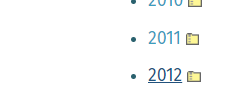
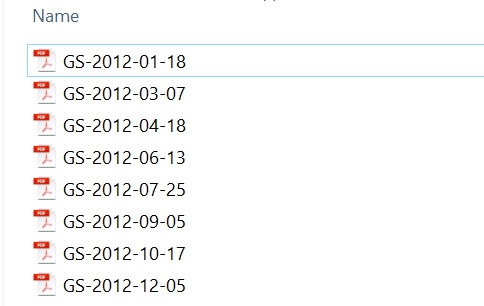
# **NLP – PDF Text Analytics and Visualization**

**Step 1** – Store all PDF files in accessible windows folder

I have downloaded PDFs from this dataset available freely on Internet. All PDFs are searchable.

Data Source - <https://www.philadelphiafed.org/research-and-data/real-time-center/greenbook-data/pdf-data-set>

**Step 2** – Set folder path in Python file in this format

directory = 'C:\\Users\\l1jxp04\\AI Notebooks\\NLP\\NLP\_Demo\\GreenSheets-2012\\GreenSheets-2012'

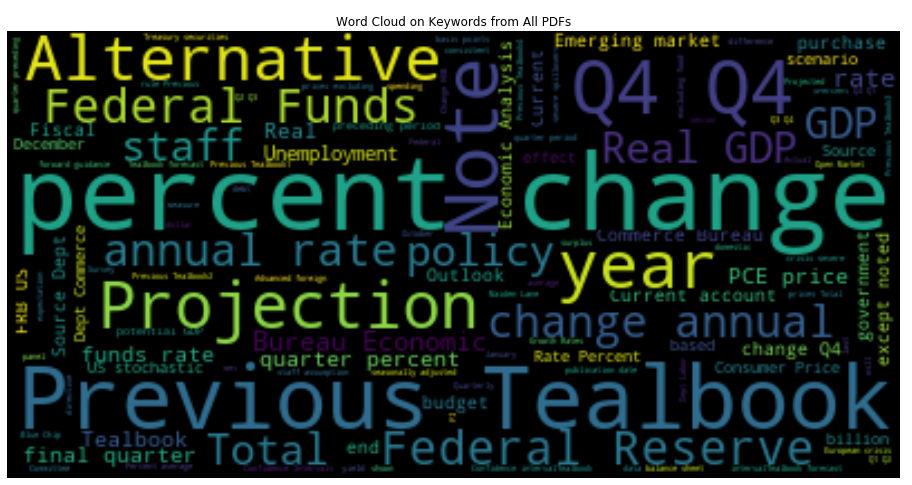
**Step 3** – Execute Python file

**Text Preprocessing** – Removed Punctuations, [Stopwords](https://pythonspot.com/nltk-stop-words/), Numbers to extract keywords

**Visualizations** –

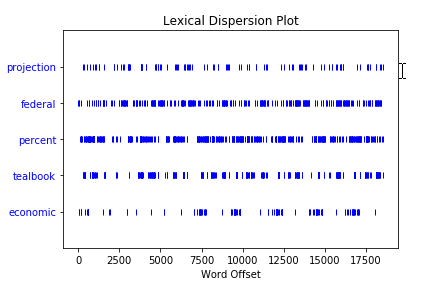
## **Word Cloud**

Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance.



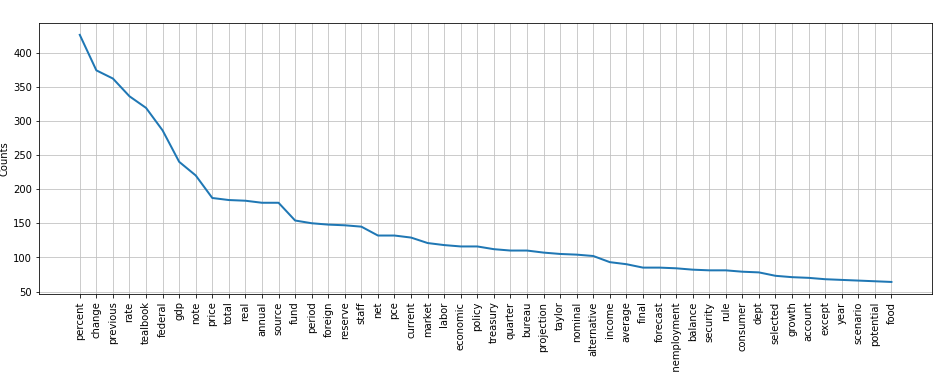
## Lexical dispersion plot

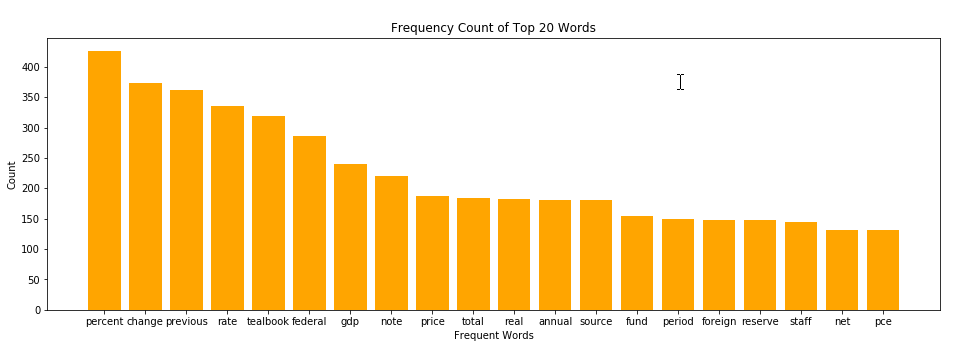
This is the plot of a word vs the offset of the word in the text corpus. The positional information can indicate the focus of discussion in the text.



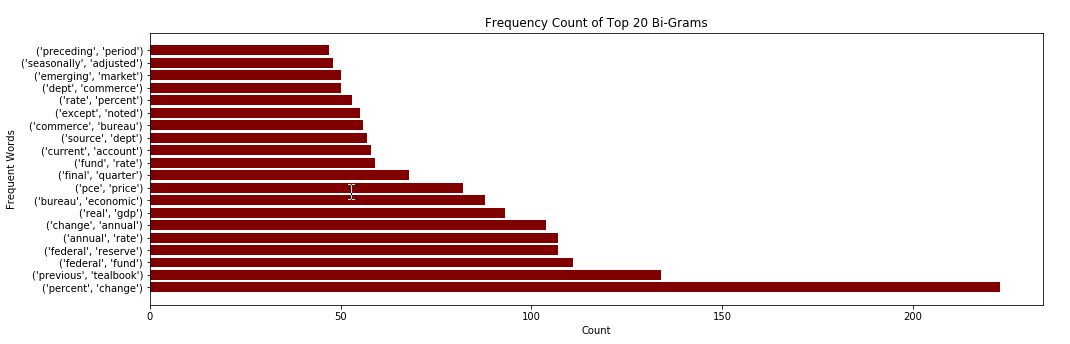
## Frequency distribution plot

Most Frequent 50 words across the corpus formed from all PDFs

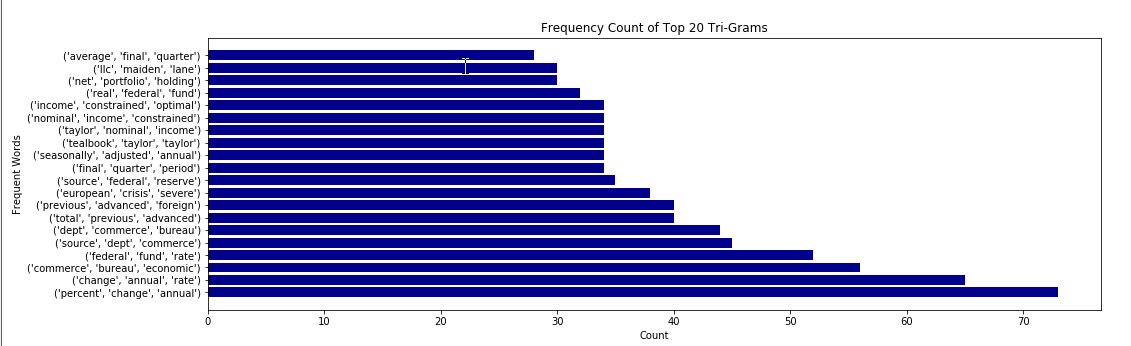




Bi-Grams : Two words that are used together most of the time (eg. Percent change)

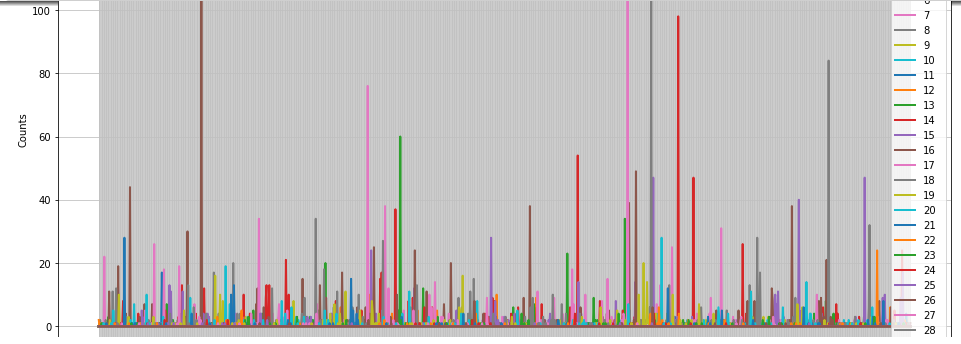


Tri-Grams : Three words that are used together most of the time (eg. Annual Percent change)

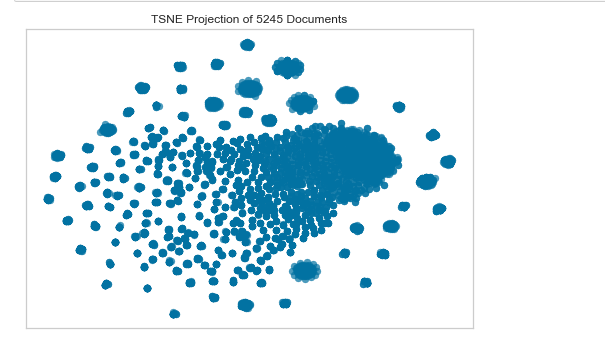


## 4. Word Length Distribution Plot

This plot is word length on x-axis vs number of words of that length on the y-axis. This plot helps to visualize the composition of different word length in the text corpus



## 5. t-SNE Corpus Visualization

Visualizing document similarity is to use t-distributed stochastic neighbor embedding 

What I did not work well due to insufficient details in data –

1. TFIDF – to know the importance of words by feature extraction
2. Sentiment Analysis
3. POS Tagging – as no grammatical structure
4. Did not perform Textract – reading text from Images through OCR