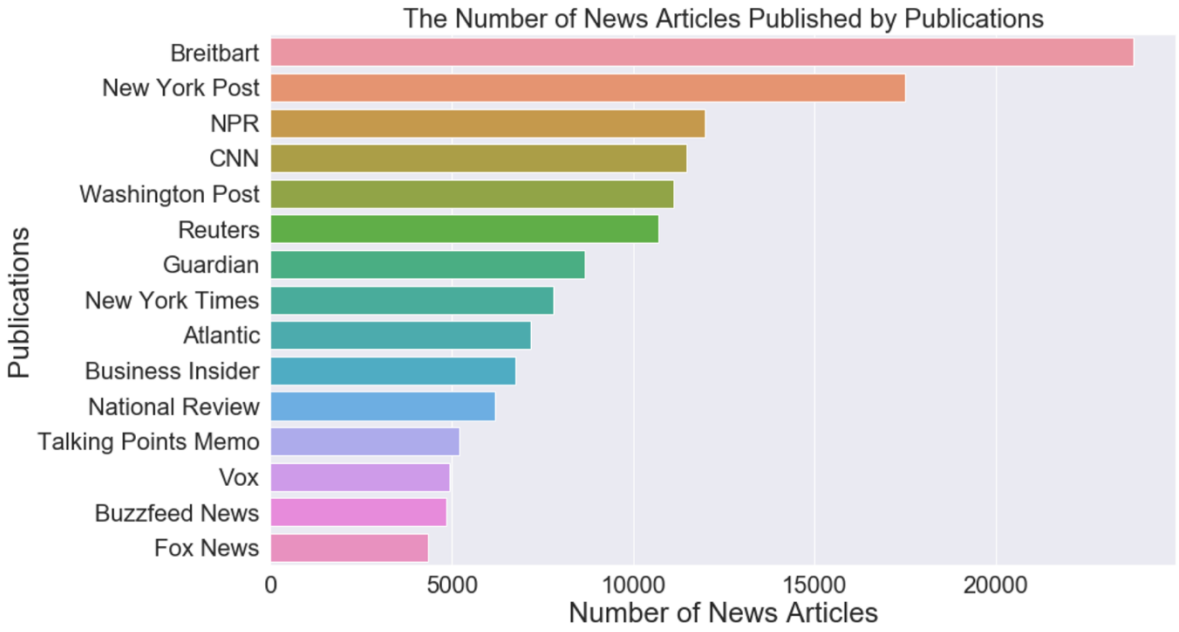
Executive Summary

Today the Internet continues to grow day by day and there are approximately more than 3 billion Internet users worldwide. The information overload has become a problem for a long time because of the growth of the internet. To help people to live their lives easier is my idea of building a text summarizer which is also the main objective of this project. In the meantime, the natural language processing (NLP) will be performed on the news articles to gain some insight of the dataset. This dataset contains 142,570 news articles published in 2016 and 2017 by different publishing companies and is accessible on Kaggle[1]. The features in the dataset are listed below.



- **Title:** The headline of the news article
- **Publication:** The name of the publishing company
- **Author:** The news journalist who wrote the article
- **Date:** Published date
- **Year:** Published year
- **Content:** The whole content of the news article

[1] https://www.kaggle.com/snapcrack/all-the-news/kernels

Text Summarizer

To build the summarizer, I used extractive text summarization which is a combination of using the count vectorizer and cosine similarity. The count vectorizer will convert sentences into vectors and the cosine similarity will return similarity scores between sentences. My last step was to extract the sentences with the highest scores to generate the summary. The advantage of this method is its ease of use but the downside of it is that it can't perform paraphrasing on the sentences to do a more flexible work.

Clustering

Since the original dataset doesn't have labels on the news articles about their categories or genre, I performed clustering on the dataset to try to gain some insight of these articles. I first applied both LSA and NMF to reduce the dimension of the dataset and then used the elbow method to find the cluster number. To be more accurate about the number I got from the elbow method, I also performed tSNE to check how many clusters I might have in the dataset. But I couldn't find a good cluster number in the end. Although the unfortunateness on finding the cluster number, I checked the keywords by changing different number manually and found that generally all the articles can be divided into four categories, Lifestyle, Election, Government Policy, and Business.

Multiclass Classification

Out of my personal curiosity, I wanted to know that if there's any writing styles or patterns in different publishing companies. Since the names of the companies are provided in the dataset, I was able to run a multiclass classification to try to capture that information. I first converted the publications column into a categorical one and then set it as my target variables. My next step was to transform all the articles into vectors using count vectorizer. The results I got from different classifiers are listed below.

Algorithm	F1_Micro Score
Decision Tree Classifier	0.3874
Extra Trees Classifier	0.1894
Gaussian Naïve Bayes	0.3231

As we can see, all the scores are too low for me to interpret the result, which means I might need further cleaning or transforming on the dataset or there might not be a certain writing pattern in each publishing company.