PROJECT REPORT

LAB2: DATA AGGREGATION, BIG DATA ANALYSIS AND VISUALIZATION

Divya Shree
50292944
dshree@buffalo.edu

Table of Contents

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.

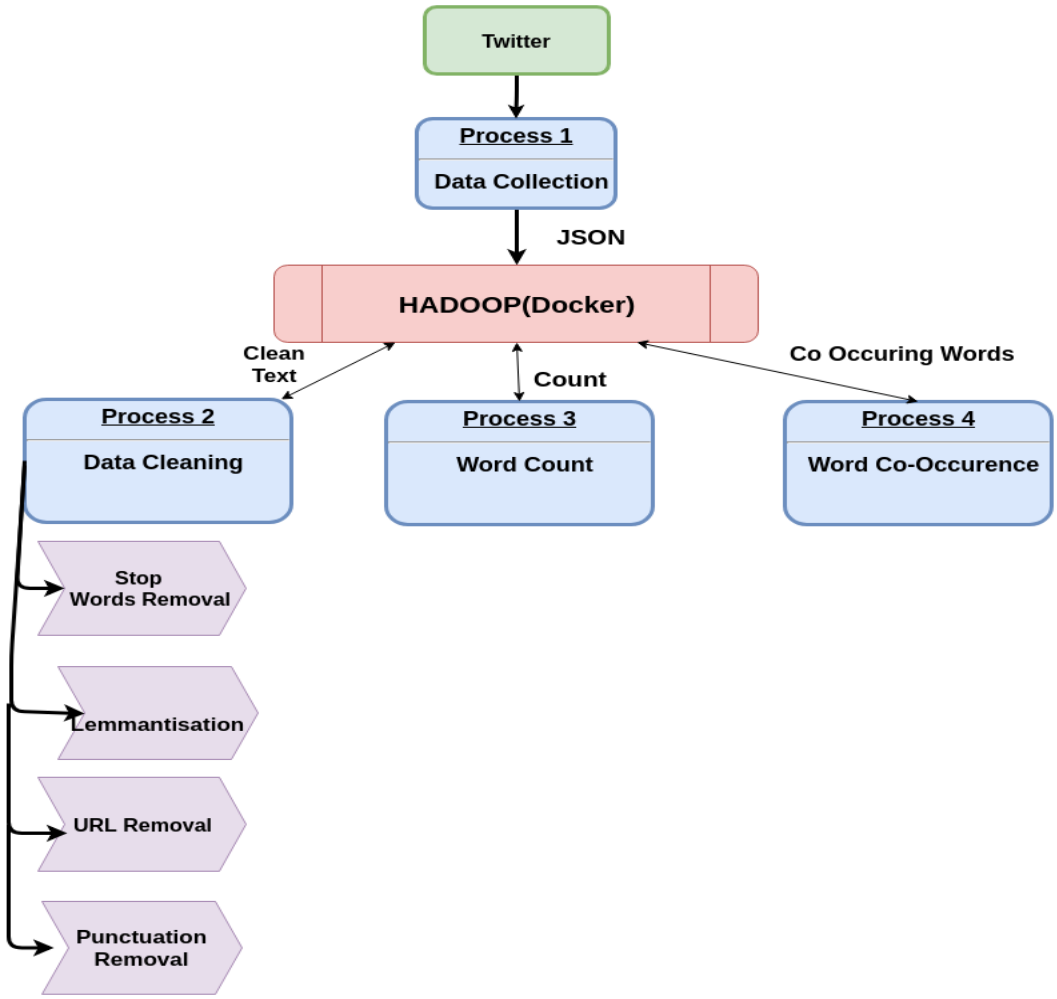
DATA SOURCES

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.

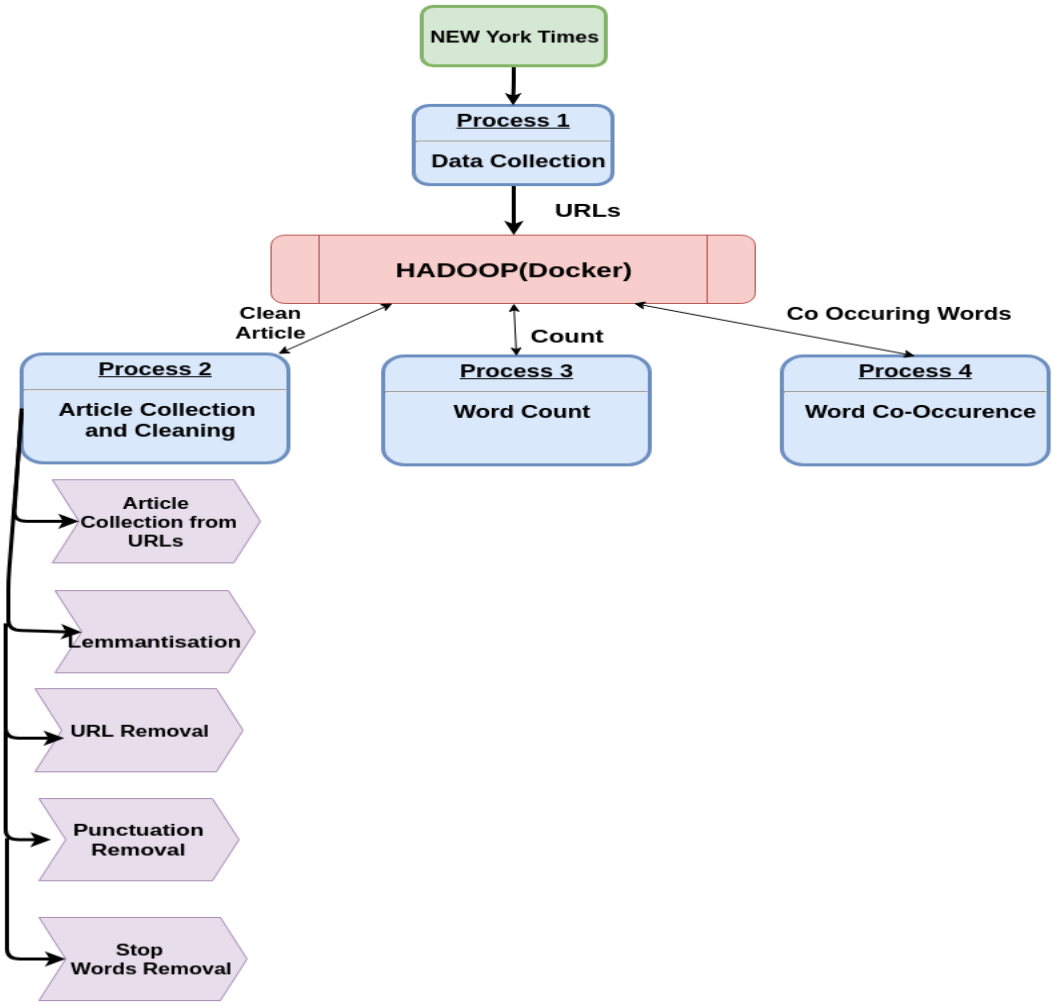


Data Size

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.



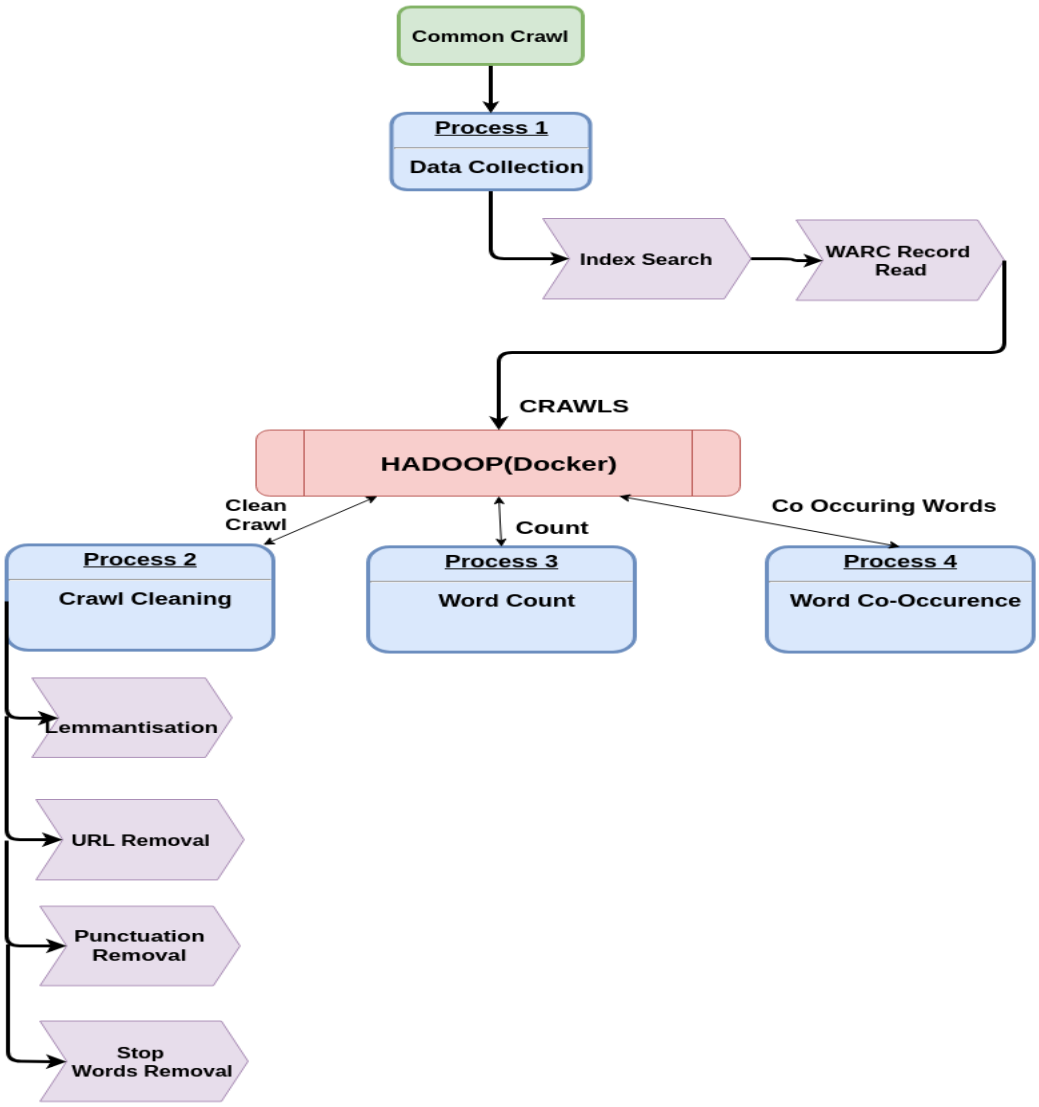
Sub Topics	Articles Count
Design Art	923
Music	1393
Painting	176
Photography	190
Visual Art	631
Writing	455

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.com**
c) **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

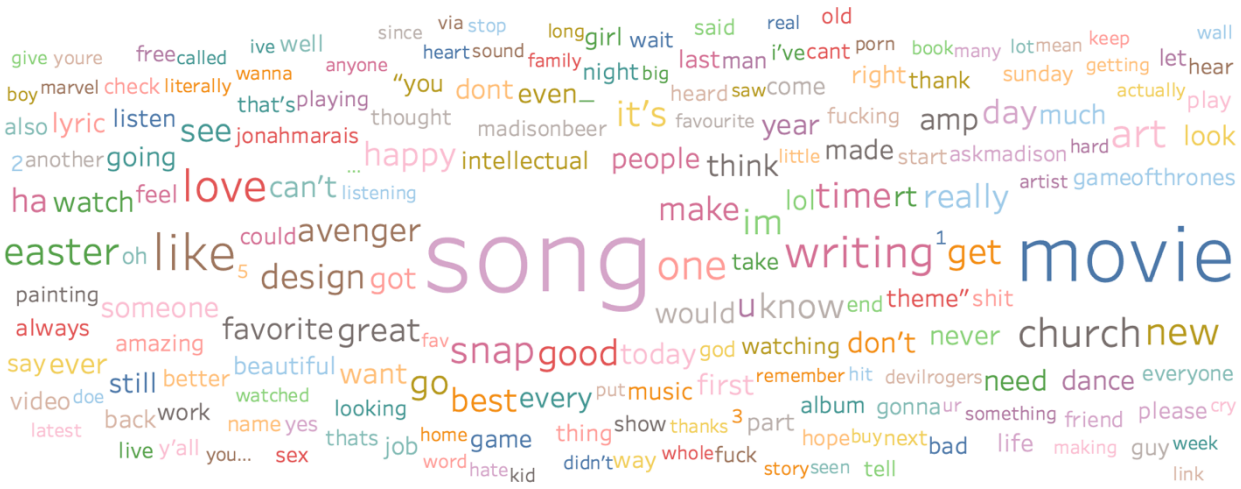
b.) NYT

c.) Common Crawl

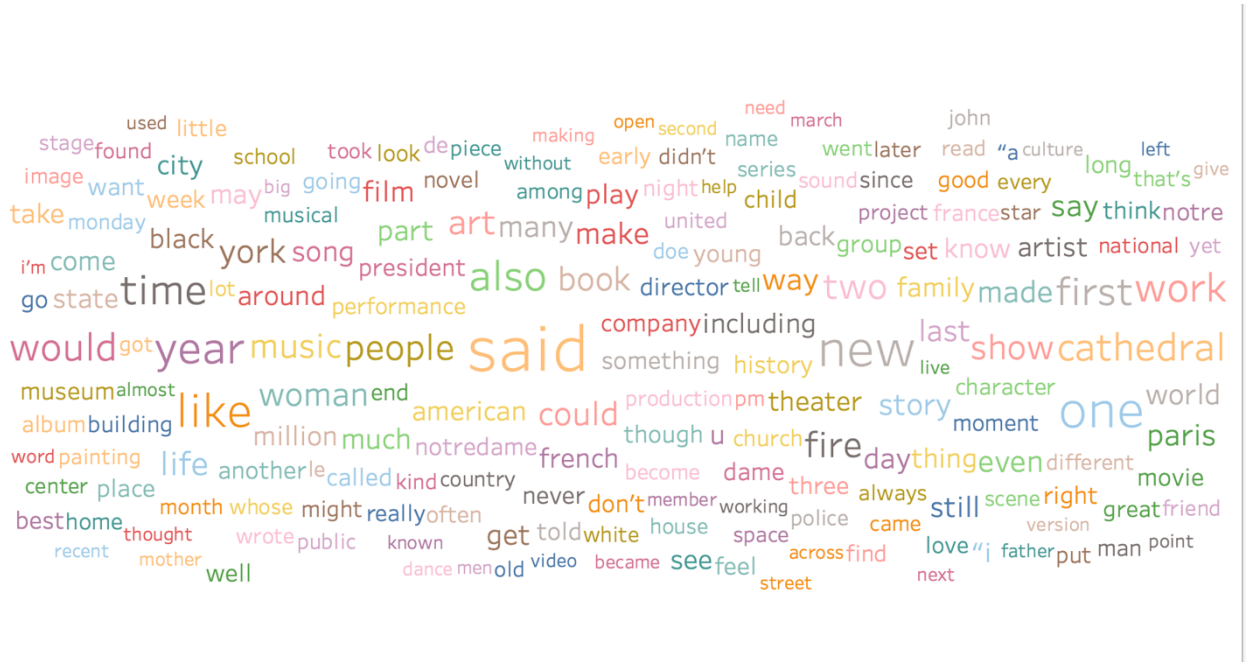
2) Large Data Set

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

a) Twitter



b) NYT

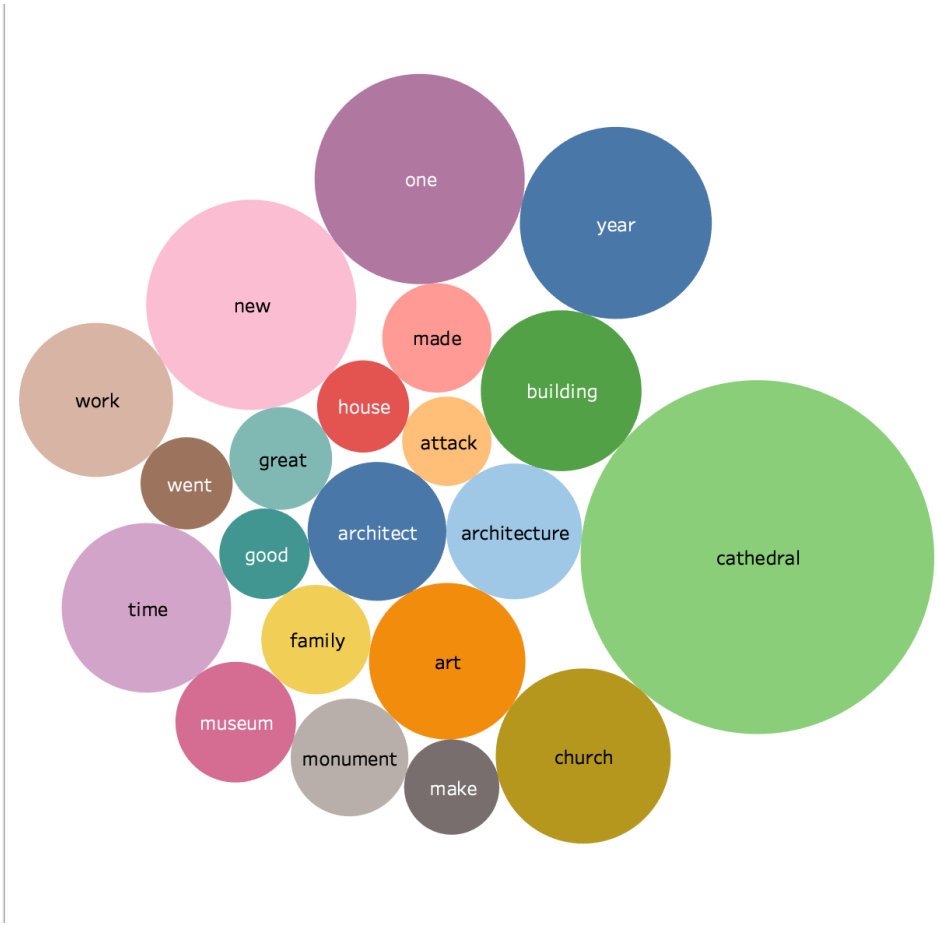


c) Common Crawl

Word Convergence

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

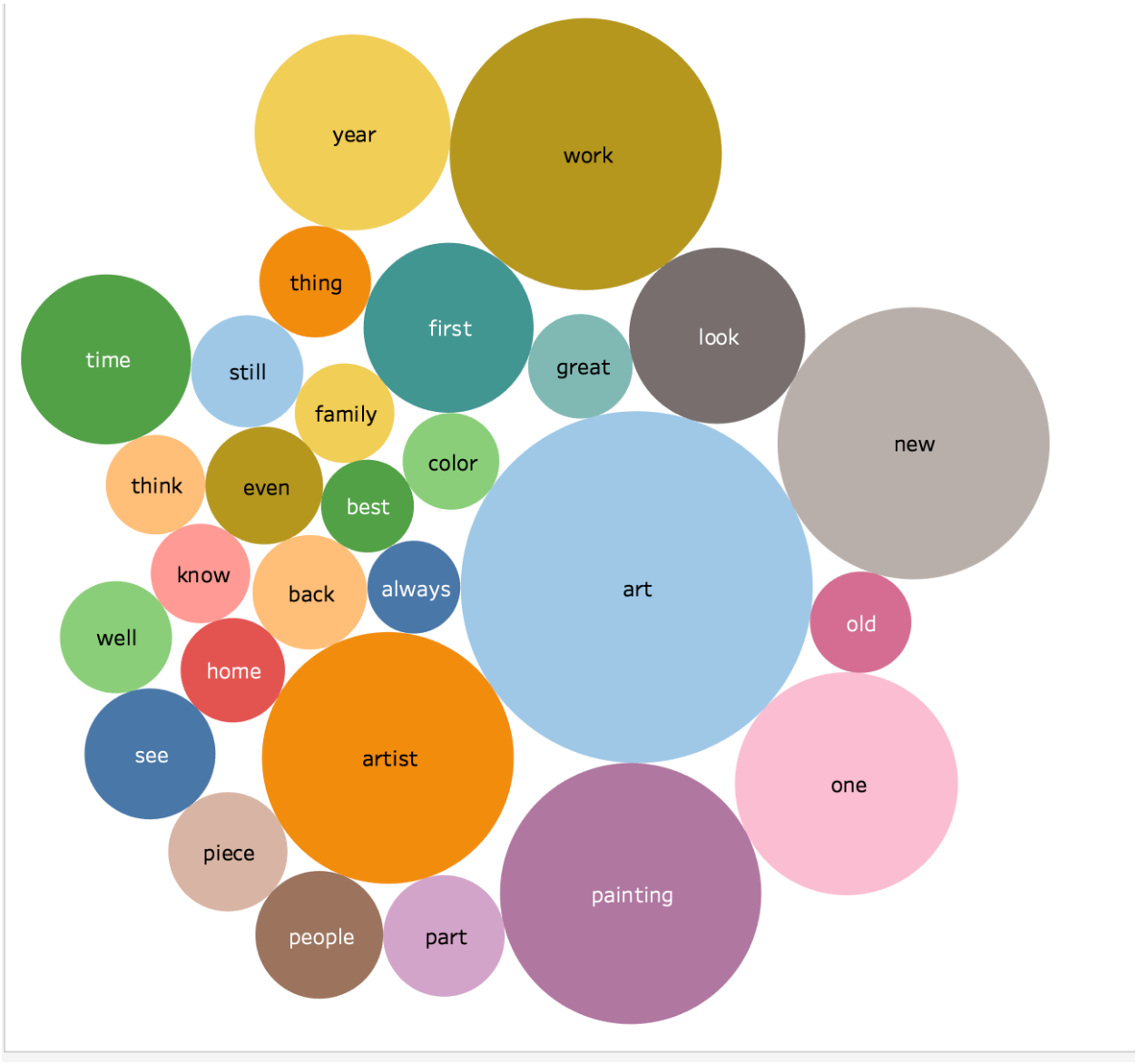
a) Design Art



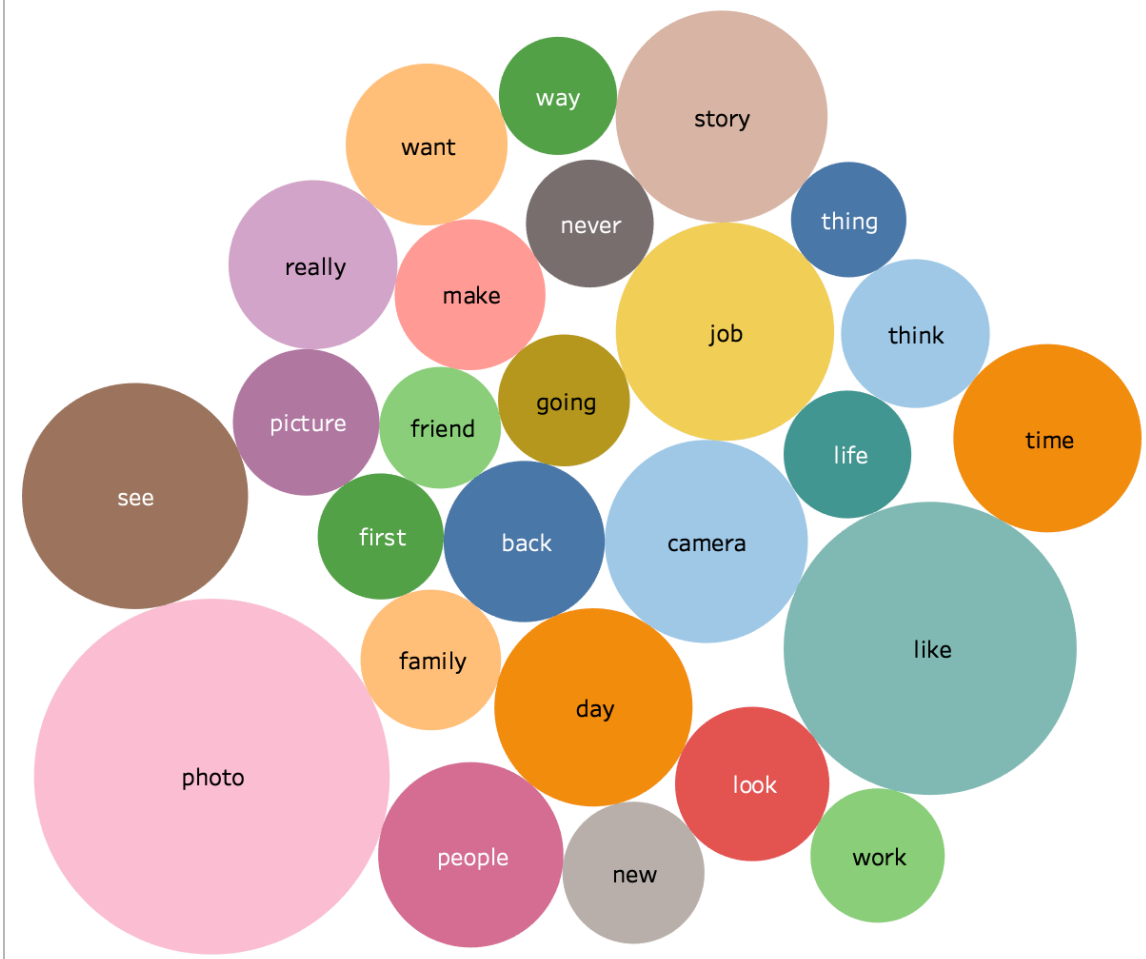
b) Music


song
thing
show
like
live
night
right
part
best
band
last
played
new
album
artist
dance
hit
people
feel
good
back
music
love
first
play
voice
time
wrote
video

c) Painting

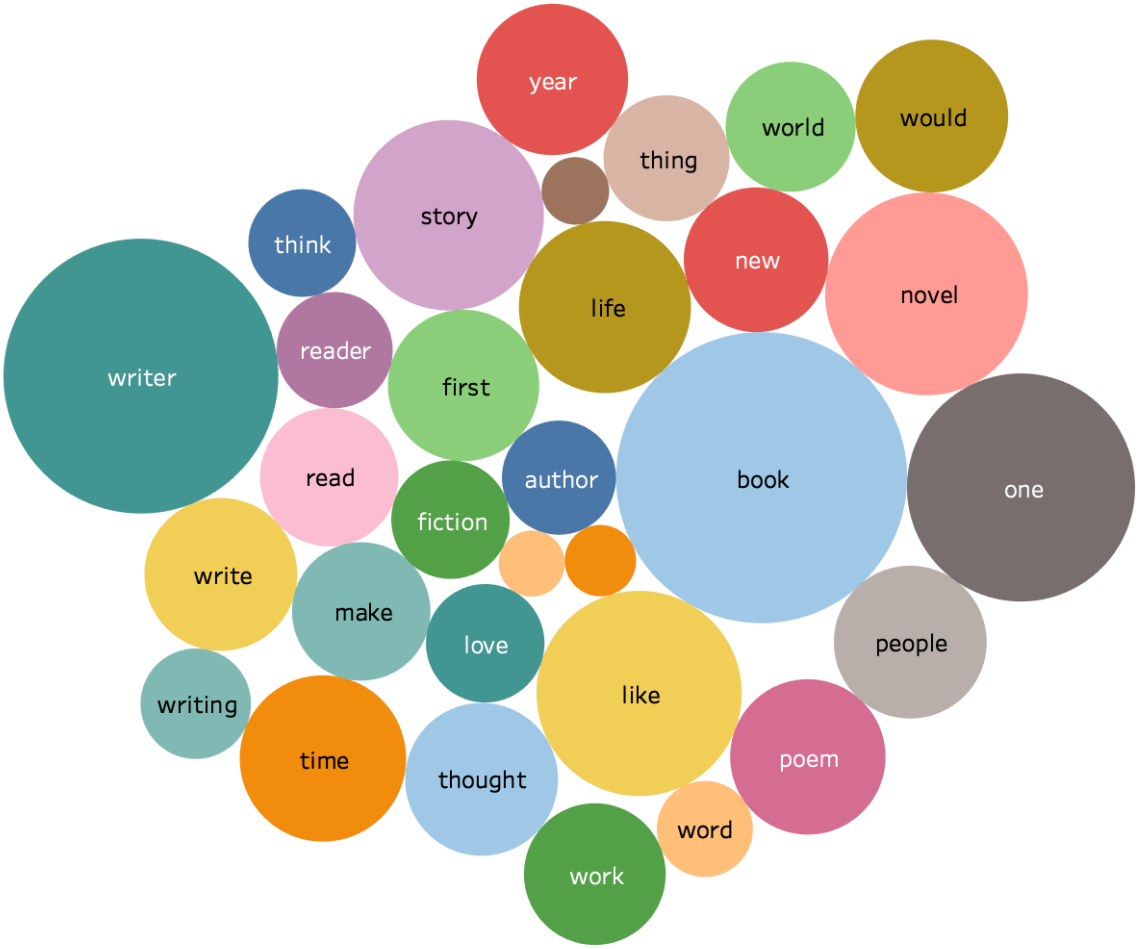

year
work
thing
first
look
time
still
great
new
family
color
think
even
best
know
back
always
art
old
well
home
see
artist
one
piece
painting
people
part

d) Photography



e) Visual Arts

f) Writing


year
world
would
thing
story
think
new
life
novel
writer
reader
first
read
author
book
one
fiction
write
make
love
people
like
writing
time
thought
poem
word
work

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

3) Common Crawl