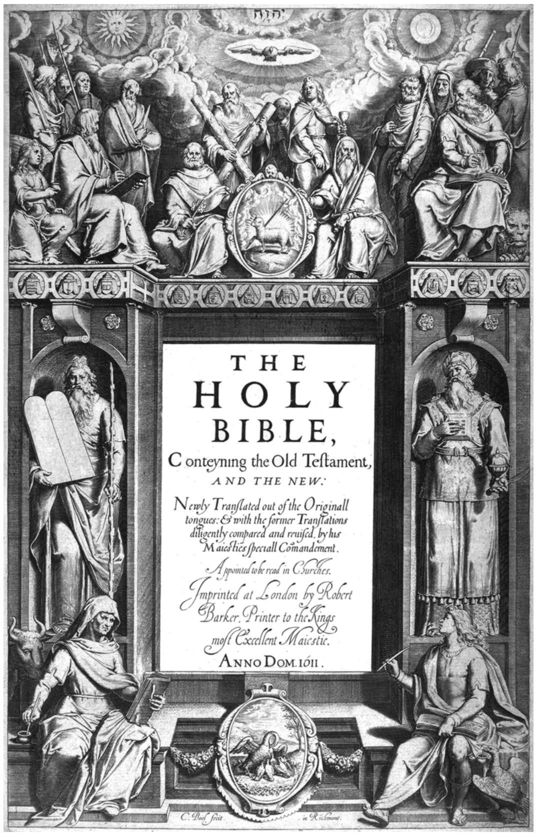
King John’s Bible - text data analysis

Tomáš Kasl

KJB is an Early Modern English translation of the Christian Bible for the Church of England, which was commissioned in 1604 and published in 1611, by sponsorship of King James VI. Here, we only focus on the biblical books belonging to either the Old Testament (also known as the Hebrew Bible, written before Christ), and the New Testament (based on the life and teachings of Jesus).



THE
HOLY
BIBLE,
Conteyning the Old Teſtament,
AND THE NEW:
Newly Tranſlated out of the Originall
tongues: & with the former Tranſlations
diligently compared and reuiſed, by his
Maieſties ſpeciall Cōmandement.
Appointed to be read in Churches.
Imprinted at London by Robert
Barker, Printer to the Kings
moſt Excellent Maieſtie.
ANNO DOM. 1611.

Figure 1: KJB (from wikipedia.org)

Can we tell the Old-testament verses apart from the New-testament ones?

Our question is, in other words: what are theirs verses most pronounced differences? How precise can we become in guessing, from which testament a verse is from?

More Precisely:

Dataset:

We are given a textfile with the KJB content transcribed from the physical book, where each line has exactly one verse; in this format:

Bookname section_number:verse_number verse_text.

for example:

```
Nahum 2:10  She is empty, and void, and waste: and the heart melteth,
        and the knees smite together, and much pain is in all loins,
        and the faces of them all gather blackness.
Nahum 2:11  Where is the dwelling of the lions, and the feedingplace
        of the young lions, where the lion, even the old lion,
        walked, and the lion's whelp, and none made them afraid?
Nahum 2:12  The lion did tear in pieces enough for his whelps, and strangled
        for his lionesses, and filled his holes with prey, ...
```

The dataset only contains the Old and New Testament, no books from the collection of biblical apocrypha.

All the books present are complete. All of the books have the correct number of chapters and verses. Comparing the content of a random verses against other online sources on the Internet, it appears to be a correct transcription, yet it is beyond this project's scope to confirm the correctness of all of the verses.

The task:

We are facing a binary classification problem of text strings, we look for

$h : strings \to \{0, 1\}$ (0 is for the Old Testament, 1 for New)

s.t. it minimizes the 0/1 loss function, i.e. $l(y^*, y) = 1 - [y^* == y]$, where $y^* := h(string)$, on the dataset of isolated verses.

Additionally:

Because of the nature of the data (basically raw text), no numerical or statistical exploration can be done right away, everything must be mined from the text artificially.

Before any analysis can be done, the textfile must first be parsed into more usable data structures. Since the text is in ‘modern’ English, the whole text is represented using only the ASCII characters, meaning no errors in UTF or other character encodings might happen.

Moreover, depending on a specific question, the biblical text should be looked upon either as a text, or as collections of tokens (~normalized words), or as sets of unique tokens.

For this, there is also a list of socalled stopwords leveraged (provided via the assignement website), usually short words, which bear little meaning, like prepositions. I have included few archaic forms to that list (such as “thou”), so it now comprises of 620 words.

Some basic observations:

Number of books: 66

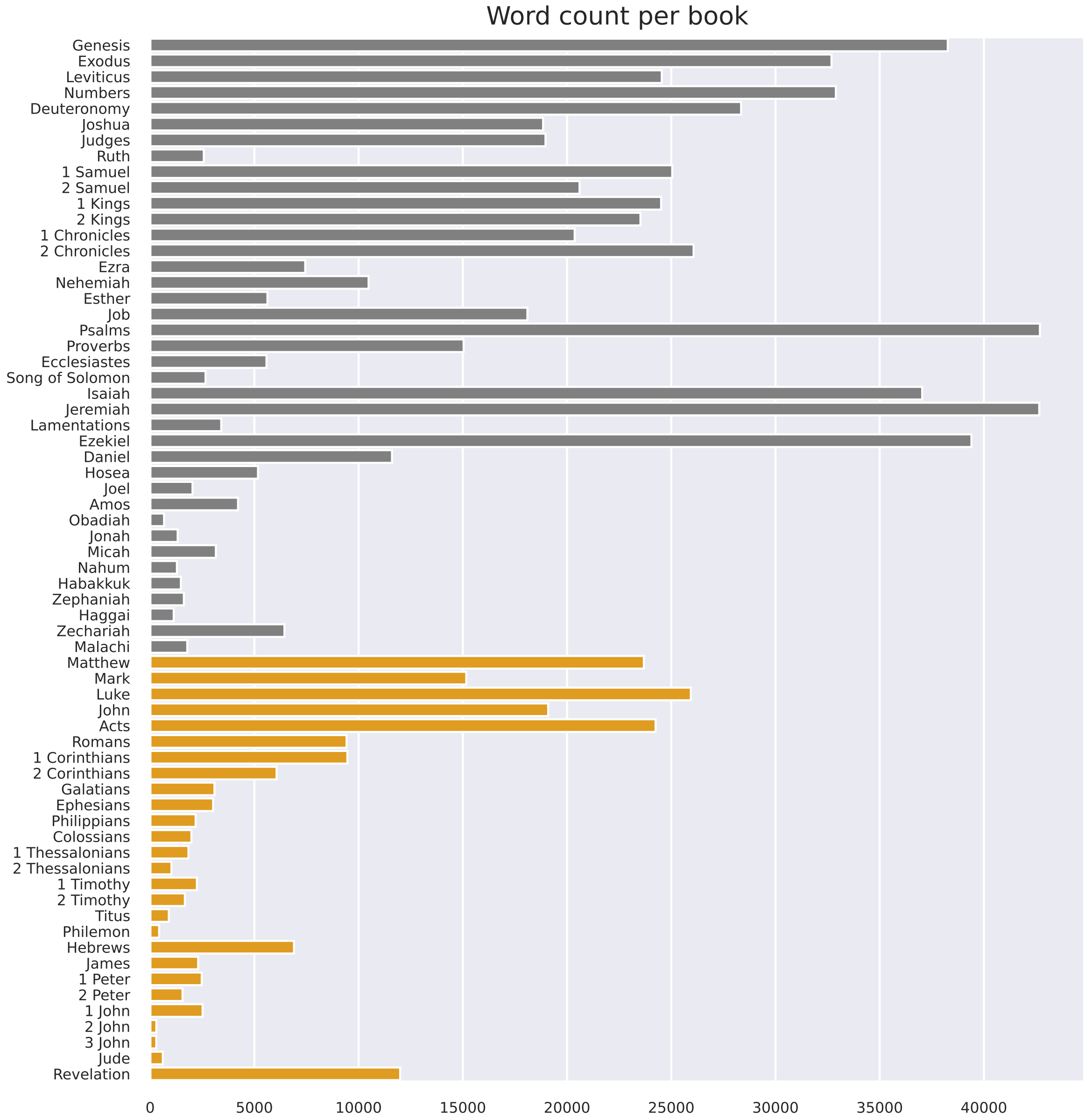
Old Testament: 39

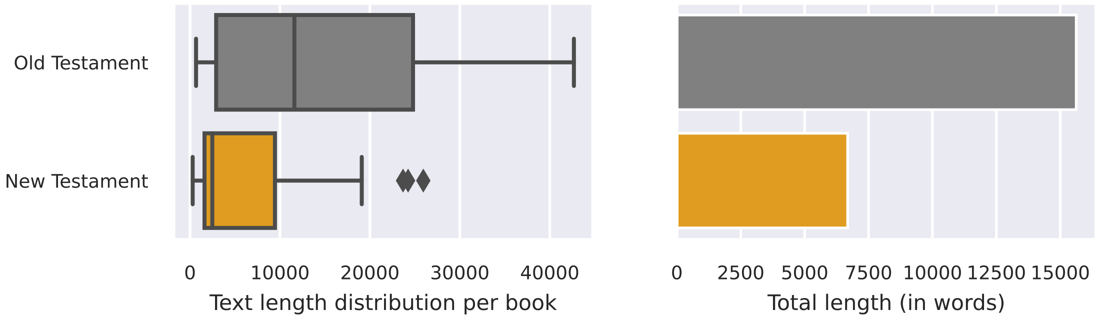
New Testament: 27

Total number of words: 789669

Number of unique words: 26467

Comparison of book lengths. Which testament, Old or New, contains more text/content (i.e. word count)?





Observations:

We already see that the Old Testament text is larger by a huge margin. This may be somewhat surprising, yet it actually makes sense, since the authors of the Old Testament had way more time (over a millenium) to write-out the whole set of books (compared to the New Testament, which was written in only few decades).

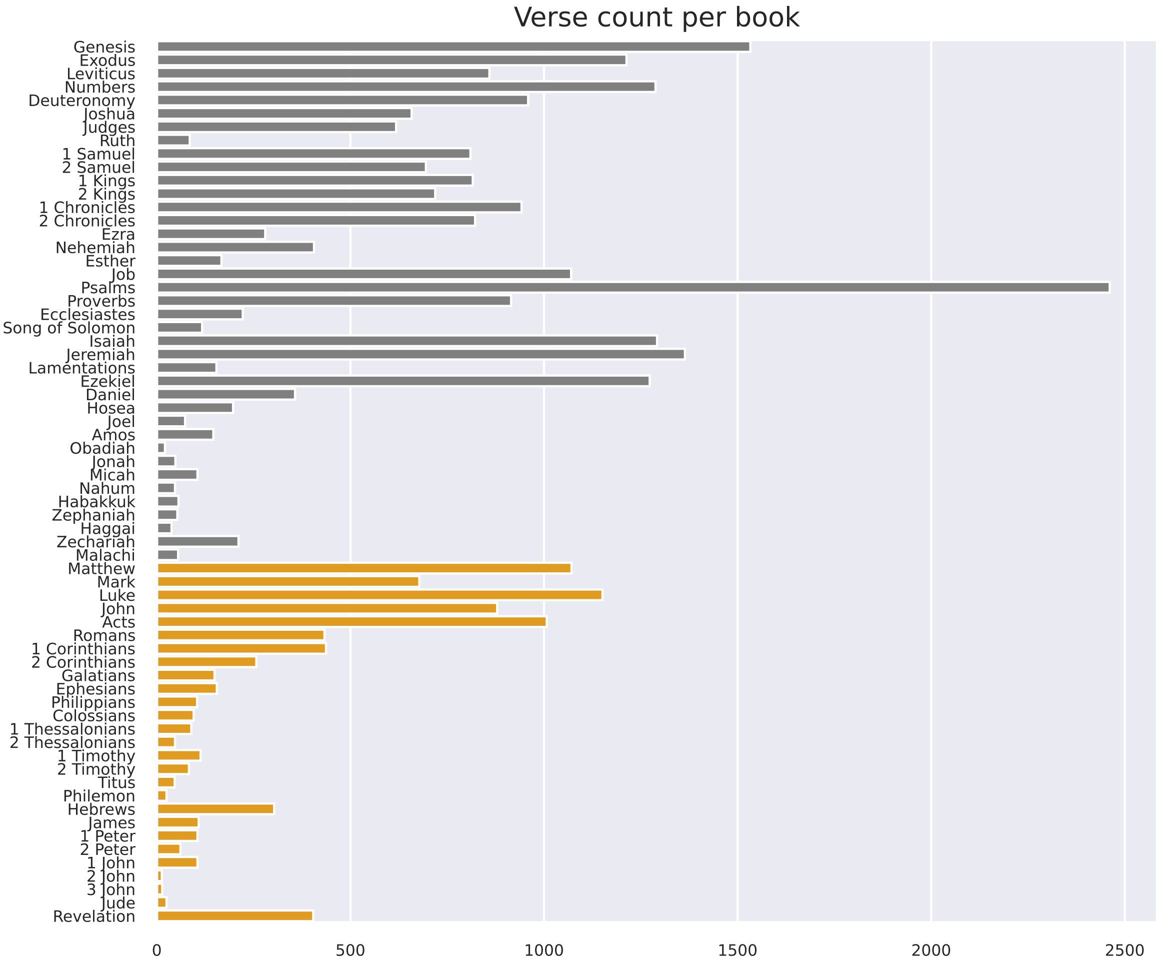
Therefore, just by naively expecting every verse to be from the Old Testament, we can probably get quite high accuracy already. But more on that later.

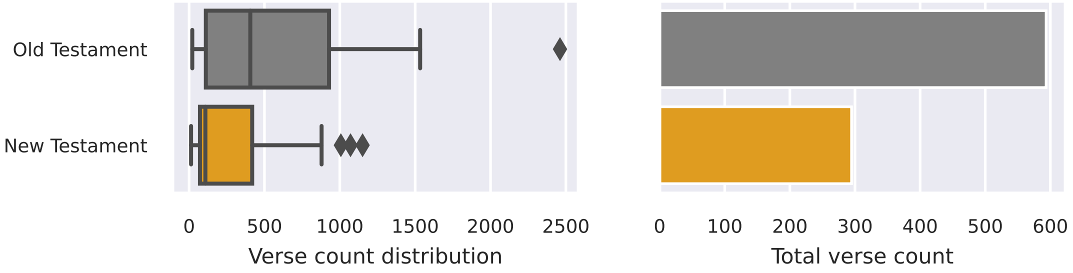
Not only is the Old Testament much larger, the dispersion from the books' length mean is also much larger for the Old Testament. Only 6 New Testament books surpass 10000 words, and only 3 books surpass 20000 words. Majority of New Testament books are shorter than 10000 words, while majority of Old Testament books are larger than 10000 words.

How well does this finding translate into the amount of verses in each book?

Is the difference in size between the testaments caused mainly by the amount of verses per book, or by the Old Testament just having longer verses?

Note that any of these findings might be helpful to improve our predictions.

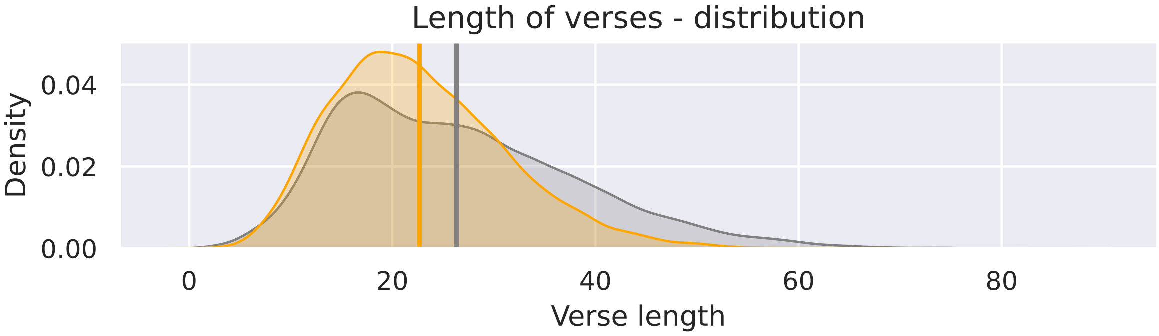




The counts relatively well match the total text lengths, overall. Old Testament books have generally more verses in them. There are outliers, too, like Obadiah with 21 verses total.

Let's now compare the average length of verses among the books.

Can we get a clear indication of the testament by looking at the length of the specific verse? Since the number of verses (even just in the KJB dataset) is so high, I model the distribution as continuous.



The length distributions of both testaments appears to be quite similar, with the verses from the Old Testament being both extremely short and long more often. Both of the testaments have the mean (the vertical line) value just over 20 words per verse, even thought the New Testament's is a bit shorter.

Therefore, usability of the verse length is not such a good indicator on its own. It still could improve the precision of compound classifiers.

What about words unique to their testament. Can we find words, which can pick the correct Testament for us? Yes!

Note: we count a plural and a singular of a word as two distinct words, for the sake of simplicity.

Number of words only present in the Old Testament / New Testament:

6775 / 1881

Number of words only present in the Old Testament / New with at least 3 occurences (aka "Extreme evidence"):

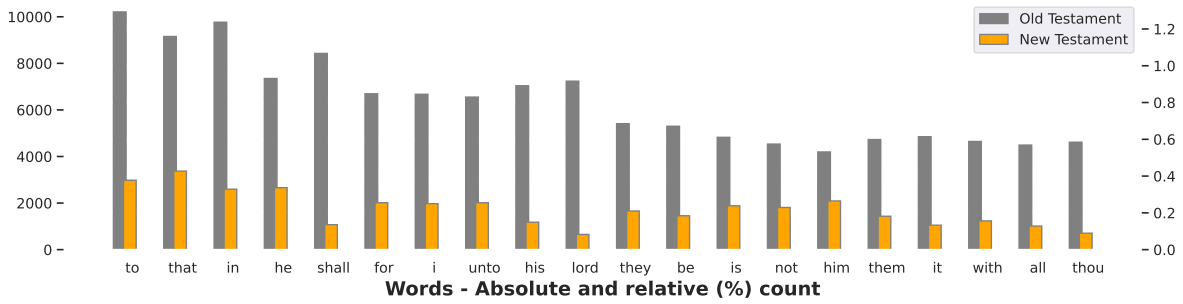
2629 / 464

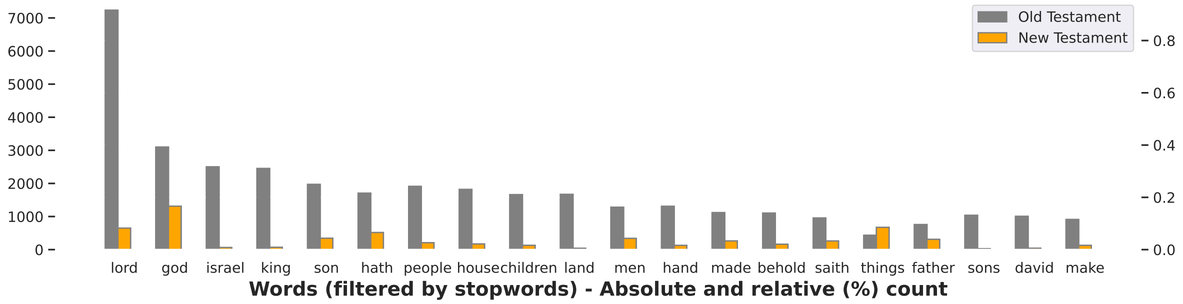
Top 10 Old Testament words	**Top 10 New Testament words**
hosts (299; noun, or verb?)	Jesus (803)
Philistines (250; a nation)	Christ (554)
Joshua (216; a conqueror)	Paul (162; an apostle)
families (174)	Peter (138; an apostle)
Moab (168; a kingdom)	John (107; a name)
Zion (153; a place)	gospel (99)
Manasseh (147; a king)	church (74)
Jeremiah (147; a prophet)	Simon (67; an apostle)
Joab (145; a general)	Pharisees (57; a sect)
statutes (132)	apostles (54)

Notice that many of those are actually names (of either people, or places), in some sense the main characters of the whole book. Naturally, many of the least-used words (used only once or twice throughout) are also real names, participants of random encounters with the main characters.

Distribution of twenty five most common words:

These are the words we are most likely to encounter when given a random verse. If they are unequally distributed between the testaments, they can give us good idea of which testament they belong to. For obvious reasons, we are ignoring the most trivial words (e.g. particles etc. (“a”,“an”,“the”,“and”,“of”)).





Interestingly, all of the overall most common words have very strong majority of their occurences in the Old Testament, too, with the lowest ratio in favor of the Old Testament is for the word “him” and “things”.

Classification process details and results

From all of the verses of the KJB, randomly chosen 20% of them are used as the testing data, the remaining 80% is used as training data for the Bayes. All of the classification metrics are computed on classifying the testing data.

Moreover, I have also added a second dataset, Smith’s Literal Translation (SLT, or JST in some sources) in the same format, which is a translation of KJB into more modern English (1867). From this SLT dataset, randomly chosen 20% are again used for classification as the testing data.

Link:

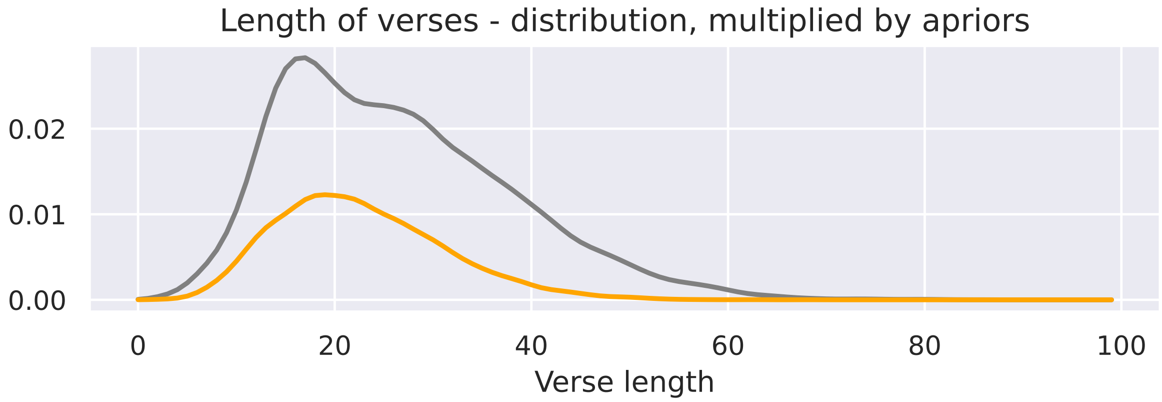
https://openbible.com/texts.htm - Smith’s Literal Translation - Text

Classifiers:

Trivial

Expect every verse to be from the Old Testament. This is based on the observation, that majority of all verses are from the much larger Old Testament.

ERM:



In the density plot, find a threshold s.t. it minimizes the expected error. That means a vertical line, for which the integrals (areas under the density lines) of incorrect classes (on the opposing sides of the threshold) are equal.

In this setting, because of the priors, it leads to the same strategy as the trivial classifier. Let's skip the implementation, then.

Simple

Let's try a simple classifier, which does not require much of computing power, but should provide more accuracy than the trivial classifier.

If a verse contains the word "Jesus", "Christ", or "Apostles", classify it as being from the New Testament, otherwise classify it as the content of the Old Testament.

Of course, other words could be used, but I decided to use these as the most fundamental to the New Testament text.

Bayesian by words

A classifier based on the likelihood of seeing a certain word in either of the Testament. If we see words more likely for the Old Testament, then the verse probably is from the Old Testament.

Lets compute log odds of seeing a word if looking on a New Testament verse, that is:

$$LO(t)_N = log_2(\frac{COUNT(t)_N}{COUNT(t)_O})$$

where $COUNT(t)_N$ is the number of occurrences of token t in the New Testament.

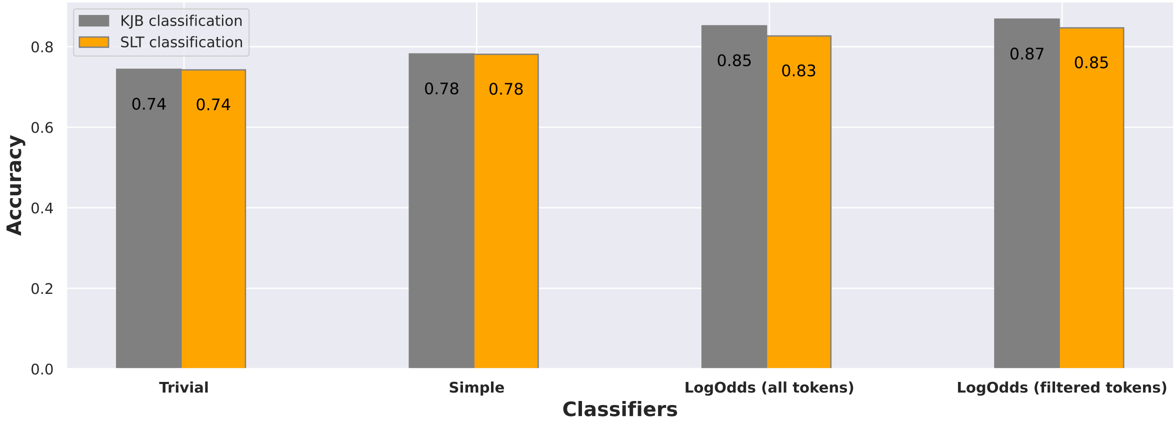
Afterwards, classify the verse as from New Testament if the sum of the log odds per words in verse surpasses 0. That is

$$y = [[\sum_{token \in verse} LO(t)_N > 0 \ \ ]]$$

Finally, there are two outcomes for this classification method, as we test both classification

1. over all words from the verse,
2. and then only filtered words - that means filtering the words of each verse using the list of stopwords explained before.

Classification outcomes



Conclusion

Firstly, even the trivial classifier gives us quite high accuracy. And for both datasets, naturally.

Because of that, accuracies of other classifiers do not rise much with increasing complexity of the models. The best one achieves around 87% of accuracy, which is not bad at all.

Since the Old Testament is so much larger than the New Testament, using the length of the verse does not help with the classification at all. New Testament has different mean value of the verse length, but Old Testament's prior is just too high. This also applies to other possible (similar) attributes.

Surprisingly, however, the classifiers trained for KJB work quite well when extrapolated to other datasets (albeit very similar), too.

Thank you for your time and atention,

Tomáš Kasl