**COP26 Linguistic Watchdog**

The goal of this project is to create clustering and topic modeling algorithms to determine the topics of discussion at the recent Glasgow Climate Change Conference (COP26), so that these topics may be compared against the 2021 People’s Climate Vote.

**Solution Path**

TO do this, I have scraped 421 English-language documents from the [UNFCCC's database](https://unfccc.int/documents?f%5B0%5D=conference%3A4301&f%5B1%5D=conference%3A4301&f%5B2%5D=conference%3A4301&f%5B3%5D=conference%3A4301&f%5B4%5D=language%3AEnglish); these documents generally take one of two forms: “high level statements” given by national delegates and other speeches, and reports put out by various committees and groups within the UNFCCC framework. These documents have a very wide range of lengths, from singular pages to more than 200 pages; splitting each document by paragraph (for speeches) and section + paragraph (for reports) has generated a corpus of almost 23,000 tokens, consisting of a total of ~1.273 million non-unique words.

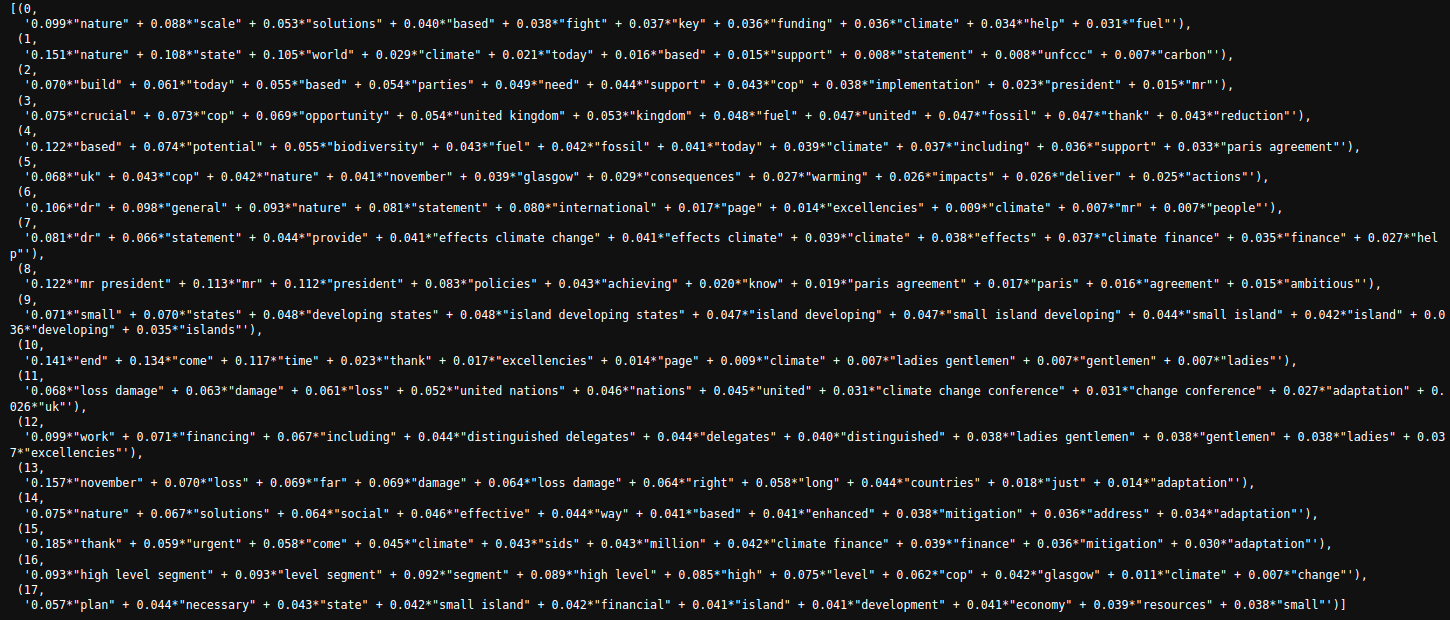
**Work Completed**

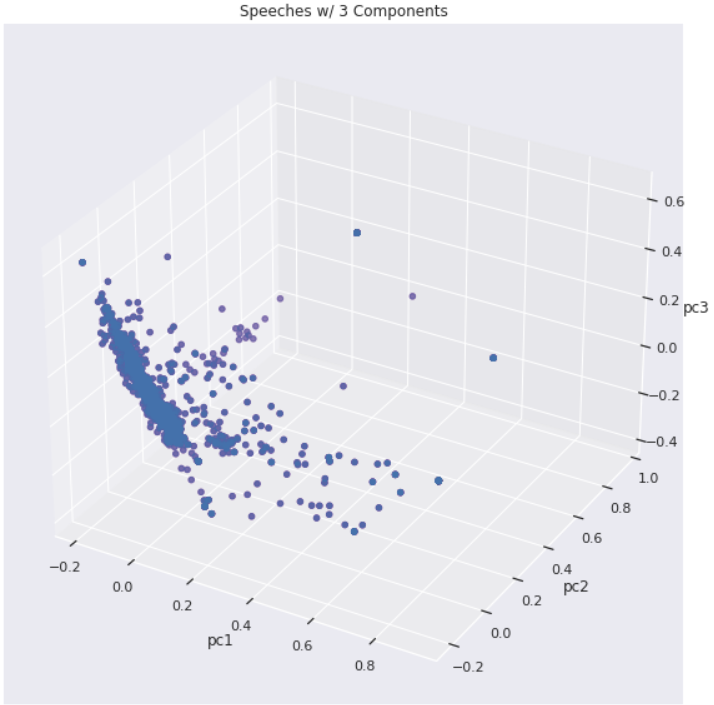
In pre-processing the data, the documents were sorted into ‘speeches’ and ‘reports’ corpora. For speeches, I removed all digits and punctuation, and change all text to lower case before tokenizing the text by paragraph. For reports, removed all punctuation and digits that didn’t precede a section identifier (alphanumeric and roman numerals). The texts were then tokenized by subsection and again by paragraph, before converting all text to lower case. Both corpora had all English stop words removed, as well as words with min\_df <= .01 or max\_df >= 0.99, and were lemmatized.

To begin topic modeling, I created a few baseline models using both CountVectorizer and TfidfVectorizer on both corpora, though for this submission I forged ahead with the tfidf model, as it appeared to have slightly more differentiated cosine similarity values. I applied PCA models with 3, 4 and 18 principal components to each corpus. Additionally, to the speeches corpus I applied an LSA, NMF, and LDA with 18 components each.

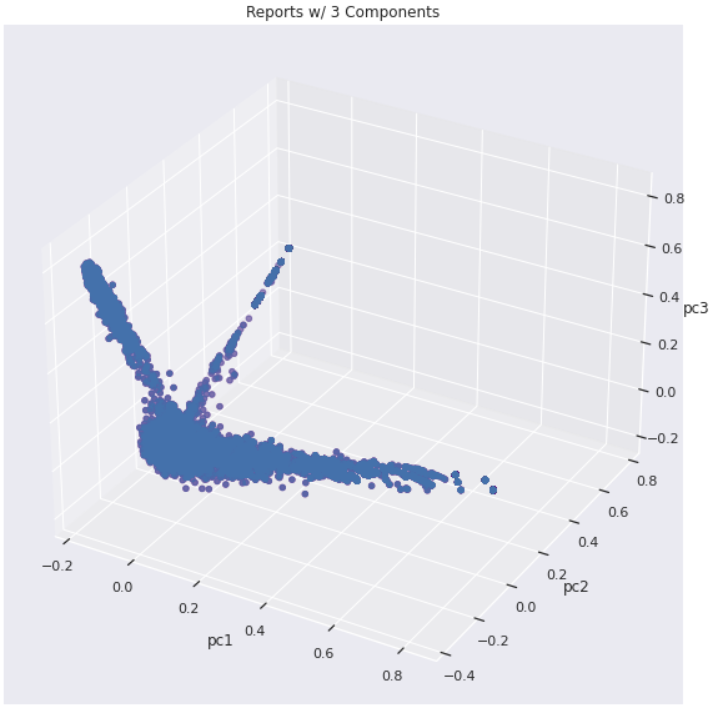
**Recent Findings**

At this stage, I am still looking over preliminary results to determine specific model choices and parameter values to tune my overall model. That said, I have achieved some interesting initial findings:





Reducing the speeches corpus to three components shows a great deal of clustering around center of the first principal component, which suggests that the speeches analyzed, the majority of which are high level statements from various national delegates to the conference, follow an very common format, with many delegates using words from the same lexicon. This theory is supported by the LDA topic modeling output of the speeches corpus, as well as the fact that 2,050 total tokens in the corpus contained a mere 343 unique words after extremely rare/common words were removed, and the corpus was lemmatized.



Putting the reports into three components generated what I call the “dragonfly graph”, which very clearly describes a great deal of topical similarity between reports of essentially three different flavors. More investigation into the reports corpus will be required to suss these relationships and topic identifiers out.

**Moving Forward**

I will continue to apply various topic models to each corpus independently, before combining the two processed corpora into a singular body. I need to make sure that my PCA parameters match up with the number of topics I am seeking to identify. More tuning of the number of topics modeled will be necessary, though it is challenging to decide if the number should be something plottable (n\_components <= 3) or reflective of the number of topics in the 2021 People’s Climate Vote, which is serving as the ultimate standard of ‘delegate diligence’ in this project. Finally, if time allows, I’d like to apply CorEx anchoring of topic words, but that will first require learning about CorEx!