omkarsun (50290136) abhave (50289049)

DATA AGGREGATION, BIG DATA ANALYSIS & VISUALIZATION REPORT

An important and critical phase of the data-science process is data collection. Several organizations including the federal government (data.gov) have their data available to the public for various purposes. Social network applications such as Twitter and Facebook collect enormous amount of data contributed by their numerous and prolific user. For other businesses such as Amazon and NYTimes data is a significant and valuable byproduct of their main business. Nowadays everybody has data. Most of these data generator businesses make subset of their data available for use by registered users for free. Some of them as downloadable data files (.csv, .xlsx) as a database (.db, .db3). Sometimes the data that needs to be collected is not in a specific format but is available as a web page content. In this case, typically a web crawler is used to crawl the web (pages) and scrap the data from these web pages and extract the information needed. Data generating organizations have realized the need to share at least a subset of their data with users interested in developing applications. Entire data sets are sold as products.

We have focused on the current issue which is grappled by the current US government and which has induced a lot of debates as to whether the decision of the Trump Government is correct and indeed required. Yes, we are going to analyze the **"BORDER WALL"** issue. We have also analyzed the issues surrounding it such as 'illegal immigration', 'daca', 'government shutdown', etc.

We collected data from three sources, one opinion-based social media in twitter, research data in New York Times, and the third is the common crawl data. We processed the three data sets collected individually using classical big data methods and compared the outcomes using popular visualization method, Tableau. The activities done for this project are described in detail below.

1) Data Aggregation:

a. Data Collection

We have used python as a programming language for scrapping data from the internet.

Twitter:
We used the ‘tweepy’ library to collect tweets. To get peoples opinion about the issue, we used some popular hashtags such as “Border wall”, “Government Shutdown”, “DACA”, “Immigration”, etc.
The main obstacle while collecting this data was removing the Retweets and duplicate tweets. We removed all the tweets that began with ‘RT@’ to ensure that each tweet was unique. We have **collected** around **27,000 tweets** related to our topics.

New York Times API
For collecting data from New York Times, we have used the New York Times API. We have used the ‘nytimesarticle’ library available for python to get all the URL related to our search query. We also used the ‘requests’ library in python to get access of all the data. Finally, we used ‘Beautiful Soup’ to get the contents of the New York Times article.
We have **collected** around **1000 articles** related to our topic.

Common Crawl
To scrape the data from common crawl, we have used multiple domains like cnn.com, foxnews.com, washingtonpost.com, heraldonline.com, etc.
We have hit these domains through common crawl and then extracted the data for the following time period: ["2019-04","2019-09","2019-13"]
We have received the links to our articles and then using Beautiful Soup, we have extracted the data. We have **scrapped** through around **500 links.**

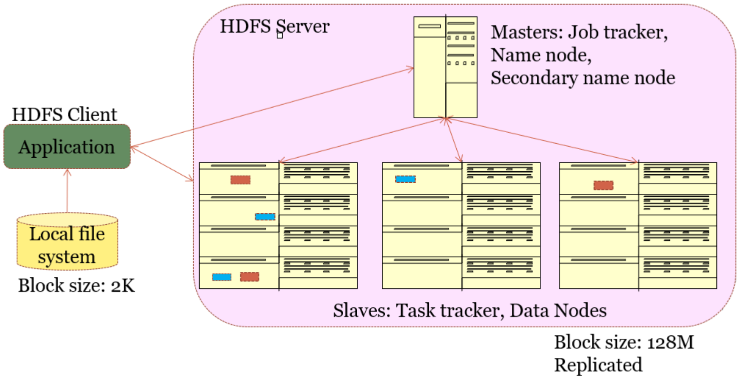
b. Data Processing:

Before executing the Map Reduce job, we had to do data pre-processing to normalize the data and make it consistent.
First, we read all the files from each of the data sources and then merged all the csv files from each of the data sources to create a consolidated csv from each data source. Next, we removed all emojis from the file. We also performed word tokenization on it to remove all stop words like “this”, “and”, etc. We also removed all digits and special characters as well as all ‘https link’ from the data. Finally, we converted the file to a ‘.txt file’ to be fed to Map Reduce job.

2. BIG DATA ANALYSIS

Hadoop Infrastructure:

The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware. It has many similarities with existing distributed file systems. However, the differences from other distributed file systems are significant. HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets. HDFS relaxes a few POSIX requirements to enable streaming access to file system data. HDFS was originally built as infrastructure for the Apache Nutch web search engine project.



We downloaded the required Hadoop files like docker toolbox, and installed the same on our local Windows Machine.

We followed the procedure mentioned in the ClouderaDockerHadoopMR pdf file provided by professor. We increased the RAM to be provided to the virtual box and ran the sample word count program successfully.

Map Reduce

We wrote a MapReduce code for word count and word co-occurrence in python. Word count is basically counting the frequency of each word in the document. The output gives us very little idea about the issue as only the keywords are highlighted.

To get better idea of the issue, we perform word co-occurrence which gives us a better inference about the problem.

For word co-occurrence, mapper was not able to understand if <Word1, Word2> are in the same order or not. E.g. Mapper cannot differentiate between <" trump", 'wall"> and <" wall"," trump">, so we wrote a script for the same which aggregates all the result.

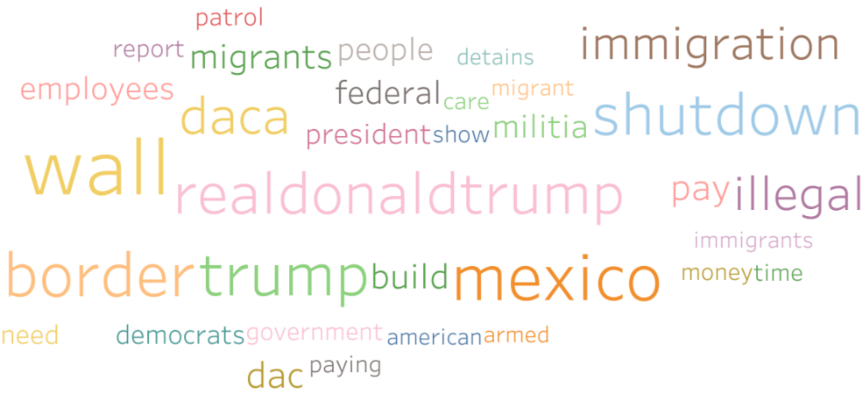
Final step was to visualize our outputs in a word cloud.

3. Visualization

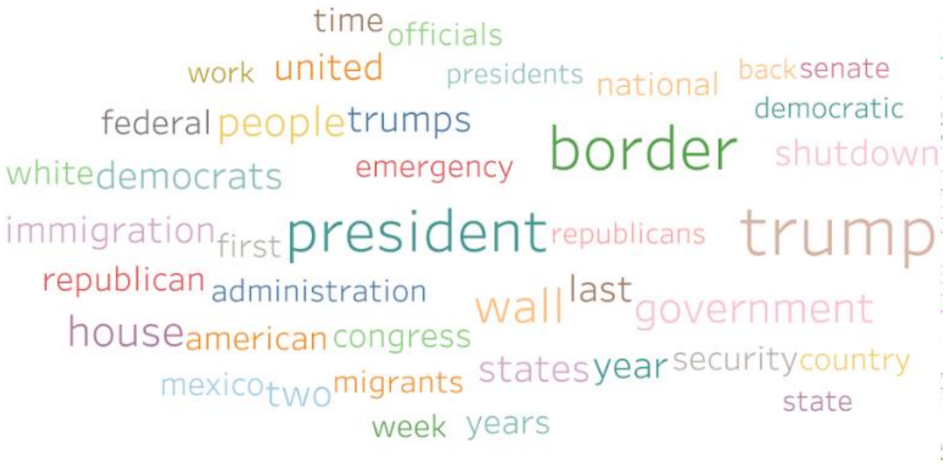
Data visualization is viewed by many disciplines as a modern equivalent of visual communication. In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions. Data visualization is another form of visual art that grabs our interest and keeps our eyes on the message. When we see a chart, we quickly see trends and outliers. If we can see something, we internalize it quickly. It’s storytelling with a purpose. If you’ve ever stared at a massive spreadsheet of data and couldn’t see a trend, you know how much more effective a visualization can be.

Tableau is one way to visualize data. Tableau can help anyone see and understand their data. The following were the word cloud outputs from our data. For word count we have taken the top 50 words and for word co-occurrences we have taken the top 20 words.

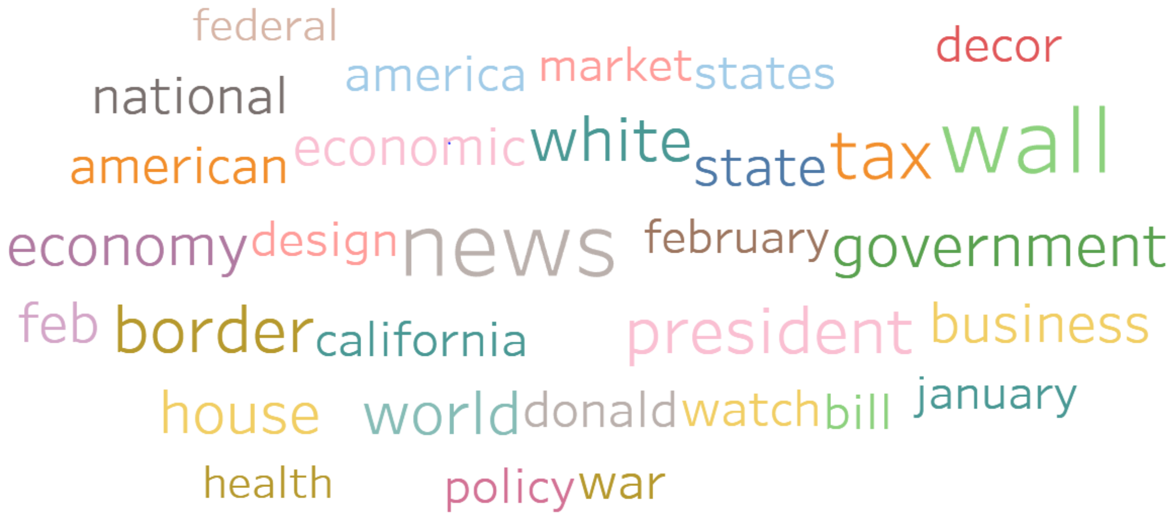
Word Cloud for Word Count



Twitter

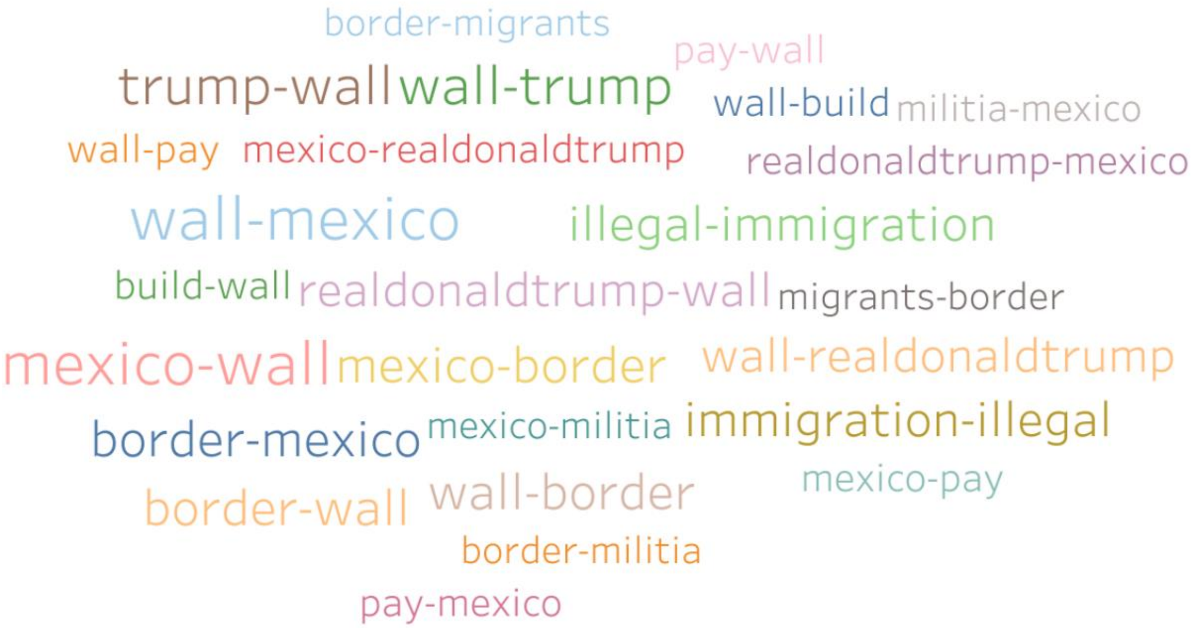


New York Times

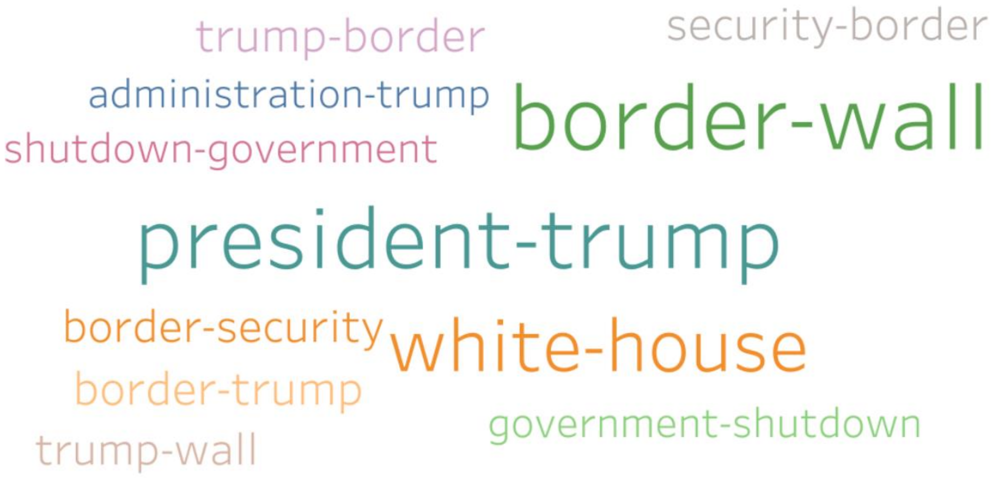


Common Crawl

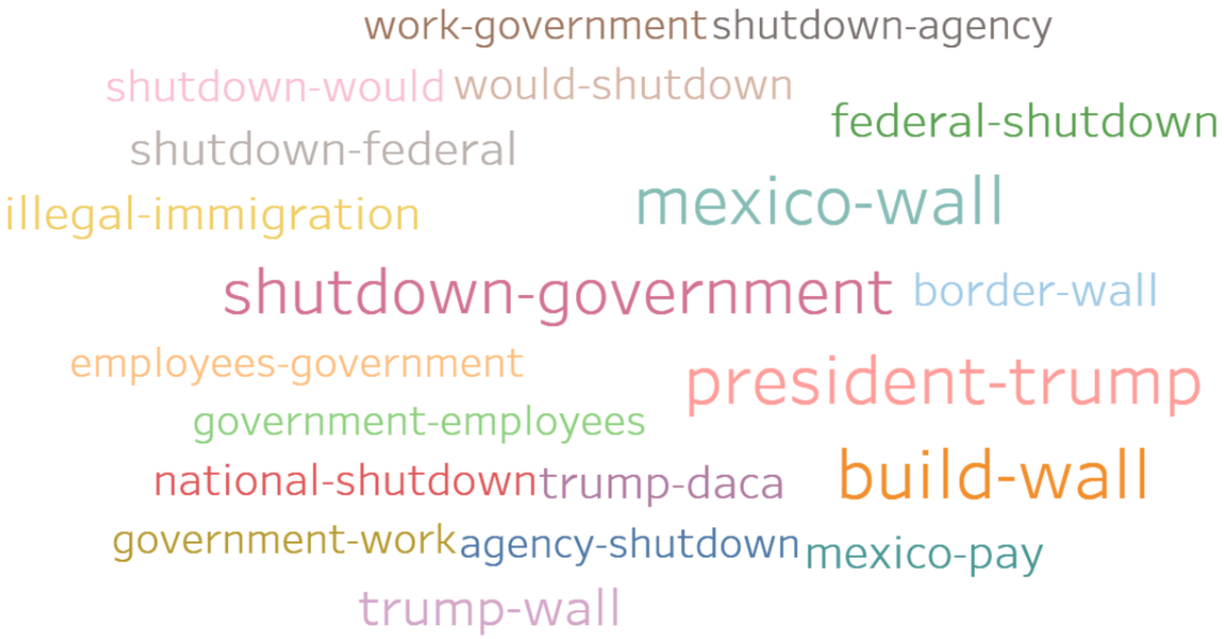
World Cloud for Word Co-occurrence



Twitter



New York Times



Common Crawl

Note:

You can view these results using the following link. Just need to login using your credentials.
https://us-east-1.online.tableau.com/#/site/dic587lab2/projects/157196?order=contentTypeOrder:asc,name:asc

REFERENCES:

[1] https://www.tableau.com/learn/articles/data-visualization

[2] https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html

[3] Lecture Notes

[4] Lab Manual