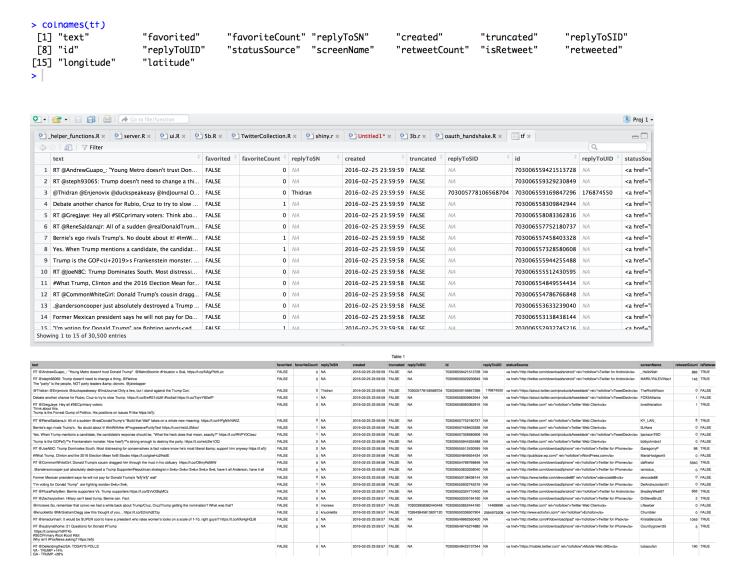
Introduction

In this project, exploratory data analysis (EDA) was conducted on large, structured and semi-structured datasets. With the help of EDA, important analysis about data and its relation with various parameters was studied. The different types of datasets analyses were twitter data set, New York Times ad click dataset, Real Direct dataset and live twitter data.

Problem 1: Data Acquisition

For this problem, the data source used was Twitter data using Twitter Search API, collected over a period of ten days. Twitter Streaming API is used in Problem 5. TwitteR and ROAuth were the packages that were used to carry out twitter data collection. TwitteR provided functions like setup_twitter_oauth which creates a connection to Twitter's Search API. The resulting list of retrieved tweets is stored in a data frame by using the function twListToDF(). Over a period of ten days, I created different data frames that stored that day's tweets. At the end, I combined all these data frames into one unit and stored this collective data as both cvs file and Rdata by using write.cv and save commands respectively. It is the local data frame containing collected tweets. These are the column variables it has.



This is the cvs file with tweets, stored on disk.