DIC LAB – 2 REPORT : DATA AGGREGATION, BIG DATA ANALYSIS AND VISUALIZATION

Muthuvel Palanisamy – 50246815

Akshay Chopra – 50248989

TOPIC :

“Facebook and Cambridge Analytics Data Breach”

**Part – 1: Data Collection**

* R is used as the language for data collection and cleaning
* A simple block diagram is shown in figure 1

**NY TIMES articles:**

* **Rtimes** packages is used to extract the url of articles using keyword like “*cambridge annalytica”, “ facebook scandal”, “facebook data leak”…*
* **Contentscrapter in Rcrawler** package is used to crawl the webpages containing the url and extract the article content
* A total of around 250 articles is present in the data collected after removing duplicated articles and articles that does not belong to the topic
* The Articles collected is present in directory Part2/Data/articlesTotal.txt

**TWITTER tweets:**

* A total of 10,500 tweets is collected on the topic using the keyword described above and hashtags like *“#deletefacebook”*
* **TwitteR** package is used to extract tweets
* Tweets are filtered using the text, tweet ids to remove duplicates
* Data is cleaned by remove non-ASCII characters, symbols as preprocessing step
* The tweets collected in present in Part2/Data/tweetsTotal.txt

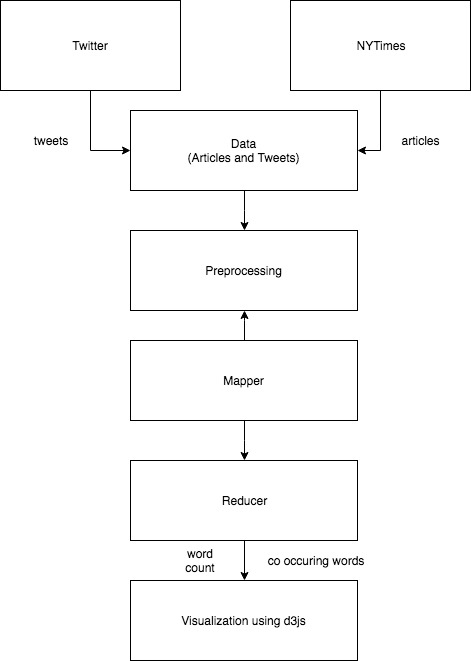


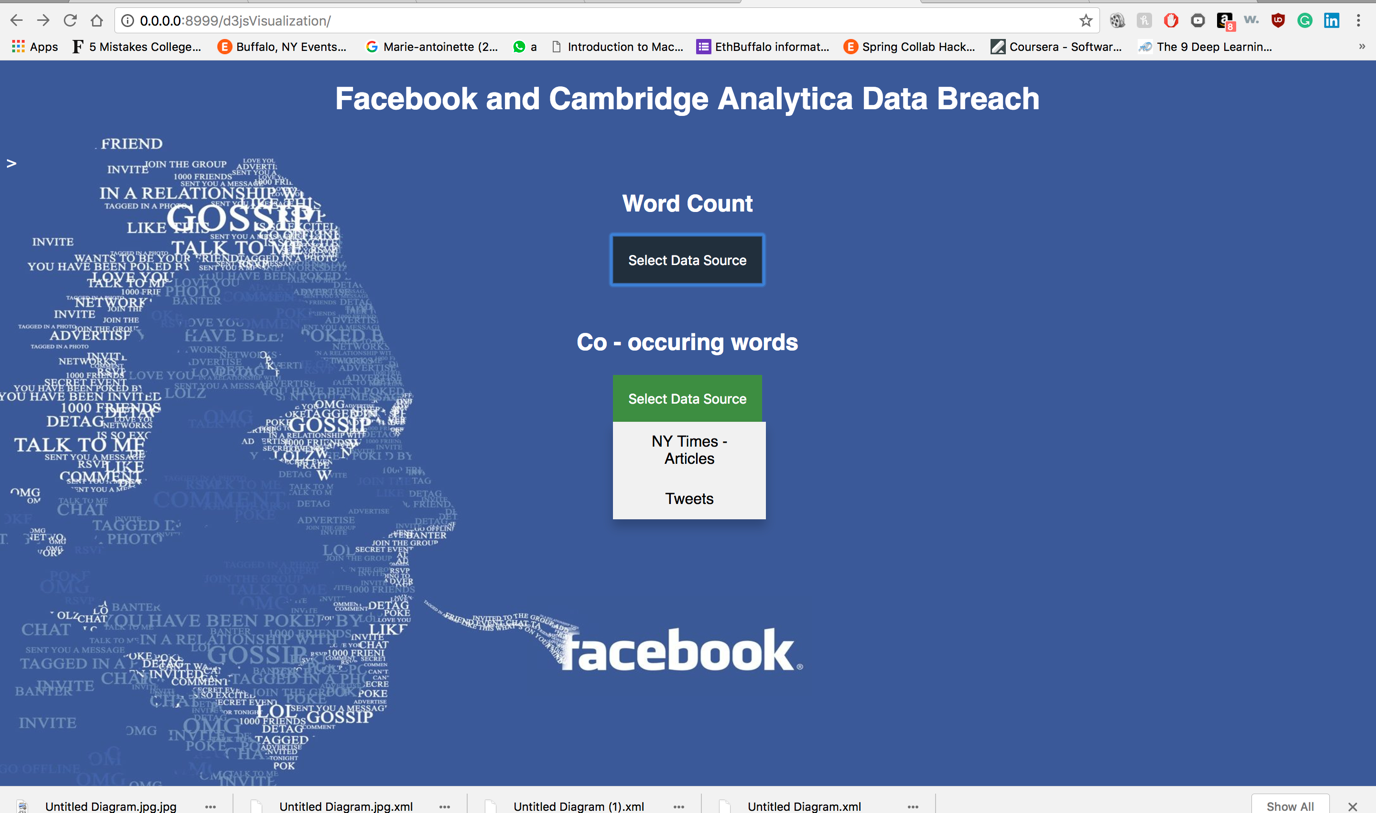
Figure 1: Block diagram of our implementation

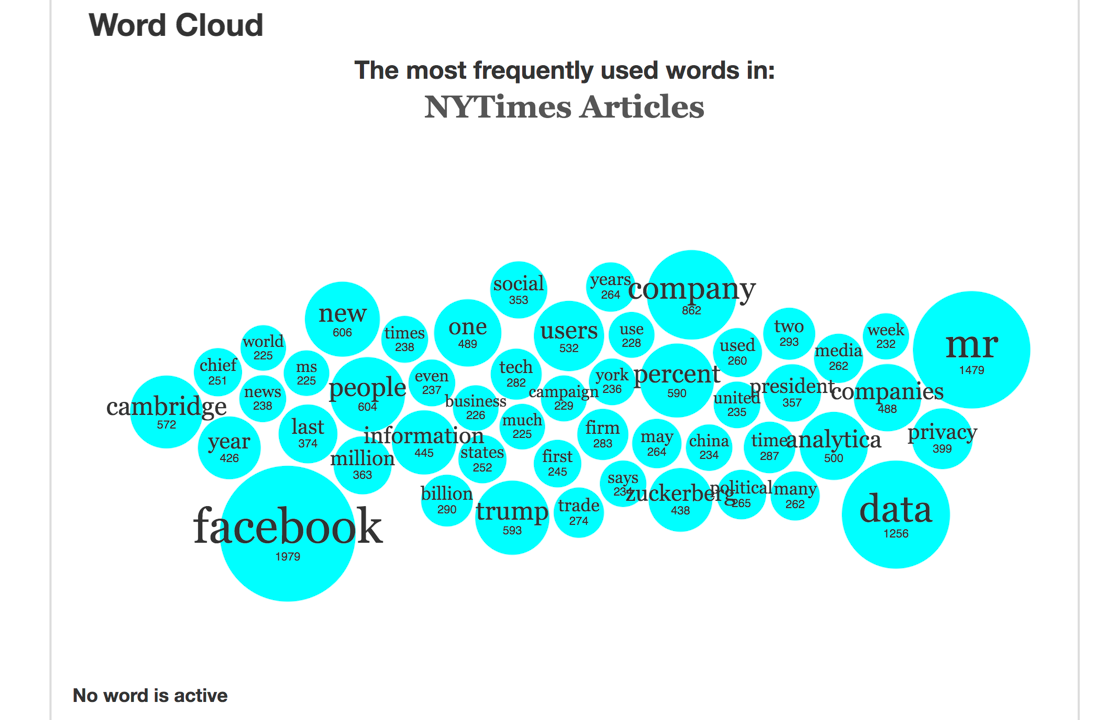
**Part – 2: Word Count using map reduce in Hadoop**

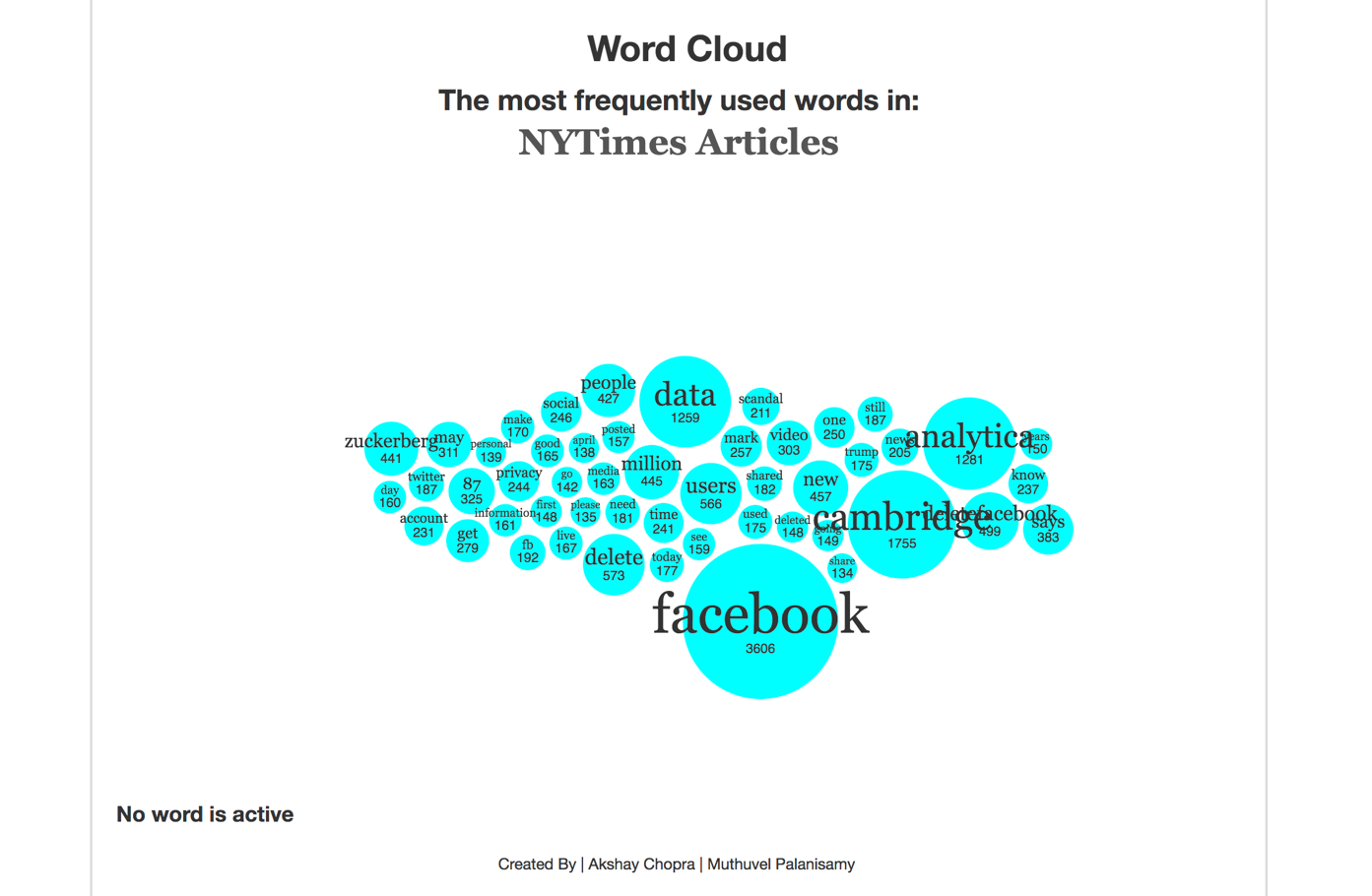
**Part – 3: Visualization of word count**

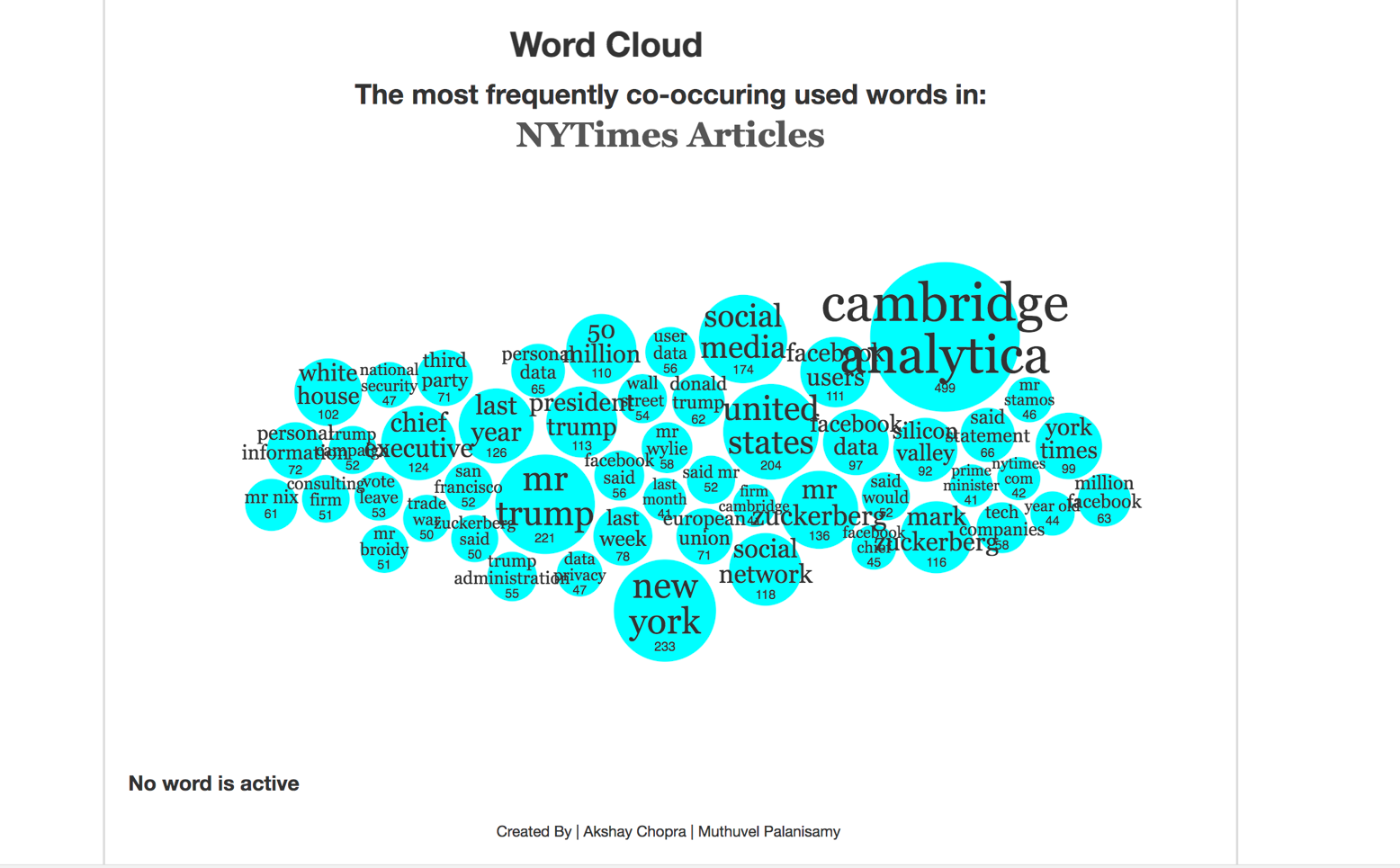
* Used **D3.JS** to visualizing the word cloud for the **top 50** word count and co- occurring words
* The output is visualized in a webpage as follows
* Each blob visualized displays the count and the word/co-occuring word, with the size of the blob depending on the count

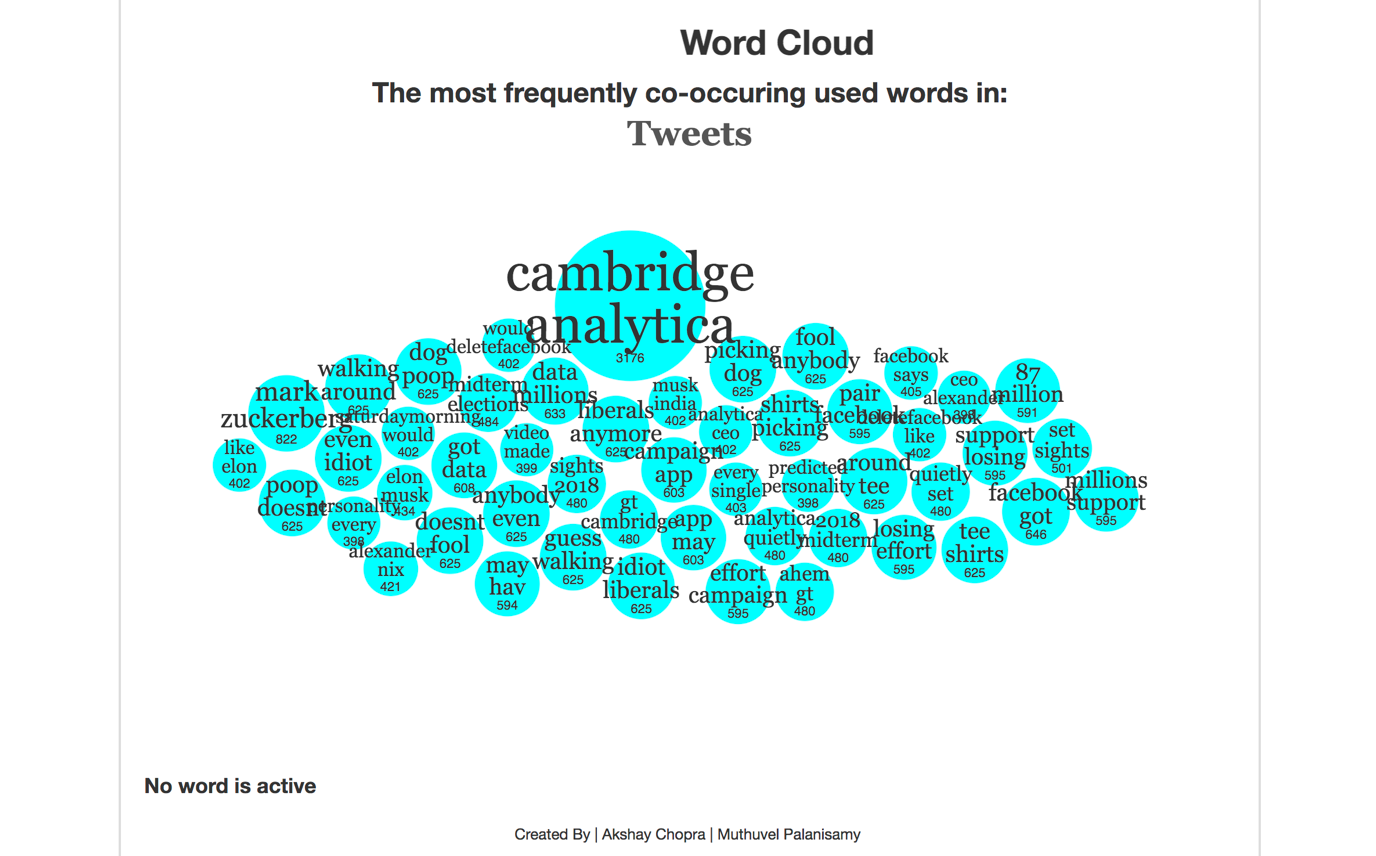
Homepage:

****









**Analysis**

* Words like “cambridge”, “analytica”, “facebook”, “Zuckerberg”, “data” were having very high frequency and were found to be common in both tweets and articles data