Documentation for DIC Lab 2

Kshitij Goel – 50246430 Roshni Murali – 50247191

https://buffalo.box.com/s/awi6dbcp5zu4tean5h4rv1wms7w739ei (Video Link)

Introduction

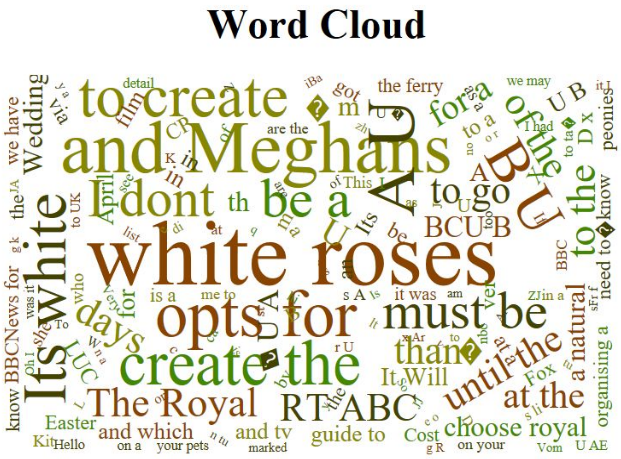
This lab allows us to to get a deeper understanding on how to handle big data for processing and visualize them for better portrayal to the user. This gives us the skill set to work in the real world environment to collect data and process it so that better judgements could be made after taking into consideration the predicted/portrayed model from the data. This improves efficiency and helps make informed decisions. The following step were followed for the lab:

Steps

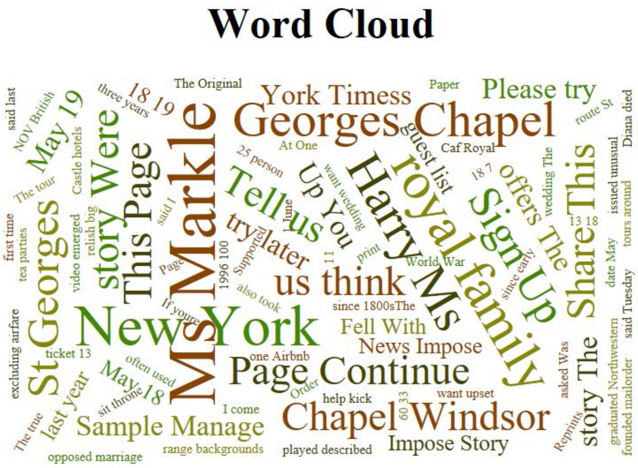
1. We collected tweets on the specific topic of “Royal Wedding” for over a week using the Twitter API via R Studio and saved them into a series of files.
2. We also collected articles by the New York Times (NYTimes) using their API on different days of the week on the same topic.
3. We removed the punctuations and stop words from the input files using the NLTK package in python and also formatted it for easier processing by the Hadoop MapReduce.
4. We then, ran the processed files through the Hadoop MapReduce using the following commands:\
    a. i). hadoop fs -put /home/hadoop/DIC_Lab_2/article.txt

    b. ii). hadoop jar /home/hadoop/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.6.4.jar -file /home/hadoop/DIC_Lab_2/mapper.py -mapper /home/hadoop/DIC_Lab_2/mapper.py -file /home/hadoop/DIC_Lab_2/reducer.py -reducer /home/hadoop/DIC_Lab_2/reducer.py -input /home/hadoop/DIC_Lab_2/article.txt -output home/hadoop/DIC_Lab_2/article_uni.csv

    c. iii). hadoop dfs -getmerge home/hadoop/DIC_Lab_2/article_uni.csv /home/hadoop/DIC_Lab_2/article_uni.csv
5. We converted the obtained data from (.csv) to (.json) for easier manipulation and reading for the javascript file.
6. The converted javascript files were then used as an input for a WordCloud visualization of the data. This step is a very important step of the process as it is the final product which will help the user to make decisions upon.



Word Cloud

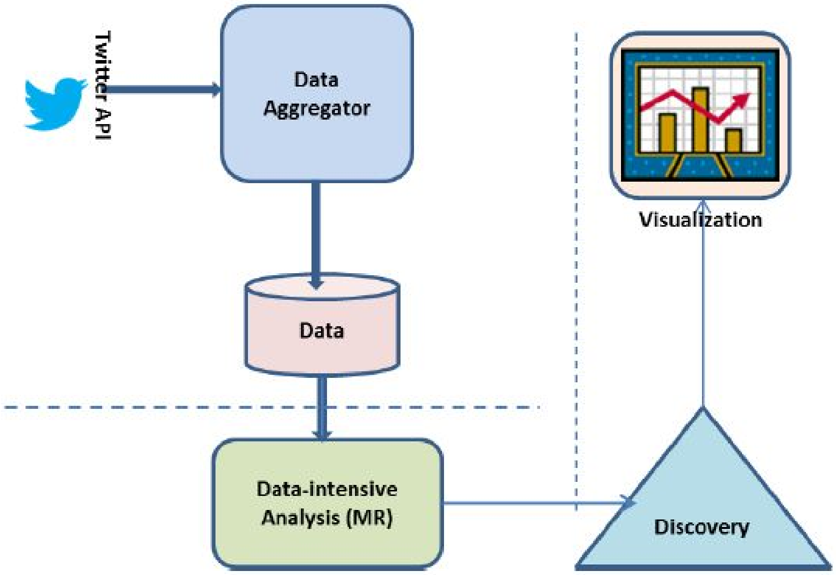


Word Cloud

Description

We used the Hadoop MapReduce for processing the data for counting the most used words in the tweets and the NYTimes articles. We also found the co-occurrence words for the most used words and mapped those to the WordCloud so that the user can easily figure out the patterns over the world regarding that topic.

The Hadoop MapReduce uses a cluster based data structure for processing the data as shown below in the figure.



As we can see from the diagram, The data was taken via the Twitter API and then processed by the data aggregator. The data was then passed on to the MapReduce which was then processed for better visualization. The data was then visualised using the D3.js library for an interactive display of the data. The interactiveness help the user in understanding the trends in a better way and can be modified as per the needs.

Conclusion

With the help of this lab, we learned many skills which help us in the real world for data analysis and gave a deep understanding on the usage and processing of big data via Hadoop MapReduce.

References

https://github.com/wvengen/d3-wordcloud

http://www.michael-noll.com/tutorials/writing-an-hadoop-mapreduce-program-in-python/