**Customer Feedback:**  
**What Matters Most to Bank Customers**

**MSA 8040 Data Management for Analytics**

**Team Members**

**Bruce Hart**

**Keerthi Bojja**

**Siddhartha Kille**

**Yishan Xu**

**Table of Contents**

**1.** Objective………………………………………………………………………… 3

**2.** Web Scraping…………………………………………………………………… 3

**2.1.** Source………………………………………………………………………. 3

**2.2.** Sample Data……………………………………………………………….. 4

**3.** Data Exploration………………………………………………………………… 5

**4.** Pre-Processing…………………………………………………….……………. 7

**5.**  Approach………………………………………………………………………… 8

**6.** Sentiment Analysis……………………………………………………………… 8

**6.1.** Methodology……………………………………………………..…………. 8

**6.2.** Model Performance………………………………………………………… 9

**6.3.** Results Analysis……………………………………………………………. 10

**7.** Topic Modeling……………………………………………………………………13

**7.1.** Methodology………………………………………………………………… 14

**7.2.** Model Performance………………………………………………………… 14

**7.3.** Results Analysis……………………………………………………………. 16

**8.** Conclusion…………………………………………………………………………. 17

**9.** Next Steps and Future Work…………………………………………………….. 17

**1.** Objective

More than likely, you have some type of account with a bank – it could be a checking or savings account, a Certificate of Deposit, or some other way that you wish to manage your personal finances. With the advent of mobile banking, in conjunction with traditional brick-and-mortar bank offices, there are a tremendous amount of bank options to select where you choose to manage your money. But with so many banks, both virtual and traditional in-person locations, how do you know the best options for your needs?

“No fee” bank accounts are a thing of the past – presently, practically every bank charges its customers for the use of their bank accounts. Fees like overdraft protection, checking and savings account service charges, and fees for certificates of deposit and minimum deposit balances are normal. Where is a consumer to know what banks provide the customer service and benefits that they desire? Many bank review websites exist, but they may lack a good representation of customers reviews or may contain a small amount of customer reviews.

Our goal is two-fold: 1) to perform sentiment analysis on the customers reviews of the three banks that we selected, and 2) to conduct topic modeling on the same set of customer reviews. Our intention is to predict the most meaningful words for positive and negative bank reviews and to use these most meaningful words to predict the topics of customer reviews that matter most to customers. The flow of these analysis allows us to feed the results of the sentiment analysis (positive and negative reviews) into the topic modeling so that banks have a better understanding of what customers like and dislike most about their services and to also ascertain the topics of customer reviews (I.e., customer service, account fees, interest rates) that are the most meaningful to customers.

**2.** Web Scraping

To do our analysis and start working towards our goal we need to get the data. We chose to get this data from the internet by using web scraping method using BeautifulSoup library in python.

**2.1.** Source

The website we found abundant reviews to scrape the data from is [www.depositaccounts.com](http://www.depositaccounts.com)

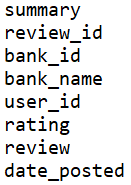


**2.2.** Data

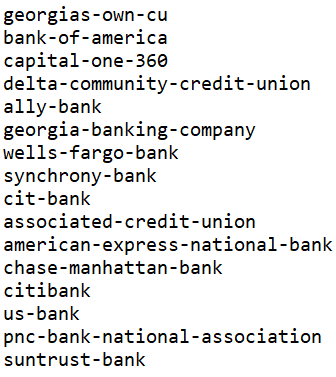
We managed to scrape a total of 3368 reviews from 16 different major banks in the United States. The below is a sample of 5 data points from the scraped data.



List of features in each data point:

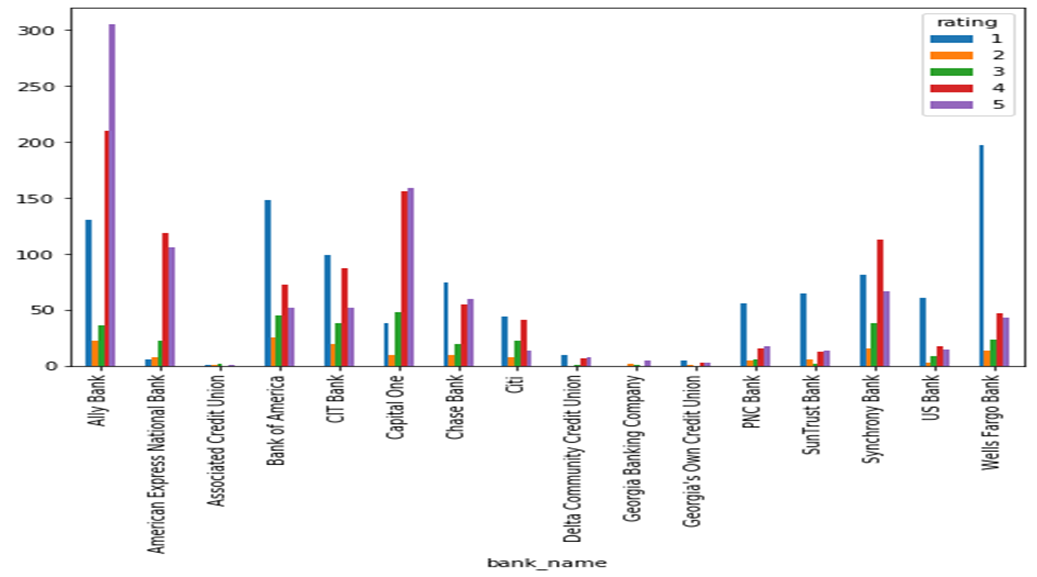


List of banks:

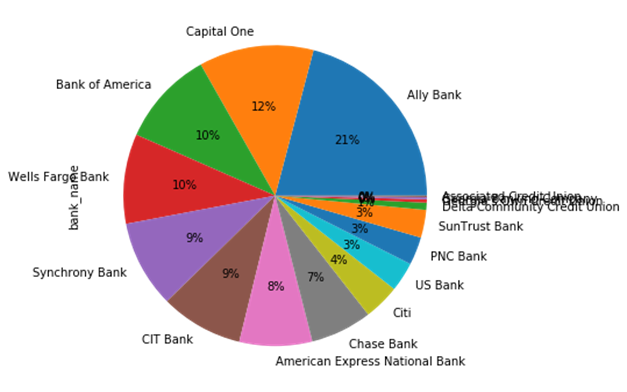


**3.** Data Exploration

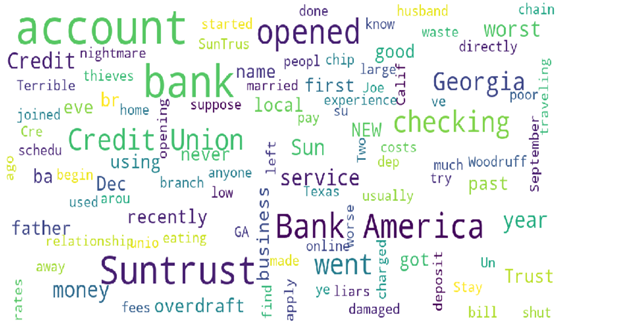
In this case study, we analyzed customer reviews from banks on [www.depositaccounts.com](http://www.depositaccounts.com), the most used website in the United States for bank reviews. Our goal was to gather review ratings and the associated text review of numerous bank customers to determine what matters most to bank customers. After scraping this website, we undertook exploratory data analysis to determine what banks have the highest number of customer reviews and the banks that had an equal distribution of reviews (in order to avoid reviewing banks that had skewed positive or negative reviews). We selected three banks for this exercise: Ally Bank, Bank of America and Capital One based on these selection criteria (see graph below).



Our exploratory data analysis identified that Ally Bank, Bank of America, and Capital One contain 21%, 12%, and 10%, respectively of the entire bank reviews on [www.depositaccounts.com](http://www.depositaccounts.com).



Using Python, our team created a dataframe from the .csv file that was created at web-scraping. We utilized this dataframe to perform a preliminary analysis of the most common words among all the bank reviews:



**4.** Pre-Processing

The goal of text preprocessing is to remove noisy words, and transfer text from human language to machine readable format. Below are steps taken to clean review text.

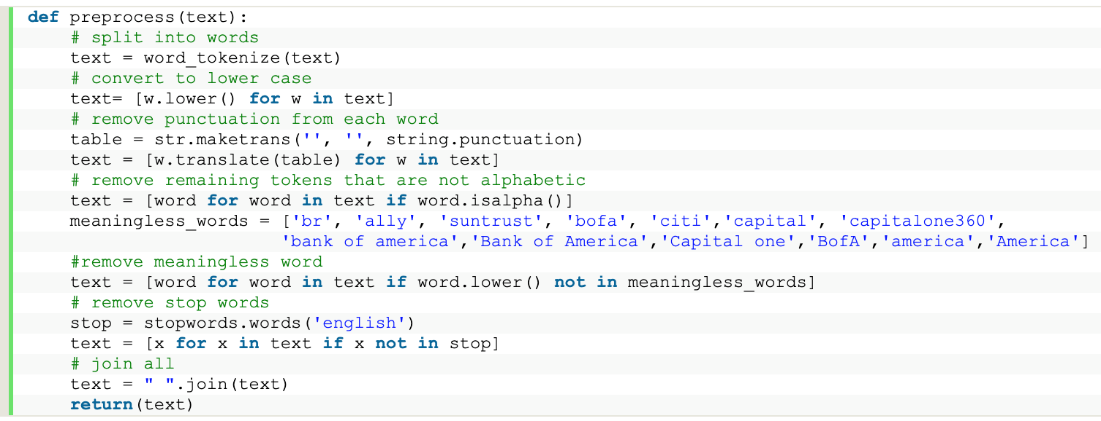
* Split text into words
* Convert text into lower case
* Remove punctuation from each word
* Remove remaining tokens that are not alphabetic
* Remove meaningless words (e.g.: any combination of the name of banks)
* Remove stop words

Some sample code would be show below:

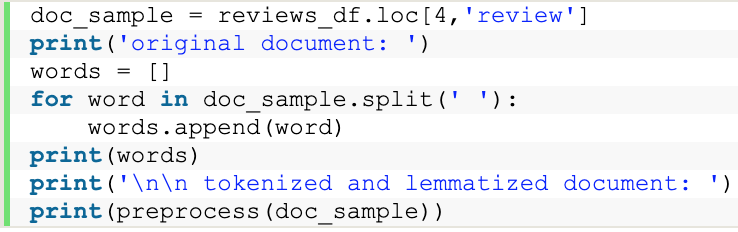
Import library

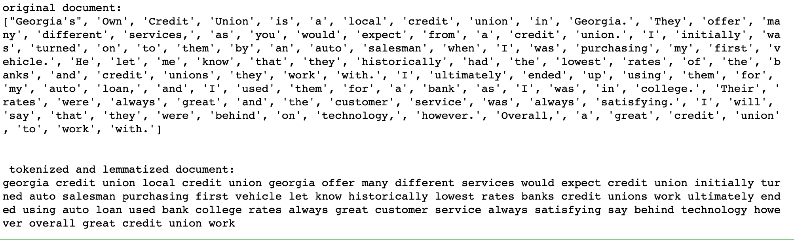


Define a function to clean up review text and join steps together



Checking with sample review





**5.**  Approach

The reviews we collected from the banks which are texts need to be classified based on their polarity scores and then we need to extract the topics that are largely talked about. Then through sentiment analysis we extract the reason behind such topics appearing in the reviews.

**6.** Sentiment Analysis

To identify the sentiments of the customers in their reviews we scraped we decided to start with sentiment analysis and then further investigate on what they are talking about in their reviews.

**6.1.** Methodology

1. utilize doc2vec to implement vectorization

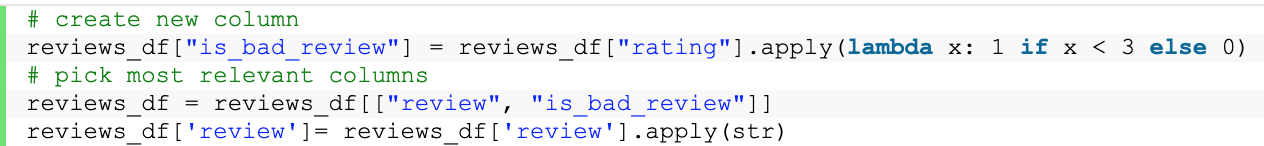
2. utilize Vader to generate polarity score

3. determine the length of review word

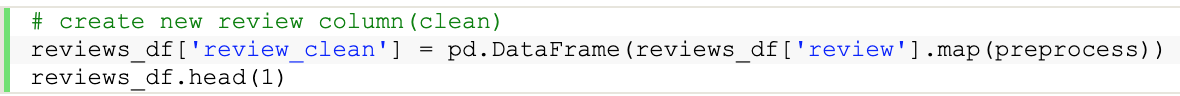
4. determine the length of review character

Below are detail steps of feature engineering

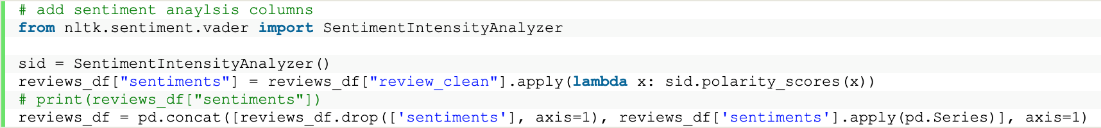
Base on the rating column from data frame, classify ratings of 1 and 2 as bad review, 3,4, and 5 as good review. Select only relevant columns from data frame

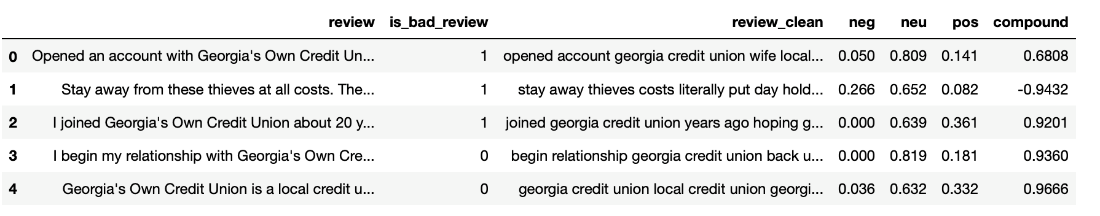


Add a new column with cleaned review text

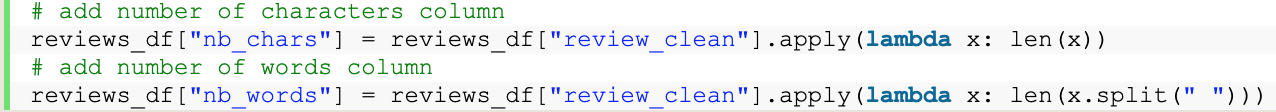


Add sentiment analysis columns. The output generates 4 scores for each review base on the context of the review: positive, negative, neutral, and compound score.

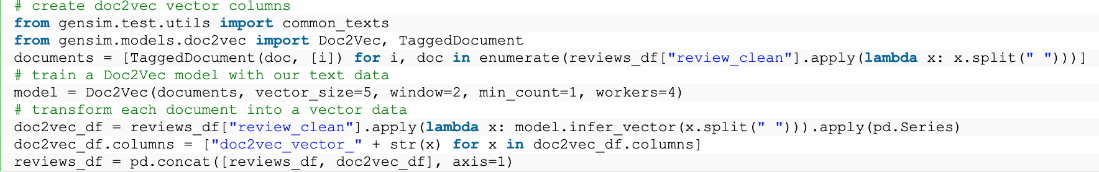




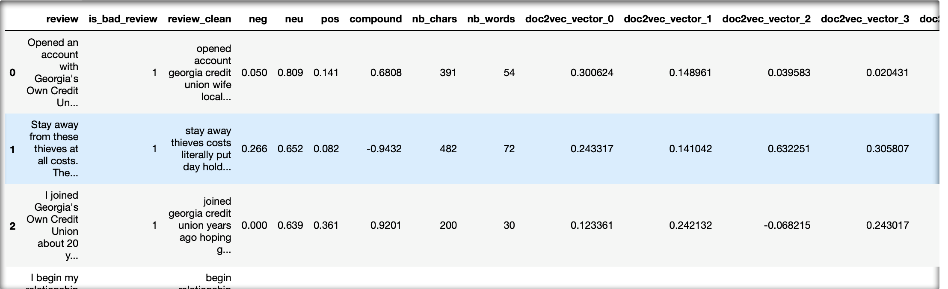
Add number of characters and number of words column



Add doc2vec vector columns

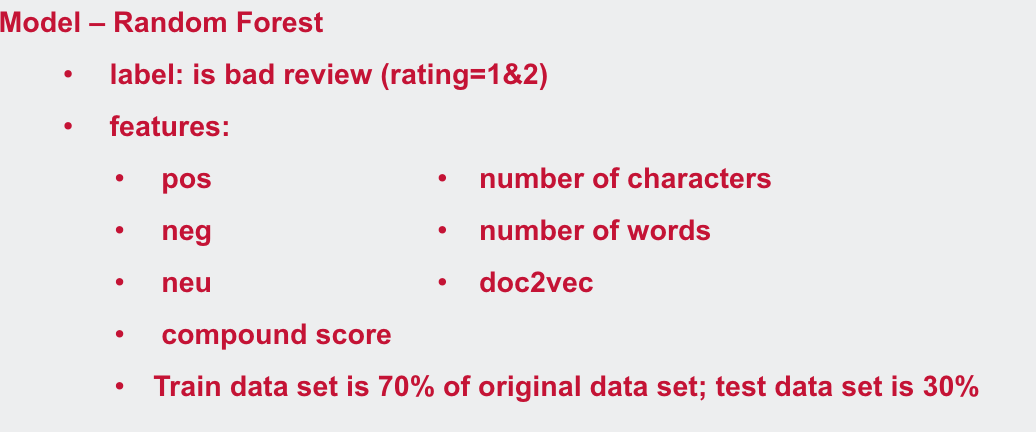


Here is the sample output of how data frame look like

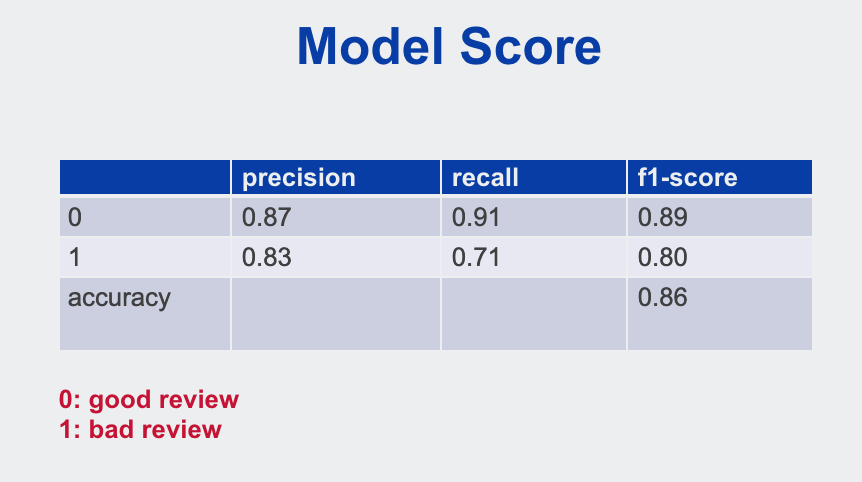


**6.2.** Model Performance

The final model used is Random Forest Classifier with features and label show below

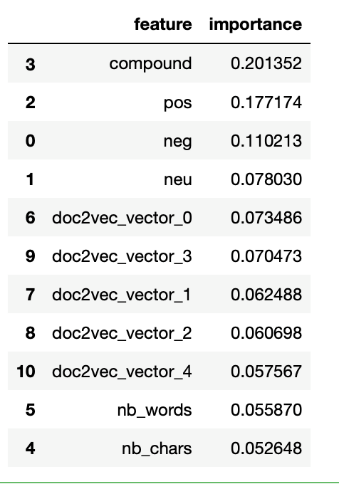


The overall accuracy is 86%.



The feature importance is as follow





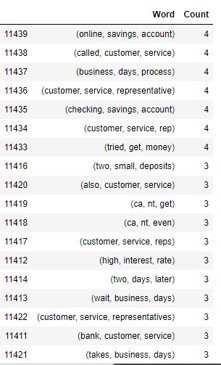
**6.3.** Results Analysis

Using the model, we can classify the data into positive and negative reviews. Since the analysis is done on banks individually to understand the customer feedback and conclude on their strengths and things to work on, we picked banks with highest number of reviews and extracted the topics that are highly spoken in positive and negative reviews separately.

These are the lists of trigrams and their count in their respective sentiment reviews:



Positive Reviews Negative Reviews



Positive Reviews Negative Reviews



Positive Reviews Negative Reviews

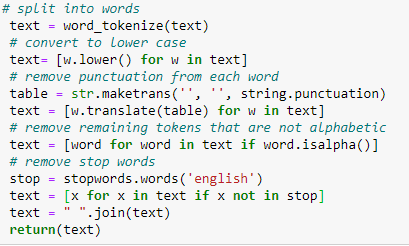
It is very insightful to see how several sets of words appear more times than any other in negative and positive reviews respectively. Which translates as the most times these things are talked about in negative or positive reviews. This gives us the areas an entity is lagging or doing great but doesn’t give the reason behind them. This is where topic modeling can help us understand the reason for which customers are expressing their concern and can work on to improve it.

**7.** Topic Modeling

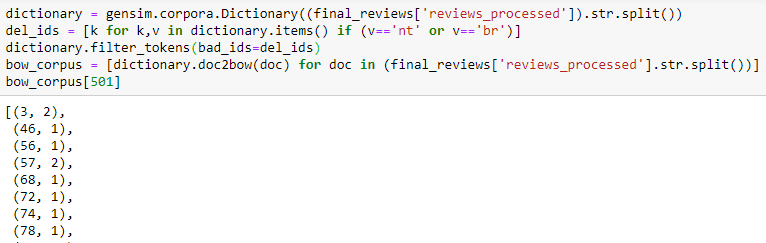
Topic Modeling attempts to classify documents into a distribution of predefined topics. After identifying set of words that fall under positive or negative reviews through sentiment analysis, do we exactly understand what is happening in that particular review? What if we want to know more about it? We first attempted to classify reviews into various topics by implementing Gensim LDA topic modeling algorithm. Then we defined labels to each topic that gave more meaningful topics. It also makes identification of particular review easier based on bi-gram or tri-gram.

**7.1.** Methodology

The preprocessing of raw data is similar to that of sentiment analysis that includes removing stop words, lemmatization of text, removing punctuations, etc.



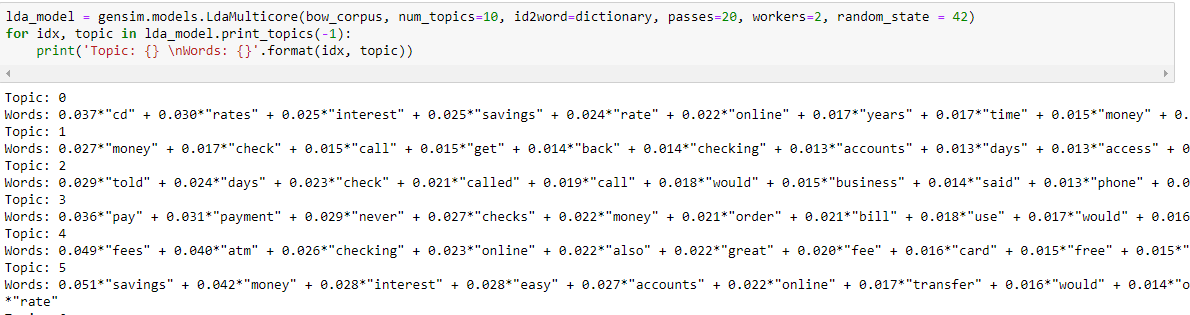
We used **Bag of Words** to convert the text into vectorized format so that we get the corpus that can be fed into the model.

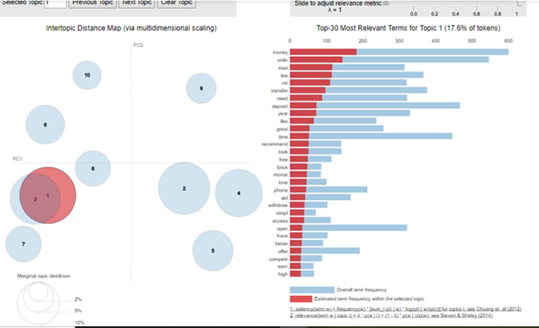


**7.2.** Model Performance

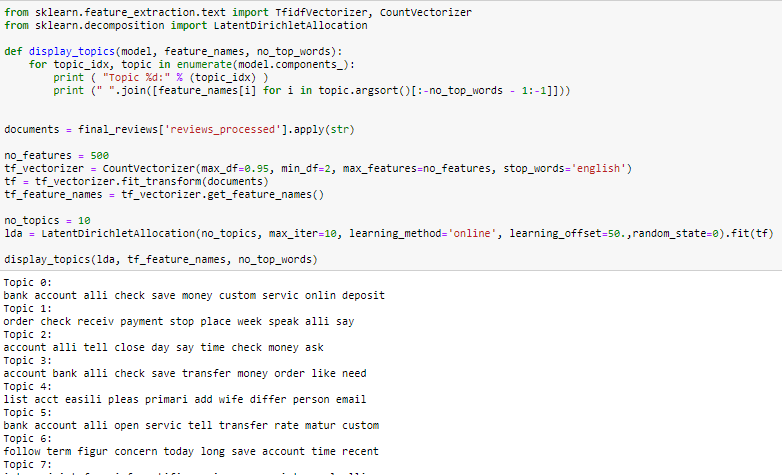
Then we ran two algorithms: Gensim LDA model and SKLearn LDA fitting the above corpus

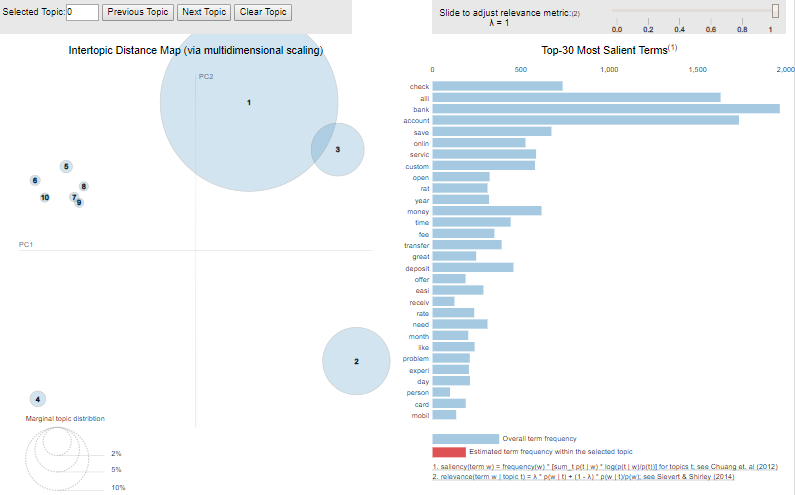
Gensim LDA:





SKLearn LDA:



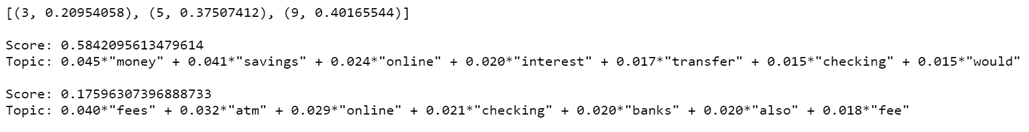


**7.3.** Results Analysis

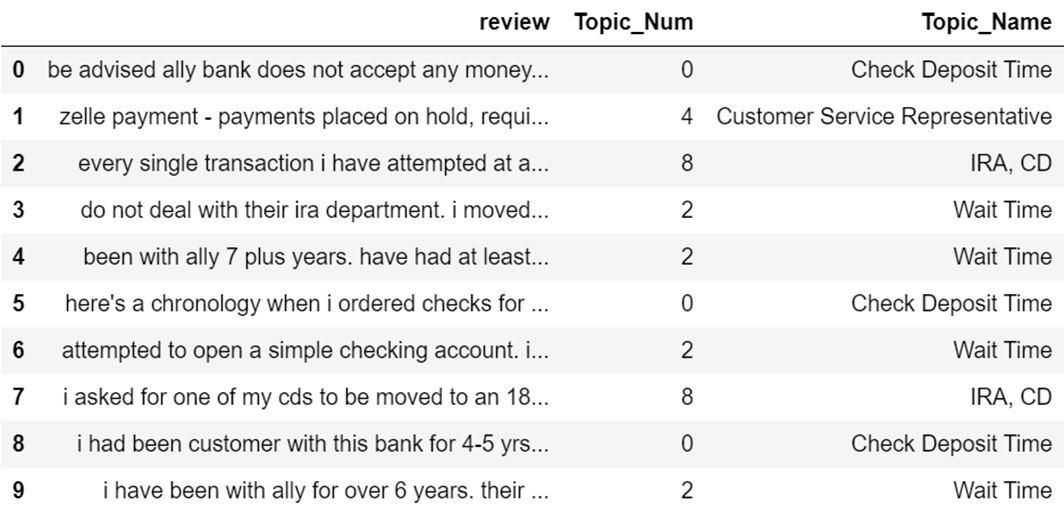
Let’s see the actual review and how model categorized it into topics

'deposits are easy, and online transactions in most cases, are available within 24 hours. the mobile app makes depositing checks easy and convenient, and the mobile app is easy to navigate. you can't ask for much more from a banking institution.**'**

**Topic Assigned:** Mobile App



Below table shows how reviews are assigned to respective labels based on Topic\_num which is the maximum probability topic that review can belong to



**8. Conclusion**

Our hypothesis that parallel workflow of sentiment analysis and topic modeling would produce insightful results proved to be true. It also gives us scope to dig deeper to get more insights.

**9. Next Steps and Future Work**

1. We can make use of summary of reviews column that we got from web scraping to see if we can explore more
2. We can also explore more on how we can label each review to more than one labels.