Exploratory Data Mining of Literature Describing Antiquity

Gregory Walsh
gw2g17@soton.ac.uk

ABSTRACT

This paper describes a method for generating document embeddings for use in exploratory data mining analysis. Using this embedding method, a small set of documents concerning classical antiquity was analysed using a variety of clustering techniques. Possible explanations for the clusters found are then presented.

1 INTRODUCTION

Finding suitable representations of documents for the purposes of performing information retrieval and document classification is an ongoing area of research. Simple representation schemes such as bag-of-words (BoW) and term frequency-inverse document frequency (tf-idf) can be used to encode documents as vectors in a term space, and from there, clustering techniques can be applied for the purpose of understanding inter-document relationships. In this paper, an embedding method using a word2vec model [3], which was found to produce more insightful results when combined with typical clustering techniques than BoW or tf-idf representations (with and without stemming and lemmatisation), is presented. Furthermore, an analysis of the relationships between a set of 24 documents on ancient Rome and Greece is given, based on the visualisations generated using clustering techniques on document representations generated using this word2vec based model.

2 APPROACH

Word2vec models generate vector representations of words which can be manipulated with standard arithmetic operations [3]. For example subtracting the vector for "man" from "king" and adding the vector for "woman" yields a vector which is closest to that of "queen". By taking advantage of this behaviour, a holistic semantic embedding $\boldsymbol{y}_d$ for a document d may be calculated as follows:

$$\boldsymbol{y}_d = \sum_{t \in T_d \cap W} f_{t,d} \boldsymbol{x}_t \tag{1}$$

where: $\boldsymbol{x}_t$ = the word2vec vector representation of term t
$f_{t,d}$ = number of occurrences of t in document d
T_d = the distinct set of terms in document d
W = the set of terms in the word2vec dictionary

A limitation of this approach, as with BoW and tf-idf, is that the order of words has no effect on the vector representation, so inevitably much of the meaning of the documents is lost.

To generate representations for the 24 documents, first textual data for each was extracted from the HTML files. Document embeddings, referred to here as mean semantic embeddings (MSEs), were then generated using equation 1. Each MSE was then normalised to emphasise semantic similarity over document length when performing comparisons based on distance. Since the number of dimensions required to describe a document in this embedding is much lower than with BoW or tf-idf, it may explain the cleaner clustering which was observed [2]. In the case of the pre-trained model used here [4], the embedding space comprised just 300 dimensions. In other words, documents are represented conceptually, rather than on a term by term basis.

K-means was the first clustering method performed on the data, using the normalised MSE representations of the documents. An optimal number of clusters k was determined by generating many clusterings for $k \in [1, 23]$, selecting the run with the smallest within-cluster sum of square errors for each k, and from these runs identifying the model with the largest silhouette coefficient [5]. This gave $k_{optimal} = 6$ which was backed up by the elbow method [1].

When performing hierarchical clustering, the MSE representations (more so than other representations) showed well separated clusters and super clusters, as shown in fig. 1. Ward's linkage method [7], with a Euclidean distance measure, was found to produce similar clusterings when compared to the k-means cluster method with six clusters, and was selected on this basis.

Across a variety of multidimensional scaling methods tested, t-distributed stochastic neighbour embedding (t-SNE) [6] gave the most clearly defined clusters, see fig. 2. The method was run several times with an increasing number of iterations (interval 500), from the same initial state, controlled with a constant random seed, until no further convergence was observed (3500 steps). Equivalent experiments were run with different random seeds, most produced extremely similar results. In all cases, the perplexity of the algorithm was set to 6, in line with the number of clusters determined by the k-means analysis. Using Euclidean distance, exact values (rather than estimates) of the gradient were computed since the number of documents is small.

3 RESULTS AND DISCUSSION

Figures 1 and 2 provide evidence that the MSE method can generate representations which, when investigated using clustering techniques, give insight into the relationships between documents.

From fig. 1, we see the texts written by the ancient historian Livy, which describe one thousand years of Roman history up to his death in 9BC, and those by two other ancient authors, Thucydides and Tacitus, are grouped together. It is possible that the overlap in time of the periods they describe, the first hand experience of the authors, and the colourful writing styles they share influence the spacial positioning of the documents and therefore cause them to cluster together. The group also includes two documents from "History of Rome", a modern work by Mommsen, likely a result of the subject matter it shares with Livy's later works.

The next most closely related cluster to the red group, after the ancient works of Josephus, which describe joint Roman Jewish history, is the the yellow cluster, which includes all volumes in the corpus from Gibbon's "The History of the Decline and Fall of the Roman Empire". Since Gibbon's works were written in the nineteenth century, from second hand sources rather than from first hand experience, and because they focus on a later period of Roman history, this may explain the separation of these clusters. In green,

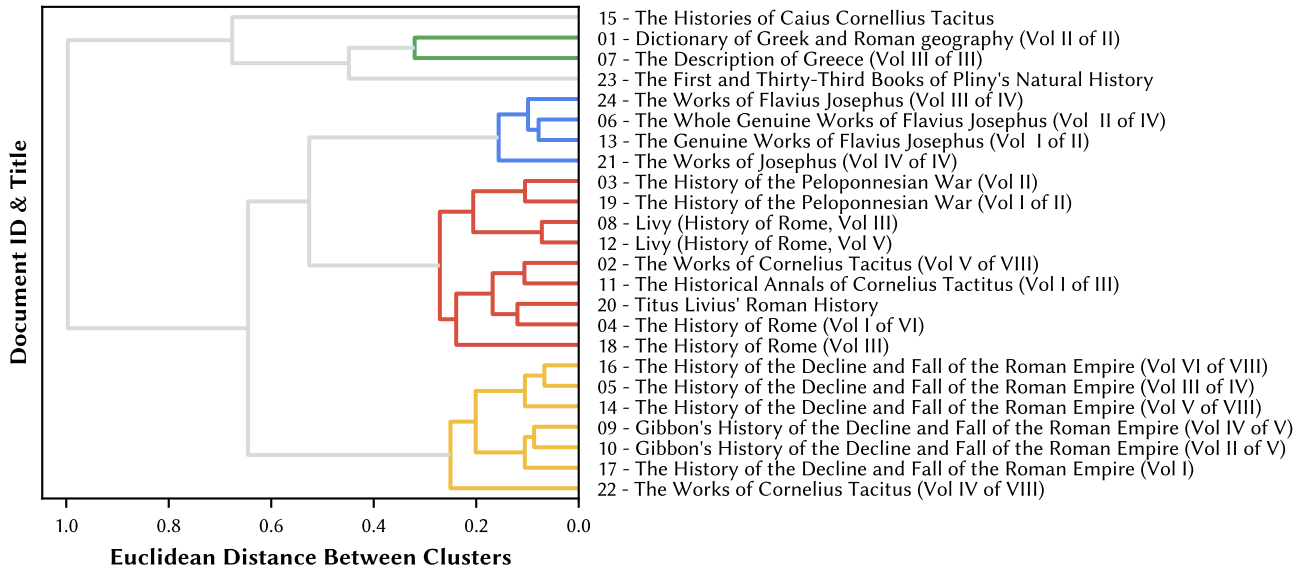

Figure 1: Hierarchical clustering of documents, as represented by the MSE model, using "Ward's minimum variance method"

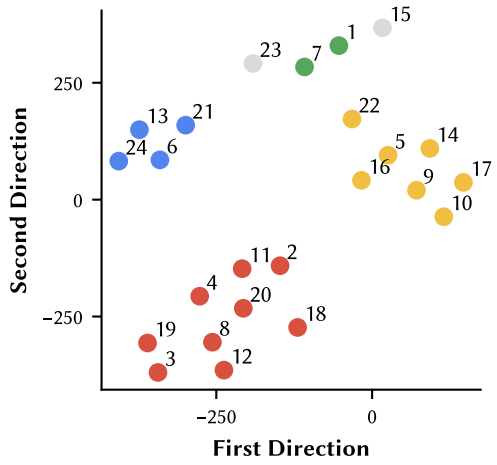

Figure 2: Multidimensional scaling of MSE data from 300 dimensions to 2 using t-SNE (3500 steps, perlexity set to 6). Numbering and colouring shared with fig. 1.

"The Dictionary of Greek and Roman Geography" is linked with Pausianus' "Description of Greece", and at the level above, with the excerpts of Pliny's "Natural History". Unlike the previously mentioned political and historical narratives, these titles describe the natural world, so it is encouraging that they are also quite closely grouped.

Considering that most other authors' works are grouped together, it was surprising to see "15 - The Histories of Caius Cornellius Tacitus" separated from other works by Cornellius. However, upon further inspection, it was found this book comprises the original Latin with supporting English notes, unlike the others which are translations. This has likely occurred because the pre-trained word2vec model used [4] was trained on news articles written in English, rather than bi-lingual documents, hence no embedding of Latin words exist in the model. On a related note, since many proper nouns are not included in the model, documents describing similar concepts are likely to be grouped together, even if the particular actors and places described in the documents are different.

From these observations, it appears that the origins, styles, concepts and narratives of documents all influence their representations. However, one could look at this combination of influencing factors as a limitation. In future work, if one were interested in generating purer representations of the subject matter (and thereby reduce biases towards writing style which may have caused the clustering of Thucydides' and Livy's works), it might be possible to find two generalised vectors representing ancient and modern texts, and subtract these averaged representations from individual document representations according to their origins. The inverse could also be examined, with a view to group by style over subject.

REFERENCES

[1] M.S. Aldenderfer and R.K. Blashfield. 1984. *Cluster Analysis.* Number no. 44 in Cluster Analysis. SAGE Publications. https://books.google.co.uk/books?id=ZIARBoJQxzcC

[2] Pedro Domingos. 2012. A few useful things to know about machine learning. *Commun. ACM* 55, 10 (2012), 78. https://doi.org/10.1145/2347736.2347755 arXiv:cs/9605103

[3] Google. [n. d.]. Google Code Archive - word2vec. ([n. d.]). https://goo.gl/TPWj9r

[4] Google. [n. d.]. GoogleNews-vectors-negative300.bin.gz. ([n. d.]). https://goo.gl/FJV4Qd

[5] Peter J Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20, C (1987), 53–65. https://doi.org/10.1016/0377-0427(87)90125-7 arXiv:z0024

[6] L J P Van Der Maaten and G E Hinton. 2008. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research* 9 (2008), 2579–2605. https://doi.org/10.1007/s10479-011-0841-3 arXiv:1307.1662

[7] Joe H Ward. 1963. Hierarchical Grouping to Optimize an Objective Function. *J. Amer. Statist. Assoc.* 58, 301 (1963), 236–244. http://www.jstor.org/stable/2282967http://www.jstor.org/http://www.jstor.org/action/showPublisher?publisherCode=astata.