DE-35(CE)

NUST COLLEGE OF

ELECTRICAL AND MECHANICAL ENGINEERING

SOCIOLOGY ANALYSIS USING DATA ANALYTICS

DE-CE (DCE)

Submitted by

RAFAY ULLAH CHOUDHARY

USAMA ZAHID

MUHAMMAD USAMA

BACHELORS

IN

COMPUTER ENGINEERING

YEAR

2017

PROJECT SUPERVISOR

BRIG. DR. SHOAB AHMED KHAN

DR. WASI HAIDER BUTT

YEAR 2017

NUST COLLEGE OF

ELECTRICAL AND MECHANICAL ENGINEERING

SOCIOLOGY ANALYSIS USING DATA ANALYTICS

DE-CE (DCE)

Submitted by

RAFAY ULLAH CHOUDHARY

USAMA ZAHID

MUHAMMAD USAMA

BACHELORS

IN

COMPUTER ENGINEERING

YEAR

2017

PROJECT SUPERVISOR

BRIG. DR. SHOAB AHMED KHAN

DR. WASI HAIDER BUTT

DECLARATION

We proclaim that in this report no piece of the work done has been submitted in support of an application for another capability or degree of this or some other organization or college of learning. In the event that any counterfeiting discovered, we will be completely in charge of each move made against us; contingent on the earnestness of the demonstrated offense, notwithstanding prompting the cancelation of our degree.

COPYRIGHT STATEMENT

TABLE OF CONTENTS

Contents

TABLE OF FIGURES

ACRONYMS

API **A**pplication **P**rogramming **I**nterface
VADER **V**alance **A**ware **D**ictionary and s**E**ntiment **R**easoner
JSON **J**ava**S**cript **O**bject **N**otation
BSON **B**inary **J**ava**S**cript **O**bject **N**otation
NaN Not a Number
HDF5 **H**ierarchical **D**ata **F**ormat **F**ile
OAuth **O**pen **A**uthorization
noSQL no **S**tructured **Q**uery **L**anguage
URL **U**niversal **R**esource **A**llocator

ACKNOWLEDGEMENTS

In the name of Allah, the Most Beneficent, the Most Merciful

ALHAMDULILLAH, all gestures of recognition to ALLAH for the strength He gave us, and gifts He showered upon us, which prompted the convenient fulfillment of this project. Without His incalculable gifts, we could never have possessed the capacity to come this far.

We might want to offer our sincerest appreciation to our supervisor Dr. Shoab Ahmed Khan, for his supervision and enormous help. He has the state of mind and a substance of virtuoso. He helped us at all times, tolerance and nonstop tutoring. Without his consistent direction and affirmation, this venture would not have been finished in the given time span; neither would we have gained as much ground as we have figured out how to. Likewise, we might want to thank our co-supervisor Dr. Wasi Haider Butt for his specialized help and direction all through our project.

What's more, we as a whole are unbelievably appreciative to our families, for unending help and consolation.

ABSTRACT

Sociology is the investigation of social conduct of society, including its inceptions, advancement, association, systems, and foundations. It is a sociology that utilizes different techniques for observational examination and empirical investigation.

The customary concentrations of sociology incorporates social stratification, social class, social portability, religion, secularization, law, sexuality and aberrance. As all circles of human action are influenced by the interaction between social structure and individual organization. Sociology has gradually expanded its focus to further subjects, such as health, medical, military, music and food.

Our project focuses on: Food, Music and Social sentiments. Qualitative research has been done on some of these issues, while our project does quantitative research which makes it unique and peculiar. Our data sets are huge and complex which helps in finding the precise and most suitable trends of the complex and ever changing society. The end results are displayed elegantly and simply which helps the end users to understand it easily.

It will help in understanding trends about sociology which will make it easier to understand a society and will help the stakeholders to pitch the idea in that particular society which will help in increasing the probability of desired results. The results are visualized in the form of graphs.

CHAPTER 1
INTRODUCTION

1.1 Background

Sociology is a vast field, it covers aspects from religion, secularization to individuals in a society, how a particular society thinks, and their nature towards a particular subject. The traditional focus for the study of society is bases on qualitative analysis meaning to study a society, a particular sample of population is considered and mostly a set of questions are developed and are than analyzed.
In this project, we study limited fields stated Food, Music and Social Sentiments about any particular subject globally. We automate this process with the help of data science by considering datasets that are huge and precise.

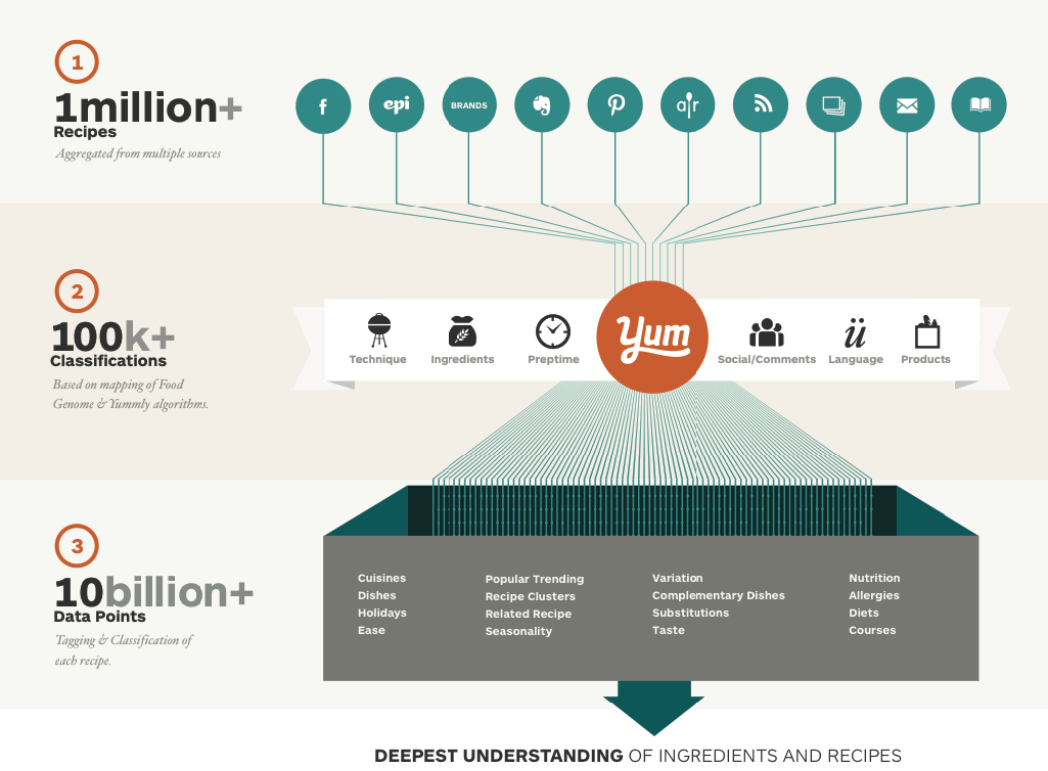


Figure 1.1 Food dataset provided by Yummly

There are nearly **2,00,000 ingredients** in this world, probably an understatement. Using the data provided by Yummly, the details of dataset are shown in figure 1.1, we can investigate that which cuisines are similar to one another, on average how healthy particular cuisines are, what is the correlation between nutrients usage and IQ value of a particular region.

There are nearly **97 million official songs**, we can use the songs dataset provided by various sources to extract interesting trends that how the music evolved over time by looking at their tempo, music duration and lyrics etc.

Similarly, micro-blogging is the future, social media apps like Facebook and Twitter are getting huge amount of user base and the data is being generated every second, growth of tweets are shown in figure 1.2. As of now there are **500 million tweets/day** generated and approximately **7.5k tweets** are tweeted **every second**. We can use these tweets to gather sentiments about a particular subject and at which frequency a particular subject is active in a particular region in real-time.

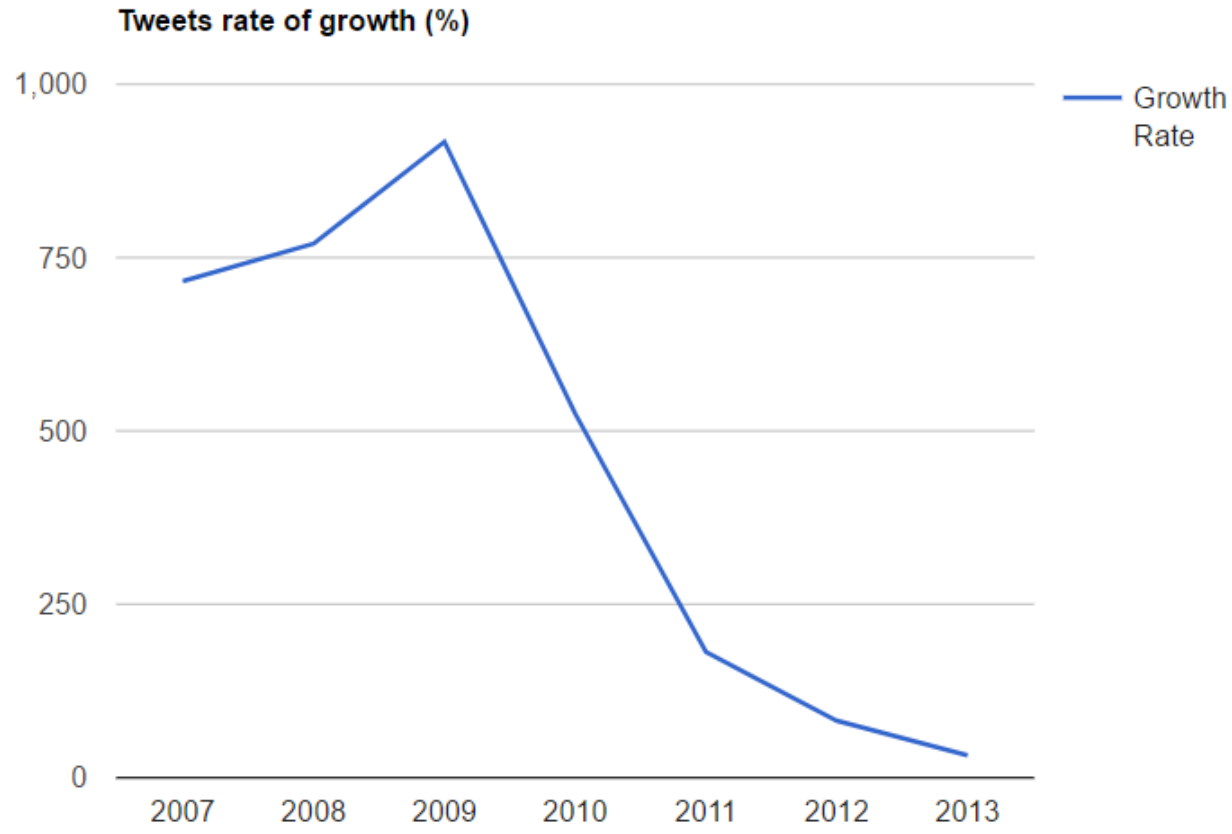


Figure 1.2 Twitter Growth Rate

1.2 Overview

When it comes to Data Science, one of the most important factor is the data and its legitimacy. As stated, our project relies on multiple dataset, each for their particular domain of study, figure 1.3 shows the flow diagram of how every component in interconnected:

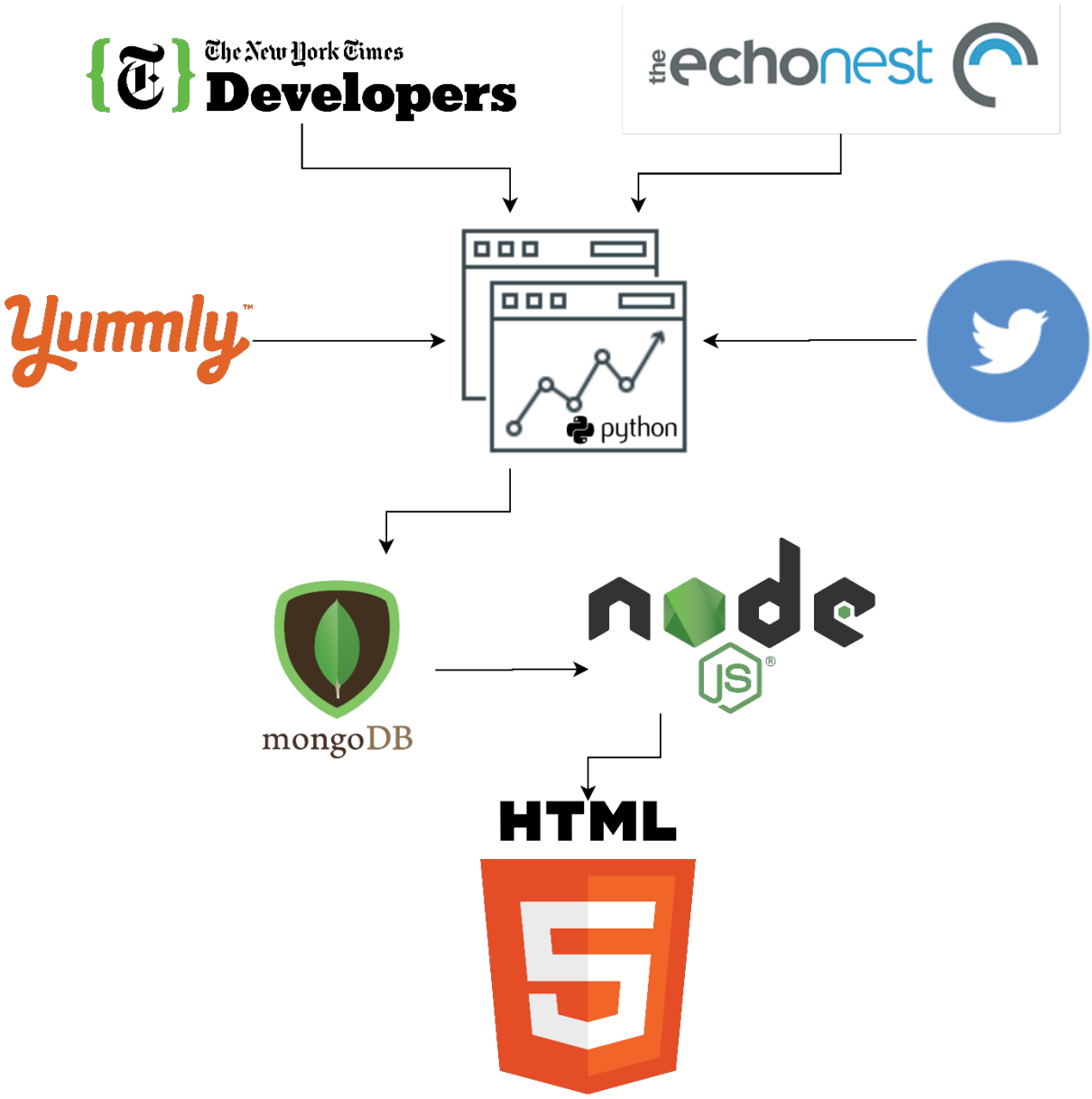


Figure 1.3 Flow Diagram

All of the data sources are legit, accurate and are big names.

Following are the datasets used in the respected fields:

- Yummly, recipes dataset used for Food
- Echonest, songs dataset used for Music
- Twitter, real-time data stream for calculating social sentiment

Figure 1.4 shows the steps performed in nearly every domain at a certain level of abstraction:

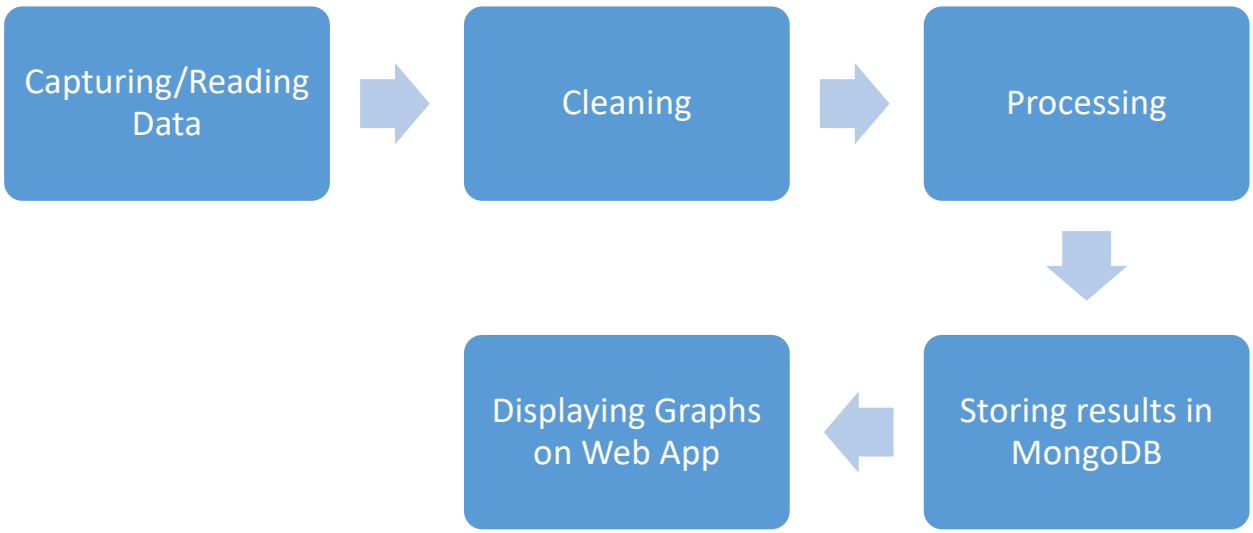


Figure 1.4 Schematic of System level diagram

Scripts for data, retrieving, processing and querying databases are written in python and all the server and client side scripting for web application is written in JavaScript with Node JS.

1.3 API's used

API stands for Application Programming Interface is a set of functions that allow the creation of applications which access the features of OS or other apps. It also provides a layer of abstraction over the implementation. For data acquisition in our project, following API's are used:

1.3.1 Yummly API

Yummly API aggregates 2 Million+ recipes from all over the world. It contains the ingredients used in a recipe, their corresponding nutrition value which can help track for meals and calorie intake, also helps in developing meal plans. This API is used in many food applications to develop food recommendations based on likings of the audience. Yummly also understands diets, allergy, taste technique and more. Response of this API is in JSON.

1.3.2 Echonest API

The Echonest API provides broad data on millions of artists and songs that contains Artist and song hotness, Year, Song popularity, category, Tempo, music duration, loudness and dance ability etc.

Many popular applications have built popular and powerful applications using this dataset, suggesting soundtracks to people based on the stats of songs they already like

and also recommends them similar artists based on the similarity among different artist.

They provide a Million Song Dataset which contains data of a million songs and its subset of 10,000 songs. Our objective here is to study how the music has evolved over the period of time by considering some of the aspects available in the dataset.

1.3.3 Twitter API

Twitter API is used via Python's tweepy library, it communicates with the twitter's API to use its functionalities, in this project we are using the streaming mode which has the ability to constantly request for tweets in real time, figure 1.5 shows the architecture of streaming API:

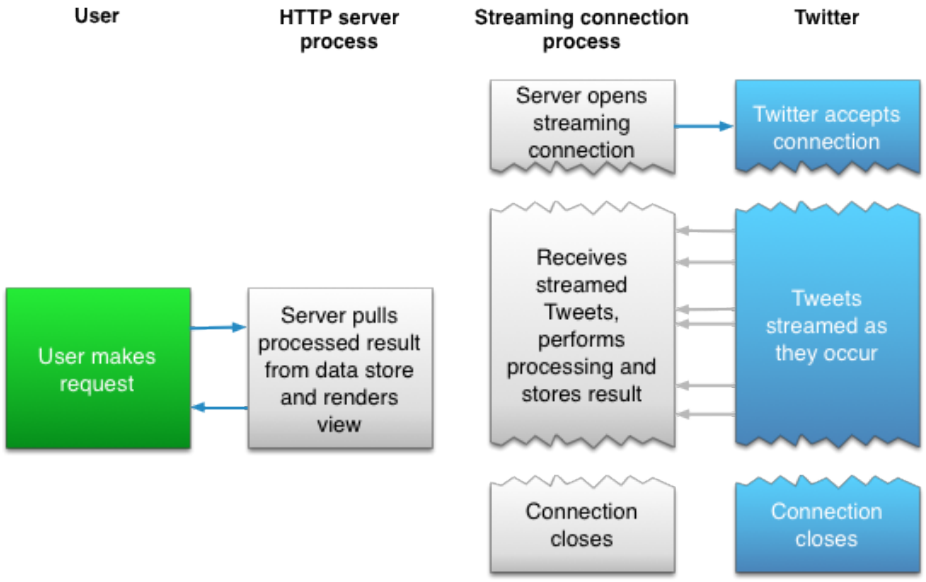


Figure 1.5 Social Media Activity

Result of the API is in JSON format, which makes it very easy to dump in noSQL databases which natively support JSON/BSOM themselves.

1.3.4 NewYorkTimes API

New York Times is an American daily newspaper founded in New York Times in 1851, since then it has won 122 Pulitzer Prizes, more than any other newspaper.
NYTimes provides their API that helps the developers to retrieve news from as early as 1851 and develop their apps over it or perform data analytics. NYTimes suites comprises of some of the following APIs as listed in table 1:

API	Functionality
Archive	Provides the list of NYTimes articles
Article Search	Provides search for a given subject returning article's headline, abstract, date and location of publication etc.
Community	Provides access to comments from registered users and what their views on articles
Semantic	Provides the access to the long list of people, places, organization and other entities that makes up the news

Table 1 NYTimes API list

We want to take advantage of the Article Search API, its Meta data consists of hits that a particular subject has for a certain date, our objective is to visualize the popularity of a certain subject over a period of time.

1.4 Motivation

Given the fact that internet is the future, and data is only increasing every second, interesting facts about societies can be extracted by converting that data into information, any social aspect can be studied.

This data can not only be used to gather the fact findings but it can also help people establishing their businesses and increase the profit. Many interesting trends and correlations can be extracted and hence here in this very project, we also try to explore them.

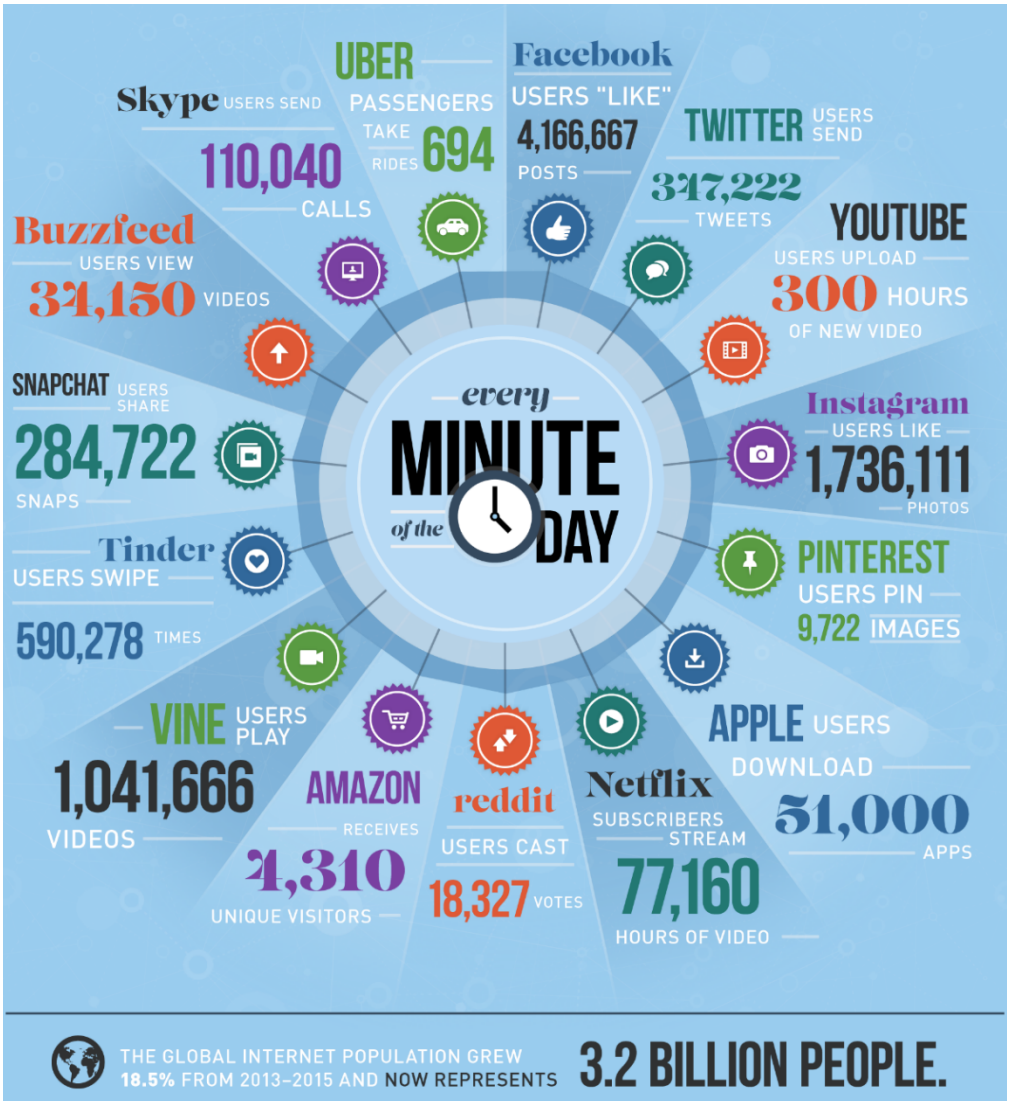


Figure 1.6 Social Media Activity

Twitter is a micro blogging site where interestingly people are very subjective about how they feel over a particular subject and also majority of the accounts are open to public as there is less trend towards more personal sharing sites such as Facebook, it makes it easier to understand the audience. Figure 1.6 shows the internet population and division of it across different platforms.

One of the coolest motivation was during BrExit, British Exit, when interestingly the finding predicted that there was a more chance for it to happen and it actually did, similarly same was the case during US elections, this was all possible due to social media analysis and data science **[1]**.

1.5 Objectives

Project objectives are listed below:

- To find the social sentiments of audience on a given particular topic in real time based on their location of tweet.
- To visualize the amount of activity from a particular region.
- Display a real-time graph visualization on an elegant and easy to understand GUI.

- To find the ingredient based similarity of a particular food cuisine from the cuisines of rest of the world.
- To find the frequency in which a particular ingredient is used different parts of the world for example the usage of salt in Pakistani, Italian or Afghani cuisines etc.
- To find the average cooking time and amount of Fats, Proteins, Sugar or Carbohydrates etc. found in cuisines from all over the world.
- To find the correlation between IQ value and Protein etc. found in cuisine of a particular region.
- To find how the music evolved until now based on the Music Duration, Tempo, and Hotness.
- To find the correlation of Song popularity with Artist popularity, song's loudness, duration tempo and energy.

1.6 Structure of Report

Chapter 2 gives a detailed background of what sociology is and how it is related to data analytics.

Chapter 3 gives a detailed literature review of the work done previously.

Chapter 4 gives the implementation detail of each field: music, food and sentiment analysis.

Chapter 5 shows the visualization and results of our work done in every field and correlation obtained.

Chapter 6 discusses the future work and how the current work can be expanded to different directions.

CHAPTER 2

BACKGROUND

2.1 Sociology

Figure 2.1 Social Media Activity

Sociology is the study of human social relationship and institution as shown in figure 2.1. It is so diverse that we can study anything related to society ranging from crime to religion, family to state, division of race and social classes to shared beliefs of a common culture and from social stability to radical change in society. Society is divided on three level i.e. personal lives, communities and world. Sociology analyze and explain all important matters of society on all level. At personal level sociology investigates personal matters like gender identity, gender equality, individual behaviors, family matters, aging, and religion. At community level, it explains matter like law, crime, poverty, wealth, discrimination, education and social movements. At world level, it explains phenomena like population growth, relation of countries with each other, war, peace and economic development.

The History of sociology may be as old as ancient Greeks. The traces of proto-sociological observation are to be found in texts of western philosophy. Even the word sociology is the mixture of Latin and Greek word “socious” with suffix logy. But the main evidence that show the start of sociology are from 14th century. Ibn Khaldun a 14th century Muslim scholar from Tunisian, North Africa is considered the father of sociology. His book Muqaddimah was the first socio-scientific book published in history of world. Although information gathering from individual (surveys) have older origin than this book, as we can find traces of surveys in a book named Domesday published in 1048.

2.2 Data Science

Also known as data driven science is an inter-disciplinary field which is used to extract insights and trends from large dataset. It is a concept that combine statistics, computing, programming and graphics to analyze and understand phenomena using data. The basis of this process is data without it all algorithms and computing power is useless. It is considered the fourth paradigm of science, the first three are empirical, theoretical, computational and now 4th one is data driven After Harvard Business Review called it "Sexiest job of 21st century" the term become a buzzword. Now the question here is why there is so much hype or so much buzz about this field. The answer is quite simple: Data analysis is an old job, Economist and Accountant analyze data to make decision for the firm or place they are working. They got the data and they analyze it. The amount of data at that time was not so much so they compute the whole thing on paper or excel sheets. But with the passage of time technology advances and the world started to shift toward digital side. This made data generation rapid. According to survey the whole word generated 5 Peta Byte of data till 2005 and onward from 2006 till now the estimated data generation daily is 5 Peta Byte. This is why data become essential for progress and we rely on it. Internet is one of the main reason behind this much data. Micro blogging sites such as Facebook, Twitter, Instagram and WhatsApp are the main source of data generation on internet, Facebook daily generate around 1.2 Tera byte of data alone. Not only these sites are contributing E-Commerce is also playing a vital role in this process. Alibaba Group and Amazon are the most illuminated source in this industry. In 2006-07 when the trend of micro blogging site and e-commerce started, mathematician and data scientist starts thinking what they can do with this data. As data science is all about to extract hidden insights and trends from data they cope this idea and started making different algorithms using stats and machine learning. This data gave them solutions with which they can develop their business in much efficient way. Example of how companies make smarter decision using data science:

- Netflix data mine the pattern to understand user interest so that they can make series related to interest of users. One example is Game of Cards
- Procter & Gamble uses time series data to make marketing decision and to understand future demand.
- Retail or medical shops use customer data to market target customer for products.
- Amazon or other e-commerce sites use user data to give them recommendation according to their taste and their past behavior on site
- Gmail spam filter is data product
- Self-driving cars are data driven products.
- IBM Watson can cure cancer with precision.

Figure 2.2 shows the skill required by a data scientist:

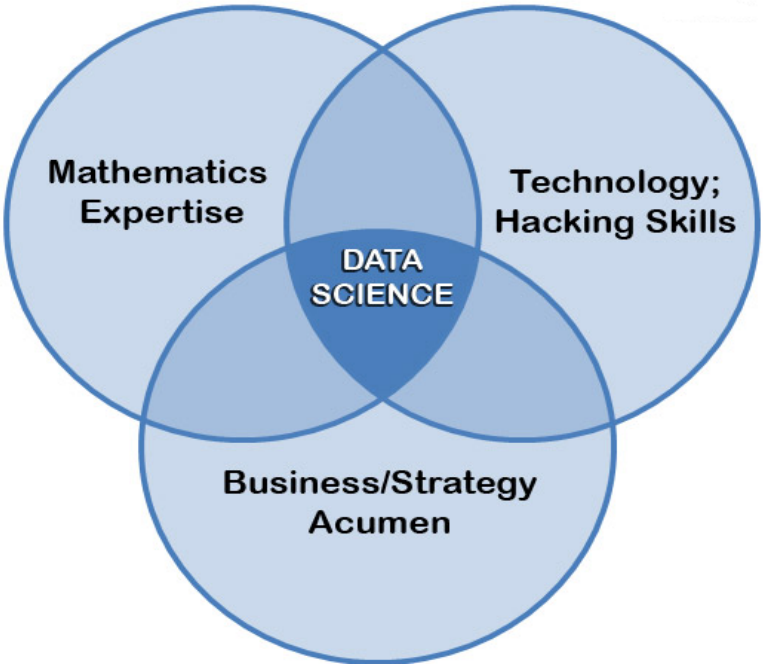


Figure 2.2 Data Scientist skill set

Data scientist create mathematical model and then implement them in any technology either software tool or some programming language to analyze the data and visualize it. After analyzing it make decision according to business.

2.3 Sociology and Data Analytics

Sociology is the study of anything that have impact on society. Society is basically a specific group of people living in specific location geographically sharing same culture, traditions and thinking. To study their culture, religion, gathering, law and crime etc. is sociology of that specific society. Social scientist performs this task in their lab by analyzing some questionnaire. These questionnaires are created by these scientists and then filled by random people of society. The answers to these questionnaires tells psychology of every single person who filled this and collectively answers the sociology of society respected to aspect under study. This is qualitative research and required a lot of time to complete the whole process. For example, due to some situation, people of society have anger. It took too long to study reasons until then sentiments of people changed. That's where data analytics comes. Now a day we can get data of people sentiment on spot. This is just because people use Facebook and twitter to show their sentiment. They share their thoughts, thing they love and even their food. These micro blogging webs become more than just entertainment.

What we have done in this project is similar that sociologist do but the difference is we used digital data to analyze sociology. Our work is not just qualitative but its mixture of quantitative and qualitative analysis. Our dataset is whole globe not like sociologist who work on just hundreds of people as dataset. We study sociology on spot, as sentiment of person change our

results change at the same time. We used Microblogging sites and some other sites to get our data. The aspect under study are Food Sociology, Music Sociology and Activities.

2.3.1 Food Sociology

Food Sociology is one of the worldview of sociology, which is the investigation of food as it identifies with the history, movement and future improvement of society. This incorporates generation, utilization, dispersion, strife, therapeutic application, custom, profound, moral, and social applications, natural and work issues.

Since the start of humankind, food is a standout amongst the most vital part of life with the end goal of sustenance. As primates strolled the earth, they exclusively devoured food for the wellspring of vitality; they chase and look generally for food as food was not effectively available. Early human battled that they require high vitality eating routine to get by in this world. This thing drives them to collect and made food for themselves to satisfy their need. This begins work for food and monetary piece of Sociology of Food. As years went off, food turns out to be increasingly of uniting society and individuals. Many societies are wound with each other in view of food. This conveyed for a considerable length of time. From homo-sapiens chasing to social occasion, to the pilgrim to New world imparting a devour to local American (custom named as Thanksgiving) to the promotion of eateries or eating out from most recent quite a few years and the harmony that accompanied eating indicates how the correspondence and availability that originated from it.

Sociologists separated diverse gatherings of foods because of their motivation and significance. These are social super food that are steady for **Culture**, the distinction food reflects **Economic status** and in conclusion there is physiological gathering which are expended for particular classification like what should a pregnant lady eat for solid pregnancy. These classifications help specialist and humanist to ponder culture in planned food. It regularly indicates how food develop, shape and change with society. Case is Homeopathy that goes under thoughtful food or physiologic food. As they are expended on the conviction of their properties that what benefits they give on the off chance that one uses them. Another case is caviar or clam for eminence food as they are costly and the customer utilizing these foods demonstrate financial status.

Sociological Prospective symbolize the sociology of Food. In many culture food unites individuals and interface them on various level. For instance, the custom of eating with family together. It gives them opportunity to speak with each other. While individuals utilized fast food in America that show occupied family culture as they require moment supper. Food itself could symbolize something more prominent than what it is. McDonaldization hypothesis based on the possibility of American culture of food utilization. As food identifies with support, the early hypothesis of food tells that the fittest can get by as everybody needs to chase or create food for them around then. The hypothesis is same today with the change that one could survive in the event that he has employment to make is living. Food sociology likewise symbolizes brain science and confusion, this speaks to one's control over himself to incorporate or avoid something they have to live from their life. A few people consider food reward or utilize it as solace however a few people consider it contrary thing and maintain a strategic distance from it. These sorts of individuals have scatter like Anorexia or bulimia, they have dread of getting weight or being fat by expending food.

All the above exchange demonstrates the connection of food and society which in logical way you can state sociology of food or to local American (custom named as Thanksgiving) to the promotion of eateries or eating out from most recent quite a few years and the harmony that accompanied eating indicates how the correspondence and availability that originated from it.

2.3.2 Music Sociology

Music have very strong impact on society. Thinking of people and music have reversible effects on each other. Music can affect thinking and emotion of a person. People who listen romantic songs always think about romance and those who listen patriotic songs are full of patriotism. Similarly thinking affect music choice. Music preference can also determine the personality of a person **[2].**

CHAPTER 3

LITERATURE REVIEW

3.1 Sociology & Data Science

The term "Sociology Analysis using Data Analytics" refers to the integration of two fields that are sociology and data science. A lot of research has been done on these two fields, but integration of these two is a new sub-field. As the data is growing tremendously, and society is responsible for the increase of data hence it shows information about society. But we face a lot of challenges as the data produced has to be mined and analysis is to be performed in order to make data meaningful **[3]**. Now we are doing research on 3 topics that are food, music and issues. Food and music are new fields in sociology and have enjoyed a notable boom in the final decade of 20^{th} and in the early years of 21^{st} century **[4] [5].** Currently, sociologists face two big problems which need to be addressed here. First is obtaining good data, as people are reluctant to fill questionnaires and secondly it requires a lot of money. The other problem is determining any kind of casual relationships among humans. Now we have addressed these two problems and have come up with an idea of performing sociological research with data science. Data science is growing exponentially **[6]** yet it faces few problems, the biggest of them is data preparation, and it is increased when performing analysis on raw data. Another issue is to choose the perfect algorithm to solve the problem **[7]**.

3.2 Tools

Now the most popular tools for statistical analysis are:

3.2.1 Mean:

The average of a set of numerical values, as calculated by adding them together and dividing by the number of terms in the set. This is useful in determine the overall trend of dataset. Also it is easy to understand and use.

$$\bar{x} = \frac{\sum X}{n}$$

Pitfall:

Taken alone, mean can be dangerous as sometimes mean is closely related to median and mode.

3.2.2 Standard deviation:

A quantity expressing by how much the members of a group differ from the mean value for the group.

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

Pitfall:

Like mean if taken alone it cannot be accurate. For example if the data has a strange pattern such as non-normal curve it won't give all the information expected.

3.2.3 Regression:

Regression models the relationships between dependent and explanatory variables, which are usually charted on a scatterplot. The regression line also explains whether those relationships are strong or weak.

$$Y = a + bX$$

$$b = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum X^2 - (\sum X)^2} \qquad a = \frac{\sum Y - b\sum X}{N}$$

Where,
N = number of observations, or years
X = a year index (decade)
Y = population size for given census years

Pitfall:

Regression is not very nuanced. Sometimes the outliers matter significantly.

3.2.4 Sample size determination:

Sample size determination is the act of choosing the number of observations or replicates to include in a statistical **sample**. The **sample size** is an important feature of any empirical study in which the goal is to make inferences about a population from a **sample**.

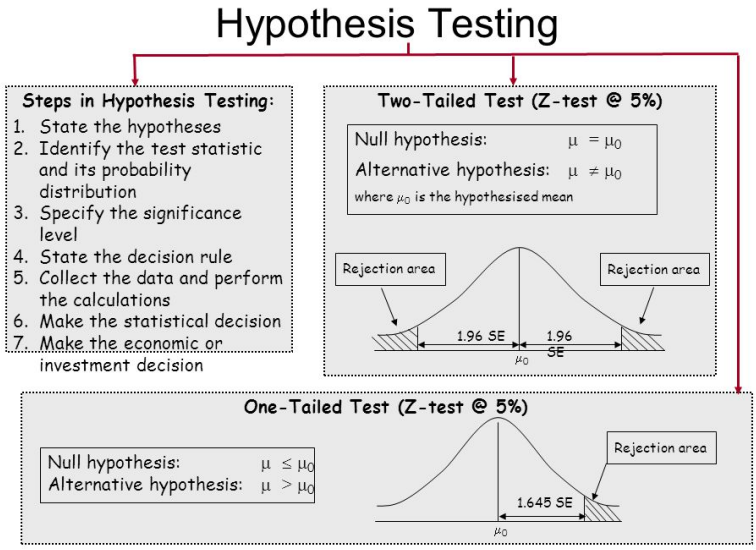
$$Sample\ Size = \frac{\frac{z^2 \times p(1-p)}{e^2}}{1+(\frac{z^2 \times p(1-p)}{e^2 N})}$$

Pitfall:

When contemplating another, untested variable in a populace, your extent conditions may need to depend on specific presumptions. Be that as it may, these suppositions may be totally off base. This blunder is then passed along to your example sample size determination and after that onto whatever remains of your statistical data analysis

3.2.5 Hypothesis testing:

A statistical **hypothesis** is an assumption about a population parameter. This assumption may or may not be true. **Hypothesis testing** refers to the formal procedures used by statisticians to accept or reject statistical **hypotheses**.



Pitfall:

It needs to watch out for common errors like Hawkthrone effect, which is when participants skew results because they know that they are being studied.

3.2.6 Correlation

A mutual relationship or connection between two or more things

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

Pitfall:

A relationship between two variables is sometimes taken as evidence that one causes the other. This is, however, often not true, and hence the popular statistical adage: "Correlation does not imply causation."

CHAPTER 4
METHODOLOGY

4.1 Food

As food changes its importance from nourishment to entertainment in last few decades, it become important part of modern life. Sharing of food and recipes on internet indicates civilization and culinary attitude in different countries. The ingredients and flavors of these recipes shows preference of individuals around the world. Thousands of recipes with millions of ingredients are shared on web. To clear the vague understanding of preferences we are analyzing these recipes to find followings:

1. Ingredient base similarity between cuisines of different countries
2. Average cooking time of cuisines of different countries
3. Relation of nutrition values with diabetes and IQ level

4.1.1 Data

Our study relies on many data sources including Yummly, BBC food and country health statistics by World Bank. In this section, we describe the detail of all datasets. All this data scraped through crawler or using Public API of websites. We wrote our crawler and access these APIs using python's different libraries including BeautifulSoup4, urllib, request and PyMongo.

4.1.1.1 Yummly:

Yummly is a recipe recommendation website on the basis of user taste. User can search recipes. It also provides suggestions on the basis of users past search experience. Yummly also provides user friendly API which we use to collect data. First, we crawled Wikipedia to search cuisine existing in world. On the basis of results, we queried Yummly API to get recipes belonging to that particular cuisine. Yummly have more then million recipes but we limited our dataset due to API limitation plus get better results (some countries like Ethiopia shared very few recipes due to lack of internet access there). We on average stored 250 recipes per cuisine in MongoDB. Response of API is in JSON format, figure 4.1 shows the response. Figure 4.2 shows the API usage. These recipes have following attributes:

1. **Ingredients:** Each recipe contain ingredients that are used to make it with proportion. As Yummly act as aggregator (it scrapes recipes from all around internet) we standardized these ingredients because all web has language according to the country from where that published.

2. **Cooking Time:** Cooking time of all recipes are mentioned with them.

3. **Nutrition:** Yummly provide nutrition value of recipes by aggregating all ingredients used in the recipe. We can't get nutrition value directly by API that's why we wrote a crawler for this.


```
JSON
  criteria
    q : "pakistani"
    allowedIngredient : null
    excludedIngredient : null
  matches
    0
      imageUrlsBySize
        90 : "http://lh3.googleusercontent.com/L_nJ-wQ26sb8dQ64zkggX39n6omOFLbNO1SCDbQzncz6e5f9mh3cUAtC74b-Nfs2U71KHA8znzyc3Me45hD8lA=s90-c"
      sourceDisplayName : "Global Table Adventure"
      ingredients
        0 : "water"
        1 : "milk"
        2 : "cinnamon sticks"
        3 : "cardamom pods"
        4 : "sugar"
        5 : "coffee"
      id : "Pakistani-Coffee-with-Cinnamon-_-Cardamom-1164213"
      smallImageUrls
        0 : "http://lh3.googleusercontent.com/KKRruLsyujFljdMpuifQChYcSH_SRJ8QT6_9iNWFrkornhVO6hBll1ONoqBd0Ujzvproc_nwL2IYE1KpA48P=s90"
      recipeName : "Pakistani Coffee with Cinnamon & Cardamom"
      totalTimeInSeconds : 1200
      attributes
        course
          0 : "Beverages"
      flavors : null
      rating : 4
```


Figure 4.1: Recipes Data with all fields

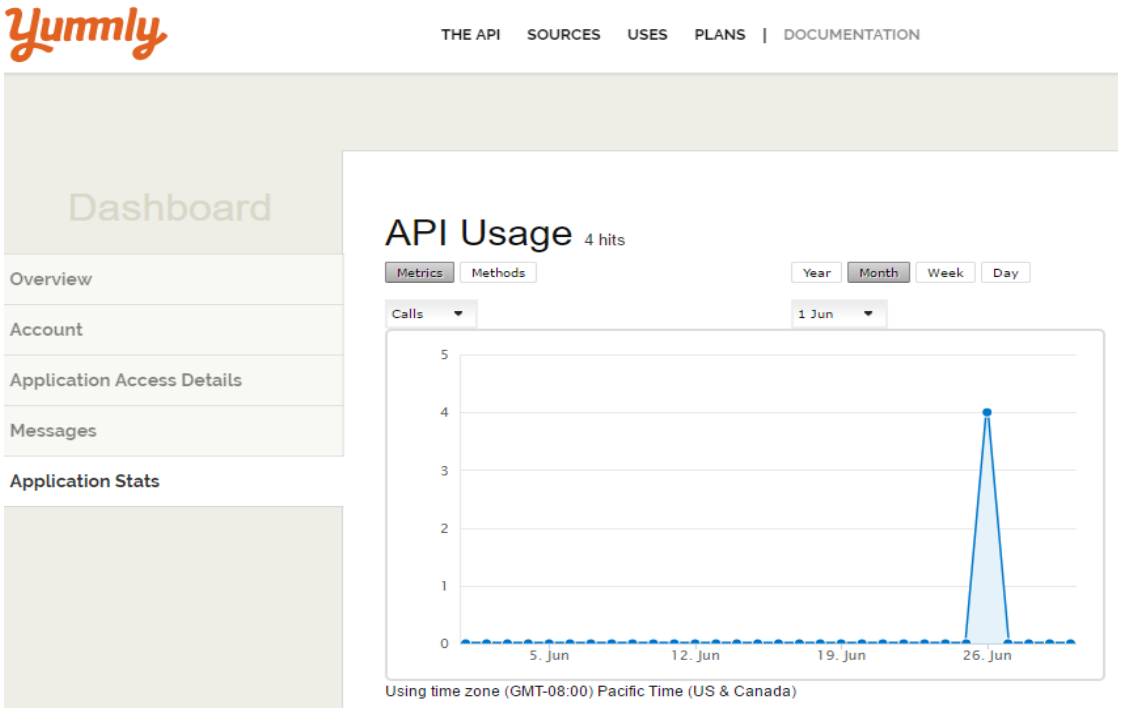


Figure 4.2: Yummly API Dashboard

4.1.1.2 BBC Food Data:

BBC Food provides different information about food like chef names, cuisines and other information. Yummly is recipe aggregator, sometime recipe published by some person have ingredients names as pronounced in their own

country. So, we standardized our dataset by crawling food information and replacing it with our existing dataset.

- We translated different multi lingual words existing our dataset using Google translate API.
- Then we mapped all possible words from BBC to Yummly to standardize this.

4.1.1.3 Country Health Stats:

To get different stats related to diabetes and IQ we crawled **World Bank** website. These stats then used to find correlation of diabetes and IQ with nutrition. They don't have the public API so we crawled the data by their permission (as per education or research usage)

4.1.2 Explanation

4.1.2.1 Cleaning:

Data cleaning is the most essential part of data analysis. Data cleaning is basically removing extra attribute or null values from data. While analyzing ingredient base similarity our data should be free from attributes like flavors, rating and URLs. Figure 4.3 shows a cleaned example.

```
[u'water', u'milk', u'cinnamon sticks', u'cardamom pods', u'sugar', u'coffee']
[u'fresh mint', u'green chilies', u'garlic', u'salt', u'low-fat yogurt']
[u'vegetable oil', u'cumin seed', u'salt', u'chili powder', u'lemon pepper', u't
omatoes', u'garbanzo beans', u'lemon juice', u'onions']
```

Figure 1.3 Ingredients from collection of Pakistani cuisine

4.1.2.2 Ingredient Base Similarity:

It is basically the measure of similarity between cuisines on the basis of ingredients used in them. We have collections of cuisines and each cuisine containing around 250 recipes. We convert each cuisine in vector for where each element of vector shows an ingredient. This made each cuisine ingredient based feature vector. Which we used to find similarity in each cuisine. To apply machine learning and statistical analysis algorithm we have to convert each ingredient word into number or "Boolean beg of words". This is achieved by using TF-IDF approach.

- **TF-IDF**: Term frequency and Inverse document frequency shows how much a word has worth in a document or corpus. It is also known as weighting factor in text mining. It is basically the count of word in a corpus means it increases proportionally as word appear in document.

$$TF(t) = \left(\frac{\textbf{Num of times term t appear in doc}}{\textbf{Total num of term in doc}}\right)$$

Term frequency is just a count, the worth of a word is determined by IDF. It tells how much a word is important in document. Words like is, of and that may appear a lot of time but they don't have much importance. So, we scaled up the importance of those words that are not appearing most frequently in document.

$$IDF(t) = \log e\left(\frac{\textbf{Total number of doc}}{\textbf{Num of doc with term t in it}}\right)$$

To measure similarity, we used a simple statistical algorithm known as **Cosine Similarity.** It is the angle measurement between two cuisine TF-IDF vectors. This angle shows how much one vector is shadowing on second vector. If angle in between two vectors is zero, it means they have full similarity, if angle is 90 degree it means zero similarity and if angle is 180 degree it means both are opposite in nature. Figure 4.4 shows cosine similarity of vector A with vector B.

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|_2 \|\mathbf{B}\|_2} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

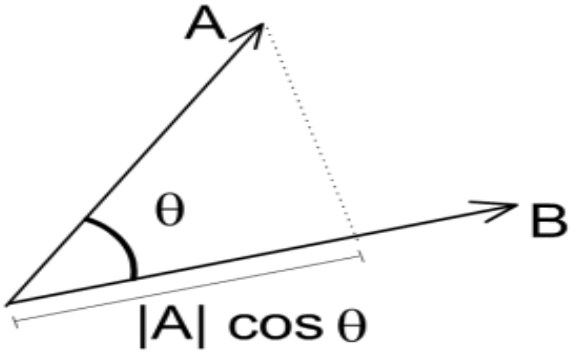


Figure 4.4 Cosine Similarity

4.1.2.3 Average Cooking Time:

Cooking time is the total time consumes to prepare a recipe. We took the mean of cooking time of all recipes in a cuisine to calculate the average cooking time of all cuisine respectively. This cooking time shows us different trends example cooking time in countries of Europe lies on average cooking time of world. Reason is atmospheric temperature of these countries is low that increase cooking time of food and it put this

food in above average category. People in these countries like to eat food that is cooked mildly (Half cooked) this compensate the time and put them on Average category.

$$\textbf{\textit{Average Cooking Time}} = \frac{\textbf{\textit{Sum of Time of Recipes in cuisine}}}{\textbf{\textit{Total number of Recipe in cuisine}}}$$

4.2 Music

As music is now our part of lives and it affects society in various ways. It is important for us to understand how it has changed to satisfy the desires of listeners, and how society has changed music.

We are finding the average duration, popularity, tempo, loudness of songs distributed across years. We also found correlation of duration, popularity of artist, tempo and loudness with the overall popularity of song. Data is accessed using file handling, results are computed using statistical approaches of average and correlation. Then data is fed into mongodb, which then is represented in graphical form on web site. As representing it was a very critical task, hence appropriate visualization tools were used for better understanding of the end users.

4.2.1 Data

Million Song Dataset was used for dataset, it is freely available. Following are the statistics of dataset:

1. 1,000,000 songs / files
2. 273 GB of data
3. 44,745 unique artists
4. 7,643 unique terms (The Echo Nest tags)
5. 2,321 unique musicbrainz tags
6. 43,943 artists with at least one term
7. 2,201,916 asymmetric similarity relationships
8. 515,576 dated tracks starting from 1922
9. 18,196 cover songs identified

Each file contains different fields, each field is a feature extracted from audio files. Figure 4.5 shows an example of what is contained in each file.

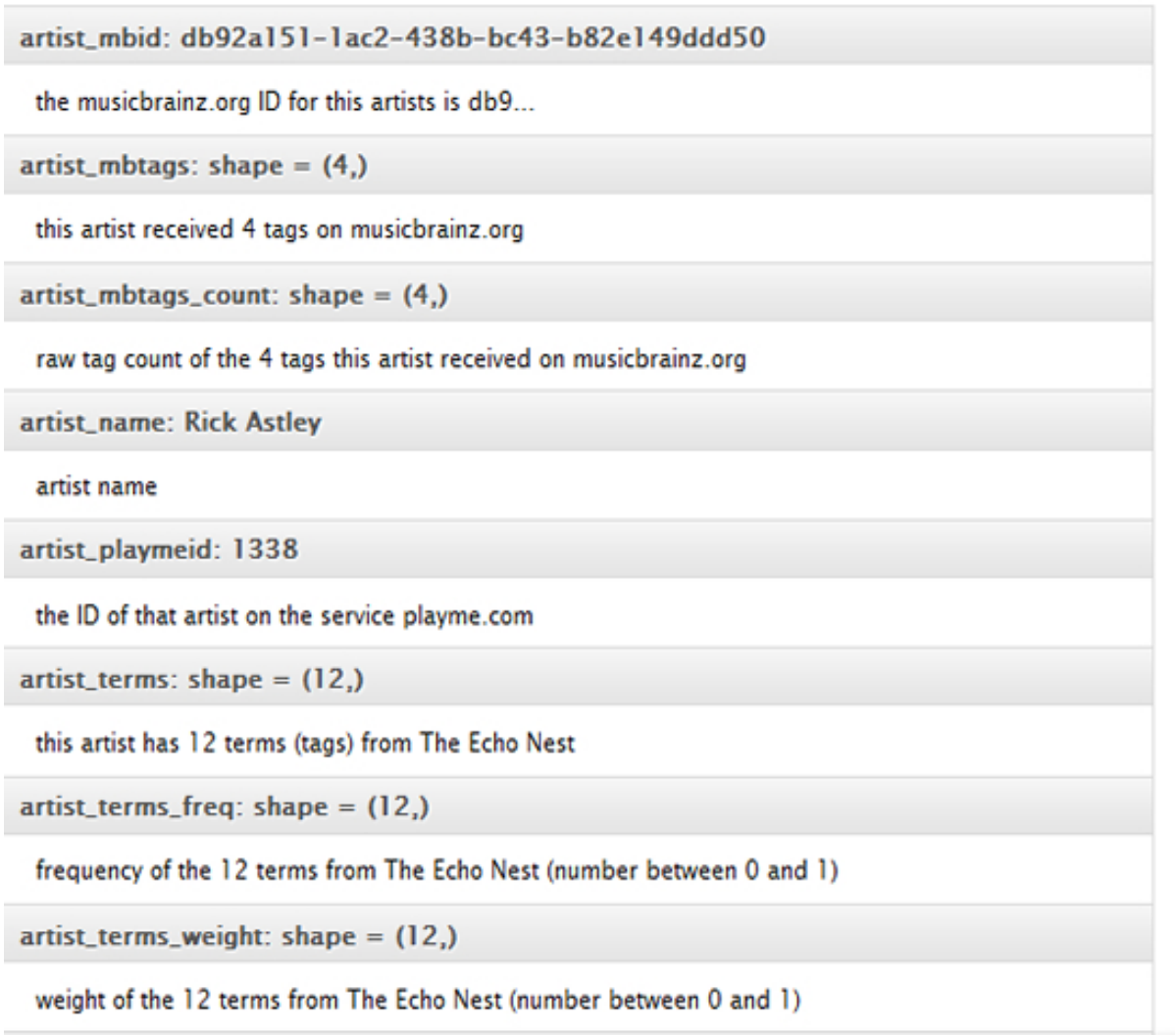


Figure 4.5 Music dataset sample

Data set is accessed using python, "os" is the library which is used to navigate through files. Results returned are stored in list and then computations are performed on it.

4.2.2 Explanation

4.2.2.1 Cleaning of Dataset:

Dataset contains many fields some of which are irrelevant for computations, hence only desired feature are taken into account. Also, dataset contains 0 values which means the field was not calculated, it also have few NaN (Not A Number), which can affect the results tremendously. Hence these undesired values were cleaned before the computations.

4.2.2.2 Applying Statistical Techniques:

Now, statistical techniques that are mean and correlation were applied on the dataset, which after intense computation produced results. Mean was used to find average duration, popularity of songs, loudness, and tempo across years. Correlation was found

among artist popularity, loudness, duration and tempo with overall popularity of that song. These results were the end values and were ready to be stored in database. Results were stored in a dictionary. Figure 4.6 is an example of result.

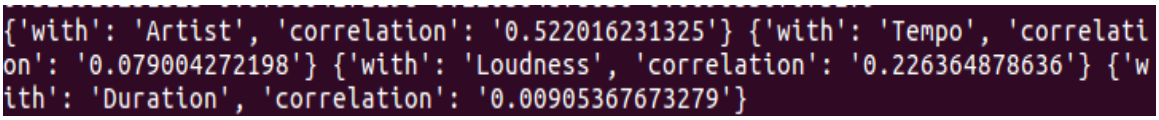


Figure 4.6 Music dataset sample

4.3 Sentiment Analysis

As the trend of microblogging site continues, we here develop a generic algorithm that calculates the sentiments of tweets posted by users in real time. This helps our users to study how positive or negative population of a certain country is feeling about a particular subject around the world.

Following are the steps used to perform sentiment analysis, they will also be explained later:

1. Establishing connection to MongoDB
2. Continuous real-time data acquisition using TweePy
3. Cleaning of tweets using regex
4. Calculating sentiment using VADER
5. Calculating the aggregate sentiment value on the basis of time zone
6. Updating the results to database

4.3.1 Data

Tweet acquisition was performed using python's library of PyMongo, it enables us to make a connection to twitter's streaming API. It can constantly pull data in real-time. It uses OAuth 2.0 to make a secure connection to the service. Each OAuth is backed up by a unique consumer secret and key allotted by twitter. The data that is streamed through this API is in JSON. It contains tweet as well as user information as well. Some of the fields are user's id, location, tweet's id, text, geo location, time of tweet, time zone, retweet count. Figure 4.7 shows an example. At one time only a single stream can be opened and results can be filtered out based on a keyword.

```
{
    "created_at": "Tue Dec 20 18:14:22 +0000 2016",
    "id": 811273554511691800,
    "text": "That's racist, said the privileged white person who's never experienced racism in their entire life.",
    "user": {
        "id": 43417981,
        "location": "WV",
        "followers_count": 8732,
        "friends_count": 6769,
        "listed_count": 126,
        "favourites_count": 79615,
        "statuses_count": 46954,
        "created_at": "Fri May 29 21:51:28 +0000 2009",
        "utc_offset": -14400,
        "time_zone": "Atlantic Time (Canada)",
        "geo_enabled": true,
        "lang": "en",
    },
    "retweet_count": 0,
    "favorite_count": 0,
    "favorited": false,
    "retweeted": false,
    "filter_level": "low",
    "lang": "en",
    "timestamp_ms": "1482257662786"
}
```


Figure 4.7 Tweet sample with limited fields

4.3.2 Explanation:

4.3.2.1 Cleaning:

Cleaning of tweets is very important because most of the tweets contains information like urls, punctuations, stop words like ‘a’, ‘the’, ‘and’ etc. and tags like RT, mentions and numbers that are not at all useful for determining the sentiments and causes huge performance hit while using machine learning based approach as well as lexicon based. We use python’s regex library to limit them. But it is also very important that we do not remove information like smileys that may determine mood of the user. Figure 4.8 shows snippet of algorithm used for cleaning of tweets.

```
#Cleaning Tweets
cl=' '.join(re.sub("(\w+:\/\/\S+)"," ",tweet['text']).split())
c2=' '.join(re.sub("(@[A-Za-z0-9_:]+)|([^0-9A-Za-z\'\"\? \t])"," ",cl).split())
c3=' '.join(re.sub("(RT)"," ",c2).split())
```

Figure 4.8 Some of the cleaning methods used

4.3.2.2 Sentiment Calculation:

For sentiment calculation, initially we used Naïve Bayes Classifier on a dataset provided by Kaggle, Naïve Bayes classifier is known to be the least resource intensive while giving nearly the best results when it comes to sentiment analysis, coupled this classifier with n-grams of 2 or 3 can make it as capable as human classification of positive negative and sarcasm. It calculates the probability of a sentence to know where it lies as positive, negative or neutral using the training dataset it was provided.

Now the problem associated with this classification was due to the dataset, a classifier is only good for what it is trained for, a classifier trained for calculating political sentiments drops its accuracy in case of subject change, like sports. Our requirement was to make a generic classification script that can give accurate classification regardless of the subject.

The solution to this problem was to use a dictionary based classification, a dictionary has a set of words that are rated for its polarity of how positive or negative a word is. Sentiment of every word is calculated that exists in the dictionary and then the overall sentiment is calculated. Some of the most popular dictionaries are SentiWordNet **[8]** and AFINN **[9].** Former has a 2477 set of words and phrases labelled by Finn Nielsen during 2009-2011 while later has the support for positivity, negativity and subjectivity.

VADER stands for **V**alence **A**ware **D**ictionary and s**E**ntiment **R**easoner, is a lexicon and rule based sentiment analysis library. It has, 9000 lexicons rated by 10 individuals. They are not only polarity rated but also intensity rated as well, as shown in figure 4.9

```
A really bad, horrible book.
    {'neg': 0.791, 'neu': 0.209, 'pos': 0.0, 'compound': -0.8211}
At least it isn't a horrible book.
    {'neg': 0.0, 'neu': 0.637, 'pos': 0.363, 'compound': 0.431}
```

Figure 4.9 Sample Output of VADER

The dictionary includes emoticons like ":(", slang words like "duh" and acronyms like 'LOL', 'OMG' etc.

	Correlation to ground truth (mean of 20 human raters)	3-class (positive, negative, neutral) Classification Accuracy Metrics		
		Overall Precision	Overall Recall	Overall F1 score
Social Media Text (4,200 Tweets)				
Ind. Humans	**0.888**	0.95	0.76	0.84
VADER	0.881	**0.99**	**0.94**	**0.96**
Hu-Liu04	0.756	0.94	0.66	0.77
SCN	0.568	0.81	0.75	0.75
GI	0.580	0.84	0.58	0.69
SWN	0.488	0.75	0.62	0.67
LIWC	0.622	0.94	0.48	0.63
ANEW	0.492	0.83	0.48	0.60
WSD	0.438	0.70	0.49	0.56

Figure 4.10 Performance comparison of VADER [10]

Tried and tested various methods using machine learning and lexicon based approach, we came to a conclusion that VADER is 'the' most efficient and gives the most accurate classifications for micro blogging sites like twitter, figure 4.10 shows the performance comparison of VADER with other classification algorithms.

Negation handling is also a part of VADER, it can detect when to invert the polarity of a subject as we know that 'not good' is negative while 'good' is positive. Apart from English language, every language is supported as long as we are providing it the dictionary of that particular language. Figure 4.11 shows negation handling of VADER.

```
NEGATE = \
["aint", "arent", "cannot", "cant", "couldnt", "darent", "didnt", "doesnt",
 "ain't", "aren't", "can't", "couldn't", "daren't", "didn't", "doesn't",
 "dont", "hadnt", "hasnt", "havent", "isnt", "mightnt", "mustnt", "neither",
 "don't", "hadn't", "hasn't", "haven't", "isn't", "mightn't", "mustn't",
 "neednt", "needn't", "never", "none", "nope", "nor", "not", "nothing", "nowhere",
 "oughtnt", "shant", "shouldnt", "uhuh", "wasnt", "werent",
 "oughtn't", "shan't", "shouldn't", "uh-uh", "wasn't", "weren't",
 "without", "wont", "wouldnt", "won't", "wouldn't", "rarely", "seldom", "despite"]
```

Figure 4.11 Negation List

4.3.2.3 Aggregate Sentiment:

The method above provides instantaneous sentiments of real time tweets, but our objective is to aggregate the sentiments of tweets of a particular subject over a specific period of time. For this purpose aggregate sentiment is used.
The sentiment value at each point gets stored in database, while our node JS server side script updates our front end by querying it with regular interval to stay updated. In the next iteration our python script calculates the sentiment of a newly streamed tweet and aggregates it with the previous one based on its time zone. Figure 4.12 shows the aggregation of sentiments.

```
prev = collection.find_one({"timezone":tweet['user']['time_zone'],"topic":keyword_list},
                           {'sentiment':1,'count':1,'_id':0})

if prev == None:
    agSent = sentiment['compound']
else:
    #print "Previous: ",prev
    agSent = (sentiment['compound']+prev['count']*prev['sentiment'])/(prev['count']+1)
```

Figure 4.12 Aggregate Sentiment

4.4 Web Application Development

Our project heavily relies on the visualization of results, without proper visualization one cannot deduce results based on our findings, hence here is our approach for a fast, interactive, clean and easy to use GUI dashboard.

Following are the parts of web dev. Process:

4.4.1 Server-side Scripting

Web application have two sides, server and client. In this section we will discuss the server side scripting in detail. The part of script that runs on web server and is capable to facilitate the transfer of data from the web server to a web browser is called server side script. All the database connection and serving of queries and request is also a part of this script. These scripts can be written in variety of language like PHP, JavaScript, Java and Python etc. PHP being the most popular one while JavaScript is also emerging really fast as discussed later.

4.4.1.1 Node JS

Node JS is a JavaScript built on Google Chrome's V8 JS Engine. Its event driven and non-Blocking I/O model makes it light weight, efficient and exceptionally suitable of real time applications which can serve many connections at once compared to PHP.

Following are the characteristics of Node JS that makes it unique and suitable to our project.

I. It is Asynchronous:
Asynchronous means not waiting for an operation to finish. Node JS uses several threads but only one thread is dedicated for execution. This means that not all the code needs to occur at the same time as we know that different tasks can take different time, therefore this resolves the need for doing all of them at the same time.

II. It uses non-blocking I/O:
This means that code can run while the background threads are blocked by the I/O while waiting for the data. Figure 4.13 shows event driven nature.

III. Node Package Manager:
Node comes with NPM, this enables developers to easily and efficiently install libraries/dependencies or packages, it comes with its own console and server.

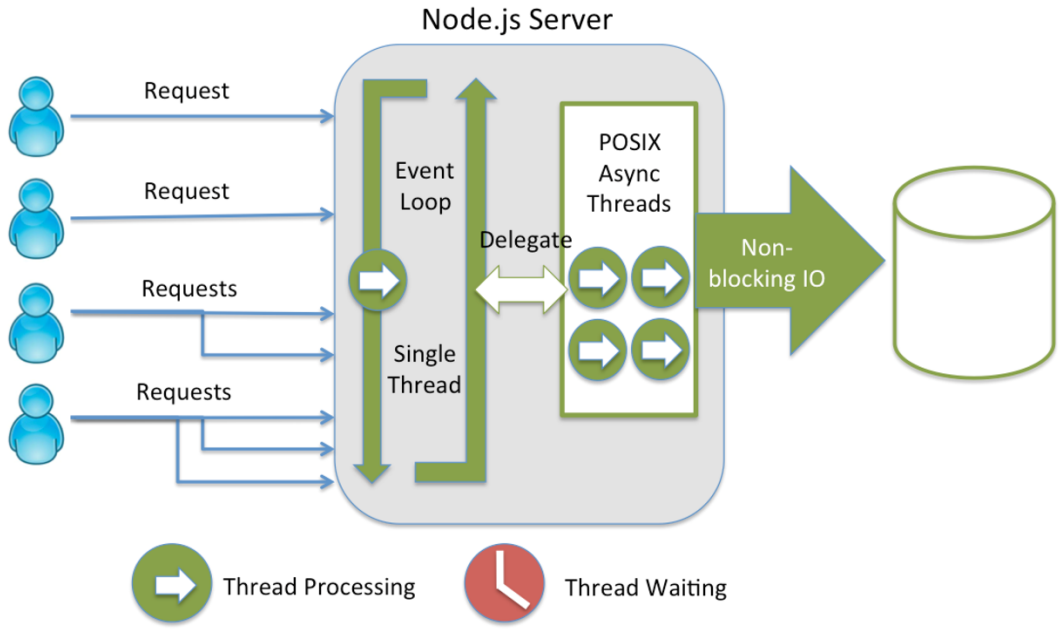


Figure 4.13 Event driven nature of Node JS

Following are the tasks our Node JS script is performing:

- Serving all "GET" requests that our client makes.
- Serving all "POST" requests made by client.
- Maintaining mongoDB local and cloud connections.
- Using packages like express and mongoDB driver.
- Express is used for handling all the routes and listening to ports.
- Handling all the errors and exceptions.

4.4.1.2 MongoDB as a Database:

For storing results and processing, we are using mongoDB. MongoDB is a noSQL database which means that it is document oriented rather than traditional Relational DBMSs.

As we know that every data transaction and communication over the internet is in JSON, JavaScript Object Notation, this makes it very useful to dump not only results but also data acquired from APIs like Yummly and Twitter. The response of these APIs is in JSON hence it can be easily parsed, processed and stored.

4.4.2 Client-side Scripting

The processing that takes place on the clients/users computer is known as client side scripting. Usually it runs scripts on a web browser sent from web server to clients computer over the internet. It may including HTMLs, Java Scripts and visualization libraries etc.

4.4.2.1 Chart JS for Visualization

Chart JS is an open source visualization library used for creating charts. It is built on html5 canvas element and is very interactive, clean, responsive and easy to use. It has many types of charts like line, bar, area, polar and radar charts etc. Figure 4.14 shows example of charts.

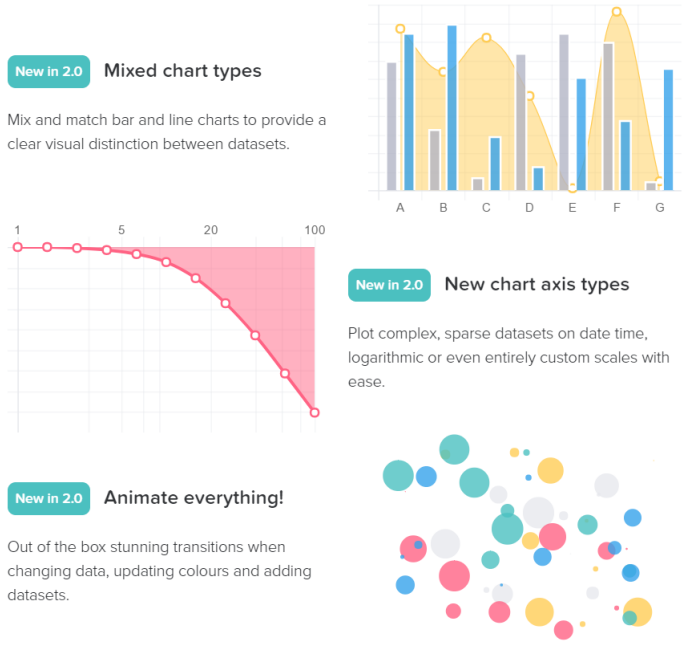


Figure 4.14 Chart JS visualization example

4.4.3 Atom

Atom provides an environment to write our scripts, as shown in figure 4.15. Following are the features that makes it the most suitable for our use case:

- It is cross-platform as it is a desktop application that runs o Electron, a framework for building cross-platform apps.
- Built-in package manager, it makes it very easy to download and install open source packages like in our case: express, mongoDB native driver and bodyparser.
- It is open source and highly customizable to its core, from themes to any changes in GUI.

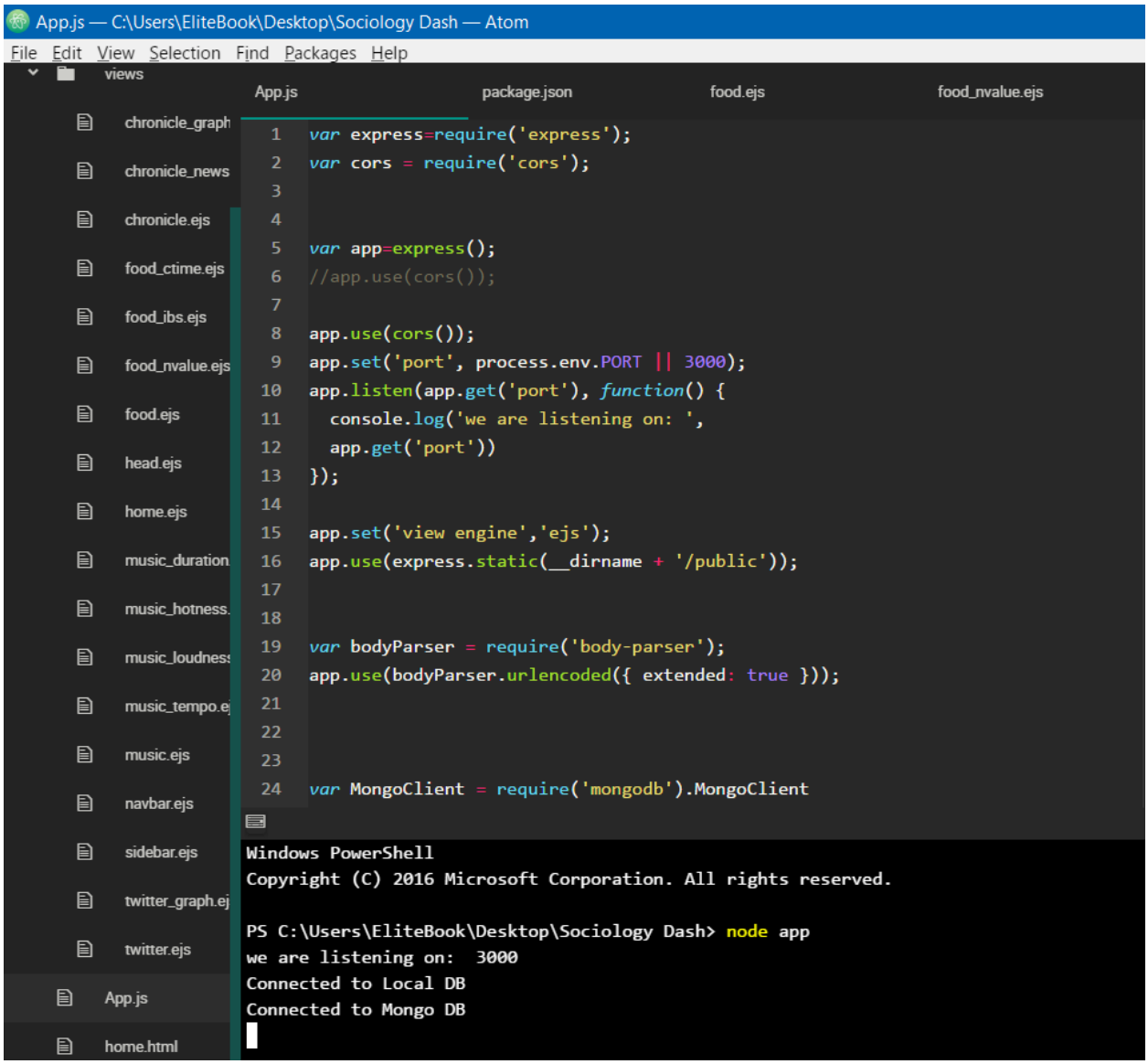


Figure 4.15 Node JS working environment

CHAPTER 5

RESULTS

5.1 Dashboard

Our dashboard built on html, CSS, bootstrap and JavaScript provides the most fluent and easy to use GUI. It is fast and elegant. As mentioned before, visualization in such projects is the most important part as without this users will not be able to understand that information. Figure 5.1 shows and example of a section of chronicle.

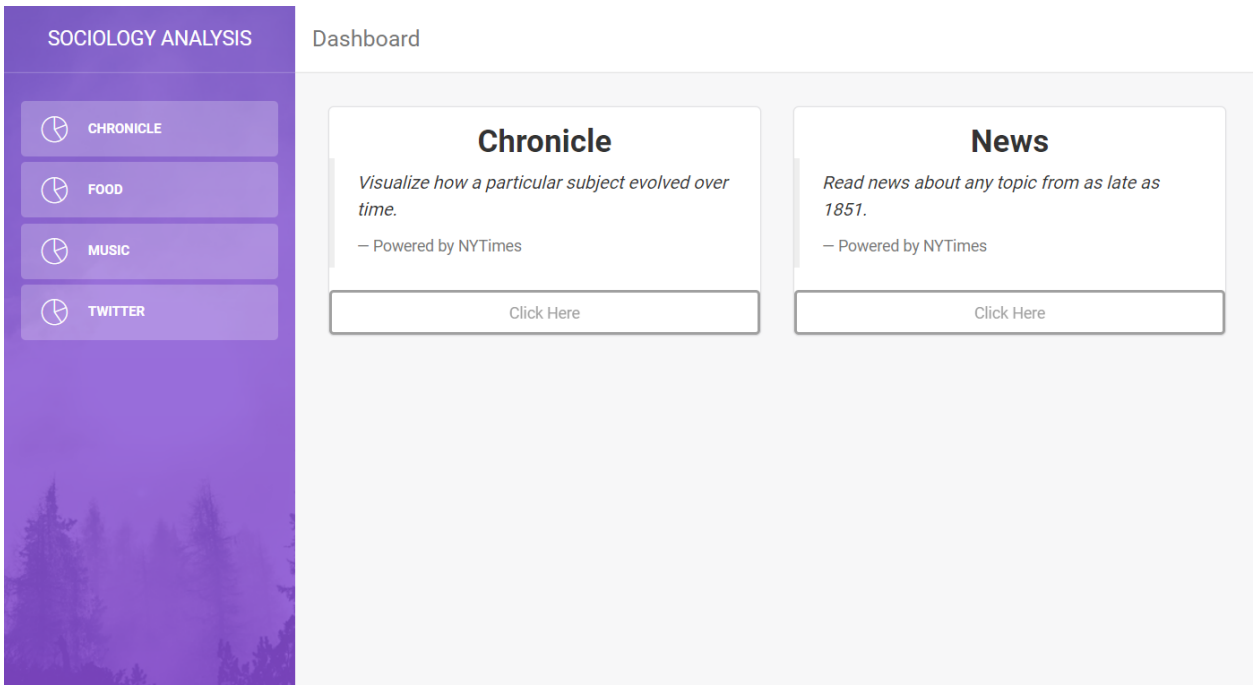


Figure 5.1 Dashboard featuring Chronicle section

5.2 Sentiment Analysis

As Obama is a controversial subject which generates both positive and negative sentiments across different countries as shown in figure 5.2, whereas racism is regarded as a negative subject, hence most of the results are negative which justifies the theory as shown in figure 5.3.

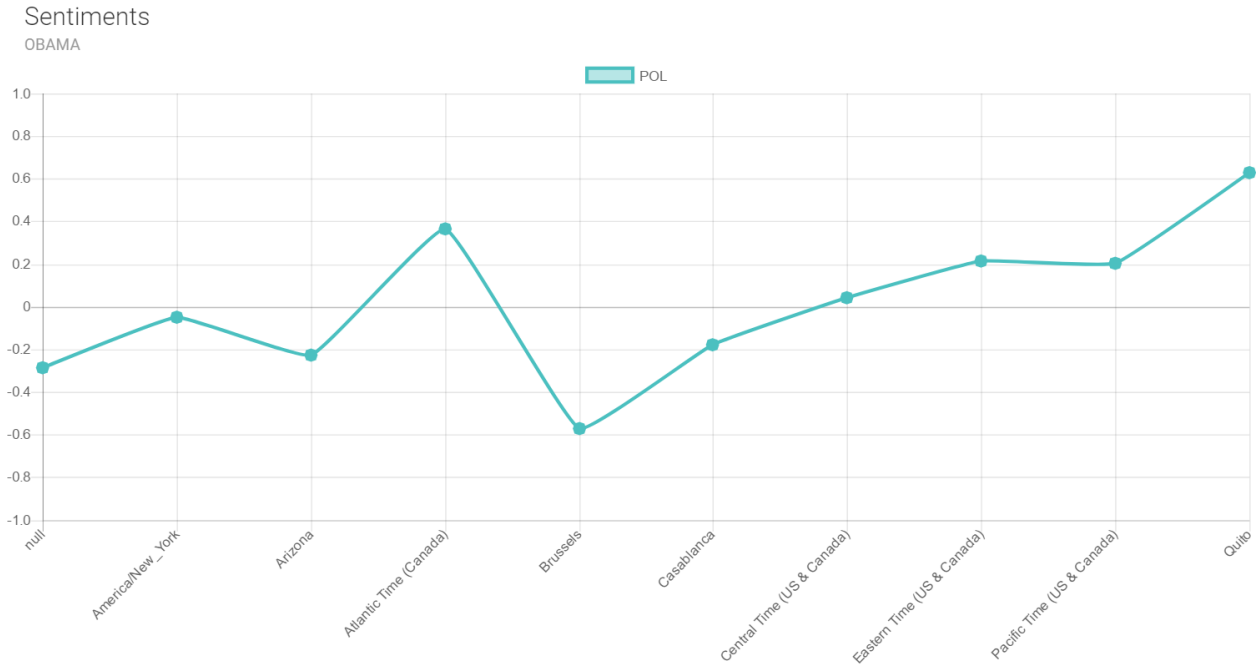


Figure 5.2 Sentiments for Obama (real time tweets gathered for 5mins)

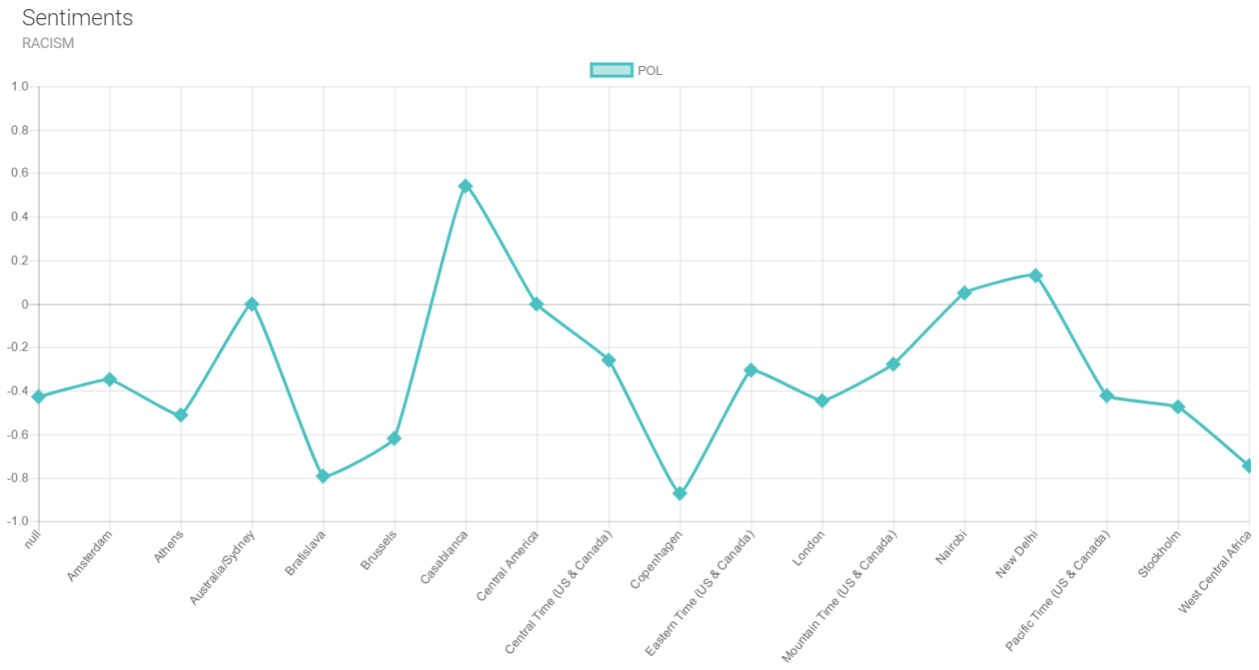
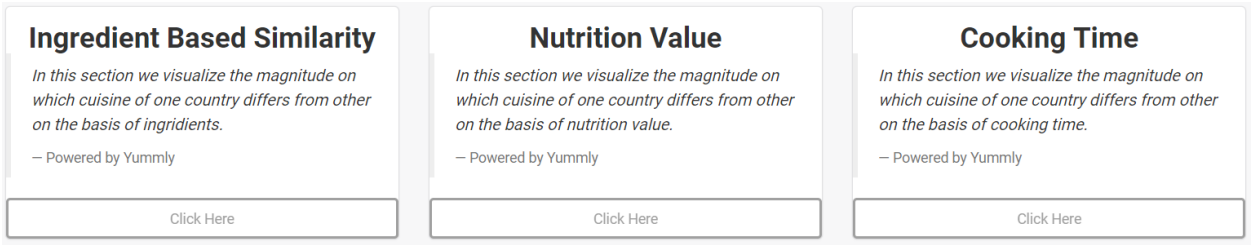


Figure 5.3 Sentiments for Racism (real time at the time of writing this)

5.3 Food



Detailed implementation of every step is discussed in detail in Chapter 4.

5.3.1 Average Nutrition Value

ANV is the average consumption/usage of a specific nutrition in cuisines from all over the world. We can visualize the healthiest as well as the unhealthiest cuisines on the bases of Proteins, Calorie, Carbohydrates, Sugar and Fats etc. as shown in figure 5.4.

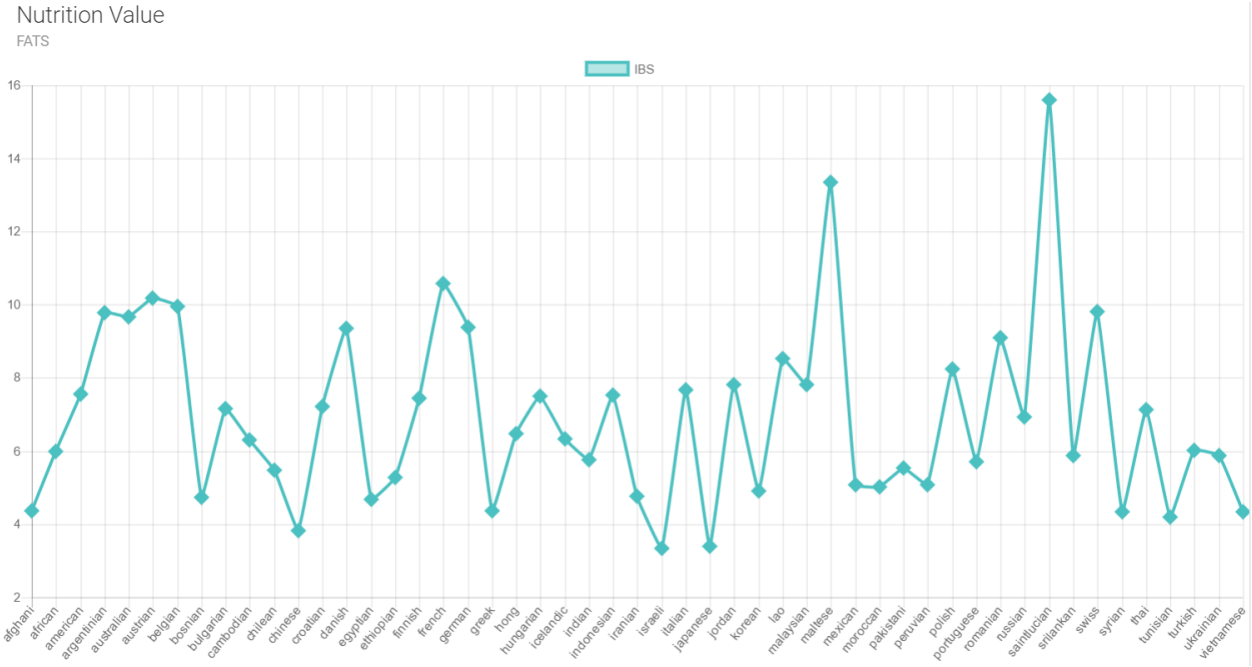


Figure 5.4 AVN of Fats among different cuisines

5.3.2 Average Cooking Time

Average cooking time differs among different cuisines, we had average cooking time of 84 cuisines as shown in figure 5.5. This is of most significance to small business owners to opt for those cuisines which have less cooking time to minimize their operational cost and time.

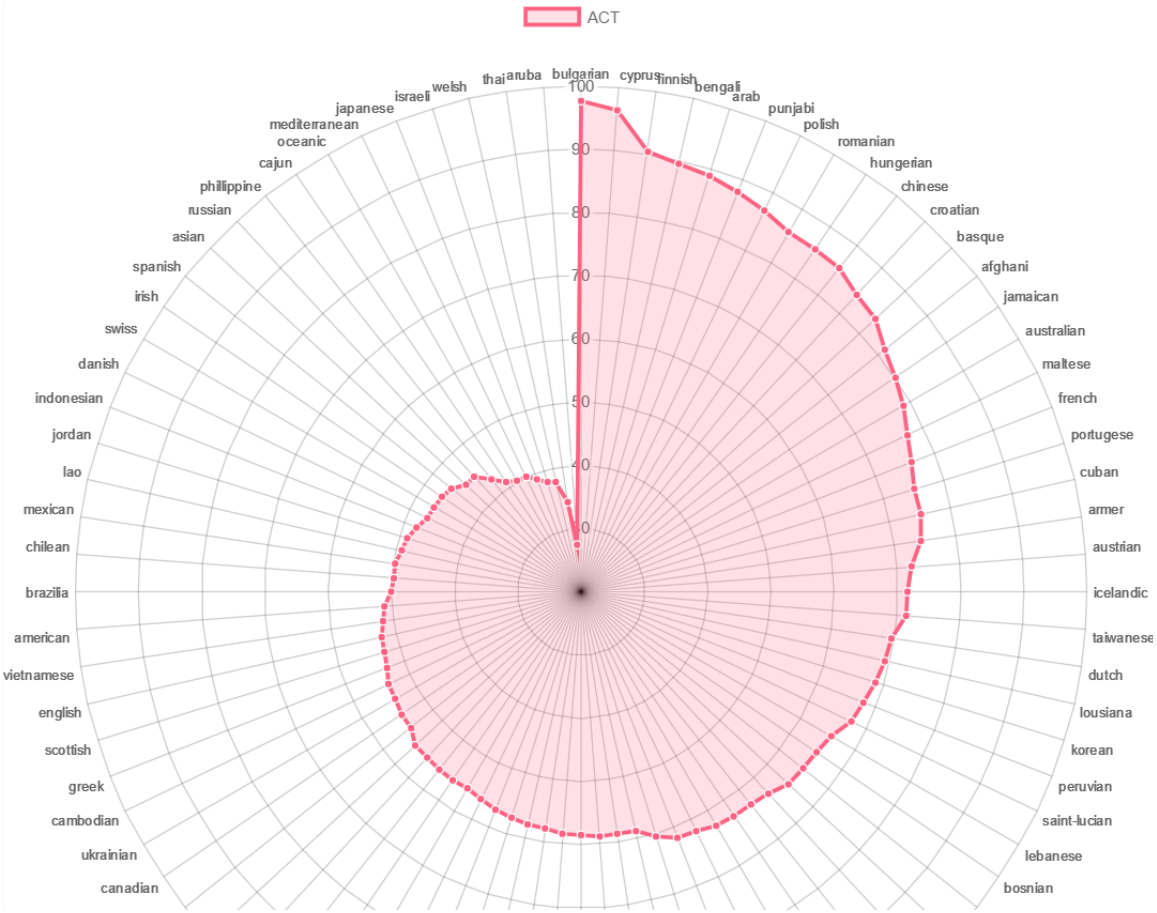


Figure 5.5 ACT of 84 cuisines

5.3.3 Ingredient Based Similarity

IBS of a country shows how similar a country is from other countries based on the similarity of ingredients used in the average of their all recipes. As shown in figure 5.6 and figure 5.7. It helps to determine the preference of other cuisines as alternates.

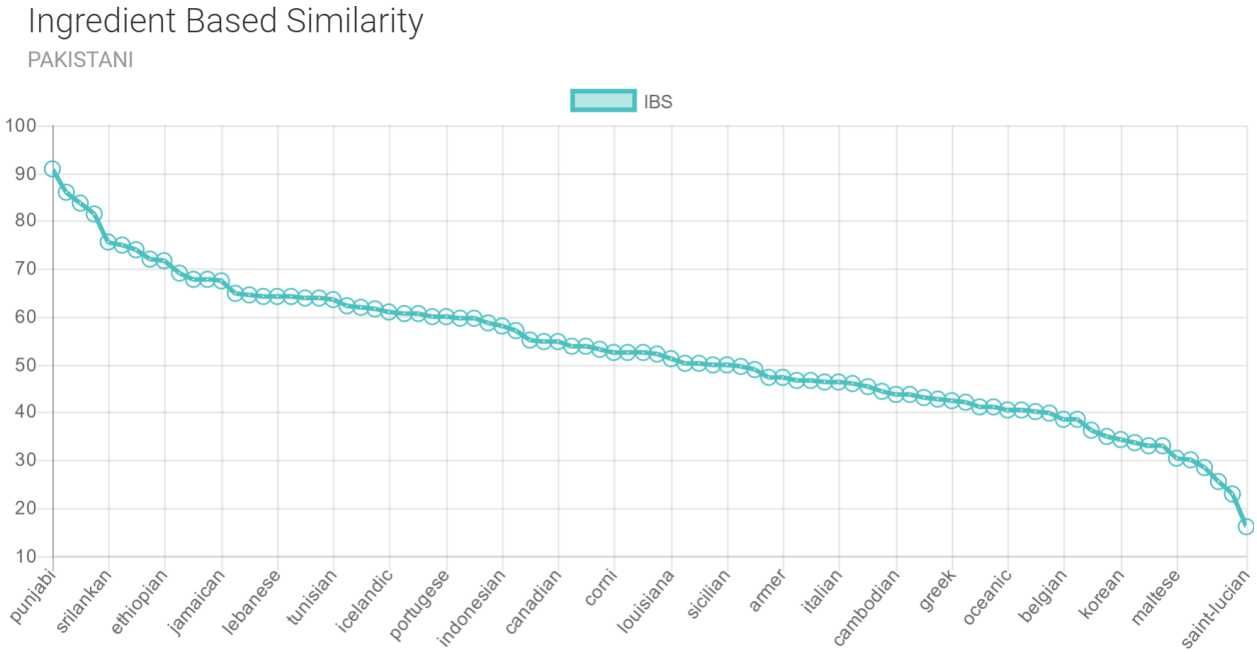


Figure 5.6 IBS of Pakistan from 86 countries

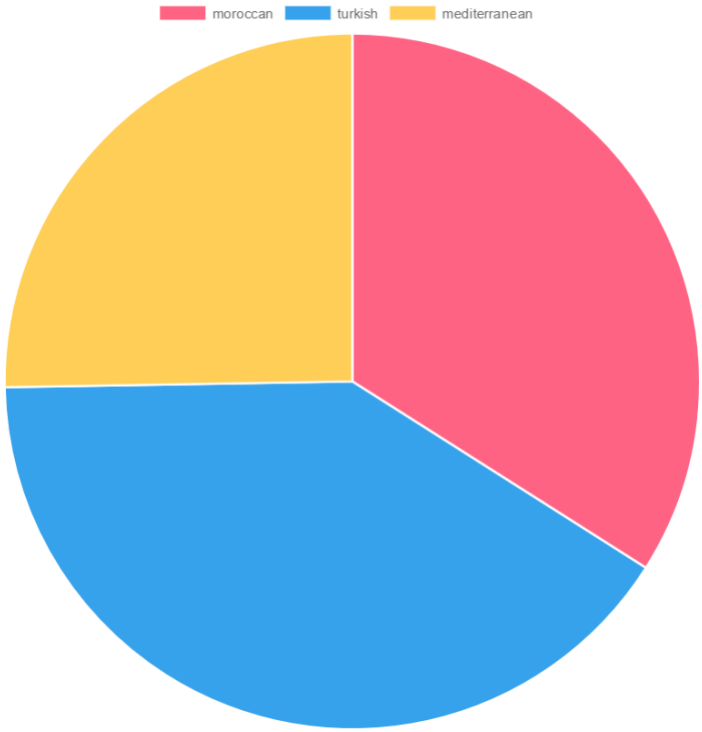


Figure 5.7 IBS of Pakistan from 3 random countries

5.3.4 Correlation of Diabetes &IQ with Nutrients

We have found a positive correlation of carbohydrates, sugar and calorie intake while a negative correlation of saturated fats, proteins and cholesterol with diabetes as shown in figure 5.8.

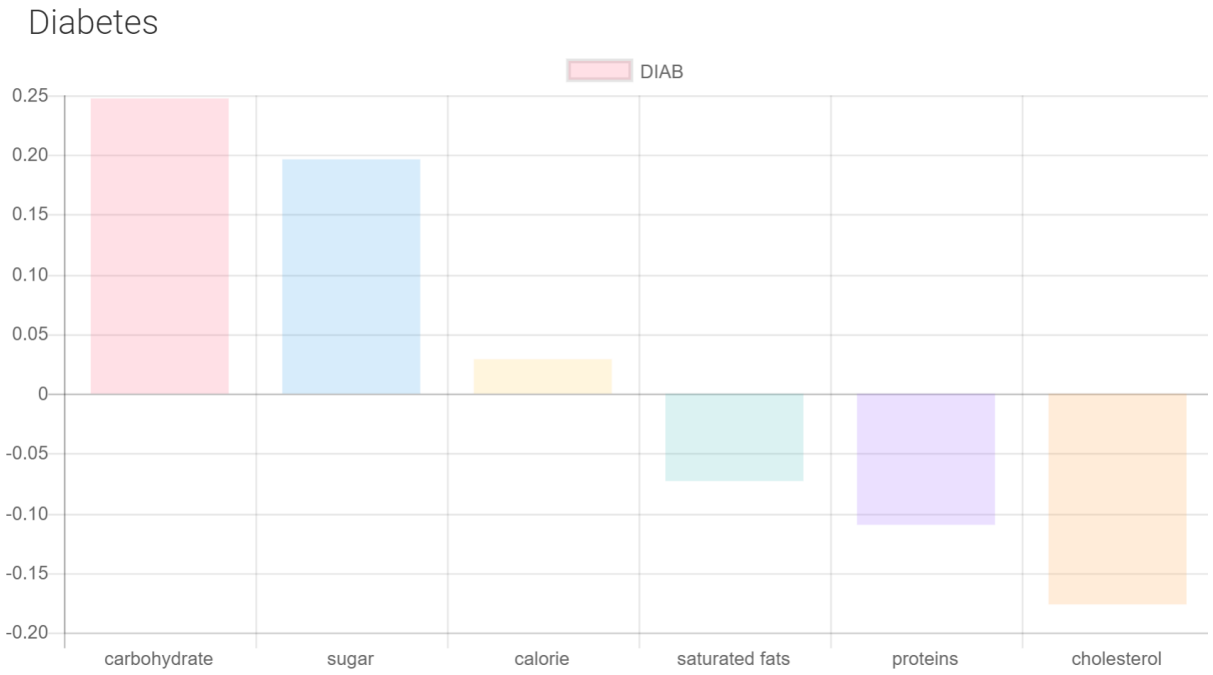


Figure 5.8 Correlation of Diabetes with Nutrients

Following correlation of six nutrients with IQ was found as shown in figure 5.9.

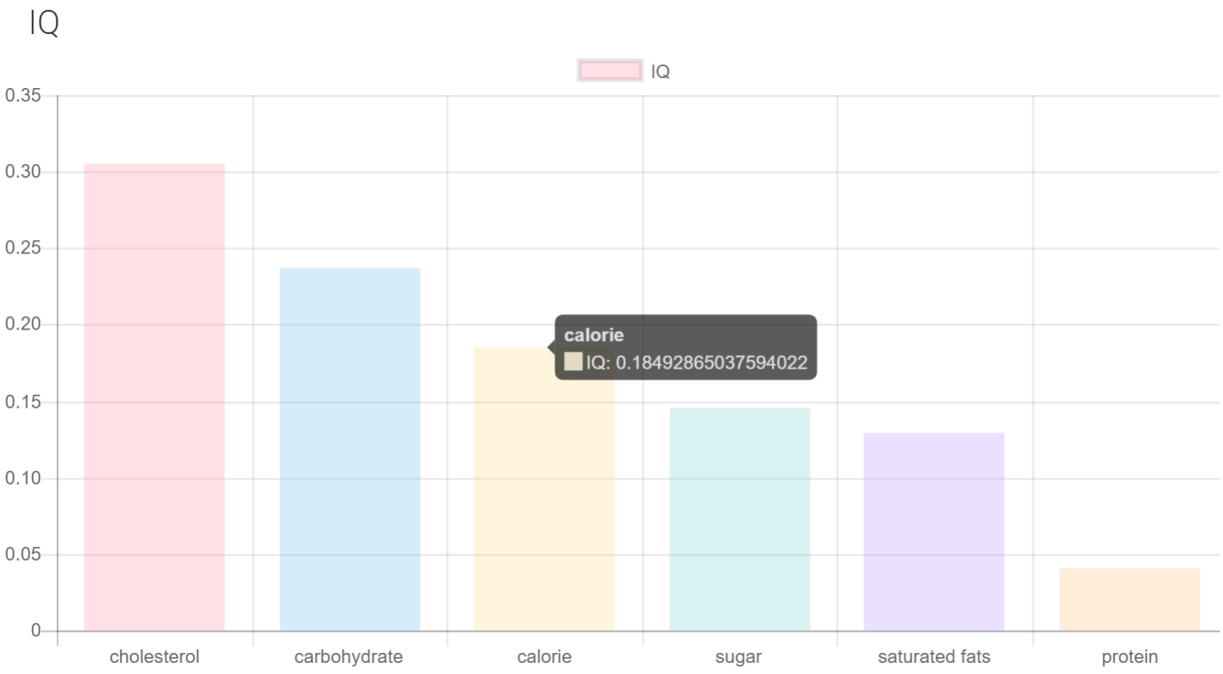
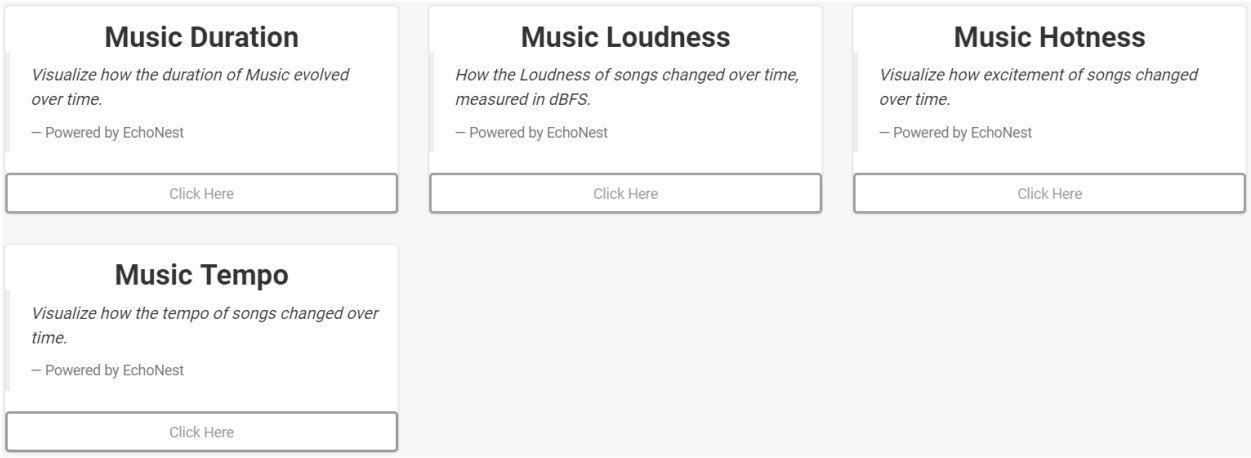


Figure 5.9 Correlation of IQ with Nutrients

5.4 Music



Detailed implementation of every step is discussed in detail in Chapter 4.

5.4.1 Hotness, Tempo, Loudness and Duration

We have found an overall increase in change of tempo which has evolved over 80 years as shown in figure 5.10

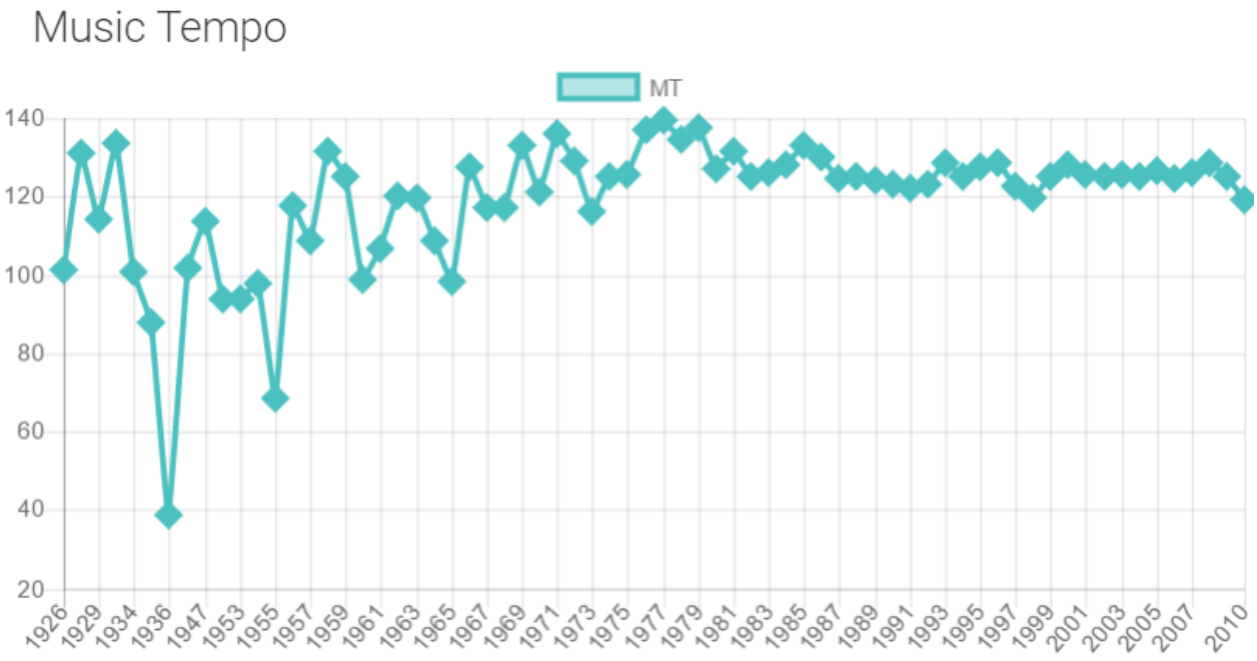


Figure 5.10 Change in Music Tempo since 1926

We have found a gradual increase in the duration of songs after 1960s as shown in figure 5.11

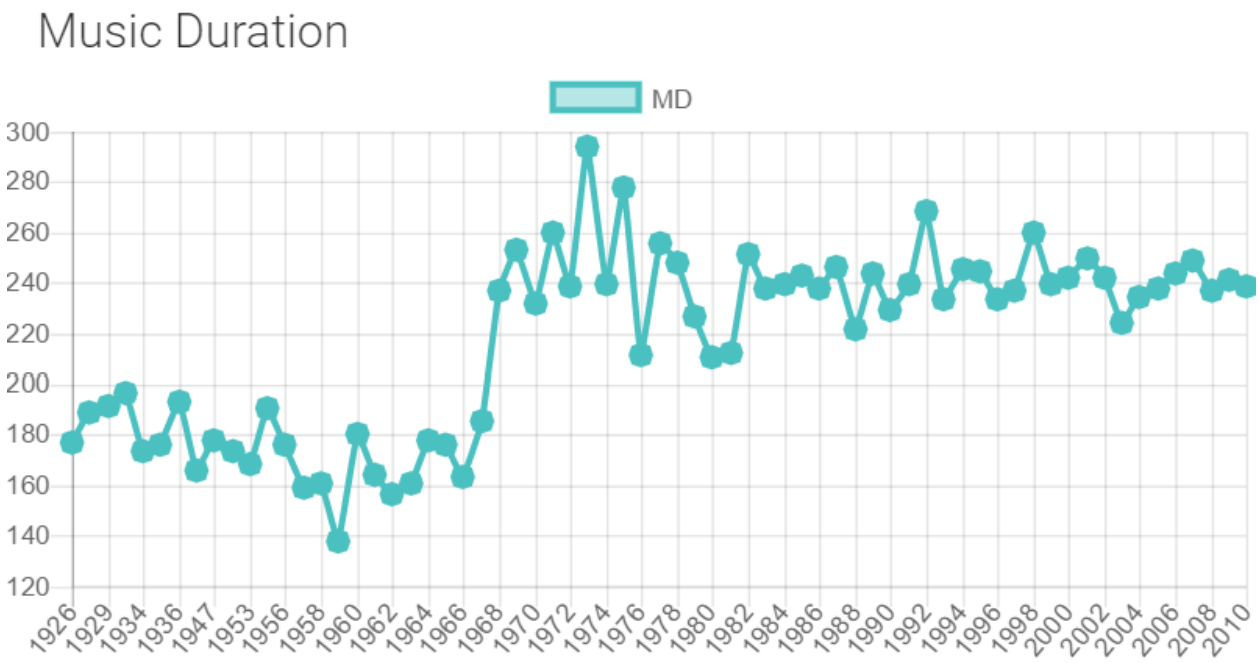
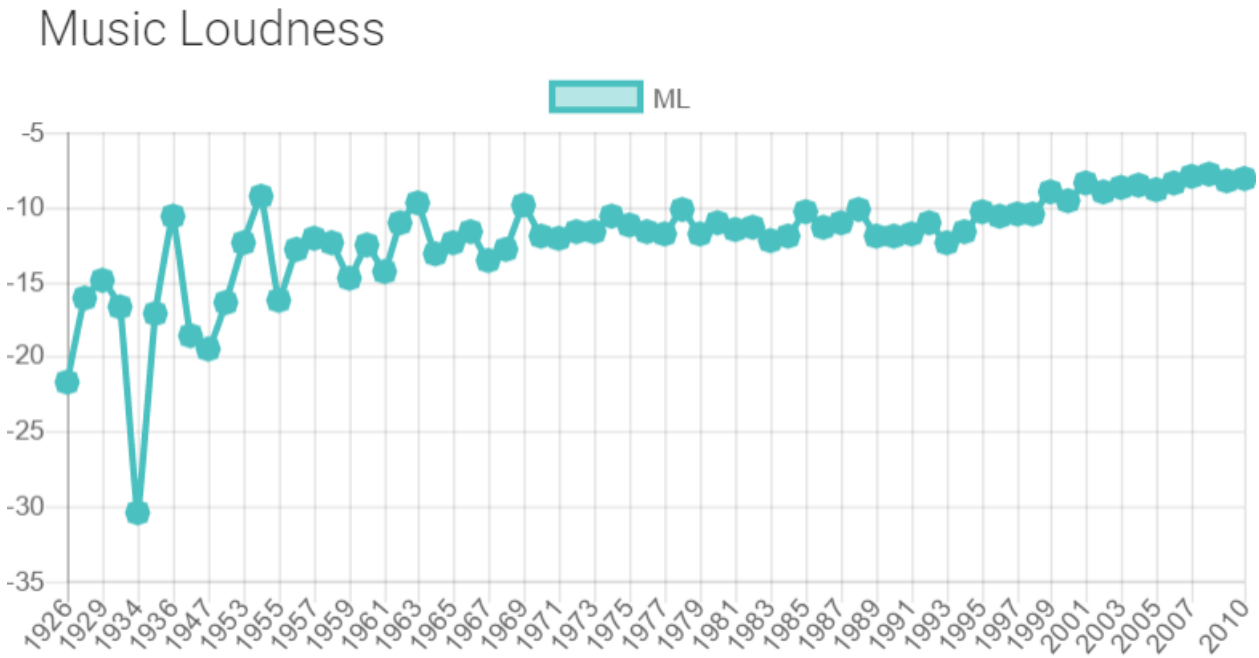


Figure 5.11 Change in Music Duration since 1926

We have found significant increase in music loudness since 1926 as shown in figure 5.12.



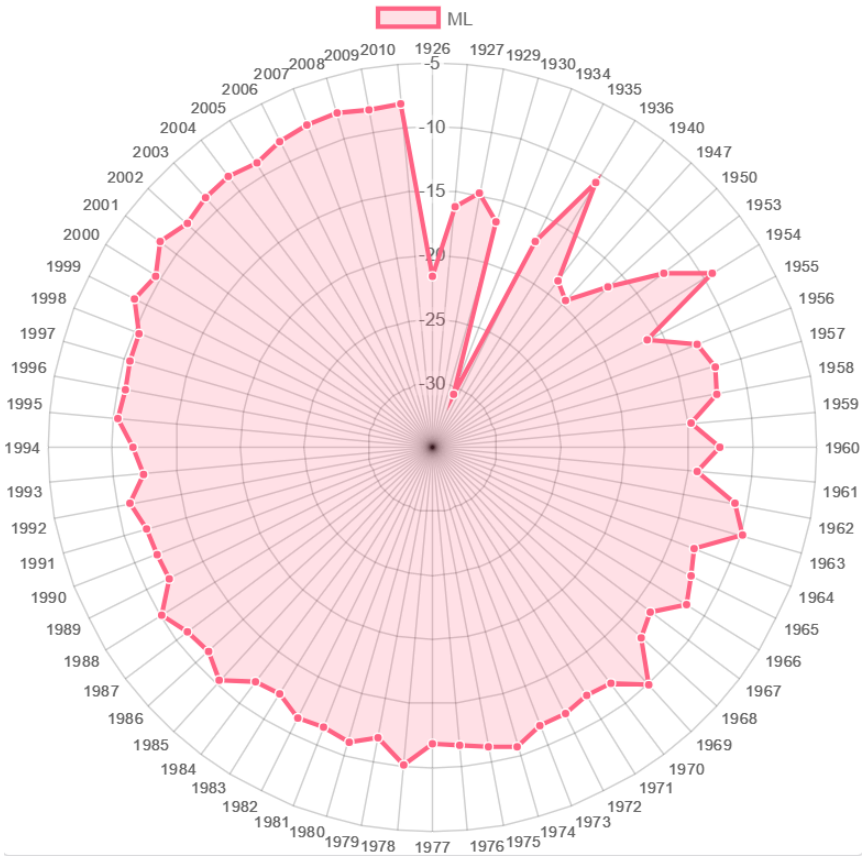


Figure 5.12 Change in Music Loudness in Db. since 1926

Strongest relation of song popularity is founded with the artist's popularity as show in figure 5.13.

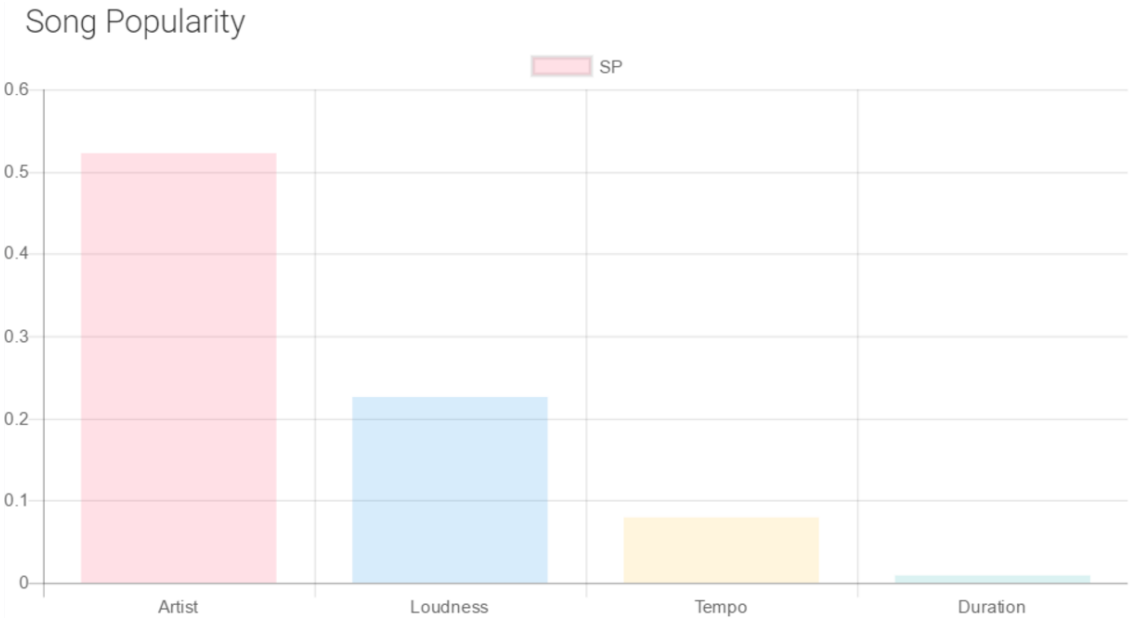


Figure 5.13 Correlation of Song popularity with different entities.

CHAPTER 6
CONCLUSION AND FUTURE WORK

6.1 Conclusion:

Our project is basically a data science based project that uses data to convert it into meaningful information and visualization. Data is increasing every second, hence sky is the limit. Sociology is a wide field, it covers the study of law, culture, traditions to law, secularization etc. hence many undiscovered areas can be contributed by further development.

News from around the globe is available and we can study the most frequent issues that appear regarding nations, we have already established New York Times as our global news provider as they have articles from as early as 1851.

One most interesting field of study is moods of a society, there are number of ways for determining the mood of a nation as a whole as shown in figure 6.1, but our study reflects the attraction towards a certain genre of movies and music can determine the mood at best. Some theories **[11]** show that humor is a born of depression and alienation from the general culture and people who are comedians or have love for comedy mostly come from a place of tragedy while people who are energetic and creative like fiction, patriotic people like war movies and people of heroic nature like superheroes. According to a study **[12]**, countries where heavy metal is popular are more content with life. Study **[13]** shows cultural analysis with respect to music.

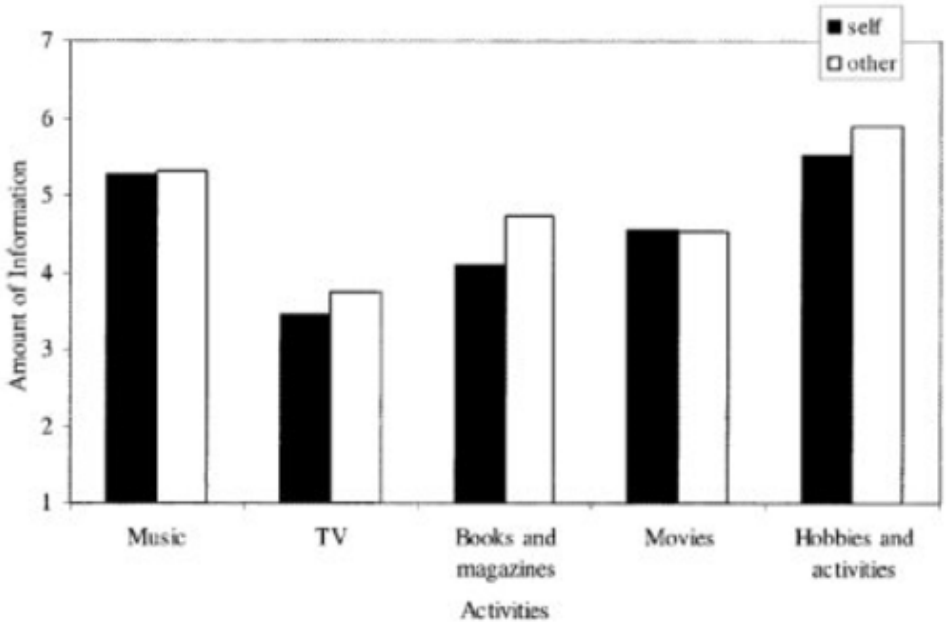


Figure 6.1 Amount of information various activities reveal about the personality of oneself.

6.2 Future Work:

Many of our goals are accomplished but there is always a room for improvements and innovations. Some of the aspects are there which could not be entertained due to time constraint. We propose to make a subsection in our application that can develop a correlation between two subjects on the go on basis of news and articles for example a correlation of Pakistan and independence can show the associated date of 1947 automatically to the user as 1947 must be the year at which both the words 'Pakistan' and 'independence' must co-exist.

We also want to categorize the mood of society based on the likings of their music and movies as discussed in previous section. Figure 6.2 shows our vision for future work.

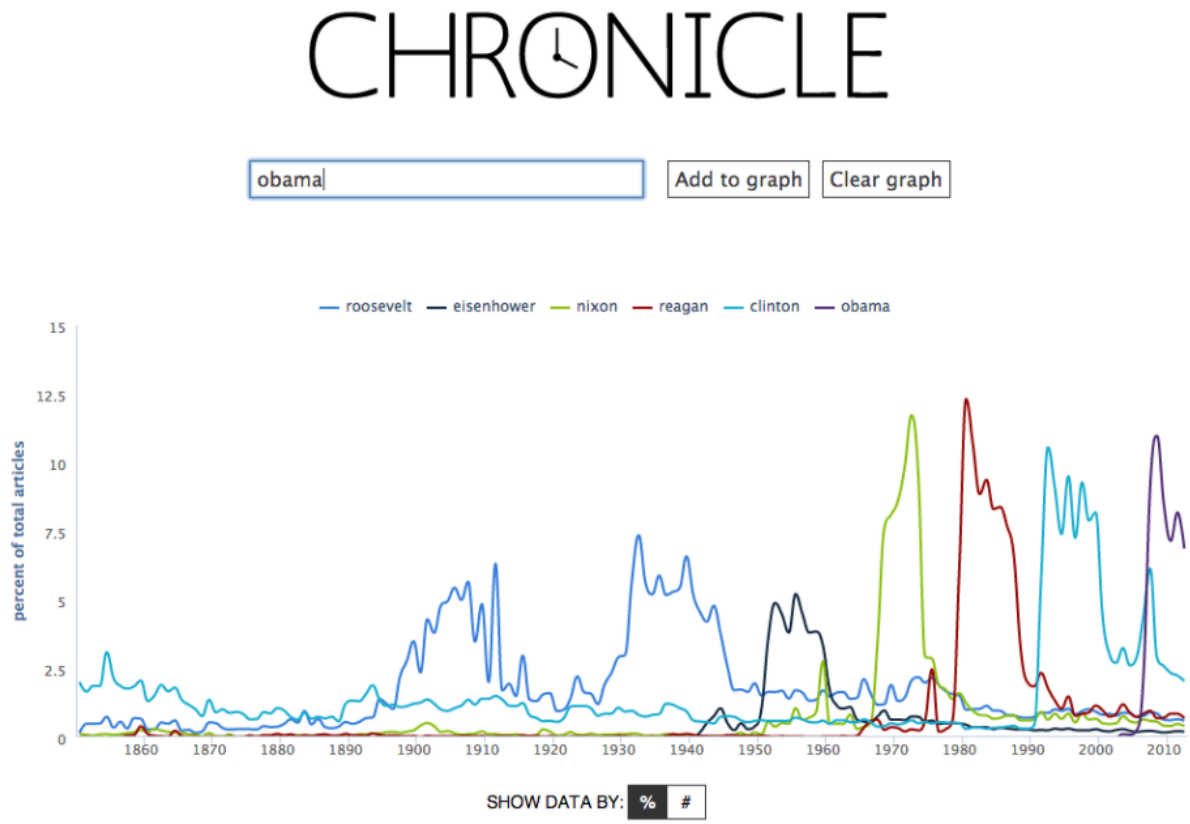


Figure 6.2 Chronicle

REFERENCES

[1] https://www.fastcompany.com/3066903/after-trump-and-brexit-data-scientists-have-one-main-job-in-2017

[2] https://www.16personalities.com/articles/music-preferences-by-personality-type

[3] "ISSUES, CHALLENGES, AND SOLUTIONS: BIG DATA MINING", Jaseena K.U. & Julie M. David

[4] http://research-paper.essayempire.com/sociology-research-paper/sociology-of-food-and-eating-research-paper/

[5] http://research-paper.essayempire.com/sociology-research-paper/sociology-of-music-research-paper/

[6] "REALIZING THE POTENTIAL OF DATA SCIENCE", National Science Foundation Computer and Information Science and Engineering Advisory Committee, Francine Berman and Rob Rutenbar, December 2016

[7] Dr. Kousar Jaha Begum, Dr. Azeez Ahmed "The Importance of Statistical Tools in Research Work", IJSIMR

[8] http://sentiwordnet.isti.cnr.it/

[9] http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010

[10] C.J. Hutto, Eric Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text", Georgia Institute of Technology, Atlanta

[11] http://time.com/3104938/depression-comedy-connection/

[12] https://consequenceofsound.net/2014/06/countries-where-heavy-metal-is-popular-are-more-wealthy-and-content-with-life-according-to-study/

[13] JARL A. AHLKVIST, "MUSIC AND CULTURAL ANALYSIS IN THE CLASSROOM: INTRODUCING SOCIOLOGY THROUGH HEAVY METAL", Johnson State College