CSE 587: Data Intensive Computing

Lab 2 - Report

Data Aggregation, Big Data Analysis And Visualization

Daniel Amirtharaj

Navaneethakrishnan Ramanathan

Objective

The objective of this project is to aggregate data from multiple sources – Twitter, New York Times and the Common Crawl, apply a big data analytic method such as MapReduce to analyse unstructured text data obtained from these sources, and build a data visualization product.

Introduction

Topic selected

The topic chosen for this project was – **Music**. Music is a subject that is talked about widely in all kinds of media, and was a good candidate for “Big Data”.

In order to collect as much data related to music as possible, several keywords/subtopics were used along with the word music.

The following subtopics, which are different genres of music, were used to aggregate data from all 3 data sources.

Subtopics **- Jazz, Pop, Classical, Rock, Rap and Country**.

Due to the limited news content related to music in the New York Times, additional genres, **Blues and Folk** were added, specifically to get data from this source.

Directory Structure Used

The root folder has the following folders:

1. **NYT** – *All data and code used to collect, analyse and visualise NYT data*.
2. **Data** – *All data and code used to collect the data*.
3. **Extracted NYT Articles** – *Distinct raw NYT articles*.
4. **Pre-Processed NYT Articles** – *Text files after stemming, lemmatization and removing stop words.*
5. **nyt\_client.**ipynb *– Code to collect NYT articles.*
6. **WCooccurence\_MR** – *Mapper and Reducer used for co-occurrence.*
7. **WCooccurence\_Output** – *Output after running co-occurence MR.*
8. **WCount\_MR** – *Mapper and Reducer used for co-occurrence.*
9. **WCount\_Output\_small** – *Output after running count MR on small dataset.*
10. **WCount\_Output\_big** – *Output after running count MR on big dataset.*
11. **Common\_Crawl** – *All data and code used to collect, analyse and visualise Common Crawl data*.
12. **Data** – *All data and code used to collect the data*.
    1. **Extracted Articles** – *Distinct raw Common Crawl articles*.
    2. **Pre-Processed Articles** – *Text files after stemming, lemmatization and removing stop words.*
    3. **Common Crawl Collector.**ipynb *– Code to collect Common Crawl articles.*
13. **WCooccurence\_MR** – *Mapper and Reducer used for co-occurrence.*
14. **WCooccurence\_Output** – *Output after running co-occurence MR.*
15. **WCount\_MR** – *Mapper and Reducer used for co-occurrence.*
16. **WCount\_Output\_small** – *Output after running count MR on small dataset.*
17. **WCount\_Output\_big** – *Output after running count MR on big dataset.*
18. **Tweets** – *All data and code used to collect, analyse and visualise Twitter data*.
19. **Data** – *All data and code used to collect the data*.
    1. **Extracted NYT Articles** – *Distinct raw tweets*.
    2. **Pre-Processed NYT Articles** – *Text files after stemming, lemmatization and removing stop words.*
    3. **Tweet Collector.**ipynb *– Code to collect tweets.*
20. **WCooccurence\_MR** – *Mapper and Reducer used for co-occurrence.*
21. **WCooccurence\_Output** – *Output after running co-occurence MR.*
22. **WCount\_MR** – *Mapper and Reducer used for co-occurrence.*
23. **WCount\_Output\_small** – *Output after running count MR on small dataset.*
24. **WCount\_Output\_big** – *Output after running count MR on big dataset.*

Data Collection

Data was collected from three sources, Twitter, New York Times and the Common Crawl dataset.

1. **Twitter**

The twitter API was used along with twython, a helper library specifically to aid tweet collection fot Python. Twitter was queried using the keywords - jazz, “pop music”, “classical music”, “rock music”, rap and “country music”. The word music was added while querying subtopics like pop, country etc. since irrelevant tweets which had the words that were not related to music were returned.