DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
UNINVERSITY AT BUFFALO

CSE 587: DATA INTENSIVE COMPUTING
LAB-2

DATA AGGREGATION, BIG DATA ANALYSIS AND VISUALIZATION

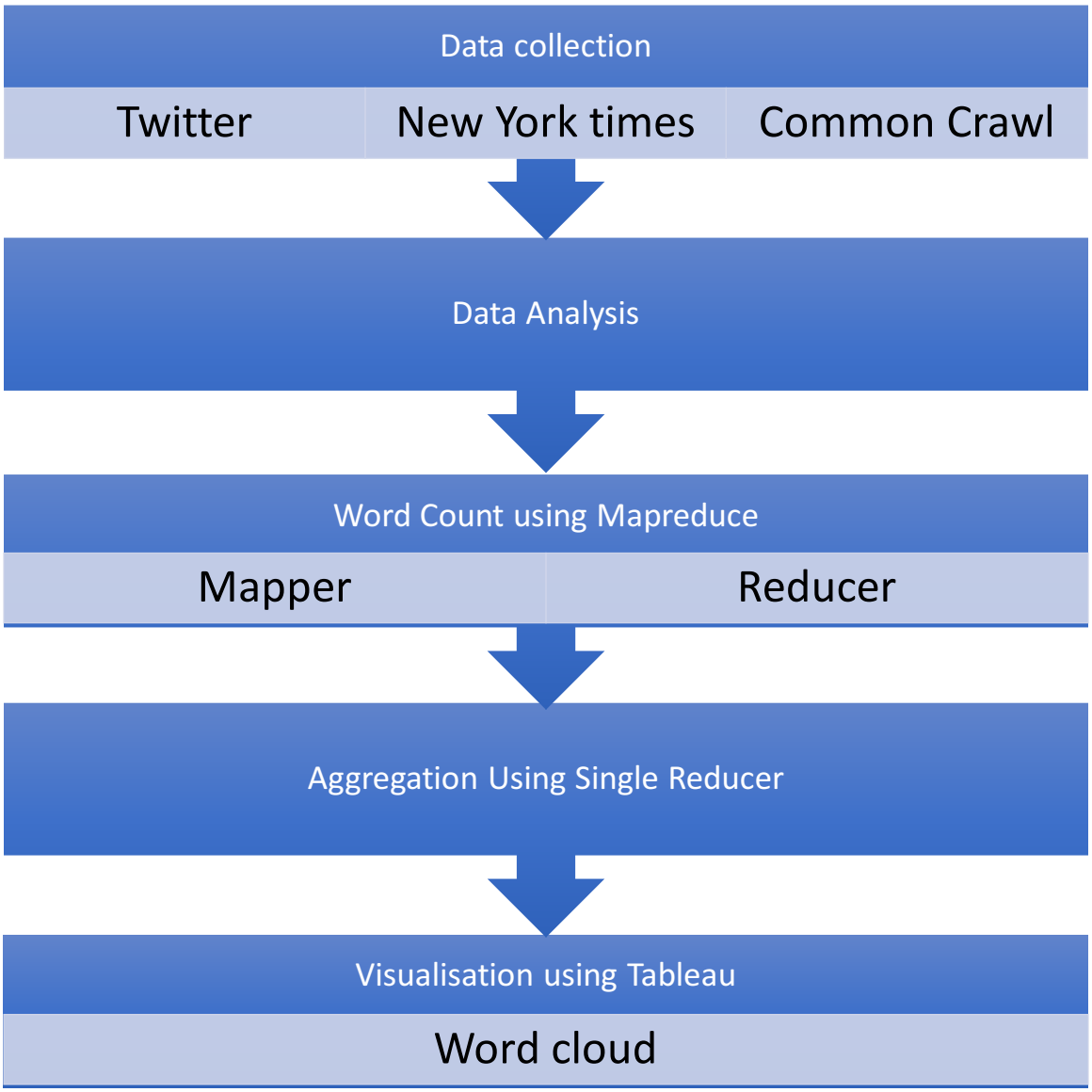
NIKHIL LALA (nlala@buffalo.edu)

MAYANK KULSHRESTHA (mkulshre@buffalo.edu)

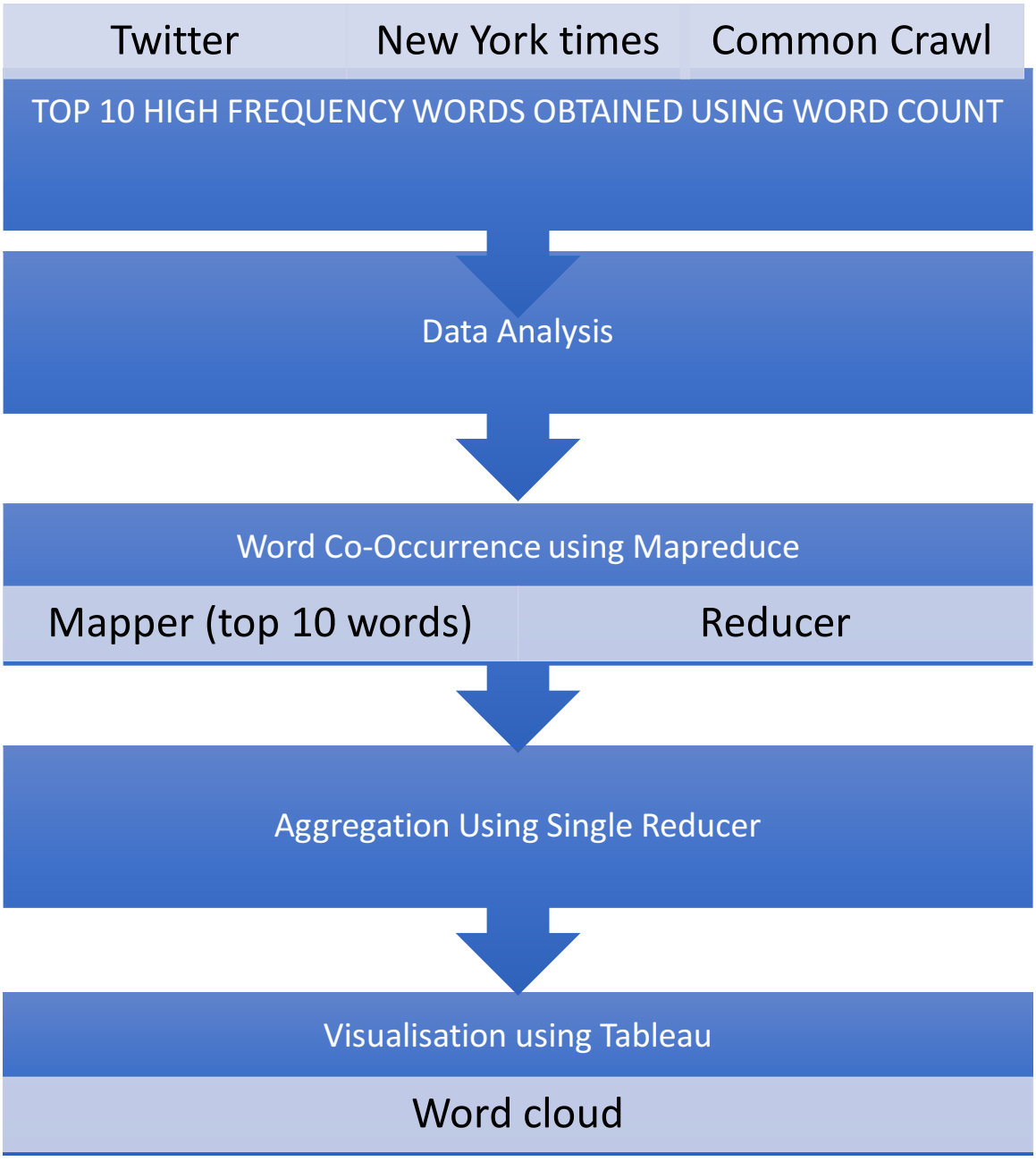
Website Hosted Link:

https://dictableau.000webhostapp.com/main%20copy.html

Word Count Pipeline



Word Co-occurrence Pipeline



UNDERLYING INFRASTRUCTURE

- ## DATA COLLECTION

  Python 2.7 & python 3.7.1

  Libraries and packages used:

  - Tweepy
  - Nytarticles
  - Json
  - Requests
  - Nltk (natural language processing Tool Kit) – Data cleaning

- ## MAP-REDUCE TASK (word count and co-occurrence)

  For the purpose of making use of Hadoop architecture, we have chosen AMAZON AWS cloud services for running Map-Reduce model.

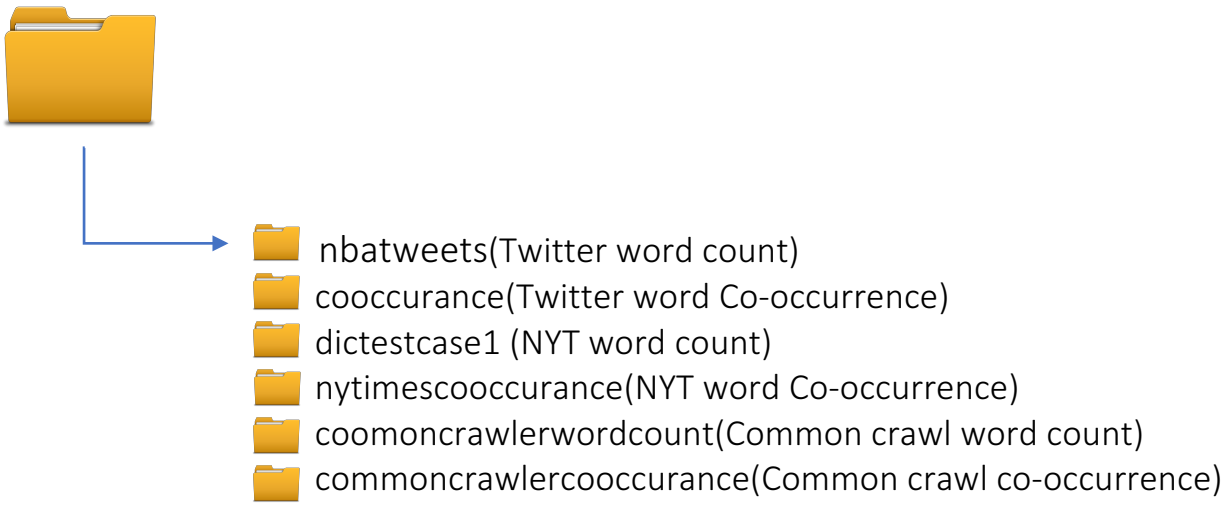
  To dive a bit deeper, we made use of Amazon EMR service. Amazon EMR is a managed cluster platform that simplifies running big data frameworks, such as Apache Hadoop and Apache Spark, on AWS to process and analyze vast amounts of data. By using these frameworks and related open-source projects, such as Apache Hive and Apache Pig, you can process data for analytics purposes and business intelligence workloads. Additionally, you can use Amazon EMR to transform and move large amounts of data into and out of other AWS data stores and databases, such as Amazon Simple Storage Service (Amazon S3).

AWS SERVICES USED

- S3 – Amazon S3 bucket is used to store huge amounts of data (probably Big data). All the data that we collected from 3 sources i.e. Twitter, New York Times and Common Crawl was stored in 3 individual S3 buckets. We also used common crawl S3 to retrieve URL list for further data collection.

- EMR – Amazon provides us with an interface that simplifies running big data frameworks, such as Apache Hadoop and Apache Spark, on AWS to process and analyze vast amounts of data. It provides the user with an option to provide mapper and reducer and other option to log and include other services.

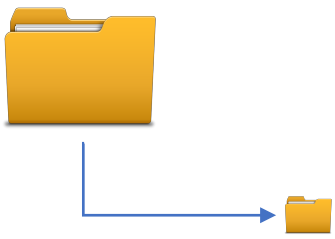
DIRECTORY STRUCTURE



Structure of each of the subdirectories

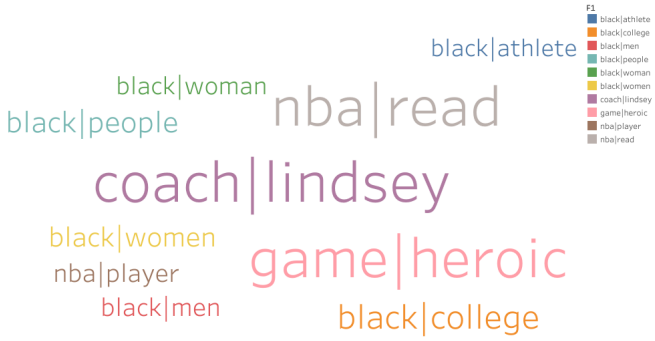
Each of the above-mentioned directories contain:

- 3 subdirectories
  - Input
  - Multiple output files
  - Single output file
- 2 files
  - mapper.py
  - reducer.py



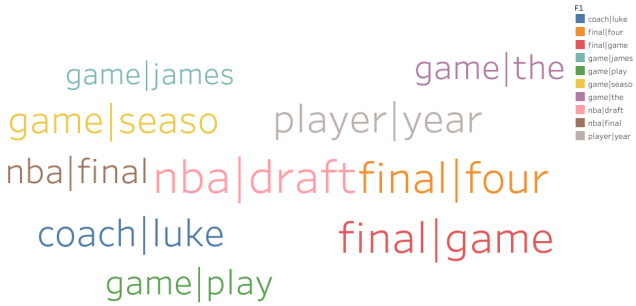
Input
Multiple output files
Single output file
Mapper.py
Reducer.py

CommonCrawlCoOccuracne



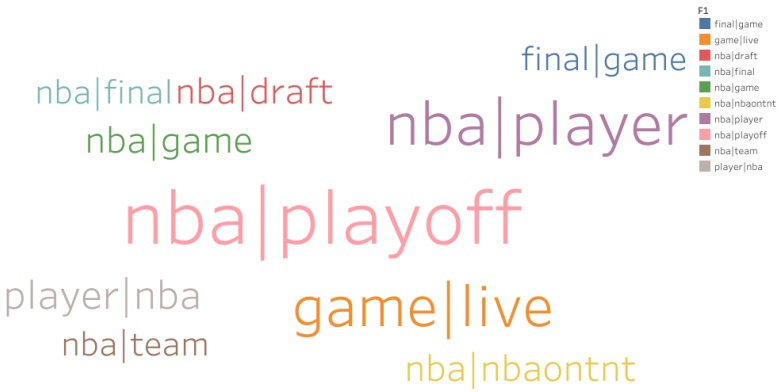
CommonCrawlWordCount

NyTimesCoOccurance



NyTimesWordCount

TwitterCoOcc



TwitterWordCount