Automatic Index Generation for Religious Texts

Jeff Chase

December 12, 2008

1 Motivation

Many religions advocate the study of a canonical text on a regular basis. As a consequence there are many study aids available to the religious student. Serious scholars have identified important passages that the the more casual student should focus on. They also provide explanations and historical details as notes to accompany study. It is often recommended to the student that he/she study the text by topic, rather than simply reading straight through. To this end topical guides and indexes are available. These guides allow a student to quickly find passages of interest in the text.

More recently, software packages have been developed that provide easy access to large amounts of published study materials and research related to the Bible[1]. This project shows how machine learning could be included in these software packages to further aid study. In particular, it demonstrates the capabilities of machine learning to perform automatic index generation.

2 Training Set

2.1 Source Text

The source text used for this project was the King James Version of the Bible. The Bible is composed of over 30,000 small passages referred to as verses. The verses are organized into chapters and the chapters into books. The source text was obtained from Project Gutenberg[2]. The text was prepared by removing punctuation, converting to lower case, and assigning each verse a unique ID. A stemmer[3] was used to reduce each word to its stem. Finally, the vocabulary was extracted and each word replaced by a unique ID. There were a total of 31,102 verses and 9,598 unique words.

2.2 Topic Index

The labels for the training set were obtained from bible-topics[4], a website containing a topical index for the Bible. It supplies a list of relevant verses for almost 700 different topics. A representative subset of 39 topics was chosen for this project[5]. In order to convert the html index files to sets of labels they were first modified slightly to be readable XML files. From there the verse references were extracted and converted to the same unique verse IDs used in the source text. Note that the same verse may appear under multiple topics.

3 Initial Implementation

3.1 Training

Training was performed using SVMs by applying a multivariate Bernoulli event model to the data: the presence or absence of each word in the vocabulary was viewed as a separate feature. A separate classifier for each topic was trained by applying a different set of labels to the same training set. To significantly speed up training the training set was limited to those verses that were referenced at least one time in any topic.

Table 1: Results Summary

	Average Error Rate	Best Error Rate
No optimization	55.7%	10.5%
C Weighting	45.6%	13.7%
Feature Selection	26.1%	1.8%

This limited the training set to 2,467 verses. This also limited the vocabulary to 3,144. The verses referenced under a topic were given positive labels and the rest in the limited training set were given negative labels. This was done under the assumption that the other verses were considered and rejected for classification under the current topic.

The training was done using PyML[6], a machine learning library for Python.

3.2 Testing

The most important metric for measuring the success of a classifier was the false negative rate, which will be referred to as simply the error rate. This refers to the verses that were originally listed under the topic but not recognized during testing. False positives were not considered an issue. First, the assumption under which the negative examples were labeled was questionable; some negative examples may in fact be positive ones. In addition, as the purpose of this project was to discover additional positive verses, it was desirable that the classifier was biased towards positive assignments.

The initial classifiers were tested using 10-fold cross validation. The average error rate for the topical classifiers was 55.7%. The best error rate was 10.5%. However, the error rate when training and testing using the entire training set was 0%, indicating low bias but high variance.

4 Improvements

In order to bias the classifiers towards positive assignments more weight was given to positive example errors than to negative example errors. This was done by setting two different values for C in the SVM, as described in [7]. This improved the average error rate to 45.6%. The best error rate increased slightly to 13.7%.

To avoid over-fitting the data, feature reduction was applied to the training set before training by recursively eliminating the features with the smallest corresponding weights in the classifier. This reduced the number of features from 3,144 to 176, on average. After feature reduction the average error rate was 26.1%, computed by leave-one-out cross validation. The best error rates, however, were just 1.8%, 5.5%, 7.7%, and 8.8%.

These results are summarized in Table 1.

5 Qualitative Results

Although the above numbers give an idea of the performance of this system, its real performance will be determined by the subjective opinion of the user. Here are presented some of the qualitative results.

5.1 Key Words

After training a particular topic, key words can be identified by by choosing the features with the highest corresponding weights in the classifier function. Table 2 shows some random examples.

Table 2: Key Words

Topic	Key Words
Prayer	prayer, prai, holiest, pour, ask
Faith	believeth, believ, faith, thoma, lot
Atonement	aton, sheep, blood, foreordain, reconcil
Creation	creat, visitest, new, morn, calleth
Sin	sin, sinneth, forgotten, destroy, rebel
Peace	peac, quiet, peacemak, desireth, depart

5.2 Verse Discovery

As already stated, the purpose of this project was to discover additional verses that belong to a given topic. The most likely verses may be found by comparing the output of the classifying function for each verse before the class decision is made. See the appendix for some examples of the most likely verses.

A limited subjective survey showed that the identified verses are relevant to the topic. It is difficult to assess, however, whether these are the most relevant verses in the entire source text.

6 Conclusion

This project showed how machine learning may be applied to expand existing topical indexes of the Bible. It was successful at finding related verses that were not already listed in the index. However, there are still many ways in which it could be improved: A different model of the data could be tried, such as the multinomial event model. Additional features could be added, such as proximity to other verses with positive labels. A separate classifier could be trained to identify interesting versus non-interesting verses in order to filter the additional results found. An improved stemmer could be used to take into account archaic suffixes such as "-eth" and "-est".

Performance also suffered from a lack of enough positive training examples for some topics. This project would benefit from a careful analysis of the number of training examples needed and then improving the existing index to meet that requirement.

Notes

[1] See `http://www.e-sword.net` and `http://www.logos.com`

[2] `http://www.gutenberg.org`

[3] `http://tartarus.org/~martin/PorterStemmer/index.html`

[4] `http://www.bible-topics.com`

[5] altars, angels, anger, atonement, baptism, charity, chastity, church, covenants, creation, devil, eternallife, faith, fall, giftsofgod, god, gospel, heaven, hope, humility, idolatry, joy, justice, light, mercy, miracles, oaths, peace, prayer, pride, resurrection, sabbath, sanctification, scriptures, selfishness, sin, truth, visions, war

[6] `http://pyml.sourceforge.net`

[7] `http://pyml.sourceforge.net/doc/howto.pdf`

A Most Likely Verses

Prayer Daniel 6:10 - Now when Daniel knew that the writing was signed, he went into his house, and his windows being open in his chamber toward Jerusalem, he kneeled upon his knees three times a day, and prayed, and gave thanks before his God, as he did aforetime.

Faith 1 Peter 1:21 - Who by him do believe in God, that raised him up from the dead, and gave him glory; that your faith and hope might be in God.

Atonement Isaiah 35:10 - And the ransomed of the Lord shall return, and come to Zion with songs and everlasting joy upon their heads: they shall obtain joy and gladness, and sorrow and sighing shall flee away.

Creation Isaiah 45:7 - I form the light, and create darkness: I make peace, and create evil: I the Lord do all these things.

Sin 1 Corinthians 6:18 - Flee fornication. Every sin that a man doeth is without the body; but he that committeth fornication sinneth against his own body.

Peace 1 Chronicles 22:9 - Behold, a son shall be born to thee, who shall be a man of rest; and I will give him rest from all his enemies round about: for his name shall be Solomon, and I will give peace and quietness unto Israel in his days.