Research Hypotheses: Topic Modeling and SNR

The goal of this project is to determine whether or not modification of the input corpus has a noticeable effect on topic modeling. We chose six measurable corpus properties that can easily be modified to test this question: number of documents, average document length, vocabulary size, lexical diversity, presence of stopwords, and readability. In initial experiments, six corpora were chosen so that a broad distribution of these six properties could be explored. The corpora and their properties are shown in Table 1.

	NUMBER OF DOCS	AVG. DOC LENGTH	VOCAB SIZE	READABILITY	LEXICAL DIVERSITY	STOPWORDS PRESENCE
WINE	1230	25.4943	3417	276.148	0.108967	0.261305
BROWN	500	2322.38	66939	11.3273	0.0576468	0
ABC	2	383406	31885	12.1194	0.0415813	0.349936
GENESIS	8	39408.5	25841	8.46133	0.0819652	0.375455
INAUGURAL	56	2602.41	9754	23.4806	0.0669297	0.449878
STATE OF THE UNION	65	6151.11	14591	17.6455	0.0364937	0.388803

Table 1. Corpora and their properties.

To measure the effect of these properties on topic modeling, we fixed our topic model and chose four topic-specific metrics to measure topic quality which in turn measures performance of the model as a whole. Our fixed approach to topic modeling was Latent Dirichlet Allocation (LDA) with symmetric priors and 100 topics. Because it is a generative model, LDA models topics as a distrubution of likelihoods over the words in the vocabulary. That is, for each word *w* in the corpus vocabulary and for each topic *t* in the provided number of topics, there is an associated likelihood of *w* being chosen to represent *t* whenever *t* is chosen during document generation. This generative approach to topic modeling allows us to calculate the average word length of the most-likely words from each topic, the distance of topics from the uniform distribution over the topic space, and the number of times top words from a given topic were seen as top words in other topics. LDA also provides topic distributions over the given documents. This allows us to also measure the number of times a topic was ranked as the most popular topic in the given input documents. We call this last metric Rank1. Results from initial explorations showed multiple trends between properties and the resulting topic metrics. We converted the most noticeable of these trends into causal hypotheses.

Hypothesis: A more diverse corpus vocabulary correlates with an decreased average distance of topics from the uniform distribution over the vocabulary space.

LDA uses observed word counts in a set of documents together with the document topic and topic word priors to determine word likelihoods for topics. Our experiments are primarily concerned with properties of the corpus as a whole and not properties of the individual documents within that corpus. To test a causal relationship between vocabulary size and increased topic distance from the uniform distribution we would need to control the distribution of the words within each individual document. This is not something we intend to experient with in this project, though it would be interesting future work. Therefore, this hypothesis remains correlational and not causal.

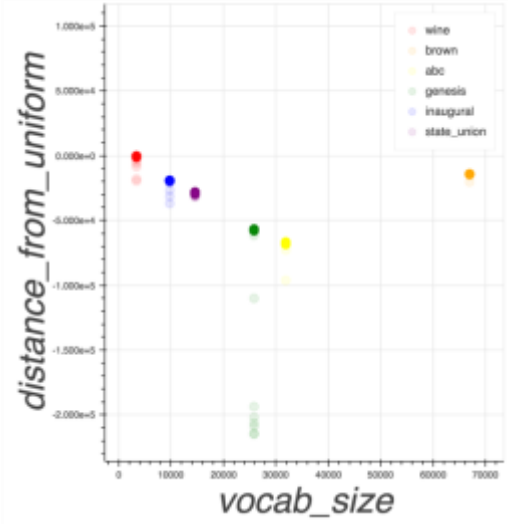

Figure 1. As vocabulary size increases, average topic distance from the uniform distribution decreases.

Hypothesis: Increases in both lexical diversity and readability correlate with a slight increase in average topic distribution from the uniform distribution.

As with vocabulary size, we can only suggest at least a correlation for these properties. The trend was not quite as strong with lexical diversity and readability as with vocabulary. The logical reasoning behind higher readability leading to better topics is hard to ignore. However, LDA is a bag of words model which means it ignores almost every aspect of the equation for calculating readability. These two properties may be causing an increase in the vocabulary size which in turn has a direct effect on the topic distributions.

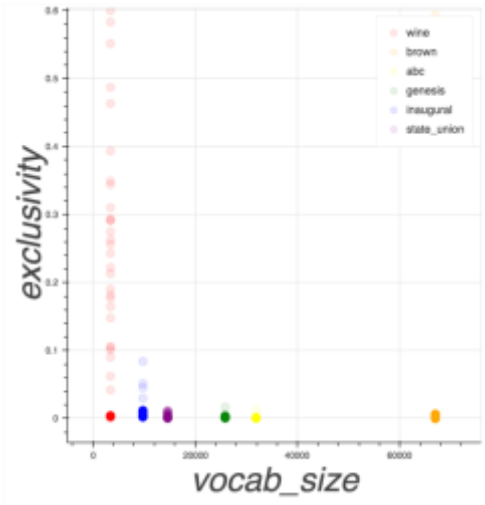

Figure 2. A more diverse vocabulary results in more exclusive topics.

Hypothesis: A smaller corpus vocabulary causes increased topic exclusivity.

This causal hypothesis suggests that as the number of unique words in a corpus increases, the resulting topics will become more exclusive. That is, resulting topics will have less instances of top words also appearing in the top words of other topics. Results to support this hypothesis can be seen in Figure X of the initial exploration. Note, smaller values indicate higher exclusivity. For example, a topic that produces 1 exclusivity is more exclusive than a topic that produces 10 exclusivity.

Hypothesis: As the number of documents increases, the potential for a topic to be seen in more documents also increases, regardless of stopword presence.

This causal hypothesis is straightforward when considering the method of calculating rank1 for topics. The metric is, in effect, another measurement of the number of documents in the corpus since it counts the number of documents for which a given topic is ranked the most popular. For example, if a corpus had a “cat” topic,

and one of the documents was a news article about cat breeds, that document would likely rank the “cat” topic as the most likely topic.

Hypothesis: Removing corpus-specific stopwords directly impacts the quality of topics produced.

This is a causal hypothesis in that it predicts removing the two most-frequently seen topics will directly, and positively, affect the resulting topic metrics. For the rank1 metric, topic quality is expected to increase with a lower score. By removing the two highest scoring topics by rank1 score, and then re-calculating the other metrics, this hypothesis can easily be dis/proved.