Declaration on Plagiarism

Name:	Saumitra Das
Student Number:	19211286
Programme:	Msc in Computing – Data Analytics
Module Code:	CA682
Assignment Title:	Data Visualisation
Submission Date:	15 Dec 2019
Module Coordinator:	Dr Suzanne Little

I declare that this material, which I now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. I understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. I have read and understood the Assignment Regulations. I have identified and included the source of all facts, ideas, opinions, and viewpoints of others in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged and the source cited are identified in the assignment references. This assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study.

I have read and understood the referencing guidelines found at http://www.dcu.ie/info/regulations/plagiarism.shtml, https://www4.dcu.ie/students/az/plagiarism and/or recommended in the assignment guidelines

Name: Saumitra Das Date: 15/12/2019

Top Tweeted Food Items by Top Nutrition Experts on Twitter

Twitter has become the source of information in the age of the internet. Many influencers share their expertise for their respective domains on Twitter, and people can achieve potential benefits by simply following them. Having proper nutritious food is one of the primary fitness goals of human in this century, and the best way to get some essential tips on health is to follow the top Nutrition Experts on Twitter.

The purpose of this project is to find the most commonly tweeted food items by top Nutrition Experts on Twitter assuming the food item shared by all the Nutrition Experts in their tweets multiple times should have a vital nutrition factor. The food items' tweet frequency is then measured in percentage to show the relative difference of food items tweeted among the Nutrition Experts. The number of tweets counts and the overall percentage of tweets for particular food items are to be measured to compare the items individually. The aim is also to find the most tweeted words related to food, health and nutrition other than the top tweeted food items to understand what nutritionist experts are talking about on twitter.

1. Dataset

Tweets of top 17 Nutrition Experts [1] are collected from Twitter using Twitter API and R. The Libraries used in R for this process are as follows

- httr
- rtweet
- twitteR
- SnowballC
- tm
- syuzhet.

Three thousand two hundred tweets (rows) of each Nutrition Experts are gathered along with ninety attributes (columns). 60616 tweets are collected in total with 90 columns and hence contains volume aspect of big data. As the data is collected from 17 different sources and merged together consistently, the data also contains the variety aspect of big data. The important attributes present in the data are screen name, creation time, text, source, location, retweets and number of followers. The data types present are integers, date-time, characters and strings.

2. Data Exploration, Processing, Cleaning and/or Integration

Data preparation is the most difficult part of this project as the tweets need to be cleaned properly and perfect set of stop words should be put in place to get the right word count of the required data. Also, singular and plural form, and lowercase and uppercase form of a particular food item should be considered as one and their summation of count should be taken.

Data Integration, exploration, cleaning and processing are done using Hadoop platform on cloud environment. Attributes which are completely unnecessary for the project are removed using Excel before uploading in Hadoop clusters.

The Hadoop utilities used in this process are:

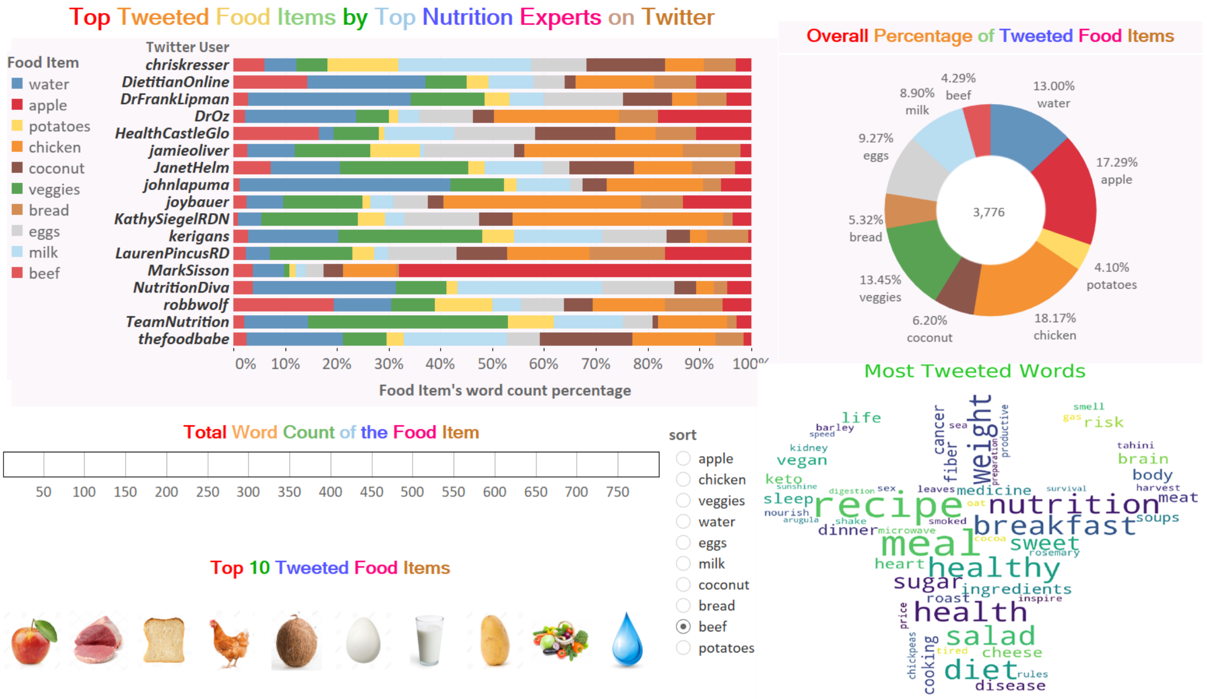
- Pig for exploration, integration and cleaning

- The 17 datasets are uploaded on Hadoop cluster and segregated using Pig.
- The whole data is explored and irrelevant attributes are removed from the final dataset.
- The tweet attribute is cleaned by removing unnecessary punctuation, new line, html links, commas etc and prepared for processing.

- Mapreduce for processing
  - One mapper code and one reducer code written in python is used to get a word count on all the words tweeted by users [2]. The stop words and irrelevant words are removed using the mapper code (the set of stop words are changed a lot of time depending on the output to get required results).
- Hive for producing the result datasets
  - Top food items tweeted by all the users commonly and top words tweeted with regards to health are achieved by creating tables and using SQL queries. The final dataset is then saved in csv format for visualization.

Using some functions in Excel, the output data is refined to get a better insight.

3. Visualisation



Choice of chart or graph type

- A stacked percentage bar chart is used to show the percentage of the top ten tweeted food items by top nutrition experts. As the goal is to show relative difference within each user for various food items, therefore stacked percentage chart is used [3]. An interactive bar chart is used to show the total number of counts for a particular food item. Whenever a food item is clicked from the “Top 10 Tweeted Food Items”, the bar chart shows the tweet count of that item.
- A donut chart is used to show the overall word count percentage of the top 10 tweeted food items.[4]. The chart interacts with “Top 10 Tweeted Food Items” and provides highlighted pie chart view for a particular food item. Although, pie chart is not believed

to present better information in most of the cases, but when highlighted for a particular category, it seems to be effective as shown in the video [5].

- A word cloud is used to show the most frequent words tweeted by the nutrition experts apart from the top 10 food items. The word cloud is shaped like a fit person showing biceps to provide relativity to this topic.

Design choices

- Nutrition rainbow colours are used for all the title of the charts as the graphs are talking about nutrition diet. Even though Comic Sans MS font was looking better while presenting colourful fancy title, it is not used because the childlike handwriting is not suitable for the intended audience.[6]. Instead, Microsoft San Serif font is used for the titles to provide a professional structure. Calibri font seems suitable for rest of the labels.
- Shapes of respective food items are used to provide relativity and colour of the food item’s label are chosen as per the colour of the food item. For example, red for apples and so on.
- For the word cloud, the dark to light shading of colour represents the frequency of the tweeted word. The intensity and boldness of colour are based on the word count. The darker and bolder the word, the most tweeted the word is.

Interactivity or animation

- A highlighted view for all the charts is presented by clicking on one of the food items. It will highlight the stacked bar graph for that particular food showing Nutrition experts’ percentage of word count in their tweets. It will also highlight the donut chart showing the overall percentage of the tweeted food item. The count bar chart will show the number of times the food item is tweeted. Therefore, it provides comprehensive information on the selected food item.
- Using the sort parameter, the stacked bar chart is sorted to get a better insight on percentage and compare the results of all the Nutrition Experts.
- Hovering over any chart will present information with pictures for that group or item. As for example, picture of the twitter influencer and food item along with other general information are displayed when hovered over the stacked bar graph.

List of tools or libraries used

- Tableau is used to build the stacked bar chart and the donut chart.
- Python is used to build the word cloud inside a shape. The libraries used for this process are:
    - Pandas
    - Numpy
    - Matplotlib
    - Matplotlib.pyplot
    - PIL
    - Wordcloud

4. Conclusion

The outcome of the visualization gives an overview of the top food items being tweeted by famous Nutrition Experts. After visualizing, it is found that chicken is the most tweeted food item (18.17% overall / wordcount: 686) out of the top 10 food items. Joy Bauer and Kathy Siegel have tweeted about chicken in majority i.e. 38.07 % and 40.83 % respectively whereas Kerigans and NutritionDiva have tweeted about chicken in minority i.e. 3.39% both. The second most tweeted item is apple and it is interesting to see that MarkSission's tweet percentage of apple is 68.01 % which is comparatively a lot higher as the second most tweet percentage of apple is 17.92 % by DrOz. So if MarkSission's apple tweet count is treated as an outlier, then apple will not be the second most tweeted food item.

The word cloud presents that Nutrition, health, sugar, weight, breakfast etc are the most tweeted word by the nutrition experts as these words are dark coloured and bold in nature. Although some words like "Recipe" displayed in bold but not with dark shade, are catchy to the eye before the higher frequency words than "Recipe".

Due to twitter's API restriction on collecting only 3200 recent tweets per user, the data used for visualization is considered as a sample of the population data. If the population data could be collected, the visualization would have given a comprehensive statistic on the data.

Were there aspects that you think could be improved upon?

- The overall model can be improved by connecting the "Top 10 Tweeted Food Items" chart with the sort parameter so that every time a user click on the food item, the sort parameter sort the stacked bar graph accordingly. An extra step can be excluded by doing so.
- The word cloud can be improved by working on detailing, boldness and colour significance in order to provide more detailed information.

Were there effects or functionality that you were technically unable to achieve?

- When clicking on a food item, the highlighted part of the donut chart turns the chart into a pie chart. The aim is to highlight it as a part of the donut chart which is not achieved technically in tableau [7].

References

1. http://blog.effifoods.com/philanthropy/the-30-top-nutritionists-on-twitter-and-why-you-should-follow-them/
2. https://github.com/devangpatel01/TF-IDF-implementation-using-map-reduce-Hadoop-python-
3. https://visual.ly/blog/how-groups-stack-up-when-to-use-grouped-vs-stacked-column-charts/
4. https://www.datarevelations.com/tag/donut-chart
5. https://interworks.com/blog/tmccullough/2014/03/06/tableau-cookbook-donut-charts/
6. https://designforhackers.com/blog/comic-sans-hate/
7. https://github.com/devangpatel01/TF-IDF-implementation-using-map-reduce-Hadoop-python-