An Examination of Federal Reserve Meeting Minutes

Dr. Irina Matveeva
Department of Computer Science
Illinois Institute of Technology
Chicago, IL 60616, USA
E-mail: imatveeva@iit.edu

Harish Gandhi Ramachandran
Department of Computer Science
Illinois Institute of Technology
Chicago, IL 60616, USA
E-mail: hramachandran@hawk.iit.edu

Dan De-Rose Jr
Department of Finance
Illinois Institute of Technology
Chicago, IL 60661, USA
E-mail: dderose@hawk.iit.edu

Abstract-In this paper we implemented Latent Dirichlet Allocation (LDA), to understand the relative proportions of concepts through time. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. We have chosen LDA because in the context of text modeling, the topic probabilities provide an explicit representation of a document. We have also used Latent Semantic Analysis (LSA) to determine the influence of different policymakers on the minutes of meeting data.

Keywords- Unsupervised learning, Topic-modeling, Text-analysis,Latent Dirichlet Allocation, LDA, Latent semantic analysis, LSI, Natural Language Processing(NLP)

1. INTRODUCTION

The Federal Reserve often referred to as "the Fed" is the central bank of the United States. Congress created the Fed in 1913 to help promote a safe and sound monetary and financial system for our nation. The Fed includes the Board of Governors in Washington D.C. which has seven members including the Chairman and Vice Chairman. All of the members of the Board are appointed of the President of the United States and confirmed by the United States Senate. The Fed also includes 12 regional Federal Reserve Banks located in cities throughout the country. The Reserve Banks serve as the central banks' operating arms and also gather economic information from all over the country to help the Fed both monitor the economy and get the broad input necessary to develop and implement effective U.S. monetary policy.

Monetary policy is the Federal Reserve's actions, as a central bank, to achieve three goals specified by Congress: maximum employment, stable prices, and moderate long-term interest rates in the United States.

The Federal Reserve conducts the nation's monetary policy by managing the level of short-term interest rates and influencing the availability and cost of credit in the economy. Monetary policy directly affects interest rates; it indirectly affects stock prices, wealth, and currency exchange rates. Through these channels, monetary policy influences spending, investment, production, employment, and inflation in the United States.

The last decade has seen central banks gain prominence in markets. The financial crisis required central banks to ride to the rescue of the financial system. Since then, Fed watching, the art of discerning the Fed's intentions, has moved from the confines of the urbane to front page news. This transition was prominently declared when Ben Bernanke, the previous Chairman of the Federal Reserve Board, gave his last speech on January 3rd, 2014 in which he insisted, "**The Fed must continue to find ways to navigate this changing environment while providing clear, objective, and reliable information to the public**." [1]

Text analysis is an ideal way to discern the exact intentions of a corpus, or body of documents. It's possible, as often happens, when one read's a text, their perception is inhibited by their current biases [2]. As discussed prominently in behavioral economics, confirmation bias is one of the strongest biases and most difficult to overcome. In this way, LDA and LSA can be tools to understand with precision what the exact implications of a text are. In the following sections we will talk about the data exploration and text analysis of the meeting documents, which are called minutes. In addition, we explore the speeches of the policy makers and how it impacts the minutes.

2. Description of Dataset:

Every year there are 8 meetings conducted across the nation and a written report called **"**Minutes of the Federal Open Market Committee report" [3] is submitted. The data as textual information is available over the web from the year 1993. The documents are obtained in both pdf and html web format. The format differs mainly because of the unavailability of the pdf format before the year 2007.

A python script has been developed to extract the data from the web page using the Beautiful soup and urllib packages. The script automatically stores the extracted data to a folder with year as the folder name. Thus documents inside the folder are monthly reports of that given year. There are 197 documents which are scrapped from the website. Since most of the documents are in the text and pdf format, the size of the data is 50MB, which quite sufficient for the in memory processing for any modern laptop. The data present inside has the information of the key attendees including the chair, which is present in the first page of the document. The rest of the document has a textual information of what happened inside committee meeting.

3. Preliminary Experiments:

3.1 Preprocessing Stage:

Even though the extracted raw data from web has precise content of the minute's information, is not clean enough to perform the analysis. So only the relevant information is extracted for analysis. For instance, every report contains an initial salutation to the committee members, thus that section of the report is removed across all the documents. A corpus has been built upon based on the documents of the given year. The words in the corpus is converted into a matrix format for easy computation. This matrix called Document Term matrix, performs the frequency count of each term and maps it to the corresponding document. The columns are the terms while the rows are the documents. The tm_map function from tm package is used to preprocess the corpus. We performed remove stop words, changed to lower case, punctuations and numbers has been removed, and transforming to lowercase characters. In the final stage we preformed stemming. The sparsity of the matrix is further reduced with the help of remove sparse terms of threshold 0.93. This gives a 61% sparsity of matrix with more significant terms shown in the table below.

	Before Preprocessing	**After Preprocessing**	**Remove Sparseness**
Number of unique terms	10836	5198	1856
Sparsity	90%	85%	61%

The words that are not significant are all removed from the corpus, ie the "Federal Reserve System", "Federal Open Market Committee", "meeting", "FOMC", "Present", "year" are some of the commonly occurring words across the document. To get a proper data of precise to the content of the speeches we followed an iterative approach of preprocessing by finding the irrelevant words and updating the stop word list.

Corpus is converted into document term matrix and term document matrix. Word cloud as shown in Fig.1 has been created across each year, to get an intuition of what the document is talking about. The word frequencies bar plot shows us a proper term counts. The tf_idf is used to find out the important words rather than just getting the frequency of words in the documents. As shown in the Fig.2 below, the top important words that occur over the documents.

Fig.ure 1: Word Cloud of important terms

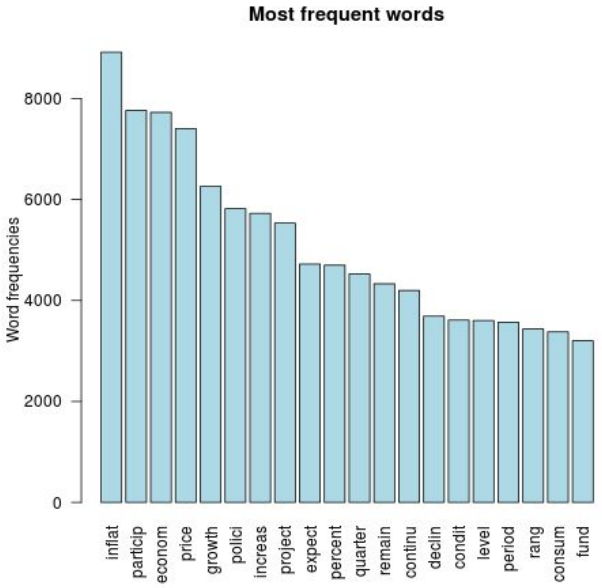

Fig 2: Bar Plot of most frequent words

3.2 Clustering:

The minutes that we use here talks more about the financial monitory policy and it is a challenging problem to provide the cluster numbers for these documents just alone based on terms (most of which are repetitive). Unsupervised learning helps in finding the optimal cluster number that reduces the error function. K-means is a popular algorithm in document clustering, which is fast and efficient. And the NbClust package is used to find the best number of clusters as 4 as shown in the table, which is representative of the concepts that is used in later part of the experiment.

Method	**Number_clusters**	**Value_Index**
NbClust(k-means)	4	0.1263

There are various methods for deciding the optimal number of clusters. As shown in the Fig.3, the Silhouette method and elbow method, the best cluster number agrees with the best nbclust cluster number.

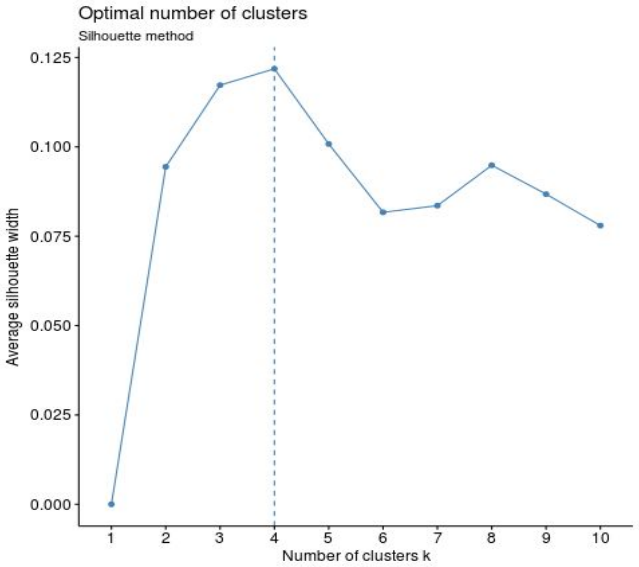

Fig 3: Optimal number of clusters using NbClust

4. Analysis

4.1. Variation of Topics across time using LDA

A first step in identifying the content of a document is determining which topics that document addresses. We describe a generative model for documents, introduced by Blei, Ng, and Jordan [4], in which each document is generated by choosing a distribution over topics and then choosing each word in the document from a topic selected according to this distribution. LDA algorithms try to classify the text into "unobserved topics" or groups, and then map each word of a text into those unobserved topics. Principal component analyses (PCA) of empirical data might offer a clarifying analogy, in that both LDA and PCA ignore any specific context or pre-conceived categories in order to statistically and unbiasedly identify relationships to reduce dimensionality or complexity. Thus, while the work that we discuss in the current paper focuses on simple "bag-of-words" models, which lead to mixture distributions for single words (unigrams).

Latent Dirichlet Allocation (LDA), the parameters of the posterior topic distribution for each document is calculated through the gamma parameter in topic modeling package [5]. We use Gibbs sampling approach to perform the LDA. Gibbs Sampling is one member of a family of algorithms from the Markov Chain Monte Carlo (MCMC) framework [6]. The MCMC algorithms aim to construct a Markov chain that has the target posterior distribution as its stationary distribution. In other words, after a number of iterations of stepping through the chain, sampling from the distribution should converge to be close to sampling from the desired posterior. Gibbs Sampling is based on sampling from conditional distributions of the variables of the posterior. The collapsed Gibbs sampler for LDA needs to compute the probability of a topic being assigned to a word, given all other topic assignments to all other words [7]. Thus based on the Gibbs sampling approach LDA is performed to obtain the topic distribution over time. The y-axis of the plot in Fig.4 represents the probability of the topic while the minutes as per the time flow is represented on the x-axis. The top terms which represent these concepts are also tabulated below,

Topic 1	**Topic 2**	**Topic 3**	**Topic 4**
growth	inflat	price	econom
price	project	growth	inflat
rang	econom	inflat	project
increas	percent	econom	price
period	polici	quarter	bank
polici	fund	increas	financi
econom	expect	spend	credit
direct	price	consum	increas
quarter	rang	polici	expect
month	labor	busi	remain

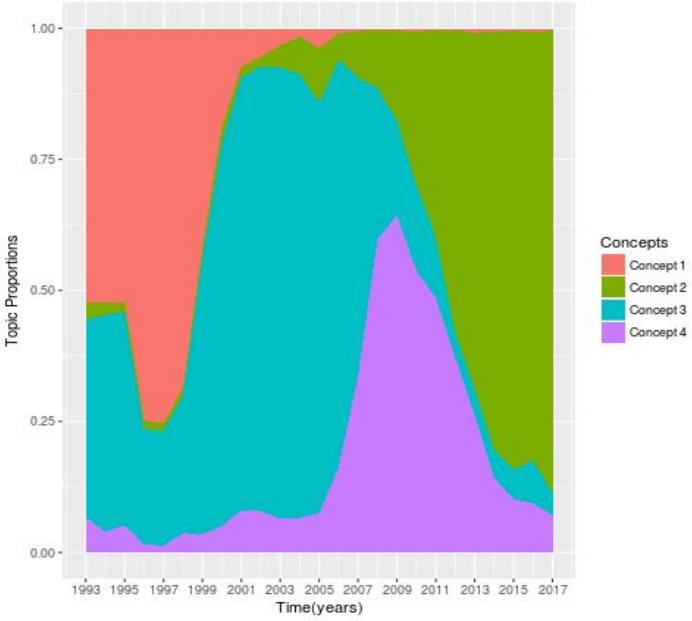

Fig 4: Variation of Topic distribution over time

Some of the Inference from the topic distribution and top terms representing those topics are:

1. There are some overlap between the most represented terms of each concepts (for eg: economy appears in all topics)
2. However the context upon which these common terms appears plays different role (for eg: in Topic 1, the econom, fianci, bank, security etc plays more into monetary aspect than the Topic2 which has

growth, expansion, policies etc which is more into the development of the nation)

3. The role of the federal bank decisions varies as time progresses. For instance consider the concept 4(topic 4) is more representative of role of money in the economy(ie economy, inflat, fianci, credit etc.), there is peak in the probability value from the year 2007 – 2010, that is when financial crisis hit the nation [8]. One hypothesis is that fed is concerned about bringing stability to the value of money.

4.2 Quantifying the Influence of Policy makers on minutes using LSA:

Latent semantic analysis (**LSA**) is an oldest among topic modeling techniques, in particular distributional semantics, of analyzing relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms. LSA assumes that words that are close in meaning will occur in similar pieces of text (the distributional hypothesis). A matrix containing word counts per paragraph (rows represent unique words and columns represent each paragraph) is constructed from a large piece of text and a mathematical technique called singular value decomposition (SVD) is used to reduce the number of rows while preserving the similarity structure among columns.

Usually we need not only analyze a fixed dataset, but also apply model to new data. For instance we may need to embed unseen documents into the same latent space in order to use their representation in some downstream task (for example here we used for constructing the correlation matrix). This feature helps in our analysis by identifying where our new minute data in the LSA space created using the speeches of policy makers.

It is to be noted that additional documents can be mapped into a pre-exisiting latent semantic space without influencing the factor distribution of the space as shown in Fig.5. Thus the fold_in function is used, when additional documents must not influence the calculated existing latent semantic factor structure. It Folds a set of minutes into the LSA space of speeches set by policy makers.

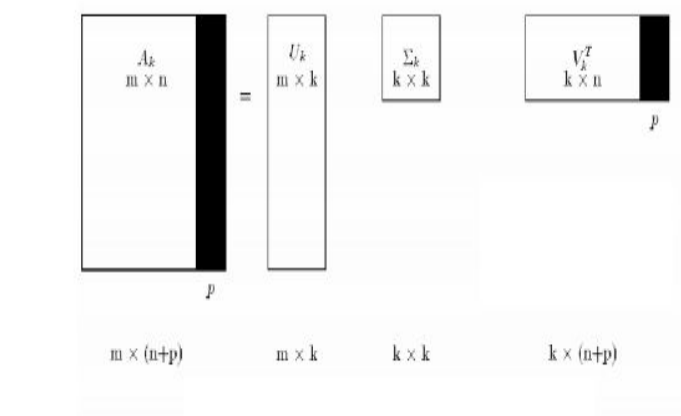

Fig 5: Mathematical representation of folding-in p documents

Let us explore the Mathematical explanation of the fold_in function specific for our application. We know the SVD matrix can be decomposed into respective terms and documents in the k dimensional space. The matrix for the LSA space is calculated with k concepts chosen by dimcalc() function of the LSA package. As shown in the equation, the document term matrix can be decomposed as shown,

$$A_{m*n} = U_K \sum_K V_K^T$$

We have the current minutes of the meeting of a particular month (here we used dec 2016 and nov 2017 on the respective latent space). Thus the [1xk] document vector in the space is obtained by projecting on the space,

$$d_{1*k} = v_{1*n}^T V_{n*k} \sum_{k*k}^{-1}$$

Now using this document vector d, we recomputed the term vector.

$$m_{n*1} = V_{n*k} \sum_{k*k} d_{k*1}^T$$

It is to be noted that we apply the weighting scheme to the matrix before performing the LSA. Here we consider the speeches document from each key members of the committee (eg: Yellen, Brainard etc.). The Latent space obtained is used for embedding the minute's data. It is to be noted that the minute's data was in December 2016 while the Latten space created is based on their speeches on November 2016. The similarity of the speakers with minutes' data is compared using the correlation matrix obtained using cosine as similarity measure [9]. The more closer the vectors (ie angle between the data points), the more similar they're to each other.

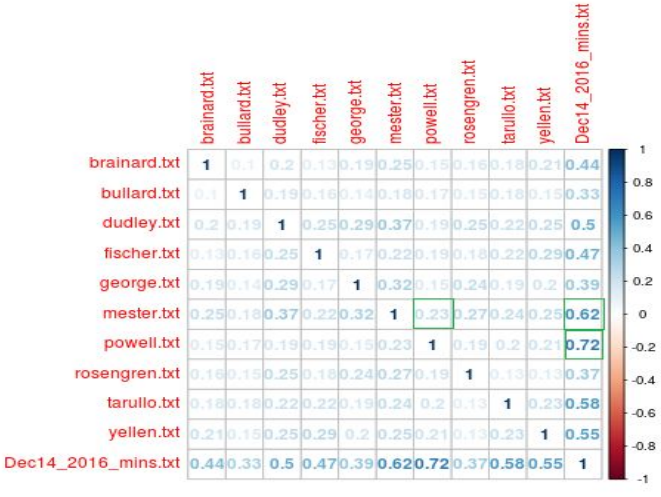

Fig 6: Correlation Matrix of minutes and speeches data

Similarly, for the relevant speeches before September 2017 is used for creating the latent space for the September minutes data. In this way a similarity search is performed to see correlation of these members. As shown in Fig.7, most representative dimension is used for plotting the speeches on a 2-dimension space.

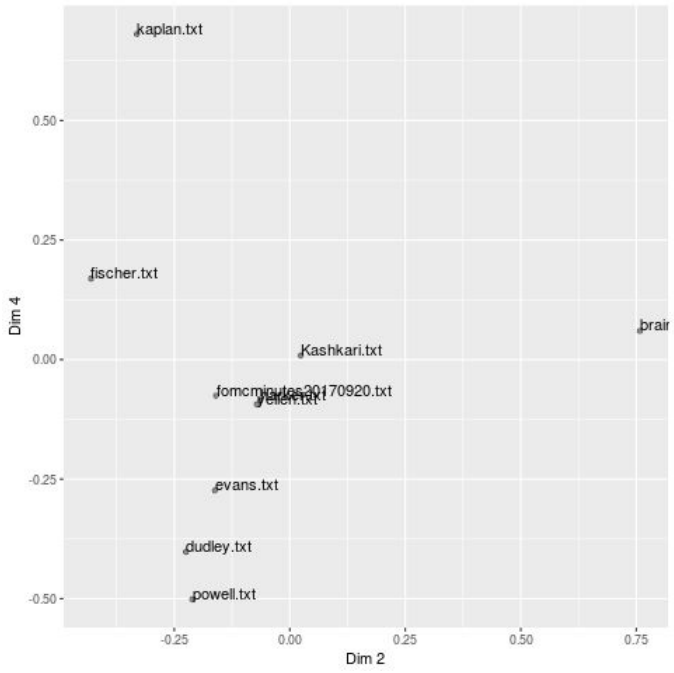

Fig 7: 2D visualization of minutes and speeches of Dec2017 data

The key Inference from LSA:

1. The created LSA for the speeches in the time frame of 2016, the speeches of Mr. Powell and Mr. Mester contributes more to the minute's data.
2. Considering the corrplot Fig.6 above, both Mr.Powell and Mr.Mester speeches are more correlated than with other speakers. Thus it is sufficient enough to infer where the minute's decision is going to be based on either one of their speeches.
3. An Easier way to visualize is using 2d - visualization plot, and for the recent (December 2017) release of minutes' data, we can see that the Chair of the committee Janet Yellen plays a significant role in deciding the outcome of the minutes.

5. RESULTS AND DISCUSSION

The Experiment carried out in this paper involves data from the web. The raw data as such is not significant enough to show any interpretable results. To avoid 'Garbage in Garbage out' scenario, we have spent a quite an amount of time in data extraction and preprocessing stage.

The topic selection is performed using the LDA approach, a bag of words model to identify the probabilities of each topic for each document. It is necessary to choose the number of concepts prior to the LDA analysis. So to perform an unbiased representation of the topic numbers, we used NbClust with kmeans algorithm to iterate over all possibilities of cluster numbers to get the optimal cluster value.

Latent semantic Analysis (LSI) is performed to mainly identify the latent space and embed with the new minutes data to quantify the similarity metrics with the respective month's speeches. In future work, this approach can be used as a predictive approach of finding whose speeches will more contribute to the outcome of the upcoming meetings. And also other advance approaches like Hierarchical Dirichlet Process (HDP) [10] can be used in future works mainly to address the number of topics.

6. REFERENCES

[1] "https://www.federalreserve.gov/newsevents/speech/bernanke20140103a.htm".

[2] "Saret, Jeffery, and Subhadeep Mitra. "An AI Approach to Fed Watching." Two Sigma Street View, May 2016, 1-6.".

[3] "https://www.federalreserve.gov/monetarypolicy/files/fomcminutes20171101.pdf".

[4] "Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003) J. Machine Learn. Res. 3, 993-1022".

[5] "https://cran.r-project.org/web/packages/topicmodels/topicmodels.pdf".

[6] "W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. Markov Chain Monte".

[7] "http://u.cs.biu.ac.il/~89-680/darling-lda.pdf".

[8] "https://www.economist.com/news/schoolsbrief/21584534-effects-financial-crisis-are-still-being-felt-five-years-article".

[9] "http://statweb.stanford.edu/~jtaylo/courses/stats202/restricted/notes/distances.pdf".

[10] "https://people.eecs.berkeley.edu/~jordan/papers/hdp.pdf".