PROJECT REPORT

LAB2: DATA AGGREGATION, BIG DATA ANALYSIS AND VISUALIZATION

Divya Shree 50292944 dshree@buffalo.edu

Table of Contents

| INTRODUCTION | |
|--------------------|----------|
| TOPIC | |
| | |
| SETUP | 3 |
| DATA SOURCES | 4 |
| 1. Twitter | |
| 2. NYT Article | |
| 3. Common Crawl | |
| WORD COUNT | |
| 1) Small Data Set | 8 |
| a.) Twitter | 8 |
| b.) NYT | |
| c.) Common Crawl | <u>C</u> |
| 2) Large Data Set | 10 |
| a) Twitter | 10 |
| b) NYT | 10 |
| c) Common Crawl | 11 |
| Word Convergence | |
| a) Design Art | |
| b) Music | |
| c) Painting | 14 |
| d) Photography | 15 |
| e) Visual Arts | 16 |
| f) Writing | 17 |
| Word Co-occurrence | 18 |
| 1) Twitter | 18 |
| 2) NYT | 19 |
| 3) Common Crawl | 10 |

INTRODUCTION

The aim of the project is to work with Big Data and compare the outputs and trends via different data sources and visualizations.

This Lab concentrates on 4 defined tasks which are as below:

- 1. Data aggregation from more than one source using the APIs (Application programming interface) exposed by data sources
- 2. Applying classical big data analytic method of MapReduce to the unstructured data collected
- 3. Store the data collected on WORM infrastructure Hadoop
- 4. Building a visualization data product

TOPIC

The topic used to collect the data is: **ARTS**. It is one of the most common frequently used topics and it involves everyday life.

The subtopics that have been considered under it are:

- Painting
- Photography
- Music
- Visual Arts
- Design Art
- Writing

SETUP

Docker setup has been used to run Map Reduce.

In the provided docker image, following steps were executed/done to enable the different functionalities:

- 1) Upgraded python to 2.7 required for Stanford nltk for data cleaning
- 2) Downloaded all the required packages:
 - a) nltk
 - b) warc
 - c) emoji
 - d) bs4

3) New Docker image was created with above packages installed to ensure that these packages don't have to be installed on other containers that will be created later.

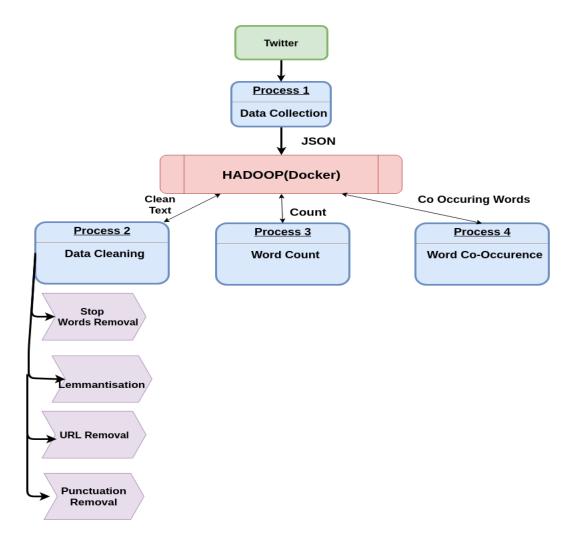
DATA SOURCES

Three data-sources were used:

1. Twitter

Twitter is one of the prominent social websites that provides a lot of data to parse and visualize. Twitter Stream API was used to extract the data corresponding to above mentioned subtopics.

The extracted data is in json format. The text from each tweet is extracted and parsed. All the cleaning of Tweets is done in Hadoop by using Map Reduce.

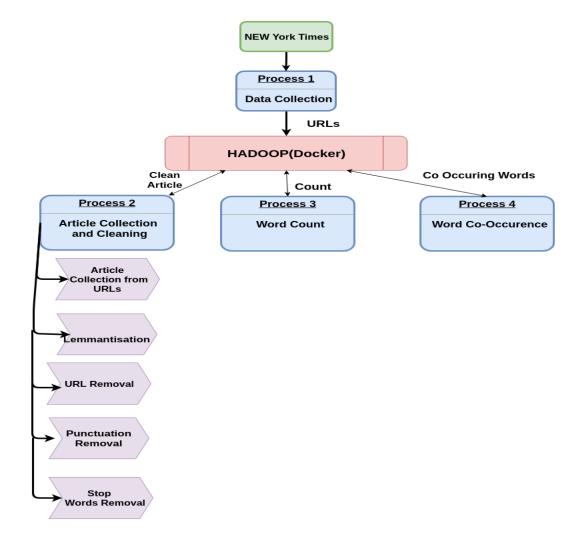


Data Size

| Sub Topics | Tweets Count |
|-------------|--------------|
| Design Art | 4292 |
| Music | 13285 |
| Painting | 6829 |
| Photography | 6240 |
| Visual Art | 8353 |
| Writing | 5662 |
| Total | 44661 |

2. NYT Article

NYT Article search was done through Article Search API provided by NY Times. The response of the API provides URLs of the actual article, hence to get the full article we have to call each URL and then get the actual article. Here, Article extraction and data cleaning (remove URLs, punctuation, lowercase, stop words and lemmatization) was done on Hadoop using MapReduce.



| Sub Topics | Articles Count |
|-------------|----------------|
| Design Art | 923 |
| Music | 1393 |
| Painting | 176 |
| Photography | 190 |
| Visual Art | 631 |
| Writing | 455 |

3. Common Crawl

Common Crawl is a functionality that accesses everyday crawls and provides it free for use to everyone. It returns index search API that provides the related WARC/WET/WAT files. The warc file is taken from the index search response and the data is retrieved using the warc file. The warc file is downloaded and records are read.

Different domains were used for different subtopics to get the appropriate data. The domains used were:

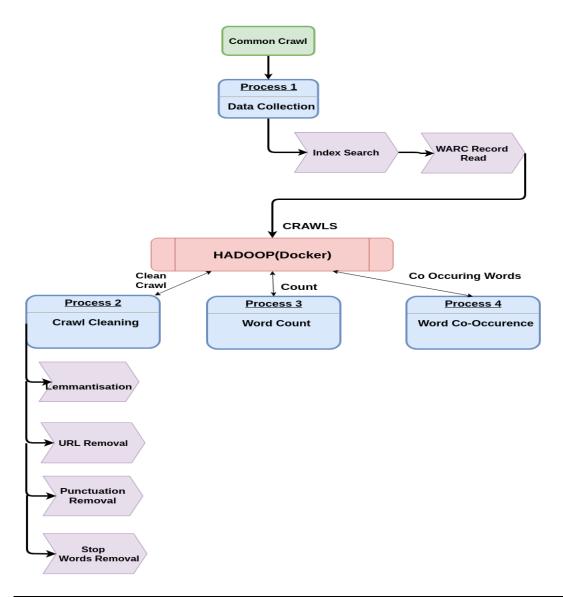
a) Design Art: archDaily.com

b) Music: billboard.comc) Painting: theartnewspaper.com

d) Photography: photographylife.com, art-is-fun.com

e) Visual Art: screenrant.com

f) Writing: lithub.com



| Sub Topics | Crawls Count |
|-------------|--------------|
| Design Art | 500 |
| Music | 500 |
| Painting | 500 |
| Photography | 500 |
| Visual Art | 500 |
| Writing | 500 |
| Total | 3000 |

^{*}Since the Crawls are big, code was written to manually exit after getting 500 crawls.

WORD COUNT

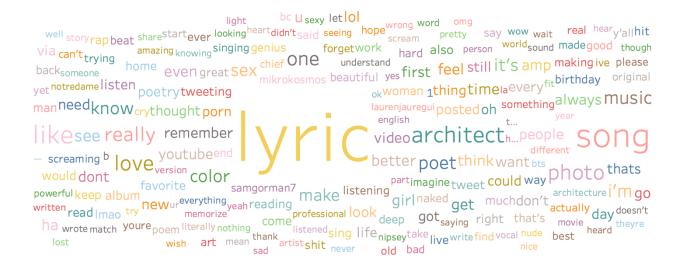
The word count was done on following data sets:

- Small Dataset
- Large Dataset

1) Small Data Set

The data from 3 sources was collected for a few days. The Word Cloud generated for top 200 words using Tableau are shown below:

a.) Twitter



```
kind version play night name war big among piece day career right great don't last great don't last next series u building sometimes different also best womannever people something put york film man much came character part would far One like moviemuseum director years ald make work new feature going want feature going want figure group latertold studio award role criticold every month power end group attention big among doesn't find star writer feelalmost since "ago doesn't find get oscar place images ayaudience wrote president life picturethink book whose know around it's collection doe know around it's collection doe know around it's collection doe wrote president life picturethink book whose become know around it's collection doe wrote president life picturethink book whose become know around it's collection doe wrote president life picturethink book whose become know around it's collection doe wrote president life picturethink book whose become tilfe picturethink book whose know around it's collection doe wrote president life picturethink book whose become tilfe picturethink book whose know around it's collection doe wrote president life picturethink book whose become tilfe picturethink book whose president life picturethink book whose wrote president life picturethink book whose the president life picturethink book whose wrote president life picturethink book whose the president life picturethink book whose wrote president life picturethink book whose the president life picturethink book whose wrote president life picturethink book whose the president life picturethink book whose wrote president life picturethink book wrote president life picturethink book wrote president life picturethink life
```

c.) Common Crawl

```
think course really novel based question know scene part mightyoull always using canvas last right type ideaday large give email well come within something used feel USe thaneeya medium picture year architecture without even though shape much would different design many spacetext water color favorite room image colored form paint still work life drawing site movie music Way time it's pencil film tv create coloring painting essential bill board to business on the life water architecture without even though shape much would different design many spacetext water color favorite room image colored form paint water color image colored form paint still work life drawing site movie music Way time it's pencil film tv create coloring painting essential bill board to business on the life water all around page story functions architect provided thing technique thing technique city support side the process main natural point along often level wall step pen
```

2) Large Data Set

The data from 3 sources was collected for 4 months. The Word Cloud generated for top 200 words are shown below:

a) Twitter

```
give youre freecalled ive well since via stop longgirl wait said real old boymarvel check literally wanna anyone "you dont even—that's playing also lyric listen see jonahmarais thought madisonbeer it's favourite year fucking amp day much heardsaw come right thank sunday getting let hear actually play heardsaw come right thank sunday getting let hear actually play heardsaw come right thank sunday getting let hear actually play heardsaw come right thank sunday getting let hear actually play heardsaw come right thank sunday getting let hear actually play heardsaw come right thank sunday getting let hear actually play heardsaw come right thank sunday getting let hear actually play heardsaw come right thank sunday getting let hear actually play heardsaw come right thank sunday getting let hear actually play heardsaw come right thank sunday getting let hear actually play actually play actually play actually play and getting let hear actually play heardsaw come right thank sunday getting let hear actually play heardsaw come right thank sunday getting let hear actually play heardsaw come right thank sunday getting let hear actually play heardsaw come right thank sunday getting let hear actually play heardsaw come right thank sunday getting let hear actually play heardsaw come right thank sunday getting let hear actually play intellectual people think little made startask madison hard art look make; ma
```

b) NYT

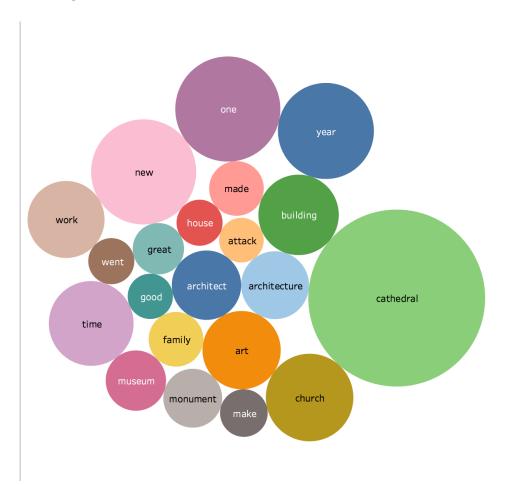
c) Common Crawl

```
said next art digital start star doe keep openman setting often love sure subject little thought shooting might season quality woman right however review book using movie photo better site take without set two really sensor house feelcharacter let like yearkind make also know space need landscape thing people get could one film image life building photographer want time new has article save question 2 architect big great used point experience bit back well never any series system world day say going aperture room shot canon filtergot shoot wall give word lidea form loss first find create black last old able
```

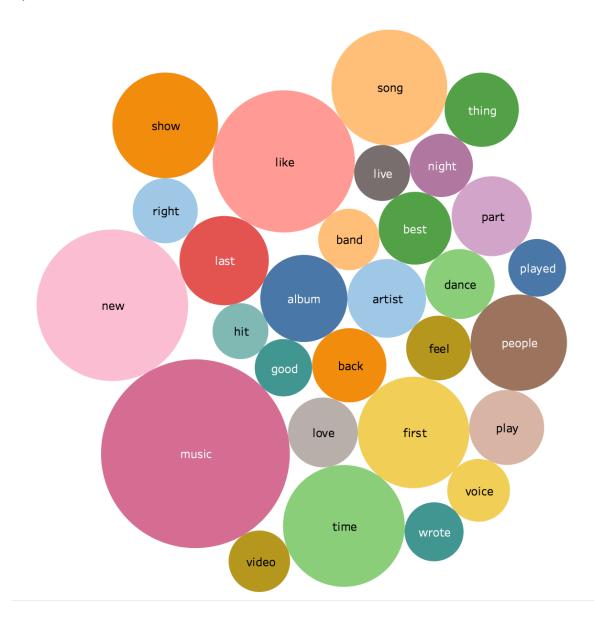
Word Convergence

Word convergence was generated on Large dataset for each subtopic. Tableau was used to generate the visualization as depicted below.

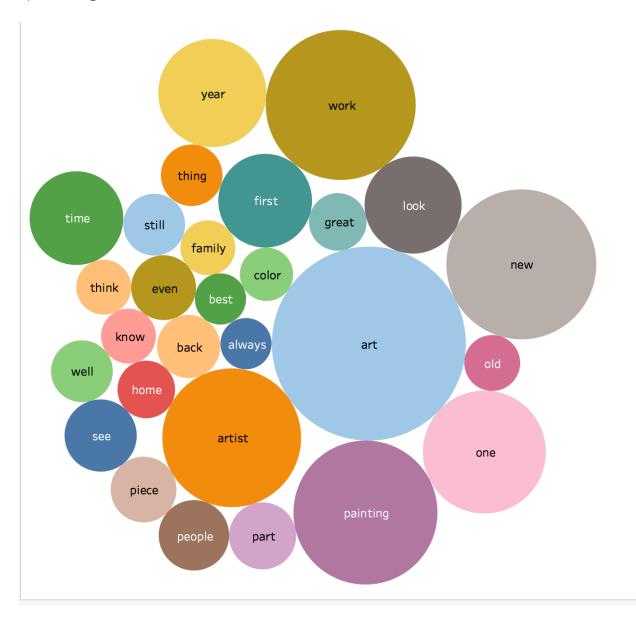
a) Design Art



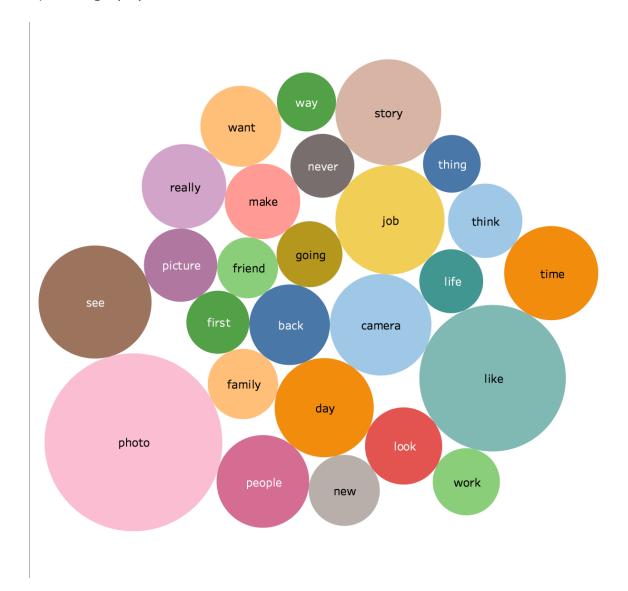
b) Music



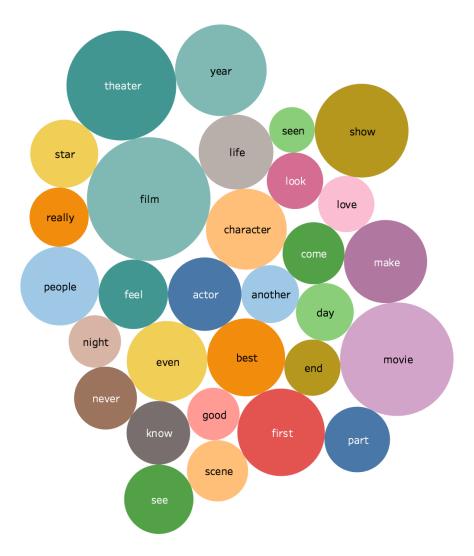
c) Painting



d) Photography



e) Visual Arts



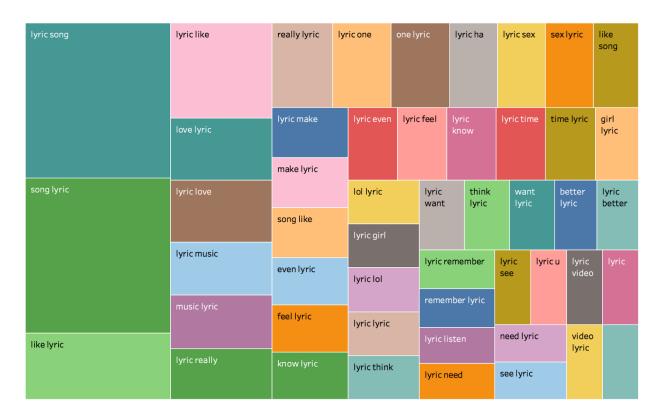
f) Writing



Word Co-occurrence

Word co-occurrence was generated on Small Dataset for all 3 data-sources. Tableau was used to generate the visualization as depicted below.

1) Twitter



2) NYT

| art museum | art wa | wa also | film wa | wa fil | lm ; | art show | show art | | | new | |
|------------|---------|------------|-------------|-----------|-------------|-------------|--------------------------|---------|-------------|-------------|---|
| museum art | wa art | mrsaid | | | | | | | | | |
| | | said mr | film film | | new said | said new | art | | | ha one | е |
| wa year | пеw пеw | | new work | | | | | | | | |
| year wa | | art work | work new | | one ha | | halike | like ha | | one one | |
| | art new | work art | work work | | museun | n work | | | | | |
| wa work | new art | | WO IN WO IN | | work museum | | like like vork museum | | new york | yorl new | |
| | | art artist | one said | like said | | i | art exhibit | bition | | | |
| work wa | also wa | artist art | said one | | said like | e | exhibition | art | year | said | |

3) Common Crawl

