How Does the State Speak about Globalisation? A Quantitative Text-Mining Approach

Justin Murphy

University of Southampton

j.murphy@soton.ac.uk

Abstract

Scholars argue that the concept of "globalisation" is strategically deployed by governments to rationalise their actions (Hay and Rosamond 2011). This article is the first large-scale quantitative assessment of this argument, using text-mining and machine learning techniques to analyze more than 60,000 government web pages. Specifically, this article exploits the newly released United Kingdom Government Web Archive to analyze a random sample of web pages published across the entire UK government web system between 2000 and 2013.

I requested 150,000 web pages, received 67k and about 1k were errors. Thus, the final sample consists of a corpus of dtm

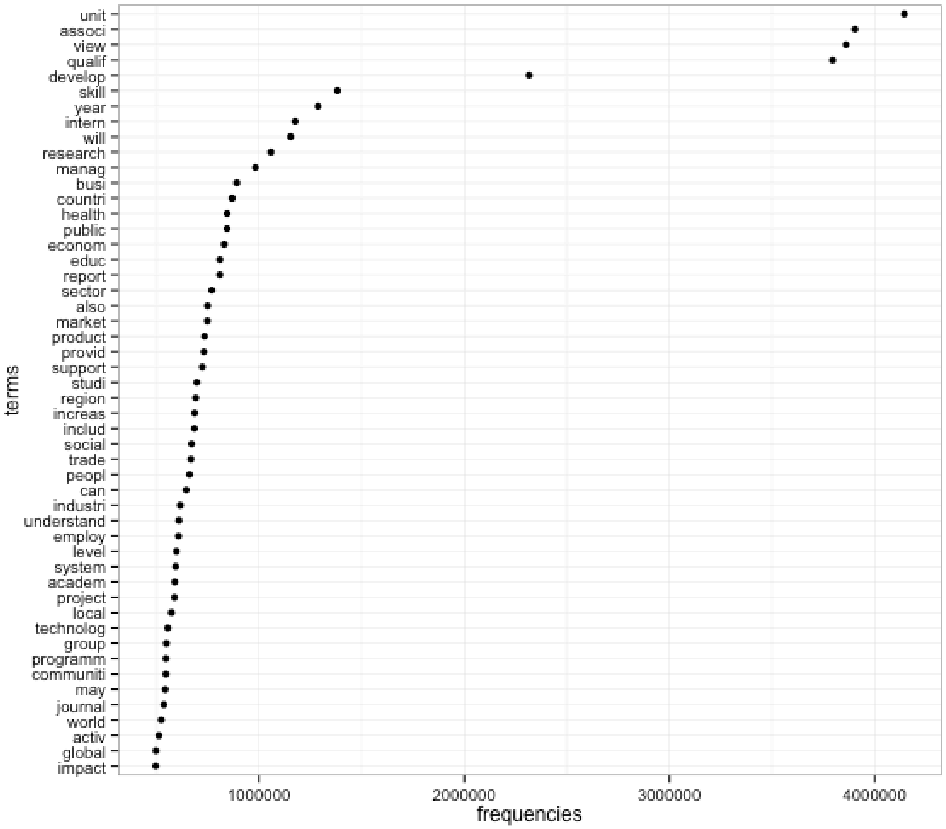

Figure 1: Most frequent terms

Descriptive Statistics

Terms	Correlation
world	0.78
countri	0.66
economi	0.63
threat	0.62
increas	0.61
goal	0.60
key	0.60
agricultur	0.59
also	0.59
develop	0.58
particular	0.57
econom	0.55
mani	0.55
privat	0.55
coher	0.54
competit	0.54
primari	0.53

Terms	Correlation
educ	0.52
exampl	0.52
howev	0.52
integr	0.52
success	0.52
capac	0.50
dfid	0.50
millennium	0.50

Correlated terms

Cluster Analysis In this section, I use *k*-means clustering to partition the corpus of documents into clusters of relatively similar documents. The *k*-means algorithm, also known as Lloyd's algorithm, is a non-parametric technique for partitioning *n* observations into the *k* clusters which minimize within-cluster variance.[1].

K-means cluster analysis requires the analyst to define *k* in advance. As the number of clusters is typically not known in advance, the analyst executes the algorithm with several different values for *k* and compares the within-cluster sum of squared error for each. The *k* which results in the clusters with

[1] Specifically, within-cluster variance refers to the within-cluster sum of Euclidean distances from the centroids, or simply within-cluster sum of squared error (SSE)

the lowest SSE, or is not significantly improved by additional k, is selected as the optimal k. This procedure showed that the 66,400 documents are optimally partitioned into about 200 clusters.[2]

[2] See the Appendix.

Appendix

Diagnostics

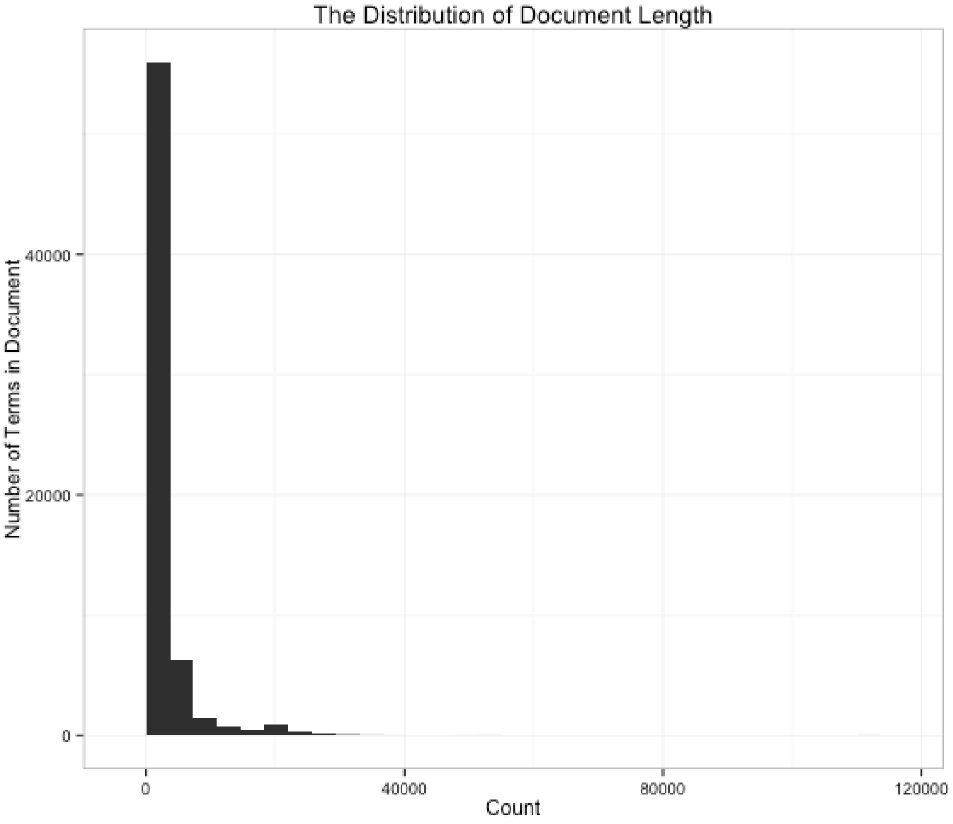

Figure 2: Document lengths

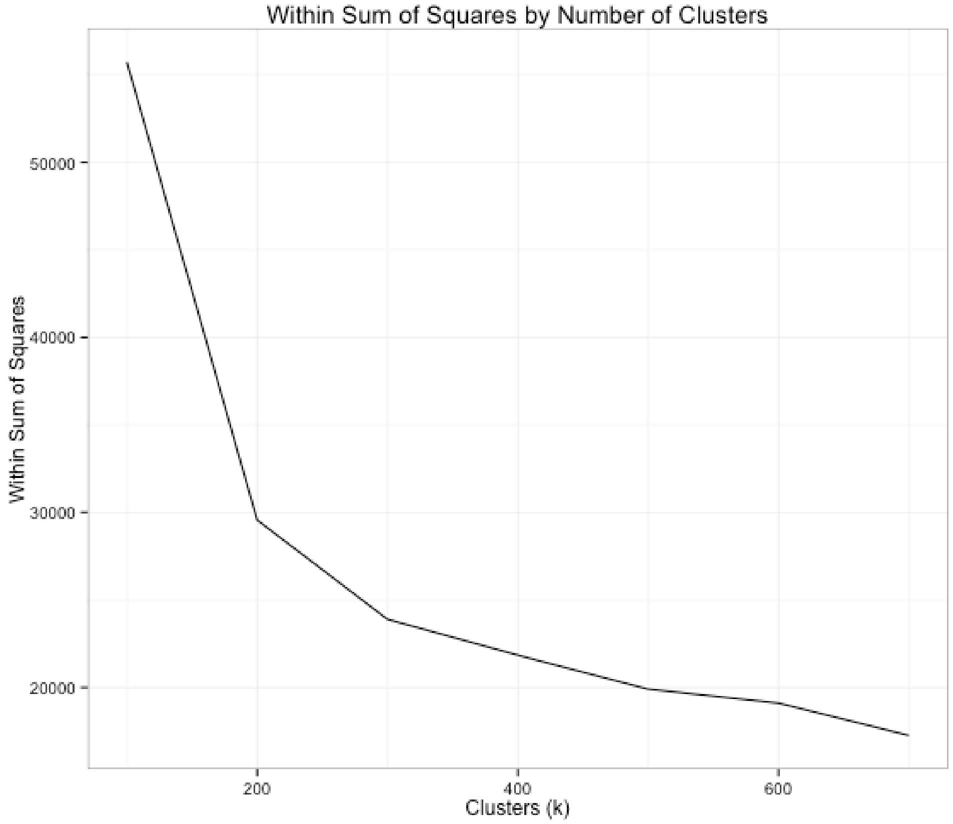

Figure 3: plot of chunk Cluster-Diagnostics

References

Hay, Colin, and Ben Rosamond. 2011. "Globalization, European Integration and the Discursive Construction of Economic Imperatives." *dx.doi.org.libproxy.temple.edu* 9(2): 147–67. http://www.tandfonline.com/doi/abs/10.1080/13501760110120192.