Text Mining

Alun Meredith

1. INTRODUCTION

The corpus analysed is a collection of 24 books about antiquity, written in the 18-19th Century. Many are translations of older texts and often volumes of a larger book. The data is html documents of OCR scans from the Google Books Library Project, separated into each page.

Manually extracting some metadata about each book (title, author, translator, year written, year translated) we can see some themes worth investigating: If it is written about/by a Roman or not, written in/about antiquity, volumes within a series, written by the same author, effect of different translators.

2. PRE-PROCESSING

To extract the text from each html page, xml2[6] was used. Using html tags to isolate the second paragraph (class = 'ocr_par'), which contained the body of text without the title. The 'ocr_cinfo' class contained each word separately to avoid an error where words across newlines were merged with a naive approach. Pages were concatenated to produce books.

Using the 'tm' package[1] a raw corpus was built from these vectors. After casefolding; whitespace, non-alphabetic characters and stopwords were removed. Stopwords removed were tm's "english", numbers and common artefacts of the OCR. A regular expression was used to remove Roman numerals and words were stemmed.

To produce a term-document matrix (tdm) of 1-3 ngrams: sparse terms where removed, a restriction that terms must occur at least 5 times was introduced, to reduce the effect of random occurances of words and reduce size of the tdm. Term frequency - inverse document frequency was used to weight terms based on their background frequency, although our corpus is a biased judge of background frequency considering large portions written by same few individuals.[1] It is important that the document term matrix is normalised, taking into account the different sizes of the documents. Finally a cosine dissimilarity matrix was computed. Many of the algorithms act directly upon this matrix rather than the tdm.

3. ENTROPY

The Shannon empirical entropy [5][2] was computed on each book using the "entropy" package before the stopwords had been removed using the entropy package conducted on corpus without stop word removal.

Within the 5 terms with the highest entropy are the dictionary/encyclopedia style books which are expected to have the highest information density. In addition volumes of the same book are generally ranked close together with the exception of Josephus IV and Tacitus History IV, the later of which is quite a remarkable outlier. Cases where the translator changes for volumes of the same book doesn't have a noticeable impact (E.g. Decline & Fall 4 vs. 1,3,5). This suggests there isn't a strong component of things being translated in totally different ways.

Figure 1: Shannon Entropy of each document, coloured by book each volume belongs to. Full names on github

4. CLUSTERING

Agglomorative hierarchical clustering was evaluated against the cosine dissimilarity matrix using the cluster package [3][4](fig. 2c). "Ward" method was used although other methods were tested with the same results (except single linkage).

Unlike the entropy measure which grouped volumes of a book together well, hierarchical clustering with TF-IDF is grouping on topic rather than language style. There are 3 main clusters, looking at their highly weighted terms we can see that they can be described as **Jewish**: Jew, Herod, etc; **Non-Roman**:Arab, Constantinople, Peloponnesan; **Roman**: loosely clustered.

These clusters can be unstable to small changes, upon identifying and merging two spellings of constantnopl Livy I and Josephus I joined the non-Roman cluster. Decline & Fall I and Tacitus Hist III are linked solely from the trigram "see geographic table".

Before TF-IDF many highly weighted words are similar to stopwords such as "there". However there are useful defining words highly weighted as well such as "citi" and "town" under the geographic dictionary [2]. Therefore an attempt was made to access the language style of the author (differing use of common terms)

[1]Different levels of sparsity used for some purposes, value of 0.95 used if not stated, i.e. term must be present in at least 2 documents.

[2]A list of the 10 highest weighted words being referred to for different weighting schemes and ngrams on github.com/alunmeredith

more by restricting terms to those that appear at least 100 times in the document and reducing the sparsity so they appear in at least half of the corpus. Before computing hierarchical clustering as before. This didn't yield any dissernable patterns and the largest cluster was grouped around the term "this".

4.1 K-Means & PCA

K-means clustering requires number of clusters to be dictated, although we believe there are 3 clusters from figure 2c an elbow plot helps make this decision (fig. 2a). This shows in the "elbow" of the curve the place where the spread of the clusters is minimised with respect to number of clusters.

After computing K-means clustering using the Partitioning Around Medoids (PAM) algorithm in the cluster package[3][4]. PAM is a similar algorithm to K-means but minimises sum of dissimilarities instead of euclidean distances. This makes it appropriate to use with our cosine dissimilarity based document vectors.

PCA plots the three main clusters from hierarchical clustering quite distinctly. Within the PCA the documents in cluster 2 are very densely packed but previous comments have been in reference to the other (relatively sparse) clusters. There is likely a unifying theme within these texts which is more spread out and consistent than the 'constantinopl' style themes previously identified.

PAM is relatively consistent with these clusters but includes Tacitus Annals in cluster 3 and Geo Dictionary, Tacitus Hist 3 and Livy 3 in cluster 1. Cluster 2 is very dense so we suspect the algorithm finds it difficult increasing the size of this cluster to accommodate only one or two additional points. E.g. the width of cluster 2 would have to be doubled to accomodate Geo Dictionary.

5. CONCLUSION

There are two main aspects in which this work can be clustered, on the language/style used within the document or on the topic. We have seen clustering and dimensionality reduction techniques, especially using TF-IDF can extract topical information whereas entropy better groups language style of volumes within a book. We also saw how TF-IDF clustering was sensitive to singular words (e.g. Constantinople or 'see geographic table').

Further work on this type of clustering would include use of a thesaurus and more advanced class equivalence techniques to try to reduce this sensitivity. Additionally as TF-IDF weights names so heavily (due to their sparsity in documents they are not important in) further work proposed[3] is to use these bigram terms to query dbpedia, if matching a person then extract a year of death/birth and plot the distribution of these for each book.

We have seen no clear patterns to distinguish works written in vs. about antiquity or effects of different translators of volumes of the same book because language style proved difficult to access.

6. REFERENCES

[1] I. Feinerer and K. Hornik. *tm: Text Mining Package*, 2015. R package version 0.6-2.

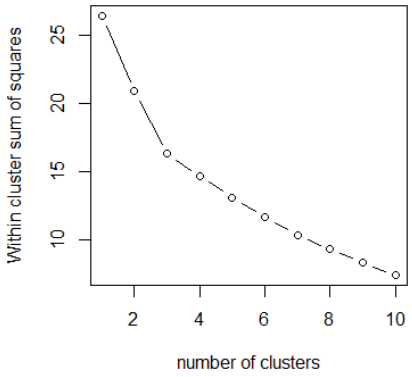
[2] J. Hausser and K. Strimmer. *Entropy*. R package version 1.2.1.

[3] L. R. Kaufman and P. Rousseeuw. Pj (1990) finding groups in data: An introduction to cluster analysis. *Hoboken NJ John Wiley & Sons Inc.*

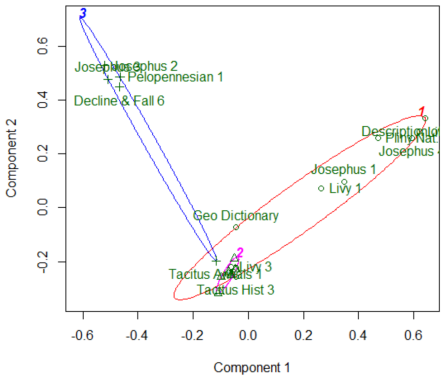
[4] M. Maechler, P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik. *cluster: Cluster Analysis Basics and Extensions*, 2015. R package version 2.0.3 — For new features, see the 'Changelog' file (in the package source).

[5] C. E. Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

[6] H. Wickham. *xml2: Parse XML*, 2015. R package version 0.1.2.

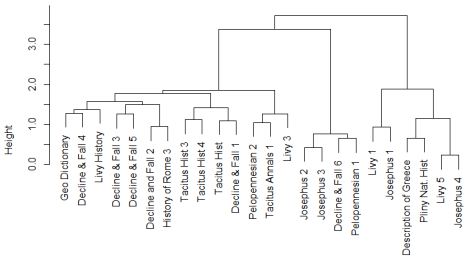


(a) Elbow plot varying number of clusters vs. within cluster spread



(b) K-means clustering plotted on First two principal components. Using cluster package in R. Documents in clusters 1 and 3 are labelled[a]

[a]Other dimensional scaling techniques such as HDS also computed but with little additional information gained w.r.t. PCA



(c) Dendrogram of agglomerative hierarchical clustering using "Ward" method and cosine distance. Produced from cluster package in R.

[3]started but unable to complete in a timely fashion