Data analysis, yelp dataset sentiment analysis

Dimple Sharma

Friday, December 12, 2014

This analysis provides Sentiment analysis of user reviews for the business. The sentiment analysis will involve following steps: * A list of words categorized by sentiments * A function which estimates sentiment score * Sentiment analysis of user reviews * What are the review counts for business? Is there a trend

Get both positive and negative words list for opinion mining

For opinion text analysis, we need a source that categorises words by sentiments.

Hu and Lius have a list of approximately 6800 positive and negative words categorized by sentiment which is also called opinion lexicon. The opinion lexicon list had been used in many research for sentiment analysis. The list is available on Bing Lui's website : Bing Liu's web site: http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar

```r
#setwd("c:/Dimple/RStats/")
#rm(list=ls())

hu_liu_pos <- scan('datayelp/opinion-lexicon-English/positive-words.txt',
                  what='character', comment.char=';')

hu_liu_neg <- scan('datayelp/opinion-lexicon-English/negative-words.txt',
                  what='character', comment.char=';')
```

Estimate sentiment score for each review

This function estimates a sentiment score for reviews.

The sentiment score function is written by Jeffrey Breen in an article which analyze twitter data: http://www.inside-r.org/howto/mining-twitter-airline-consumer-sentiment

The score sentiment function is modified to use for user review sentiment analysis.

```r
library(plyr)
library(stringr)
```

```
## Warning: package 'stringr' was built under R version 3.1.1
```

```r
score.sentiment = function(reviews, pos.words, neg.words, .progress='none')
{
        # Parameters
        # sentences: vector of text to score
        # pos.words: vector of words of postive sentiment
        # neg.words: vector of words of negative sentiment
        # .progress: passed to laply() to control of progress bar
        # create simple array of scores with laply
        scores = laply(reviews, function(review, pos.words, neg.words) {

        # clean up sentences
        review <- gsub('[[:punct:]]', '', review)       # Remove punctuaitons
        review <- gsub('[[:cntrl:]]', '', review)       # Remove control characters, if any
        review <- gsub('\\d+', '', review)              # Remove digits and spaces
        review <- tolower(review)    # Convert the text to lower case

        # split the sentences into words.
        word.list = str_split(review, '\\s+')
        # sometimes a list() is one level of hierarchy too much
        words = unlist(word.list)

        # compare our words to the dictionaries of positive & negative terms
        pos.matches = match(words, pos.words)
        neg.matches = match(words, neg.words)

        # The match methods returns a logical array of the matched term or NA
        # Keep the logical TRUE /FALSE values and discard the NAs
        pos.matches = !is.na(pos.matches)
        neg.matches = !is.na(neg.matches)

        # Get an estimate sentiment score, TRUE/FALSE will be treated as 1/0 by sum():
        score = sum(pos.matches) - sum(neg.matches)

        return(score)
        }, pos.words, neg.words, .progress=.progress )

        #scores.df = data.frame(score=scores, text=reviews)
        #return(scores.df)
        return(scores)
}
```

Sentiment analysis of user reviews

In this section we look at yelp user reviews data to know more the data and its attributes. We plot a review trends to know more about the no. of reviews the business has received.

We will deep dive into user reviews to get an estimate sentiment score. We make an estimate of sentiment score for each user review by subtracting the number of occurances of negative words from the positive words.

We will plot a histogram of scores to learn more about the distribution of scores.

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.1.1
```

```
##
## Attaching package: 'dplyr'
##
## The following objects are masked from 'package:plyr':
```

```
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(zoo)
```

```
##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.1.1
```

```
filename <- "datayelp/yelp_academic_dataset_review.csv"

data_review_all <- read.csv(filename,
                    header = T, stringsAsFactors = F, na.strings = c("","NA"))

# Analyze the data of yelp review dataset
dim(data_review_all)
```

```
## [1] 1125458      10
```

```
colnames(data_review_all)
```

```
##  [1] "user_id"      "review_id"    "text"         "votes.cool"
##  [5] "business_id"  "votes.funny"  "stars"        "date"
##  [9] "type"         "votes.useful"
```

```
# Get some sample rows of the data and summarize the columns
head(data_review_all)
```

```
##                  user_id              review_id
## 1 Xqd0DzHaiyRqVH3WRG7hzg 15SdjuK7DmYqUAj6rjGowg
## 2 H1kH6QZV7Le4zqTRNxoZow RF6UnRTtG7tWMcrO2GEoAg
## 3 zvJCcrpm2yOZrxKffwGQLA -TsVN230RCkLYKBeLsuz7A
## 4 KBLW4wJA_fwoWmMhiHRVOA dNocEAyUucjT371NNND41Q
## 5 zvJCcrpm2yOZrxKffwGQLA ebcN2aqmNUuYNoyvQErgnA
## 6 Qrs3EICADUKNFoUq2iHStA _ePLBPrkrf4bhyiKWEn4Qg
##
## 1
## 2
## 3
## 4
## 5
## 6 I don't know what Dr. Goldberg was like before  moving to Arizona, but let me tell you, STAY AWAY
##   votes.cool            business_id votes.funny stars       date   type
## 1          1 vcNAWiLM4dR7D2nwwJ7nCA           0     5 2007-05-17 review
## 2          0 vcNAWiLM4dR7D2nwwJ7nCA           0     2 2010-03-22 review
## 3          1 vcNAWiLM4dR7D2nwwJ7nCA           0     4 2012-02-14 review
## 4          0 vcNAWiLM4dR7D2nwwJ7nCA           0     4 2012-03-02 review
## 5          1 vcNAWiLM4dR7D2nwwJ7nCA           0     4 2012-05-15 review
## 6          0 vcNAWiLM4dR7D2nwwJ7nCA           0     1 2013-04-19 review
##   votes.useful
## 1            2
## 2            2
## 3            1
## 4            0
## 5            2
## 6            0
```

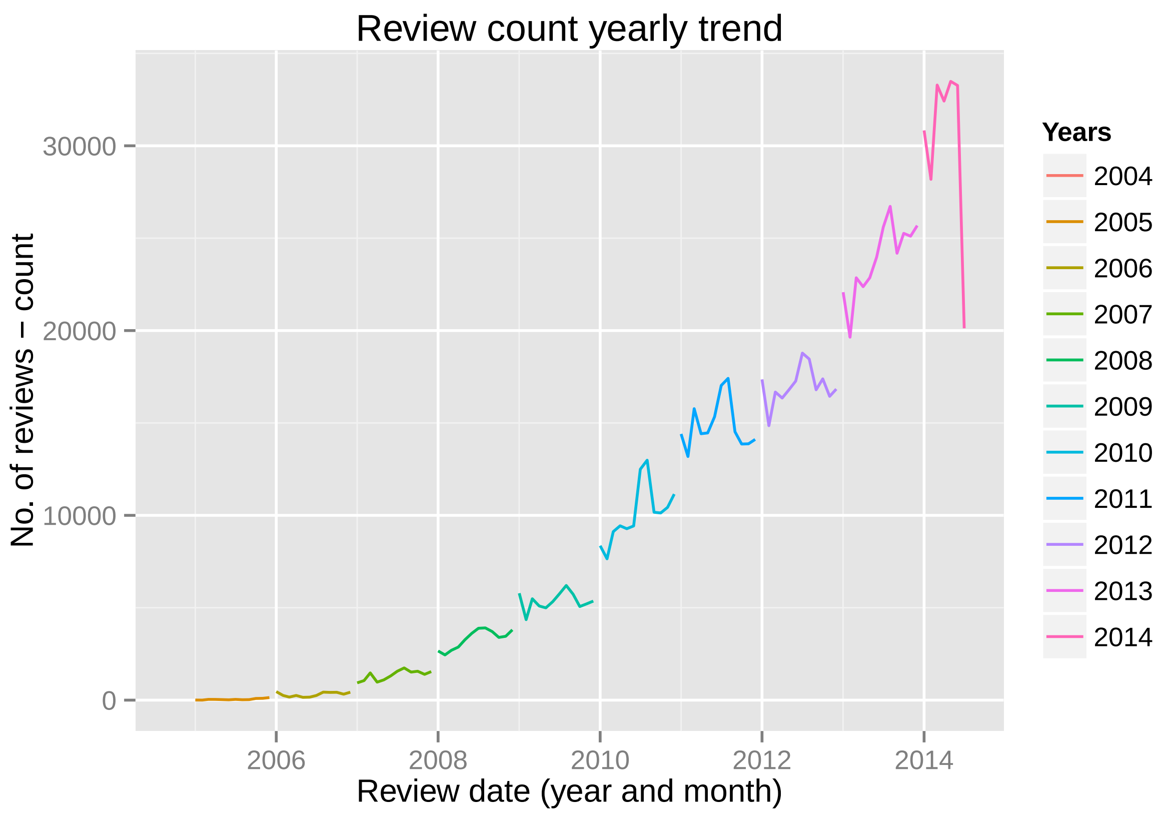
```r
tail(data_review_all)
```

```
##                         user_id              review_id
## 1125453 ZVkaYRCTFSSt-Ch9uwigOQ FOIB5Cx_iIY-FWKMH45VRw
## 1125454 lhMo-dGq0V2iKqBIiwUJSg eujuvkGqy2ssZ9zjdPJrMA
## 1125455 TTrzXCtB2MZA8Azw56bRlw vFA5KXUGEH-oMcM6WTC-8w
## 1125456 rtS7mDof5d-cEPBsmVuUJw 0sVK4VUxvj3cy78W0DlvWQ
## 1125457 tZs84cKAUSOtP_nAiSdreQ Nx88b_tCsP7Oja3PvhR5tQ
## 1125458 LfnB4N7SVSAIPOM3If_kDA p2KCsYuLHDtXqh1BpR9dGQ
##
## 1125453
## 1125454
## 1125455
## 1125456 Perfect little shop to go to if you want to pick up some Asian food real fast. Has a little
## 1125457
## 1125458
##         votes.cool            business_id votes.funny stars       date
## 1125453          1 BMjggIgOghBMEXPo8q7q3w           0     5 2014-07-16
## 1125454          0 BVxlrYWgmi-8TPGMe6CTpg           0     5 2010-08-11
## 1125455          1 BVxlrYWgmi-8TPGMe6CTpg           2     5 2012-06-15
## 1125456          1 BVxlrYWgmi-8TPGMe6CTpg           0     3 2013-09-17
## 1125457          1 BVxlrYWgmi-8TPGMe6CTpg           1     4 2013-09-18
## 1125458          0 BVxlrYWgmi-8TPGMe6CTpg           0     3 2013-12-17
##           type votes.useful
## 1125453 review            0
## 1125454 review            1
## 1125455 review            1
## 1125456 review            2
## 1125457 review            1
## 1125458 review            1
```

```r
summary(data_review_all)
```

```
##    user_id           review_id             text             votes.cool
##  Length:1125458     Length:1125458     Length:1125458     Min.   :  0.00
##  Class :character   Class :character   Class :character   1st Qu.:  0.00
##  Mode  :character   Mode  :character   Mode  :character   Median :  0.00
##                                                           Mean   :  0.65
##                                                           3rd Qu.:  1.00
##                                                           Max.   :137.00
##  business_id         votes.funny         stars          date
##  Length:1125458     Min.   :  0.00   Min.   :1.00   Length:1125458
##  Class :character   1st Qu.:  0.00   1st Qu.:3.00   Class :character
##  Mode  :character   Median :  0.00   Median :4.00   Mode  :character
##                     Mean   :  0.53   Mean   :3.74
##                     3rd Qu.:  0.00   3rd Qu.:5.00
##                     Max.   :141.00   Max.   :5.00
##      type            votes.useful
##  Length:1125458     Min.   :  0.00
##  Class :character   1st Qu.:  0.00
##  Mode  :character   Median :  0.00
##                     Mean   :  1.13
##                     3rd Qu.:  1.00
##                     Max.   :166.00
```

```r
# Get a trend for review counts
# To get a trend of reivew count, we summarize frequency of review by month and year
data_review_all$date <- as.Date(data_review_all$date, "%Y-%m-%d")
review_count_data <- data_review_all %>%
                        group_by(review_date = as.Date(as.yearmon(date))) %>%
                        summarise(review_count = n())

# What is the yearly trend of reviews for business.
# The following creates a yearly trend for reviews
review_graph <- ggplot( data = review_count_data, aes( x = review_date, y = review_count,
                        colour = factor(format(review_date, "%Y") )))
review_graph <- review_graph + geom_line()
review_graph <- review_graph + labs( title = "Review count yearly trend",
                                     x = "Review date (year and month)",
                                     y = "No. of reviews - count")
review_graph <- review_graph + scale_color_discrete(name="Years")
review_graph
```



```
# Estimate sentiment score for each reivew
scores <- score.sentiment(data_review_all$text, hu_liu_pos, hu_liu_neg, .progress = "text")
```

```
##
  |
  |                                                                      |   0%
  |
  |                                                                      |   1%
  |
  |=                                                                     |   1%
  |
  |=                                                                     |   2%
  |
  |==                                                                    |   2%
  |
  |==                                                                    |   3%
  |
  |==                                                                    |   4%
  |
  |===                                                                   |   4%
  |
  |===                                                                   |   5%
  |
  |====                                                                  |   5%
  |
  |====                                                                  |   6%
```

```
|
|====                                                             |   7%
|
|=====                                                            |   7%
|
|=====                                                            |   8%
|
|======                                                           |   8%
|
|======                                                           |   9%
|
|======                                                           |  10%
|
|=======                                                          |  10%
|
|=======                                                          |  11%
|
|=======                                                          |  12%
|
|========                                                         |  12%
|
|========                                                         |  13%
|
|=========                                                        |  13%
|
|=========                                                        |  14%
|
|=========                                                        |  15%
|
|==========                                                       |  15%
|
|==========                                                       |  16%
|
|===========                                                      |  16%
|
|===========                                                      |  17%
|
|===========                                                      |  18%
|
|============                                                     |  18%
|
|============                                                     |  19%
|
|=============                                                    |  19%
|
|=============                                                    |  20%
|
|=============                                                    |  21%
|
|==============                                                   |  21%
|
|==============                                                   |  22%
|
|===============                                                  |  22%
```

```
|
|===============                                                  |  23%
|
|===============                                                  |  24%
|
|================                                                 |  24%
|
|================                                                 |  25%
|
|=================                                                |  25%
|
|=================                                                |  26%
|
|=================                                                |  27%
|
|==================                                               |  27%
|
|==================                                               |  28%
|
|===================                                              |  28%
|
|===================                                              |  29%
|
|===================                                              |  30%
|
|====================                                             |  30%
|
|====================                                             |  31%
|
|====================                                             |  32%
|
|=====================                                            |  32%
|
|=====================                                            |  33%
|
|======================                                           |  33%
|
|======================                                           |  34%
|
|======================                                           |  35%
|
|=======================                                          |  35%
|
|=======================                                          |  36%
|
|========================                                         |  36%
|
|========================                                         |  37%
|
|========================                                         |  38%
|
|=========================                                        |  38%
|
|=========================                                        |  39%
```

```
|
|==========================                                       |  39%
|
|==========================                                       |  40%
|
|==========================                                       |  41%
|
|===========================                                      |  41%
|
|===========================                                      |  42%
|
|============================                                     |  42%
|
|============================                                     |  43%
|
|============================                                     |  44%
|
|=============================                                    |  44%
|
|=============================                                    |  45%
|
|==============================                                   |  45%
|
|==============================                                   |  46%
|
|==============================                                   |  47%
|
|===============================                                  |  47%
|
|===============================                                  |  48%
|
|================================                                 |  48%
|
|================================                                 |  49%
|
|================================                                 |  50%
|
|=================================                                |  50%
|
|=================================                                |  51%
|
|=================================                                |  52%
|
|==================================                               |  52%
|
|==================================                               |  53%
|
|===================================                              |  53%
|
|===================================                              |  54%
|
|===================================                              |  55%
|
|====================================                             |  55%
```

```
|
|====================================                             |  56%
|
|=====================================                            |  56%
|
|=====================================                            |  57%
|
|=====================================                            |  58%
|
|======================================                           |  58%
|
|======================================                           |  59%
|
|=======================================                          |  59%
|
|=======================================                          |  60%
|
|=======================================                          |  61%
|
|========================================                         |  61%
|
|========================================                         |  62%
|
|=========================================                        |  62%
|
|=========================================                        |  63%
|
|=========================================                        |  64%
|
|==========================================                       |  64%
|
|==========================================                       |  65%
|
|===========================================                      |  65%
|
|===========================================                      |  66%
|
|===========================================                      |  67%
|
|============================================                     |  67%
|
|============================================                     |  68%
|
|=============================================                    |  68%
|
|=============================================                    |  69%
|
|=============================================                    |  70%
|
|==============================================                   |  70%
|
|==============================================                   |  71%
|
|==============================================                   |  72%
```

```
|
|===============================================                  |  72%
|
|===============================================                  |  73%
|
|================================================                 |  73%
|
|================================================                 |  74%
|
|================================================                 |  75%
|
|=================================================                |  75%
|
|=================================================                |  76%
|
|==================================================               |  76%
|
|==================================================               |  77%
|
|==================================================               |  78%
|
|===================================================              |  78%
|
|===================================================              |  79%
|
|====================================================             |  79%
|
|====================================================             |  80%
|
|====================================================             |  81%
|
|=====================================================            |  81%
|
|=====================================================            |  82%
|
|======================================================           |  82%
|
|======================================================           |  83%
|
|======================================================           |  84%
|
|=======================================================          |  84%
|
|=======================================================          |  85%
|
|========================================================         |  85%
|
|========================================================         |  86%
|
|========================================================         |  87%
|
|=========================================================        |  87%
|
|=========================================================        |  88%
```

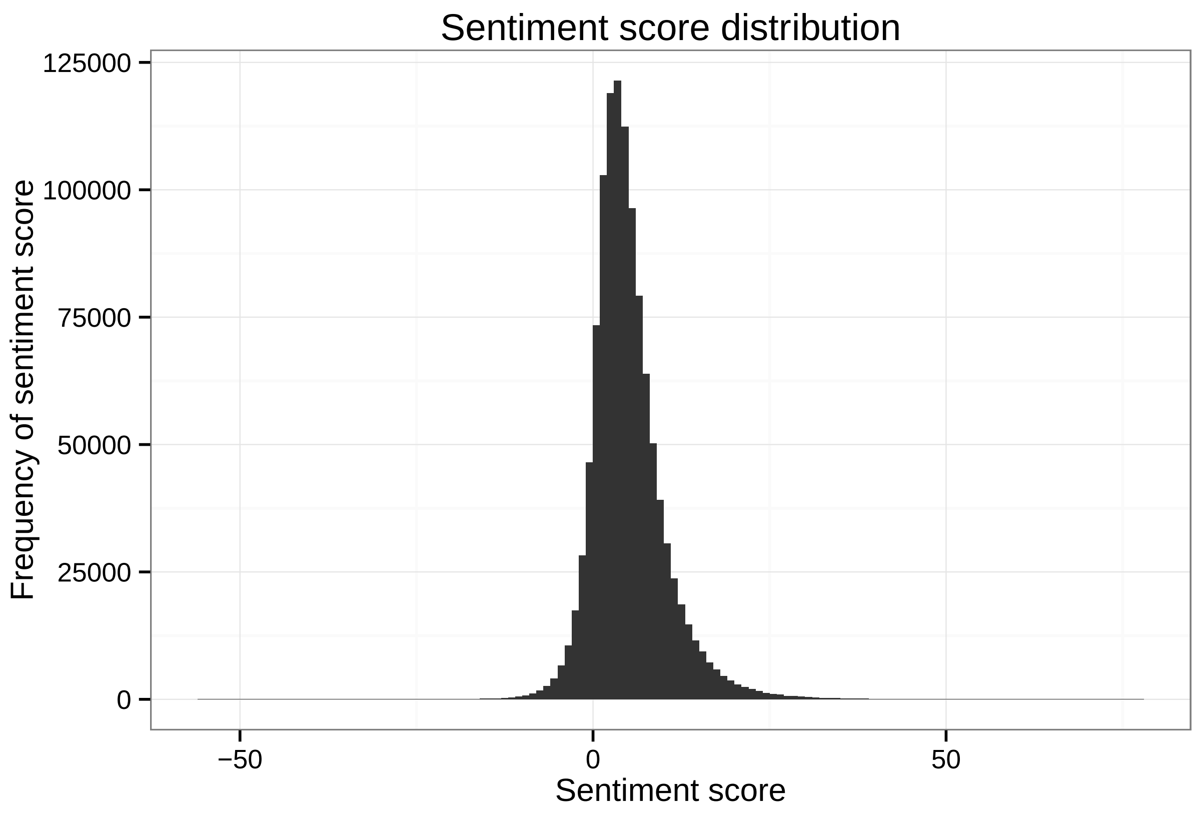
```
  |
  |==========================================================       |  88%
  |
  |==========================================================       |  89%
  |
  |==========================================================       |  90%
  |
  |===========================================================      |  90%
  |
  |===========================================================      |  91%
  |
  |===========================================================      |  92%
  |
  |============================================================     |  92%
  |
  |============================================================     |  93%
  |
  |=============================================================    |  93%
  |
  |=============================================================    |  94%
  |
  |=============================================================    |  95%
  |
  |==============================================================   |  95%
  |
  |==============================================================   |  96%
  |
  |===============================================================  |  96%
  |
  |===============================================================  |  97%
  |
  |===============================================================  |  98%
  |
  |================================================================ |  98%
  |
  |================================================================ |  99%
  |
  |=================================================================|  99%
  |
  |=================================================================| 100%
```

```
# Merge the scores with reviews data frame
data_review_all$scores <- scores

# What is the summary of scores, analyze the min, max, mean and quantile values
summary(scores)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -55.00    1.00    4.00    4.48    7.00   77.00
```

```
# Save the scores along with original data.
# Scores will be used for linear regression analysis
write.csv(data_review_all, file = filename, row.names = FALSE)
```

```r
# For memory efficency remove review text from reviews
#data_review <- subset(data_review_all, select = -c(text, date, user_id, review_id, type))

# A histogram will allow us to see a distribution of the score
graph_score <- ggplot(data_review_all, aes( x = scores)) + geom_bar(binwidth = 1)
graph_score <- graph_score + labs(x = "Sentiment score", y = "Frequency of sentiment score",
                                  title = "Sentiment score distribution")
graph_score <- graph_score + theme_bw()
graph_score
```


Sentiment score distribution
Frequency of sentiment score
Sentiment score