I590 - Final Project

Anagha Vyas / Angel Rangel

# Twitter Social Media Analytics

## Goals and Objectives:

Social media analytics is used to gather information from social media sites like Twitter and analyze that data to make decisions. Business goals determine which data to gather and mine. Typical objectives are increasing revenue, reducing customer service costs, getting feedback, and understanding public opinion. Most large ‘born in the cloud’ product companies like Salesforce use annual conferences as the platform to launch new versions of products or new product lines. These conferences draw employees, customers and competition in large numbers. The attendees tweet about their experiences and opinions throughout the conference. Mining these tweets can help the decision makers to determine what worked and what did not.

For this final project we are trying to find what were the most discussed keywords at the conference? And also determine how popular they were.

The deliverables for the project are:

1. Project Report
2. Working code
3. Documentation to compile, deploy and execute the solution.

Software Used:

* R programming software
* library("shiny"): Web Application
* library("tm"): Text Mining Framework
* library("SnowballC"): Word Steaming Algorithm
* library("wordcloud"): Used to generate the Wordcloud plot
* library("ROAuth"): Authentication Library
* library("twitteR"): Twitter Library
* library("memoise"): Cache the results of a function that is executed with the same parameters.

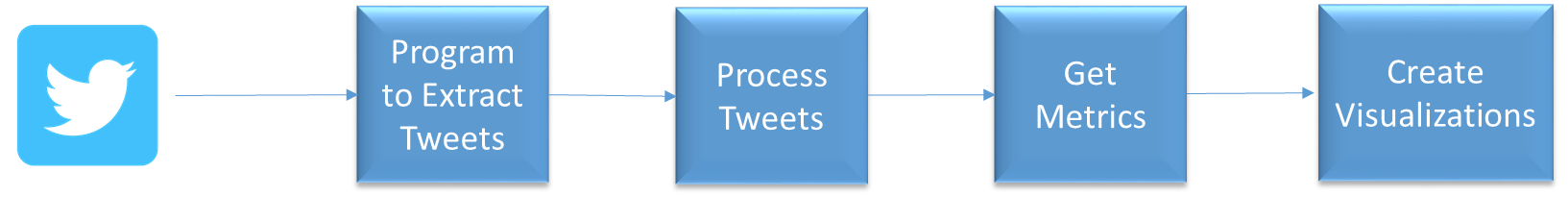
Tasks executed to develop the solution are –

1. Install R and RStudio (IDE)
2. Install the twitteR and ROAuth package to connect with Twitter.
3. Install the text mining package (tm)
4. Install SnowballC library.
5. Install the wordcloud package that will display the most popular twitter discussion terms in a word cloud
6. Install the Shiny package to create a web application where the hashtag of the conference can be entered and corresponding word cloud displayed in the web client window.
7. Create a Twitter App to access the Twitter API
8. Code the solution. Steps in the code -
   1. Establish an authenticated connection to Twitter.
   2. Search tweets based on a hashtag or search string
   3. Load tweets into R
   4. Clean the data and build a document-term matrix
   5. Find the frequency of the words
   6. Create the word cloud to visual important words
9. After successful code completion, create deployment bundle and deploy.

## Software Solution Details:

We considered Python and R for implementing the solution. We identified R as the programming language for the solution. We found that R has an advantage over Python for Data Analytics. It has a huge number of packages. There were packages available for creating a web application, data manipulation and visualization. Also the visualization packages in R were varied, rich and easy to customize.

Below the flow of information in the solution:



Step 1: Extract tweets-

We used the twitteR package to access the twitter API. ROauth was used for authentication. Once the authentication was successful a connection was established. Using this connection we searched for the tweets using a search string or a hashtag. These tweets were stored in-memory.

Step 2: Data Cleansing-

We used the tm (text mining package) for data transformation. The text was converted to lower case. All numbers, punctuations and URLs were removed from the data. English stopwords i.e. words that are very common were also filtered out. Examples of stop words – The, a, an, what, which. The final step of data transformation was ‘Stemming words’. Similar words like example and examples are converted to their radical – exampl. After data is cleansed we used Stem Complete to return the radical words back to words that were in the original text.

Step 3: Built the term matrix and calculated the frequency of terms

Step 4: Using Document matrix and the associated frequency, created a word cloud.

## Project Results:

As mentioned before, the application is created using RStudio . It has been published on following URL for review:

<https://anaghaangelprojecti590.shinyapps.io/AnaghaAngelProjectI590>

The application screenshot -

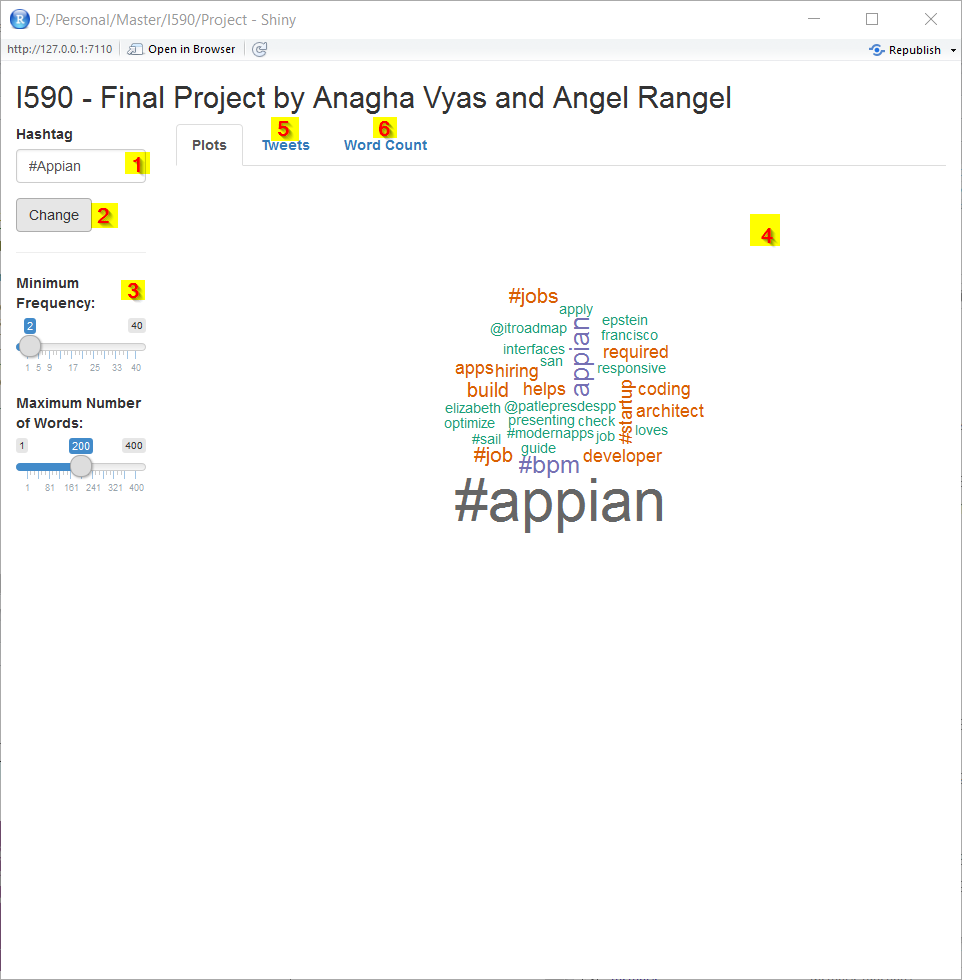


Fig 1. Main Screen

1) Text Field to change the Hashtag used to extract tweets.

2) In order to update the hashtag the user needs to click on “Update”.

3) Minimum Frequency and Maximum Number of Words Sliders.

4) Plot Tab will show the wordcloud generated from the word count.

5) Tweets Tab.

6) Word Count Tab.

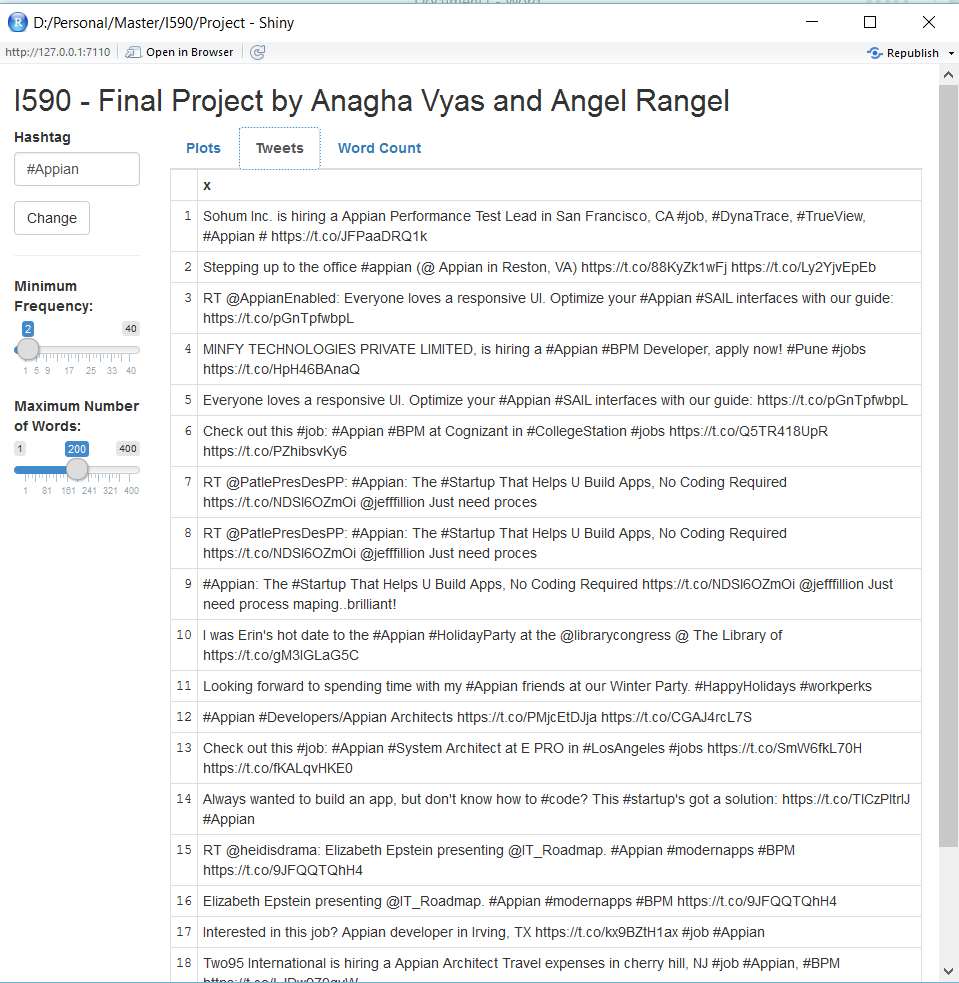


Fig 2. Tweets Tab will show a table with all the extracted tweets

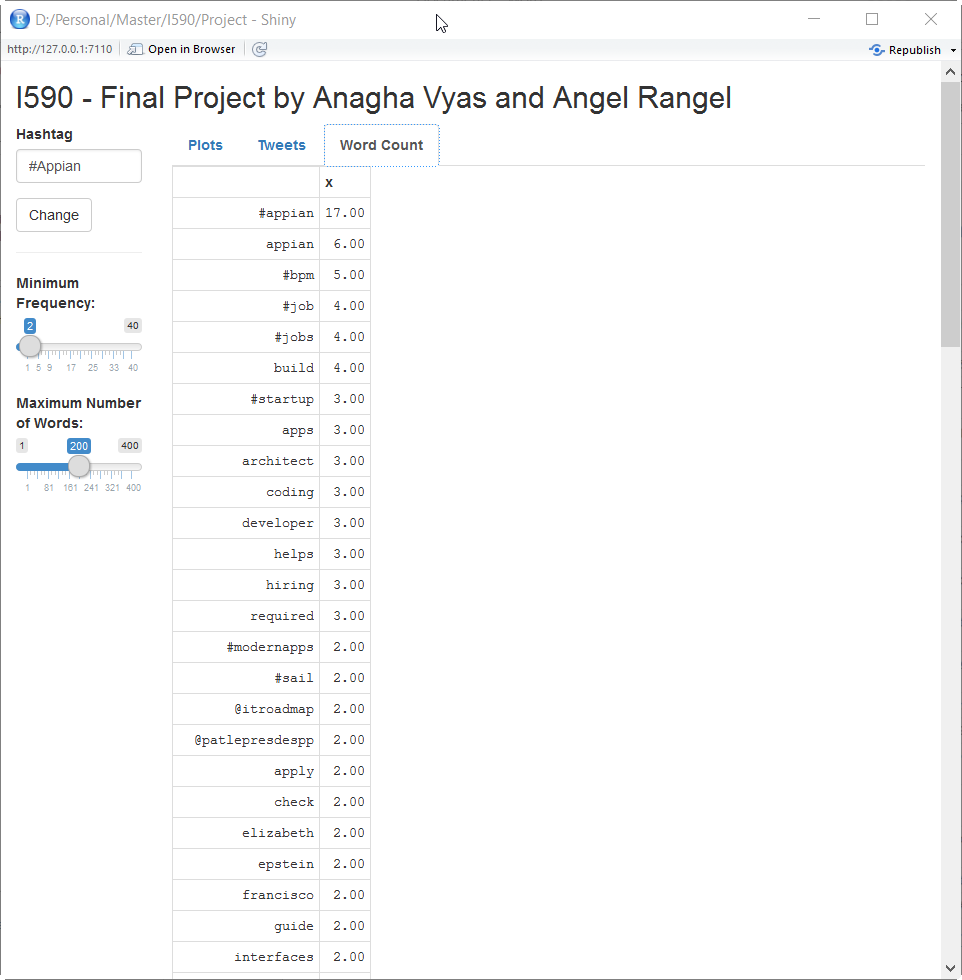


Fig. 3. Word Count will show a table with the results of the word counting after the data cleansing with the frequency of occurrence for each word.

## Reproducing the code

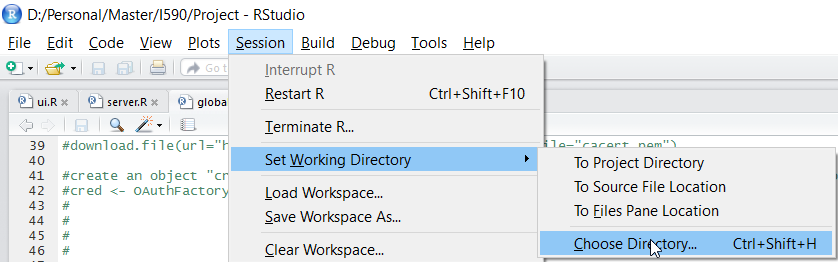
Source Code

Three files have been submitted along with this report -

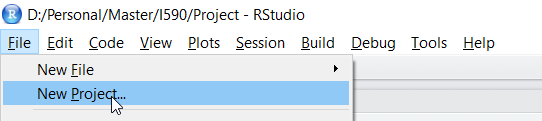
1. globalR: R program that is used by Shiny library and it contains the connection to Twitter and also the function used to extract the tweets.
2. server.R: R program that is used by Shiny library to handle all the operations from the user, and it calls the function to extract the tweets. Cleansing instructions are handle in this program.
3. ui.R: R program that is used by Shiny Library to create User interface code used to render HTML.

**Installation Guide for Windows User:**

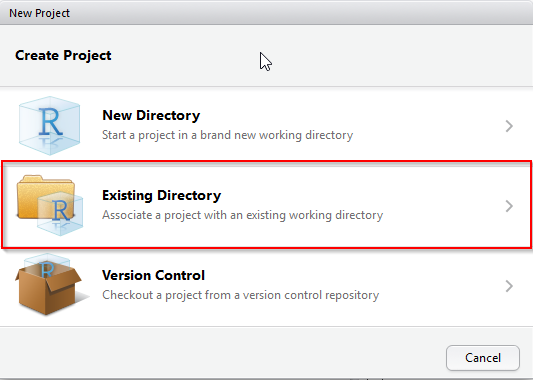
1. Download and Install RStudio from follow site: https://www.rstudio.com/
2. Save the Source Code in a Directory in your local machine.
3. Set the working directory to the folder where the files are located:



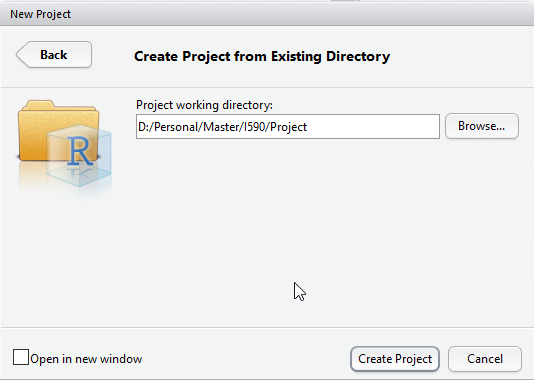
1. Create a new project in RStudio



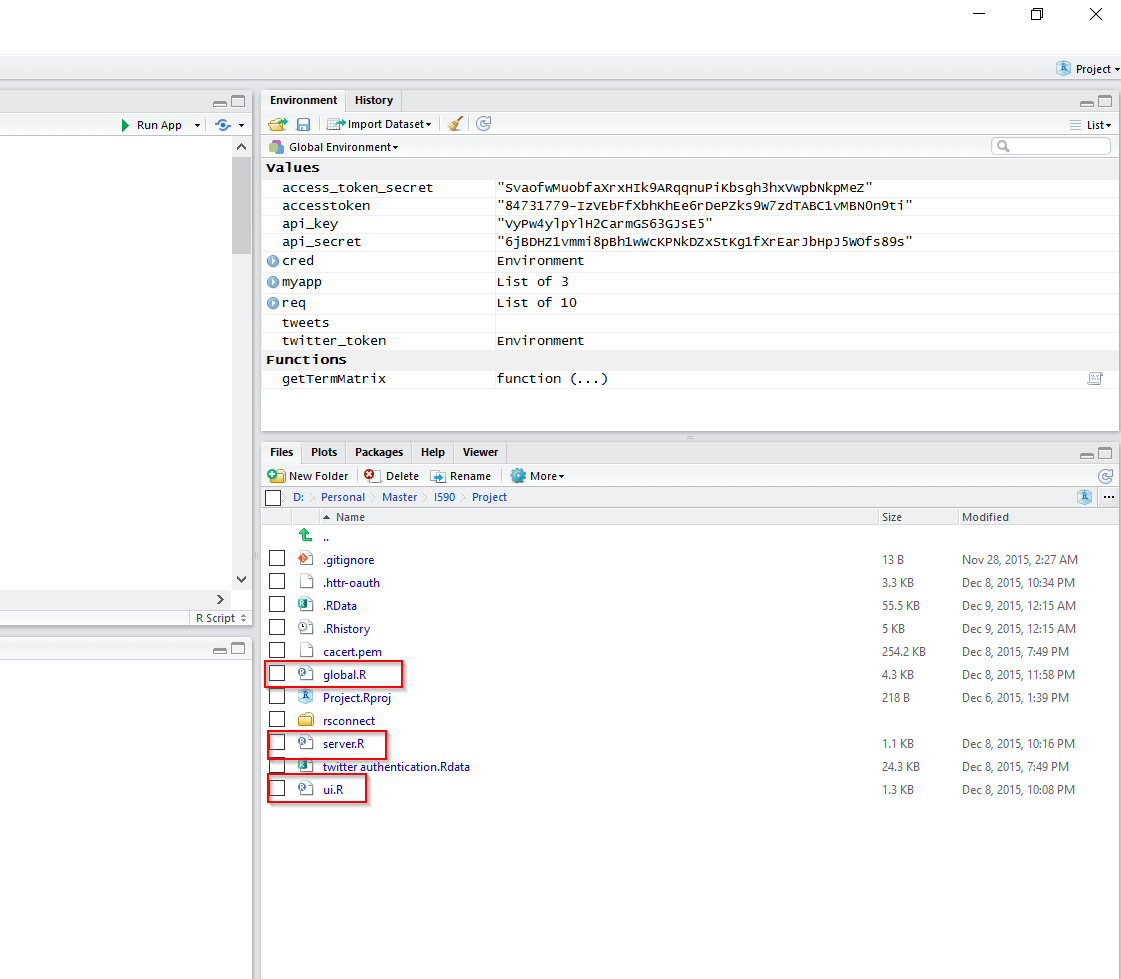
1. Select Existing Directory:



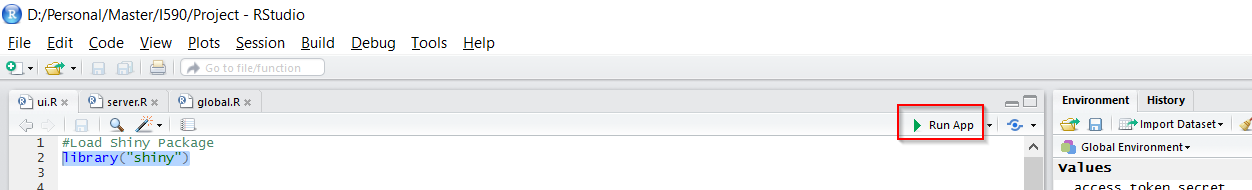
1. Select the folder that it was just created:



1. Click on Create Project
2. Execute follow commands (one by one):
   1. install.packages("ROAuth")
   2. install.packages("twitteR")
   3. install.packages("memoise")
   4. install.packages("devtools")
   5. install.packages("rjson")
   6. install.packages("bit64")
   7. install.packages("base64enc")
   8. install.packages("httr")
   9. install.packages("tm")
   10. install.packages("SnowballC")
   11. install.packages("wordcloud")
   12. install.packages("shiny")
3. Open the three files



1. Click on Run App:



**Datasets Used:**

The datasets will be extracted directly from twitter and it will search for the most recent tweets. The default hashtag provided is #Appian. Appian is one of the leading Business Process Management software. It recently celebrated Appian World, their annual marketing event gala.

**Future Work:**

With the current implementation we are just showing the wordcloud based on the word counting and after applying some cleansing. As future work we can add sentiment analysis to determinate the sentiment of the most recent tweets about a hashtag. Also the current implementation is limited as the tweeter policies about tweet search and extraction, so as future work we can create a process that will extract the tweets of the same hashtag during one week, so we will be able to have an additional dimension about how the hashtag has changed across the time.

References:

<http://dataconomy.com/r-vs-python-the-data-science-wars/>

<https://georeferenced.wordpress.com/2013/01/15/rwordcloud/>

<https://rstudio-pubs-static.s3.amazonaws.com/66739_c4422a1761bd4ee0b0bb8821d7780e12.html>

<http://shiny.rstudio.com/>