Yusiqi Wang

Professor Diego Klabjan

Data Science and Analytics

March 20, 2020

# Executive Summary

**Process Taken:**

This study aims to process **Topic Modeling** and **Document Clustering** from daily business news for 2013 and 2014.

**Latent Dirichlet Allocation** is performed based on the documents in each quarter. Each document of 8 quarters is named by **14 topics**. This process is also performed on the collection for all the articles.

Documents for the first quarter of 2013 are selected to be clustered. Top 500 Tf-idf vectors and LDA-based features are combined as features for documents. **Principal Components Analysis** is applied to further reduce the number of features. Then the documents are clustered based on the resulting features.

**Report**

# Data Preprocessing

The first step is data preprocessing.

Names of documents in each quarter were extracted from TXT files and put into a list.

Quarterly documents were put in one file and read.

Punctuations were removed and error data was ignored. Each TXT file was tokenized into words and stop words were removed.

# LDA Model Implementation

# Gensim creates a unique id for each word in the document. Here we create the two main inputs of the dictionary(id2word) and the corpus for the LDA topic model.

# Documents for one quarter was selected. We define validation set and train set and create dictionary for them. Thus, we can find the right number of topics and verify its accuracy in the validation set.

# In order to determine the optimal number of topics, perplexity is applied here. In the train set, we calculate perplexity for different numbers of topics and draw a line diagram to show the relationship between the two elements. A model with lower perplexity is considered to be good. With the perplexity seems to keep decreasing with the number of topics, it may make better sense to pick the model that gave the lowest perplexity before flattening out or a major drop. In this case, therefore, we picked 14 as the number of topic (in the range 1-20). Test accuracy of perplexity in the validation set.

# When the number of topics is 14, we’re going to define the Keyword for each topic by the weights of words. We selected the topic name from the top 10 keywords that are representative of the topic. Below chart shows the topic names for each quarter and the collection for all the articles.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Topic** | **Topic** | **Topic** | **Topic** | **Topic** | **Topic** | **Topic** |
| **1** | **2** | **3** | **4** | **5** | **6** | **7** |
| **2013 Q1** | Tax | Stock | China | Percent | Business | Recession | Investor |
| **2013 Q2** | Gold | Bank | Policy | Financial | Company | Obama | Business |
| **2013 Q3** | Housing | Debt | Growth | Apple | Fund | Rate | Price |
| **2013 Q4** | Index | Government | Data | Company | Economy | Job | Hong Kong |
| **2014 Q1** | Free | Shoe | New York | Government | Street | Appdownload | Market |
| **2014 Q2** | Growth | Free | Company | Business | Market | Rate | Appdownload |
| **2014 Q3** | Growth | Appdownload | Apple | Oil | Dollar | Russia | Company |
| **2014 Q4** | ECB | Company | Bank | Hamm | Market | Free | Reuters |
| **All** | Russia | European | Percent | Policy | Tax | Oil | Gold |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Topic** | **Topic** | **Topic** | **Topic** | **Topic** | **Topic** | **Topic** |
| **8** | **9** | **10** | **11** | **12** | **13** | **14** |
| **2013 Q1** | Bank | Market | Economic | Bailout | Economy | Financial | Foreclosure |
| **2013 Q2** | Fed | Bitcoin | Fund | Percent | Market | Investment | China |
| **2013 Q3** | House | Index | Fed | Emerging | Mortgage | Business | Bank |
| **2013 Q4** | Market | Bitcoin | Debt | Gold | Currency | Inflation | Reuters |
| **2014 Q1** | Business | Australia | Bitcoin | Company | Apple | China | Unemployment |
| **2014 Q2** | Free | Inflation | Sales | AMP | Business | Bank | Investment |
| **2014 Q3** | PIMCO | Argentina | Tesla | Government | Food | Funds | Fed |
| **2014 Q4** | Stocks | Oil | Financial | Appdownload | Credit | Price | Amazon |
| **All** | People | Growth | Bank | New York | Appdownload | Stock | Economy |

# In the last row (Topics for all documents), the previously repeated words have been highlighted.

# In conclusion, Topics for all documents are similar with Quarterly Topics.

# TF-IDF

# Documents for the first quarter of 2013 are selected to be clustered. Read articles day by day and Remove stop words.

# Now we will import the vectorizer sklearn and then call fit and transform over it to calculate the TF-IDF score for the text.

# Sort words by value in descending order and put them into dataframe. The words have more significance with higher TF-IDF score. Then, top 500 TF-IDF scores were picked.

手机屏幕截图

描述已自动生成

**Top 500 TF-IDF and LDA-based features should be combined as features** for documents. Then we add LDA-based features. Ignore it if the word with top 500 TF-IDF score duplicate with LDA-based topic names. We have **505 features in total**.

手机屏幕截图

描述已自动生成

Process the same steps on all documents in 2013 Q1. If there is no such word in certain article, fill in the value of 0. Below is the features for all the **90 documents**.

手机屏幕截图

描述已自动生成

# PCA

**Principal Components Analysis** is applied to further reduce the number of features.

The horizontal axis in below plot is number of components and the vertical one is length of array. When **number of components** equal to **0.85,** the slope is highest, which means the effect of dimensionality reduction is the best.

手机屏幕截图

描述已自动生成

In order to create a 2D plot for the clusters, we set the **number of components as 2**. We finally got 2D array and plot a scatter diagram.

手机屏幕截图

描述已自动生成

# Cluster

Finally, the documents are **clustered** based on the resulting features.

# The Silhouette could be used to assess the quality of clustering. After the Silhouette was calculated and the plot was created. Look at the max value. When the number of clusters is 15, the Silhouette most close to 1, the best each element would fit into the cluster.

手机屏幕截图

描述已自动生成

When projecting data point from high dimension to low dimension, some data point might be overlapped. Thus, some clusters were blocked and invisible.