Final exam assignment

Question 1

The goal of question one was to use LDA to determine similar topics per quarter, in 2013 and 2014.

After importing the data, a few things were done in order to ensure that the analysis could be done in a systematic way. Firstly, I ensured that the files were named in the "Y-M-D" format. This was done to more accurately subset the data into quarters later on in the assignment. A data frame was then created, with a column for "date" and a column with the information from the document. By doing this, the dates became unorganized, so I reset the indices before continuing. Finally, I cleaned the data to remove punctuations and to convert the information to lowercase.

After this, I started looking into how to separate the data into quarters, according to the dates. First, I manually inputted the end dates for the quarter, and located the index at which they were located in the data frame. Then, I created 4 separate data frames holding the document information for each quarter.

Next, I decided to stem and tokenize the words in the documents. I used the Snowball Stemmer algorithm to remove the suffixes of the words in the documents. I decided to strip off the suffixes, as differences in suffixes are extremely common for words that have the same stem. Additionally, I lemmatized the words using WordNetLemmatizer(). In the tokenizer function, I stemmed and lemmatized the words. I also restricted to include words that were only greater than 3 characters. I did this because during initial experiments with the code, I noticed that the topic generation resulted in a lot of garbage phrases that were about 3 characters long. Hence, I restricted the output to only include words that had more than 3 characters.

For the number of topics and the number of words in each topic, I chose 3 and 5 respectively. I chose 3 topics because when writing business reports, people often talk about the "top three topics" of discussion. Hence, to align more closely to a managerial insight, I chose to include 3 topics. Additionally, I chose to include 5 words in each topic, to allow easy comparison of common words between topics.

The top 3 topics for the 4 quarters can be seen in the output file. These are summarized in the two tables below:

Topics in 2013:

Quarter	Topic 1	Topic 2	Topic 3
1	year market report percent said	year market bank percent said	alcoa gerhartsreit solar iwatch greenberg
2	shaft eurochem snowden sinker soni	suspect boston tsarnaev polic bomb	market year said report rate
3	rate report year market increas	market year said like time	trump yahoo board blackberri bank
4	market year said like price	ahrendt amazoncouk cva/dva burberri gurley	year market said report like

Topics in 2014:

Quarter	Topic 1	Topic 2	Topic 3
1	market year said free appdownload	koch nonprofit caterpillar dealer disregard	market year said report bank
2	year said market bank compani	market year said report compani	said year market compani free
3	said year market report compani	said year compani market report	said year market compani report
4	year said market report compani	climax intergener 1235 clan reutersback	year said market price bank

From here, it can be seen that the trends in the topics are relatively similar over the two years. Most of the common topics are “market”, “company”, “year”, “report” and “bank”. This is expected of the data set, as the articles used were from Business Insider. However, it is seen that some of the topics are not common and remain as a relevant topic only for the quarter. This can be seen in quarter 2 in 2013, where some topics were “police”, “boston” and “bombing”, or in quarter 3 in 2013 where a topic is “trump”. These topics reflect the sudden shocks around the world. For example, the topics selected in quarter 2 in 2013 highly reflect the Boston bombing that occurred on April 15th, 2013. This notion can be seen again when LDA was done on all of the 2013 and 2014 data, i.e. the main topics are surrounding business instead of big, yet short termed shocks. These topics can be seen below:

Topic 1: said free appdownload year market
Topic 2: report said year market price
Topic 3: year market said bank company

Question 2

The goal of this question is to cluster the documents according to tf-idf features and LDA based features.

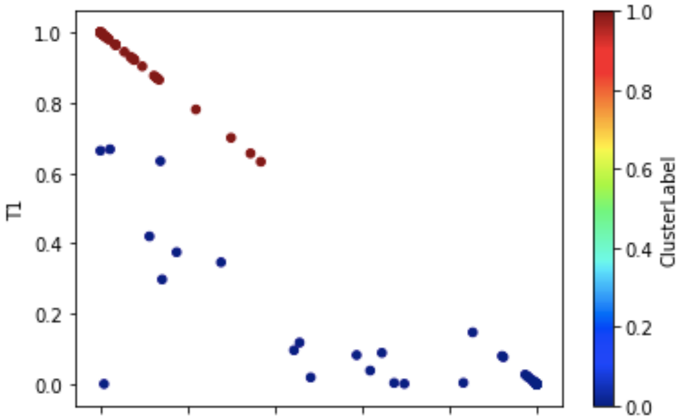
To create a tf-idf model, I used the TfidVectorizer function and I labeled the feature names with the variable "vocab". As it can be seen form the data frame, the individual occurrences of each word are relatively low and it would be difficult to isolate clusters according to this matrix.

Next, I created the lda model and created feature vectors based on that. The probabilities of each document appearing in a particular topic group are much higher than the probabilities observed in the previous model.

For clustering, I decided to use hierarchical clustering instead of a conventional method like Kmeans. I did this because this type of clustering can use methods other than distance to calculate similarities between items. Since distance cannot be used to measure the difference between documents, this method of clustering is used.

After producing a dendogram, it was seen that the optimal amount of clusters are two clusters. The following clusters were created:

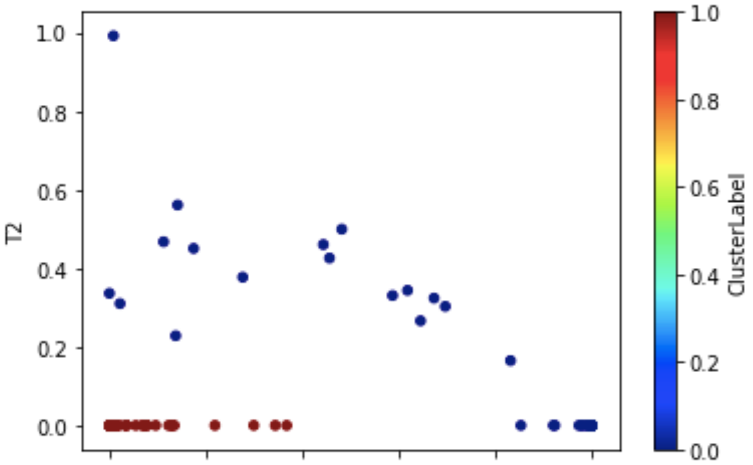
Comparing Topics 1 and 2



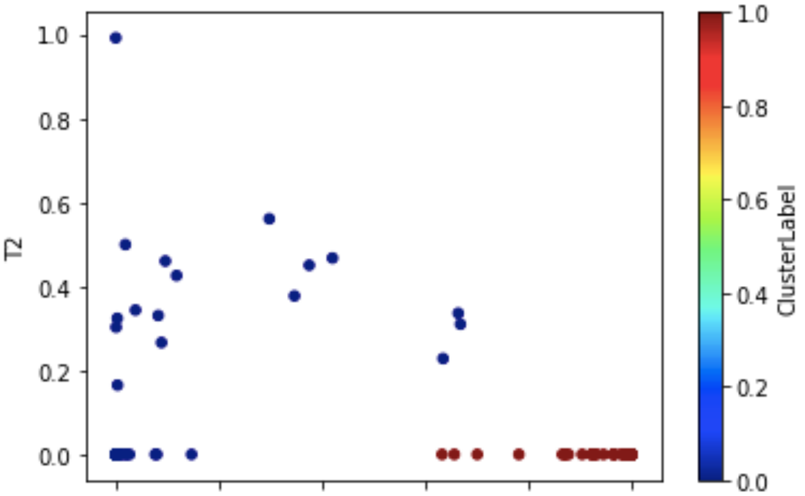
Here it can be seen that features that are more highly classified into topic 2 possess similar qualities.

Comparing Topics 1 and 3

Here it can be seen that features that are less highly classified into topic 1 possess similar qualities.



Comparing Topics 2 and 3



Here it can be seen that features that are less highly classified into topic 3 possess similar qualities.

The blue clusters (i.e. the clusters labeled 0) are the clusters that have the least visibility in all of these graphs.