Text Mining on Twitter: Finding frequently used words.

From: Smoking Aces

Abstract

We investigated the relation between the words and the frequency of the times they are used, from a number of tweets, for a given trending hashtag.

Introduction

Internet has fundamentally changed the way we interact with each other. Today, social networking tools like Twitter alone have billions of active users. Performing analysis on the message texts is not only challenging, but also a powerful tool that has been employed in diverse fields such as business, humanities and health sciences.

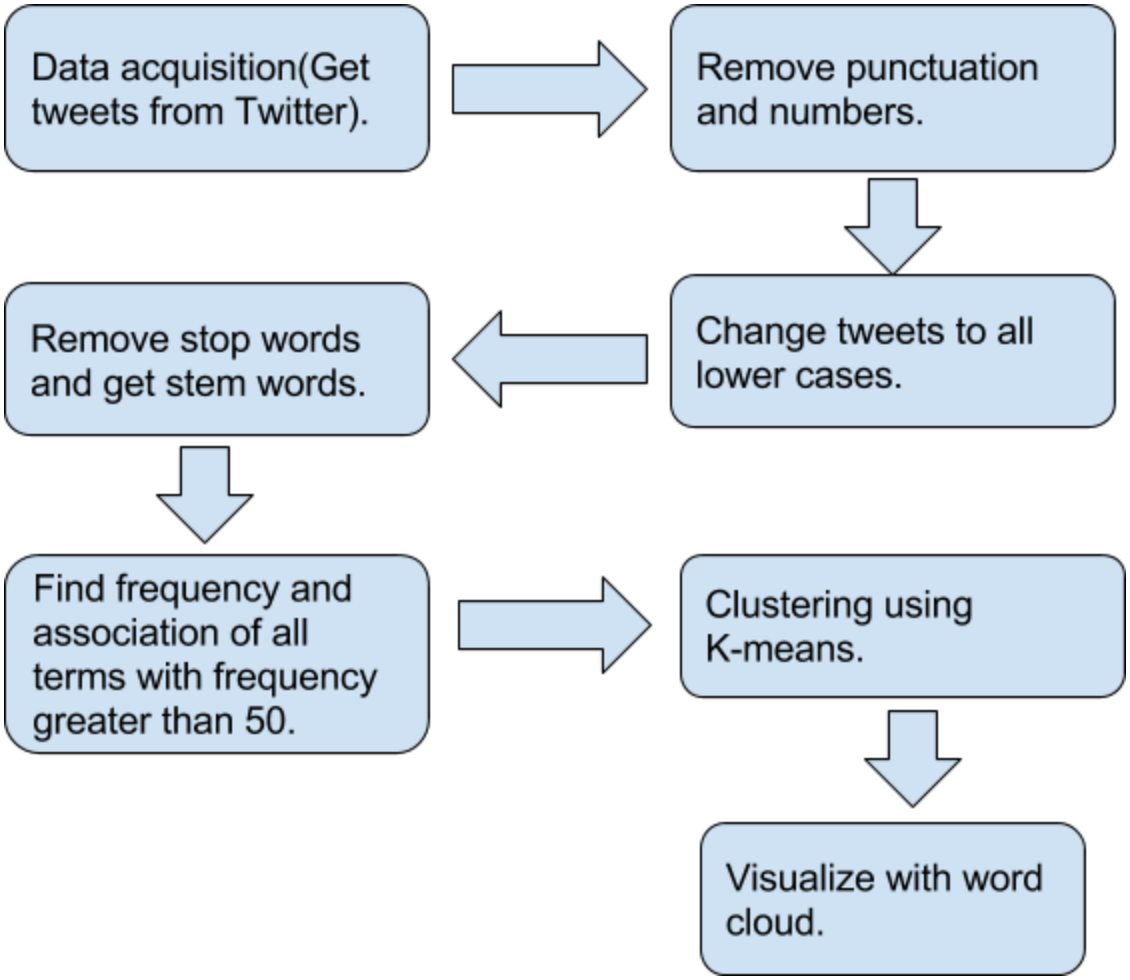
Given that, we used R programming language to retrieve tweets on trending hashtags, get the terms with the highest frequency and cluster them to get a sense about the most used words in the given context.

Design

The design phase was an important one, since the structure of our project depended on it.
The steps (in order) for the design phase was as follows:

- Get the most trending hashtags (say 5) from Twitter.

- Get tweets (say 1000) for each of the trending hashtags.
- Remove punctuation and numbers from it.
- Convert the tweets to all lowercase letters.
- Remove stop words and get only the stem words from the processed tweets.
- Count the number of times each of the terms are repeated (find the frequency terms).
- Get all the terms whose frequency is at least 50.
- Cluster the terms using K-means algorithm.
- VIsualize the results using word cloud.



Implementation

To implement the design, so as to complete our project, we proceeded as follows.
First off, we installed all the necessary softwares for the project.
To implement the first stage of design, we did as follows:

- Created a twitter app, to retrieve the tweets.
- We used
  > setup_twitter_oauth("API key", "API secret", "Access token", "Access secret") command to connect to Twitter. (Creating a Twitter app was necessary to get the API and Access keys and secrets).
- Then, we used the getTrends() function to get the latest trending hashtags.
- Finally , to get the tweets, we used the searchTwitter() function.

To begin with our second step, we proceed:

- We first convert the tweets to data frames.
- With the help of tm_map(docs, removePunctuation), we remove the punctuation in the tweets.
- Then, with the help of tm_map(docs, removeNumbers), we remove the numbers in the tweets.
- Later, to convert the tweets to all lowercase, we use the tm_map(docs,toLower) function.
- Finally to remove all the stop words and get the stem words, we use the tm_map(docs, removeWords, stopwords(“english”)) and tm_map(docs, stemDocument) respectively.
- To finish the data preprocessing, we used the tm_map(docs, stripWhitespace) to remove white spaces.

Next, continuing to the third phase, we do as follows:

- Determining the optimal number of clusters predicted from the function fviz_nbclust(), we call kmeans function to fit the data.
- Finally, pam() function and its plotter visualize the clusters.

Conclusion and challenges

From our project, we conclude that there is a lot of untapped potential in data mining from social networking sites to understand the public in a better way.

Looking at the run time of some of the algorithms and facing frequent type conversion errors, we had to use helper functions to understand what was going on behind the scenes.

With the amount of data collected, with the correct analysis, a lot of information about the public can be collected and this can be used in marketing.

References

1. http://www.sthda.com/english/wiki/cluster-analysis-in-r-unsupervised-machine-learning
2. https://rstudio-pubs-static.s3.amazonaws.com/31867_8236987cf0a8444e962ccd2aec46d9c3.html