Project Report for IAL620-01: Text Mining & Natural Language Processing

Chatbot, Speech Recognition, UNCG Emails, NYCTimes data for Text Analytics, sentiment analysis

Bala Mallampati Friday, December 10, 2021
Master's in informatics and Analytics
University of North Carolina, Greensboro, USA
b_mallampat@uncg.edu

Abstract:

After pandemic (Covid-19) started, every organization tracks their work from home employee activities digitally to predict and improve their business operations and employee's work time & increase their wellness time, retention rate.

In this project, I seek to answer the following data research questions:

A. Create chatbots and speech to text functionality to digitalize organization operations and better customer experiences as part of Data Generation process.

B. Web scrape NYCTimes data & Extract UNCG emails data to analyze and convert to numbers as part of Text Analytics PCA (Principal Component Analysis)

C. Run exploratory analysis on UNCG Email data and show comparisons using Business Intelligence tools and also do sentiment analysis on subjects.

D. Clean and Visualize NYCTimes Top Words, Bigrams, Correlate words and run sentiment analysis, Topic Modeling to get top topics in news.

E. Integrate NLP (Natural Language Processing) to BI tools (Power BI, Tableau) and present extracted insights to organization to improve business operations.

Introduction and Motivation:

Every organization wants to digitalize their operations and minimize human tasks, improve their employee satisfaction rating and retention rate. Every Text analytics and NLP(Natural Language Processing) needs accurate & digitalized data to predict employee's & customers satisfaction rate.

I love to play with data, now "The world's most valuable resource is no longer oil, but data" as per The Economist article (https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data).

Chatbots are most useful in day to day operations of Business, mainly Rule based, Self-Learning(Retrieval based , Generative) Chatbots.

Text data will be useful for employee's sentiment analysis and predict their satisfaction rate. Email's classification using ML models for Spam, Important emails, collaboration, Network emails, Communication habits.

In future, I will to extend this project for web scrap organization complaints from websites (google reviews, company portals, Face book Pages..) and internal employees chats analysis to determine employee satisfaction rate, retention rate.

Data Source and description:

I have used Chatbot input text with fixed labels and for exploratory analysis & sentiment analysis from below sources.

Data extracted from below NYC Times news website. I had extracted latest top articles using html_nodes in R.

https://www.nytimes.com/ → It has total 5 features and 68 observations Text column data.

NYC Times

Attribute	Description	Type	Data Sample
Rownum	Serial Number	Int64	1,2,3..
URL	Uniform Resource Locator	String	https://www.nytimes.com/2021/12/10/us/politics/texas-abortion-supreme-court.html
Title	NYC Website Title	String	Supreme Court Allows Challenge to Texas Abortion Law but Leaves It in Effect - The New York Times
Date	Date, News posted	Date	12/10/2021
Text	NYCTimes News description	String	Advertisement Supported by The law, which bans most abortions after about six weeks of pregnancy, was drafted to evade review in federal court and has been in effect since September. Send any friend a story As a subscriber, you have 10 gift articles to give each month.....

https://mail.google.com/mail/u/0/#inbox → It has total 13 features and 2123 observations Text column data.

UNCG Email Outlook:

Attribute	Description	Type	Data Sample
Received	Email Received Date	Date	10/28/2021 11:39:00 AM,9/27/2021 12:25:00 PM
Created	Email created date	Date	Thu 10/28/2021 10:23 PM, Thu 9/23/2021 1:38 PM

From	email Sender name	String	UNCG Athletics, Genevieve Smith
To	Email receiver Name	String	inan-students-l@uncg.edu, Bala Mallampati
Subject	email subject	String	Re: Capstone Presentation TONIGHT at 5:00 PM - Zoom
Size	Email Size in KB, MB	String	2 MB,4 KB
Conversation	If same topic discusses then its Conversation	String	Capstone Presentation TONIGHT at 5:00 PM - Zoom
Categories	Email categories	String	[bridges2-users] Bridges-2 (Including VMs and Filesystems)
Sensitivity	email Organization sensitive flag	String	Normal, High, Low
Contacts	Same as receiver email id	String	b_mallampat@uncg.edu
Cc	CC emailer names	String	Bala Mallampati; Richa Kurkure
Sent	Email Sent from sender time	Date	12/9/2021 6:01:00 PM, 12/9/2021 10:15:00 AM
Follow Up Flag	Follow-up flag enabled time	Date	10/28/2021 11:10:00 AM,9/27/2021 12:10:00 PM

Methodology:

Data processing is an important step for in the data analysis. Data science involves methods of analyzing massive amounts of data for the purposes of knowledge extraction. It evolved from statistics and traditional data management. Data comes in many shapes and forms, and many times we need to get it ready to be able to analyze it. The phrase "garbage-in and garbage-out" is particularly applicable to text mining to Train and Test Data.

In this project I have used the following languages, frameworks, tools, libraries, packages from web scrape data to do Sentiment analysis and extract from UNCG email.

Technologies & Libraries: R(ndjson, tidyverse, tm, lubridate, fliptime, stringr,ggmap, maps, quantenda, readtxt, textplots, widyr, topicmodels, ldatuning, lubridate, rvest), Python(numpy, pandas, matplotlib, seaborn, sklearn, google.colab), R Selenium

Tools: Rstudio, Colab for python, Power BI, Tableau, Outlook 365.

Chatbot creation using Python Chatterbot, Chatterbot_corpus, ChatterBotCorpusTrainer

Speech Recognition using Google Colab and Google cloud platform developer authentication.

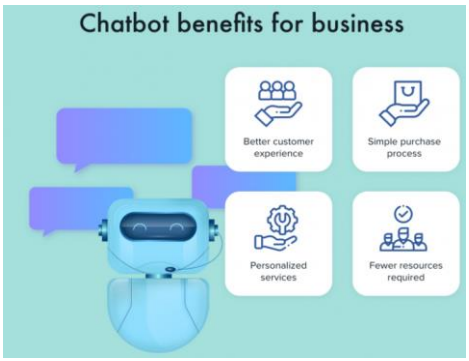
Power BI with Text Analytics – It needs Power BI Premium License, So imported Wordcloud Visualization.

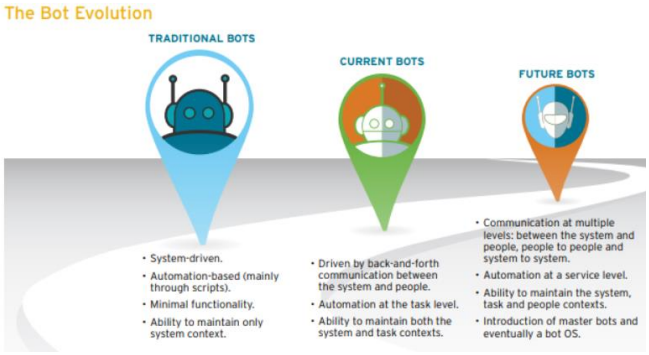
Tableau Prep Builder BI tool integrate with Text Analytics using Python Script – In progress.

Chatbot Creation:

Today, we have smart AI-powered Chatbots that use natural language processing (NLP) to understand human commands (text and voice) and learn from experience. Chatbots have become a staple customer interaction tool for companies and brands that have an active online presence (website and social network platforms) .

I have created simple chatbot using python files and below libraries.






```
from chatterbot import ChatBot
from chatterbot.trainers import ListTrainer
my_bot = ChatBot(name='PyBot', read_only=True,
                 logic_adapters=['chatterbot.logic.MathematicalEvaluation'
,                                'chatterbot.logic.BestMatch'])
```

Output:-

```
[ ] print(my_bot.get_response("do you know the law of cosines?"))

    c**2 = a**2 + b**2 - 2 * a * b * cos(gamma)

    print(my_bot.get_response("what\'s your name?"))

    i'm pybot. ask me a math question, please.
```

```
[ ] print(my_bot.get_response("How is weather today?"))

    Its cold outside with 32 degree F.
```

Speech Recognition:

Google Colab speech recognition also created using Google Authentication Json File.

https://console.cloud.google.com/iam-admin/serviceaccounts/details/105018423178735813720/keys?project=balamallampatiuncg

```
!pip3 install google-cloud-speech
!curl -
LO https://github.com/mozilla/DeepSpeech/releases/download/v0.6.0/audio-
0.6.0.tar.gz
!tar -xvzf audio-0.6.0.tar.gz
```

```
!ls -l ./audio/
```

```
from google.colab import files

uploaded = files.upload()

for fn in uploaded.keys():
  print('User uploaded file "{name}" with length {length} bytes'.format(
      name=fn, length=len(uploaded[fn])))
```

```
Choose Files  No file chosen    Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.
Saving balamallampatiuncg-f201d9967188.json to balamallampatiuncg-f201d9967188.json
User uploaded file "balamallampatiuncg-f201d9967188.json" with length 2337 bytes
```

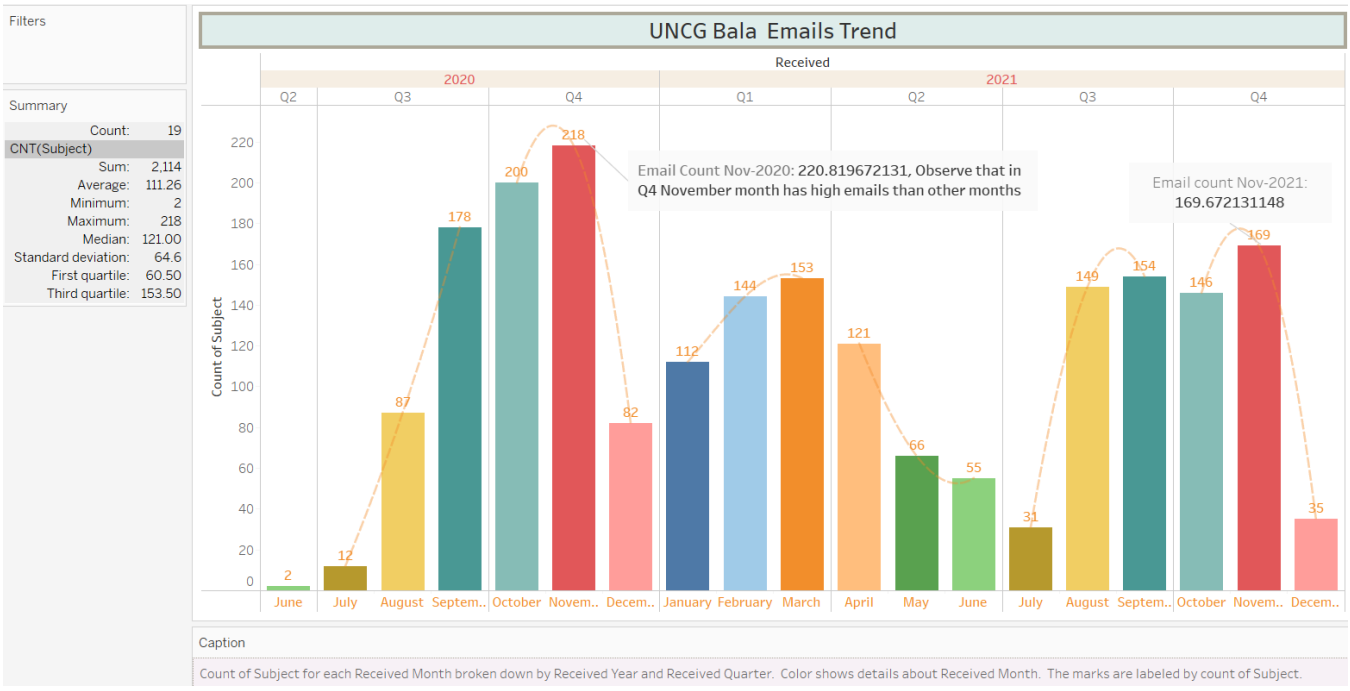
```
[ ] !pwd
    !ls -l ./balamallampatiuncg-f201d9967188.json

    /content
    -rw-r--r-- 1 root root 2337 Dec 10 19:28 ./balamallampatiuncg-f201d9967188.json
```

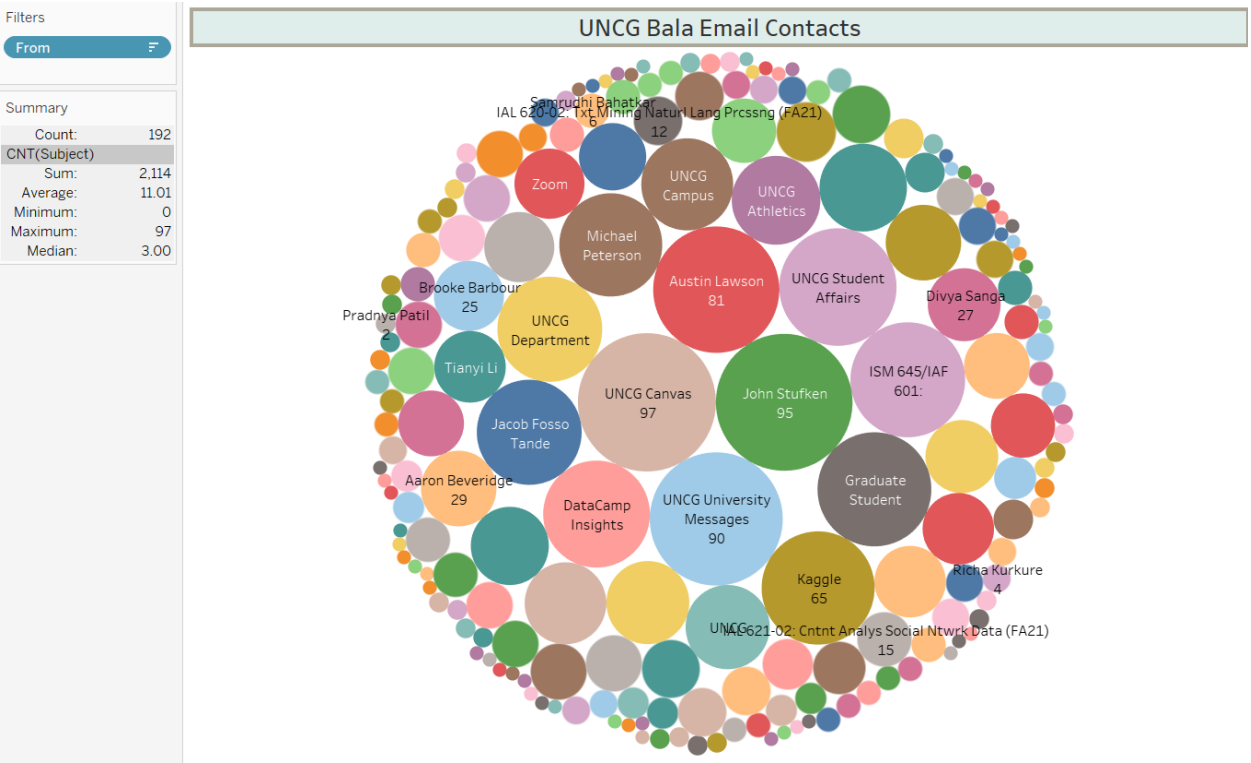
It accesses your local computer Input devices (audio and video).

Data Analysis:

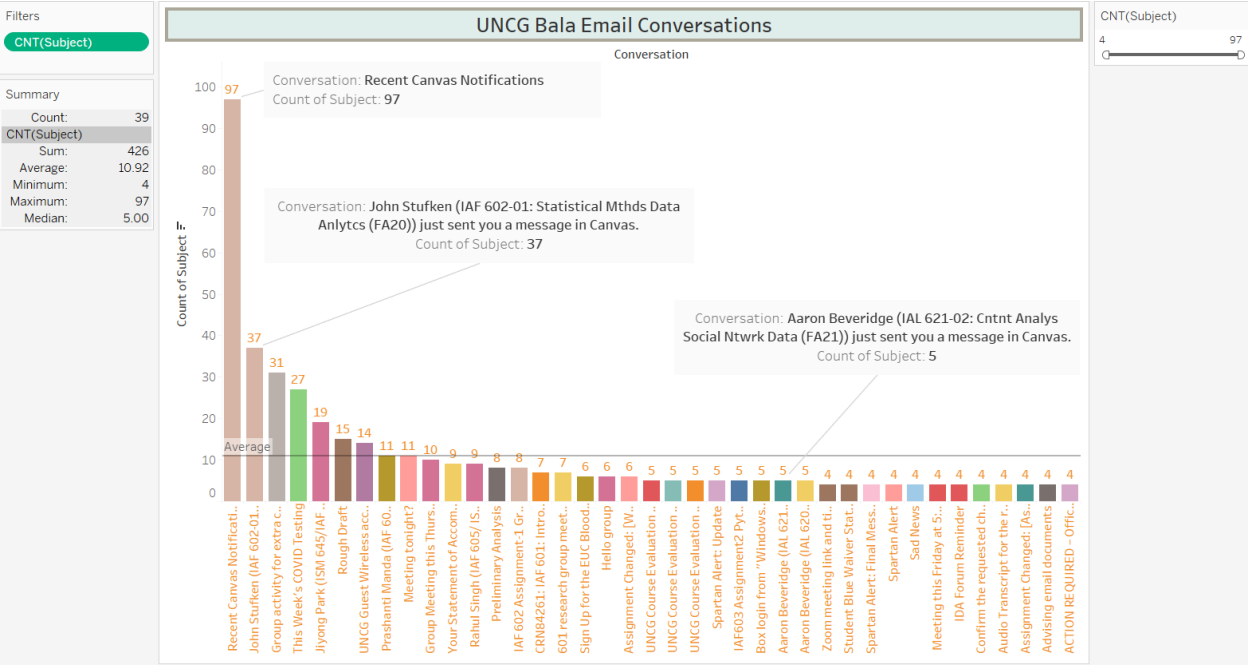
UNCG email data Exploratory analysis and comparisons of all metrics. Below Tableau graph shows that over all monthly trend and it shows Quarter4 Nov-2021 has highest email usage.



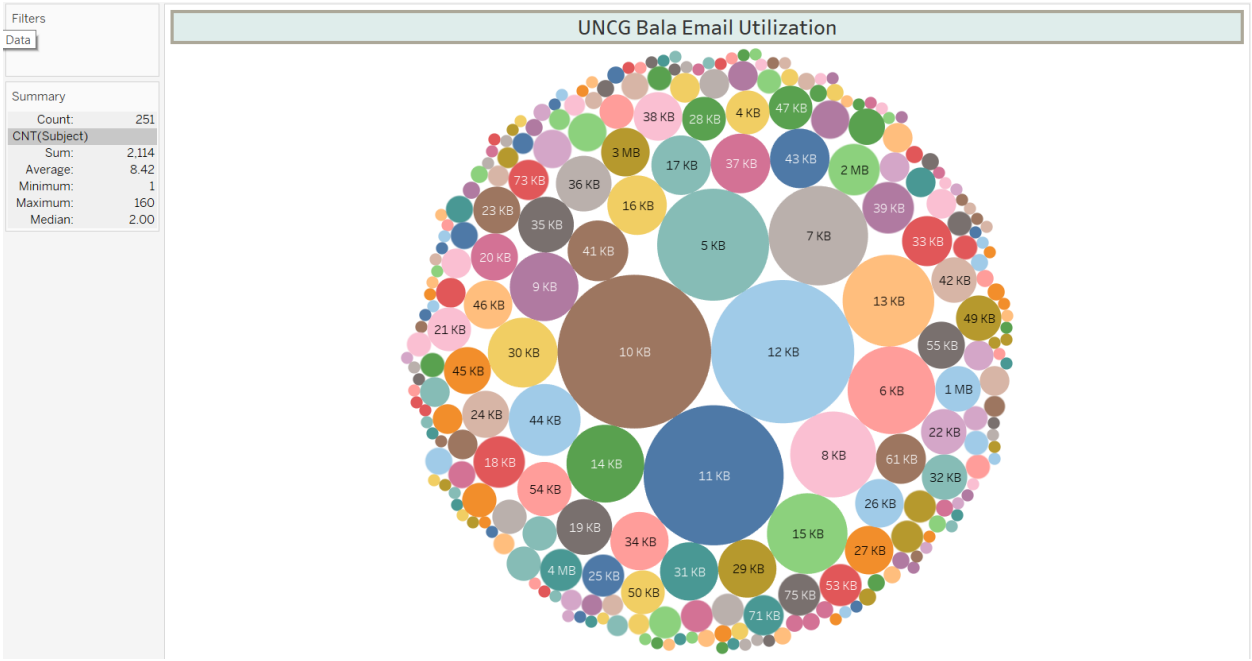
UNCG Email contacts graph shows that who are all email senders by their email count. Example: Dr Aaron sent 29 emails so far to me and Dr Stufken sent 97 emails..



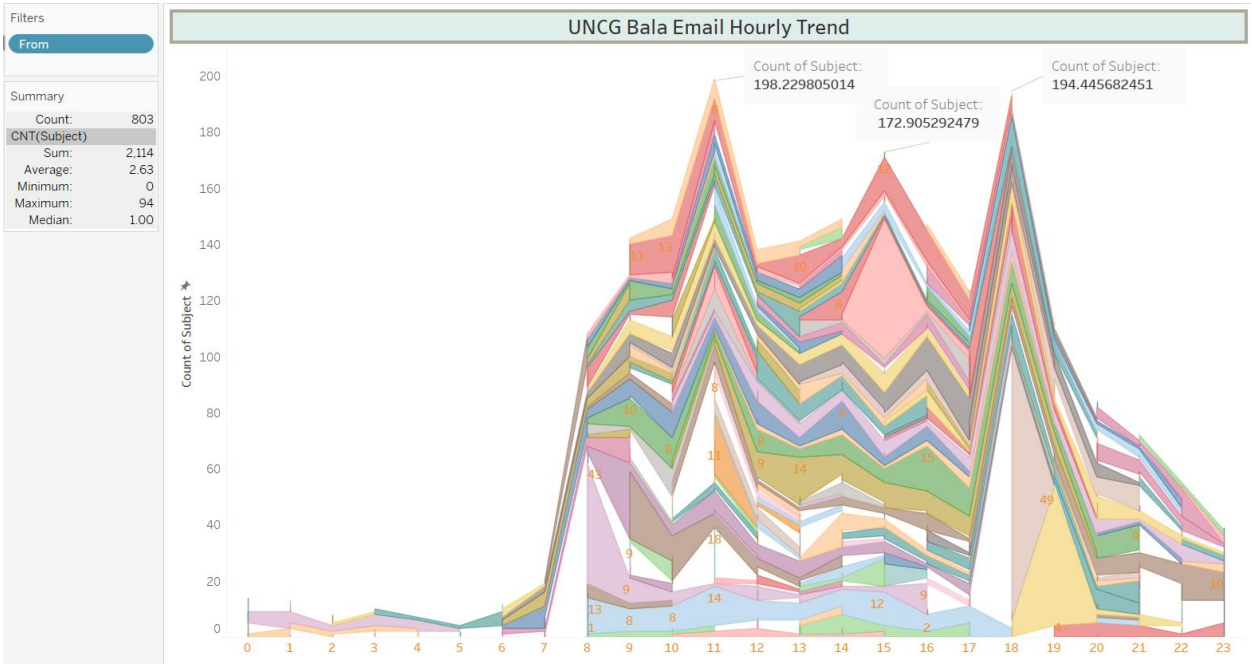
UNCG Email Conversations shows that Recent Canvas Notifications with 97 emails.



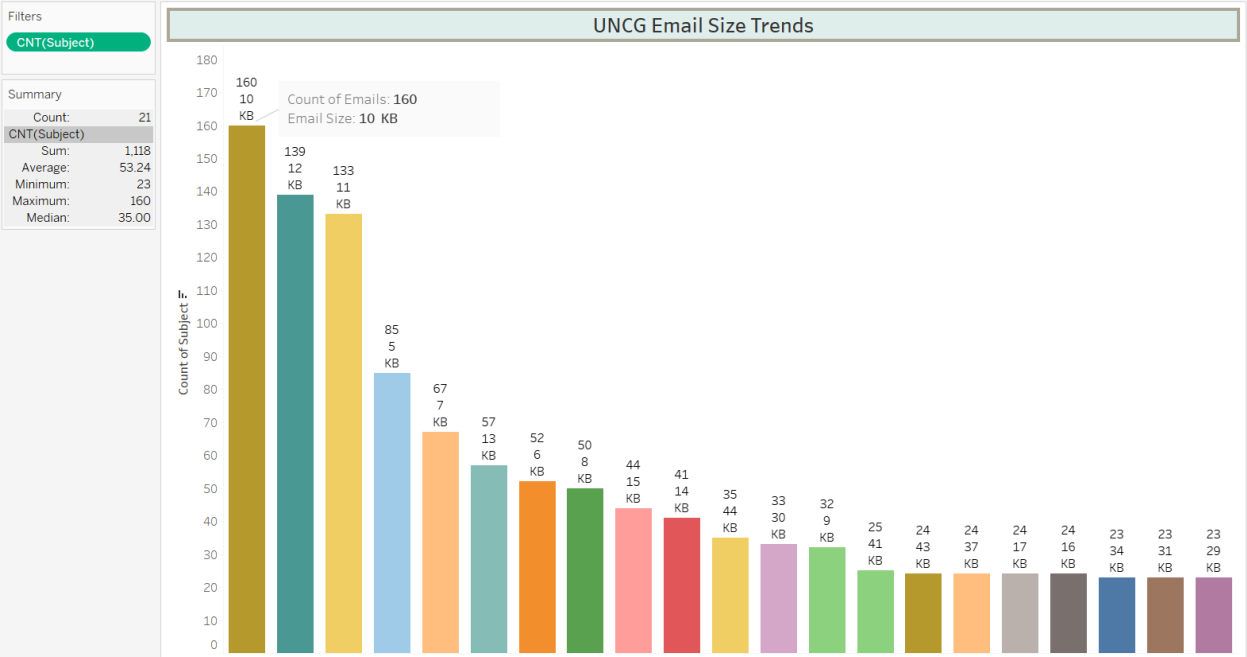
Similarly, below analysis shows email sizes and by count wise.



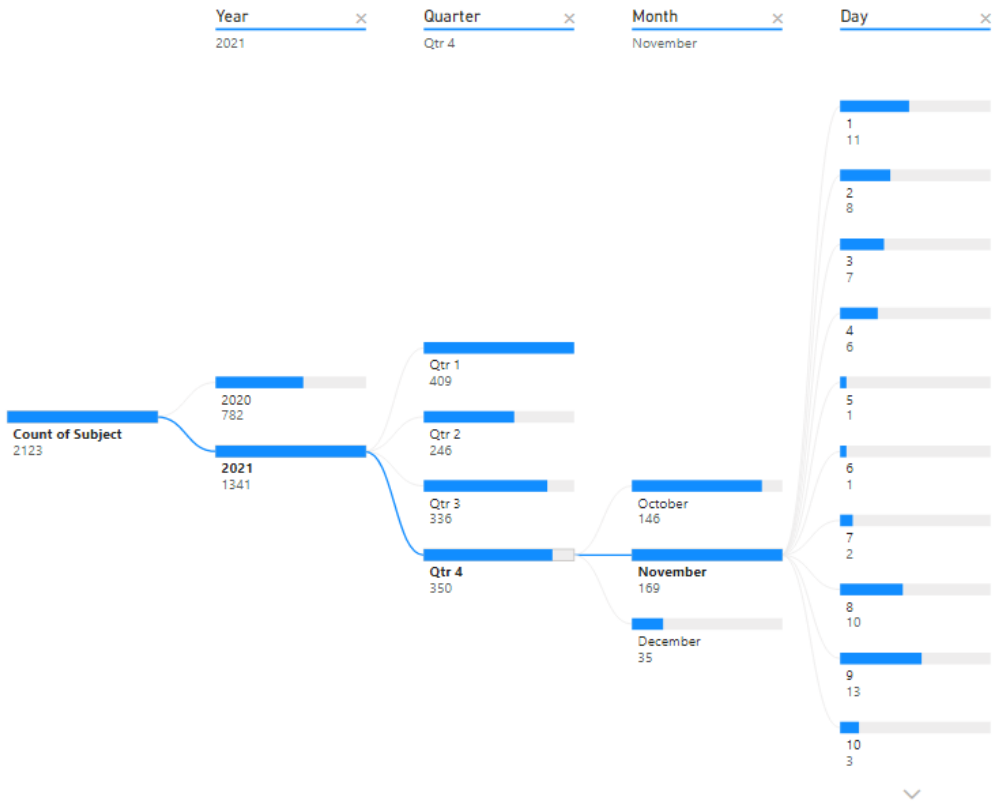
UNCG Email Hourly trend is very interesting graph to understand UNCG staff, students, employees working hours. I see that at 11AM, 3PM and 6PM most of users sending emails.



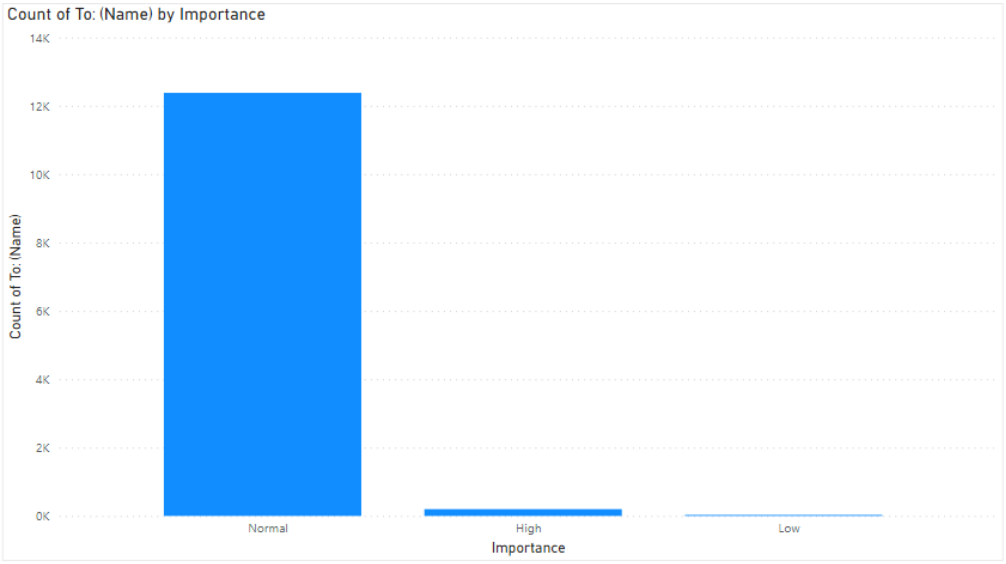
Another analysis run on email size wise counts. Most of them <20KB, so no load on SMTP/IMAP UNCG servers.



Power BI Decomposition tree for UNCG email counts track by year, quarter, month, and day.

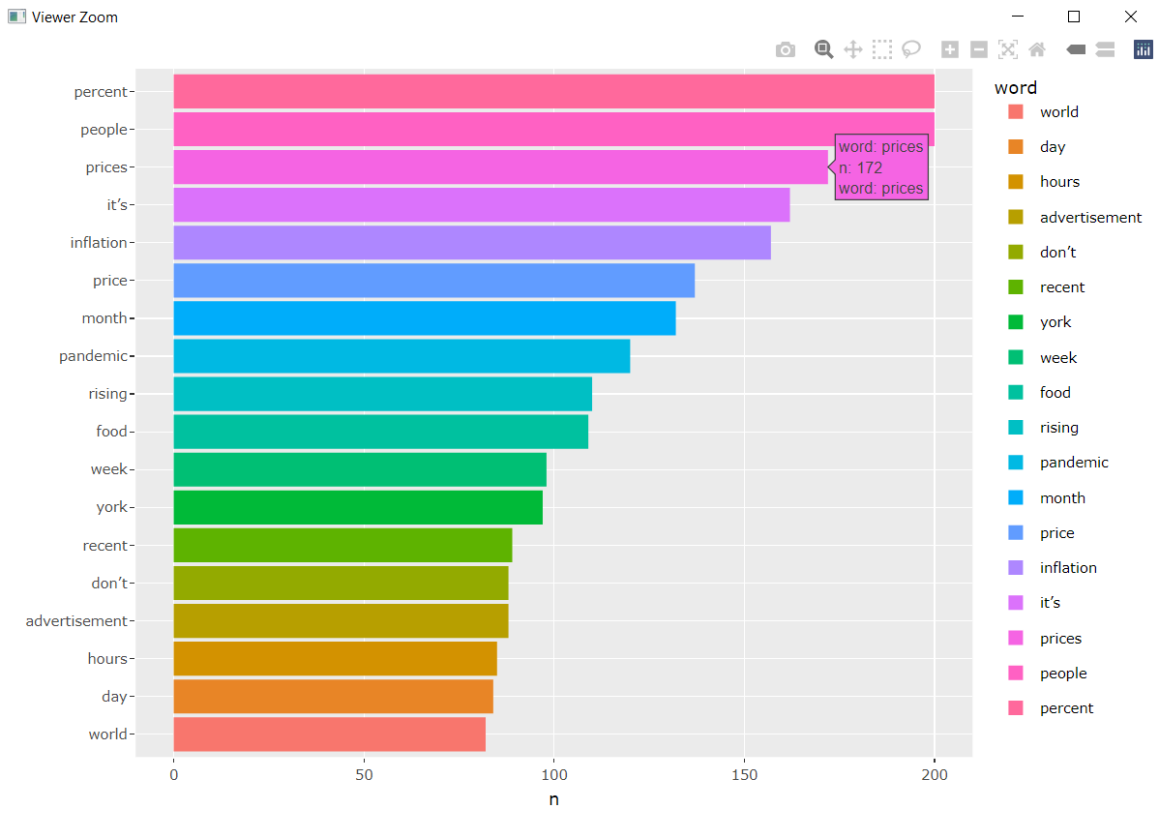


Below Power BI graph shows Email importance and their counts. Most of them are Normal emails.



NYC Times Exploratory Analysis:

Below R plot shows what are Top words in NYC Time News based on text column. Ex: Price, pandemic, food



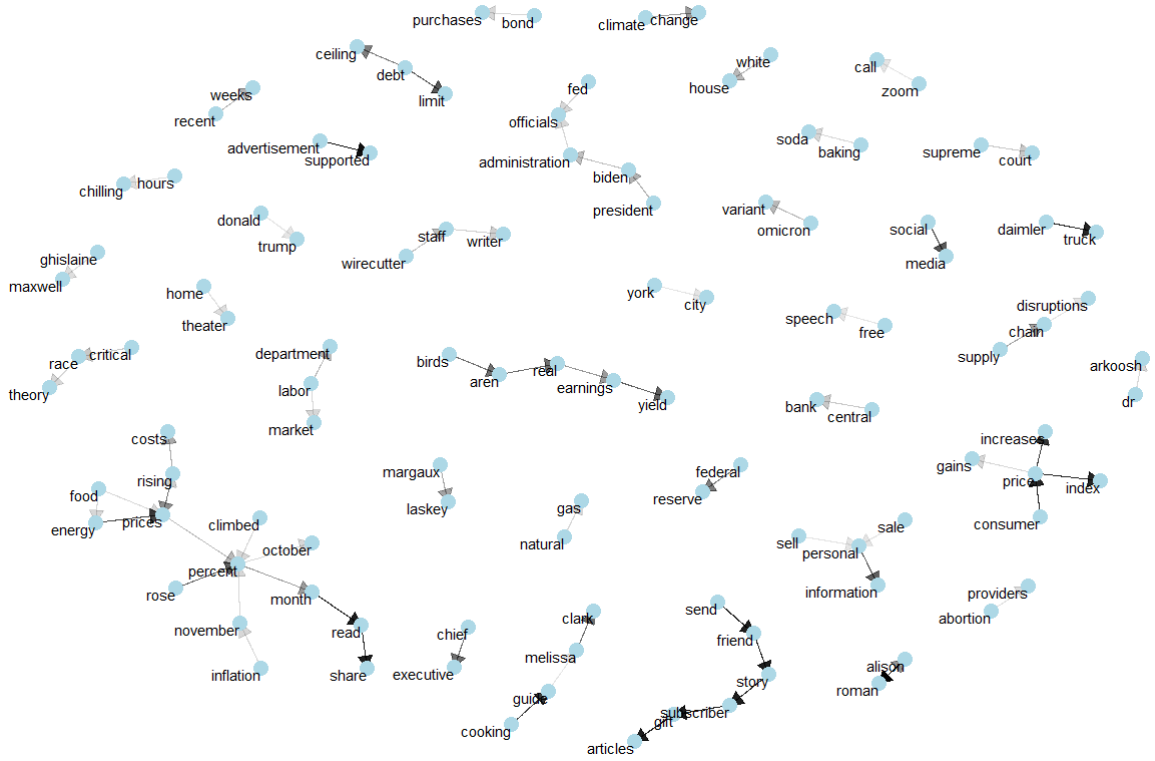
Word Embeddings:

Below code shows NYCTimes news text wordvectors and their context.

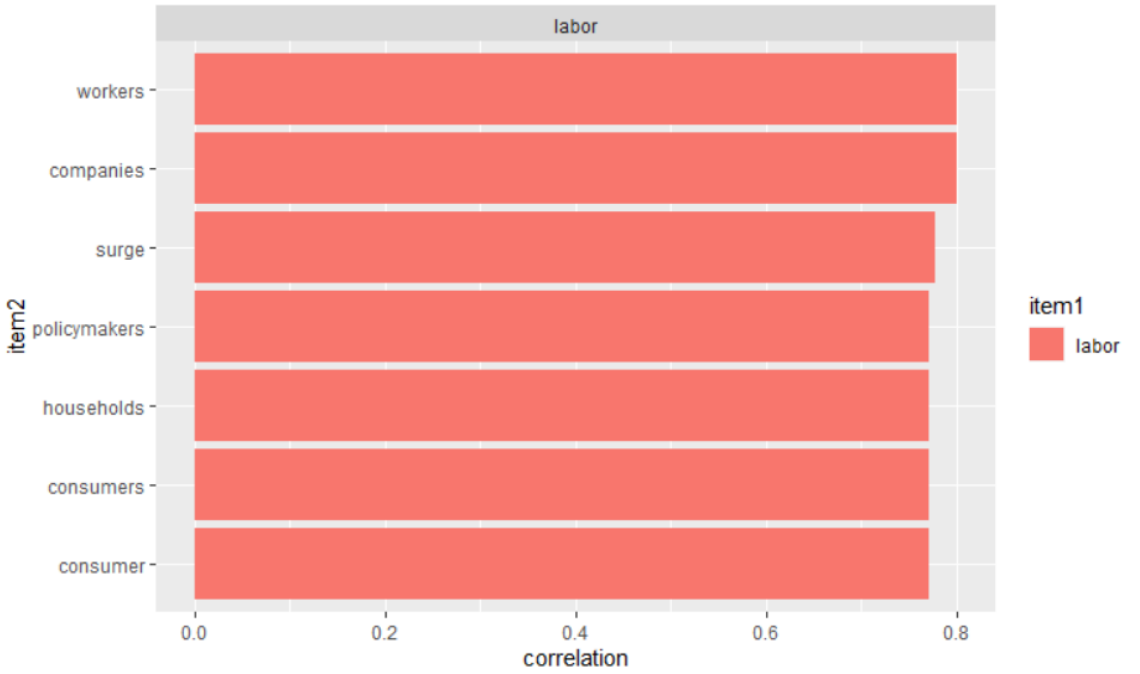
```
> ####################################
> # explore context for individual words
> ####################################
>
> cos_sim = sim2(x = word_vectors, y = word_vectors["books", , drop = FALSE], method
 = "cosine", norm = "l2")
> head(sort(cos_sim[,1], decreasing = TRUE), 5)
      books     letters approaching        week        list
  1.0000000   0.4719539   0.4374036   0.4319884   0.4233033
>
> ##############################
> # test word contexts/analogies
> ##############################
>
> test.word = word_vectors["books", , drop = FALSE] -
+   word_vectors["brush", , drop = FALSE] +
+   word_vectors["chair", , drop = FALSE]
> cos_sim = sim2(x = word_vectors, y = test.word, method = "cosine", norm = "l2")
> head(sort(cos_sim[,1], decreasing = TRUE), 5)
    chair     books     virus   receive  separate
0.6819302 0.4476177 0.4231864 0.4115812 0.4100632
>
```

Correlation:

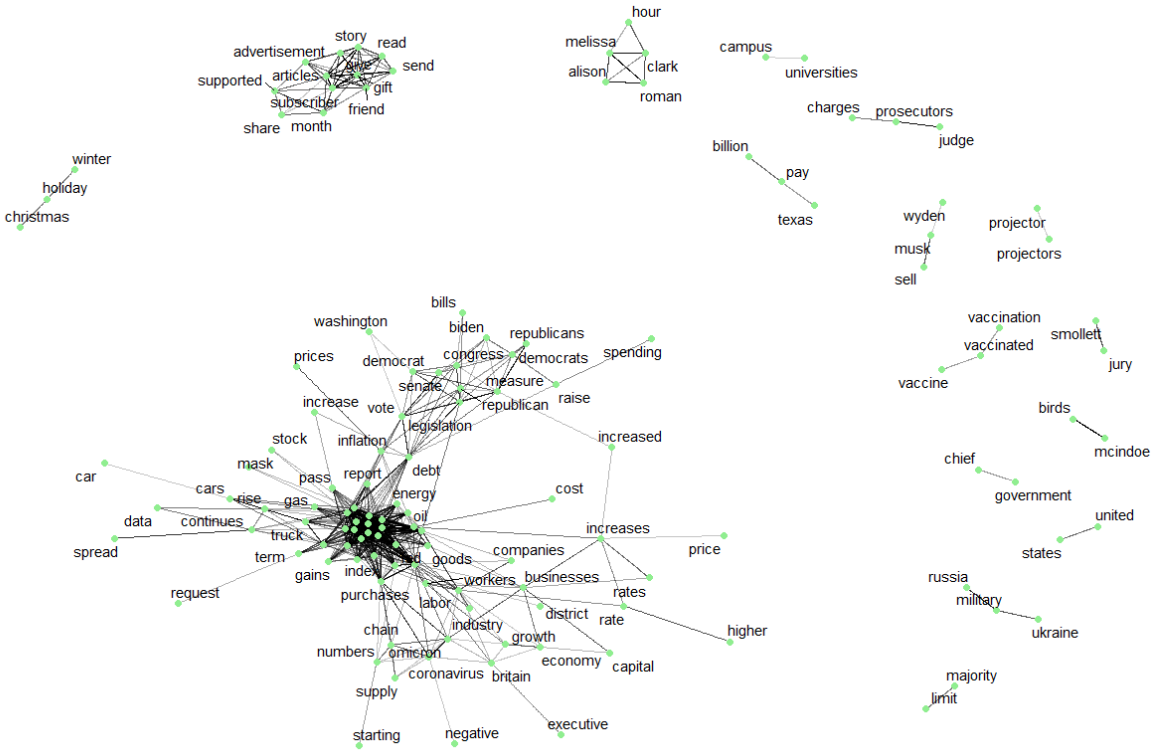
I also ran correlation between all top words in NYC Times text data, below shows highly correlated words. Example: Climate change, Donald Trump, Cost Rising.etc



Below R plot shows Labor word correlation with other words in news text. It shows it highly correlated with Workers & companies.

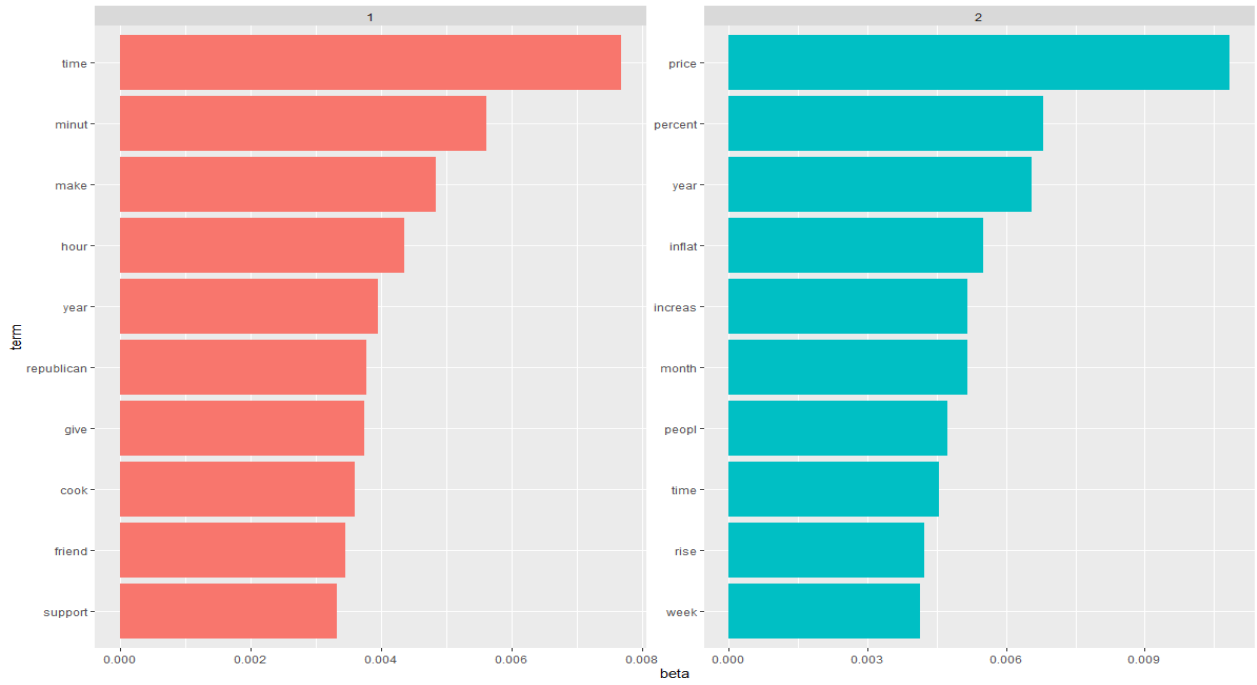


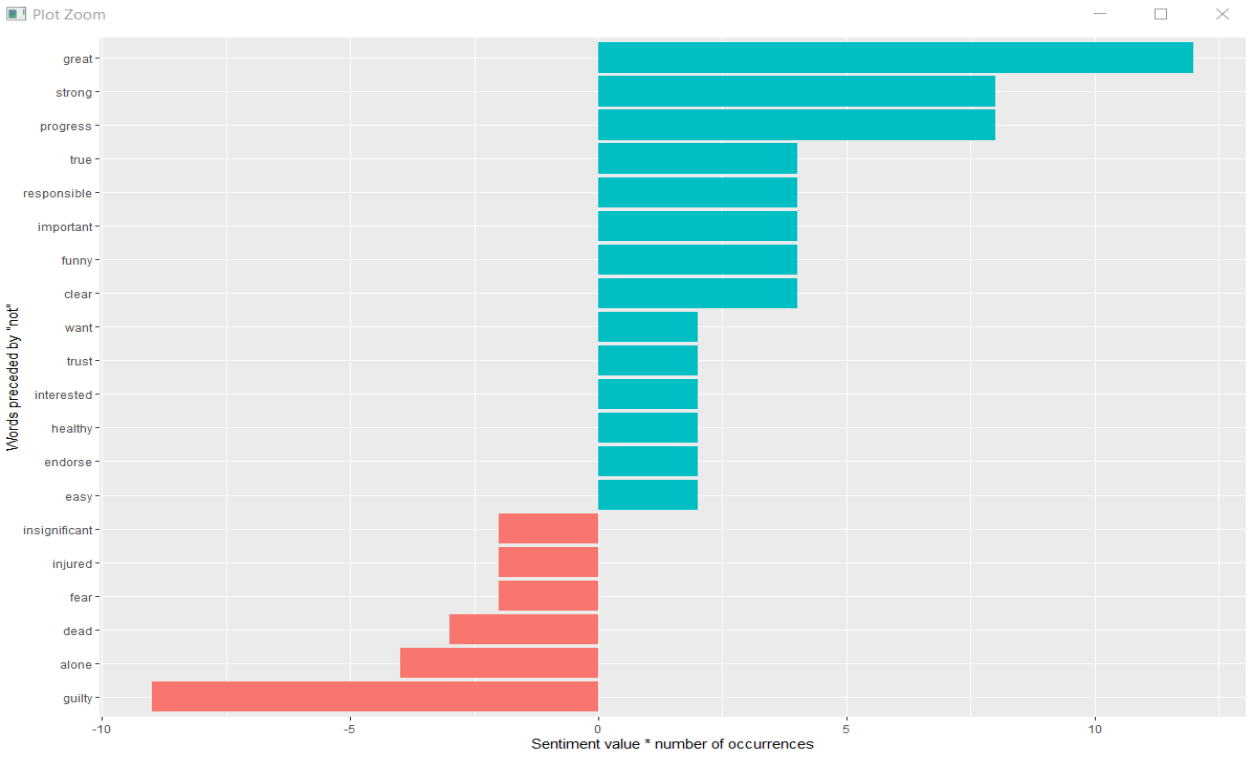
I ran correlation >0.7 to minimize words for bigrams creation.



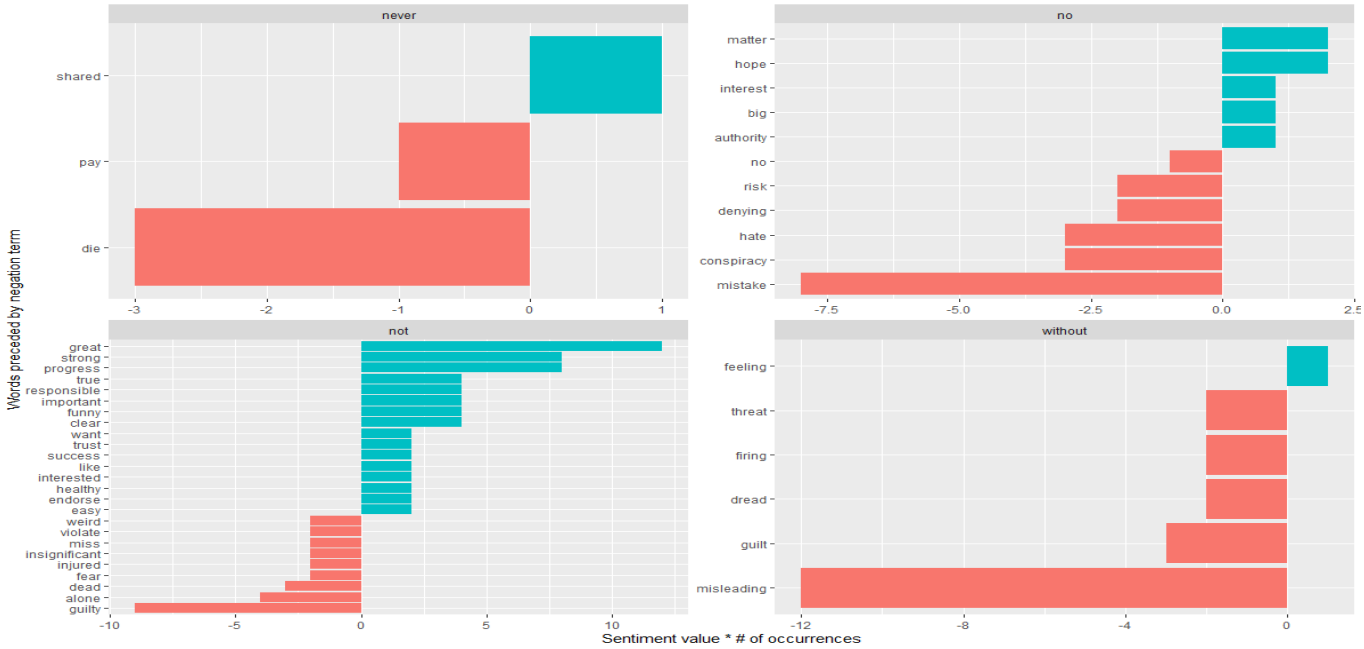
Sentiment Analysis NYC Times:

One of my research questions to run sentiment analysis on news data, below 2 plots shows that Positive and negative words almost same and little higher positive side.



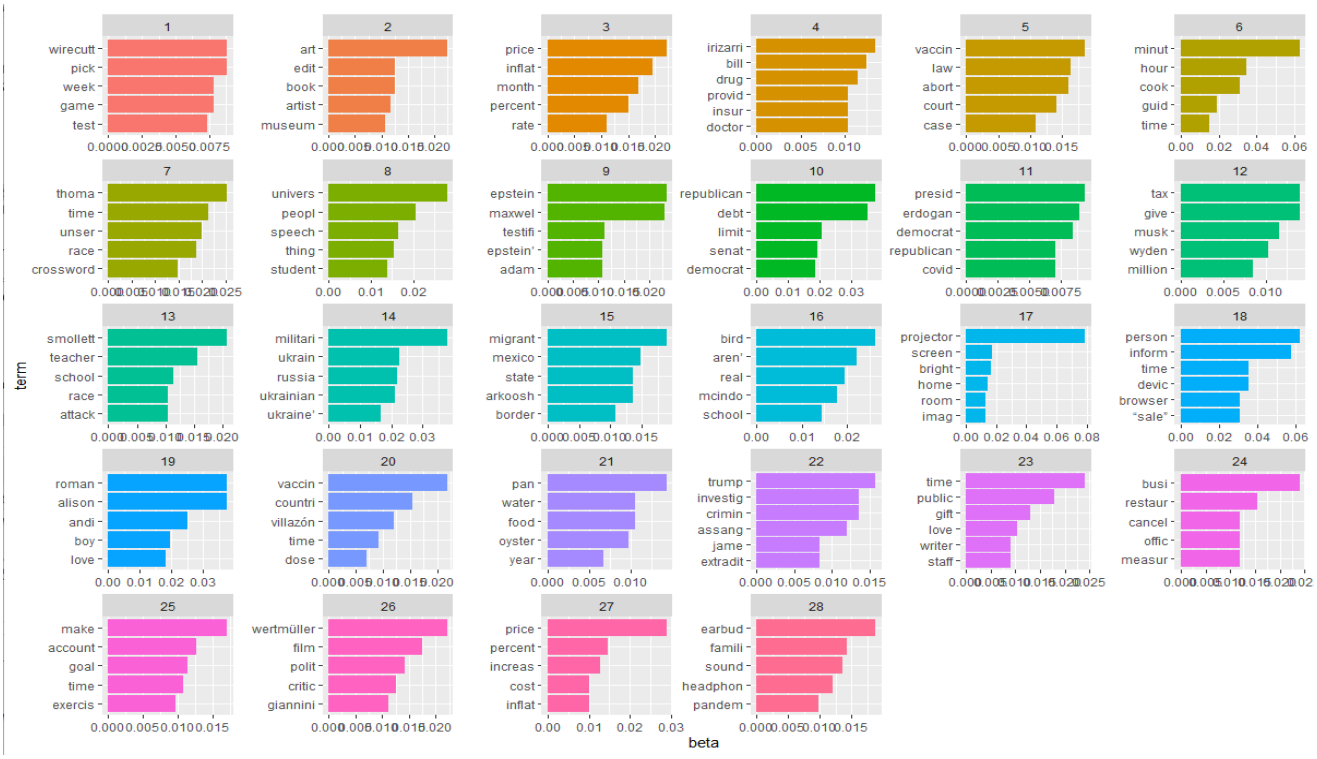


Also checked words proceeded by negation terms and observe below words.

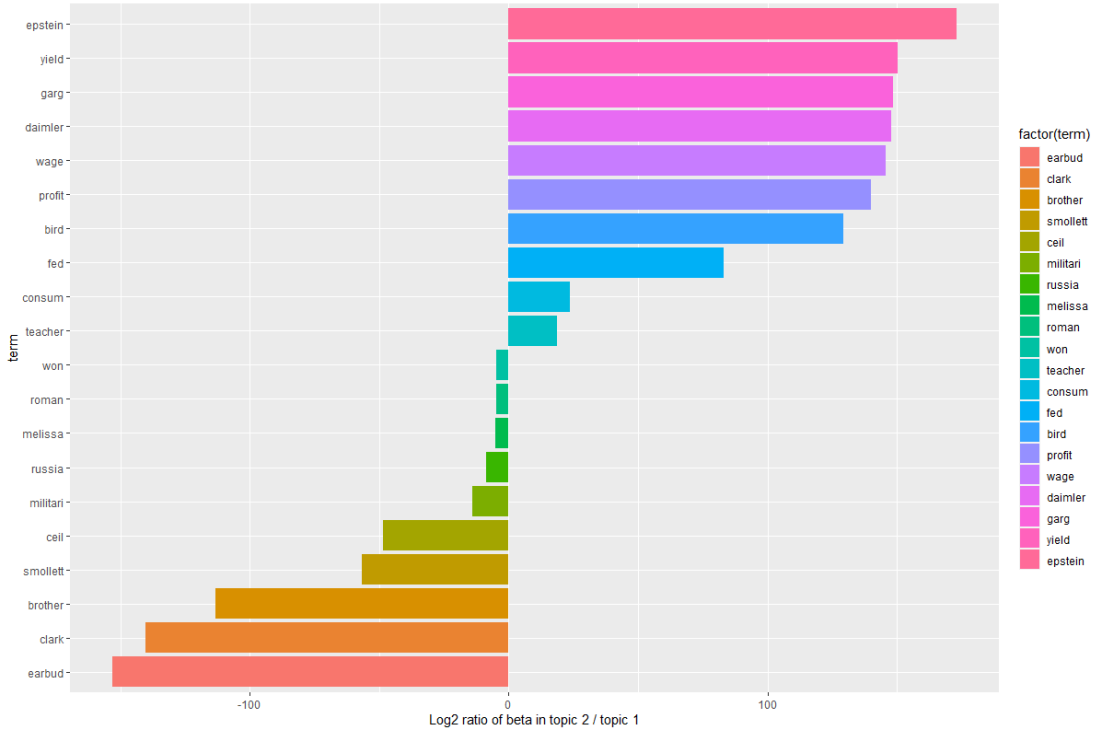


Topic Modeling:

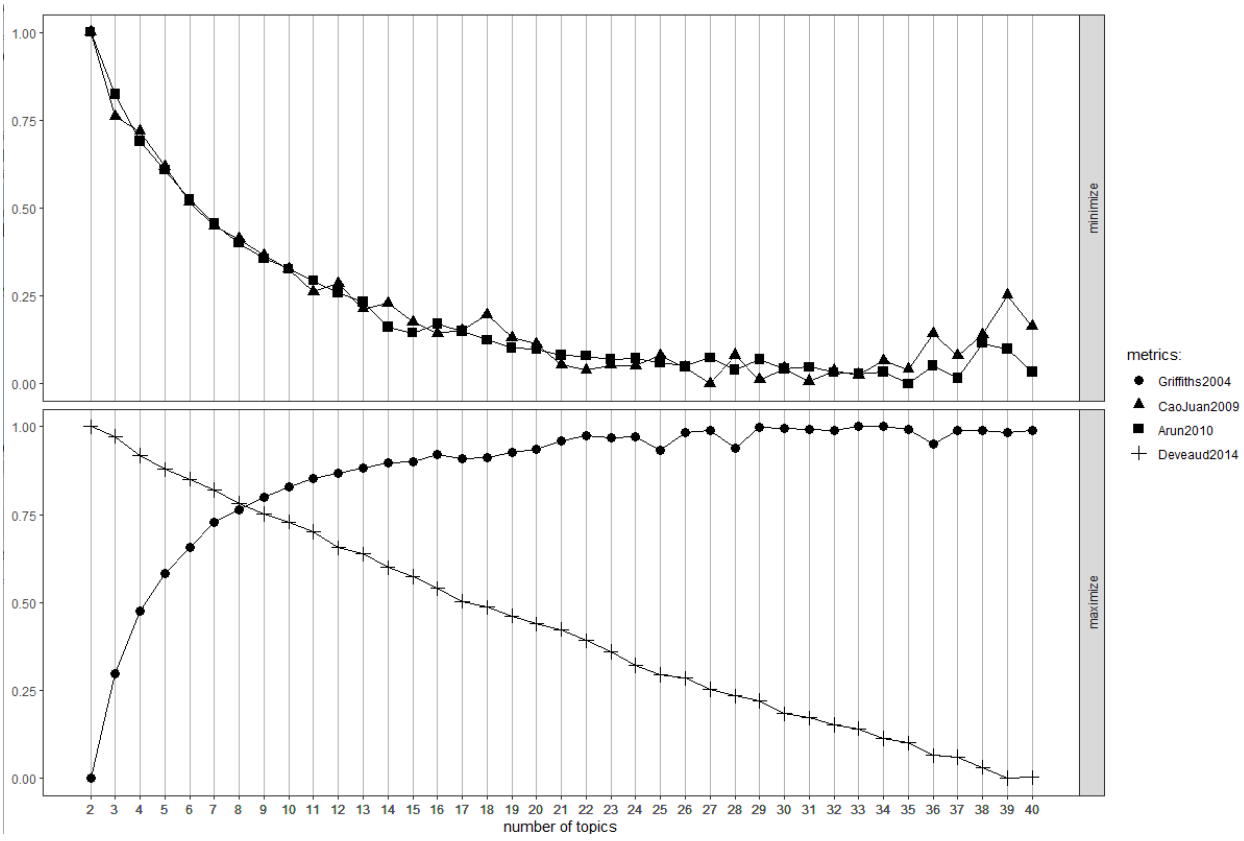
I ran Topic modeling using R topicmodels LDA(Latent Dirichlet Allocation) Tuning and Top 2 topics and their top 10 terms



Words with greatest beta spread between 2 topics shows below analysis

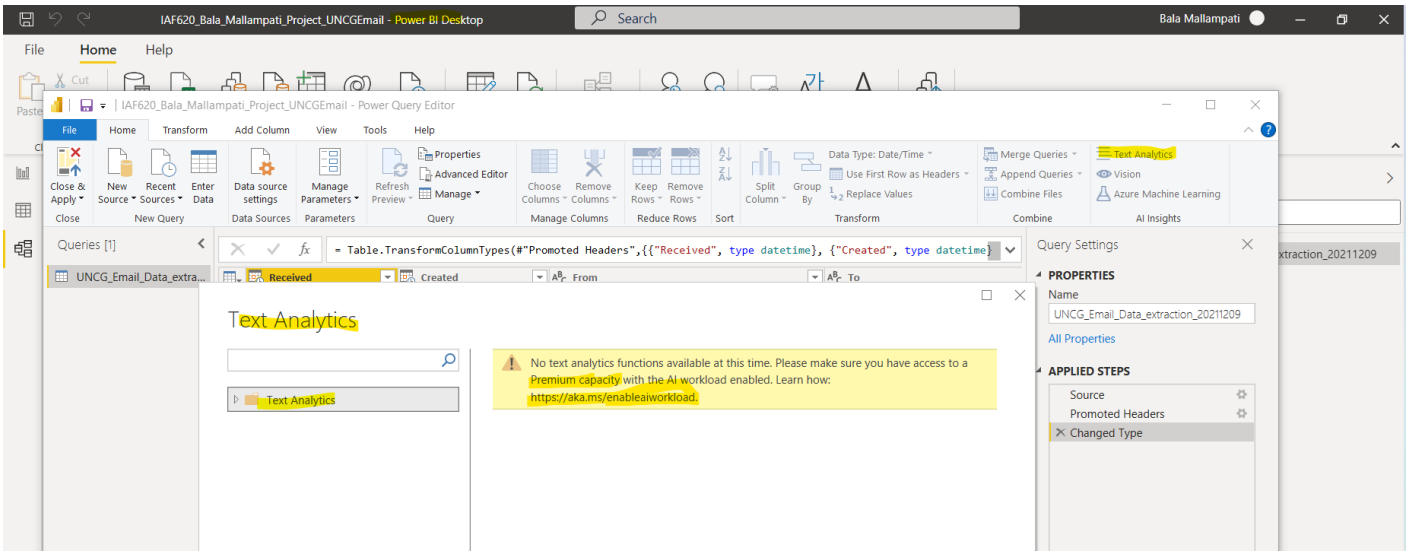


Number of Topics in NYC Times using different models:

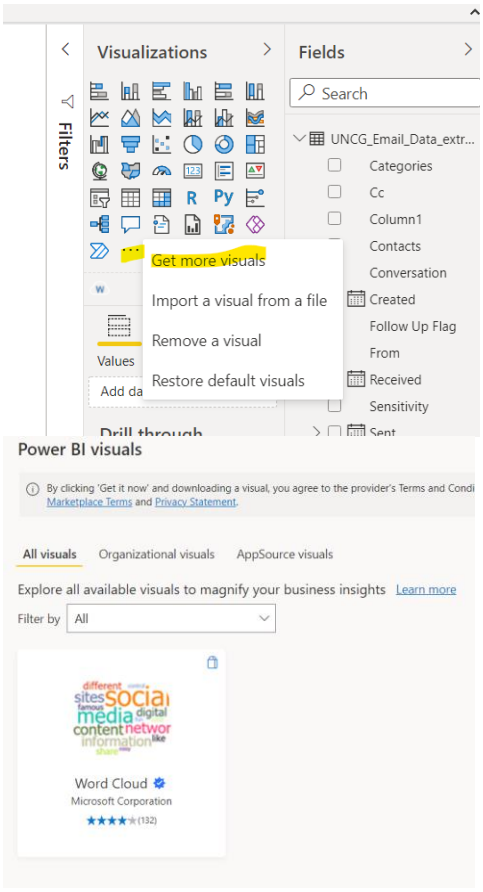


Power BI tool integration with Text Analytics:

Power BI needs Premium license to do Text analytics and Azure Machine learning. I have only Pro license.



So Imported WordCloud visualization provided by Microsoft Into PowerBI and run below visualizations.

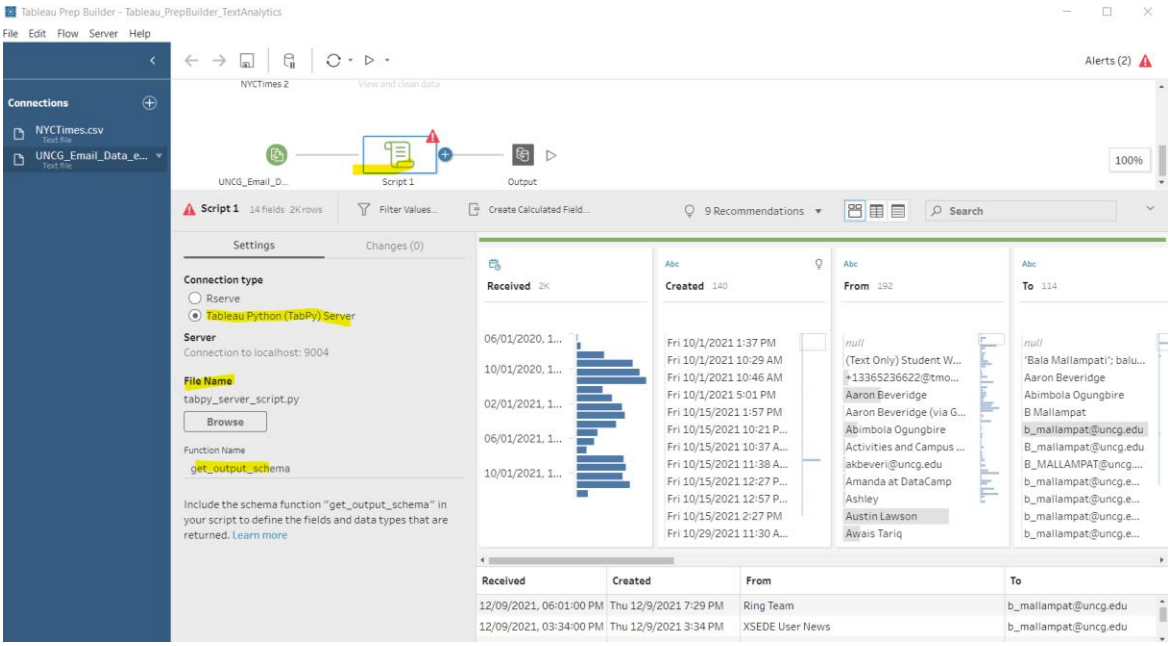


UNCG Email Subject WordCloud Using Power BI:

NYC Times World Cloud using Power BI:

***Tableau with Text Analytics using Python*:** We can do Text Analytics using Tableau Prep Builder and then feed output file to desktop

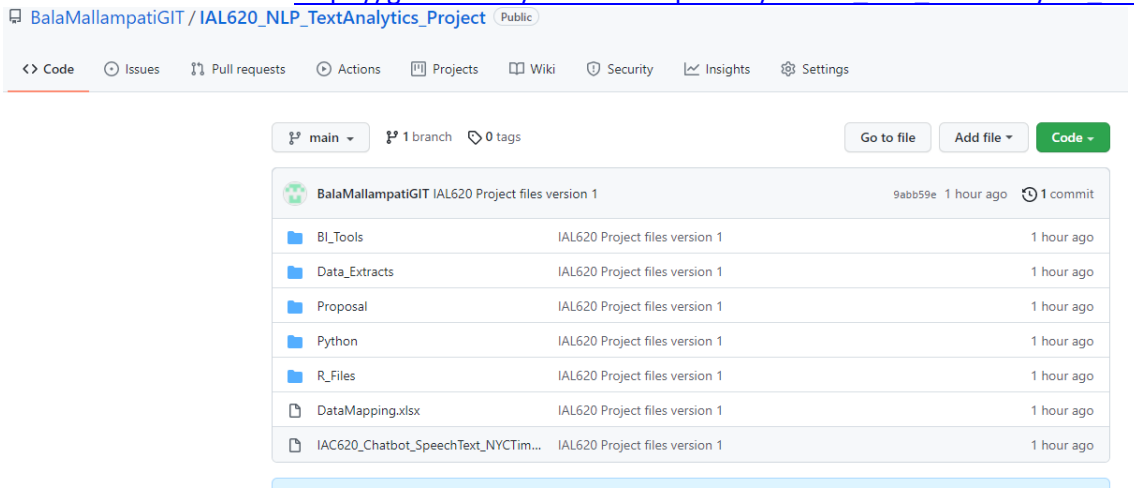
https://towardsdatascience.com/a-guide-to-integrating-text-analytics-into-tableau-94d1a1331117



From Python script we can run Sentiment Analysis and identify named entities and assign back to dataframe to export data. We can refer project files BI_Tools/tabpy_server_script.py script

```
# Importing the requisite libraries for this function
from textblob import TextBlob
import en_core_web_sm
nlp = en_core_web_sm.load()
import pandas as pd
```

Github URL for code: https://github.com/BalaMallampatiGIT/IAL620_NLP_TextAnalytics_Project



Conclusion:

I learnt new way of Text Analytics with BI tools(Power BI, Tableau) and created chatbot & Speech Recognition using Python.

Though my tools & languages ultimately did not result in an effective way to do text analytics and Sentiment analysis of UNCG Email & NYCTimes data, the improvements I am doing is alternative solutions or proved packages available in market. This is the reason started integration of Text analytics & Sentiment Analysis with Power BI & Tableau. Given more time and resources to do further research and feature engineering, I believe I could retrain this analysis to better success.

I believe that I did not have enough of the contributing tools/standard words/corpus that drive Text analysis of UNCG Email/NYCTimes news data. Additional data points, tools, Azure machine learning techniques may have improved these model results.

Challenges faced:

Time constraint and needs to do R&D on each tool. Don't have some of tools licenses(Ex:power BI Premium).

Since I choose different tools & methods to do end to end implementation, it long time get initial Skelton of project.

It has lot of scope for improvements and additions to this project. I still need to extend this scope to fulfill my initial thoughts.

References:

- Sklearn, chatterbot, chatterbot_corpus, Worldcloud, TextBlob Python library
- https://www.upgrad.com/blog/how-to-make-chatbot-in-python/
- https://towardsdatascience.com/3-super-simple-projects-to-learn-natural-language-processing-using-python-8ef74c757cd9
- https://www.nytimes.com/
- https://docs.gspread.org/en/latest/oauth2.html
- https://towardsdatascience.com/a-guide-to-integrating-text-analytics-into-tableau-94d1a1331117
- https://www.red-gate.com/simple-talk/databases/sql-server/bi-sql-server/text-mining-and-sentiment-analysis-power-bi-visualizations/