Research Design: SNR and Topic Modeling

The aim of this project is to examine the effect of corpus modification on the task of topic modeling. To do this, we explore several causal hypotheses. Each hypothesis involves observing a change in topic quality (i.e. performance of the topic modeling system) as a result of changing the input corpus. Based on preliminary behavioural observations, we believe the most meaningful results will come from corpus modifications that affect the corpus vocabulary (a.k.a the number of unique words in the corpus). This assumption is supported by a common understanding in the literature that word co-occurrence is the underlying phenomenon which produces topics.

First, we will confirm two existing hypotheses related to count and length of documents in the corpus. The first posits that increasing the number of documents will increase the quality of topics produced. The second posits that lengthening documents in the corpus also increases the quality of resulting topics. The driving factor behind topic construction is word co-occurrence (i.e. frequency of words appearing together in a corpus). By varying the number of documents and document size we will be varying co-occurrence indirectly.

We will also test two hypotheses that vary co-occurrence directly by injecting or removing specific words from documents. First, we posit that decreasing the presence of stopwords will result in improved topic quality. Particulary, we predict topic exclusivity will increase since many of the overlapping words between topics are stopwords. Second, we predict that increasing the presence of target words will improve topic quality. Increasing the presence of target words can bee viewed as an alternative approach to decreasing stopwords presence.

The process of testing these four hypotheses breaks down into several components: collecting corpora, measuring corpora properties, varying these properties, finding topics in each corpus, and measuring the quality of these topics.

1 Collecting Corpora

Constructing entirely synthetic corpora is out of the scope of this project. Instead, we will collect real-world corpora, measure their properties, and modify them to suit the needs of this experiment. To collect corpora, we rely on the Natural Language Toolkit (NLTK) [1]. There are several plaintext corpora readily available in the form of Python objects referred to as "corpus readers." Since the properties we wish to calculate are not directly available from the NLTK corpus readers, we built an subclass that we call a "properties corpus reader" (PCR). The PCR allows us to directly query corpora for the various properties we may wish to modify or measure. It also inherits all of the same functionality of the original NLTK corpus reader.

2 Corpus Properties

The corpus properties being directly modified are: number of documents, average document length, and stopwords presence. In addition, we will continue to measure vocabulary size, lexical diversity, and readability. Table 1. shows an example of these properties on the corpora that were examined in our initial exploration.

Corpus Name	# of Docs	Doc. Length	Vocab. Size	Read-ability	Lexcal Diversity	Stopword Presence
Wine	1230	25	3417	276	0.109	0.26
Brown	500	2322	66939	11	0.058	0.00
ABC (rural)	2424	143	17222	12	0.050	0.35
ABC (science)	764	551	24306	13	0.058	0.35
Genesis	8	39409	25841	8	0.082	0.38
Inaugural	56	2602	9754	23	0.067	0.45
State of the Union	65	6151	14591	18	0.036	0.39

Table 1. Corpus properties.

Calculating the number of documents is straightforward. Since our topic modeling algorithm uses a bag-of-words approach, document length is measured in number of words. Stopwords presence is measured as the total number of stopwords (from NLTK's provided list of English stopwords) divided by the total number of words in the corpus. Vocabulary size is the number of unique words in the corpus. Lexical diversity divides the vocabulary size by the total number of words in the corpus. Readability is the SMOG measure of the corpus which is given as an integer representing the years of education needed to understand a piece of writing. The number comes from a unique equation which combines the number of syllables and number of sentences. More detailed information can be found at [4] and Figure 1 gives a visual representation. Essentially, longer words in longer sentences makes a piece of writing more difficult to understand.

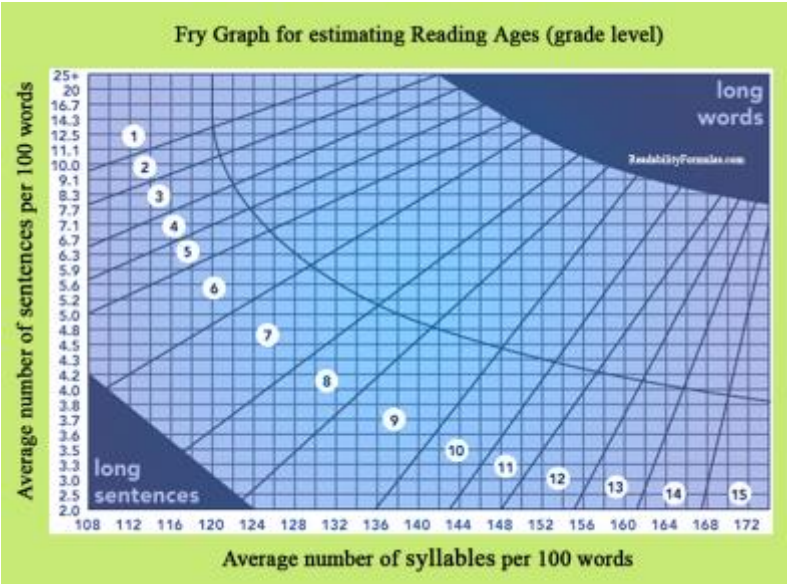

Figure 1. Visualization of text readability based on number of syllables and sentences. [3]

3 Varying Corpus Properties

In order to observe changes in topic quality, we need to make modifications to the independent variables: corpus properties. This is where elements of randomness are introduced. We vary the number of documents in the corpus by making a uniform random selection of *d* documents from the original corpus where *d* ranges from 2 to the true number of documents in the corpus. The number of words *w* in each document is varied by making a uniform random selection of *w* words from each document where *w* ranges from 10 to the average number of words per document in the original corpus.

Modifying the number of documents is straightforward. We simply limit the fileids passed to the corpus reader constructor. To modify the size of documents, however, is a bit more challenging. For this task, we constructed a method that can convert a list of string objects into an NLTK corpus reader object which can then be used to build a PCR.

To alter the presence of stopwords we uniformly remove or inject stopwords at random in the range of 0% to 50% (percent of the corpus that is composed of stopwords). The set of stopwords present in each document of the corpus is extracted using built-in Python tools. Subsets of varying sizes are chosen at random from this list and used as a reference for which words to remove from the document. Since the topic model uses a bag-of-words approach, we do not need to worry about where stopwords are being removed from in the document. We only care that they are removed.

A similar approach is used to inject target words into documents in the corpus. Target words are removed or injected at random in the range of 0% to 100%. If our hypotheses are correct and our methods for evaluating topics are useful, we should observe measurable changes in topic quality as a result of removing or injecting stopwords and target words.

4 A Fixed Approach to Topic Modeling

After calculating corpus properties, we need to extract topics. This component of the process is fixed. We use Latent Dirichlet Allocation with K=100 topics and symmetric document-topic (Alpha) and topic-word (Beta) priors both set to 1/K. We are operating under the assumption that topics converge as K increases beyond the true number of topics in the corpus. Based on related work showing this sort of convergence and the fact that determining the true number of topics in a corpus is an inherently difficult task, we believe this assumption to be reasonable.

5 Measuring Topic Quality

Extracted topics need to be measured for quality. State-of-the-art topic evaluation relies on UCI coherence and the model's perplexity on held-out documents. UCI coherence requires collecting external word co-occurrence data from Wikipedia (or other wiki-type sources) and perplexity requires having/obtaining labels for the true underlying topics. Unfortunately, these tasks are both out of the scope of this project. Instead, we propose several new topic metrics that do not rely on external information or human-labeling. The metrics we calculate are: effective size, exclusivity, rank1, average word length, distance from uniform distribution, and distance from corpus distribution. These metrics are mainly obtained from [5] which provides detailed descriptions of how to calculate each metric as well as how to interpret the value.

Effective size is equivalent to the effective number of parties measure used in politics. It weighs each word in a topic by the words relative likelihood of being found in that topic. The result is that

topics comprised of a few extremely likely words are effectively smaller than topics comprised of many equally likely words. Topics that are more broad will have a larger effective size. Therefore, we say a topic is good if it has a relatively small effective size (when compared to the other topics).

Exclusivity measures the degree to which top words in one topic also appear as top words in another topic. First, we take the likelihood of each top word in a topic and divide it by the sum of the likelihoods of that word in all other topics. The average of this fraction over all top words in a topic is that topic's exclusivity.

In LDA, documents can contain several topics. Rank1 counts the number of documents that rank a given topic as their most common topic. The higher this metric is, the worse a topic is. Topics that are very popular in a lot of documents are probably meaningless because they are so popular. They will not stand out as much as a topic that is only spoken about frequently in a small subset of documents.

Word length is measured in number of characters. The underlying assumption is that longer words are more descriptive. We take the top 20 words from each topic and calculate their average length. Topics with longer words are said to be more descriptive (i.e. "better") than topics with shorter words.

Distance from uniform and distance from corpus distributions are both calculated using Kullback-Leibler divergence. Also known as the relative entropy, Kullback-Leibler divergence measures the information gain of using the topic to construct a document versus using the uniform distribution over the vocabulary space or the observed distribution of vocabulary words in the corpus. We want topics that diverge from the corpus and uniform distributions because that means they stand out and are more descriptive than those that converge.

6 Protocol and Analysis

All experiments are written and executed in Python 3.5. Open-source code can be found at [2].

Analysis of the results This work will hopefully result in several contributions. First of all, we hope to confirm the results of related work which has shown the effects of modifying the number of documents and the average document size. In addition, we hope to extend related work by showing the effects of modifying the presence of particular words in the corpus (stopwords and target words). To our knowledge, this sort of direct approach to modifying a corpus and testing the resulting effectiveness of topic models has not been attempted.

In general, our aim is to show that modifying the input vocabulary does affect the possibility to effectively model topics within a corpus. Due to the design of our experiments, we can observe falsification of the hypothesis in two ways. First, there may be no change at all in the topic metrics when we change the input vocabulary. Second, there may be changes but not in such a way that suggests any sort of causal relationship. Since it is well-understood that the underlying phenomenon of topic modeling is word co-occurrence, it is much more likely that we would observe the latter of the two falsification possibilities (if any). We are almost guaranteed to see changes in the topic metrics since we are going to be making changes (directly and indirectly) to word co-occurrences.

[1] Bird, Steven. "NLTK: the natural language toolkit." *Proceedings of the COLING/ACL on Interactive presentation sessions*. Association for Computational Linguistics, 2006.
[2] https://github.com/cassiecorey/snr-topic-modeling
[3] Figure 1 Illustration: http://www.readabilityformulas.com/graphics/frygraph2lg.jpg
[4] Mc Laughlin, G. H. (1969). SMOG grading-a new readability formula. *Journal of reading*, *12*(8), 639-646.
[5] McCallum, Andrew Kachites. "MALLET: A Machine Learning for Language Toolkit." http://mallet.cs.umass.edu. 2002.