# CSE 587 – Data Intensive Computing
# Project 2

Aakash Gupta – aakashgu – 50289067
Sai Saket Regulapati – saisaket – 50286747

## Contents

# INTRODUCTION

This project is a combination of Data aggregation (collecting data from different sources), Big Data Analysis (deriving valuable information out of the collected data) and Data Visualisation (representing the found information in a user friendly way).

The main steps in the making of this project are given below:

1. Data Collection/Aggregation
2. Data Preprocessing
3. Performing word count and word co-occurrence
4. Visualizing the data
5. Building a web page for the visualized data

We have made use of jupyter notebook using the python kernel for data collection, analysis and tableau for the visualization part.

The above mentioned steps are further explained in the following sections.

# DATA COLLECTION

The data required for the project has been mainly collected from three main sources.

1. Twitter
2. New York Times
3. Common Crawl

We have used different API's and libraries to retrieve the data.

Twitter: We used the openly available library 'tweepy' for the retrieval of tweets related to our topics.

New York Times: We used the library provided by the New York times (NYTimesArticle) to retrieve the data.

Common Crawl: We have used the MRjob file from the common crawl website that sets up the Hadoop architecture. The data will be stored in Amazon S3. BeautifulSoup has been used to parse the json and get useful data

# DATA PREPROCESSING

We have used the various preprocessing techniques for the three data sources. Some of the processes and techniques we have used are only including alphanumerics, spaces and using stop words and stemming. All the preprocessing has been done in the python in the mapper.

Duplicate tweets have been removed and care has been taken so that all tweets are unique and there are no retweets included in the dataset.

# MAP REDUCE: WORD COUNT, WORD CO-OCCURRENCE

We have used the Hadoop infrastructure for the implementation of the map reduce part. We have used the AWS for the Hadoop infrastructure.

The mapper reduce process of the system is the most crucial stage of the project. The mapper generally takes the words (and in our case performs the preprocessing too i.e, it removes stop words and does stemming so preprocessing is a parallel process not sequential).

The reducer gets the list of words as input and basically returns the count of the words in case of word count and the co-occurrence of words in the second case.

Diagram of Map-reduce for word count.

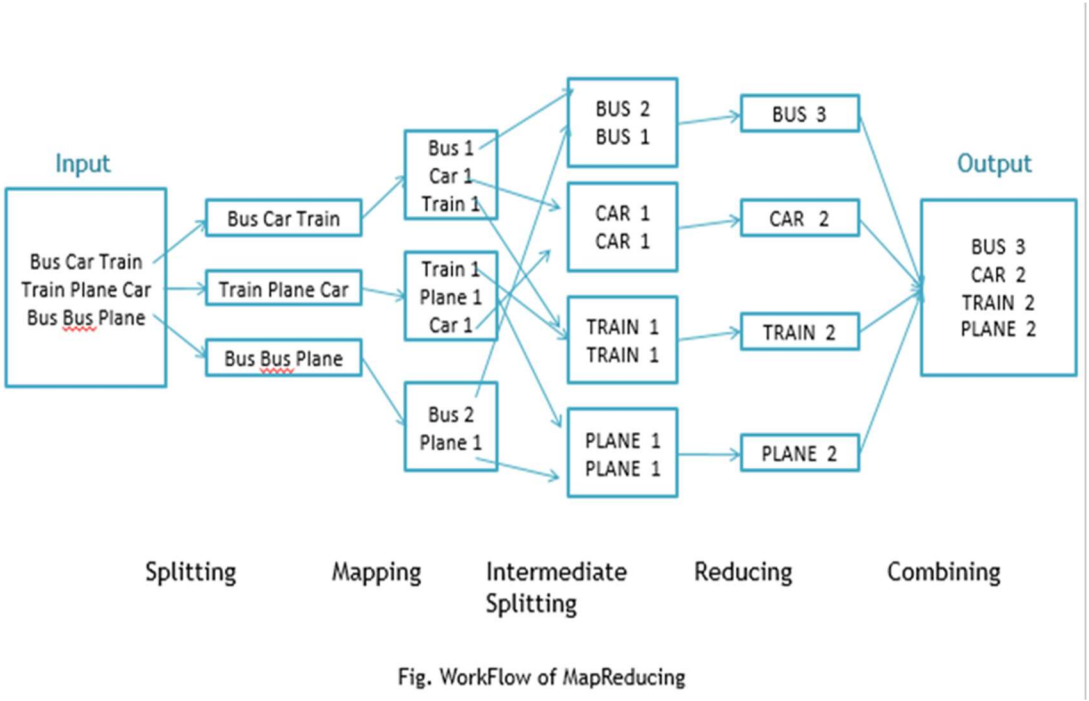
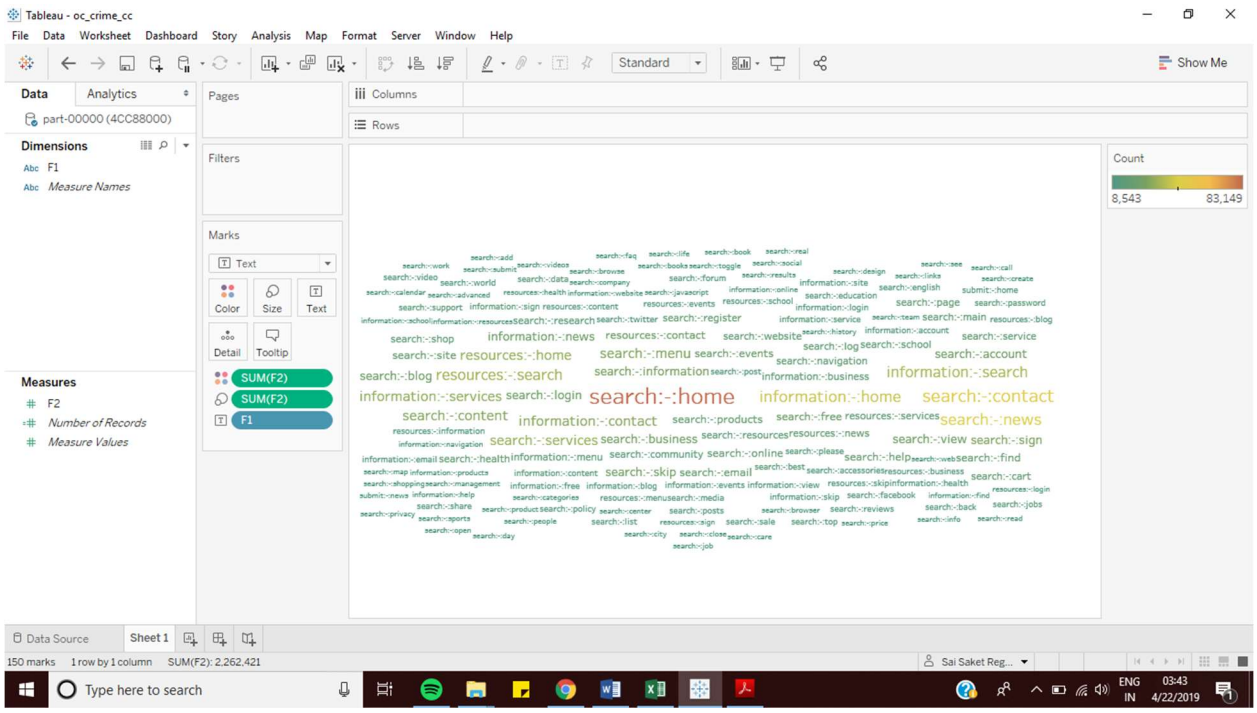


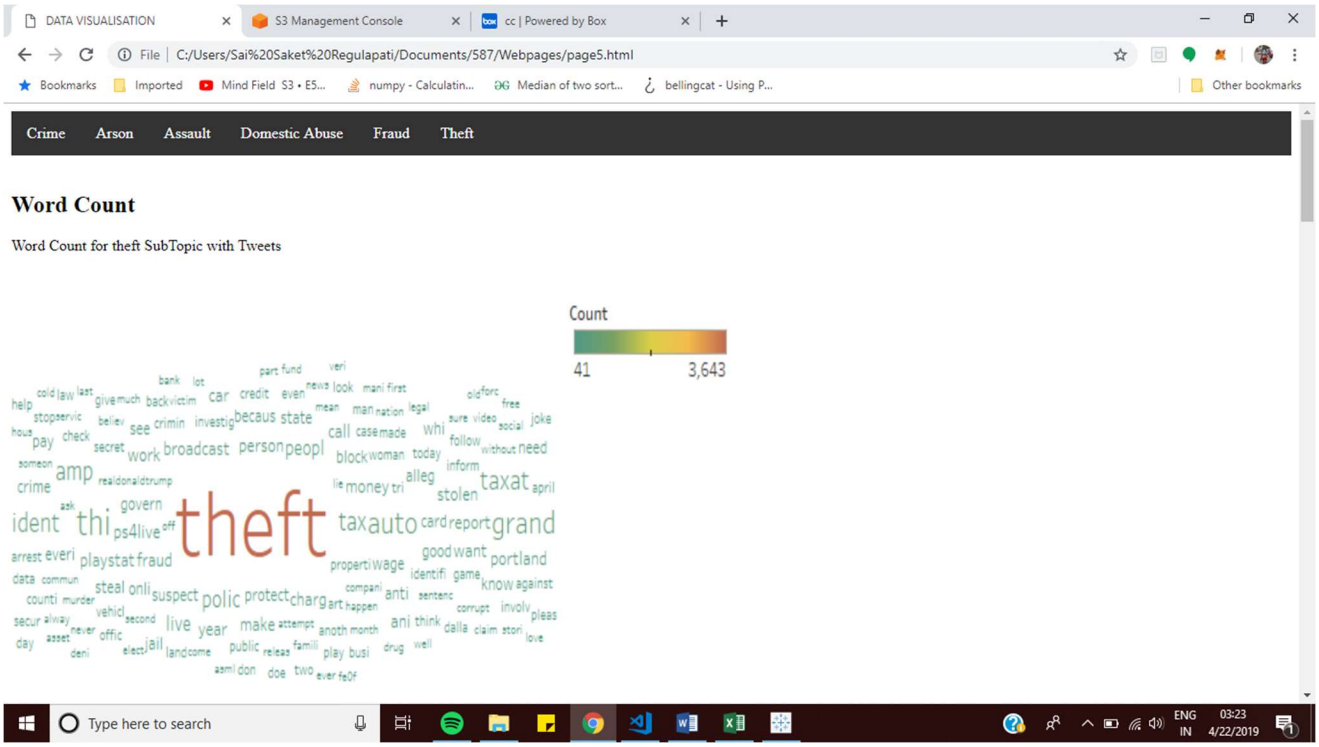

Fig. WorkFlow of MapReducing

# DATA VISUALIZATION

We have used tableau for the representation of data. Tableau is very simple to use. We just upload the excel sheet to the tableau. Drag the names to labels and the count to size and color.




Now that we are done with this let's see the webpage we have designed. We made the webpage dynamic and the webpage will have the images in them.



DATA VISUALISATION | S3 Management Console | cc | Powered by Box

File | C:/Users/Sai%20Saket%20Regulapati/Documents/587/Webpages/page3.html

Bookmarks | Imported | Mind Field S3 • E5... | numpy - Calculatin... | Median of two sort... | bellingcat - Using P... | Other bookmarks

Crime | Arson | Assault | Domestic Abuse | Fraud | Theft

## Word Count

Word Count for Domestic Abuse SubTopic with Tweets


Type here to search

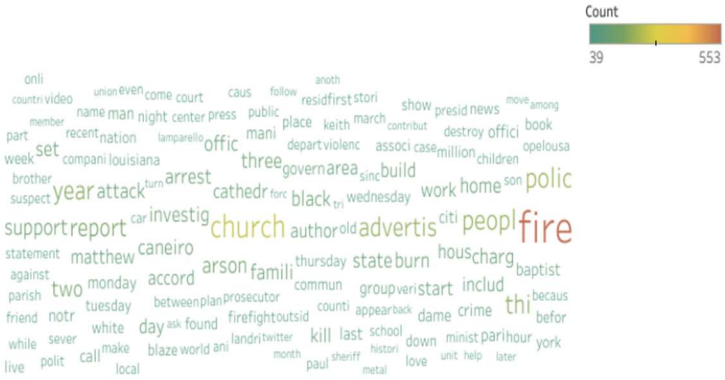
ENG IN 03:23 4/22/2019

DATA VISUALISATION | S3 Management Console | cc | Powered by Box

File | C:/Users/Sai%20Saket%20Regulapati/Documents/587/Webpages/page.html

Bookmarks | Imported | Mind Field S3 • E5... | numpy - Calculatin... | Median of two sort... | bellingcat - Using P... | Other bookmarks

Word Count for Arson SubTopic with NYT




Type here to search

ENG IN 03:23 4/22/2019