

ONLINE APPENDIX

Matthew Denny and Arthur Spirling “Text
Preprocessing For Unsupervised Learning:
Why It Matters, When It Misleads, And What
To Do About It”

Online Appendix A Google Scholar Results

To investigate the use of supervised and unsupervised methods for text analysis in Political Science over time, we collected data from Google Scholar. Google Scholar allows users to search for the number of results containing a key term in a particular year, thus giving us a sense of the use of a term in academic research over time. We collected data on five search terms over the past 9 years (since the first Wordfish results appeared on Google Scholar) to examine trends related to supervised and unsupervised learning. Figure 1 depicts the relative increase in the number of results returned by Google Scholar (with the number of results for each term in 2008 used as the baseline for that term) over time between 2008 and 2016.

We included three general terms in our search (“Supervised Learning”, “Unsupervised Learning”, and “Text Analysis”). As we can see from Figure 1, the growth in the use of these three terms tracked closely together over time. While these terms appear in papers published in a wide range of fields, they serve as a good baseline against which to compare changes in the political science literature. To examine that field specific part, we selected two unsupervised models prominent in the Political Science literature (“Topic Model”, and “Wordfish”). As we can see, the use of these key terms increased at a much higher rate over the time period than the baseline terms. These results are far from exhaustive, but they demonstrate the growth in importance of unsupervised methods in Political Science and in text analysis more broadly. We feel that they highlight the importance of taking preprocessing seriously.

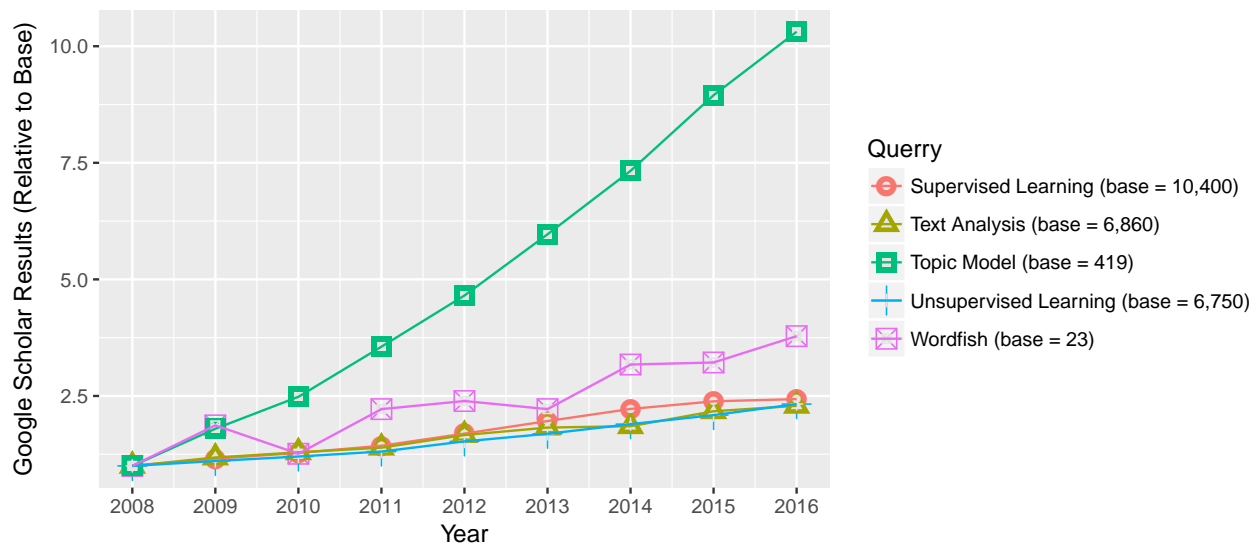


Figure 1: Google scholar results.

Online Appendix B Why Preprocessing Matters: An example and Intuition

To see why preprocessing matters, consider the following sentences dealing with Britain’s nuclear defence system, Trident. The first is from the UK Labour manifesto in 1983:

The next Labour government will cancel the Trident programme.

The second is from the same party in 1997:

A new Labour government will retain Trident.

Clearly, these represent very different positions. The question though, is in what ways preprocessing might affect our sense of how different they are. We note, to begin, that the cosine similarity of these snippets is 0.51. The relevant document frequency matrix looks as that in Table 1 (assuming we only lowercase words)

	the	next	labour	government	will	cancel	trident	programme	.	a	new	retain
1983	2	1	1	1	1	1	1	1	1	0	0	0
1997	0	0	1	1	1	0	1	0	1	1	1	1

Table 1: Document frequency matrix with stop words retained, no stemming.

Consider two researchers, *A* and *B*. Researcher *A* decides to remove stop words from the documents—‘and’, ‘the’ and so on—while Researcher *B* keeps stop words in, but decides to stem the words back to their ‘roots’. In this particular case, Researcher *B*’s decision has zero effect on the distance between the documents: this is because, the words that were stemmed (‘government’, ‘programme’) were common to both documents. Table 2 shows the relevant document term matrix: with minor column name changes, it is identical to Table 1.

	the	next	labour	govern	will	cancel	trident	programm	.	a	new	retain
1983	2	1	1	1	1	1	1	1	1	0	0	0
1997	0	0	1	1	1	0	1	0	1	1	1	1

Table 2: Document frequency matrix with stop words retained, and stemming.

What about Researcher *A*? In practice, removing stop words changes the documents in different ways. In particular, the 1983 manifesto had more incidences of ‘the’. With those removed—as pictured in Table 3—the documents now look *more similar* than before. Indeed, the cosine distance between them rises from 0.51 to 0.62. Thus, when Researcher *A* and Researcher *B* are asked how similar the documents are, their conclusions differ. This matters because document similarity is not some abstruse property: in various forms, it is at the core of almost all unsupervised techniques—be they scaling or clustering or something else.

	next	labour	govern	cancel	trident	programm	.	new	retain
1983	1	1	1	1	1	1	1	0	0
1997	0	1	1	0	1	0	1	1	1

Table 3: Document frequency matrix with stop words removed, no stemming.

Online Appendix C Held-out likelihood and Perplexity

Consider a split of documents in a corpus into a training set \mathbf{w} , and a test set \mathbf{w}' . In the case of LDA (Blei, Ng and Jordan, 2003), the predictive distribution for the model is characterized by the document-topic probability matrix Φ , and hyperparameter α (controlling document-topic distributions). The log-likelihood of the held out test set is thus:

$$\mathcal{L}(\mathbf{w}') = \log p(\mathbf{w}'|\Phi, \alpha) = \sum_d \log p(\mathbf{w}'_d|\Phi, \alpha). \quad (1)$$

This log-likelihood of unseen documents can thus be used to compare models, with a higher log-likelihood implying a “better” model. The *perplexity* of a test set is a closely related to its log-likelihood and is defined as:

$$\text{perplexity}(\mathbf{w}') = \exp \left\{ -\frac{\mathcal{L}(\mathbf{w}')}{\text{count of tokens in } \mathbf{w}'} \right\} \quad (2)$$

which is essentially a normalization of the held-out log-likelihood. Perplexity is the most commonly used metric for evaluating topic model fit. It is intractable because calculating $\mathcal{L}(\mathbf{w}')$ is intractable, however approximation methods have been developed (Wallach et al., 2009) and implemented in numerous software packages.

Online Appendix D Replication of Topic Model Results

Figure 2 displays the average percentage of topic top-20-terms which contain the stem of each of five keywords across 40 different initializations of LDA. Comparison to Figure ?? illustrates highly similar results, indicating that the potential instability of LDA is unlikely to be driving our results.

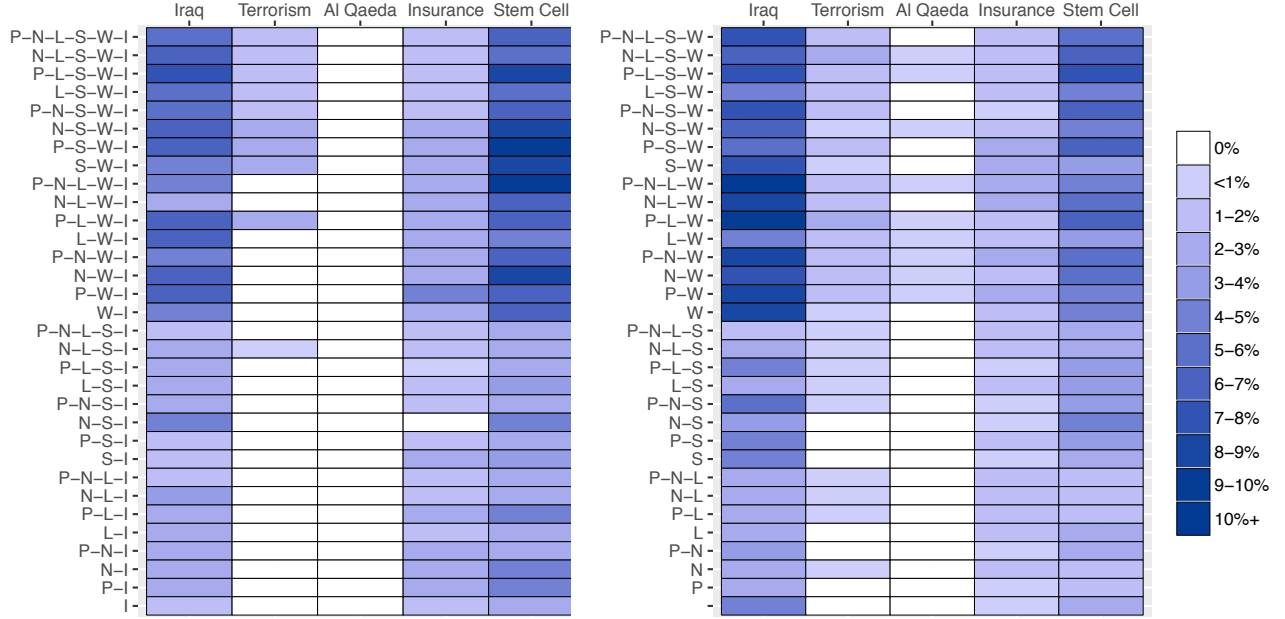


Figure 2: Plots depicting the average percentage of topic top-20-terms which contain the stem of each of five keywords, for each of 64 preprocessing steps (thus excluding those which include trigrams) across 40 different initializations of LDA. The number of topics for specifications fit to each of the 64 DFMs were determined through ten-fold cross validation, minimizing the model perplexity.

Online Appendix E Applying preText to Lowe and Benoit (2013)

In this Appendix, we replicate the Wordfish scaling results from Lowe and Benoit (2013) using the author’s preferred preprocessing specification, as well as model averaging suggested by preText regression results. Lowe and Benoit apply a Wordfish scaling model to 14 Irish parliamentary budget debate speeches from 2009, and then compare the results of their analysis to human expert coding results. The authors are very careful throughout the paper, and place a strong emphasis on validating their results.

The authors selected a relatively standard preprocessing specification of removing all punctuation, numbers, and lowercasing all text (P-N-L). The authors did not stem, or remove stopwords or infrequently occurring words, and did not include n-grams in their analysis. They also noted that they replicated their results with stemming, but this did not change their substantive conclusions at all (something that is backed up by our results). Furthermore, the authors note that they did not remove stopwords or infrequently occurring terms primarily because they did not have a-priori information about which terms might be important to their analysis. We feel that this study represents a case where conscientious and

experienced authors used their best judgement in preprocessing, but did not have the luxury of obvious theoretical guidance for all of their preprocessing decisions.

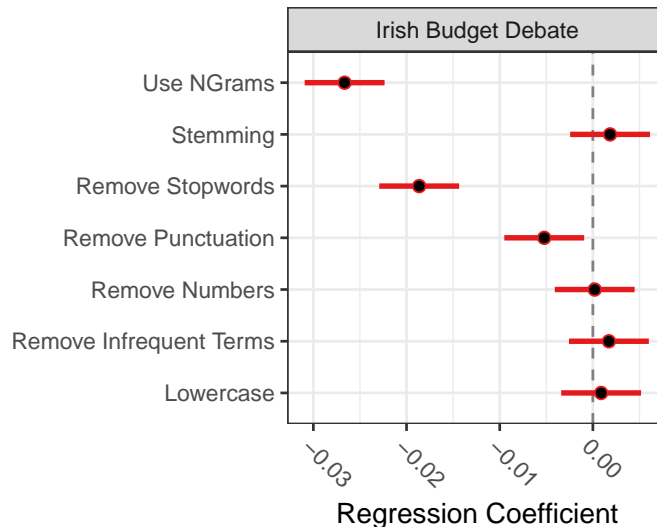


Figure 3: PreText results for 14 Irish parliamentary budget debate speeches from Lowe and Benoit (2013)

To assess the sensitivity of the findings of Lowe and Benoit (2013) to their preprocessing specification, we performed a **preText** regression analysis of the corpus. Regression results are displayed in Figure 3, and indicate that the choices of whether to use ngrams, remove stopwords, and remove punctuation all had significant effects on **preText** scores. While there was a significant effect of including ngrams (or not), we decided to focus our attention on stopwords and punctuation. The choice to include ngrams has not been standard in the literature using Wordfish, and should be further explored in terms of its consequences for the estimation procedure.

Following our own advice to practitioners (see Section ??), we averaged Wordfish estimation results over four possible combinations of preprocessing steps (P-N-L, P-N-L-W, N-L, N-L-W) implied by the **preText** regression analysis (excluding ngrams). The averaged parameter estimates are compared to those from the theoretically justified specification of Lowe and Benoit (2013) in Figure 4. Going by point estimates, we can see that the median legislator is somewhere between OCaolain and ODonnell for both the theoretical and averaged results. But, once we look at the confidence intervals, life is more interesting: for Lowe and Benoit, Gilmore is almost certainly to the ‘left’ of OCaolain, and Ryan is almost certainly to the ‘right’ of Morgan (the confidence intervals do not overlap). But using the averaged results, this need not be the case—because we can switch people’s point estimates around based on uncertainty bands: now Gilmore and OCaolain overlap, as do Ryan and Morgan. While Lowe and Benoit were primarily interested in comparing these Wordfish estimates to human

coding, our results suggest that a researcher could be led to different conclusions from the averaged Wordfish results.

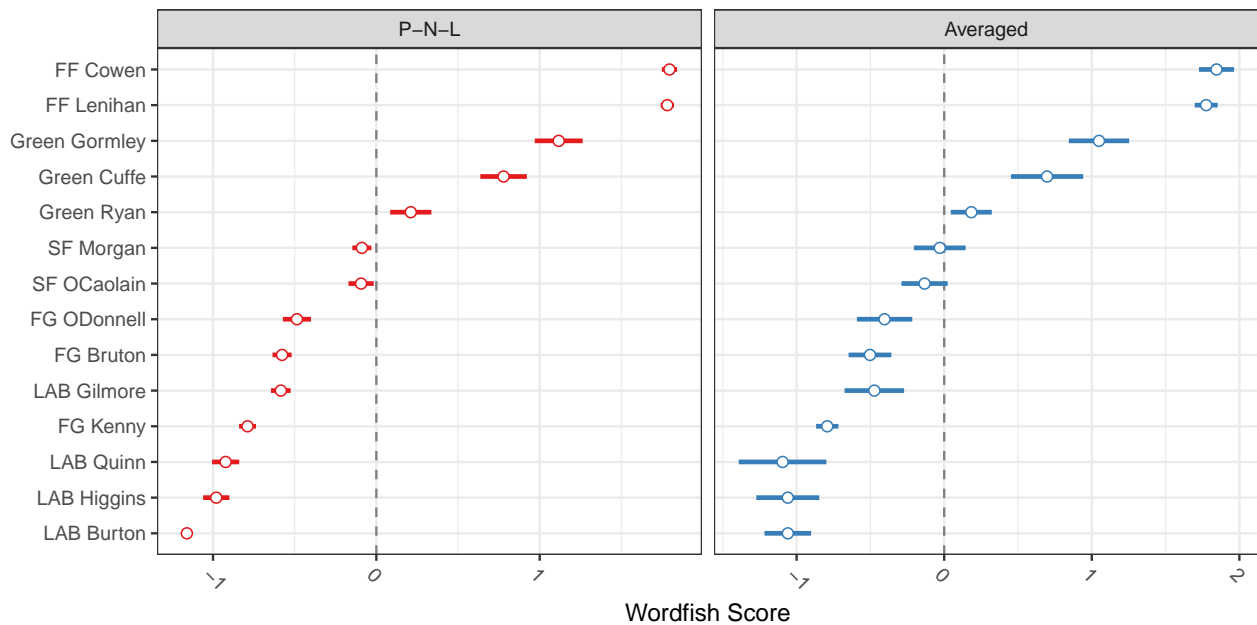


Figure 4: Wordfish scores for 14 Irish parliamentary budget debate speeches from Lowe and Benoit (2013), generated using the authors’ selected preprocessing specification (P-N-L), and averaged across the four possible DTMs generated using stopping (or not), and removing punctuation (or not). These choices correspond to the choices with parameter estimates that were significantly different from zero in Figure 3, but exclude n-grams.

References

- Blei, David M., Andrew Y. Ng and Michael I. Jordan. 2003. “Latent Dirichlet Allocation.” *The Journal of Machine Learning Research* 3:993–1022.
- Lowe, Will and Kenneth Benoit. 2013. “Validating Estimates of Latent Traits from Textual Data Using Human Judgment as a Benchmark.” *Political Analysis* 21(3):298–313.
- Wallach, Hanna M., Iain Murray, Ruslan Salakhutdinov and David Mimno. 2009. “Evaluation methods for topic models.” *Proceedings of the 26th Annual International Conference on Machine Learning - ICML ’09* (4):1–8.
URL: <http://portal.acm.org/citation.cfm?doid=1553374.1553515>