**RESULTS**

METRIC

- BASE RELATIONSHIP

- EFFECT OF INCREASING ONE FREE PARAMETER

- EFFECT OF INCREASING THE OTHER FREE PARAMETER

**1 NUMBER OF DOCUMENTS**

*1.1 Distance from Uniform Distribution*

As the number of documents increases, we observe an increase in topic distance from the uniform distribution. This increase is minor for the majority of topics but quite severe for those weaker topics that contain mostly corpus stopwords. Increasing the document length only adds to the effect, pushing topics farther from the uniform distribution. Increasing the stopword presence slightly increases variance of the scores for most topics but most severely impacts the weaker topics. At 90% stopword presence, the weakest topic is clearly distinguishable from the remainder of topics regardless of the number of documents in the corpus or their length.

Same as Rank1.

Wine is the only corpus showing positive correlation between the scores for all topics and the number of documents. The deviation of scores also increases as the number of documents increases. This suggests that the model is able to produce a wider variety of meaningful topics when it has more documents to work with.

*1.2 Effective Size*

There is a logarithmic relationship of the form y=a\*log(x)+b between the number of documents and the effective size of topics where log(x) is the natural log (base e) of x. Figure X shows an example of this relationship with fixed document length around the mean for each corpus and stopword presence fixed at 30%.

Figure X shows the effect of increasing document length on the parameters of this relationship, a and b. As the number of documents increases, the parameters for the best-fit log curve also increase.

Figure X shows the effect of increasing stopword presence on the parameters of the logarithmic relationship. As stopword presence increases,

As document length increases

Increases for all four corpora. Topics in Wine show less deviation from the mean than the other three corpora with Brown being the worst. ABC Science and Brown have step-like increases in exclusivity as the number of documents increases. ABC Rural and Wine show smoother increases.

Logarithmic relationship. b increases. a stays the same.

y=a+blog(x)

*1.3 Exclusivity*

For ABC Rural, ABC Science, and Wine we see a steady increase in exclusivity for all topics as the number of documents increases. For Brown, however, only the exclusivity of the injected stopwords seems to increase. Exclusivity of all the other topics decreases as the number of documents is increased.

Linear relationship. m and b increase as number of documents increases.

y=mx+b

*1.4 Rank1*

Wine is the only corpus showing a positive correlation between the number of documents and Rank1 score of topics. For the other three corpora, only Rank1 of the corpus-specific stopwords topic increases significantly. Though, we do see an increase in the standard deviation of topic rank1 scores for ABC Rural and ABC Science as the number of documents increases.

*1.5 Cosine Distance*

There is hardly any change observed in cosine distance between topics for any of the four corpora as the numebr of documents is increased. At around 10% of their original size, Wine and ABC Science begin to show a dip but it is still not enough to diswarrant cosine distance as a metric for distinction of topics.

*1.6 Jensen-Shannon Divergence*

Divergence of topics increases in a similar pattern for ABC Rural and Wine. It still increases but at a slower rate for ABC Science and hardly increases at all for Brown.

*1.7 Kullback-Leibler Divergence*

Same as Jensen-Shannon Divergence.

**2 DOCUMENT LENGTH**

*1.1 Distance from Uniform Distribution*

Similar to increasing number of documents. Wine shows a slight but steady increase in distance from uniform. ABC Rural and ABC Science show an even slighter increase. A few topics stand out from the rest with sharp increases in their distance from the uniform distribution as the length of documents increases. Upon investigation, the top words for these topics are corpus-specific stopwords such as “don’t” and “like.” They are essentially meaningless topics.

*1.2 Effective Size*

There is a logarithmic relationship between As the average length of documents in the corpus increases, effective size also increases logarithmically.

Logarithmic relationship between document length and effective size of topics.

Reducing stopword presence reduces variance in the scores.

Increasing

*1.3 Exclusivity*

Topic exclusivity is largely unaffected by increases in the document length. Wine and ABC Science show slight increases.

*1.4 Rank1*

Increasing the number of documents caused Rank1 and Distance from Uniform to behave similarly. Although increasing the length of documents has caused the distance from the uniform distribution to increase especially for meaningless topics, it has left Rank1 scores largely unaffected.

*1.5 Cosine Distance*

Unaffected.

*1.6 Jensen-Shannon Divergence*

Unaffected.

*1.7 Kullback-Leibler Divergence*

Unaffected.

**3 STOPWORD PRESENCE**

*1.1 Distance from Uniform Distribution*

For most topics, distance from uniform distribution decreases as stopword presence is increased. However, for meaningless topics, it increases significantly after 50% stopword presence.

*1.2 Effective Size*

For the majority of topics, effective size increases relatively slightly as stopword presence is increased. For the corpus-specific stopwords, and other stopwords not captured by our fixed pre-processing, effective size is substantially reduced as the presence of stopwords increases.

*1.3 Exclusivity*

*1.4 Rank1*

At around 50% stopword presence, the score for corpus-specific stopword topics begins to increase, drastically at first for ABC Science, and gradually increasing for Wine.

No significant R.

*1.5 Cosine Distance*

As stopword presence increases, Brown topics diverges to the two possible extremes. For ABC Rural and Wine, variance in cosine distances begins to increase drastically around 50% stopword presence. At 90% stopword presence, unlike Brown, they have converged from the two possible extremes to cover nearly the entire range of possible cosine distance scores. ABC Science remains mostly unaffected.

*1.6 Jensen-Shannon Divergence*

*1.7 Kullback-Leibler Divergence*

**GENERAL OBSERVATIONS**

We identify a threshold for the effect of stopword presence on cosine distance between topics. Topic similarity as measured by cosine distance remains largely unaffected until stopword injection reaches surpasses the originally observed presence (30-40%) at which point topics begin to become more similar to one another. Prior to this, the majority of topics have cosine distance greater than 0.5. The few topics less than this are corpus-specific stopwords.

The smallest corpus, Wine, was the most affected by changes in document length and document size. The largest corpus, Brown, was the most unaffected. These results support the findings of Tang et al. who discovered a threshold on the benefits of increasing both document length and number of documents.

Your list of stopwords matters more than you might think. Two choices: filter for stopwords before everything, but this requires being psychic. Or, filter for stopwords by choosing a large enough K to capture the corpus-specific stopwords in one or two of their own topics. Then, take these topics as a list of words to filter out until you have meaningful topics.

It is very easy to identify one or two meaningless topics even with as

Brown remains largely unaffected. Shows again the importance of removing stopwords. Since brown had stopwords removed already, there was no overlap between existing vocabulary and our injected vocabulary. Therefore, it was very easy to spot our injection.

🡪 The fact that increasing document length doesn’t affect rank1. This is significant because it means we did not identify any changes in “top topic” within a document. Means documents in existing corpora are pretty impervious to changes in their word occurrences.

**FUTURE WORK**

Removing stopwords by selecting top one/two meaningless topics.

Finding the threshold for k where corpus-specific stopwords don’t make it into their own topic?