LDA Topic Modeling and Corpus Size

Topic modeling requires a large number of good-size documents for results reliability.

And **how large is large**? What would be considered the **least number of documents** (corpus size) for LDA/SLDA topic model?

Is a corpus of 200 documents large enough? Is 20 words a good-size document?

There are no magic numbers. The general rule of thumb is the more the better. Individual documents should also be of a decent size. Some researchers split documents into smaller units for analysis with interesting results.

Comparisons of documents in the hundreds and in the thousands yield significantly different results. Results based on thousands of documents are more reliable when evaluated by humans. In case of comparing different models (through topic coherence) it tells them apart to a good degree, whereas they tend to converge when the documents are in hundreds.

A good source of information for Gensim topic modeling is Micah Saxton's Capstone - Topic Modeling Best Practices.pdf (https://msaxton.github.io/topic-model-best-practices/).

Take-away lessons

- The **number of documents** is the most important variable; a few (e.g., tens of) documents won't work even if they are long; a few hundred documents should suffice but performance stabilizes after a large number of documents (1000 documents for 100 topics)
    - **Split large documents** into subdocuments if you have large but few documents
    - **Sample** a fraction of documents if you have too many documents
- **Document length** plays a useful role too; short documents (e.g., less than 10 words) won't work even if there are many of them; performance stabilizes after some large number of words per document (100 words for 100 topics)