CSE 587 - Lab 2 report

Data Aggregation, Big data analysis and Visualization

Submitted by:

Saketh Varma, Pericherla - sakethva - 50288206

Aditya Vikram, Parakala - aparakal - 50289171

Instructions:

- Create a new virtual environment. If you are using anaconda use

```
$ conda create --name env_name
```

- Activate the created environment

```
$ conda activate env_name
```

- The project already comes with a requirements.txt file so use the below command to install all the project dependencies

```
$ pip install -r requirements.txt
```

- Create a `.env` file in the root of the project and store secrets like TWITTER_CONSUMER_KEY, NYTIMES_API_KEY etc

- Type python on the command line and type the following commands to install nltk packages:

```
C:\Users\socket_var\Desktop\dic>python
Python 3.7.3 (default, Mar 27 2019, 17:13:21) [MSC v.1915 64 bit (AMD64)] ::
Anaconda, Inc. on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> import nltk
>>> nltk.download("wordnet")
>>> nltk.download("stopwords")
>>>
```

- To get the website up and running install the npm package called `serve` by using the command:

```
$ npm install serve -g
```

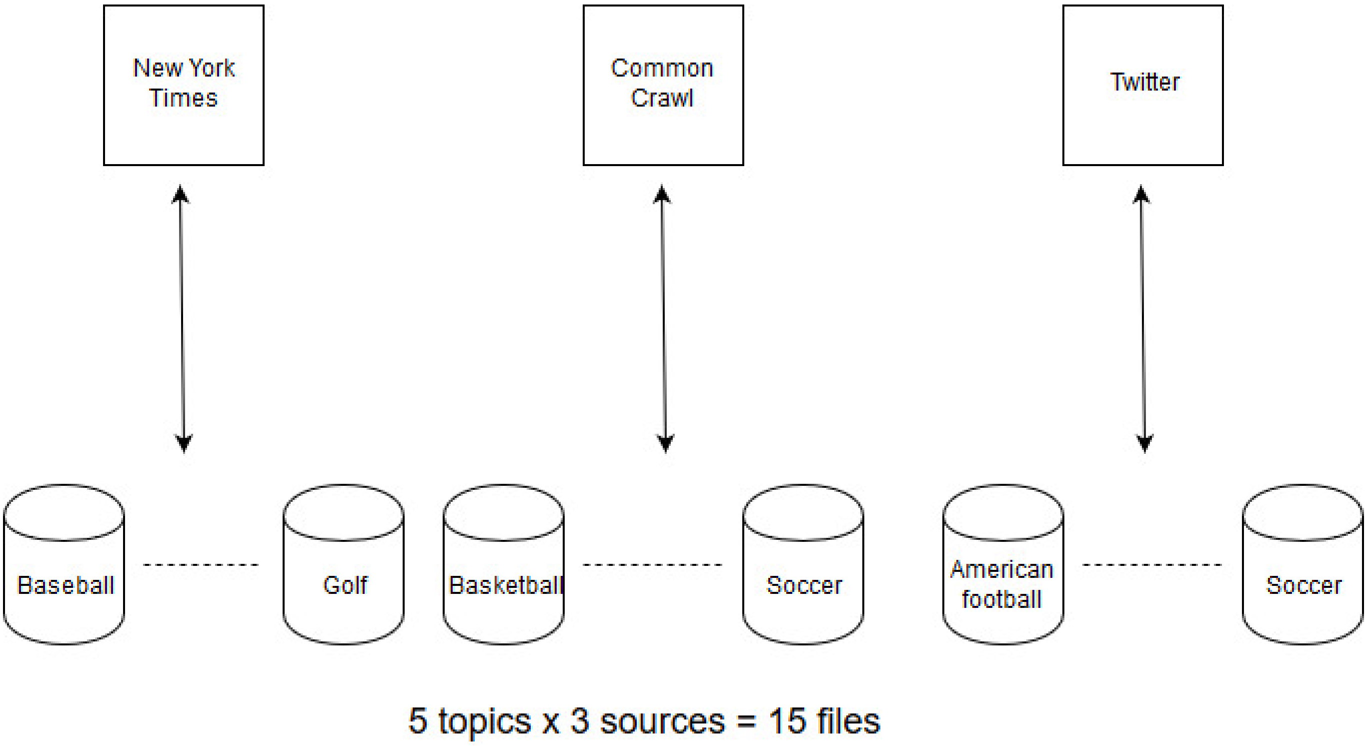
- Now go to the ((website root directory here)) and just type

```
$ serve
```

- Head over to http://localhost:5000 and browse through the visualizations.

Data collection:

The below flow chart summarizes the data collection logistics:



New York Times:

Common Crawl:

Twitter:

Data processing:

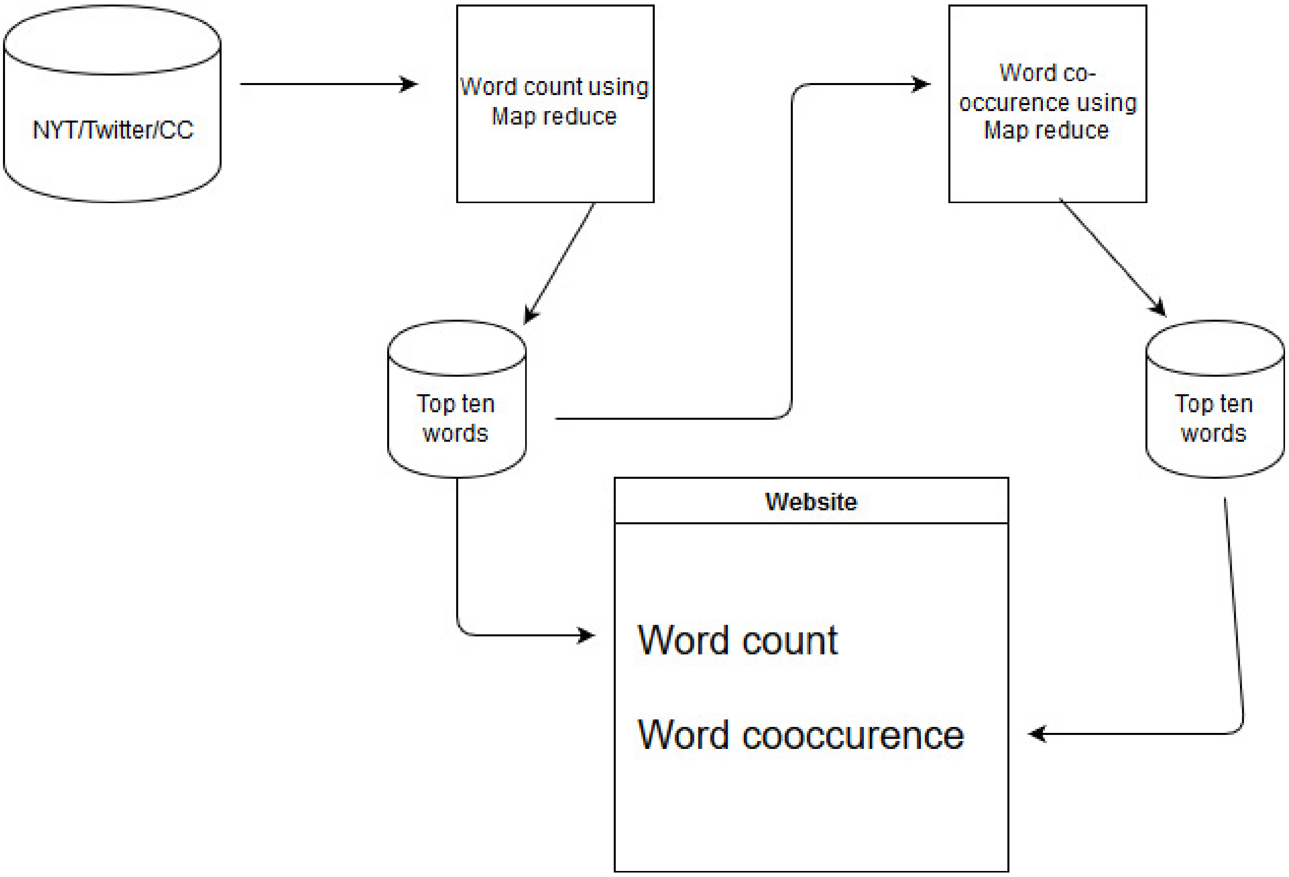
New York Times:

Common Crawl:

Twitter:

Hadoop workflow:

The below flow chart summarizes the process of performing big-data analysis for this lab:



5 topics x 3 data sources x 2 algorithms = 30 visualizations

Word count algorithm:

Word co-occurence algorithm:

Visualization using d3.js:

Results:

Baseball:

Basketball:

Soccer:

American Football:

Golf: