

Minor project 2021 Report

On

Cluster Analysis of Twitter Data

Submitted in partial fulfilment of the requirements for the award of the degree

of

Bachelor of Technology

in

Computer Science & Engineering

by

Partha Pratim Deori

(180101032)

Rahul Ranjan

(180101036)

Yashvant Singh Yadav

(180101049)

Under the esteemed Supervision

of

Dr. Rupam Bhattacharya

(Faculty, Department of Computer Science and Technology, IIIT Bhagalpur)



Department of Computer Science and Engineering

Indian Institute of Information Technology Bhagalpur


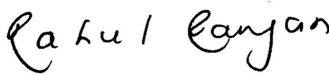

October 2021



भारतीय सूचना प्रौद्योगिकी संस्थान भागलपुर
INDIAN INSTITUTE OF INFORMATION TECHNOLOGY BHAGALPUR
An Institute of National Importance Under Act of Parliament

DECLARATION

We hereby declare that the work reported in this project on the topic “*Cluster Analysis of Twitter Data*” is original and has been carried out by us independently in the Department of Computer Science and Engineering, IIIT Bhagalpur under the supervision of **Dr. Rupam Bhattacharyya**, Assistant professor, IIIT Bhagalpur. We also declare that this work has not formed the basis for the award of any other Degree, Diploma, or similar title of any university or institution.

		
Partha Pratim Deori (<u>180101033</u>)	Rahul Ranjan (<u>180101036</u>)	Yashvant Singh Yadav (<u>180101049</u>)



भारतीय सूचना प्रौद्योगिकी संस्थान भागलपुर
INDIAN INSTITUTE OF INFORMATION TECHNOLOGY BHAGALPUR
An Institute of National Importance Under Act of Parliament

CERTIFICATE

This is to certify that the project entitled “*Cluster Analysis of Twitter Data*” is carried out by

Partha Pratim Deori

(180101033)

Rahul Ranjan

(180101036)

Yashvant Singh Yadav

(180101049)

B. Tech. students of IIT Bhagalpur, under my supervision and guidance. This project has been submitted in partial fulfillment for the award of “*Bachelor of Technology*” degree in *Computer Science and Engineering* at *Indian Institute of Information Technology Bhagalpur*. No part of this project has been submitted for the award of any previous degree to the best of my knowledge.

(Supervisor)

Dr. Rupam Bhattacharya

Assistant Professor

Dept. of Computer Science and Engineering

Indian Institute of Information Technology Bhagalpur

Abstract

Nowadays, people from all around the world use social media sites to share information. Twitter for example is a platform in which users send, read posts known as 'tweets' and interact with different communities. Users share their daily lives, post their opinions on everything such as brands and places. Companies can benefit from this massive platform by collecting data related to opinions on them. The aim of this paper is to present a model that can perform analysis of real data collected from Twitter. Data in Twitter is highly unstructured which makes it difficult to analyze. However, our proposed model is different from prior work in this field because it combined the use of unsupervised machine learning algorithms. The process of performing sentiment analysis as follows:

Tweet extracted directly from Twitter API, then cleaning and discovery of data performed. After that the data were fed into several models for the purpose of training.

Each tweet extracted classified into four categories that is “social”, “health”, “culture”, “economic”. Data was collected from my twitter Account and it is a developer account and Different machine learning algorithms were used. The results from these models were tested using various tests by our self.

Index

Abstract	i
Index	ii
Learning Objectives/Internship Objectives	iii
List of Figures	iv
List of Tables	v
Chapter 1: Introduction	
1.1 Overview	4
1.2 What are Tweets ?	5
1.3 What is a Clustering algorithm ?	
1.4 What is the K-Mean algorithm ?	6
1.5 Project Objective	
Chapter 2: Project Prerequisite	9
2.1 Twitter Developer Account	9
Chapter 3: Project Initialization	14
3.1 Import Python Libraries	14
3.2 Tweets Datasets	14
Chapter 4: Preprocessing of Tweets	16
4.1 Cleaning tweets	16
4.2 Tokenization, Lemmatization and removing stopwords	17
Chapter 5: Tweets Classification	18
5.1 What is Cosine Similarity ?	18
5.2 What is Jaccard Similarity ?	
Chapter 6: Creating Sets of Related Words	19
6.1 Economy Related Words	19
6.2 Social Related Words	19
6.3 Health Related Words	20
6.4 Culture Related Words	21
Chapter 7: Applying clustering algorithm	24
7.1 Clustered Data Frame	24
7.2 KMeans Clustering.	25
7.3 Conclusion	27
References	28

Learning Objectives/Internship Objectives

- Internships are generally thought of to be reserved for college students looking to gain experience in a particular field. However, a wide array of people can benefit from Training Internships in order to receive real world experience and develop their skills.
- An objective for this position should emphasize the skills you already possess in the area and your interest in learning more.
- Internships are utilized in a number of different career fields, including architecture, engineering, healthcare, economics, advertising and many more.
- Some internships are used to allow individuals to perform scientific research while others are specifically designed to allow people to gain first-hand experience working.
- Utilizing internships is a great way to build your resume and develop skills that can be emphasized in your resume for future jobs. When you are applying for a Training Internship, make sure to highlight any special skills or talents that can make you stand apart from the rest of the applicants so that you have an improved chance of landing the position.

List of Figures

Figure 1: Twitter Developer Account.....	9
Figure 2: Relevant Library.....	14
Figure 3: Twitter Data.....	15
Figure 4: Cleaning Tweets.....	16
Figure 5: Cleaned DataSet.....	16
Figure 6: Finding Jaccard Similarities Score.....	23
Figure 7: Tweets under different category.....	25
Figure 8: Cluster of Tweets in different groups.....	27
Figure 9: Confusion Metrics for CNN Gender Classifier.....	19
Figure 10 : Examples of Age Prediction.....	20
Figure11: Example of Gender Prediction.....	20
Figure 12: Examples of some Mis-Classification.....	21
Figure 13: UI of the inference engine.....	21
Figure 14: Prediction demo of inference engine.....	22

List of Tables

Table 1: Gender Distribution.....	9
Table 2: Summary of Models.....	19

Chapter 1: Introduction

1.1 Overview

Twitter is a social networking and micro blogging service on which users post and interact with each other through messages known as “tweets”. It’s ranked as the 6th most popular social networking site and app by Dream Grow as of April, 2020 with an average of 330 million active monthly users.

Unlike other platforms like Facebook whose main role is to play ‘catch-up’ with friends, it is where people let loose and engage with different personalities from all walks of life on all sorts of matters. This atmosphere is what makes it the ideal platform for marketers, politicians and other titles whose success depend on a deep understanding of people’s views.

Through sentiment analysis, interested parties can understand what users are talking about and from the insights, make the appropriate decisions. This post focuses on classifying tweets into 4 major categories: *Economic, Social, Cultural and Health* then performing KMeans cluster analysis on the groups.

1.2 What are Tweets ?

Twitter is a social networking and microblogging service on which users post and interact with each other through messages known as “tweets”.

a Tweet is a message sent on Twitter. To send or receive a Tweet, you have to create a free account with Twitter. You also need to have friends and contacts with Twitter accounts -- otherwise you're typing to the void. Of course, you could use Twitter as a blog and keep all of your Tweets public, meaning anyone could read them on your personal Twitter profile page. But if you want to use Twitter as a way to keep in touch with friends, you'll need to convince them to sign up, too.

1.3 What is a Clustering Algorithm ?

Clustering algorithms take the data and using some sort of similarity metrics, they form these groups – later these groups can be used in various business processes like information retrieval, pattern recognition, image processing, data compression, bioinformatics etc. In the Machine Learning process for Clustering, as mentioned above, a distance-based similarity metric plays a pivotal role in deciding the clustering.

Types of Clustering Algorithms

- Connectivity-based Clustering (Hierarchical clustering)
- Centroids-based Clustering (Partitioning methods)
- Distribution-based Clustering
- Density-based Clustering (Model-based methods)
- Fuzzy Clustering
- Constraint-based (Supervised Clustering)

1.4 What is the K-Mean Algorithm ?

“ It is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs to only one group that has similar properties. ”

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of predefined clusters that need to be created in the process, as if $K=2$, there will be two clusters, and for $K=3$, there will be three clusters, and so on.

It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.

Chapter 2: Project Prerequisite

2.1 Twitter Developer Account

The Twitter developer portal contains a set of self-serve tools that developers can use to manage their access to the Twitter API and Twitter Ads API.

In the portal, you have the opportunity to:

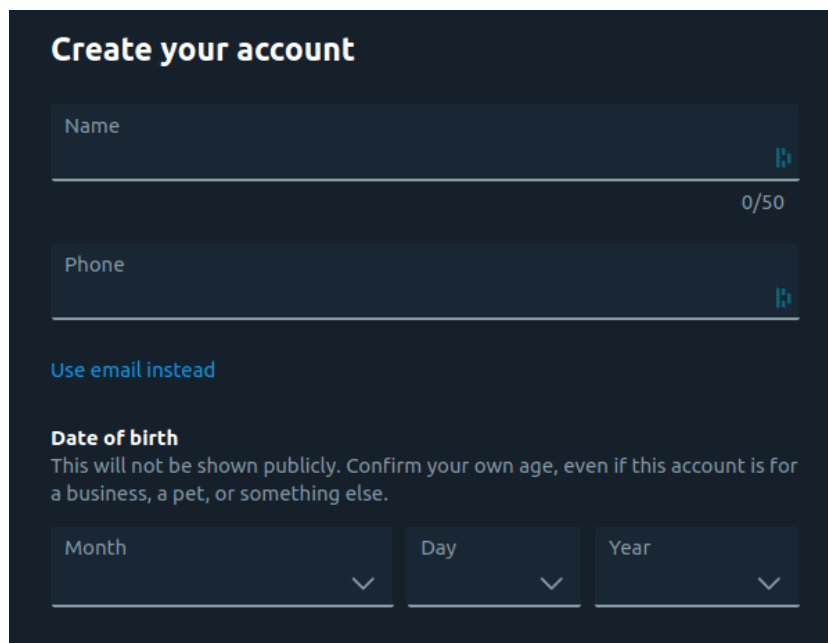
- Create and manage your Twitter Projects and Apps (and the authentication keys and tokens that they provide).
- Manage your access levels and integrations with the Twitter API premium v1.1 and v2 endpoints.
- Learn more about different endpoints and features available.
- If you have Elevated or Academic Research access, you can view team pages where you can add and manage the different handles that have access to your team's account.

How to apply for a Twitter Developer account :-

Create an account on Twitter

Go to [Twitter](#) and log in or sign up for an account which you would use for development.

It is recommended to create a *separate account* for your twitter application. Especially if you are applying for enhancing your business, for creating a bot or a web-hosted app.



The screenshot shows the 'Create your account' form on a dark background. It includes a 'Name' field with a character count of 0/50, a 'Phone' field, a link to 'Use email instead', and a 'Date of birth' section with three dropdown menus for Month, Day, and Year. A disclaimer states that the birth date will not be shown publicly and is for age verification.

Create your account

Name 0/50

Phone

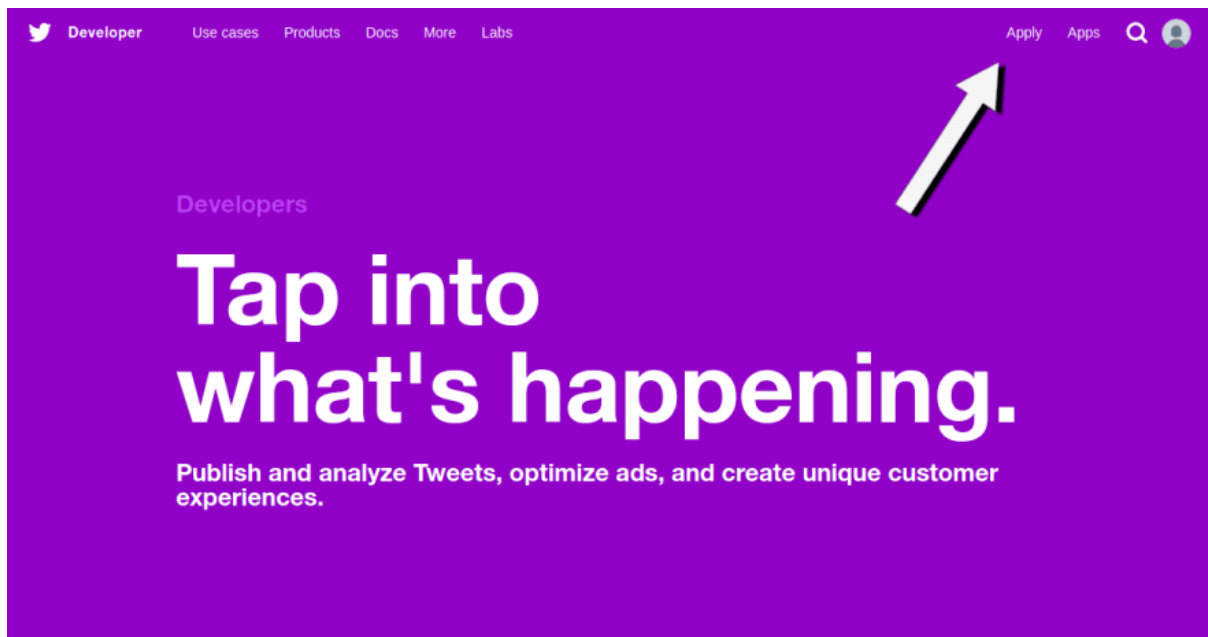
[Use email instead](#)

Date of birth
This will not be shown publicly. Confirm your own age, even if this account is for a business, a pet, or something else.

Month Day Year

Apply for access

- Go to developers.twitter.com and click on Apply



- On the next page, click on [Apply for a Developer account](#)

Get started with Twitter APIs and tools

Apply for access

All new developers must apply for a developer account to access Twitter APIs.

[Apply for a developer account](#)

[Restricted use cases >](#)

Specify the type of application

In the next step, you will have to enter the purpose of your twitter developer account

The first category is Professional which is generally used by business teams to enhance their business process.

The other 3 categories are Hobbyist, Student and Other which are for Individual developer account

The screenshot shows the Twitter Developer API application form. On the left is a purple sidebar with a '#welcome' message and instructions. The main area is titled 'What is your primary reason for using Twitter developer tools?' and lists four categories: Professional, Hobbyist, Academic, and Other. Each category has a list of use cases with radio buttons. 'Exploring the API' under the Hobbyist category is selected. A 'Next' button is at the bottom right.

Professional ...for commercial uses	Hobbyist ...for a personal project	Academic ...for education or research	Other I don't fit any of those
<input type="radio"/> Building B2B products	<input type="radio"/> Making a bot	<input type="radio"/> Doing academic research	<input type="radio"/> Embedding Tweets on a website
<input type="radio"/> Building consumer products	<input type="radio"/> Building tools for Twitter users	<input type="radio"/> Teaching	<input type="radio"/> Doing something else
<input type="radio"/> Build customized solutions in-house	<input checked="" type="radio"/> Exploring the API	<input type="radio"/> Student	
<input type="radio"/> Publishing ads programmatically			

Specify Intended use

In this step, you have to specify the use case for your application.

Be as specific as you can while specifying your intended use of the Twitter API

The more specific you are, the better are your chances of getting approval.

IMPORTANT ! READ ALL THE guidelines before applying

Make sure you have read the [Developer Agreement and Policy](#), [Automation rules](#) and [The Twitter Rules](#) thoroughly before applying.

Developer Agreement and Policy

It consists of the *DOs* and *DON'Ts* of how to use the Twitter API.
It's a pain in the *neck* to read the entire thing, so I have summarized it for you.

Summary

- Reverse Engineering - Don't try to reverse engineer the Twitter API. Do not try to sell, lease, distribute, or provide access to any licensed material to a third-party.
- Security - Never give away your account's API keys (The ones which you generate after creating an app)
- Rate limits - Don't call the Twitter API endpoints beyond the specified rate limits. In short, *don't spam*. The rate limits of individual endpoints can be found in [Twitter API reference](#).
- Location Data - This data can only be used to identify the tagged location of Twitter content like tweets, retweets, DMs, and more.
- Use of Twitter Marks - Don't use the Twitter Logo for this account
- Automation Guidelines - Automated liking, automated bulk following, and automated adding to lists or collections are discouraged.

⚠ Failing to meet these conditions will cause your application to be rejected

Review, confirm the application, and wait!




⚠ Note: Once you submit your application, it *cannot be edited*. So make sure you have reviewed it properly.

Also, *take a screenshot* of the 'review application' page, so that you can view it later.

It usually takes a day or two, or sometimes more, for your application to be reviewed by Twitter.

After you receive the E-mail:

Twitter can send you three types of emails -

-  Approved developer account
-  Application *rejected*
-  *Review* - They will ask you to review your application since it violates one or more sections of their policies.

Chapter 3: Project Initialization

3.1 Import Python Libraries

Import in Python helps you to refer to the code, i.e., .functions/objects that are written in another file. It is also used to import python libraries/packages that are installed using pip(python package manager), and you need then to use in your code.

The following libraries will be used throughout the post.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import re
from sklearn.model_selection import train_test_split
import nltk
nltk.download('stopwords')
from nltk.tokenize import RegexpTokenizer, WhitespaceTokenizer
from nltk.stem import WordNetLemmatizer
from nltk.corpus import stopwords
import string
from string import punctuation
import collections
from collections import Counter
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
import en_core_web_sm
from sklearn.feature_extraction.text import CountVectorizer, TfidfTransformer
from sklearn.metrics.pairwise import cosine_similarity
from sklearn.metrics import jaccard_score
```

```
[nltk_data] Downloading package stopwords to
[nltk_data] C:\Users\503TS\AppData\Roaming\nltk_data...
[nltk_data] Package stopwords is already up-to-date!
```

3.2 Tweets Datasets

The data used is scraped from twitter using Tweepy, a python library for accessing the Twitter API. It has 197802 tweets from different users from Kenya. Code to scrap the data is available in this repository.

The data set is called "df".

A random sample:

tweets	
0	@mansukhmandviya The government is making all ...
1	Progress of #OmicronVariant is being monitored...
2	RT @TheHinduSports: After the race in Jeddah, ...
3	RT @TheHinduCinema: Stand-up comic @@Comedy_Pr...
4	Covid-19 casualties in Maharashtra down 61 per...

Chapter 4: Preprocessing of Tweets

4.1 Cleaning tweets

Tweets contains unnecessary objects like hashtags, mentions, links and punctuation that can affect the performance of an algorithm thus they have to be rid off. All the texts are converted to lower case to avoid algorithms interpreting same words with different cases as different.

```
# remove the hashtags, mentions and unwanted characters.
def clean_text(text):
    text = re.sub(r'@[A-Za-z0-9]+', '', text)
    text = re.sub(r'#', '', text)
    text = re.sub(r'RT[\s]+', '', text)
    text = re.sub(r'https?:\/\/\S+', '', text)
    return text
df['tweets'] = df['tweets'].apply(clean_text)
```

After Cleaning the tweets our dataset df will look like

```
: df.head()
```

```
:
```

tweets

0	The government is making all possible efforts...
1	Progress of OmicronVariant is being monitored ...
2	: After the race in Jeddah, 1996 world champio...
3	: Stand-up comic @_Praveen talks about his new...
4	Covid-19 casualties in Maharashtra down 61 per...

4.2 Tokenization, Lemmatization and removing stopwords

Stopwords are commonly used words whose presence in a sentence has less weight compared to other words. They include words like 'and', 'or', 'has' et.c.

Tokenization is the process of splitting a string into a list of tokens. A sentence can be reduced to words and a word can be reduced to letters using the appropriate tokenizers.

Lemmatization is reducing a word to its root form. For instance the root form of 'rocks' is 'rock'.

Languages used in the tweets are mainly English and Swahili. The latter has no support thus we'll only work with the former . This renders the analysis crippled in a way given that the Swahili texts will be ignored.

```
nlp = en_core_web_sm.load()
tokenizer = RegexpTokenizer(r'\w+')
lemmatizer = WordNetLemmatizer()
stop = set(stopwords.words('english'))

#already taken care of with the cleaning function.

punctuation = list(string.punctuation)
stop.update(punctuation)
w_tokenizer = WhitespaceTokenizer()
def furnished(text):
    final_text = []
    for i in w_tokenizer.tokenize(text):
        if i.lower() not in stop:
            word = lemmatizer.lemmatize(i)
            final_text.append(word.lower())
    return ' '.join(final_text)
df['tweets'] = df.tweets.apply(furnished)
```

Chapter 5: Tweets Classification

This approach uses the technique of creating a set of words that can be confidently classified as belonging to a particular category for each of the 4 classes. (Economic, Social, Cultural and Health)

The tweets are each compared with the 4 sets and assigned a similarity score. There are 2 main techniques popular for computing similarity score between documents:

5.1 What is Cosine Similarity ?

Cosine similarity is a metric used to measure how similar documents are irrespective of their size. Mathematically, it measures the cosine of the angle between two vectors projected in a multi-dimensional space. This would involve creating word vectors for the set of words and all the tweets then performing the cosine similarity. TFIDF (bag of words model) Vectorizer would be ideal for the vectorization.

5.2 What is Jaccard Similarity ?

Jaccard similarity or intersection over union is defined as size of intersection divided by size of union of two sets.

Jaccard similarity takes only a unique set of words for each sentence or document while cosine similarity takes total length of the vectors. Jaccard similarity is good for cases where duplication does not matter, cosine similarity is good for cases where duplication matters. In our case, context matters more than duplication thus Jaccard similarity is the ideal technique to use.

Chapter 6: Creating Sets of Related Words

The block below represents economy related words. There's 3 other such sets (social_related_words, health_related_words and culture_related_words) for the 3 remaining groups.

6.1 Economy Related Words

““ agriculture infrastructure capitalism trading service sector technology economical supply industrialism efficiency frugality retrenchment downsizing credit debit value economize save economically economies sluggish rise rising spending conserve trend low-management decline industry impact poor profession surplus fall declining accelerating interest sectors balance stability productivity increase rates pushing expanding stabilize rate industrial borrowing struggling deficit predicted increasing data economizer analysts investment market-based economy debt free enterprise medium exchange metric savepoint scarcity capital bank company stockholder fund business asset treasury tourism incomes contraction employment jobs upturn deflation macroeconomics bankruptcies exporters hyperinflation dollar entrepreneurship upswing marketplace commerce devaluation quicksave deindustrialization stockmarket reflation downspin dollarization withholder bankroll venture capital mutual fund plan economy mortgage lender unemployment rate credit crunch central bank financial institution bank rate custom duties mass-production black-market developing-countries developing economic-growth gdp trade barter distribution downturn economist ””

6.2 Social Related Words

““ sociable gregarious cultural politics societal friendly society socialization political mixer sociality interpersonal ethnic socially party welfare economic public health community socialist societies development educational intellectual religious social network humans socialism collective personal corporation social constructivism relations of production organisms animals volition socii citizenship brute beast animal multiethnic ethnical fauna creature herding attitude rights swarming socio sociological sociopolitical socioeconomic ethics civic

multi-ethnic communal marital corporate education sociale socialized communities employment human policies industrial focus sustainable creating civil reform fringe undesirable governance issues multiparty policy emphasis employee tea unions nets perspective urban subsidised culture focuses environment parasite fiesta institutions focusing values labour poverty particular own creation focused awareness company governmental promote labor aspects organizations change critical jamboree promoting festivity create work context the changing integration poor organizational ideas reforms institutional fairness ways learning support challenges care leadership important interests kuomintang sides progressive fundamental example among attention moreover basic activism advancement experience problems understanding name life desa stock sociocultural ”

6.3 Health Related Words

“ disease obesity world health organization medicine nutrition well-being exercise welfare wellness health care public health nursing stress safety hygiene research social healthy condition aids epidemiology healthiness wellbeing care illness medical diet education infectious disease environmental healthcare physical fitness hospitals health care provider doctors healthy community design insurance sanitation human body patient mental health

medicare agriculture health science fitness health policy weight loss physical therapy psychology pharmacy metabolic organism human lifestyle status unhealthy upbeat vaccination sleep condom alcohol smoking water family eudaimonia eudaemonia air house prevention genetics public families poor needs treatment communicable disease study centers improve problems experts services benefits treating hiv agencies benefit patients concerned risk tuberculosis according protection malaria development food priority ”

6.4 Culture Related Words

“ society civilization philosophy anthropology subculture acculturation religion cultivation nationalism counterculture cultural ideology art popular culture folklore agriculture country writing music monoculture cyberculture language social class high culture cultural studies cultural anthropology cooking literature science growth tillage grow ritual perfection development metaphor concept symbol mythology gender tradition

clothing edward burnett tylor traditions traditional western culture contemporary multiculturalism elite politics ethnicity heritage sociology modernity spirituality marxism material culture low culture mass culture critical theory ethos nationality humanism romanticism finish polish refinement civilisation traditionalism genetics human learning interaction kinship heredity marriage dance technology shelter indigenous peoples of the americas growing biology starter viticulture discernment content maturation appreciation ontogeny ”

Just like the tweets, they have to undergo some preprocessing. The function furnished used on the tweets is applied on the sets.

```
In [13]: economy = furnished(economy_related_words)
social = furnished(social_related_words)
culture = furnished(culture_related_words)
health = furnished(health_related_words)
```

The duplicates are also dropped:

```
In [14]: string1 = economy
words = string1.split()
economy = " ".join(sorted(set(words), key=words.index))
economy
```



```
In [15]: string1 = social
words = string1.split()
social = " ".join(sorted(set(words), key=words.index))
social
```

```
In [16]: string1 = health
words = string1.split()
health = " ".join(sorted(set(words), key=words.index))
health
```

```
In [17]: string1 = culture
words = string1.split()
culture = " ".join(sorted(set(words), key=words.index))
culture
```

Jaccard Similarity Scores

The Jaccard index, also known as the Jaccard similarity coefficient, is a statistic used for gauging the similarity and diversity of sample sets. It was developed by Paul Jaccard, originally giving the French name coefficient de communauté, and independently formulated again by T. Tanimoto. Thus, the Tanimoto index or Tanimoto coefficient are also used in some fields. However, they are identical in generally taking the ratio of Intersection over Union. The Jaccard coefficient measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

```
In [18]: #Jaccard Similarity Scores
def jaccard_similarity(query, document):
    intersection = set(query).intersection(set(document))
    union = set(query).union(set(document))
    return len(intersection)/len(union)
def get_scores(group,tweets):
    scores = []
    for tweet in tweets:
        s = jaccard_similarity(group, tweet)
        scores.append(s)
    return scores
e_scores = get_scores(economy, df.tweets.to_list())
s_scores = get_scores(social, df.tweets.to_list())
c_scores = get_scores(culture, df.tweets.to_list())
h_scores = get_scores(health, df.tweets.to_list())
```

There might be a thin line between the economic and social scores depending on the sets of words used.

Chapter 7: Applying clustering algorithm

7.1 Clustered Data Frame

We wish to create a data frame containing the total number of tweets per category per person. A 4D data frame with the index column populated with users, and 3 other columns containing the total number of the user's tweets under social, cultural, health and economic classes.

This can be achieved first by creating a data frame containing Jaccard scores for each tweet for each category, then assigning a tweet to a category depending on the highest score and finally grouping the tweets by user names and sum of the tweets.

The final data frame:

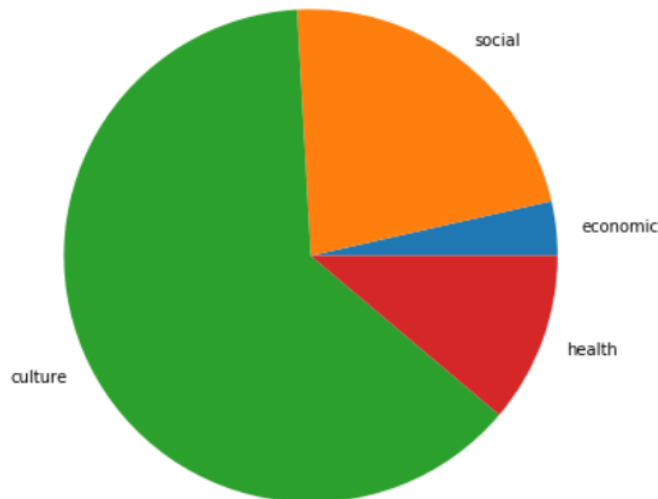
	names	economic_score	social_score	culture_score	health_scores
0	government making possible effort country achi...	0.656250	0.677419	0.724138	0.700000
1	progress omicronvariant monitored daily: healt...	0.689655	0.714286	0.769231	0.740741
2	race jeddah, 1996 world champion damon hill to...	0.666667	0.687500	0.677419	0.656250
3	stand-up comic @_praveen talk new special kanc...	0.647059	0.666667	0.656250	0.687500
4	covid-19 casualty maharashtra 61 per cent nove...	0.645161	0.666667	0.655172	0.689655
...
192	photos: 'secret' wedding tejashwi yadav rachel...	0.633333	0.655172	0.642857	0.620690
193	meta platforms inc's cryptocurrency wallet, no...	0.677419	0.700000	0.750000	0.724138
194	grounded earlier, 737 max face mid-air emergency	0.548387	0.566667	0.551724	0.586207
195	covid19	0.166667	0.172414	0.185185	0.178571
196	covid19update 131.18 cr vaccine dos administer...	0.487805	0.500000	0.526316	0.512821

197 rows × 5 columns

Below is a pie chart to show the tweets volumes in the different categories:

```
In [21]: fig = plt.figure(figsize =(10, 7))
a = new_groups_df.drop(['total'], axis = 1)
plt.pie(a.loc['Total'], labels = a.columns)
plt.title('A pie chart showing the volumes of tweets under different categories.')
plt.show()
```

A pie chart showing the volumes of tweets under different categories.



culture has the largest percentage. This could be as a result of the current pandemic that everyone is talking about.

The data can be played with for loads of analysis and beautiful visualisations but the focus of the post is cluster analysis.

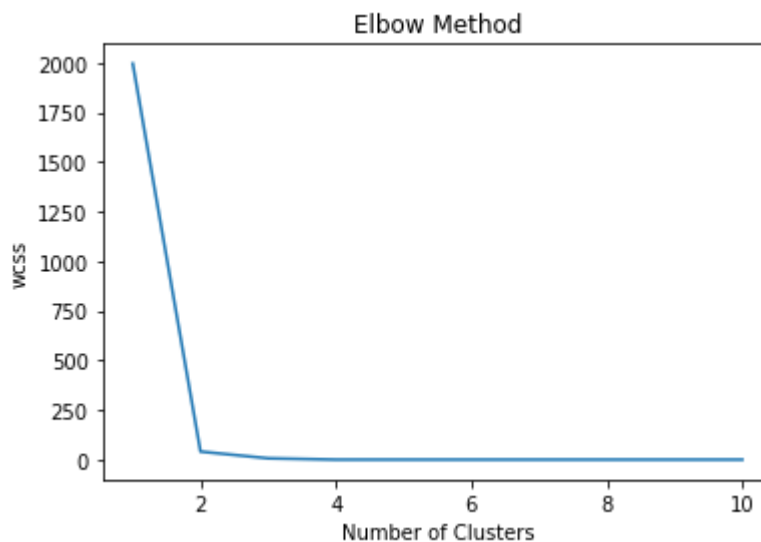
7.2 KMeans Clustering

“ Distance computation in k-Means weighs each dimension equally and hence care must be taken to ensure that unit of dimension shouldn’t distort relative near-ness of observations. Common method is to unit-standardize each dimension individually. ”

The unit for the variables of interest are the same: Number of tweets, thus no need for standardization. The code below would standardize a column ‘a’ if there was the need:

We will work with 2D clustering, i.e clustering between 2 variables. There are different methods to determine the optimal number of clusters and one of them is the elbow method. The approach consists of looking for an elbow in the WCSS graph. Usually, the part of the graph before the elbow would be steeply declining, while the part after it would be smoother.

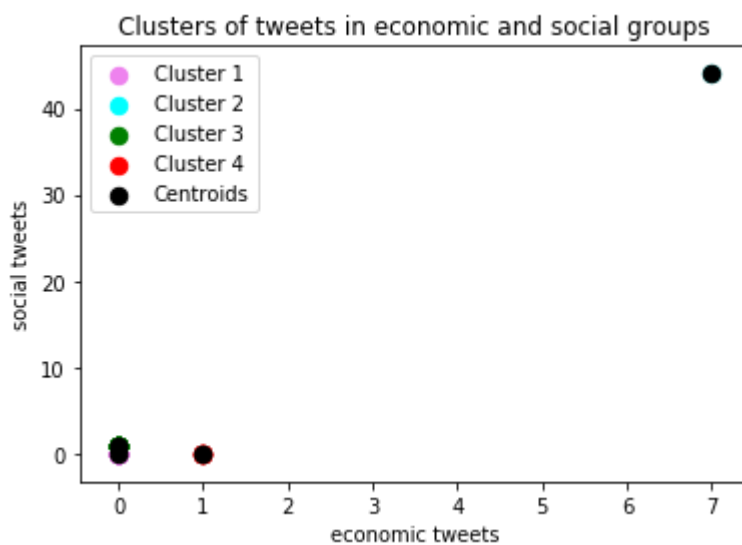
```
In [34]: X = new_groups_df[['economic', 'social']].values
# Elbow Method
from sklearn.cluster import KMeans
wcss = []
for i in range(1, 11):
    kmeans = KMeans(n_clusters=i, init='k-means++', n_init=10, max_iter=300, random_state=0)
    kmeans.fit(X)
    wcss.append(kmeans.inertia_)
plt.plot(range(1,11), wcss)
plt.title('Elbow Method')
plt.xlabel('Number of Clusters')
plt.ylabel('wcss')
plt.show()
```



Taking $k = 4$.

```
In [37]: # fitting kmeans to dataset
kmeans = KMeans(n_clusters=4, init='k-means++', n_init=10, max_iter=300, random_state=0)
Y_kmeans = kmeans.fit_predict(X)
# Visualising the clusters
plt.scatter(X[Y_kmeans==0, 0], X[Y_kmeans==0, 1], s=70, c='violet', label= 'Cluster 1')
plt.scatter(X[Y_kmeans==1, 0], X[Y_kmeans==1, 1], s=70, c='cyan', label= 'Cluster 2')
plt.scatter(X[Y_kmeans==2, 0], X[Y_kmeans==2, 1], s=70, c='green', label= 'Cluster 3')
plt.scatter(X[Y_kmeans==3, 0], X[Y_kmeans==3, 1], s=70, c='red', label= 'Cluster 4')

plt.scatter(kmeans.cluster_centers[:, 0], kmeans.cluster_centers[:, 1], s=70, c='black', label='Centroids' )
plt.title('Clusters of tweets in economic and social groups')
plt.xlabel('economic tweets')
plt.ylabel('social tweets')
plt.legend()
plt.show()
```



The plot above indicates most of the users share more economy-centred tweets compared to social tweets. There's a few who try to maintain a balance between the categories.

The same method can be implemented on the other pairs to observe how they relate and interpretations made.

7.3 Conclusion

We present a system for the acquisition, analysis and visualisation of Twitter data. Twitter messages are harvested and stored in a distributed cluster, and the data is processed using algorithms implemented. We present a clustering algorithm capable of identifying the main topics of interest in a tweet data set. Also, we designed a visualization method which allows us to follow the intensity of Twitter activity at a given geographical location.

References

- [1] Tweets Classification and Clustering in Python.
<https://medium.com/swlh/tweets-classification-and-clustering-in-python-b107be1ba7c7>
- [2] Step-By-Step Twitter Sentiment Analysis
<https://ipullrank.com/step-step-twitter-sentiment-analysis-visualizing-united-airlines-pr-crisis>
- [3] Natural Language Toolkit . <https://www.nltk.org/>
- [4] Matplotlib Documentation:
<https://matplotlib.org/stable/contents.html>
- [6] Pandas Documentation: <https://pandas.pydata.org/docs/pandas.pdf>
- [7] Seaborn Documentation: <https://seaborn.pydata.org/tutorial.html>
- [8] X. Wang, R. Guo and C. Kambhamettu, "Deeply-Learned Feature for Age Estimation," 2015 IEEE Winter Conference on Applications of Computer Vision, Waikoloa, HI, 2015, pp. 534-541, doi: 10.1109/WACV.2015.77.
- [9] MLP Classifier Sklearn/Tensorflow Method :
https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html &
https://www.tensorflow.org/recommenders/api_docs/python/tfrs/layers/blocks/MLP
- [10] Additional Reference :
<https://towardsdatascience.com/multi-layer-neural-networks-with-sigmoid-function-deep-learning-for-rookies-2-bf464f09eb7f>
- [11] C. Ranjan "Rules-of-thumb for building a Neural Network" in towards data science, 2019. [Online]. Available: <https://towardsdatascience.com/17-rules-of-thumb-for-building-a-neural-network-93356f9930af>

