Behavioural Exploration: Topic Modeling and SNR

1 INTRODUCTION

The code for this section can be found at https://github.com/cassiecorey/snr-topic-modeling. For legal reasons, it is a private repository. Please email cjcorey@umass.edu for access.

This exploration examines what effects, if any, modification of the input corpus has on the ability of a topic modeling algorithm to extract meaningful topics. First, we establish a fixed topic model in Section 2. We define several properties of corpora that may be modified during experiments in Section 3. These corpora are loaded, their properties are examined, and the generic topic model is used to extract topics in Section 4. Section 5 defines performance metrics for the topic model that may be affected by changes to the corpora. We calculate the metrics for each corpus and discuss observations in Section 6.

2 A FIXED MODEL

The model is generic Latent Dirichlet Allocation (LDA) with K=100 (number of topics), and both Alpha and Beta equal to 1/K (document_word and document_topic priors respectively). Although an asymmetric document_word prior has been shown to improve performance, we use a symmetric prior for now. Choosing an appropriate K is still somewhat up-in-the-air. Blei et al. state it should be less than the number of documents and there is limited specification on an upper bound. The general consensus has been that larger K "can't hurt." Future work includes generating asymmetric alpha and determining the upper-bound for the number of topics

3 CORPUS PROPERTIES

To characterize each corpus we chose the following properties: number of documents, average document length (in words), vocabulary size, readability, distance from uniform distribution, lexical diversity, and stopword presence. Several of these properties were chosen due to their appearance in related work on topic modeling evaluation and are straightforward (number of documents and average document length). The remainder were chosen due to their general popularity in other fields of natural language processing.

3.1 Measures of Size

We explore several properties that can be calculated directly from the corpus with minimal application of natural language processing techniques: number of documents, average document length, and vocabulary size. Tang et al. explored the effects on performance of modifying the number of documents and the average document length. Their results can be found in [3] and the same exploration is repeated in this work.

3.2 Readability

Readability is the Simple Measure of Gobbledygook (SMOG) index of the corpora. This measure estimates the years of education a person needs to understand a piece of writing. For more information on this measure and how it is calculated see [2].

3.3 Distance from Uniform Distribution

This property was chosen because it was easy to calculate and could potentially lead to some interesting observations. The distance was calculated as a cosine distance between the word distribution of the corpus and a uniform distribution over the vocabulary size.

3.4 Lexical Diversity

This property measures the complexity of a corpus. If a corpus uses similar words to describe the same concept, it will be more diverse than a corpus that repeats the same vocabulary. We expect this to influence the topics extracted from a model because of the fact that topic modeling operates on word frequencies.

3.5 Stopword Presence

Stopwords are words that provide very little meaning to a corpus. They are frequently removed before application of topic modeling because their frequency in a corpus can often push them to the top of each topic in terms of likelihood. The presence of stopwords is calculated as a percentage of the corpora made up of any stopwords from the NLTK list of English stopwords.

4 CHOOSING CORPORA

Coropora were chosen based on their availability with slight attention to differences in properties. Since this is only an initial exploration, only three real-world corpora were chosen. In future work, it will be useful to generate synthetic corpora as well as examine real-world corpora with more variety of properties.

The three corpora chosen were Wine Reviews, Brown, and ABC. Wine Reviews was chosen as a good example of a small corpus with short-text documents. Brown is perhaps one of the most famous corpora. It was the first digital corpora and is regarded as a good example of general American English language. ABC has only two very large documents. Properties of each corpus are shown in Table 1.

CORPUS	NUMBER OF DOCUMENTS	AVERAGE DOCUMENT LENGTH	VOCAB. SIZE	READABILITY	DISTANCE FROM UNIFORM	LEXICAL DIVERSITY	STOPWORD PRESENCE
WINE	1230	25.5	3417	276.148	0.869	0.109	0.261
ABC	2	383405	31885	12.119	0.942	0.042	0.350
BROWN	500	2322.4	66939	11.327	0.971	0.058	0.000

Table 1. Corpus Properties.

5 PERFORMANCE METRICS

Metrics should adhere to the following axioms:

1. *Similar models or topics should have similar scores.*
2. *Different models or topics should have differing scores.*
3. *Scores should change when the model changes.*

We also consider metrics that have been tested in previous work on topic modeling performance evaluation in order to determine their effectiveness. Metrics are designed to be calcuable independent of the model used to create the topics. That is, they rely only on four matrices: document word count, document topic distribution, topic word distribution, and features (the literal words). When possible, the metrics are calculated for individual topics.

5.1 Average Word Length

For a given topic, this metric calculates the average length of a topic's most-likely words. The number of most-likely words is set by default to 20. Figure 1 shows how drastically the importance of words in a topic drops.

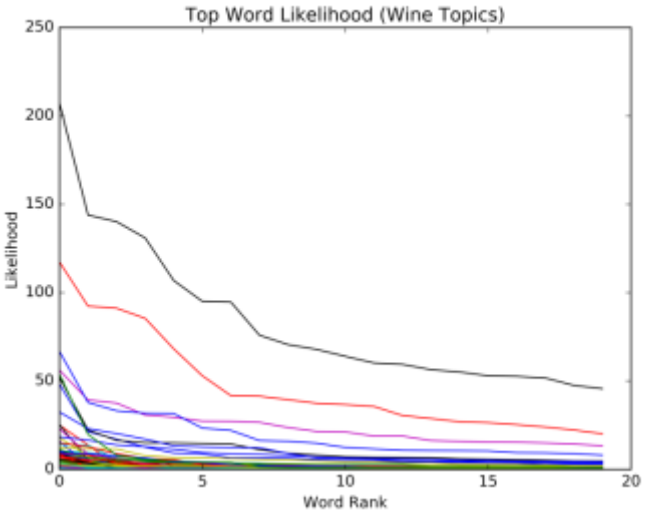

Figure 1. Word likelihood drastically decreases after the first few positions.

These top-words are the most descriptive words of the topic. We would expect topics with a higher average word length to be more meaningful than those with lower average word length as longer words are often more descriptive than short words.

5.2 Exclusivity

Exclusivity measures the extent to which top words do not appear in the set of top words for other topics. It is calculated as the average over each top word of the probability that the word appears in a topic divided by the sum of the probabilities of that word in all other topics.

5.3 Distance From Uniform

This is similar to the corpus property for distance from the uniform distribution. It measures the distance of a topic's word probabilities from a uniform distribution over the vocabulary.

5.4 Rank1

Rank1 calculates the likelihood of this topic being the most popular topic in a document. It requires examining all topics across all documents to compare popularity. If this metric is high, it means the topic is not meaningful because it is frequently found in a lot of documents.

6 RESULTS AND OBSERVATIONS

Figures 2-8 illustrate the results of fitting a generic LDA model to the three chosen corpora and comparing the resulting performance metrics.

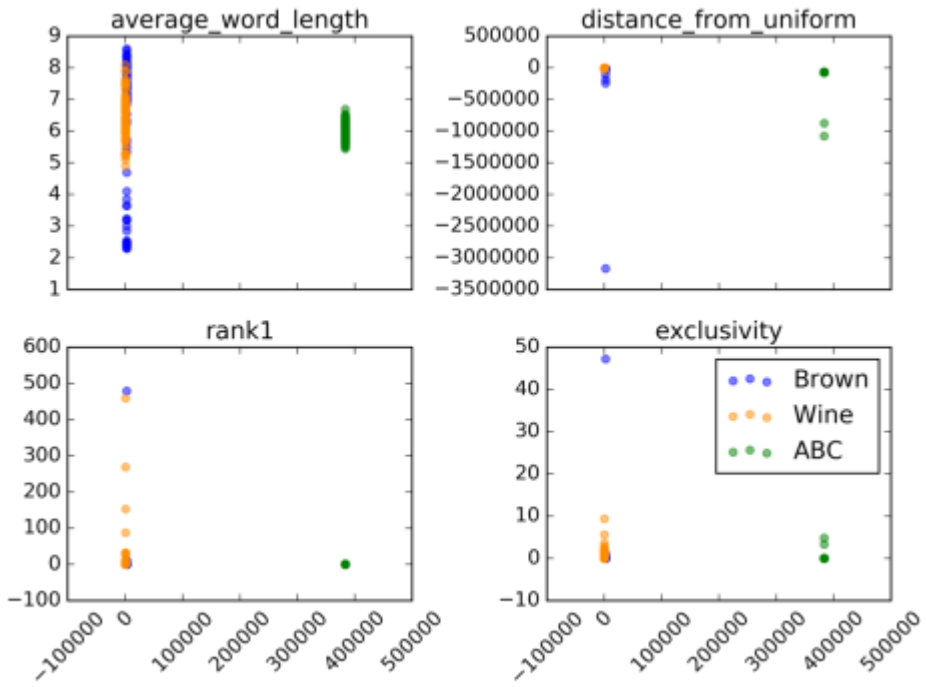

Figure 2. Average document length

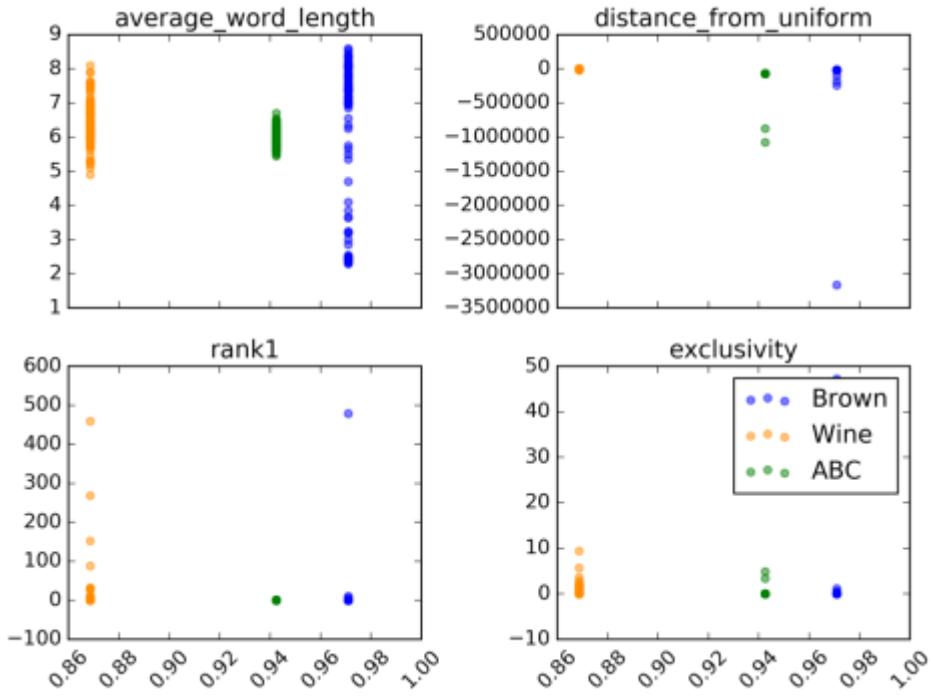

Figure 3. Distance from uniform distribution

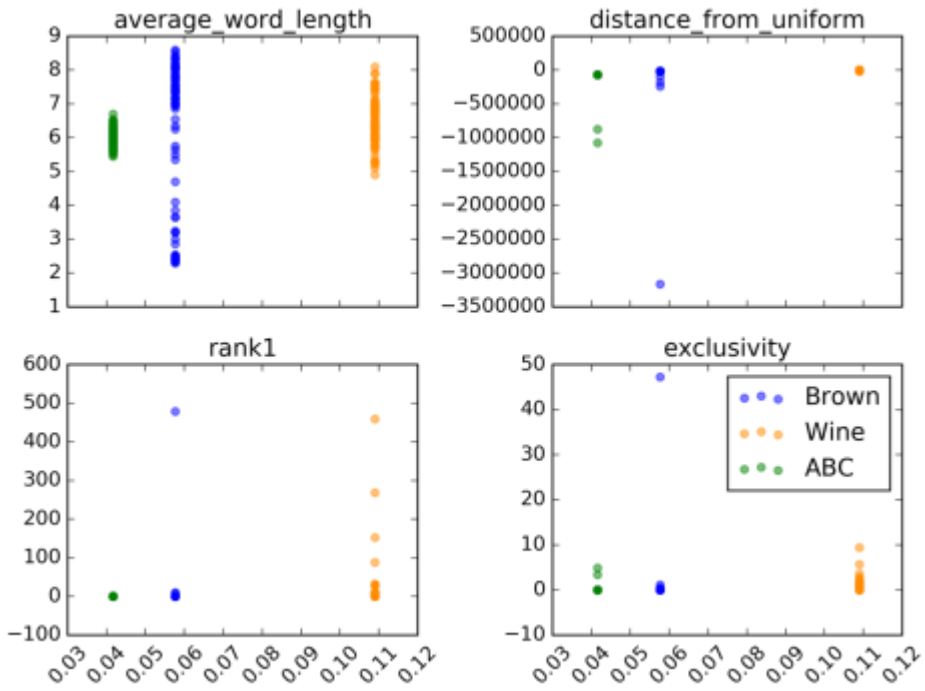

Figure 4 Lexical diversity

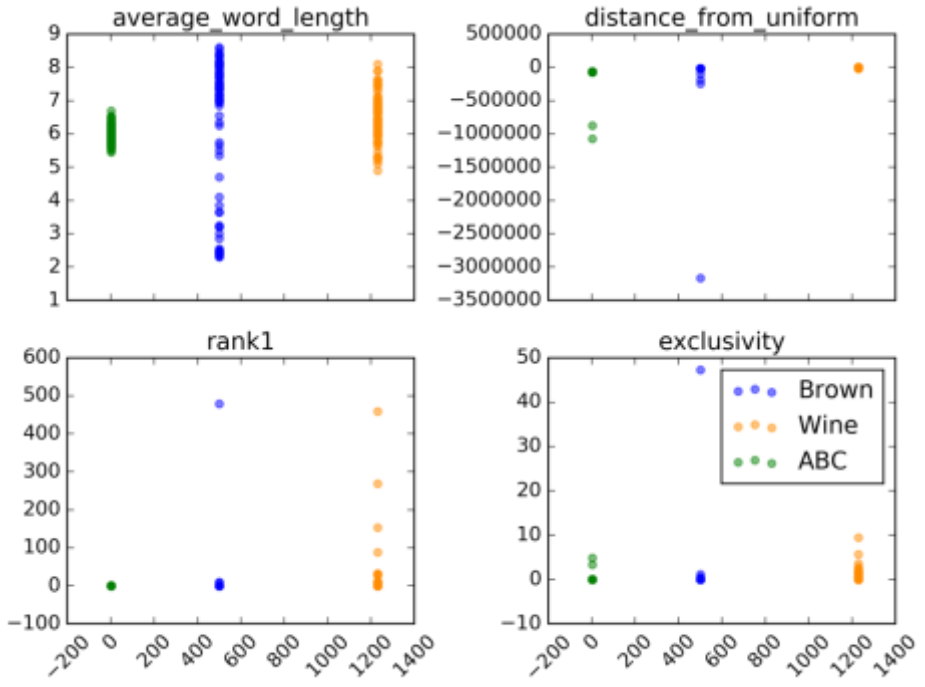

Figure 5 Number of documents

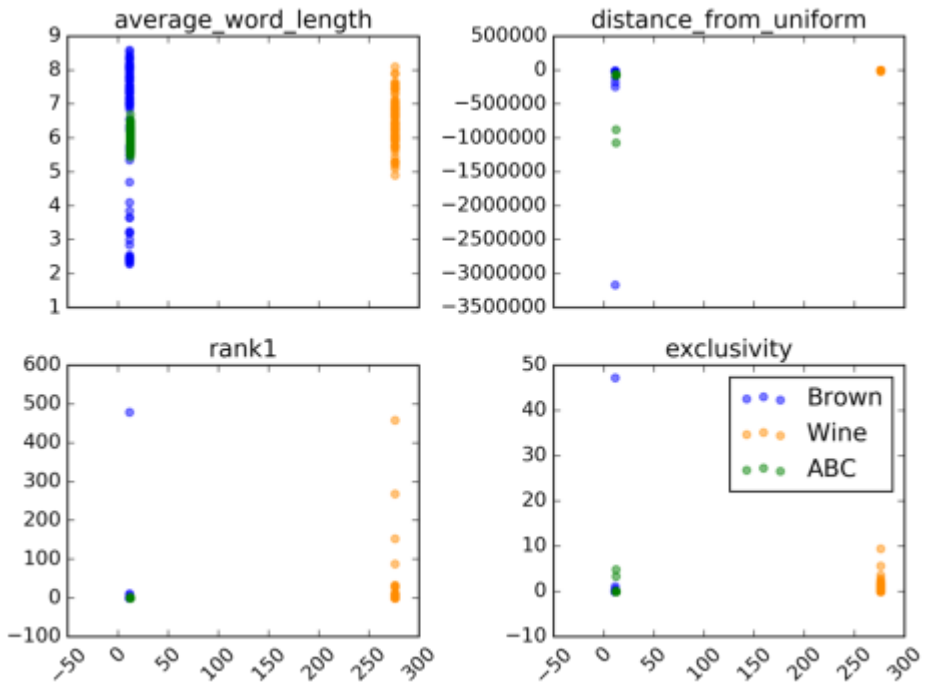

Figure 6 Readability

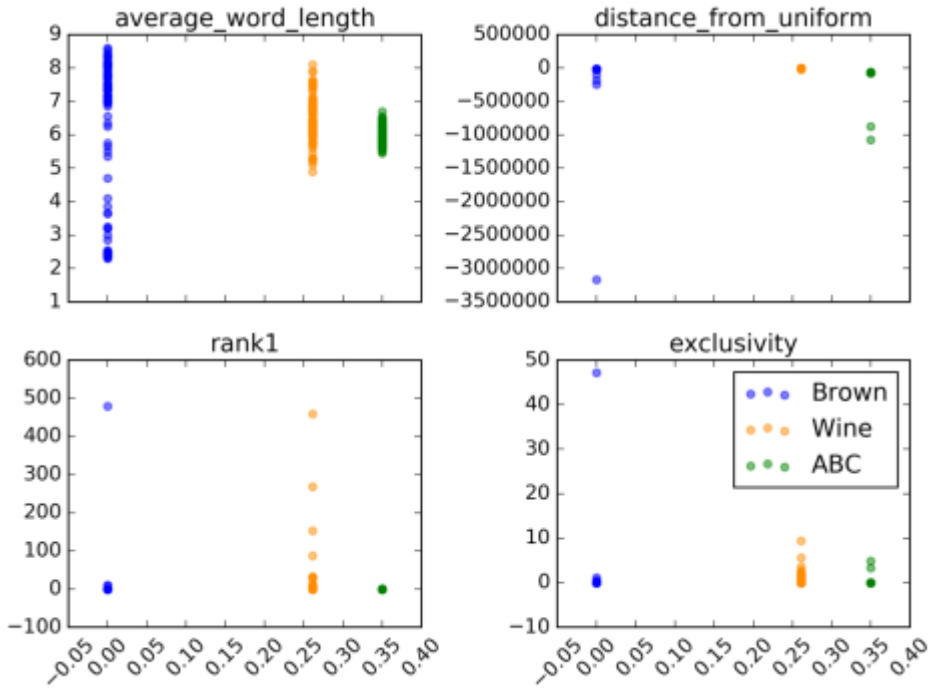

Figure 7 Stopword presence

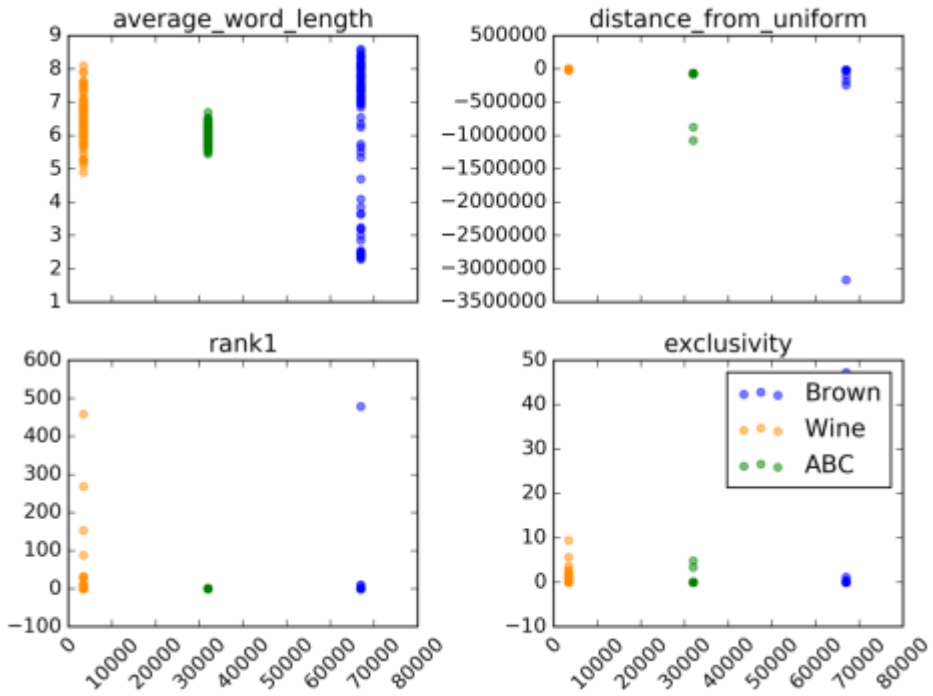

Figure 8 Vocabulary size

[1] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3* (Jan), 993-1022.

[2] Mc Laughlin, G. H. (1969). SMOG grading-a new readability formula. *Journal of reading*, *12*(8), 639-646.

[3] Tang, J., Meng, Z., Nguyen, X., Mei, Q., & Zhang, M. (2014, January). Understanding the limiting factors of topic modeling via posterior contraction analysis. In *International Conference on Machine Learning* (pp. 190-198).