New York Times Data Analysis Using Hive

Divya Pakhale,Sanchita Gawand,Siddhi Udani,Tanvi Gawade
Department of Information Systems, California State University
Los Angeles
Tel. 626-438-0899, Fax. 323-343--5209
e-mail: dpakhal@calstatela.edu,sgawand@calstatela.edu,sudani2@calstatela.edu,tgawade@calstatela.edu

Abstract: For our analysis project, our group will analyze "New York Times Comments,", a data set containing information about the comments made on the articles published in New York Times for the period January-May 2017 and January-April 2018. It illustrates the usage of Hadoop, MapReduce, and Hive on big data for easy summarization by utilizing the knowledge gained in lab sessions by querying big data, hands-on practice, extensive researches and development of HiveQL in order to generate and query about 1.5 GB of data and visualize it on Tableau, Power BI, Excel 3D Maps. The dataset 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 which include comments on articles, number of readers, headline categorization, document type, section name and many more. This data will serve the purpose of understanding and analyzing the public mood, analyzing behaviors of the top commentators such as which topics they most likely comment and the sentiment analysis of the comments. Other Elements of this project include a report paper, a tutorial on the queries, and one group presentation.

URL: https://www.kaggle.com/aashita/nyt-comments
Dataset size: 1.5GB
Cluster version: Oracle Big Data Compute Edition
No of nodes: 5 (management and data nodes)
HDFS Capacity: 147GB
CPU Speed: 2.20GHz
Storage: 678GB

1. Introduction

Newspapers have been a part of people's life for decades and even after the digital revolution, a large amount of masses read and gain awareness about the daily happenings from newspapers. New York Times has wide 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 'article' or 'blogpost' 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 to determine how many of these receive the most recommendations and thirdly to perform the sentiment analysis of these comments. Our target would be public, editors, 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.

2. Manipulating Datasets

2.1 Tools and data processing

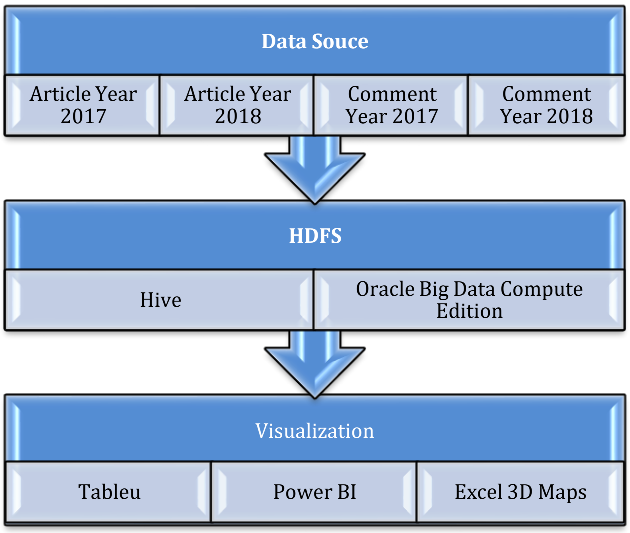


Figure 1. Data Processing

- In order to determine the readers emotional reaction towards the headlines we analyzed and performed sentiment analysis on the dataset April Year 2017, Comment Year 2017 with respect to contextual polarity to systematically segregate the positive and negative headlines.
- Based on the above results which displayed the attitude of readers with respect to specific headlines, we further analyzed the datasets April Year 2017, Comment Year 2017, April Year 2018 and Comment Year 2018 to derive the words that were commonly used in the headlines published in the New York Times to attract the greatest number of readers.
- Fundamental commands like wget, mkdir and hive were used to initiate the project. We were able to successfully download and upload the New York Times datasets from Kaggle following which we connected with Hive and built complex codes to analyze each dataset.
- Furthermore, we also conducted analyses to derive results based on below queries.

Sr No.	Analysis Topic	Categories
1	Types of Documents that were published more with respect to the year 2017 and 2018	Articles/Blogs
2	Document Type that received highest number of replies with respect to the year 2017 and 2018	Articles/Blogs
3	Which type of comment received the highest number of counts with respect to the year 2017 and 2018	Comment/User Reply/Report Reply
4	Which month received the highest count of New Desk(Section categories like culture, dining, editorial) for the year 2017 and 2018	January, February, March, April, May
5	Which is a highly recommended New Desk as per the readers with respect to 2017 & 2018	Business, Editorial, Foreign, Learning, National, OpEd
6	Recommendations were received greatly from which state in the United States with respect to 2017 & 2018	New York, California, Chicago
7	Most Popular Author with respect to reader's recommendations for the year 2017 and 2018	Katherine Schulten - 2017 Steven Buser - 2018

Table 1

2.2 Sentiment Analysis (2017)

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

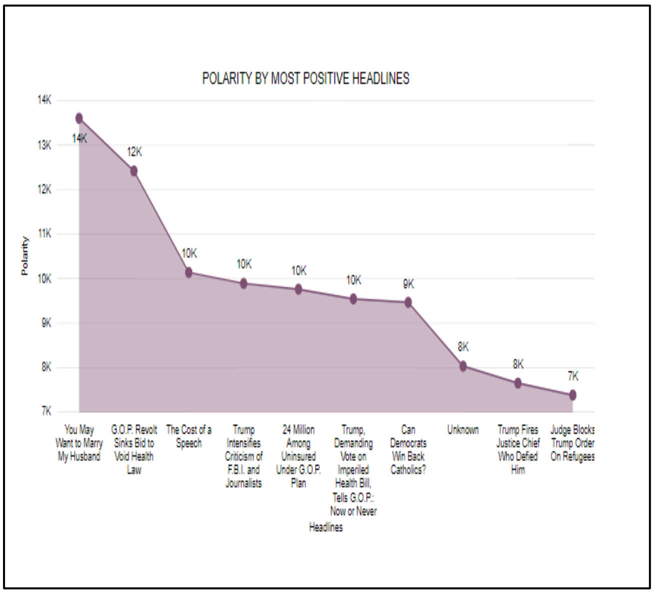


Figure 2

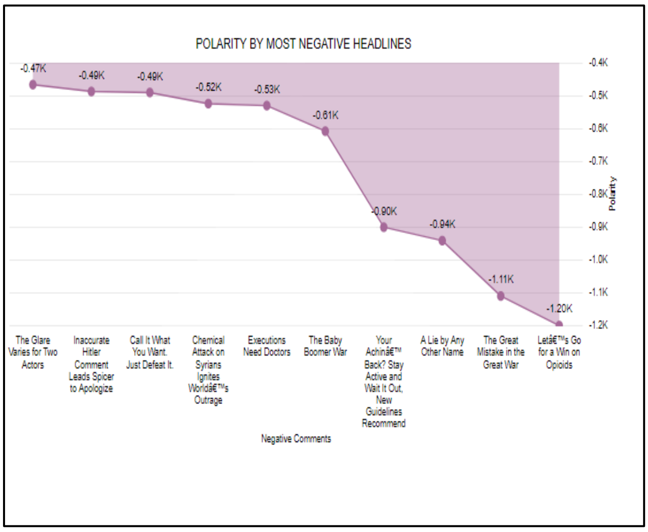


Figure 3

2.3 Word Cloud for Most Commonly Used Words in Headlines (2017 and 2018)

We were further intrigued to analyze New York Times targeted which words which indirectly pointed out to hot topics for the year 2017 and 2018, in order to attract more readers and observed that Trump was the most commonly used word in the headlines published for the year 2017 and 2018.

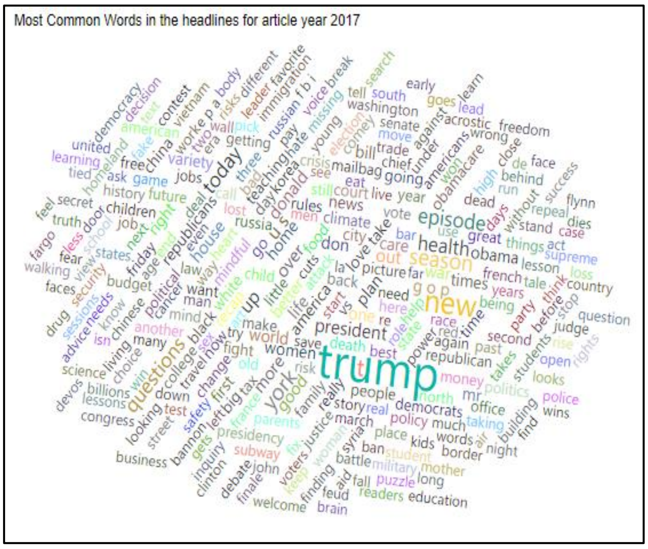


Figure 4

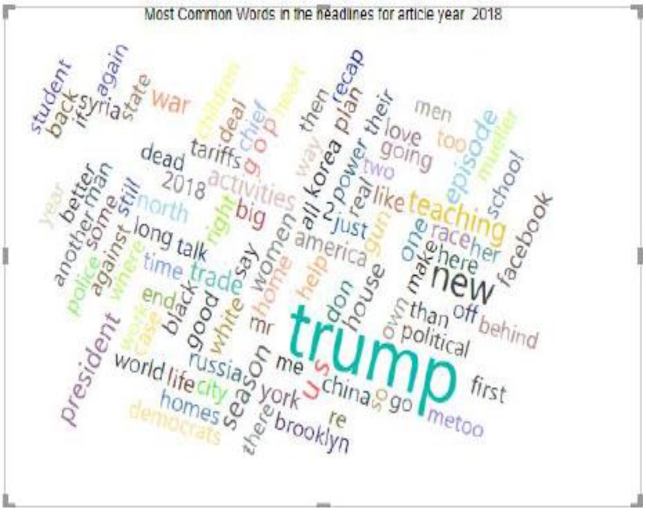


Figure 5

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

New York Times majorly published 2 types of documents that is the 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 the years compared to blogs.

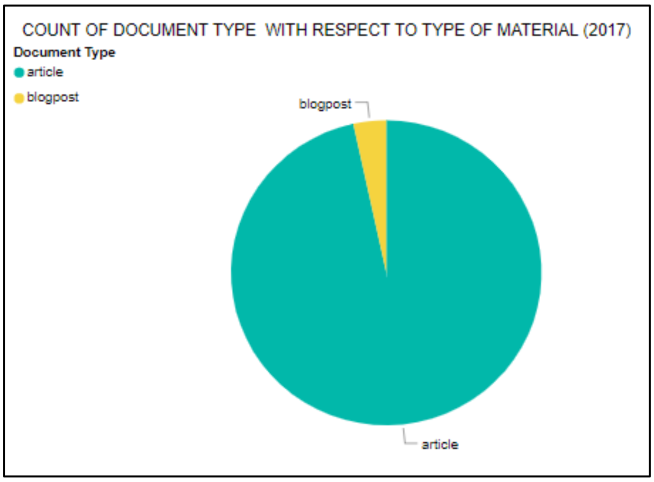


Figure 6

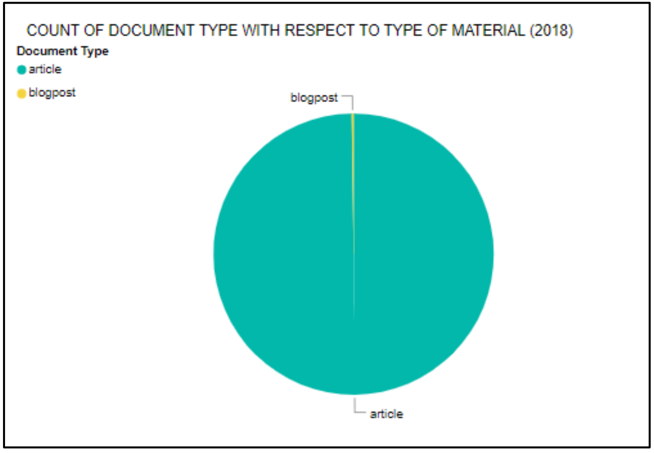


Figure 7

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 reader replies.

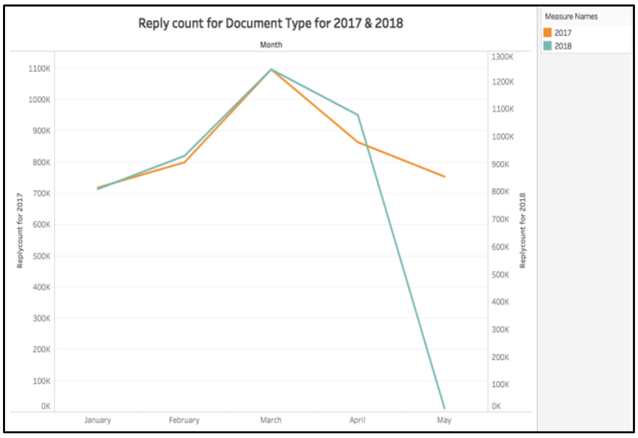


Figure 8

2.5 Highest Number of Reply Count for Comment Type (Comment, User Reply, Report Reply)

Comment types of readers are generally classified into 3 broad categories i.e. Comment, User Reply, Report Reply. Here the User Reply are associated with the registered users of New York Times and hence we were keen to know which type of comment raised the highest number of replies.

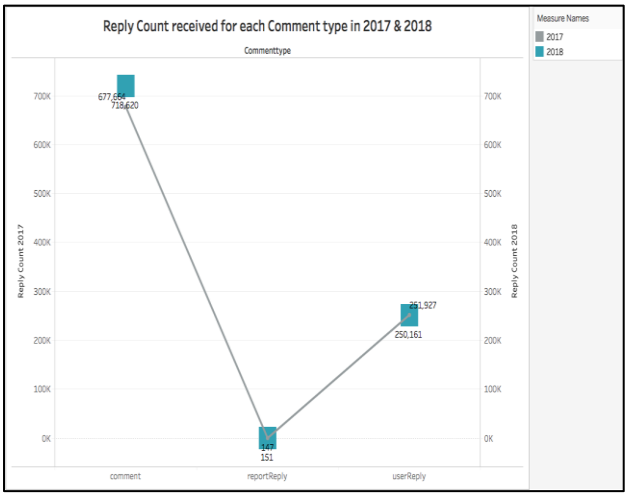


Figure 9

2.6 Highest number of New Desk 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.

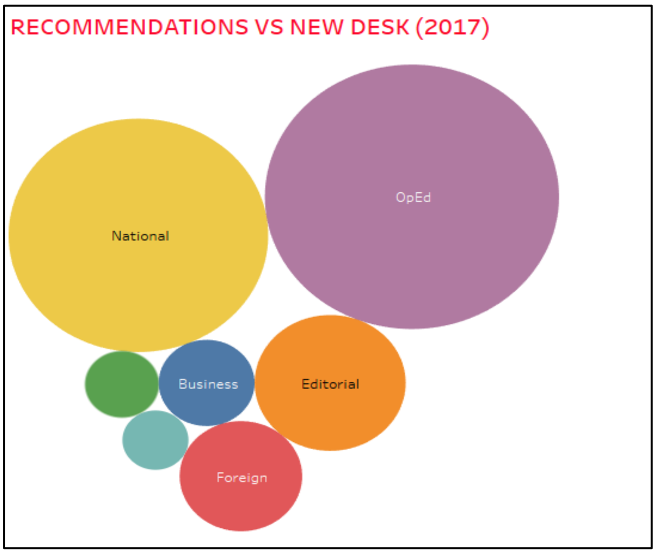


Figure 10

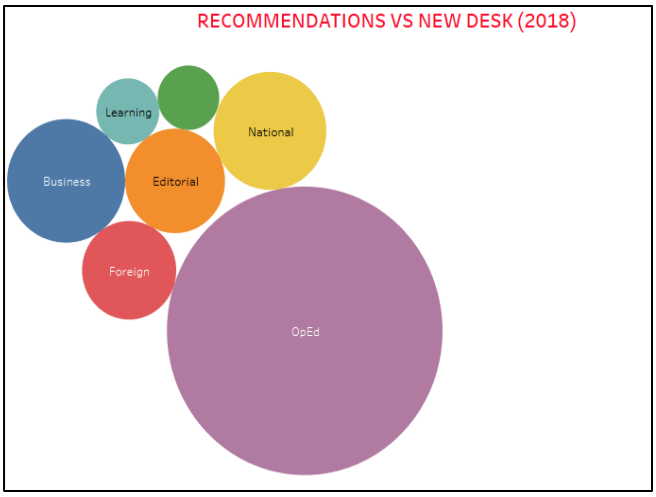


Figure 11

We observed that OpEd which is the opinions provided by the Editorial of the Newspaper were highly recommended by people.

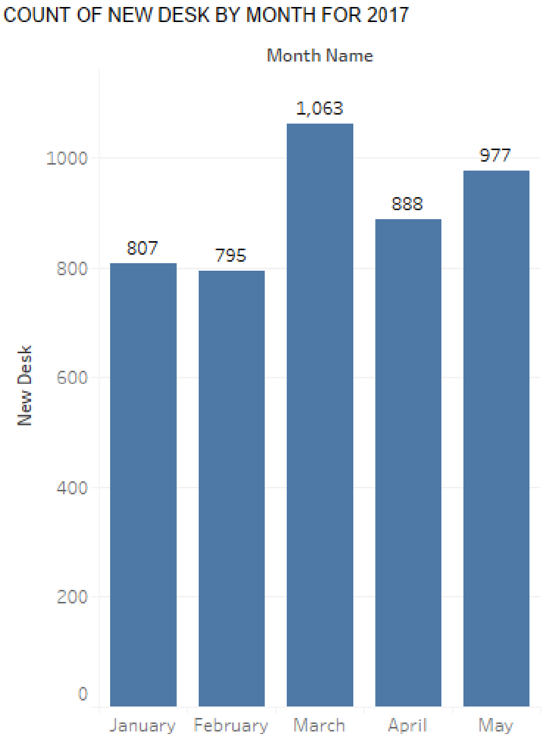


Figure 12

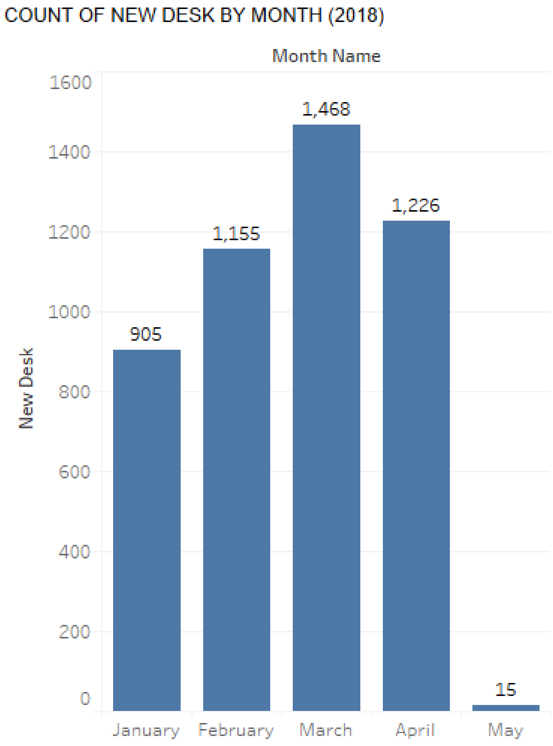


Figure 13

March was the month where New Desk for both the years were highly published.

2.6 Recommendations 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 recommendations. For the year 2017 we found the below results.

- NY: 27147
- Cal: 15273
- Chicago: 12271

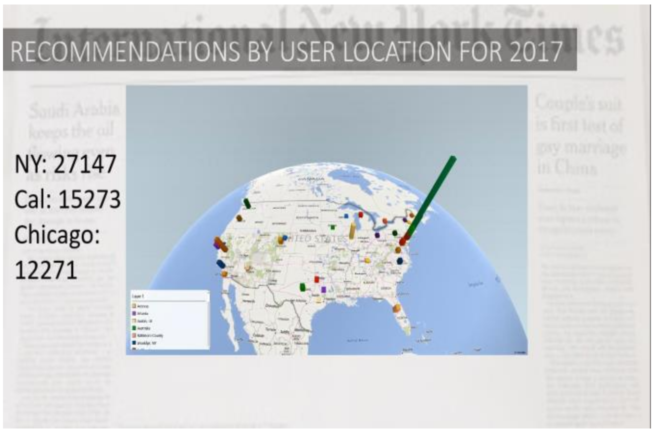


Figure 14

For the year 2018 we found the below results.

- NY: 13990
- Cal: 4251
- Chicago: 3126

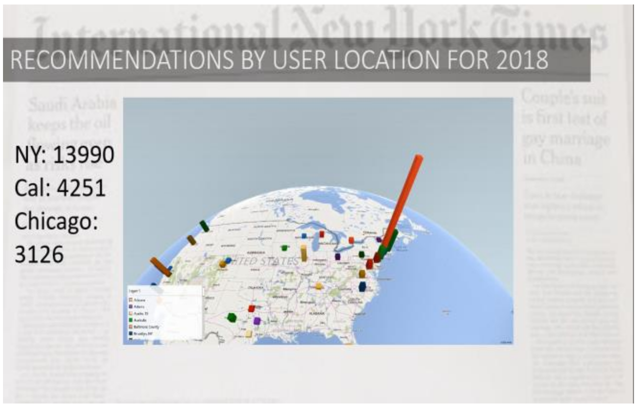


Figure 15

2.7 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

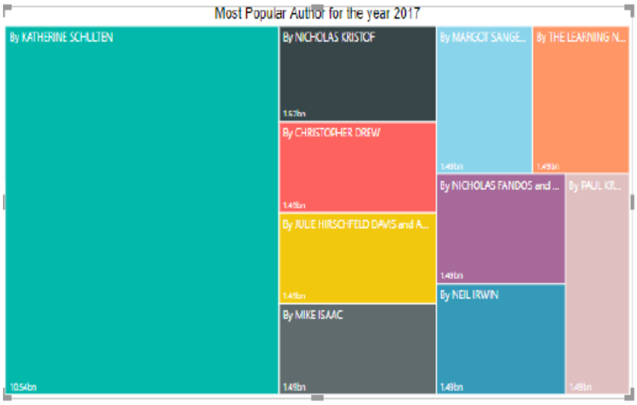
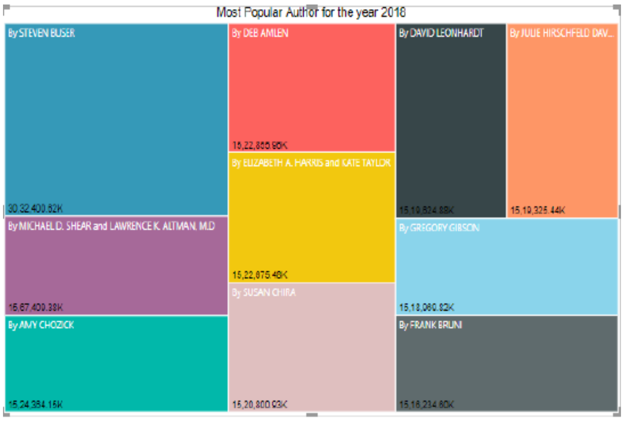


Figure 16



3.Conclusion

While exploring the New York Times Dataset, we successfully used Hadoop, HiveQL, 3D -Maps 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 May. 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. We also investigated the top most areas of people's interest on which they most likely comment as well as determined those topics 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 documents 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 both the years 2017 and 2018. While querying the data we also uncovered that 'news' was the topic that was most talked(read) about in NYT and by investigating the 'headlines' further we uncovered that 'TRUMP' was the hottest topic being discussed which can be linked to the political situation prevalent in the United States and the world.

4. GitHub Link

1. https://github.com/tanvigawade/Project5200_Group3.git

5. References

[1]https://www.kaggle.com/aashita/exploratory-data-analysis-of-comments-on-nyt/notebook

[2]https://towardsdatascience.com/predicting-popularity-of-the-new-york-times-comments-part-1-d32f26261f6f

[3]https://www.kaggle.com/aashita/word-clouds-of-various-shapes

.