**Summary**:

Our Project consists of a dataset extracted from New York Times. We have more than 9000 articles and around 2 million comments on the articles. The aim of this project is to analyze the data by doing sentiment analysis on the articles, building word clouds and see the topic distribution of both articles and comments in order to get the glimpse of readers on the matters of concerning articles.

The data contains information about the comments made on articles published in New York Times. With over 2 million comments and 34 features including editorsSelection, and recommendations, this data can serve the purpose of understanding the public mood. Since, it is very difficult to analyze so much data, we propose to use Natural Language Processing approach to solve this problem.

Analyzing the article data file and comment data file will help us realize which type of article gets most comments by users. We aim to visualize this big dataset using several approaches such as word clouds, topic clusters, sentiment graphs, extracting the hot topics and doing an overall detailed analysis.

**Dataset Overview** :

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*.

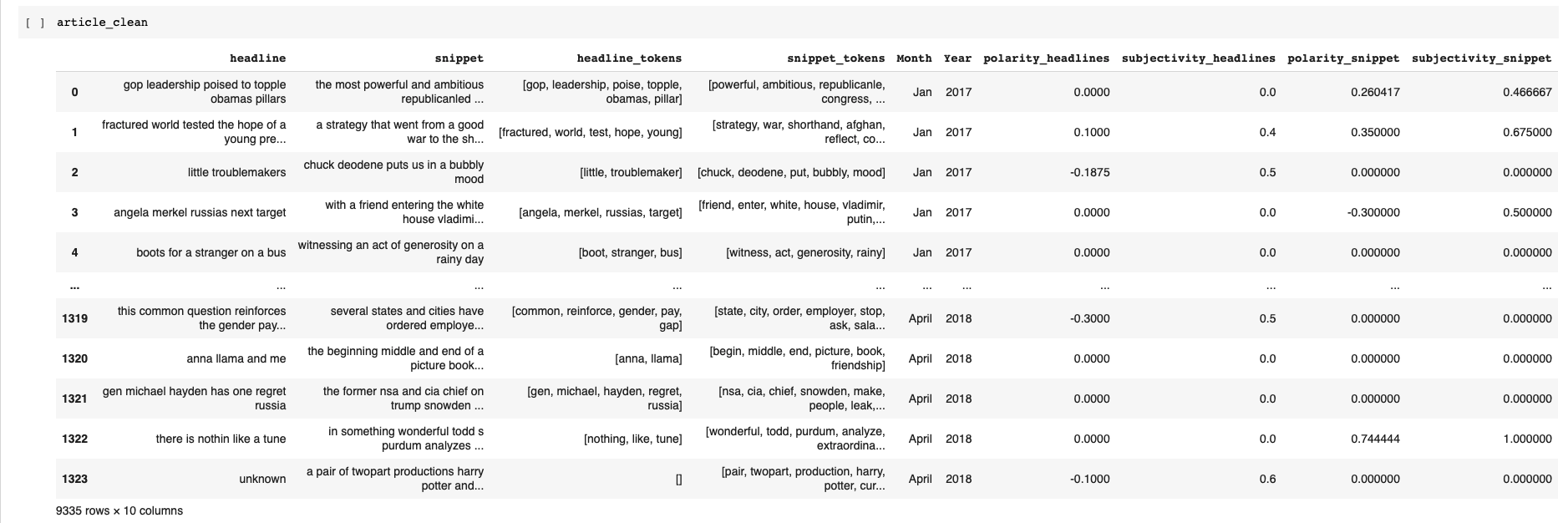
We have features such as article, word count of articles, headline of articles, keywords, editor selection and recommendation to name a few.

**Proposed Plan**:

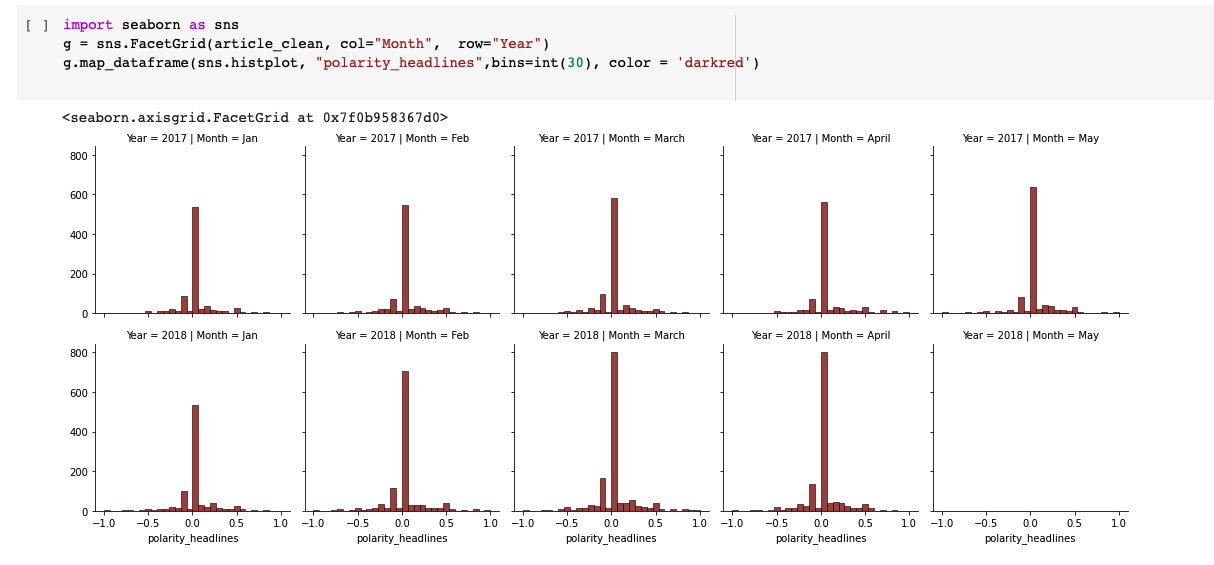
* The initial plan would be loading the data into respective data frames of different timelines according to their content type which is either article or comment. We then plan to clean the data by removing all the features that we will not be using.
* In the data cleaning part, we plan to remove all the irrelevant characters such as any non-alphanumeric characters. We will then tokenize the data and remove all the stop words and apply lemmatization to reduce the inflectional forms. We also will do some stemming in order to get highest quality dataset.
* In the data preprocessing part, we plan to create a good data representation as machine learning models take numerical values of the input, we plan to create Bag of Words and then plan to visualize the embeddings.
* We aim to analyze the behaviors of top commenters as which topic they find the most interesting by performing topic modelling using various algorithms such as LDA and visualize them using pyLDAvis library.
* We then plan to do a sentiment analysis check on the comments so gain more insight from the data.

**Preliminary Result**:

* A picture containing text, newspaper

  Description automatically generatedFirst, we visualized the frequency of words in headlines for each month(Graph 1).

With using the "Wordcloud," in January 2017, the most frequently used words in headlines are "Trumps" and "Obama". In February 2017, the most used word in headlines was "fight". However, in February 2018, the most favorite topic is "gun" and "school." This shows the NYT articles are focusing on the "Stoneman Douglas High School shooting" in that month (Graph2).

* Furthermore, by using "TextBlob," we cleaned the article data and calculated the "polarity" and "subjectivity" for each article's headline and snippet. As the graph shows, the "polarity" of each article in other months is concentrated around 0. It shows that most of the headlines of articles in the NYTimes are neutral (polarity = -1 means the statement is negative, polarity = 1 means the statement is positive)(Graph3).

**References:**

The entire dataset is retrieved using NYT api and collected at the below mentioned Kaggle instance .

<https://developer.nytimes.com/apis> : link to the api

<https://www.kaggle.com/aashita/nyt-comments> for the dataset.

<https://www.nytimes.com> for the articles and headlines

* Follow interested topics for comments data