MA415 Final Proejct Report/Technical Specification

CJ/Sijie Shan

December 14, 2017

1. Project Overview

The current project contains an interactive Shiny application, which allows users to plot a word cloud of most used words of *any* given Twitter account. The only input required from users is the user name of the account to be visualized, for example, @BU_Tweets, @BarackObama, or @YouTube. The idea behind the project is that I hope users to have more control of what is being presented to them, and Shiny, featured by its interactive interface, serves as an appropriate tool to achieve that purpose.

2. Example Word Cloud of Katy Perry's TWitter Account

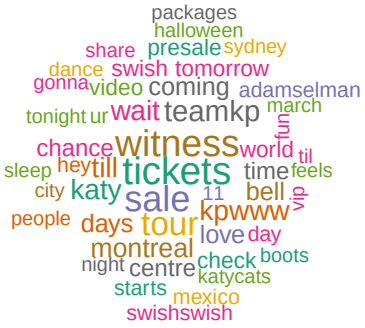


Table 1: Frequency of Most Commmon Words

Word	Frequency
witnessthetour	49
tickets	20
witness	19
sale	18
tour	15
teamkp	12
kpwww	11
wait	10
till	10
katy	10

3. How to Use the Shiny App

The word cloud visualization Shiny application can be accessed at *https://cjshan0417.shinyapps.io/Word_Cloud/*. After entering the Shiny application, please allow a few seconds for the app to finish loading. While loading, the app will show a "loading" message, and a default word cloud will appear as the app finishes loading.

Users can also control the minimum frequency of words appearing in the word cloud by changing value of the slider in the top left corner. For example, a minimum frequency of 5 means that only words that appear 5 or more times in Tweets will be included in the word cloud.

To change the Twitter account being visualized, enter a new Twitter user name to the input box under "Enter a Twitter User Name," upon finish entering, click on the "Refresh" button below the box.

Please note that a Twitter user name is different from a Twitter account name. A Twitter user name, which always starts with "@," is a unique name that is linked to a Twitter account. It cannot be changed by user. A Twitter account name, on the other hand, is not unique, and can be changed anytime. To draw a word cloud of a Twitter user, his or her Twitter **user** name - the name that begins with "@" - is required.

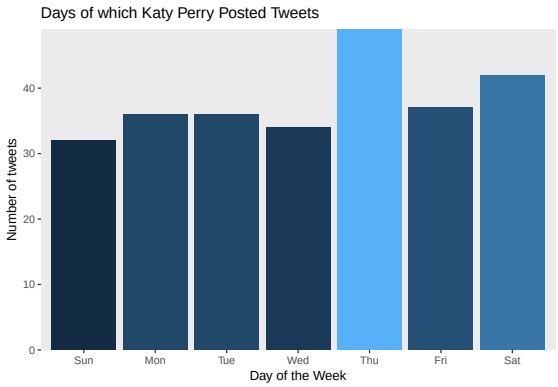
3.1. Exmaple of Visualizing a Twitter Account

To draw a word cloud of Tweets of Donald Trump, the current president of the United States, simply enter "@RealDonaldTrump" to the user name box, and click "Refresh." The new word cloud will appear in seconds. The "@" symbol need not to be included per users' preference.

4. More Visualization of @katyperry

To my surprise, most tweets are not geocoded - On Dec. 12th, 2017, I collected 279 tweets, but only 25 of them were geocoded. Out of these 25 tweets, there were 12 unique locations. Some locations are so close that they cannot be distinguished when plotted on a map. Below is a map of location of Katy Perry's tweets. It looks that she has been traveling quite frequently.

Below is a summary of Katy Perry's Twitter use by day.



5. Project Technical Specification

This section provides a brief overview of process of data manipulation in the Shiny app, and discusses some bugs in the application.

5.1. Data Manipulation in the Shiny App

The current Shiny apps adopts the following R packages: *twitteR*, *reshape*, *ROAuth*, *tm*, *wordcloud*, *shiny*, *tidyverse*, *tidytext*. Upon launching, the app fetches default data from Katy Perry's Twitter account, and normalizes text field of the Tweets. Examples of data cleaning include removing punctuation, stop words, single numbers, and single characters. Then the app converts the normalized data into a term document matrix, and uses the matrix to pick out most frequent words, and visualize those words using the *wordcloud()* function in the *wordcloud* package.

When user enters a new twitter user name, the app fetches new data from the updated Twitter account, and relaunches the above process.

5.2. Bugs and Potential Remedies

When testing the Shiny app, I noticed a few potential bugs to be fixed: when user enters a non-existing Twitter account, the app will return an error message. To avoid such prompt, some kind of output verification is needed; also, when the minimum frequency of words (which is control by the slider) exceeds the maximum frequency of words in the data, all words in the data will be plotted in the word cloud. This problem may require use of the *observeEvent()* Shiny function.

6. Summary

This Shiny app, though small, shows that I have learnt several essential abilities in data science: first, it demonstrates my ability to collect, clean, and normalize data as needed, which, in fact, is a dispensable skill to have when working with any type of data; second, it shows my ability to visualize data, and, more importantly, to allow users to participate in the process of visualization; last, it trains my ability to communicate what I have achieved to users, i.e., writing technical specification/business requirement document.

Throughout the semester, I have learnt a lot from the class - from producing PDF file using R Markdown, to collect, clean and visualize data using R Studio. I feel that my first foray into data science was quite successful, and I truly felt a sense of achievement when I finished the current project - it is already a handy app! I want to say thank you for teaching this interesting class, and for the knowledge you have taught us. Merry Christmas and Happy New Year!