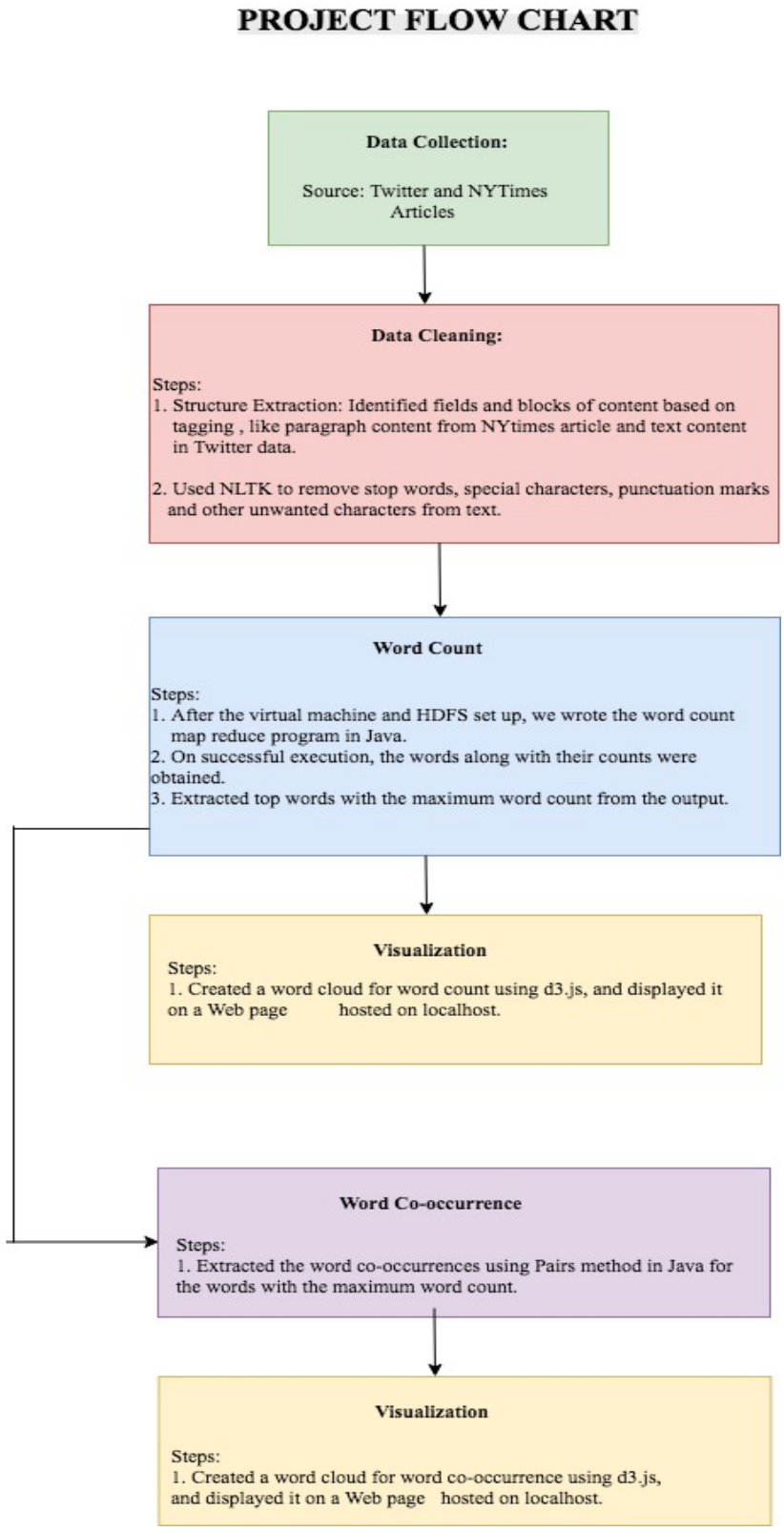
DATA INTENSIVE COMPUTING: LAB 2

The flowchart below explains the steps followed in completing the tasks for this LAB.



Extensions:
The whole project as a whole can be reused for analysis of other datasets by just inputting the respective data source in the first step of Data Collection. The Input data can be from any source, which can then be preprocessed and other functions can be performed.

Folder Structure:

The Description below describes the folder structure aligned with the steps mentioned above:

sharmaLab2:

1. **Data Collection**: Contains the collected data from Twitter and NYtimes.
2. **Data Cleaning**: Contains the python code to preprocess the data obtained.
   a) sharma_Part2_twitter_processing.ipynb – Contains code to preprocess twitter data.
   b) sharma_Part2.ipynb – Contains code to preprocess NYTimes article data
3. **Word Count**:
   a) Contains WordCount.java file used to calculate word counts using Map Reduce
4. **WordCount Output:**
   a) Contains the output word counts obtained from the Map reduce.
   b) Top words - Word count output:
      i) This folder contains the top word counts extracted from the output and stored in json format to give as an input for d3 visualization.
5. **Co-occurrence:**
   a) Contains Pairs.java to find out the word Co-occurrences based on the collected data.
6. **Word Co-occurrence output:**
   a) This folder contains the output obtained from Pairs.java, i.e json file required for d3.js
7. **Visualization:**
   a) Contains index.html and d3.layout.cloud.txt required to visualize all the above obtained results.
   b) sharmaTopWordCount.ipynb – Contains the code to extract top word counts using python to display.
   c) sharmaTopWordCooccurrence.ipynb – Contains the code to extract the top word cooccurrences to display.
8. **Part 1:**
   a) Contains three files, containing each of the code for Chapter 1, 2 and 3.