Thea Pre-processing

Starting our exploration of the World Bank Annual Reports data, we ran some basic preliminary exploratory analysis on about 20 selected World Bank annual reports after manually cleaning them to investigate the potential difference that cleaning the reports had on our model results. The results did not improve significantly to justify the extra time spent cleaning. Aware of the time constraints on the project, we, in consultation with our instructor, agreed that the best course of action would be to proceed with downloading the reports from 1947 from the World Bank website and load it into R to begin the data exploration. This initial analysis revealed tendencies, indeed, agreeing to our hypothesis that the report discourse complexifies over time. However, the data progression analysis also uncovered outlier reports with extreme values.

Manually investigating such reports like the 2008 report showed that throughout the report larger sections of words where continuously lumped together due to formatting error with lacking necessary spaces to indicate the different words apart in the txt-file. This led to extreme spikes when graphed in the readability measures misleadingly showing that the reports were of lower complexity than they really were. In these instances, we ensured to correct the formatting error to limit the bias and better the validity in our results.

Taking the summary of the reports corpus reveals that the number of types, tokens and sentences for each report varies considerably. Comparing the same unit of analysis in the World Bank Annual Reports using the tesseract method later explained to heighten the comparability of the reports in different combined measures, the variation should not negatively impact our model results and validity.

Running initial exploratory analysis on the Flesch measure finds the 1947-96 report scores laying between the values of 30 and high 50s meaning that the reports would be understandable to people under university level over the age of about 15 years; after 1996, the reports became much more difficult to understand for people below university level. This finding is in line with our hypothesis that the language of the reports over time develops to become more difficult to understand by general publics as intended for the World Bank Annual Reports.

Preforming the preliminary Flesch-Kincaid measure finds most of the report scores to level between 10 and mid-20s. This, again, indicates that they are very difficult to read and best understood by university graduates in line with our hypothesis.

The preliminary Gunning Fog measure supports our findings in the other measurements that the reports become increasingly complex over time where they increase from initial scores of academic-level papers to scores denoting complex texts.

With this insight into the data, we continued the preprocessing by removing typical stopwords in the English language, punctuations and numbers, creating tokens and converting to DFM. We decided to not covert to lowercase as we found in our preliminary exploratory analysis that many of our collocations were using capitalizations.

To get an overview of the most frequently used words in the reports, we created a word cloud showing the most frequently used words to be in order of decreasing frequency bank, million, develop, countri, project, total, will, year, loan, fiscal, ida, world, program, increas, sector, ibrd, cost, financ, provid and credit. In line with the presented state of the art, this shows that the reports are highly technical in their language and unsurprisingly focused on strategies for how to develop countries every year within sectors through loans, credit and projects.

Further exploring the data, we preformed a KWIC test to more closely investigate how words or phrases - like develop or implement - were used in the corpus texts to see if the corpus contained words or phrases that we should be especially aware of for our later analysis.

Graphing the lexical dispersion allowed us to track the development of especially technocratic or complex words like South-South cooperation difficult to wide audiences to understand the meaning of. The analysis revealed that such words began appearing sporadically in the reports from the early 1980s and then every year since 2014 except for 2017 testifying to the increased complexity in the discourse. The lexical dispersion was also comparatively investigated in a side-by-side graph displaying the change in discourse over the years where words like project appear consistently in the reports throughout the period in comparison to the word cost, which would be likely to go with projects as all projects have a cost.

We also applied collocation analysis to identify multi-word expressions frequently used in our corpus. This is relevant is something that is likely making the reports more complex and difficult to understand for wider audiences is the frequent usage of subject specific multi-word expressions, terms or abbreviations that are not adequately defined or explained but used often in the reports making them difficult to understand. This shows that the most frequently used collocations are in order, unsurprisingly, World Bank and United States but, then, interestingly, Executive Directors, East Asia, Financial Statements and South Asia, which shows a top-down approach of the reports to development aid from the donor countries headed by USA and the World Bank but also a surprising focus on Asia.