Empirical Results II: Corpus Modification and Effects on Topic Modeling

1 INTRODUCTION

1.1 Primary Goal

The intended goal of this project is to explore the effects of corpus modification on the task of topic modeling. Since the goal is not to produce the absolute best topics possible, we have constructed a bare-bones model using SciKit-Learn's Latent Dirichlet Allocation library. We then defined easy-to-measure, easy-to-modify corpus properties and added them as an extension to NLTK's built-in corpus reader object. Our experiments in the previous report involved separately modifying the properties of each corpora and testing the model on the adjusted corpus to compare the effects of the changes. We compared effects using previously defined topic-quality metrics.

1.2 Review of Topic Metrics

As a brief reminder, the goal of topic modeling itself is to produce meaningful topics. Since there is no concrete definition of what constitutes a meaninful topic, and we would rather not attempt to rate each topic by manually inspecting the word likelihoods, we use topic metrics to rate meaningfulness. The four topic metrics are: distance from uniform, effective size, exclusivity, and rank1.

Distance from uniform is the Euclidean measure of distance between a topic's word likelihoods and a uniform distribution over the vocabulary space. If the topic vector is $\boldsymbol{t}$, this measures is the distance between $\boldsymbol{t}$ and a vector of the same length with all values equal to $1/|\boldsymbol{t}|$. As a topic's distance from the uniform distribution grows, that topic's meaningfulness also grows.

Effective size is the measure of a topic's size when word likelihoods are taken into account as weights. A small effective size would suggest that there are only one or two words which define this topic. Therefore, more meaningful topics yield larger effective sizes.

Exclusivity measures the degree to which top words of one topic do not appear as top words of another topic. Less overlap (smaller exclusivity value) signifies a more meaningful topic.

Finally, Rank1 is simply a count of the number of times a given topic appeared as the most frequent topic in documents from the entire corpus. Somewhat counterintuitively, more meaningful topics will have smaller Rank1. Exploring the word-likelihoods in topics that score high on this metric often show that the most frequent words are ones like: *say, does, like, think*. These are referred to as *corpus-specific stopwords*.

2 HYPOTHESIS

2.1 Original Hypothesis

Our hypothesis for the original experiment was that results would follow those observed by Tang et al. That is, we expected to find both upper and lower bounds on topic quality when changing the number of documents, document length, and the number of topics

2.3 Changes to Original Hypothesis

In our new hypothesis, we believe that the underlying cause of the changes observed by Tang et al. is a result of more than just an increase in the number of documents or their length. More specifically, we believe that the phenomenon of word co-occurrence shown in related work

suggests that the words behind the increase in document length and quantity matter at least as much as, if not more than, the document length and quantity themselves.

For this reason we chose to investigate the effect of modifying the presence of stopwords. Our new hypothesis is that we believe the same results observed by Tang et al. when increasing corpus and document size will be observed by our experiments unless that corpus and document size increase is due to an increase in only stopwords (i.e. words that don't contribute to the extracted topics).

3 DESIGN

3.1 Original Design

Originally, we were exploring the effects of modifying the number of documents, the length of documents, and the presence of stopwords. Each of these variables was changed individually. The choice to perform our experiment in this manner was primarily due to the ease of calculation afforded by changing each variably separately from the rest.

3.2 Changes to Original Design

In the second round of results, we have explored not only the effects of modifying these three variables separately but also modifying them in combination. Each run of the experiment yields 15k results. For each corpus, the experiment is set up to test 10 different quantities of documents, 10 different document lengths, and 0% through 90% stopword presence.

4 RESULTS

4.1 Increasing the Number of Documents

Figures 1a-d show the effect of increasing the number of documents being used to fit our generic topic model. Document length has been fixed to be the mean length of all documents in the original corpus. Stopwords presence is fixed at 30% (approximately what we observed in the real-world corpora).

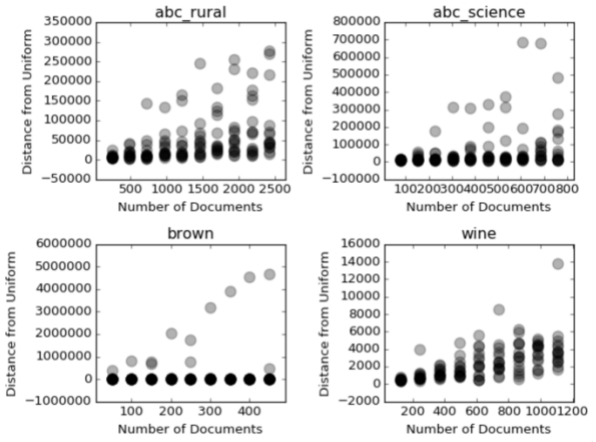

Figure 1a Distance from Uniform Distribution

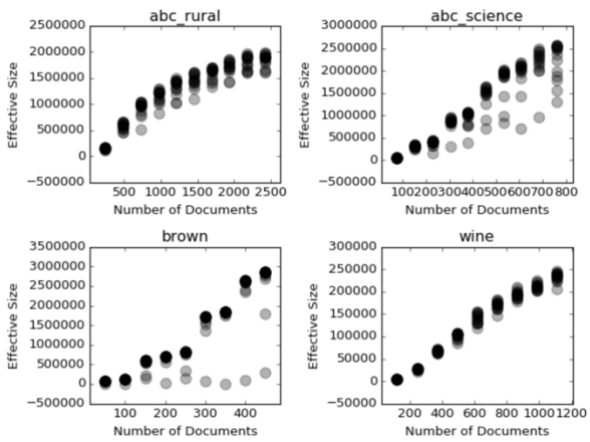

Figure 1b. Effective Size

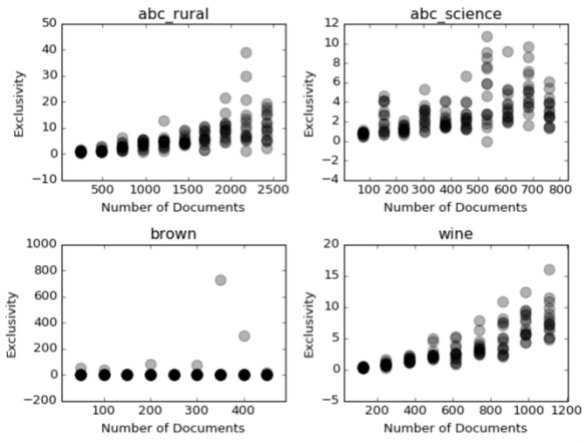

Figure 1c. Exclusivity

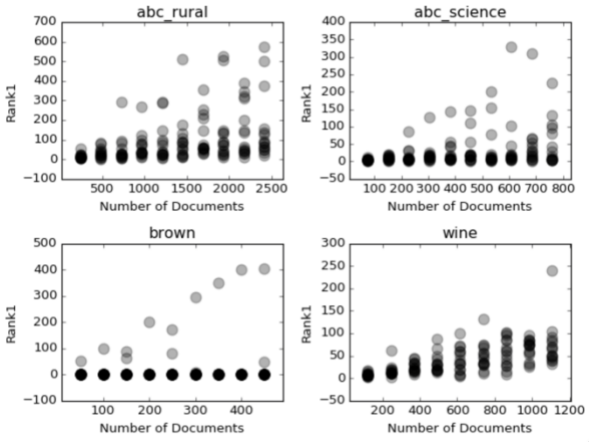

Figure 1d. Rank1

4.1 Increasing Stopword Presence

Figures 2a and 2b show the effect of increasing stopword presence. For brevity, only the minimum and maximum stopword presences are shown (0% and 90%). The change between these two extremes was gradual.

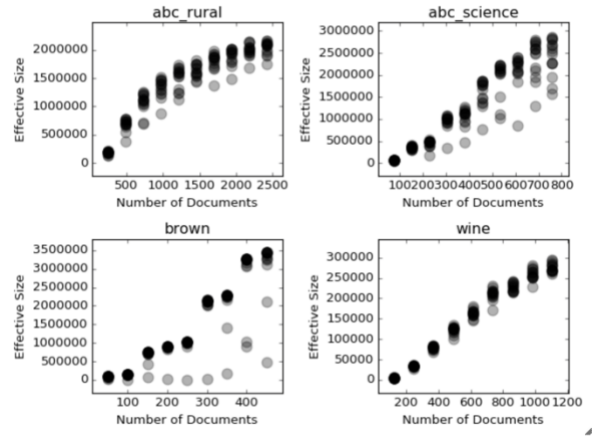

Figure 2a. Stopwords presence at 0%

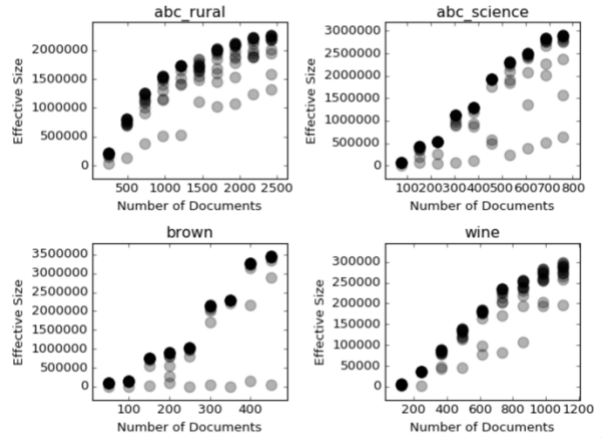

Figure 2b. Stopwords presence at 90%

4.2 Increasing Document Length

Figures 3a and 3b show the effect of increasing or decreasing the document length while increasing the number of documents in the corpus.

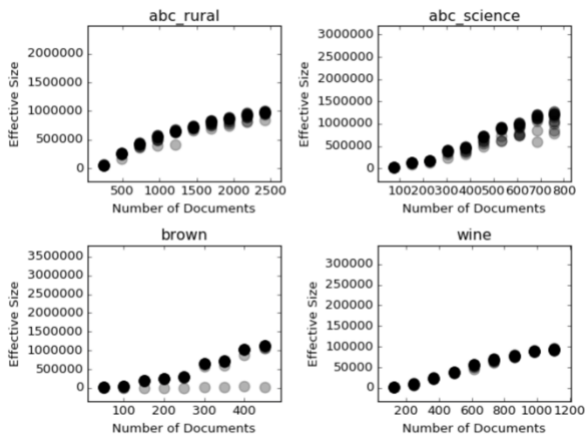

Figure 3a. Decreased document length

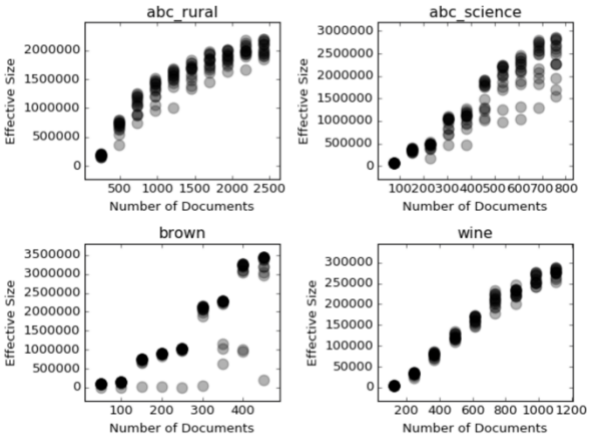

Figure 3b. Increased document length

5 CONCLUSION

4.1 Number of Documents

Our results show that increasing the number of documents has a slight positive effect on topic's distance from uniform distribution, exclusivity, and rank1 measure. It has a larger impact on the effective size of topics.

By our definitions of meaningfulness, topic quality is reduced as the number of documents increases by all metrics except for distance from uniform distribution. Upon manual exploration of the top-words in each topic, we can see that the topics containing corpus stopwords are the ones most affected by increasing the number of documents. This may explain why distance from uniform distribution is the only metric that increases, because we are injecting words that were not originally found in these documents. Essentially, we're inserting a new topic into the corpus. This may be a sign that we need a better method for injecting stopwords.

4.3 Document Length

We find that modifications to document length produce roughly the same effect on topic metrics as modifications to the number of documents overall. This is an interesting result because Latent Dirichlet Allocation relies mostly on word co-occurrence within documents. Increasing document length should have a different effect on co-occurrences than increasing the number of documents.

4.2 Stopword Presence

Injecting stopwords into the documents only negatively affected the quality of those topics containing corpus stopwords.

Interestingly, distance from uniform distribution of the corpus stopword topic increases as the presence of stopwords increaeses. This contradicts our original hypothesis that meaningful topics will be farther from a uniform distribution over the vocabulary. It may be due to the fact that we are artificially injecting words from a very small set that may becoming its own topic.

Overall, we find that most topics extracted from a variety of corpora are quite resistant to alterations of the original corpus. The only topics which were significantly affected by our modifications to the original corpus were those made up almost entirely of corpus stopwords.