Just few months ago, the online AlphaGo[1] Go Game computer model beat all the top Go gamers, the news shocked the world in such a way that computer now can show the advantage in many ways than human, some even can say that computer is smarter than human being now. Go Game has been considered one of the toughest games reflecting the power of the human brain. Not until half year ago, there was not such a computer model can beat human in the Go game; however, this has been the history.

The winning logic behind is the advance of the human technology, such as in the way of the evolution of the machining learning, where the deep learning and deep mind algorithm is the new extension of machine learning and creates the AlphaGo.

It has been estimated that the computer can't win the Go Game at least within 5~10 years; however, the computer technology evolves so quick that it is faster than human's imagination. For example, just within less than six months, the initial AlphaGo model is able to learn so quick that initially it only beat the middle tie Go gamers, then suddenly it can beat all the top Go gamers in next 6~12 months.

Lectured by Professor Surendra Sarnikar [2] at California State University, East Bay, I am about finishing the graduate course – ITM 6285, Data Mining, the final project for the class will be focused on some topics of the driverless car. A driverless car (also known as autonomous car, auto, self-driving car, robotic car) is a vehicle that is capable of sensing its environment and navigating without human input [3], the developments of the driverless car require extensive machine learning, data mining and all other related high-tech elements, where it is very similar to the AlphaGo development. Though occasionally we can see the google driverless car on the road, most of us still think the mature of the driverless car will be in a long future; however, just by learning from the AlphaGo instance, driverless car in everywhere might just happen within next 10 years. 10-year is a short period for car usage, most of the car can last 15 years and very commonly many car can last more than twenty years. In some opinions, a 10-year-old car is just like a new car since all the components, especially many metal and mechanical parts inside the car are still in the mid-way of their life and in good shape.

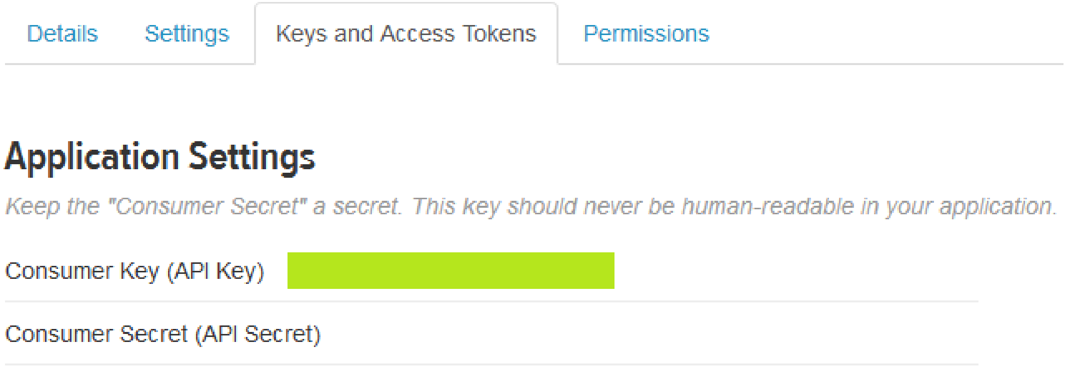
Then just imaging this scenario, in 2027, your 10-year-old car would be still considered a "new" car, but surrounding your car are more and more driverless cars, how do you think? My car is the antique now?

The technology always changes the ways we are living and thinking, imaging how about we have no internet and cell phone today, this will be the same feeling after the driverless car becomes popular and enter each family, then people can't live without the driverless car. So how do people feel about the driverless car toady? How would the people feel in the near future, for example, in 2027, a ten-year period from now on? And what would be every millstone the driverless car is going to develop?

Unlike the traditional way, nowadays we can do the survey over the internet and gain our insight regarding the topics we are searching for (this is also another proving that technology is changing our way of living and thinking in every respect). Besides we google the feedback and find every relevant forum, we can find our answer from twitter.

We can use the twitter API application program interface to harvest our searching result. Here is the step:

1. Sign up twitter account if none, and associate your phone number to the account;
2. Go to https://apps.twitter.com/ sign in using the same account name and password for the twitter account in 1;
3. Click create new apps (top right), fill in your new apps name, etc.;
4. Click on keys and Access Tokens, and find your consumer Key (API Key), Consumer Secret (API Secret), Access Token and Access Token Secret, they are total 4 long strings.



Application Settings

Keep the "Consumer Secret" a secret. This key should never be human-readable in your application.

Consumer Key (API Key) [redacted]

Consumer Secret (API Secret)

5. Run the following code in RStudio (R version 3.3.2 (2016-10-31))

```
requestURL <- "https://api.twitter.com/oauth/request_token"

accessURL <- "https://api.twitter.com/oauth/access_token"

authURL <- "https://api.twitter.com/oauth/authorize"

consumerKey <- "copy and paste the full string in Consumer Key (API Key)"

consumerSecret <- "copy and paste the full string in Consumer Secret (API Secret)”

accessToken <- " copy and paste the full string in Access Token"

accessSecret <- "copy and paste the full string in Access Token Secret "

setup_twitter_oauth(consumerKey,    #Please install.packages(“twitteR")

                consumerSecret,

                accessToken,

                accessSecret)
```

6. If success, you can see the prompt: [1] "Using direct authentication"

<u>Data Preparation:</u>

After the hand shake between Twitter and local PC is established, we can search for Twitter and save the search result to the local PC. We search and store 1000 twitter to the R list file.

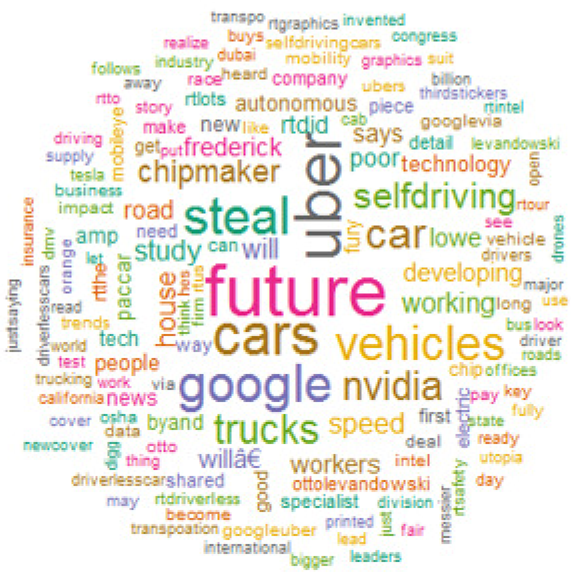
dlcar <- searchTwitter("Driverless", n=1000)

# where class(dlcar) [1] "list" and #dlcar for driverless car.

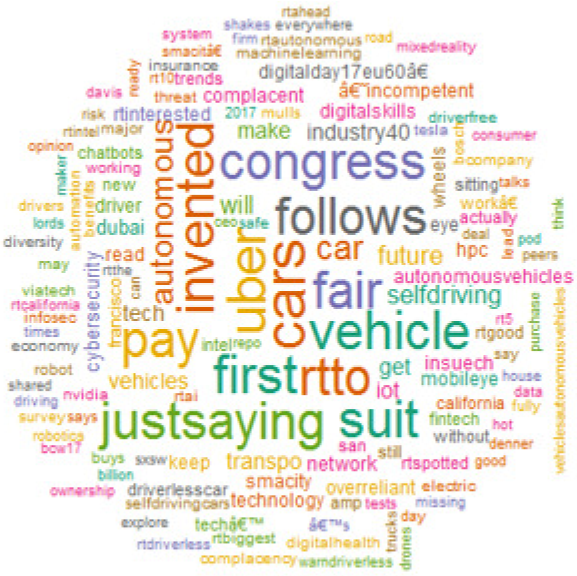
We can use the hash tag search option, "#Driverless" for more topic orientated; However, in this case study, we more focus on giving the twitter feedback not specific on certain interesting group so we can give a related fair sampling data. We can specify the result type by resultType argument specifying the type of search results received in API response. Default is mixed. Allowed values are mixed (includes popular + real time results), recent (returns the most recent results) and popular (returns only the most popular results). In our case, the resultType is mixed since we believe that the mixed option can provide the feedback from variative sources, the mixed option can even the data from the beginning. The since until option as following: since='2017-03-01', until='2017-03-02' so we can narrow down to a certain period, our case is to use the current. Please use the R helper ??twitteR for more options.

Data Analysis:

Now we can more further to the text mining for the 1000 twitter. Let's first look at the word cloud from the twitter searching without hashtag (#) and with hashtag (#).



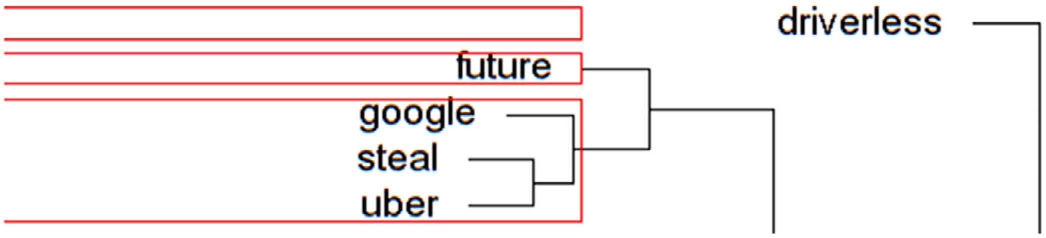
without Hash tag #



with Hash tag #

Twitter searching without hashtag # (left picture) have more general interesting orentitated. The twitter groups from general interest still think driverless car is the future (the biggest future word in pink); however, the special interest group (may including the driverless car manufacture) think the driverless car is the current issue such that it is the time for the congress for drivelss car guidline. Also the word invented is in the past tense, meaning that the driverless car now in the second phase not just building for the prototype, but it is the current issue for getting the congress pass some laws for the driveless car volumn production. The above is the quick insight from the word cloud comparison.

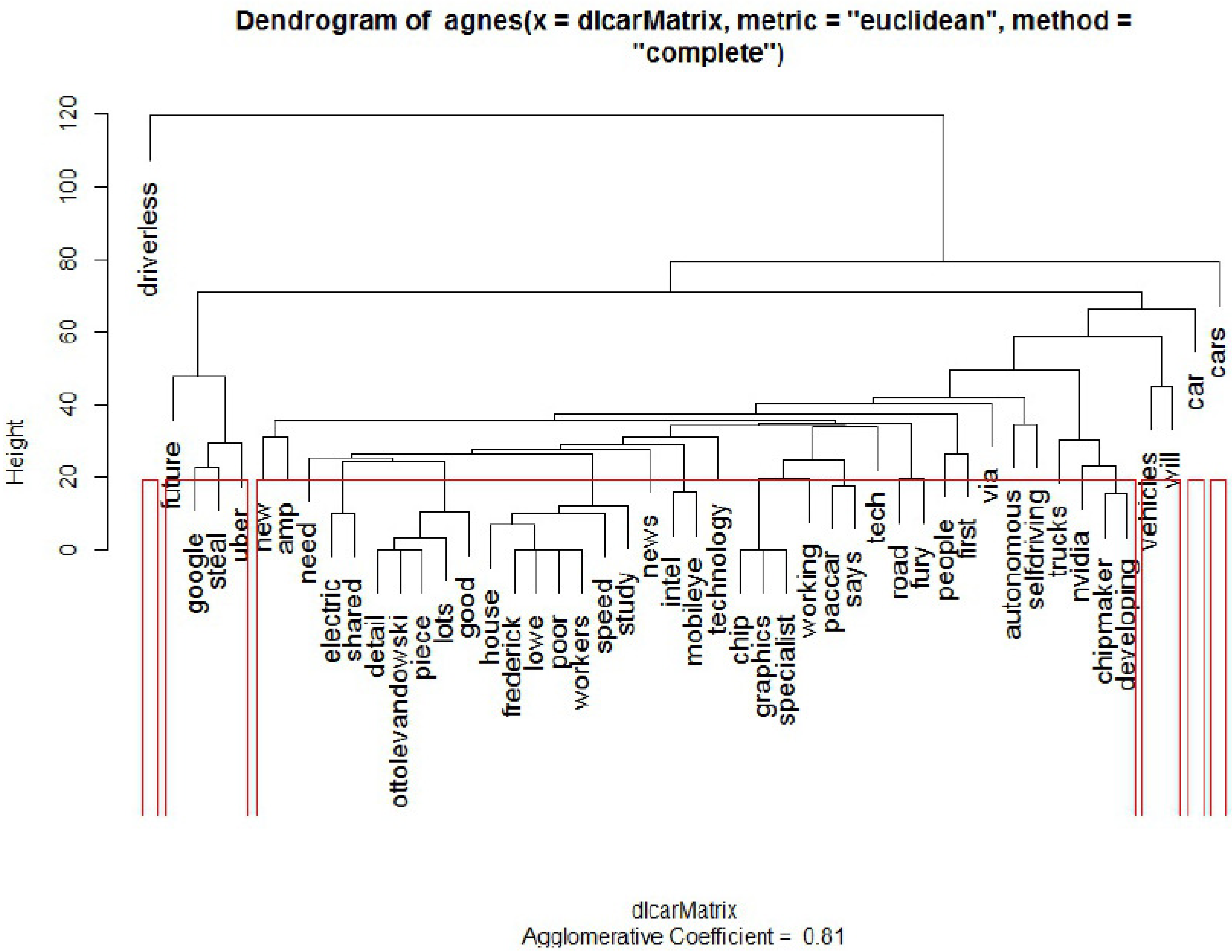
Another bring us the attention is big word “Steal”. By doing the cluster analysis, for example, once we get the dendrogram, our intuition will tell us that something happen between google and uber: we find out the google, uber and steal are linked in the nearby group, and this is corelated to a recent lawsuit between Google and Uber, where Google claims that Uber is stealing its driverless car technology.



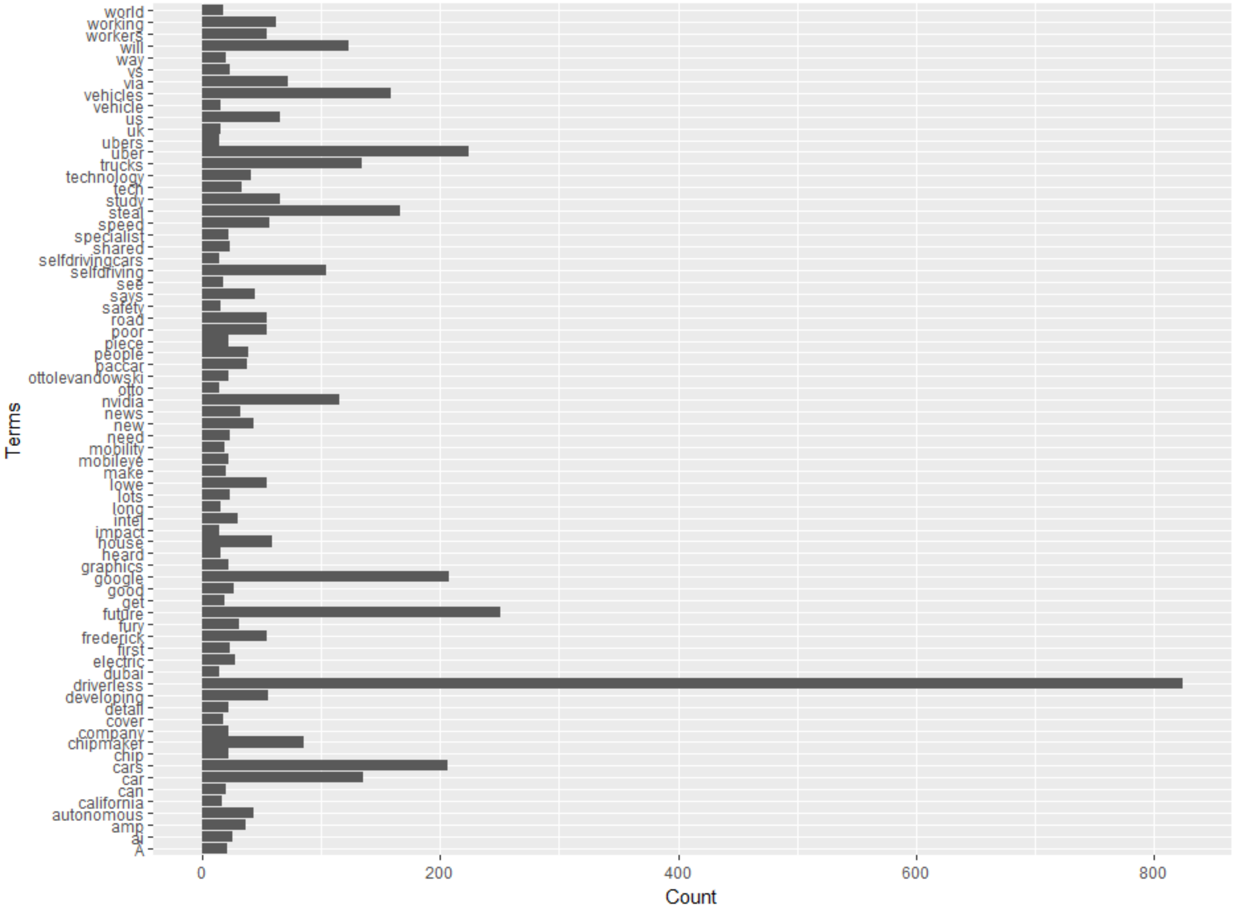
However, we won’t see the google and uber pattern in the dataset from the right wordcloud. This tells that the general public is acting faster to the general event. We can use agnes and hclustis clustering methods, and we use agnes for the presentation of dendrogram.

```
dlcarclusters <- agnes(dlcarMatrix, method = "complete", metric = "euclidean")

plot(dlcarclusters, which.plots=2, cex = 1.2)
```

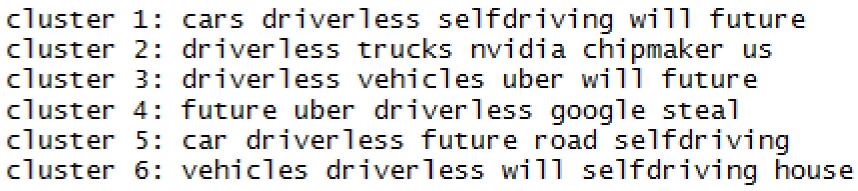


We can also merge some terms, such as car and cars to car, however, keeping them separate still understandable here since car is the Singular Noun form, some users may use the Singular Noun form car to describe the driverless car, so it may contain serval combination meanings for the car, that is, there will be some subtle differences between car and cars in this dataset.



The frequent word plot

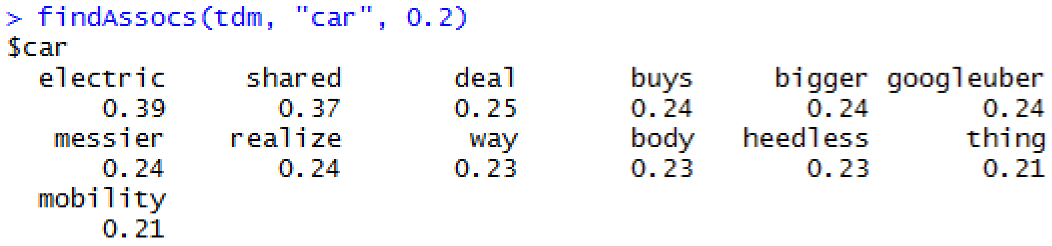
The cluster word group by k means method. Depending on the application, we can use the software to get the best k value by the elbow method, etc.; in this case study, the information is good if clustering by k=6 after try-and-error for the different k value. Below is the word grouping:



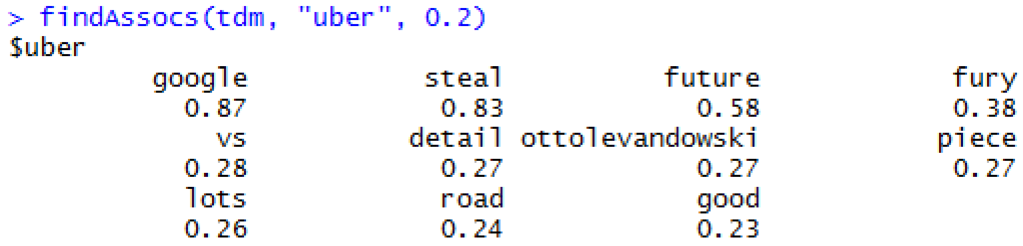
```
cluster 1: cars driverless selfdriving will future
cluster 2: driverless trucks nvidia chipmaker us
cluster 3: driverless vehicles uber will future
cluster 4: future uber driverless google steal
cluster 5: car driverless future road selfdriving
cluster 6: vehicles driverless will selfdriving house
```

Please note that cluster 4 is aligned the finding from the dendrogram clustering.

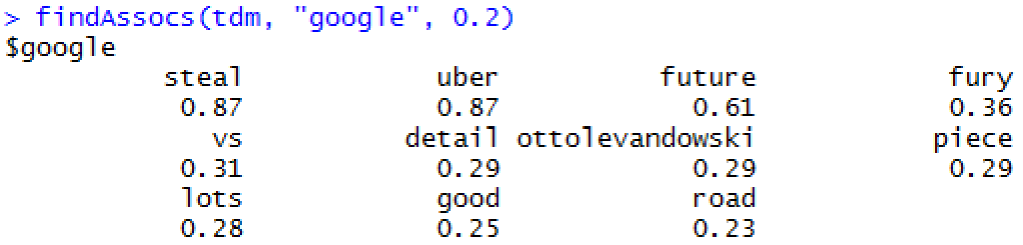
We then move further to take a look what is the association between the word uber, google, future and etc.



```
> findAssocs(tdm, "car", 0.2)
$car
  electric     shared       deal       buys     bigger googleuber 
      0.39       0.37       0.25       0.24       0.24       0.24 
   messier    realize        way       body   heedless      thing 
      0.24       0.24       0.23       0.23       0.23       0.21 
  mobility 
      0.21 
```



```
> findAssocs(tdm, "uber", 0.2)
$uber
         google           steal          future            fury 
           0.87            0.83            0.58            0.38 
             vs          detail ottolevandowski           piece 
           0.28            0.27            0.27            0.27 
           lots            road            good 
           0.26            0.24            0.23 
```

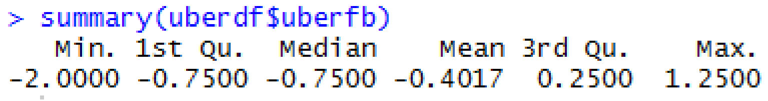


```
> findAssocs(tdm, "google", 0.2)
$google
          steal            uber          future            fury 
           0.87            0.87            0.61            0.36 
             vs          detail ottolevandowski           piece 
           0.31            0.29            0.29            0.29 
           lots            good            road 
           0.28            0.25            0.23 
```

We can easily tell that there are some relations between google and uber, for example, all the associated words are the same except the google and uber itself. Without looking the detail sentence of the twitter, we can get our institution of some events between google and uber, and by search the web, we find that there is an article with title of “Fury Road: Did Uber Steal the Driverless Future From Google?” [4] published about 10-hour ago before we collected the “driverless” related data from twitter.

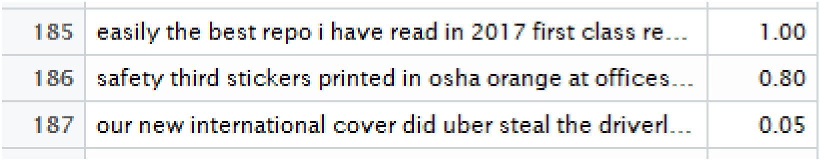
By using the text data mining, or using the term of text analytics, we can easily to find some pattern alert and do the further investigation. After the word cloud has been created for the dataset we collected, we confused with why the work steal appears in the word cloud picture, and by using the tm text mining package from R and by using the package of cluster, we can generate the cluster dendrogram and catch the relation between the google and uber from one fork of the dendrogram cluster tree (the picture shown above), and by using the findAssocs function from the tm package, we get more details regarding the related word between google and uber (the words shown above).

Furthermore, we can find the public feeling and response regarding the event between google and uber by using the function of get_sentiment.



```
> summary(uberdf$uberfb)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
-2.0000 -0.7500 -0.7500 -0.4017  0.2500  1.2500 
```

With some sample twitter shown below:

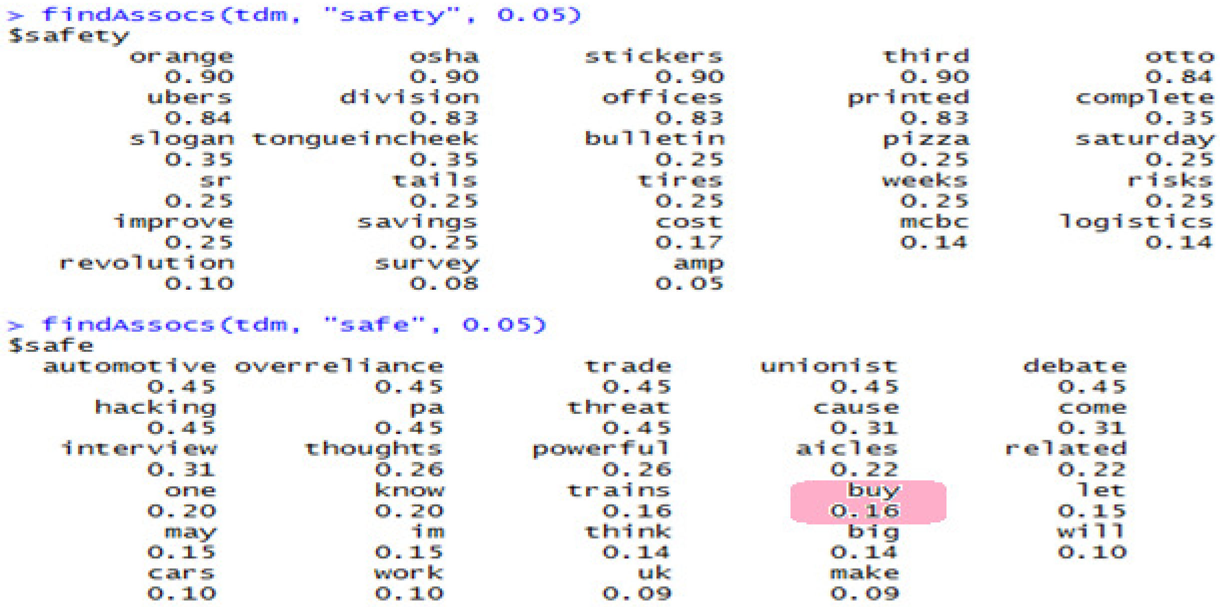


185	easily the best repo i have read in 2017 first class re...	1.00
186	safety third stickers printed in osha orange at offices...	0.80
187	our new international cover did uber steal the driverl...	0.05

uberfb <- get_sentiment(abc) #sample R code

By looking the subset of the data of 1000 twitter using word uber, we get 242 twitter as the subset. And from this subset, we have mean score for all the subset twitter sentiment estimation, the value is negative 0.4017. By looking at the subset data, we find that some twitters are repeated many times which might impact the score. However, by using the retweet function, the news can catch the public attention, and once we get more of the diversified feedback, the score should reflect the real feeling of the bigger public group correctly. From the statistic standpoint, we should have enough sample so that the sample standard variation is approx. equal to the population standard variation, for example, if we want to get a fair public response to the google and uber case, we can search the twitter again with difference search key word, such as google, uber, google+uber and etc. the twitter amount should as large as possible to reduce the sample variation, then we can use sentiment sore for the related fair estimation for the public opinion.

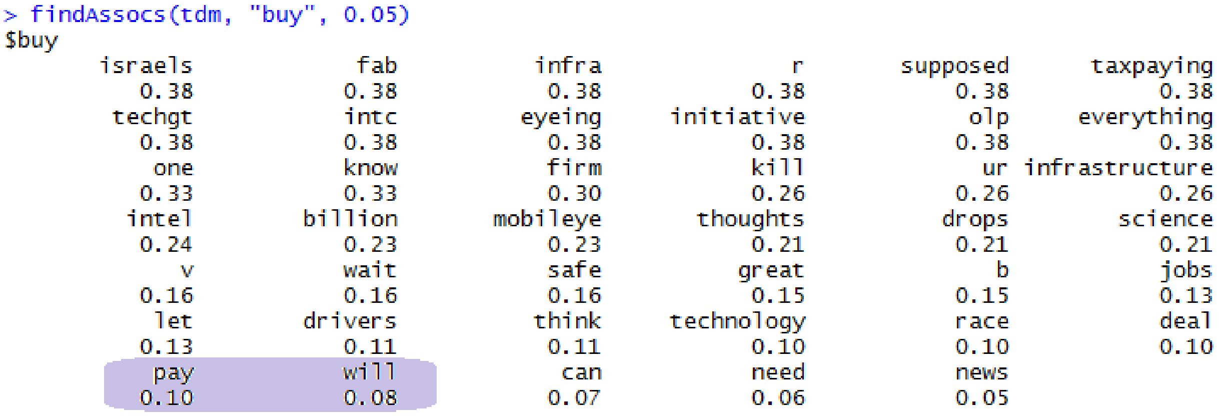
I am sure if we collect the data with the keyword of driverless in the different time from twitter, we will have the different result, this dynamic social behavior is common since everyday there are so many new events regarding this fantastic progress of the driverless car. The advantage of driverless car is obvious, we can have more time to do our own stuffs during the car is running, in this perspective, the driverless car is very similar to the public bus, where we can read our books, watch movies or do our works, etc. However, the major concern is how safe the driverless car will be and should be. From the word association mining from the same dataset, we can further understand the future consumer’s concern.



```
> findAssocs(tdm, "safety", 0.05)
$safety
         orange          osha      stickers         third          otto
           0.90          0.90          0.90          0.90          0.84
          ubers      division       offices       printed      complete
           0.84          0.83          0.83          0.83          0.35
         slogan tongueincheek      bulletin         pizza      saturday
           0.35          0.35          0.25          0.25          0.25
             sr         tails         tires         weeks         risks
           0.25          0.25          0.25          0.25          0.25
        improve       savings          cost          mcbc     logistics
           0.25          0.25          0.17          0.14          0.14
     revolution        survey           amp
           0.10          0.08          0.05

> findAssocs(tdm, "safe", 0.05)
$safe
  automotive overreliance        trade     unionist       debate
        0.45         0.45         0.45         0.45         0.45
     hacking           pa       threat        cause         come
        0.45         0.45         0.45         0.31         0.31
   interview     thoughts     powerful      aicles      related
        0.31         0.26         0.26         0.22         0.22
         one         know       trains          buy          let
        0.20         0.20         0.16         0.16         0.15
         may           im        think          big         will
        0.15         0.15         0.14         0.14         0.10
        cars         work           uk         make
        0.10         0.10         0.09         0.09
```

First all, the word orange osha stickers third rank the top association list, and there are several article related to it when searching by the google, for example, ‘Safety Third' Is The Running Joke At Uber's Self-Driving Car Unit [5]. And for the marketing section of the driverless car manufacture, and for the general public, the low association of safe and buy (0.16) meaning driverless car still have some big improvement for the safety. The below buy association also tell the low demand regarding the driverless car.

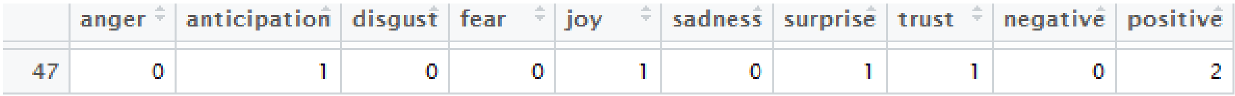


```
> findAssocs(tdm, "buy", 0.05)
$buy
        israels            fab          infra              r       supposed      taxpaying
           0.38           0.38           0.38           0.38           0.38           0.38
         techgt           intc         eyeing     initiative            olp     everything
           0.38           0.38           0.38           0.38           0.38           0.38
            one           know           firm           kill             ur infrastructure
           0.33           0.33           0.30           0.26           0.26           0.26
          intel        billion       mobileye       thoughts          drops        science
           0.24           0.23           0.23           0.21           0.21           0.21
              v           wait           safe          great              b           jobs
           0.16           0.16           0.16           0.15           0.15           0.13
            let        drivers          think     technology           race           deal
           0.13           0.11           0.11           0.10           0.10           0.10
            pay           will            can           need           news
           0.10           0.08           0.07           0.06           0.05
```

Understanding that the text analytics is relatively subjective, I highly respect your opinion. The goal for the blog is to develop some practical ways to analysis some current social issues not just limited to driverless technology, the word cloud, cluster and word association analysis techniques can apply to most of the twitter based on user’s favorited topic, besides your own opinion, you can view the other’s concurrently, if you want to research the pass event, you can change the twitter search and harvest option so you have the flexibility for your search target.

Computer technology, especially data mining, AI, BI, machining learning and deep learning with utilizing the structured or unstructured big data are changing our daily life, providing us better heathy life styles.

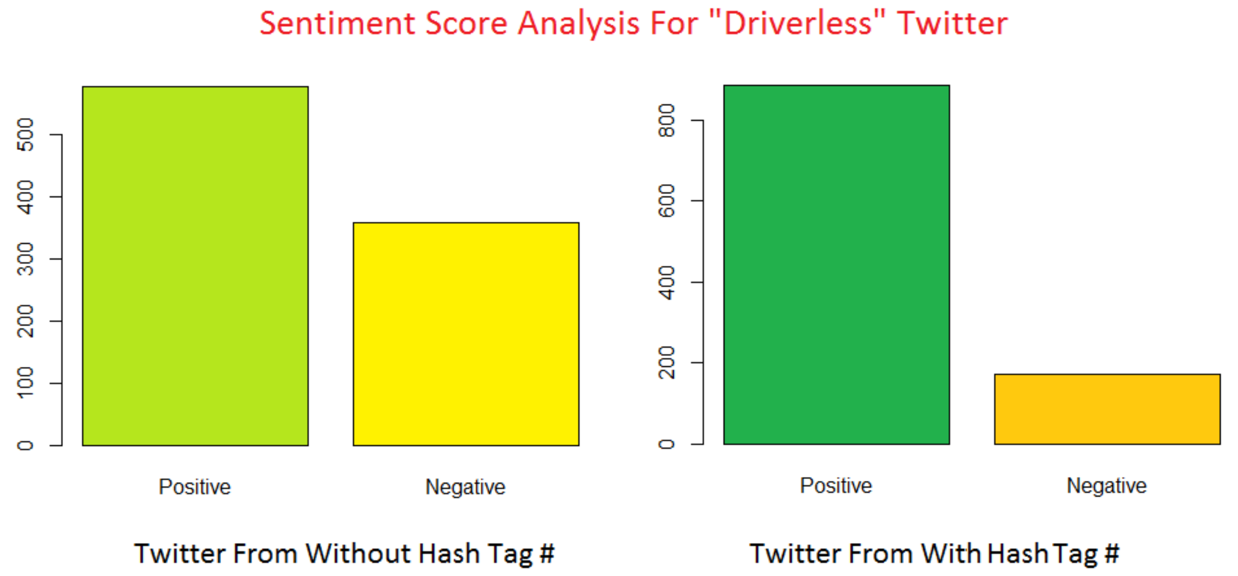
For the chart below, we have the analysis result further proving that the driverless technology is promising. The left graphic is for the twitter from without hash tag #, impacted by the google and uber event, the public opinion can change the attribute momentary. The left is for the twitter from hash tag#, it reflects a relatively stable view of the topic. The method is to use the R package syuzhet to get the score for each twitter and count for the total positive or negative for each dataset. Basically, the function get_nrc_sentiment from syuzhet works as the pre-built-in classifier for the text mining. It will evaluate each twitter and get the score value regarding the sentiment of the twitter, positive, negative or neutral. For example, the twitter #47 has the positive value of 2, and it ca have negative 1 together or all 0 on negative and positive. The chart doesn’t include the neutral count.



	anger	anticipation	disgust	fear	joy	sadness	surprise	trust	negative	positive
47	0	1	0	0	1	0	1	1	0	2

Code sample:

```
mySentiment <- get_nrc_sentiment(dlhatagchar)
```



Sentiment Score Analysis For "Driverless" Twitter

Twitter From Without Hash Tag #

Twitter From With Hash Tag #

Thank you very much for reviewing this blog. I also specially appreciate Professor Surendra Sarnikar, his excellent lectures of machine learning, data mining and data warehouse in California State University, East Bay.

References

1. AlphaGo | DeepMind https://deepmind.com/research/alphago/
2. Professor Surendra Sarnikar, Ph.D CSU, East Bay http://www.csueastbay.edu/directory/profiles/mgmt/sarnikarsurendra.html
3. Autonomous car https://en.wikipedia.org/wiki/Autonomous_car
4. Fury Road: Did Uber Steal the Driverless Future From Google? https://www.bloomberg.com/news/features/2017-03-16/fury-road-did-uber-steal-the-driverless-future-from-google
5. 'Safety Third' Is The Running Joke At Uber's Self-Driving Car Unit http://jalopnik.com/safety-third-is-the-running-joke-at-ubers-self-driving-1793368132