DTSC 3472 Term Project

Authors:
Harshada Phadol - 1272847
Dilip Verma - 1273084

Instructor:
Dr. Houwei Cao

Sentiment Analysis on New York Times Comments Data

Abstract:

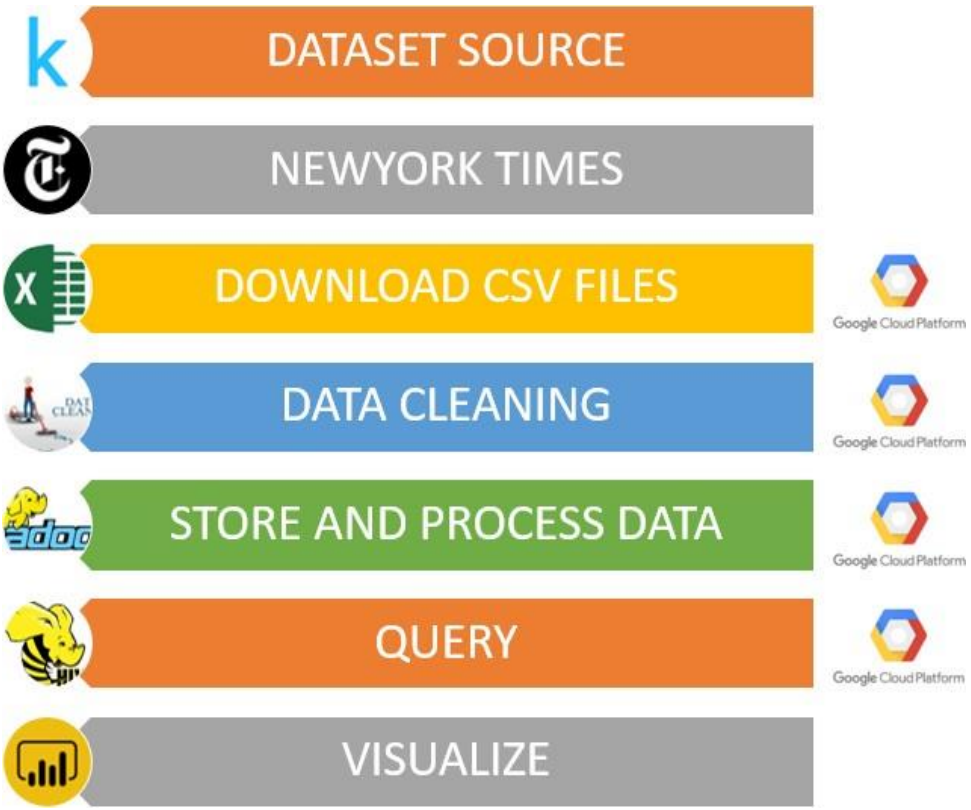
In this project, we will analyze "New York Times Comments,", a data set having information about the comments made on the articles published in New York Times. It demonstrates the usage of Hadoop, MapReduce, and Hive on big data. The dataset files for comments comprise of over 2 million (Approx.) comments in total with 34 features. This data will help the purpose of understanding and analyzing the public reading interests, analyzing behaviors.

Introduction

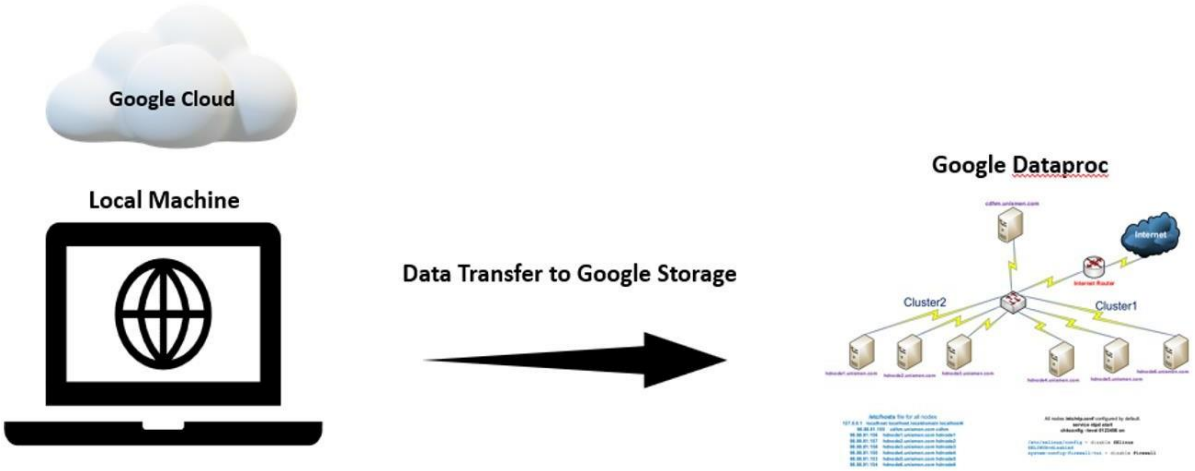
New York Times has extensive audience and plays important role in shaping people's opinion about current affairs, especially in United States of America. The comments sections for articles in the NYT are quite active and give insights to readers' opinions on the subject matter of the articles. Each comment can receive other readers' recommendations in the form of upvotes. Our First aim is to classify a given piece of material in NYT as an 'article' or 'blogpost' and then further categorize in that which is most read material in articles and then to find the article on which the commenters such as people/NYT users/editors most likely comment. Second step is to analyze the public response over these articles and blogposts by seeing the number of reply count and to determine how many of these receive the most recommendations and thirdly to perform the sentiment analysis of these comments and lastly to analyze which is the best author for that particular year which furthermore gives us insight on author likes to write which type of articles

more. Our target would be public, editors, authors, users and the commentators here. Performing the steps above would help us to analyze the trending topics of people's interest and the ones which receive a lot of response and recommendations based on the reply count and most active area from USA which replies to these articles and blogposts.

Flowchart:



Architecture:



PLATFORM SPECIFICATIONS:

Hive on Google Cloud Platform

- ➢ Cluster version: Hadoop 2.6.5
- ➢ Cluster number of nodes: 1 Master, 3 worker nodes
- ➢ Memory size (CPU): 8 cores CPU, 30 GB Memory
- ➢ HDFS Capacity: 200 GB
- ➢ Storage: 500 GB
- ➢ Hive Version: 2.6.5
- ➢ YARN Cores: 12
- ➢ YARN Memory: 36 GB

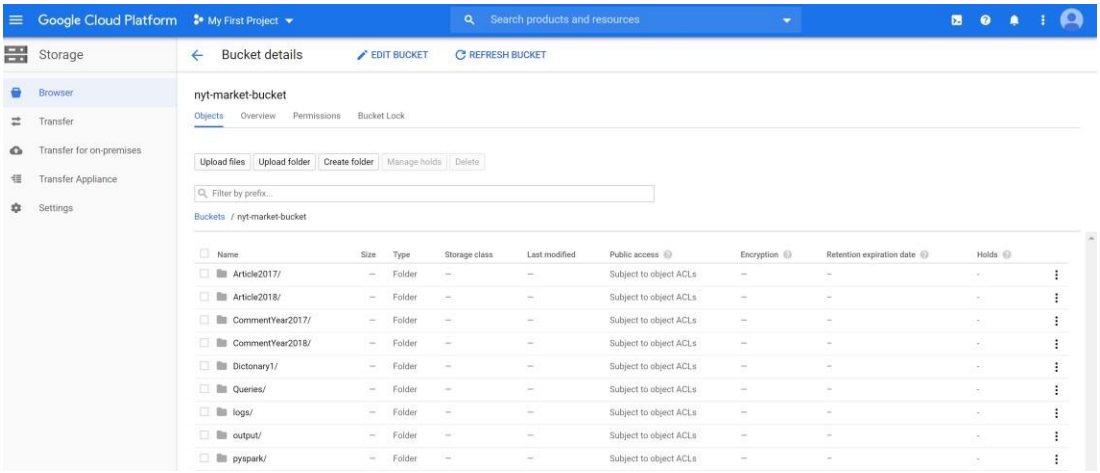
PREREQUISITES:

- ➢ You must have Microsoft Excel 2010, 2013 or 2016 installed.
- ➢ You must have your Excel 3D-Map enabled.
- ➢ Tableau 2019.4 installed for visualization of the analyzed data.
- ➢ Power BI Desktop Version
- ➢ You must have your Excel 3D-Map enabled.
- ➢ Tableau 2019.4 installed for visualization of the analyzed data.
- ➢ Enable following APIs on Google Cloud Platform:
    - Google Cloud Dataproc API
    - Create cluster with 1 master and 3 worker nodes
    - Google Cloud Storage and create bucket

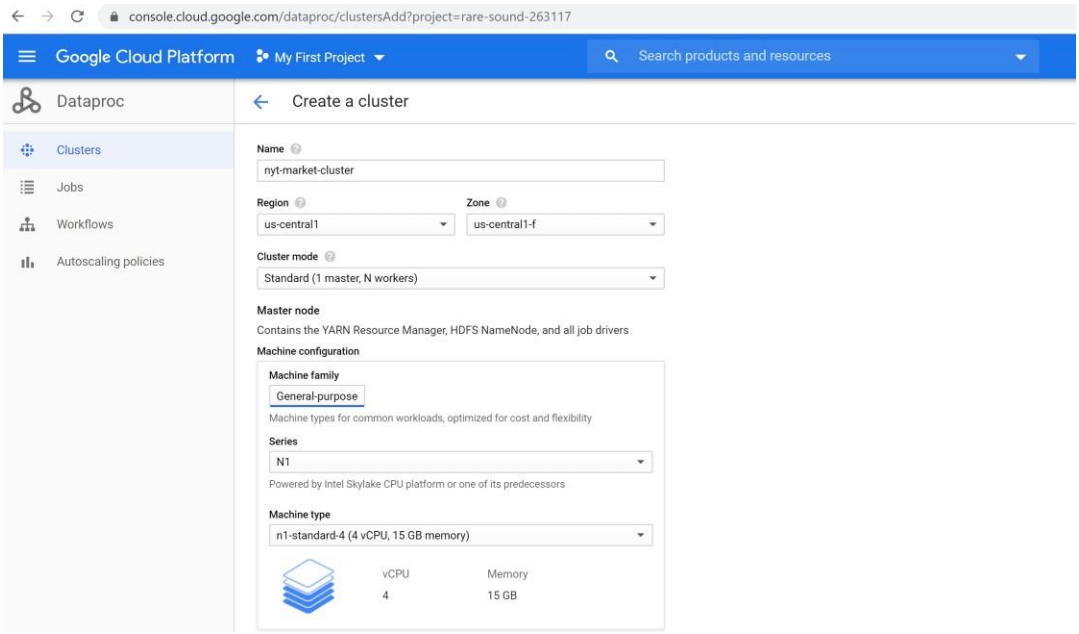
Data Description:

The data contains information about the comments made on the articles published in New York Times in Jan-May 2017 and Jan-April 2018. The month-wise data is given in two csv files - one each for the articles on which comments were made and for the comments themselves.
The csv files for comments contain over *2 million comments* in total with *34 features* and those for articles contain *16 features* about more than *9,000 articles*.

Cloud Storage - To store the data in the bucket



Dataproc – To create the cluster with 1 master and 3 worker nodes.



Output files:

```
0: jdbc:hive2://localhost:10000/default> Closing: 0: jdbc:hive2://localhost:10000/default
dverma01@nyt-market-cluster-m:~$ gsutil ls gs://nyt-market-bucket/
gs://nyt-market-bucket/Article2017/
gs://nyt-market-bucket/Article2018/
gs://nyt-market-bucket/CommentYear2017/
gs://nyt-market-bucket/CommentYear2018/
gs://nyt-market-bucket/Dictonary1/
gs://nyt-market-bucket/logs/
gs://nyt-market-bucket/output/
dverma01@nyt-market-cluster-m:~$ gsutil ls gs://nyt-market-bucket/output/
gs://nyt-market-bucket/output/
gs://nyt-market-bucket/output/query1-2017-article/
gs://nyt-market-bucket/output/query1-2017-blogpost/
gs://nyt-market-bucket/output/query1-2017/
gs://nyt-market-bucket/output/query1-2018-docType/
gs://nyt-market-bucket/output/query1-2018/
gs://nyt-market-bucket/output/query10-2018/
gs://nyt-market-bucket/output/query2-2017-art-commt/
gs://nyt-market-bucket/output/query2-2017-article/
gs://nyt-market-bucket/output/query2-2018-art-commt/
gs://nyt-market-bucket/output/query2-2018-article/
gs://nyt-market-bucket/output/query3-2017-comment/
gs://nyt-market-bucket/output/query3-2018-comment/
gs://nyt-market-bucket/output/query4-2017-article/
gs://nyt-market-bucket/output/query4-2017-blogpost/
gs://nyt-market-bucket/output/query4-2018-article/
gs://nyt-market-bucket/output/query4-2018-blogpost/
gs://nyt-market-bucket/output/query5-2017-recomnd/
gs://nyt-market-bucket/output/query5-2017/
gs://nyt-market-bucket/output/query5-2018-recomnd/
gs://nyt-market-bucket/output/query5-2018/
gs://nyt-market-bucket/output/query6-2017/
gs://nyt-market-bucket/output/query7-2017-negative-senti/
gs://nyt-market-bucket/output/query9-2017/
dverma01@nyt-market-cluster-m:~$
```

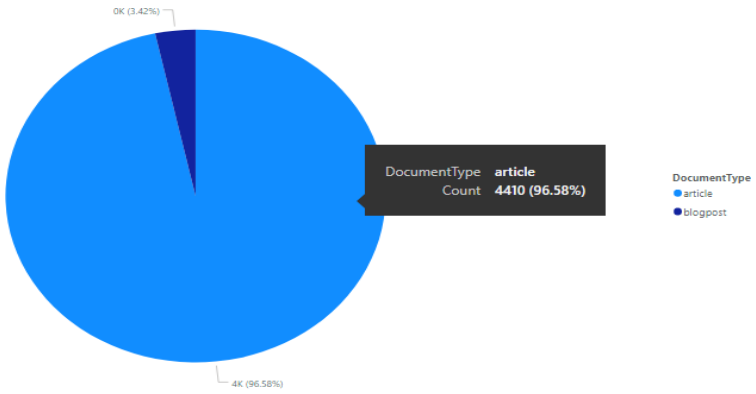
DATA VISUALIZATION

Visualizing Data: (In order to visualize the data, we have used power BI as well as Excel 3D Maps)

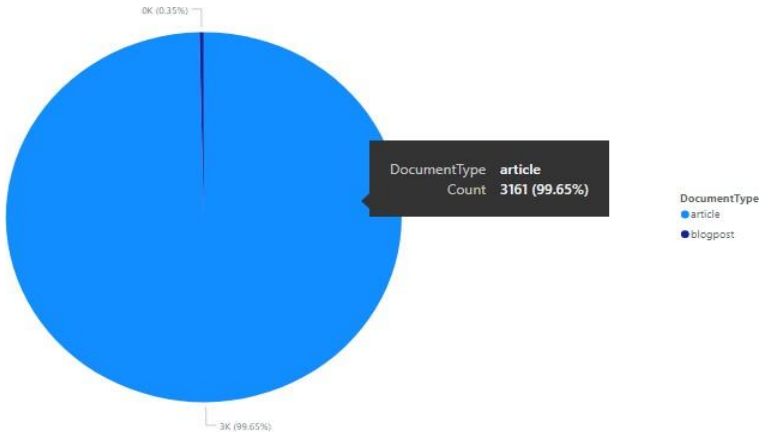
Type of Document that were Highly Published and in which month of the year (2017,2018) did the type of Documents Receive Highest Number of Replies.

New York Times majorly published 2 types of documents which area Articles and Blogs. We wrote queries on hive to analyze which type of documents were published more in 2017 and 2018 and acquired the result that Articles were highly published in both years compared to blogs. Also, we categorized articles furthermore into type of material, which was News, Op-Ed, Review, Editorial, Briefing, and Letter. We concluded that most read was news and least read was letter in 2017 and Interview in 2018.

2017

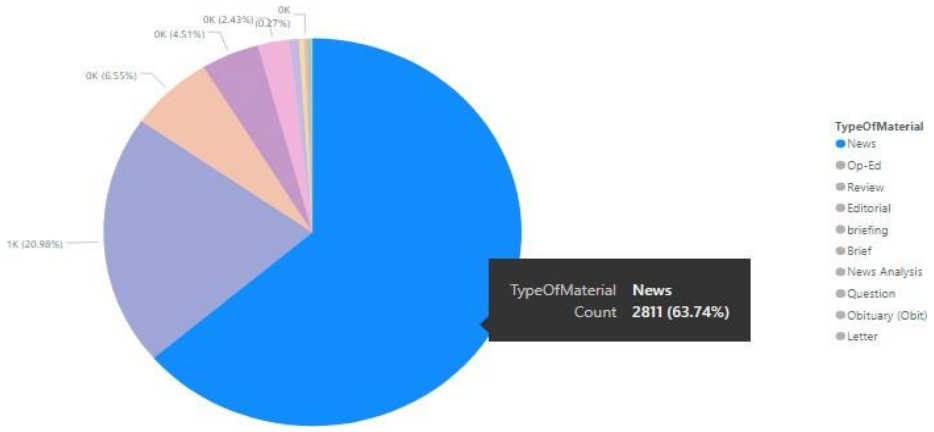


2018

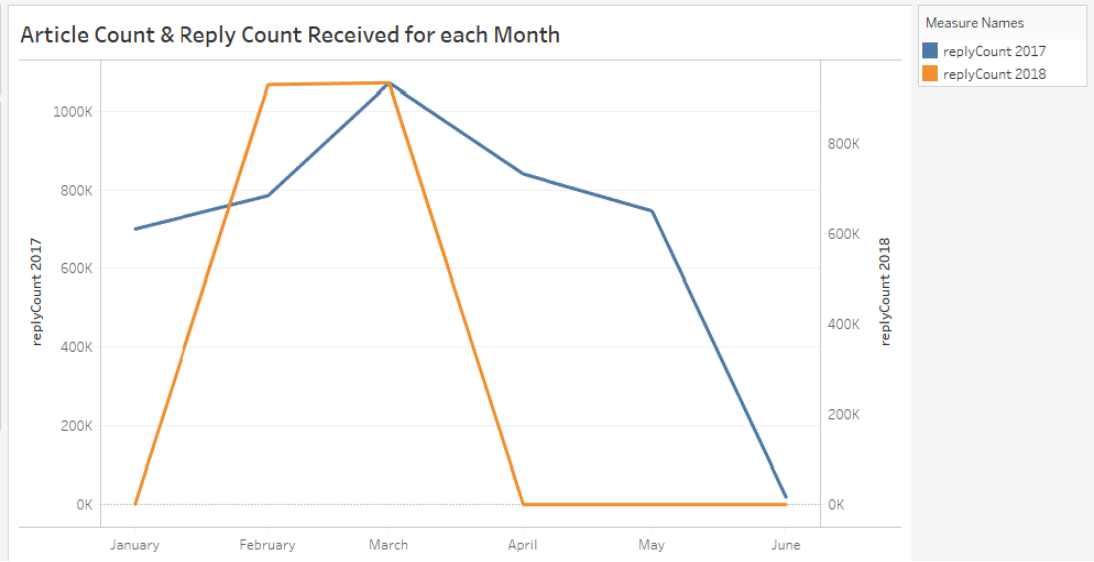


Material Visualization:

Count of Type of Material by Article 2017



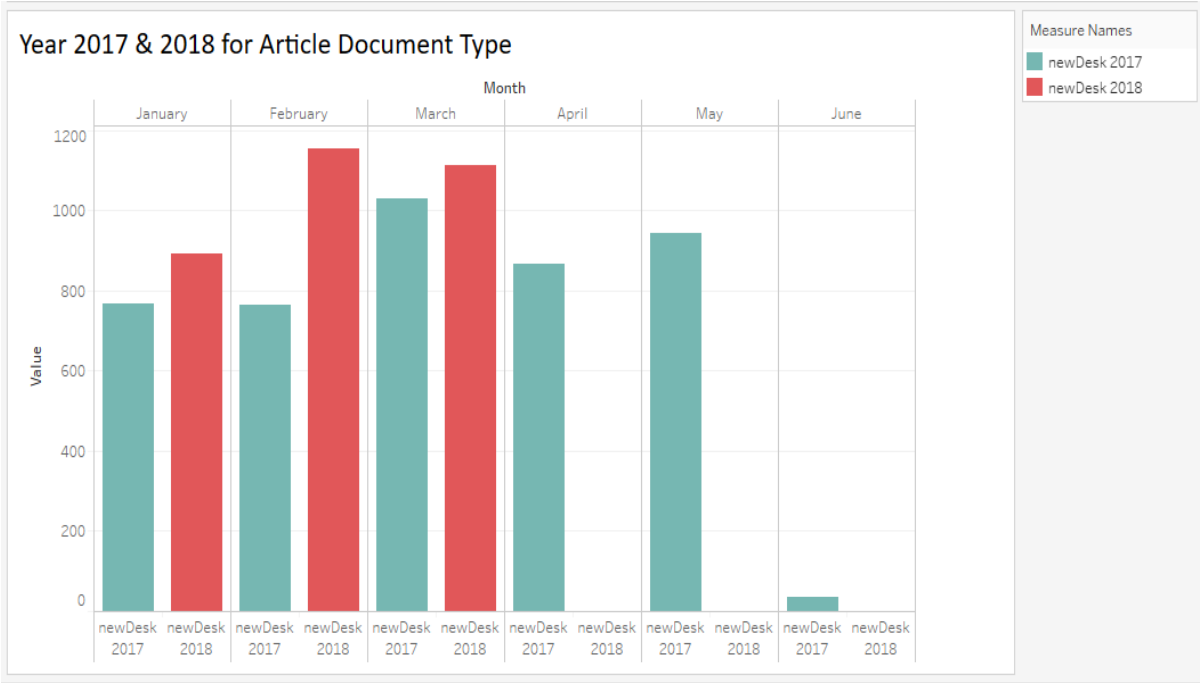
Furthermore, we found out with the help of hive queries, same month of both years i.e. March 2017 and March 2018 received the highest number of count and reader replies. For document type articles and for the blogpost document type March was highest in 2017 and in 2018 it was only January after which there was no response for consecutive 4 months.



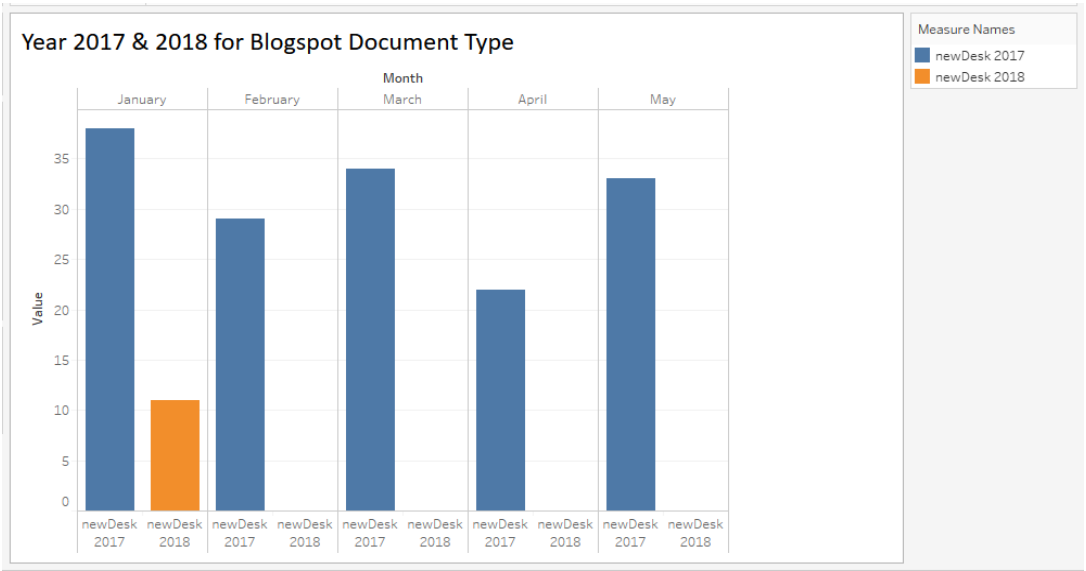
What is the count of new desk month wise.

For the year article: 2017 & 2018 year

March was the month where New Desk for both the years were highly published for the article document type and January was the month where New Desk for both the years were highly published for the Blogpost document type.



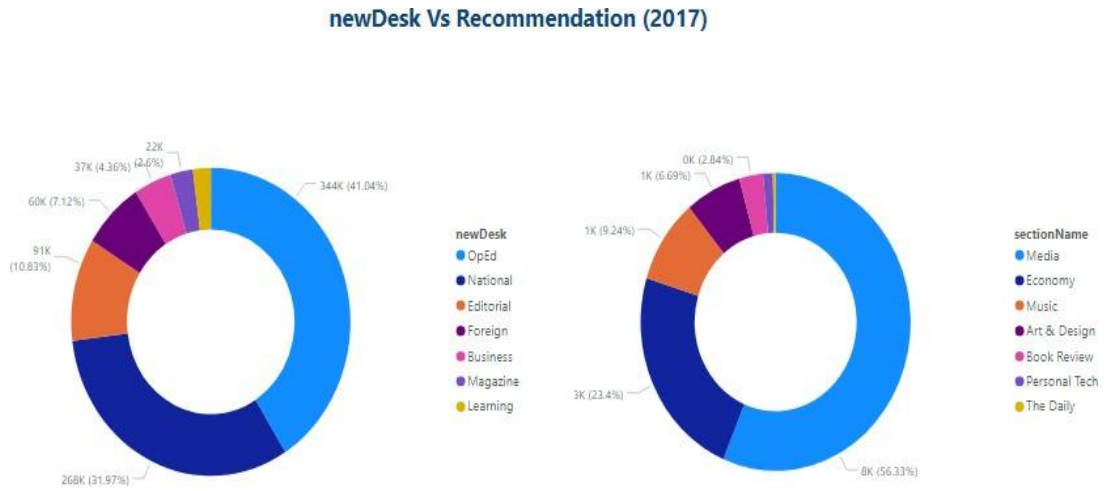
For the Blog post year 2017 & 2018:



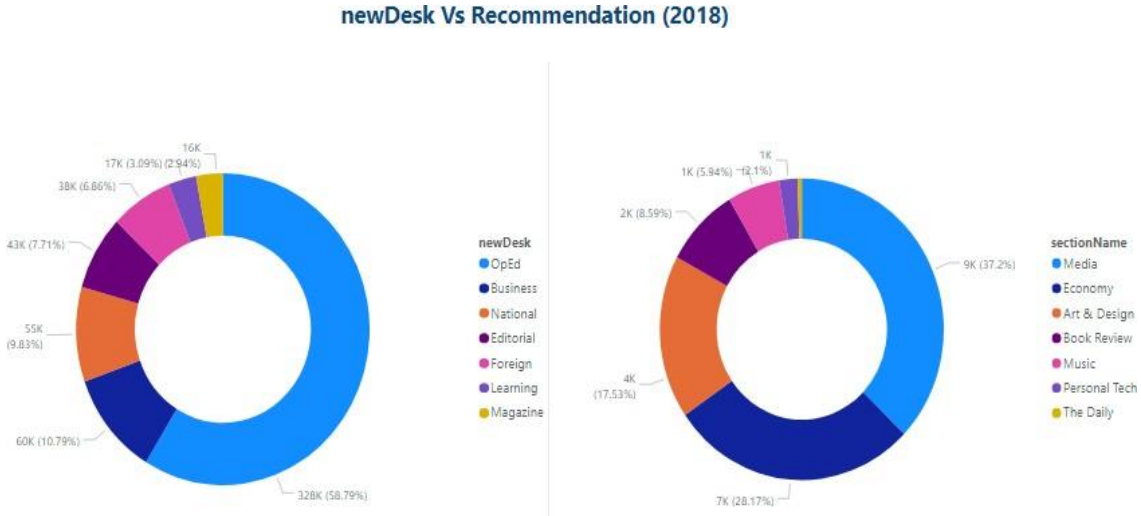
Highest number of New Desk for the document type: Articles/Blogposts published for year (2017, 2018) and Highly recommended New Desk as per Readers Interests.

Articles are categorized based on the interests and likings of the readers where the column named New desk involves such categories. Some of the categories under New Desk are Art & Leisure, Business, Dining, Culture, OpEd (Editorial Opinion), Learning, National, Foreign. We wrote queries to know what type of New Desk the readers highly recommended for both the years and in which month of both the years was New Desk highly published for both the document types.
We observed that OpEd which is the opinions provided by the Editorial of the Newspaper were highly recommended by people for articles and Politics for blogposts.

For the year 2017



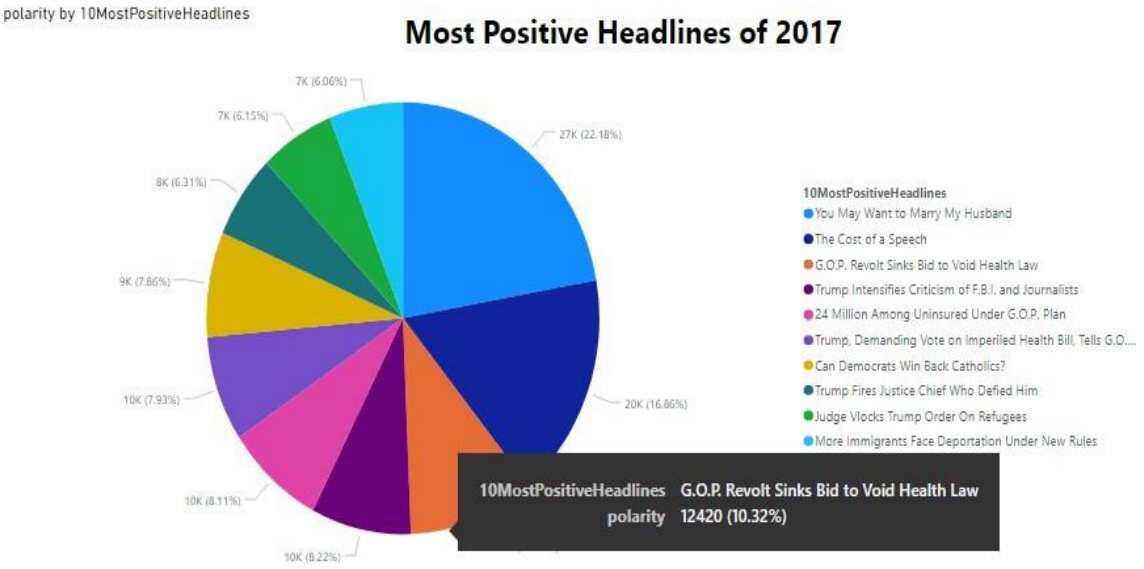
For the year 2018



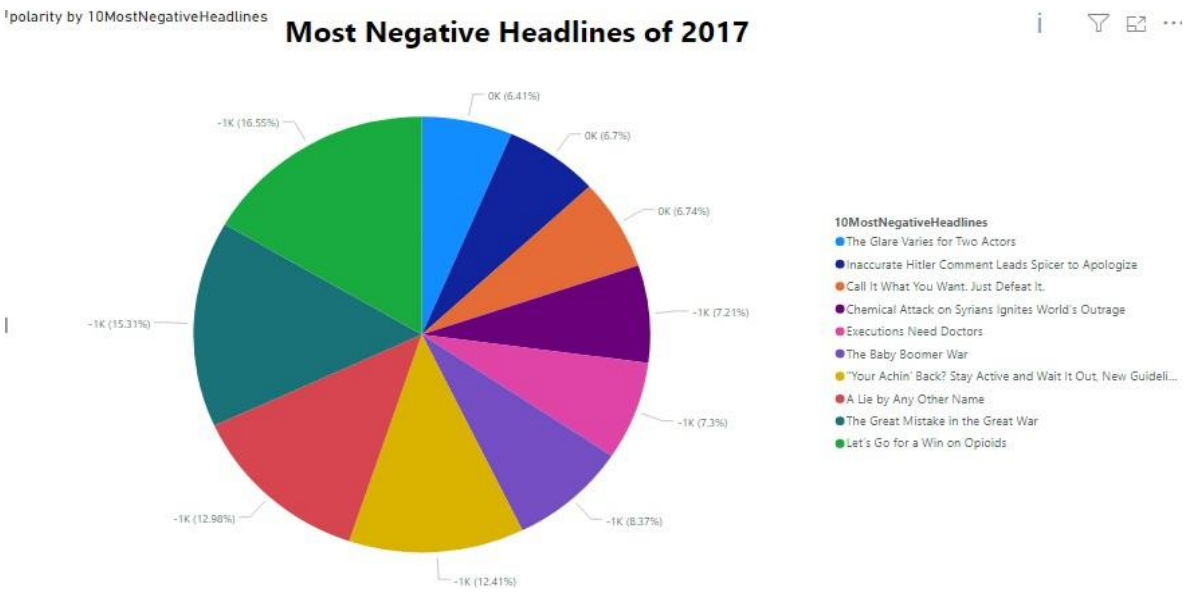
What is the degree of polarity by most positive headlines

We performed the sentiment analysis by evacuating appropriate degree of polarity on the datasets Article Year 2017, Comment Year 2017 and derived the top headlines which received most positive as well as most negative reactions from readers by analyzing their comments.

Polarity calculated based on positive and negative words in reader's comments.



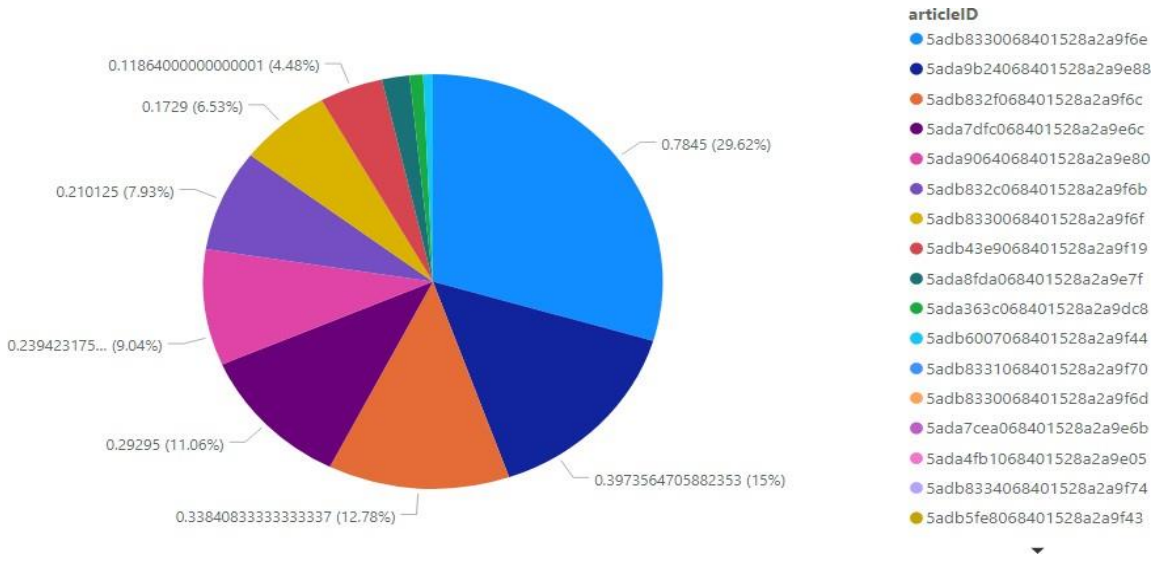
1. What is the degree of polarity by most negative headlines?



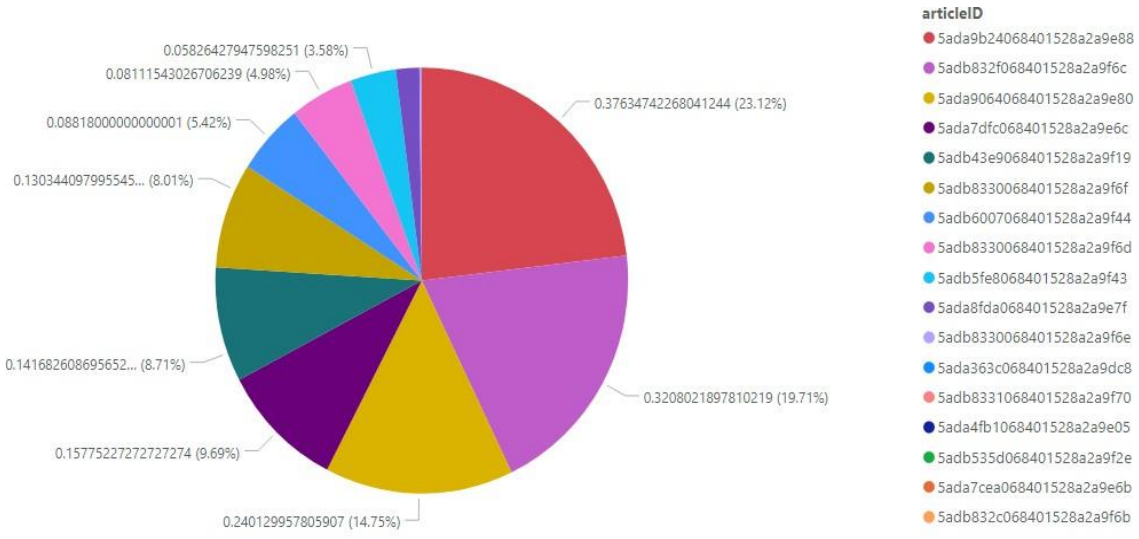
Degree of polarity for each article by comments :

Polarity calculated based on the positive and negative comments on each article.

2017



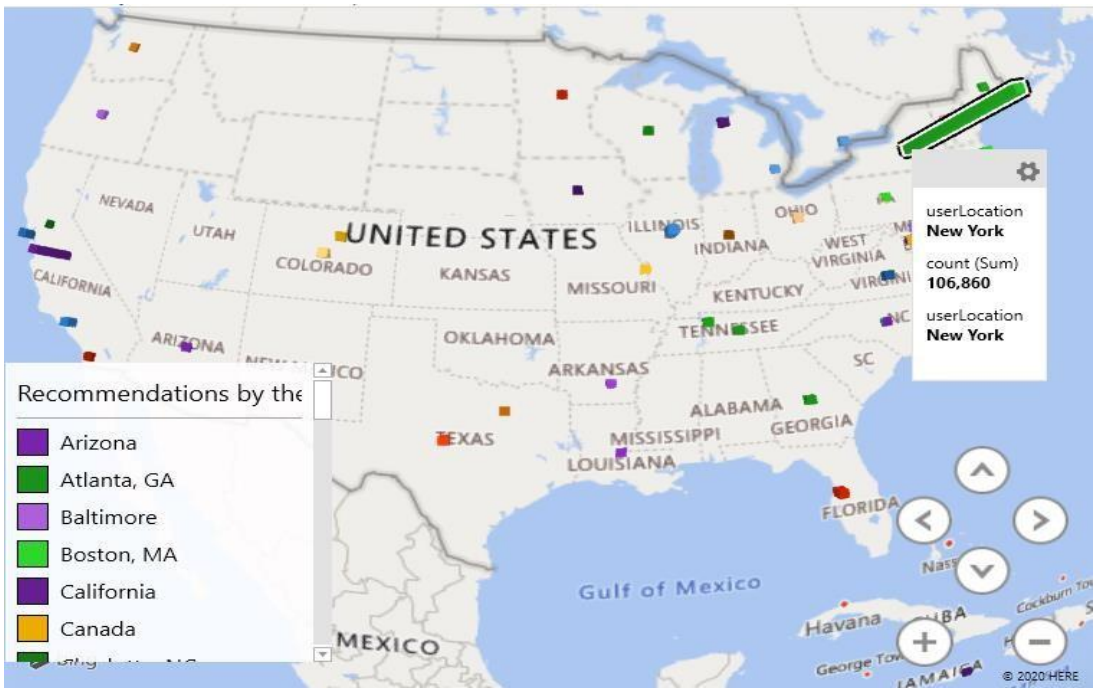
2018



Reply Count by User Locations for Year (2017,2018).

We analyzed to know from which state of United State users were most active and providing the greatest number of replies. For the year 2017 we found the below results.

- NY: 30,638
- Cal: 15,273
- Chicago: 12271



Most Popular Author with Respect to Reader's Recommendations for the Year 2017 and 2018.

We were intrigued to analyze which authors were highly recommended by the readers in the year 2017 and 2018. We evaluated the below results using the hive queries.

- Highly Recommended author 2017: KATHERINE SCHULTEN
- Highly Recommended author 2018: STEVEN BUSER

2017

2018

Challenges:

- Configurations in Google Cloud.
- Initially took much time to download the data to virtual machine and implement the further process.
- RAM issues with Virtual machine.

Contribution:

Both the team members contributed in this project equally. Since we implemented this project in two ways (HDFS and Google Cloud), we divided the work accordingly. Also, in data visualization, we used multiple platforms like PowerBI, Tableau and Excel 3D Maps. We had about 10 data visualizations, so we distributed partly so both of us can contribute and learn multiple platforms.

Future Scope:

- On Google Cloud platform, huge dataset (of any size) can be stored in google storage and practice machine learning concepts directly.
- Since analysis is done on HIVE and data is stored on GCP, implement the same with machine learning concepts on the same platform. This would be beneficial in terms of working in real world problems.

Conclusion:

While exploring the New York Times Dataset, we successfully used Hadoop, Google cloud, HiveQL, 3D -Maps, PowerBI and Tableau to store and manipulate the data in order to gain the maximum insights from it. We analyzed the dataset, being provided with the data for 2 years, that is, 2017 and 2018, right from January to June for 2017 and January to May for 2018. We were able to draw conclusions by analyzing the sentiments of people as positive or negative. Also, we found out that the type of materials being published in NYT were more of the 'article' type than the 'blogpost' types and in that most used type of material for articles was News. We also investigated the topmost areas of people's interest on which they most likely comment as well as determined those topics for each document type: article/Blogpost that received the highest recommendations from the public as well as the users and editors of NYT. Moderators can focus on these categories when moderating comments added by readers. We even interpreted month wise that the articles received much more replies(responses) in the month of March as compared to other months with a significant decrease of replies for the month of May, for year 2017 and for 2018 and we also observed that the greatest number of replies for Blogpost was in March for year 2017 but for 2018 it was only in January and no one commented after that. While querying the data we also uncovered that 'news' was the topic that was most talked(read) about in NYT and by investigating the reply

counts for user more we found that users from New York are more active followed by California and Chicago. Also, we found out the best author for both the years from the articles and comments data.

References:

[1] Set up hadoop on google cloud dataproc
http://holowczak.com/getting-started-with-hive-on-google-cloud-dataproc/

[2] https://medium.com/@cuongdo.uconn/how-to-hive-on-google-cloud-platform-dataproc-and-storage-d141536644cd

[3] Upload multiple files to google cloud storage
https://www.youtube.com/watch?v=ji1DWCTI05A

[4] Hive data to CSV file
https://community.cloudera.com/t5/Support-Questions/how-to-download-hive-data-into-csv-format/td-p/59591

[5] https://xebia.com/blog/sentiment-analysis-using-apache-hive/

[6] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7122945/

[7] https://dzone.com/articles/data-analysis-using-apache-hive-and-apache-pig

[8] https://fdocuments.in/document/basic-sentiment-analysis-using-hive.html