna, R., & Fei-Fei, L. (2017). A hierarchical approach for generating descriptive image paragraphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 317-325).

Lau, J. H., Baldwin, T., & Cohn, T. (2017). Topically driven neural language model. *arXiv preprint arXiv:1704.08012*.

LDS Scriptures - Database and Text Exports. (2018). Retrieved from https://scriptures.nephi.org

Lenhart, P., & Herzog, D. (2016, September). Combining Content-based and Collaborative Filtering for Personalized Sports News Recommendations. In *CBRecSys@ RecSys* (pp. 3-10).

Li, C., Xu, B., Wu, G., Zhuang, T., Wang, X., & Ge, W. (2014, May). Improving Word Embeddings via Combining with Complementary Languages. In *Canadian Conference on Artificial Intelligence* (pp. 313-318). Springer, Cham.

Li, F., Zhang, M., Fu, G., & Ji, D. (2017). A neural joint model for entity and relation extraction from biomedical text. *BMC bioinformatics*, *18*(1), 198.

Li, H. (2011). Learning to rank for information retrieval and natural language processing. *Synthesis Lectures on Human Language Technologies*, *4*(1), 1-113.

Li, J., & Jurafsky, D. (2015a). Do multi-sense embeddings improve natural language understanding?. *arXiv preprint arXiv:1506.01070*.

Li, J., Luong, M. T., & Jurafsky, D. (2015b). A hierarchical neural autoencoder for paragraphs and documents. *arXiv preprint arXiv:1506.01057*.

Liu, L. (2015, October). Semantic topic-based hybrid learning resource recommendation. In *Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications*(pp. 55-60). ACM.

Liu, Y., Liu, B., Shan, L., & Wang, X. (2018). Modelling context with neural networks for recommending idioms in essay writing. *Neurocomputing*, *275*, 2287-2293.

Lu, Z., Wang, H., Mamoulis, N., Tu, W., & Cheung, D. W. (2017). Personalized location recommendation by aggregating multiple recommenders in diversity. *GeoInformatica*, *21*(3), 459-484.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).

Mimno, D., Wallach, H. M., Naradowsky, J., Smith, D. A., & McCallum, A. (2009, August). Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2* (pp. 880-889). Association for Computational Linguistics.

Mullen, L., America's Public Bible: Biblical Quotations in U.S. Newspapers, website, code, and datasets (2016): <http://americaspublicbible.org>.

Nguyen, D. Q., Billingsley, R., Du, L., & Johnson, M. (2015). Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics*, *3*, 299-313.

Niemann, K., & Wolpers, M. (2013, August). A new collaborative filtering approach for increasing the aggregate diversity of recommender systems. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 955-963). ACM.

Nister, D., & Stewenius, H. (2006). Scalable recognition with a vocabulary tree. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)* (Vol. 2, pp. 2161-2168). Ieee.

Parapar, J., Bellogín, A., Castells, P., & Barreiro, Á. (2013). Relevance-based language modelling for recommender systems. *Information Processing & Management*, *49*(4), 966-980.

Paul, M. (2009). Cross-collection topic models: Automatically comparing and contrasting text. *Bachelor Degree. Advisor: Girju, R. Department of Computer Science. University of Illinois at Urbana-Champaign*.

Project Gutenberg. (n.d.). Retrieved February 12, 2018, from www.gutenberg.org

Rendle, S. (2010, December). Factorization machines. In *2010 IEEE International Conference on Data Mining* (pp. 995-1000). IEEE.

Ricci, F., Rokach, L., & Shapira, B. (2011). Introduction to recommender systems handbook. In *Recommender systems handbook* (pp. 1-35). Springer, Boston, MA.

Riggin, K., Greve, J., Lindberg, E., Mathias, J. (2018). Topic-focused Summarization via Decomposition. *Department of Linguistics. University of Washington*.

Ritchie, A., Robertson, S., & Teufel, S. (2008, October). Comparing citation contexts for information retrieval. In *Proceedings of the 17th ACM conference on Information and knowledge management* (pp. 213-222). ACM.

Rosati, J., Ristoski, P., Di Noia, T., Leone, R. D., & Paulheim, H. (2016). RDF graph embeddings for content-based recommender systems. In *CEUR workshop proceedings* (Vol. 1673, pp. 23-30). RWTH.

Sarwar, G., O'Riordan, C., & Newell, J. (2017). Passage Level Evidence for Effective Document Level Retrieval.

Schofield, A., Magnusson, M., & Mimno, D. (2017). Pulling out the stops: Rethinking stopword removal for topic models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers* (Vol. 2, pp. 432-436).

Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830 (2011): <http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html>.

Semantic Experiences. (n.d.). Retrieved July 10, 2018, from https://research.google.com/semanticexperiences/about.html

Serban, I. V., Lowe, R., Charlin, L., & Pineau, J. (2016). Generative deep neural networks for dialogue: A short review. *arXiv preprint arXiv:1611.06216*.

Shafer, Gryphon. (n.d.). Bible-Reference. Retrieved from https://github.com/gryphonshafer/Bible-Reference

Silveira, T., Rocha, L., Mourão, F., & Gonçalves, M. (2017, April). A framework for unexpectedness evaluation in recommendation. In *Proceedings of the Symposium on Applied Computing* (pp. 1662-1667). ACM.

Smith, Stephen. (n.d.). OpenBible.info. Retrieved from www.openbible.info

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems* (pp. 3104-3112).

Suzuki, K., & Park, H. (2009). NLP-based Course Clustering and Recommendation.

Tai, K. S., Socher, R., & Manning, C. D. (2015). Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*.

Tamura, A., & Sumita, E. (2016). Bilingual segmented topic model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vol. 1, pp. 1266-1276).

Tan, J., Wan, X., & Xiao, J. (2015, February). Learning to recommend quotes for writing. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Tan, J., Wan, X., & Xiao, J. (2016, October). A neural network approach to quote recommendation in writings. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* (pp. 65-74). ACM.

Tan, J., Wan, X., Liu, H., & Xiao, J. (2018). QuoteRec: Toward quote recommendation for writing. *ACM Transactions on Information Systems (TOIS)*, *36*(3), 34.

Tang, D., Qin, B., & Liu, T. (2015). Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1422-1432).

Tang, X., Wan, X., & Zhang, X. (2014, July). Cross-language context-aware citation recommendation in scientific articles. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval* (pp. 817-826). ACM.

Vagliano, I., Figueroa, C., Rocha, O. R., Torchiano, M., Zucker, C. F., & Morisio, M. (2016, September). Redyal: A dynamic recommendation algorithm based on linked data. In *3rd Workshop on New Trends in Content-Based Recommender Systems co-located with ACM Conference on Recommender Systems (RecSys 2016)* (Vol. 1673). CEUR.

Valcarce, D. (2015, September). Exploring statistical language models for recommender systems. In *Proceedings of the 9th ACM Conference on Recommender Systems* (pp. 375-378). ACM.

Wang, W., Gan, Z., Wang, W., Shen, D., Huang, J., Ping, W., ... & Carin, L. (2017). Topic Compositional Neural Language Model. *arXiv preprint arXiv:1712.09783*.

Wikipedia contributors. (2019a, January 11). Biblical languages. In *Wikipedia, The Free Encyclopedia*. Retrieved 00:08, April 5, 2019, from https://en.wikipedia.org/w/index.php?title=Biblical_languages&oldid=877894986

Wikipedia contributors. (2019b, March 21). Pledge of Allegiance. In *Wikipedia, The Free Encyclopedia*. Retrieved 05:57, April 4, 2019, from https://en.wikipedia.org/w/index.php?title=Pledge_of_Allegiance

Xiong, W., Wu, Z., Li, B., Gu, Q., Yuan, L., & Hang, B. (2016, June). Inferring Service Recommendation from Natural Language API Descriptions. In *2016 IEEE International Conference on Web Services (ICWS)* (pp. 316-323). IEEE.

Yu, H., Wang, J., Huang, Z., Yang, Y., & Xu, W. (2016). Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4584-4593).

Zhang, H., Chow, T. W., & Wu, Q. J. (2016). Organizing books and authors by multilayer SOM. *IEEE transactions on neural networks and learning systems*, *27*(12), 2537-2550.

Zhang, H., Wang, S., Xu, X., Chow, T. W., & Wu, Q. J. (2018a). Tree2Vector: learning a vectorial representation for tree-structured data. *IEEE transactions on neural networks and learning systems*, (99), 1-15.

Zhang, H., Wang, S., Zhao, M., Xu, X., & Ye, Y. (2018b). Locality reconstruction models for book representation. *IEEE Transactions on Knowledge and Data Engineering*, *30*(10), 1873-1886.

Zhang, S., Yao, L., Sun, A., & Tay, Y. (2019). Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys (CSUR)*, 52(1), 5.

Zheng, S., Bao, H., Zhao, J., Zhang, J., Qi, Z., & Hao, H. (2015, December). A novel hierarchical convolutional neural network for question answering over paragraphs. In *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)* (Vol. 1, pp. 60-66). IEEE.

Zhu, W., Lan, C., Xing, J., Zeng, W., Li, Y., Shen, L., & Xie, X. (2016, March). Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Zhu, X., Sobihani, P., & Guo, H. (2015, June). Long short-term memory over recursive structures. In *International Conference on Machine Learning* (pp. 1604-1612).

APPENDICES

APPENDIX A. ALL SCORES

All six tables of scores are listed here (see 5.5 for discussion). For each evaluation category (all verses, cited verses, or uncited verses) utility scores and diversity scores are in separate tables. Coverage is only shown once, over all verses, because Coverage doesn't change when only cited or uncited verses are included as gold standard recommendations. See the following tables for lists of acronyms and definitions. In each table, the best scores for each metric are bolded, and the best score that isn't one of the Random (Random) or Fixed Popularity (Popular) baselines is bolded and also highlighted yellow (where one of these baselines has the highest score, there are two bolded scores).

Method Acronyms

TF-IDF	Bag-of-words vector with Term Frequency - Inverse Document Frequency weightings
USE	Universal Sentence Encoder embedding
D	Direct Query-to-Verse Similarity
C	Context-Based Quote Representation Model (C-QRM)
WC	Weighted Context-Based Quote Representation Model (WC-QRM)
TWC	Transformed Weighted Context-Based Quote Representation Model (WC-QRM)

Metric Acronyms

@5, *@20*	Only the top 5 or 20 recommendations are considered
MRR	Mean Reciprocal Rank
NDCG	Normalized Discounted Cumulative Gain
Prec	Precision (Mean Average Precision)
Rec	Recall
F	F1 Score
Cov	Coverage

PCov	Precision Coverage
#Cov	Number for unique verses used to calculate Coverage
#PCov	Number for unique verses used to calculate Precision Coverage

Table 12: Utility metrics for all verses

Method	**MRR**	**NDCG@*5***	**NDCG@*20***	**Prec@*5***	**Prec@*20***	**Rec@*5***	**Rec@*20***	**F@*5***	**F@*20***
Random	0.005019	0.000100	0.000240	0.000763	0.000802	0.000090	0.000415	0.000143	0.000430
Popular	**0.132500**	**0.023218**	**0.032659**	**0.042620**	**0.030471**	**0.009960**	**0.031353**	**0.012850**	**0.023706**
TF-IDF-D	0.004050	0.000157	0.000236	0.000646	0.000611	0.000064	0.000316	0.000105	0.000316
TF-IDF-C	0.027770	0.003500	0.006453	0.011758	0.012149	0.001638	0.006453	0.002540	0.007000
TF-IDF-WC	0.048617	0.014472	0.016486	**0.017969**	**0.012694**	0.002905	**0.007802**	0.004250	**0.007746**
TF-IDF-TWC	0.013903	0.001615	0.002060	0.004372	0.003079	0.000461	0.001220	0.000769	0.001512
USE-D	**0.068195**	**0.019972**	**0.018188**	0.016952	0.007730	**0.003314**	0.005468	**0.004605**	0.005061
USE-C	0.009324	0.000937	0.001368	0.003248	0.002913	0.000438	0.001512	0.000718	0.001736
USE-WC	0.014136	0.001771	0.002426	0.003952	0.003590	0.000505	0.001721	0.000790	0.002018
USE-TWC	0.002881	0.000060	0.000162	0.000274	0.000448	0.000039	0.000240	0.000066	0.000264

Table 13: Diversity metrics for all verses

Method	Cov@5	Cov@20	#Cov@5	#Cov@20	PCov@5	PCov@20	#PCov@5	#PCov@20
Random	**0.913561**	**0.999976**	**38365**	**41994**	0.001810	0.007382	76	310
Popular	0.000119	0.000476	5	20	0.000119	0.000476	5	20
TF-IDF-D	0.255840	0.546660	10744	22957	0.001238	0.003667	52	154
TF-IDF-C	0.258436	0.353471	10853	14844	0.010120	0.026503	425	1113
TF-IDF-WC	0.283772	0.359781	11917	15109	0.017454	0.032123	733	1349
TF-IDF-TWC	0.139945	0.283058	5877	11887	0.003000	0.007834	126	329
USE-D	**0.508775**	**0.825860**	**21366**	**34682**	**0.025813**	**0.034957**	**1084**	**1468**
USE-C	0.202762	0.289082	8515	12140	0.003453	0.009954	145	418
USE-WC	0.245172	0.330015	10296	13859	0.005072	0.014025	213	589
USE-TWC	0.018240	0.054245	766	2278	0.000214	0.001167	9	49

Table 14: Utility metrics for cited verses

Method	MRR	NDCG@5	NDCG@20	Prec@5	Prec@20	Rec@5	Rec@20	F@5	F@20
Random	0.004596	0.000096	0.000222	0.000685	0.000717	0.000088	0.000400	0.000139	0.000412
Popular	**0.132500**	**0.023685**	**0.033380**	**0.042620**	**0.030471**	**0.010388**	**0.032767**	**0.013347**	**0.024455**
TF-IDF-D	0.003811	0.000157	0.000232	0.000587	0.000543	0.000069	0.000316	0.000114	0.000305
TF-IDF-C	0.027770	0.003547	0.006572	0.011758	0.012149	0.001731	0.006774	0.002653	0.007233
TF-IDF-WC	0.048617	0.014628	0.016770	**0.017969**	**0.012694**	0.003094	**0.008241**	0.004492	**0.008085**
TF-IDF-TWC	0.013448	0.001653	0.002077	0.004284	0.002959	0.000486	0.001256	0.000801	0.001514
USE-D	**0.064848**	**0.019104**	**0.017572**	0.016140	0.007395	**0.003387**	0.005632	**0.004667**	0.005120
USE-C	0.009324	0.000952	0.001407	0.003248	0.002913	0.000471	0.001630	0.000762	0.001816
USE-WC	0.014136	0.001798	0.002479	0.003952	0.003590	0.000532	0.001837	0.000830	0.002114
USE-TWC	0.002744	0.000069	0.000175	0.000274	0.000430	0.000043	0.000262	0.000073	0.000274

Table 15: Diversity metrics for cited verses

Method	**PCov@*5***	**PCov@*20***	**#PCov@*5***	**#PCov@*20***
Random	0.001619	0.006572	68	276
Popular	0.000119	0.000476	5	20
TF-IDF-D	0.001095	0.003167	46	133
TF-IDF-C	0.010120	0.026503	425	1113
TF-IDF-WC	0.017454	0.032123	733	1349
TF-IDF-TWC	0.002857	0.007287	120	306
USE-D	**0.024217**	**0.032742**	**1017**	**1375**
USE-C	0.003453	0.009954	145	418
USE-WC	0.005072	0.014025	213	589
USE-TWC	0.000214	0.001024	9	43

Table 16: Utility metrics for uncited verses

Method	**MRR**	**NDCG@*5***	**NDCG@*20***	**Prec@*5***	**Prec@*20***	**Rec@*5***	**Rec@*20***	**F@*5***	**F@*20***
Random	0.000611	0	0.000049	0.000078	0.000086	0.00004	0.000198	0.000036	0.000081
Popular	0.000028	0	0	0	0	0	0	0	0
TF-IDF-D	0.000439	0	0.000098	0.000059	0.000068	0.000013	0.000212	0.000018	0.000072
TF-IDF-C	0.000036	0	0	0	0	0	0	0	0
TF-IDF-WC	0.000036	0	0	0	0	0	0	0	0
TF-IDF-TWC	0.000713	**0.000049**	**0.000342**	0.000088	0.00012	0.000079	0.000449	0.000055	0.000121
USE-D	**0.003971**	0	0	**0.000812**	**0.000335**	**0.001047**	**0.001688**	**0.000701**	**0.000415**
USE-C	0.000033	0	0	0	0	0	0	0	0
USE-WC	0.000033	0	0	0	0	0	0	0	0
USE-TWC	0.000313	0	0	0	0.000017	0	0.000009	0	0.000010

Table 17: Diversity metrics for uncited verses

Method	**PCov@*5***	**PCov@*20***	**#PCov@*5***	**#PCov@*20***
Random	0.000190	0.000810	8	34
Popular	0.000000	0.000000	0	0
TF-IDF-D	0.000143	0.000500	6	21
TF-IDF-C	0.000000	0.000000	0	0
TF-IDF-WC	0.000000	0.000000	0	0
TF-IDF-TWC	0.000143	0.000548	6	23
USE-D	**0.001595**	**0.002215**	**67**	**93**
USE-C	0.000000	0.000000	0	0
USE-WC	0.000000	0.000000	0	0
USE-TWC	0.000000	0.000143	0	6

Appendix B. Example Recommendations

I present the top five examples (in order) for each method based on a common query (which I repeat for easy reference). To describe each part of the examples, here's an example and then a template example (the brackets <> and italics denote information that varies by example):

Example:

2. Volume: The New Testament
Verse 29175: Galatians 5:13—For, brethren, ye have been called unto liberty; only use not liberty for an occasion to the flesh, but by love serve one another.
Score: 0
Citations for this verse: 0
Top verse words: liberti, occas, flesh, use, brethren, anoth, serv, call, ye, love

Template:

<rank>. Volume: *<(Old Testament, New Testament, Book of Mormon, Doctrine and Covenants, or Pearl of Great Price)>*
Verse <verse index from 0 to 41994>: *<book name> <chapter number>:<verse number>—<verse text>*
Score: *<similarity score between the query and this verse, assigned by the recommendation system>*
Citations for this verse: *<A citation sum of 1 is added each term a verse is cited, or 1 / (number of verses) if the citation references multiple verses in a range>*
Top *<(context or verse)>* words: *<list of 10 preprocessed (stemmed) words with the highest scores (in descending order) in this verse's vector representation>*

> Top words are only shown for TF-IDF methods, not including the Transformation approach. For C-QRM methods, the words come from training contexts for this verse. For the Direct method, the words come from the content of the verse itself. These words are shown to give insight into what words are taken into account or why a particular verse was recommended.

Random Recommendation

Query: I pledge allegiance to the Flag of the United States of America, and to the Republic for which it stands, one Nation under God, indivisible, with liberty and justice for all.

Recommendations:
1. Volume: The Old Testament
Verse 13484: Job 27:3—All the while my breath is in me, and the spirit of God is in my nostrils;
Score: 1.0
Citations for this verse: 0.5

2. Volume: The Old Testament
Verse 655: Genesis 24:64—And Rebekah lifted up her eyes, and when she saw Isaac, she lighted off the camel.
Score: 0.9999
Citations for this verse: 0

3. Volume: The New Testament
Verse 29268: Ephesians 3:17—That Christ may dwell in your hearts by faith; that ye, being rooted and grounded in love,
Score: 0.9999
Citations for this verse: 4.658

4. Volume: The Old Testament
Verse 879: Genesis 31:6—And ye know that with all my power I have served your father.
Score: 0.9999
Citations for this verse: 0

5. Volume: The Old Testament
Verse 2159: Exodus 23:15—Thou shalt keep the feast of unleavened bread: (thou shalt eat unleavened bread seven days, as I commanded thee, in the time appointed of the month Abib; for in it thou camest out from Egypt: and none shall appear before me empty:)
Score: 0.9999
Citations for this verse: 0

Fixed Popularity

Query: I pledge allegiance to the Flag of the United States of America, and to the Republic for which it stands, one Nation under God, indivisible, with liberty and justice for all.

Recommendations:
1. Volume: The Pearl of Great Price

Verse 41398: Moses 1:39—For behold, this is my work and my glory--to bring to pass the immortality and eternal life of man.
Score: 1.0
Citations for this verse: 146.6

2. Volume: The Book of Mormon
Verse 33212: Mosiah 18:9—Yea, and are willing to mourn with those that mourn; yea, and comfort those that stand in need of comfort, and to stand as witnesses of God at all times and in all things, and in all places that ye may be in, even until death, that ye may be redeemed of God, and be numbered with those of the first resurrection, that ye may have eternal life--
Score: 0.5
Citations for this verse: 78.01

3. Volume: The Book of Mormon
Verse 32473: 2 Nephi 31:20—Wherefore, ye must press forward with a steadfastness in Christ, having a perfect brightness of hope, and a love of God and of all men. Wherefore, if ye shall press forward, feasting upon the word of Christ, and endure to the end, behold, thus saith the Father: Ye shall have eternal life.
Score: 0.3333
Citations for this verse: 73.2

4. Volume: The Pearl of Great Price
Verse 41923: Joseph Smith–History 1:17—It no sooner appeared than I found myself delivered from the enemy which held me bound. When the light rested upon me I saw two Personages, whose brightness and glory defy all description, standing above me in the air. One of them spake unto me, calling me by name and said, pointing to the other--This is My Beloved Son. Hear Him!
Score: 0.25
Citations for this verse: 72.32

5. Volume: The Book of Mormon
Verse 32869: Mosiah 3:19—For the natural man is an enemy to God, and has been from the fall of Adam, and will be, forever and ever, unless he yields to the enticings of the Holy Spirit, and putteth off the natural man and becometh a saint through the atonement of Christ the Lord, and becometh as a child, submissive, meek, humble, patient, full of love, willing to submit to all things which the Lord seeth fit to inflict upon him, even as a child doth submit to his father.
Score: 0.2
Citations for this verse: 68.47

TF-IDF BOW Direct

Query: I pledge allegiance to the Flag of the United States of America, and to the Republic for which it stands, one Nation under God, indivisible, with liberty and justice for all.

Recommendations:
1. Volume: The Book of Mormon
Verse 34827: Alma 42:13—Therefore, according to justice, the plan of redemption could not be brought about, only on conditions of repentance of men in this probationary state, yea, this preparatory state; for except it were for these conditions, mercy could not take effect except it should destroy the work of justice. Now the work of justice could not be destroyed; if so, God would cease to be God.
Score: 0
Citations for this verse: 0
Top verse words: justic, could, condit, except, destroy, state, probationari, preparatori, work, ceas

2. Volume: The New Testament
Verse 29175: Galatians 5:13—For, brethren, ye have been called unto liberty; only use not liberty for an occasion to the flesh, but by love serve one another.
Score: 0
Citations for this verse: 0
Top verse words: liberti, occas, flesh, use, brethren, anoth, serv, call, ye, love

3. Volume: The Doctrine and Covenants
Verse 39757: Doctrine and Covenants 87:3—For behold, the Southern States shall be divided against the Northern States, and the Southern States will call on other nations, even the nation of Great Britain, as it is called, and they shall also call upon other nations, in order to defend themselves against other nations; and then war shall be poured out upon all nations.
Score: 0
Citations for this verse: 0
Top verse words: nation, southern, state, call, shall, britain, northern, defend, upon, divid

4. Volume: The Old Testament
Verse 5531: Deuteronomy 24:6—No man shall take the nether or the upper millstone to pledge: for he taketh a man's life to pledge.
Score: 0
Citations for this verse: 0

Top verse words: pledg, nether, millston, upper, taketh, man, take, life, shall, rawhid

5. Volume: The Book of Mormon
Verse 34844: Alma 42:30—O my son, I desire that ye should deny the justice of God no more. Do not endeavor to excuse yourself in the least point because of your sins, by denying the justice of God; but do you let the justice of God, and his mercy, and his long-suffering have full sway in your heart; and let it bring you down to the dust in humility.
Score: 0
Citations for this verse: 0
Top verse words: justic, deni, sway, let, god, endeavor, excus, dust, humil, least

TF-IDF BOW Unweighted C-QRM

Query: I pledge allegiance to the Flag of the United States of America, and to the Republic for which it stands, one Nation under God, indivisible, with liberty and justice for all.

Recommendations:
1. Volume: The Book of Mormon
Verse 34808: Alma 41:9—And now behold, my son, do not risk one more offense against your God upon those points of doctrine, which ye have hitherto risked to commit sin.
Score: 0.1965
Citations for this verse: 0.5
Top context words: pledg, fair, love, god, republ, heritag, merci, pay, declar, self

2. Volume: The Book of Mormon
Verse 35233: Alma 54:14—Now I close my epistle. I am Moroni; I am a leader of the people of the Nephites.
Score: 0.1884
Citations for this verse: 0.125
Top context words: land, yea, behold, countri, america, man, peopl, liberti, freedom, nation

3. Volume: The Book of Mormon
Verse 35232: Alma 54:13—Behold, I am in my anger, and also my people; ye have sought to murder us, and we have only sought to defend ourselves. But behold, if ye seek to destroy us more we will seek to destroy you; yea, and we will seek our land, the land of our first inheritance.
Score: 0.1884
Citations for this verse: 0.125

Top context words: land, yea, behold, countri, america, man, peopl, liberti, freedom, nation

4. Volume: The Book of Mormon
Verse 35225: Alma 54:6—Behold, I would tell you somewhat concerning the justice of God, and the sword of his almighty wrath, which doth hang over you except ye repent and withdraw your armies into your own lands, or the land of your possessions, which is the land of Nephi.
Score: 0.1884
Citations for this verse: 0.125
Top context words: land, yea, behold, countri, america, man, peopl, liberti, freedom, nation

5. Volume: The Book of Mormon
Verse 35226: Alma 54:7—Yea, I would tell you these things if ye were capable of hearkening unto them; yea, I would tell you concerning that awful hell that awaits to receive such murderers as thou and thy brother have been, except ye repent and withdraw your murderous purposes, and return with your armies to your own lands.
Score: 0.1884
Citations for this verse: 0.125
Top context words: land, yea, behold, countri, america, man, peopl, liberti, freedom, nation

TF-IDF BOW Weighted C-QRM

Query: I pledge allegiance to the Flag of the United States of America, and to the Republic for which it stands, one Nation under God, indivisible, with liberty and justice for all.

Recommendations:
1. Volume: The Book of Mormon
Verse 36546: 3 Nephi 18:23—But ye shall pray for them, and shall not cast them out; and if it so be that they come unto you oft ye shall pray for them unto the Father, in my name.
Score: 0.2242
Citations for this verse: 1.242
Top context words: pledg, prophet, god, shall, weak, much, never, may, document, armour

2. Volume: The Book of Mormon
Verse 37449: Ether 13:3—And that it was the place of the New Jerusalem, which should come down out of heaven, and the holy sanctuary of the Lord.
Score: 0.1788
Citations for this verse: 1.241
Top context words: land, nation, constitut, decre, shall, god, ether, destini, wise, lord

3. Volume: The Doctrine and Covenants
Verse 40102: Doctrine and Covenants 98:9—Nevertheless, when the wicked rule the people mourn.
Score: 0.1732
Citations for this verse: 1.501
Top context words: wise, constitut, men, nation, liberti, govern, repres, good, righteous, elect

4. Volume: The Book of Mormon
Verse 36553: 3 Nephi 18:30—Nevertheless, ye shall not cast him out from among you, but ye shall minister unto him and shall pray for him unto the Father, in my name; and if it so be that he repenteth and is baptized in my name, then shall ye receive him, and shall minister unto him of my flesh and blood.
Score: 0.1704
Citations for this verse: 2.74
Top context words: pledg, die, age, elderli, fit, live, prepar, time, life, sing

5. Volume: The Old Testament
Verse 1499: Genesis 49:26—The blessings of thy father have prevailed above the blessings of my progenitors unto the utmost bound of the everlasting hills: they shall be on the head of Joseph, and on the crown of the head of him that was separate from his brethren.
Score: 0.1573
Citations for this verse: 0.5495
Top context words: america, land, beauti, lord, hath, whose, countri, bough, 49:22, peopl

TF-IDF BOW Transformed Weighted C-QRM

Query: I pledge allegiance to the Flag of the United States of America, and to the Republic for which it stands, one Nation under God, indivisible, with liberty and justice for all.

Recommendations:
1. Volume: The Old Testament
Verse 1495: Genesis 49:22—Joseph is a fruitful bough, even a fruitful bough by a well; whose branches run over the wall:
Score: 0.1081
Citations for this verse: 2.55

2. Volume: The Old Testament

Verse 5823: Deuteronomy 33:13—And of Joseph he said, Blessed of the LORD be his land, for the precious things of heaven, for the dew, and for the deep that coucheth beneath,
Score: 0.1045
Citations for this verse: 0.2

3. Volume: The Old Testament
Verse 19557: Jeremiah 25:23—Dedan, and Tema, and Buz, and all that are in the utmost corners,
Score: 0.1025
Citations for this verse: 0

4. Volume: The Old Testament
Verse 17883: Isaiah 10:33—Behold, the Lord, the LORD of hosts, shall lop the bough with terror: and the high ones of stature shall be hewn down, and the haughty shall be humbled.
Score: 0.09989
Citations for this verse: 0

5. Volume: The Book of Mormon
Verse 32214: 2 Nephi 20:33—Behold, the Lord, the Lord of Hosts shall lop the bough with terror; and the high ones of stature shall be hewn down; and the haughty shall be humbled.
Score: 0.09989
Citations for this verse: 0

Universal Sentence Encoder Direct

Query: I pledge allegiance to the Flag of the United States of America, and to the Republic for which it stands, one Nation under God, indivisible, with liberty and justice for all.

Recommendations:
1. Volume: The Old Testament
Verse 5660: Deuteronomy 28:49—The LORD shall bring a nation against thee from far, from the end of the earth, as swift as the eagle flieth; a nation whose tongue thou shalt not understand;
Score: 0.6759
Citations for this verse: 0.008264

2. Volume: The Doctrine and Covenants
Verse 40862: Doctrine and Covenants 124:3—This proclamation shall be made to all the kings of the world, to the four corners thereof, to the honorable president-elect, and the high-minded governors of the nation in which you live, and to all the nations of the earth scattered abroad.

Score: 0.6711
Citations for this verse: 0.02135

3. Volume: The Doctrine and Covenants
Verse 38720: Doctrine and Covenants 45:69—And there shall be gathered unto it out of every nation under heaven; and it shall be the only people that shall not be at war one with another.
Score: 0.6599
Citations for this verse: 0.1896

4. Volume: The Old Testament
Verse 14897: Psalms 67:4—O let the nations be glad and sing for joy: for thou shalt judge the people righteously, and govern the nations upon earth. Selah.
Score: 0.6507
Citations for this verse: 0

5. Volume: The Doctrine and Covenants
Verse 40638: Doctrine and Covenants 109:54—Have mercy, O Lord, upon all the nations of the earth; have mercy upon the rulers of our land; may those principles, which were so honorably and nobly defended, namely, the Constitution of our land, by our fathers, be established forever.
Score: 0.6486
Citations for this verse: 3.079

Universal Sentence Encoder Unweighted C-QRM

Query: I pledge allegiance to the Flag of the United States of America, and to the Republic for which it stands, one Nation under God, indivisible, with liberty and justice for all.

Recommendations:
1. Volume: The Doctrine and Covenants
Verse 41237: Doctrine and Covenants 134:9—We do not believe it just to mingle religious influence with civil government, whereby one religious society is fostered and another proscribed in its spiritual privileges, and the individual rights of its members, as citizens, denied.
Score: 0.6522
Citations for this verse: 0.8333

2. Volume: The Book of Mormon

Verse 35229: Alma 54:10—But, as the Lord liveth, our armies shall come upon you except ye withdraw, and ye shall soon be visited with death, for we will retain our cities and our lands; yea, and we will maintain our religion and the cause of our God.
Score: 0.6475
Citations for this verse: 0.125

3. Volume: The Book of Mormon
Verse 35228: Alma 54:9—And now behold, we are prepared to receive you; yea, and except you withdraw your purposes, behold, ye will pull down the wrath of that God whom you have rejected upon you, even to your utter destruction.
Score: 0.6475
Citations for this verse: 0.125

4. Volume: The Book of Mormon
Verse 35227: Alma 54:8—But as ye have once rejected these things, and have fought against the people of the Lord, even so I may expect you will do it again.
Score: 0.6475
Citations for this verse: 0.125

5. Volume: The Book of Mormon
Verse 35226: Alma 54:7—Yea, I would tell you these things if ye were capable of hearkening unto them; yea, I would tell you concerning that awful hell that awaits to receive such murderers as thou and thy brother have been, except ye repent and withdraw your murderous purposes, and return with your armies to your own lands.
Score: 0.6475
Citations for this verse: 0.125

Universal Sentence Encoder Weighted C-QRM

Query: I pledge allegiance to the Flag of the United States of America, and to the Republic for which it stands, one Nation under God, indivisible, with liberty and justice for all.

Recommendations:
1. Volume: The Book of Mormon
Verse 35459: Alma 60:34—And now behold, I, Moroni, am constrained, according to the covenant which I have made to keep the commandments of my God; therefore I would that ye should adhere to the word of God, and send speedily unto me of your provisions and of your men, and also to Helaman.

Score: 0.6805
Citations for this verse: 0.02778

2. Volume: The Book of Mormon
Verse 35460: Alma 60:35—And behold, if ye will not do this I come unto you speedily; for behold, God will not suffer that we should perish with hunger; therefore he will give unto us of your food, even if it must be by the sword. Now see that ye fulfil the word of God.
Score: 0.6805
Citations for this verse: 0.02778

3. Volume: The Book of Mormon
Verse 35429: Alma 60:4—But behold, were this all we had suffered we would not murmur nor complain.
Score: 0.6805
Citations for this verse: 0.02778

4. Volume: The Old Testament
Verse 4345: Numbers 21:5—And the people spake against God, and against Moses, Wherefore have ye brought us up out of Egypt to die in the wilderness? for there is no bread, neither is there any water; and our soul loatheth this light bread.
Score: 0.6805
Citations for this verse: 1.0

5. Volume: The Book of Mormon
Verse 35428: Alma 60:3—And now behold, I say unto you that myself, and also my men, and also Helaman and his men, have suffered exceedingly great sufferings; yea, even hunger, thirst, and fatigue, and all manner of afflictions of every kind.
Score: 0.6805
Citations for this verse: 0.02778

Universal Sentence Encoder Transformed Weighted C-QRM

Query: I pledge allegiance to the Flag of the United States of America, and to the Republic for which it stands, one Nation under God, indivisible, with liberty and justice for all.
D&C
Recommendations:
1. Volume: The Old Testament
Verse 2102: Exodus 21:25—Burning for burning, wound for wound, stripe for stripe.

Score: 0.4698
Citations for this verse: 0

2. Volume: The Old Testament
Verse 17390: Ecclesiastes 4:9—Two are better than one; because they have a good reward for their labour.
Score: 0.4689
Citations for this verse: 1.5

3. Volume: The New Testament
Verse 29018: 2 Corinthians 11:29—Who is weak, and I am not weak? who is offended, and I burn not?
Score: 0.4678
Citations for this verse: 0

4. Volume: The Doctrine and Covenants
Verse 40504: Doctrine and Covenants 107:28—A majority may form a quorum when circumstances render it impossible to be otherwise--
Score: 0.4673
Citations for this verse: 0.07622

5. Volume: The Old Testament
Verse 13040: Job 8:11—Can the rush grow up without mire? can the flag grow without water?
Score: 0.4663
Citations for this verse: 0

Appendix C. Most Common Scripture Citations

Here I list the verses with the most citations in the training data, first with all volumes together and then by each individual volume.

All Scripture

1. Volume: The Pearl of Great Price
Verse 41398: Moses 1:39—For behold, this is my work and my glory--to bring to pass the immortality and eternal life of man.
Citations for this verse: 146.6

2. Volume: The Book of Mormon
Verse 33212: Mosiah 18:9—Yea, and are willing to mourn with those that mourn; yea, and comfort those that stand in need of comfort, and to stand as witnesses of God at all times and in all things, and in all places that ye may be in, even until death, that ye may be redeemed of God, and be numbered with those of the first resurrection, that ye may have eternal life--
Citations for this verse: 78.01

3. Volume: The Book of Mormon
Verse 32473: 2 Nephi 31:20—Wherefore, ye must press forward with a steadfastness in Christ, having a perfect brightness of hope, and a love of God and of all men. Wherefore, if ye shall press forward, feasting upon the word of Christ, and endure to the end, behold, thus saith the Father: Ye shall have eternal life.
Citations for this verse: 73.2

4. Volume: The Pearl of Great Price
Verse 41923: Joseph Smith--History 1:17—It no sooner appeared than I found myself delivered from the enemy which held me bound. When the light rested upon me I saw two Personages, whose brightness and glory defy all description, standing above me in the air. One of them spake unto me, calling me by name and said, pointing to the other--This is My Beloved Son. Hear Him!
Citations for this verse: 72.32

5. Volume: The Book of Mormon
Verse 32869: Mosiah 3:19—For the natural man is an enemy to God, and has been from the fall of Adam, and will be, forever and ever, unless he yields to the enticings of the Holy Spirit, and putteth off the natural man and becometh a saint through the atonement of Christ the Lord, and

becometh as a child, submissive, meek, humble, patient, full of love, willing to submit to all things which the Lord seeth fit to inflict upon him, even as a child doth submit to his father.
Citations for this verse: 68.47

The Old Testament

1.
Verse 6491: Joshua 24:15—And if it seem evil unto you to serve the LORD, choose you this day whom ye will serve; whether the gods which your fathers served that were on the other side of the flood, or the gods of the Amorites, in whose land ye dwell: but as for me and my house, we will serve the LORD.
Citations for this verse: 40.45

2.
Verse 22402: Amos 3:7—Surely the Lord GOD will do nothing, but he revealeth his secret unto his servants the prophets.
Citations for this verse: 37.5

3.
Verse 17672: Isaiah 1:18—Come now, and let us reason together, saith the LORD: though your sins be as scarlet, they shall be as white as snow; though they be red like crimson, they shall be as wool.
Citations for this verse: 34.46

4.
Verse 23130: Malachi 3:10—Bring ye all the tithes into the storehouse, that there may be meat in mine house, and prove me now herewith, saith the LORD of hosts, if I will not open you the windows of heaven, and pour you out a blessing, that there shall not be room enough to receive it.
Citations for this verse: 31.55

5.
Verse 2063: Exodus 20:12—Honour thy father and thy mother: that thy days may be long upon the land which the LORD thy God giveth thee.
Citations for this verse: 27.14

The New Testament

1.

Verse 26762: John 17:3—And this is life eternal, that they might know thee the only true God, and Jesus Christ, whom thou hast sent.
Citations for this verse: 62.32

2.
Verse 26136: John 3:16—For God so loved the world, that he gave his only begotten Son, that whosoever believeth in him should not perish, but have everlasting life.
Citations for this verse: 56.88

3.
Verse 24048: Matthew 25:40—And the King shall answer and say unto them, Verily I say unto you, Inasmuch as ye have done it unto one of the least of these my brethren, ye have done it unto me.
Citations for this verse: 56.09

4.
Verse 30271: James 1:5—If any of you lack wisdom, let him ask of God, that giveth to all men liberally, and upbraideth not; and it shall be given him.
Citations for this verse: 55.12

5.
Verse 26674: John 14:6—Jesus saith unto him, I am the way, the truth, and the life: no man cometh unto the Father, but by me.
Citations for this verse: 53.36

The Book of Mormon

1.
Verse 33212: Mosiah 18:9—Yea, and are willing to mourn with those that mourn; yea, and comfort those that stand in need of comfort, and to stand as witnesses of God at all times and in all things, and in all places that ye may be in, even until death, that ye may be redeemed of God, and be numbered with those of the first resurrection, that ye may have eternal life--
Citations for this verse: 78.01

2.
Verse 32473: 2 Nephi 31:20—Wherefore, ye must press forward with a steadfastness in Christ, having a perfect brightness of hope, and a love of God and of all men. Wherefore, if ye shall press forward, feasting upon the word of Christ, and endure to the end, behold, thus saith the Father: Ye shall have eternal life.

Citations for this verse: 73.2

3.
Verse 32869: Mosiah 3:19—For the natural man is an enemy to God, and has been from the fall of Adam, and will be, forever and ever, unless he yields to the enticings of the Holy Spirit, and putteth off the natural man and becometh a saint through the atonement of Christ the Lord, and becometh as a child, submissive, meek, humble, patient, full of love, willing to submit to all things which the Lord seeth fit to inflict upon him, even as a child doth submit to his father.
Citations for this verse: 68.47

4.
Verse 31152: 1 Nephi 3:7—And it came to pass that I, Nephi, said unto my father: I will go and do the things which the Lord hath commanded, for I know that the Lord giveth no commandments unto the children of men, save he shall prepare a way for them that they may accomplish the thing which he commandeth them.
Citations for this verse: 59.37

5.
Verse 31776: 2 Nephi 2:25—Adam fell that men might be; and men are, that they might have joy.
Citations for this verse: 53.13

The Doctrine and Covenants

1.
Verse 38187: Doctrine and Covenants 20:77—O God, the Eternal Father, we ask thee in the name of thy Son, Jesus Christ, to bless and sanctify this bread to the souls of all those who partake of it, that they may eat in remembrance of the body of thy Son, and witness unto thee, O God, the Eternal Father, that they are willing to take upon them the name of thy Son, and always remember him and keep his commandments which he has given them; that they may always have his Spirit to be with them. Amen.
Citations for this verse: 59.99

2.
Verse 37997: Doctrine and Covenants 14:7—And, if you keep my commandments and endure to the end you shall have eternal life, which gift is the greatest of all the gifts of God.
Citations for this verse: 58.21

3.

Verse 37743: Doctrine and Covenants 1:38—What I the Lord have spoken, I have spoken, and I excuse not myself; and though the heavens and the earth pass away, my word shall not pass away, but shall all be fulfilled, whether by mine own voice or by the voice of my servants, it is the same.
Citations for this verse: 52.17

4.
Verse 40832: Doctrine and Covenants 121:45—Let thy bowels also be full of charity towards all men, and to the household of faith, and let virtue garnish thy thoughts unceasingly; then shall thy confidence wax strong in the presence of God; and the doctrine of the priesthood shall distil upon thy soul as the dews from heaven.
Citations for this verse: 51.2

5.
Verse 39579: Doctrine and Covenants 81:5—Wherefore, be faithful; stand in the office which I have appointed unto you; succor the weak, lift up the hands which hang down, and strengthen the feeble knees.
Citations for this verse: 44.83

The Pearl of Great Price

1.
Verse 41398: Moses 1:39—For behold, this is my work and my glory--to bring to pass the immortality and eternal life of man.
Citations for this verse: 146.6

2.
Verse 41923: Joseph Smith--History 1:17—It no sooner appeared than I found myself delivered from the enemy which held me bound. When the light rested upon me I saw two Personages, whose brightness and glory defy all description, standing above me in the air. One of them spake unto me, calling me by name and said, pointing to the other--This is My Beloved Son. Hear Him!
Citations for this verse: 72.32

3.
Verse 41994: Articles of Faith 1:13—We believe in being honest, true, chaste, benevolent, virtuous, and in doing good to all men; indeed, we may say that we follow the admonition of Paul--We believe all things, we hope all things, we have endured many things, and hope to be able to endure all things. If there is anything virtuous, lovely, or of good report or praiseworthy, we seek after these things.

Citations for this verse: 53.08

4.
Verse 41796: Abraham 3:25—And we will prove them herewith, to see if they will do all things whatsoever the Lord their God shall command them;
Citations for this verse: 51.37

5.
Verse 41984: Articles of Faith 1:3—We believe that through the Atonement of Christ, all mankind may be saved, by obedience to the laws and ordinances of the Gospel.
Citations for this verse: 33.58

Appendix D. Additional Reading

Modeling Segments and Hierarchy

See 2.1.2. Some related models for hierarchy are tree LSTM (Tai 2015; X. Zhu 2015), LSTM with regularization (to avoid overfitting) and co-occurrence exploration (W. Zhu 2016), contextual hierarchical LSTM (J. Li 2015b; Hu 2016; Ghosh 2016), hierarchical CNN+LSTM (Gan 2016), hierarchical RNN (El Hihi 1995; Y. Du 2015; Chung 2016; Yu 2016; Krause 2017), hierarchical RNN language model (HRNNLM; Lin 2015), character-level HRNNLM (Hwang 2017), gated RNN (D. Tang 2015), a layered RNN for multitask learning (P. Liu 2016), clockwork RNN (Koutnik 2014), paragraph vector (Mikolov 2014), hierarchical attention network (Yang 2016; Hassan 2018), hierarchical CNN (Zheng 2015), CNN-deconvolutional CNN (CNN-DCNN; Y. Zhang 2017), and hierarchical HMM (Fine 1998; Bui 2004).