CSE 578 Lab 2 Report

AUTHENTICATION

NYTimes: Generate API key for NYTimes.
Twitter: API key and API secret for Twitter.

WEB SCRAPING

NYTimes:

1. Use search API to get article URLs in a given time range.
2. Navigate to NYTimes URLs and parse through the data.
3. Pick only 'p' tags which contain article text and beautify it using 'beautifulsoup' package.
4. Join all the paragraph information and save it as NYTimes data. **Twitter**:

1. Collect tweet text using 'rtweet' and filter it by date.
2. Write it to Twitter data file.

MAPREDUCE

Start hadoop from CMD in the VM.

Mapper:

1. Feed mapper with Twitter data file and NYTimes data file separately in HDFS.
2. Mapper reads the input from CMD,removes stop words and special characters.
3. Read each line from the data files and emit <word,1> to reducer **Mapper for bigrams:** Take pairs of words at a time and emit <word,co-occurring word,1>

Reducer:

Combine and aggregate the <key,value> pairs and emit the key word and it's frequency.

Finding co-occurrence words:

Sort the unigram reducer output by values and find the co-occurrence words for the top 50 words.

WORD CLOUD GENERATION

Input: Processed text files from MapReduce

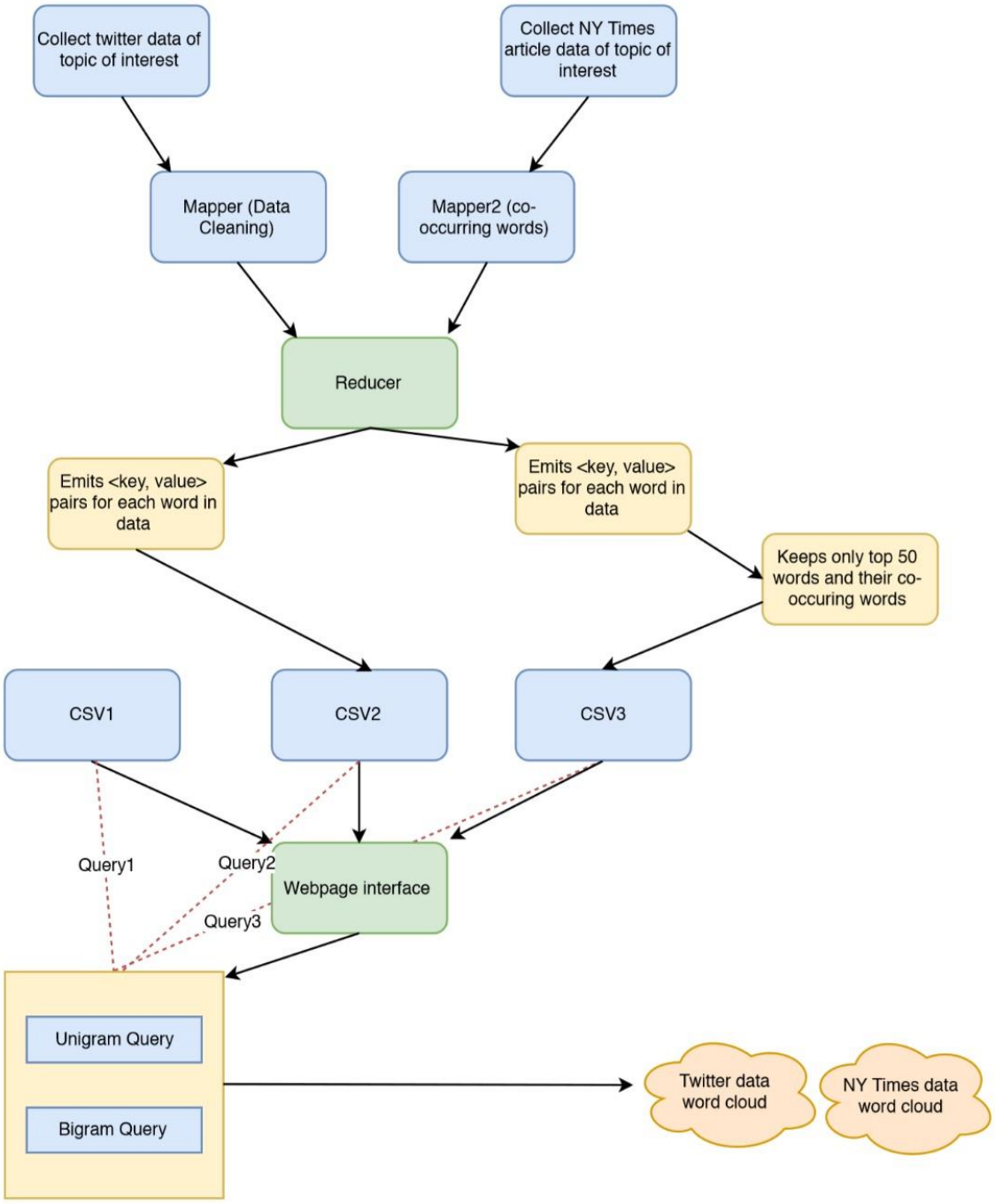
Output: Word cloud

1. Input the data file to word cloud program
2. The word cloud program has 3 types of files:
    a. Data files (csv)
    b. Html files (for the webpage)
    c. JavaScript files (for d3.js interactions)
3. The webpage is hosted on tomcat and displays three dropdowns and three buttons for unigram day, unigram week, and bigram data
4. This was done using html and css for the styling
5. Each dropdown has a set of predefined queries which the user can select and make a choice between generating the word cloud for unigram (single words) or bigram (cooccurring word pairs)
6. Based on the query, the program fetches the required data file and generates the word cloud
7. For each query, two word clouds are generated side by side (the first created using Twitter data and the second using NY Times data)
8. Words with higher frequencies of occurrence are displayed in progressively increasing sizes

FLOWCHART

We have created a flowchart depicting each step of the process we followed to generate the word clouds.

Inputs are represented in blue, outputs in yellow/orange, and processing elements in green.


Collect twitter data of topic of interest
Collect NY Times article data of topic of interest
Mapper (Data Cleaning)
Mapper2 (co-occurring words)
Reducer
Emits <key, value> pairs for each word in data
Emits <key, value> pairs for each word in data
Keeps only top 50 words and their co-occuring words
CSV1
CSV2
CSV3
Query1
Query2
Query3
Webpage interface
Unigram Query
Bigram Query
Twitter data word cloud
NY Times data word cloud