



## **Twitter Health &Diet Analysis**

### **Phase 2 Report**

**Spring 2016**

- ❖ Gupta, Arunit – agp52
- ❖ Patel, Marmikkumar – mnpw3d
- ❖ Muktevi, Vamsi Krishna –vm4q8

## Table of Contents

1. Introduction .....	3
2. System setup.....	4
3. Design .....	5
4. Implementation .....	6
5. Testing.....	17
6. References.....	18

# 1. Introduction

To visualize the Heath patterns across the globe by analyzing twitter data at real time and answers user queries by extracting useful information. Our motive is to analyze unstructured data about health, food habits, hygiene, work outs which is collected from twitter tweets done by users across the world. We have collected more than hundred thousand of data and stored in JSON format, where we can read that data in a structured format which can be read by humans easily.

Our goal is to analyze what are the eating habits of people across the world, which food they are interested in most, how many people prefer vegan food. People interested in different cuisines in United States. Interest of people in street food in United States and other parts of the world.

Research about countries more specific to gym diet across the world to know how healthy people want to stay and focus on fitness, taking into consideration if they are tweeting, possibly they are following it to stay fit. People around are more focused on vegan diet, so we want to see which cuisine comes under the vegan diet. Different kinds of Indian food liked by people in states and across world.

Popular food on which retweet is done mostly or we can say which category is more followed by people and shared on social media.

## 2. System Setup

**Environment:** Windows 10

**IDE:** Eclipse Mars

**Browser:** Google Chrome

**Language:** Java

**Charts Reference:** High Charts, Amcharts, d3.js

**Software:** Apache Spark

**Web Technologies:** HTML, JS, JSP, Servlets

### 3. Design

#### The Spark Stack

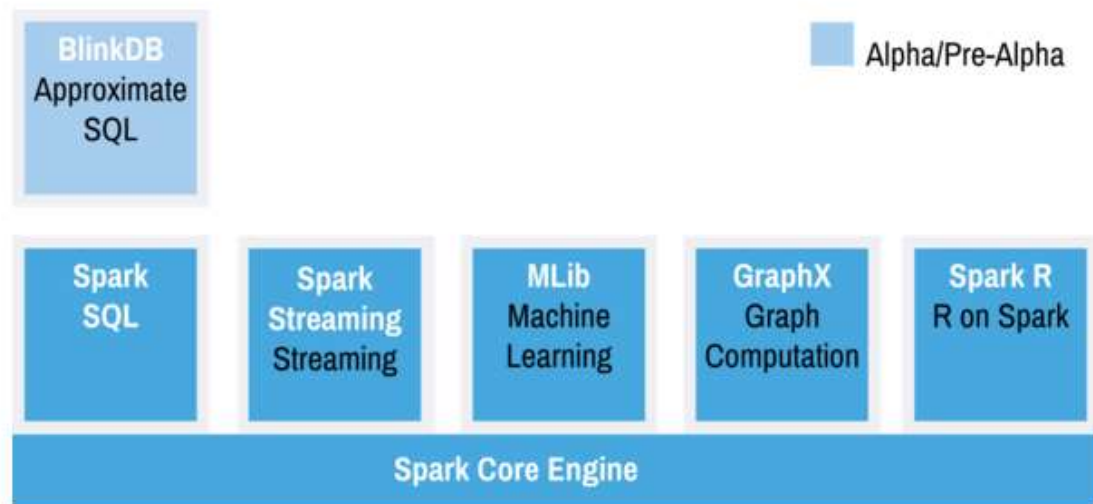


Fig 1. Apache Spark stack Diagram

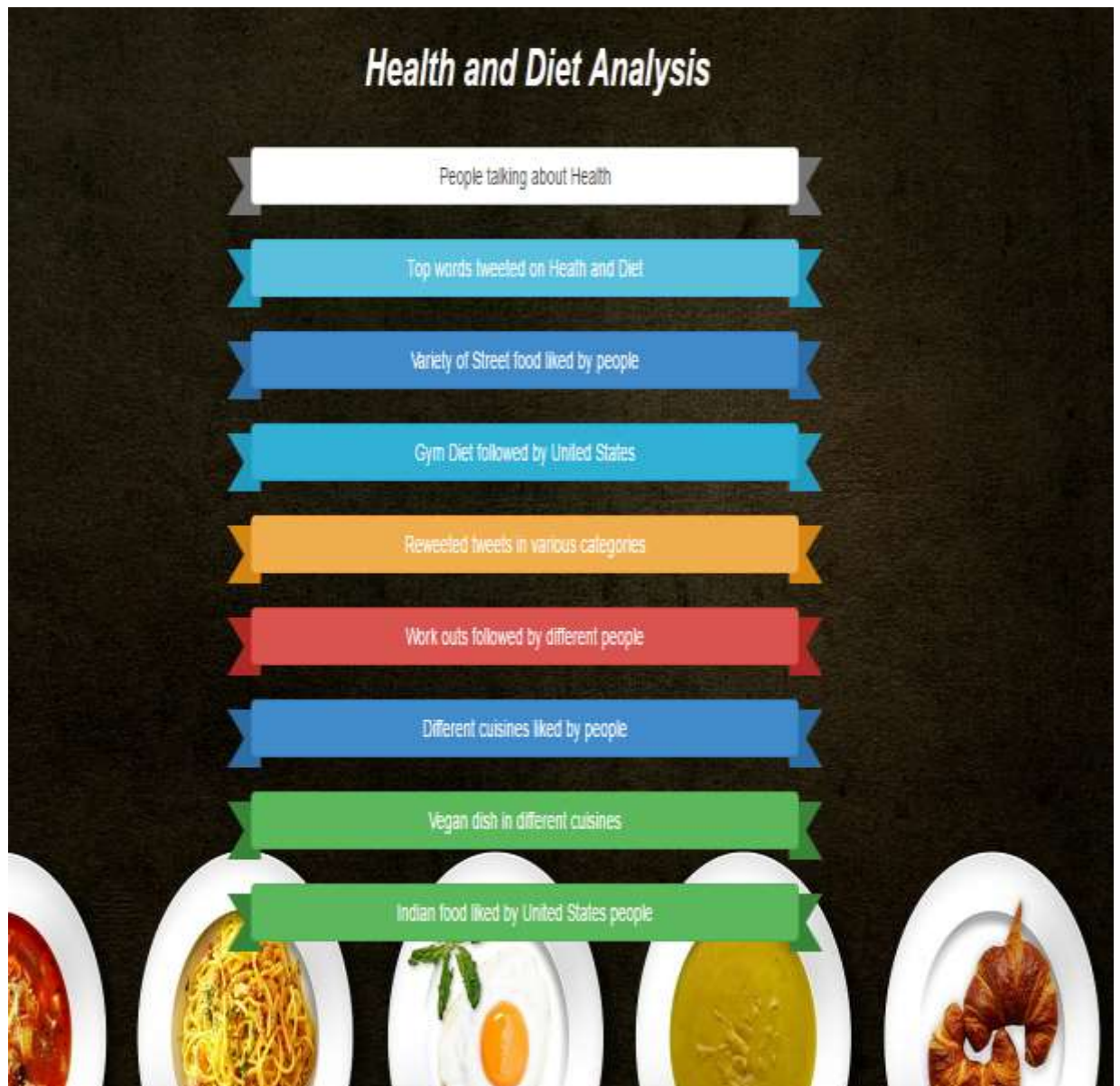
**Spark:** Spark provides programmers with an application programming interface centered on a data structure called the resilient distributed dataset (RDD), a read-only multiset of data items distributed over a cluster of machines that is maintained in a fault-tolerant way. It was developed in response to limitations in the Map Reduce cluster computing paradigm, which forces a particular linear dataflow structure on distributed programs: Map Reduce programs read input data from disk, map a function across the data, reduce the results of the map, and store reduction results on disk. Spark's RDDs function as a working set for distributed programs that offers a (deliberately) restricted form of distributed shared memory.

## 4. Implementation

### Queries:

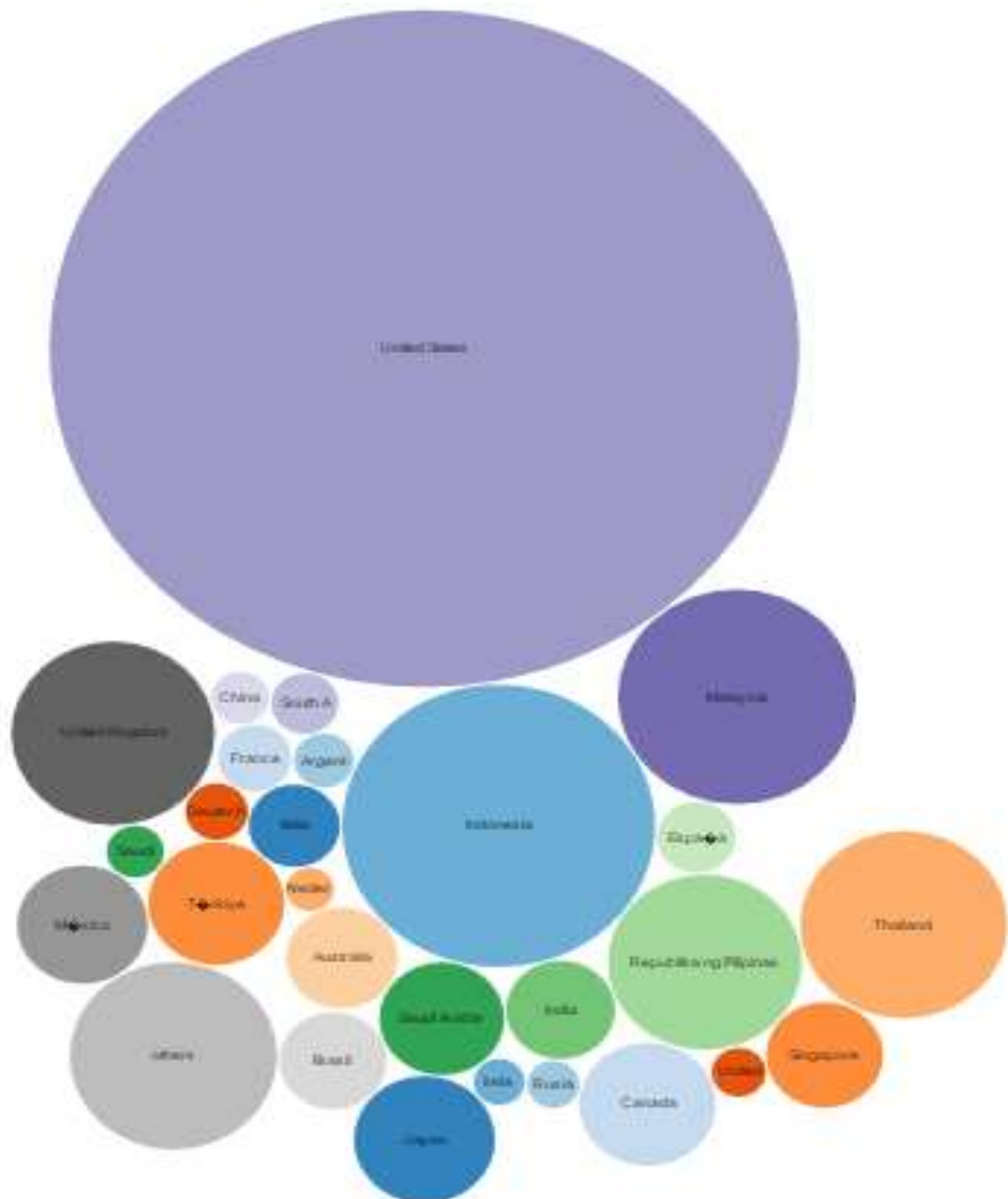
#### ❖ *Home Screen*

*ScreenShot:*



- ❖ No of people across the world talking about Health.

*Visualization:*



### Output Data:

A screenshot of a Notepad window titled "part-00000". The window contains a single line of text listing various countries followed by their respective counts in parentheses, separated by commas. The text is: (Italia,792)(Indonesia,9319)(Argentina,348)(France,517)(Deutschland,371)(Nigeria,176)(Türkiye,1773)(Nederlan, (Australia,1190)(????,1483)(????,1127)(Peru,193)(Polska,117) (Republika ng Pilipinas,3587)(Costa Rica,145)(España,590)(Malaysia,5400)(United States,53582)(South Africa,4, (Paraguay,181)(?????,262)(Canada,1777)(Việt Nam,119)(Colombia,199)(????????? ????????,283) (???,1303)(New Zealand,127)(?????????,4108)

- ❖ Popular words tweeted by people about health and diet  
*Visualization:*

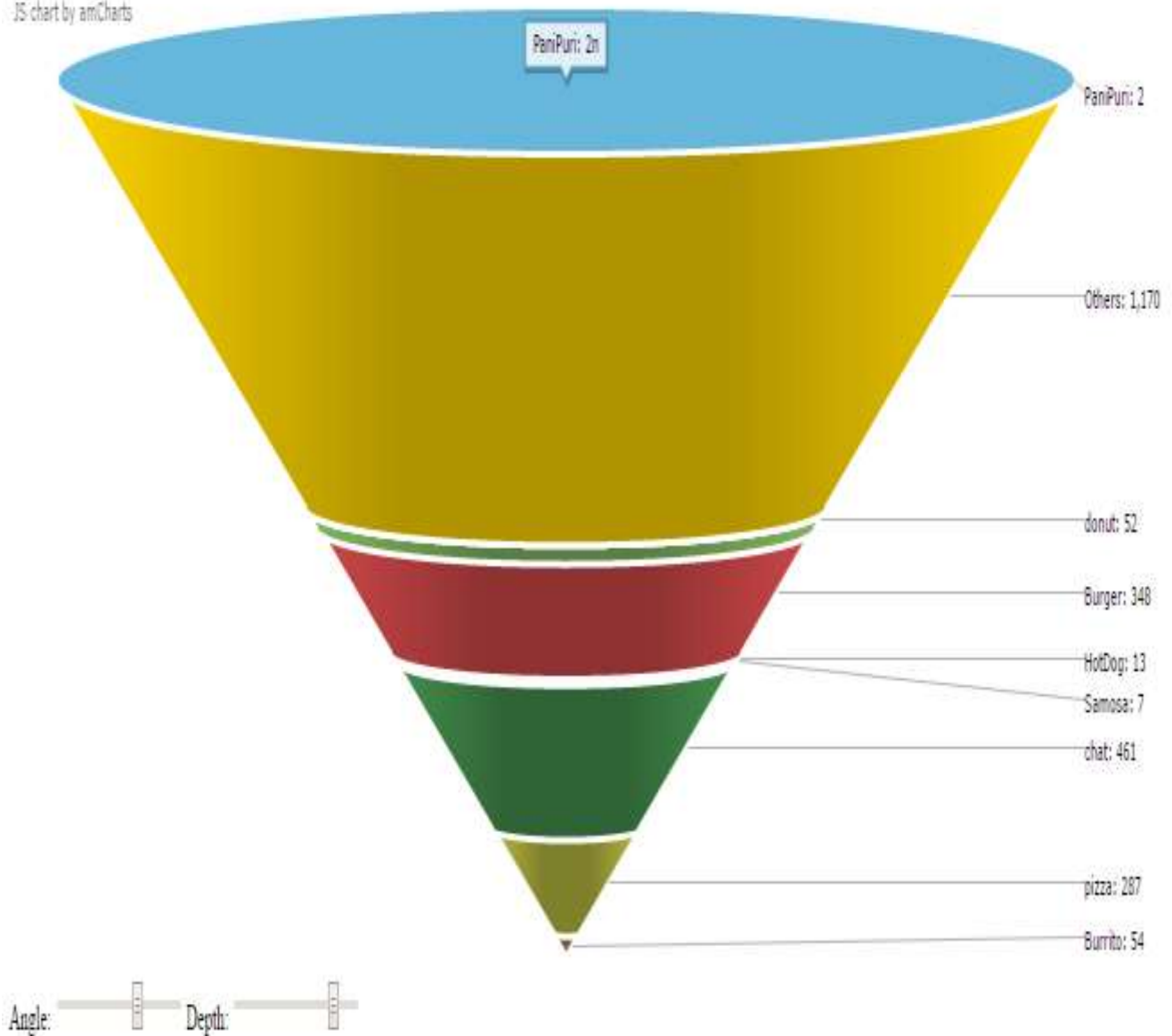




## ❖ Different kind of street food around the world liked by people.

*Visualization:*

JS chart by amCharts



### Output Data:

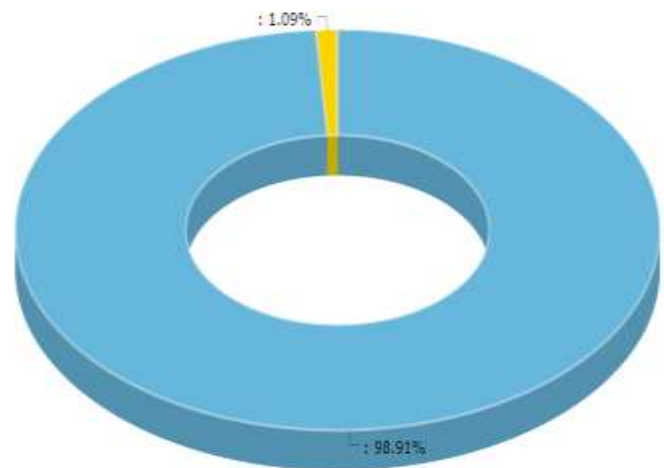
A screenshot of a Notepad window titled "part-00006 - Notepad". The window contains the following text: 

```
(Samosa,7)(chat,461)
(Burger,348)(HotDog,1)
(donut,52)(Others,818)
(PaniPuri,2)(pizza,1)
```

- ❖ Indian food in United States liked by people.

### Visualization:

JS chart by amCharts



Angle:  Depth:  Inner-Radius:

***Output Data:***

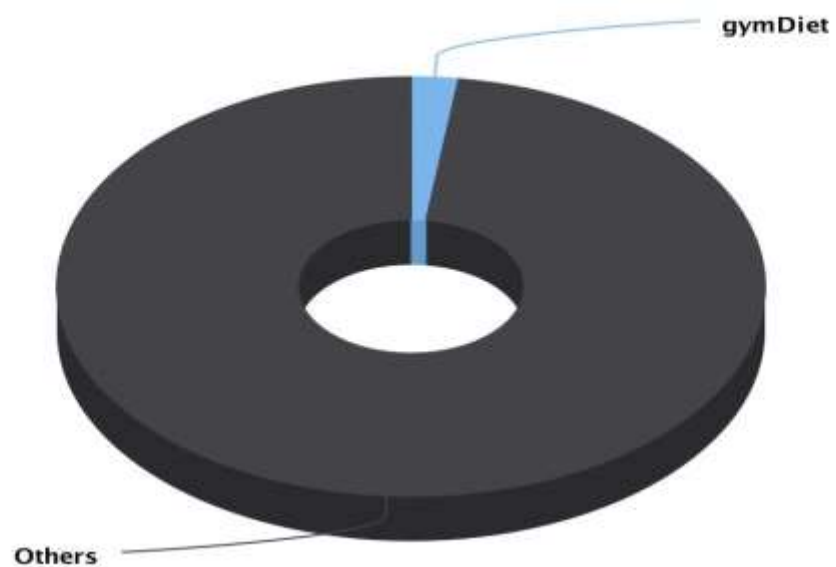
A screenshot of a Notepad window titled "part-00003 - Notepad". The window contains the following text:

```
(Indian,38)|  
(others,3399)
```

- ❖ Gym Diet followed by people in United States.

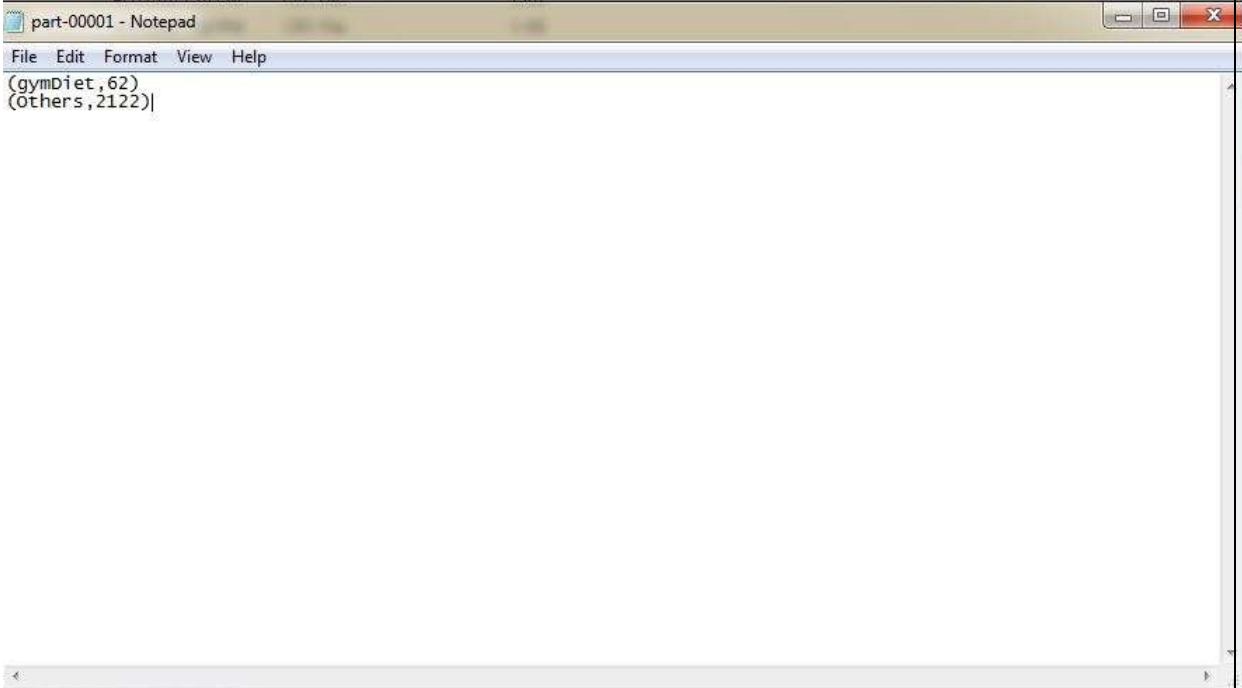
***Visualization:***

People Following Gym Diet in USA



Highcharts.com

### Output Data:



A screenshot of a Notepad window titled "part-00001 - Notepad". The window contains the following text:

```
(gymDiet,62)  
(others,2122)|
```

- ❖ Tweets retweeted by people categorized by various parameters.

*Visualization:*



### Output Data:

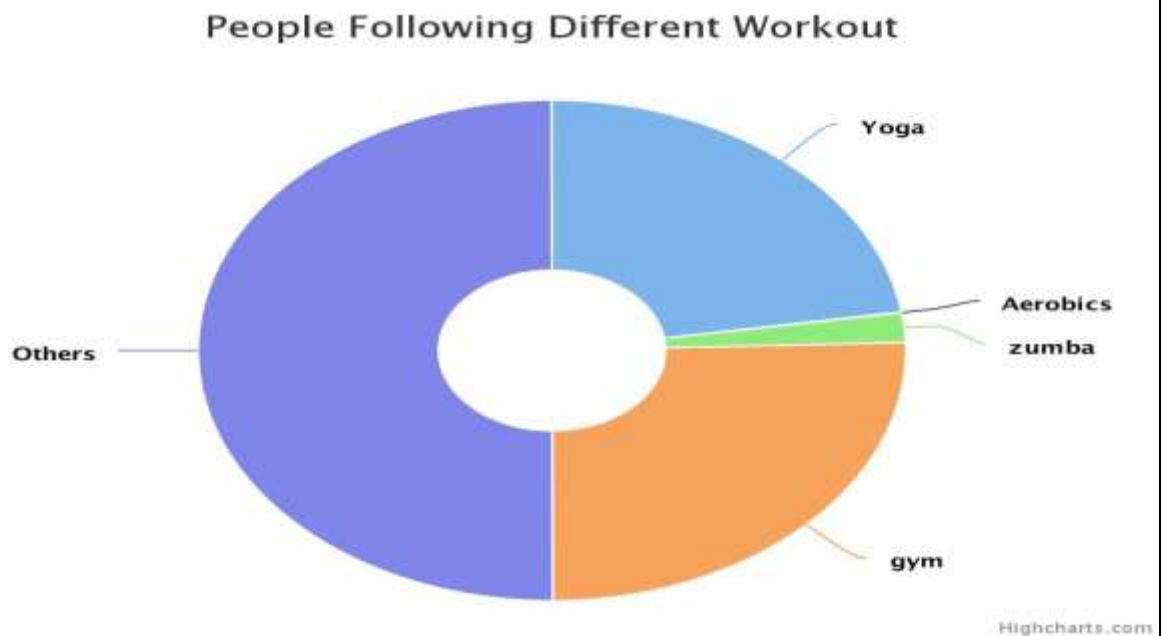


A screenshot of a Notepad window titled "part-00001 - Notepad". The window contains the following text:

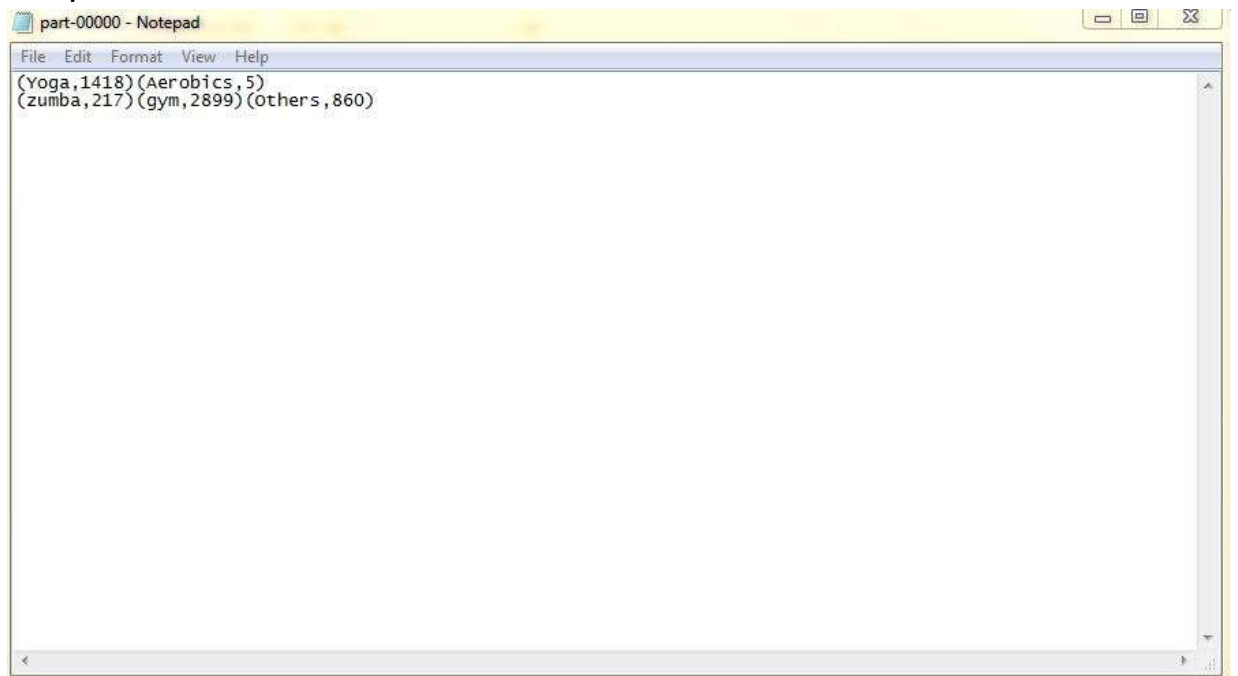
```
(FoodLovers,471)  
(FitnessFreaks,3859)  
(Others,436)
```

- ❖ Different workouts followed by people.

### Visualization:



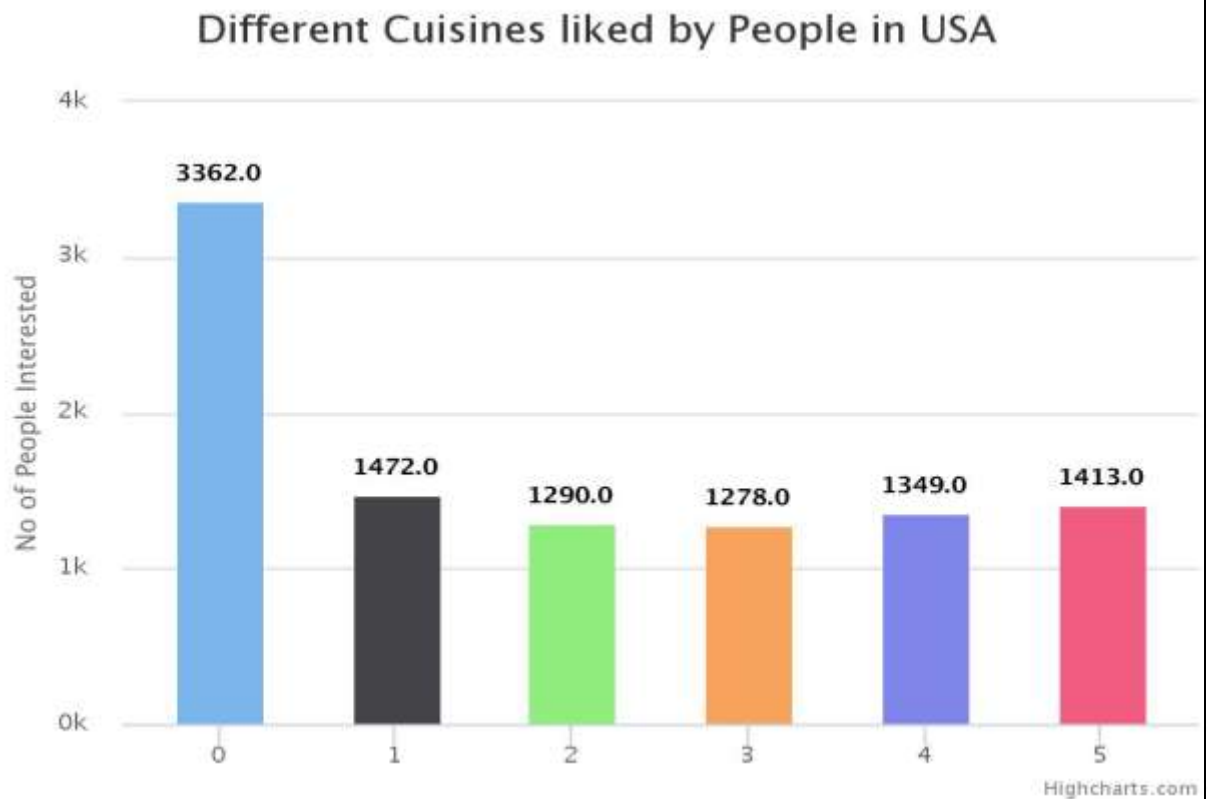
### Output Data:



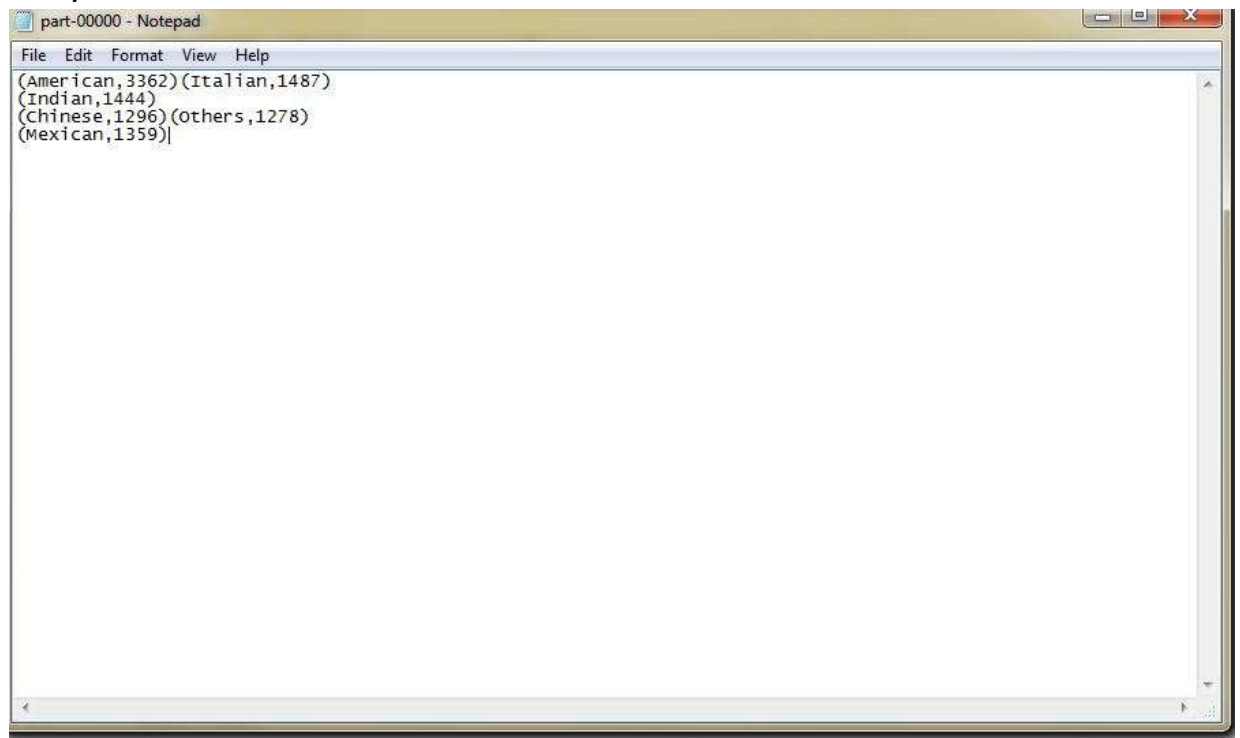
```
part-00000 - Notepad
File Edit Format View Help
(Yoga,1418) (Aerobics,5)
(zumba,217) (gym,2899) (others,860)
```

- ❖ What are the different cuisines liked by people.

### Visualization:



### Output Data:

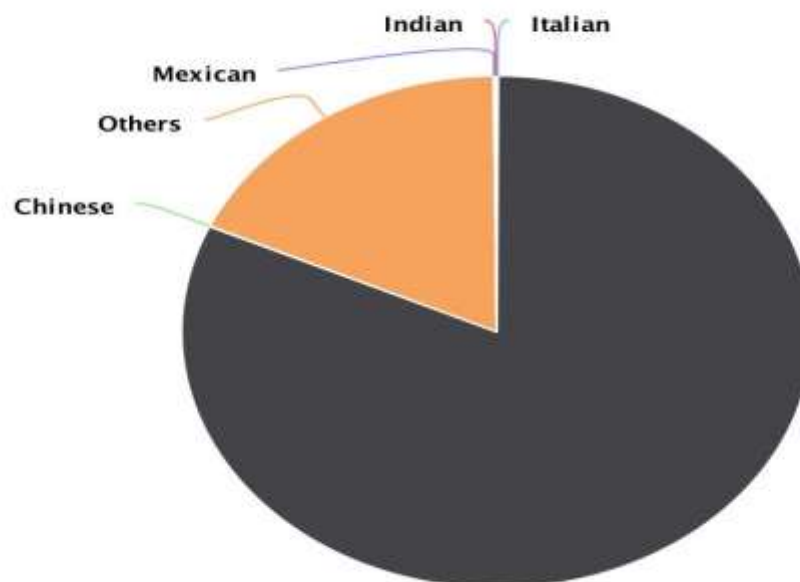


```
File Edit Format View Help
(American,3362)(Italian,1487)
(Indian,1444)
(Chinese,1296)(Others,1278)
(Mexican,1359)
```

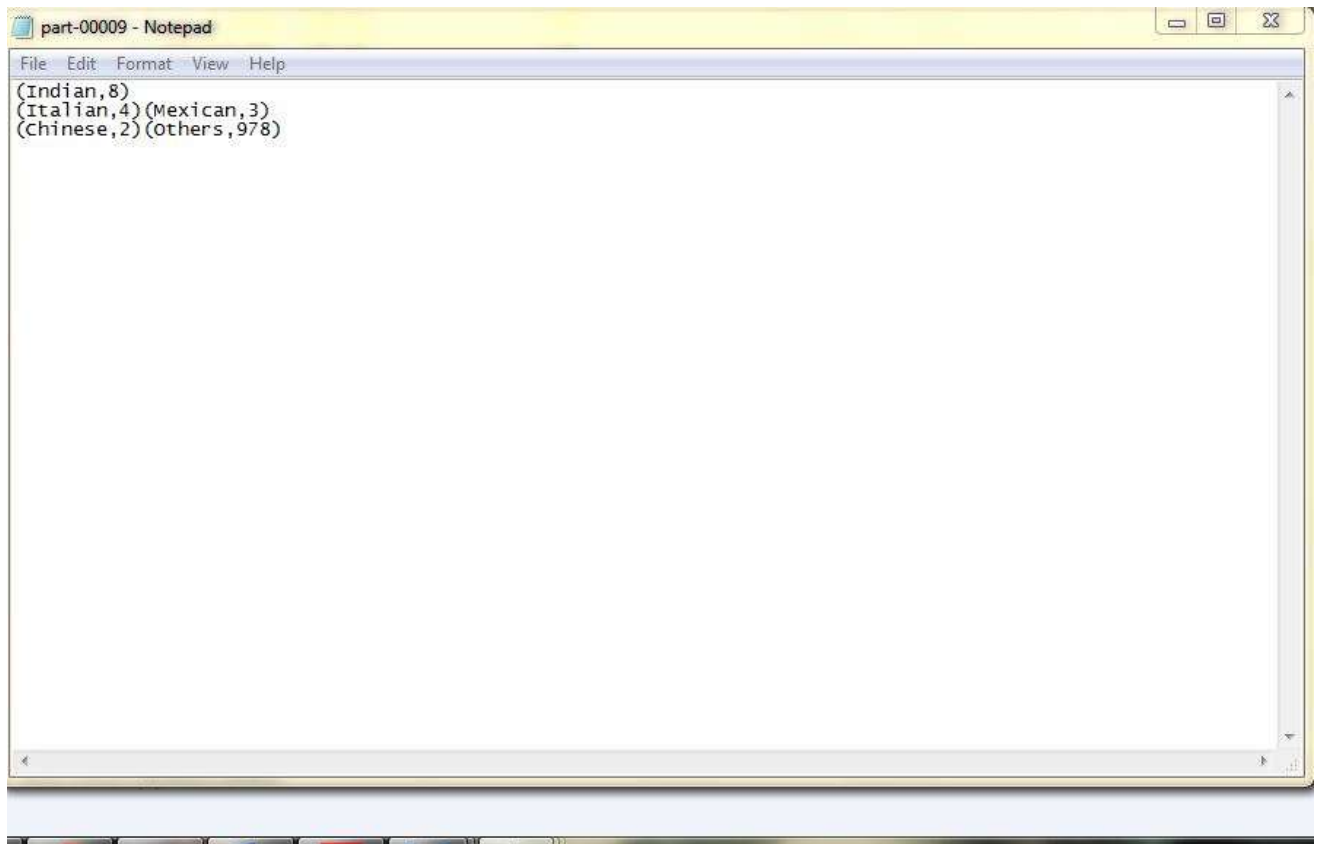
- ❖ Vegan dish in different cuisines liked by people.

### Visualization:

Vegetarian People in Different Cousine



Highcharts.com

*Output Data:*A screenshot of a Windows Notepad application window titled "part-00009 - Notepad". The window has a menu bar with "File", "Edit", "Format", "View", and "Help". The text area contains three lines of data: "(Indian,8)", "(Italian,4)(Mexican,3)", and "(Chinese,2)(others,978)". The text is black on a white background. The window has standard Windows window controls (minimize, maximize, close) in the top right corner. The taskbar is visible at the bottom of the screen.

```
(Indian,8)
(Italian,4)(Mexican,3)
(Chinese,2)(others,978)
```



## 5. Testing

### UNIT TEST CASES

S.no	Test case Title	Description	Expected Outcome	Result
1	Twitter data collection	Java Program to collect tweets based on our filters applied	Successful build and output was coming with word count	Pass
2	Query testing	Sample queries were ran on word count program	Output was displayed for the queries	Pass
3	Project queries created	All queries were created and ran on dump of twitter data.	Output count was displayed on successful build.	Pass
4	Visualization	Data was parse to display visual graphs, charts etc. on browser.	Visual charts were Displayed on browser.	Pass

### Integration testing:

Queries were ran on more than a hundred thousand tweets to display output and display output on browser for all queries.

## 6. References

- <https://spark.apache.org/docs/1.2.1/api/java/org/apache/spark/api/java/JavaPairRDD.html>
- <https://spark.apache.org/docs/0.7.2/api/core/spark/api/java/JavaPairRDD.html>
- <http://www.highcharts.com/demo/3d-pie-donut>
- <http://www.highcharts.com/demo/pie-semi-circle>
- <http://www.highcharts.com/demo/column-drilldown>
- <http://bl.ocks.org/mg1075/8bb2b55d2c5cf5667b01>
- <https://www.amcharts.com/demos/3d-funnel-chart/>