**Objective and Introduction**

Our purpose for the project was to analyze and compare sentiments and tone from the FOMC meetings with the Chairperson’s Press Conference that occurs immediately after the meetings.

For context, we web scraped FOMC Meeting transcripts from 2006 to 2017. The primary reason for stopping precisely at 2017 is the fact that the Fed releases transcripts **five years after the meeting**. On the other hand, transcripts from the Chairperson’s conference were scraped for a period ranging from 2011 to 2017.

**Data Collection**

Web scraping PDFs is an imperfect act; we encountered various issues with the original data. Through the introduction of **stopwords** (commonplace pronouns and articles) and **manual cleansing**, we minimized the errors within the dataset to an acceptable limit.

However, it is still **imperfect** – a cursory glance through the transcripts should indicate the minor problems that still persist.

**Processing**

Once we felt comfortable with the input, we started to look into how analysis could be performed on the dataset – [Latent Dirichlet allocation (LDA)](https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation) made the most sense at this step. We utilize the Gensim library within Python for all of the further work (a familiarity with the [documentation](https://radimrehurek.com/gensim/auto_examples/tutorials/run_lda.html#sphx-glr-auto-examples-tutorials-run-lda-py) would be good).

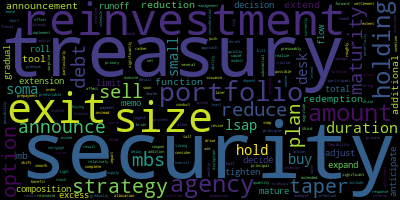


Figure 1: A wordcloud, a visual way to represent a topic

**How it works?**

Functions within the Gensim class allow for us to process our data through a **tokenizer**. The purpose for the tokenizer is to reduce certain **words to a common synonym**. For instance, words like foreign and overseas get reduced to a common synonym like international.

Once tokenized, we create a **corpus** – simply a collection of the tokenized words – and a **dictionary** that uniquely IDs the tokens to an integer.

After these steps, we can now finally run the Gensim model. First, we input how many **topics** we want the whole corpus partitioned into. Simply put, a **topic** is considered to be a set of terms (i.e., individual words or phrases) that, taken together, suggest a shared theme.

In our final results, we were able to achieve a coherence score of 0.56 ([considered permissible](https://stackoverflow.com/questions/54762690/evaluation-of-topic-modeling-how-to-understand-a-coherence-value-c-v-of-0-4)) for 32 topics for the FOMC meeting transcripts. We elected to go with 14 topics for the Chairperson’s conference, primarily due to the considerably smaller transcript size of the conferences in contrast to the meetings.

This then allowed us to create a [Jaccard Similarity Index](https://www.statology.org/jaccard-similarity/), a commonly used tool to calculate matrices for clustering and multidimensional scaling of sample sets.

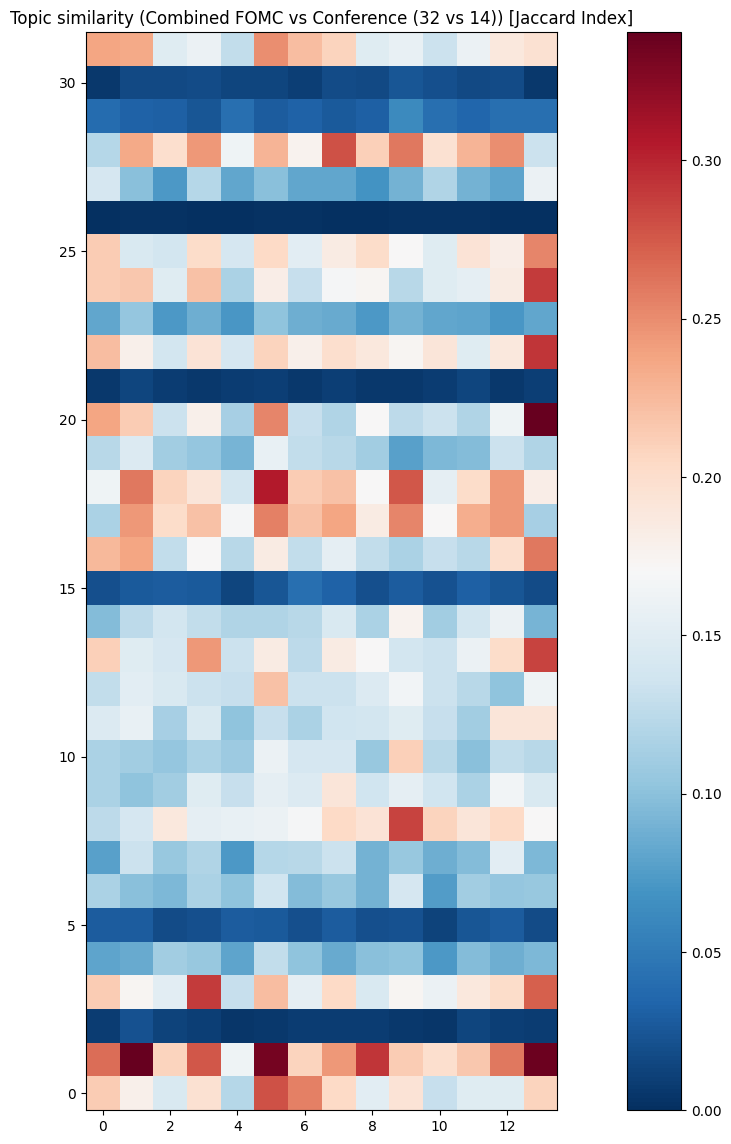


Figure 2: Topic Similarity Scores