Yelp Dataset Review

John Antony

March 15, 2017

Loading required libraries

```
library(jsonlite)
library(stringr)
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tibble)
require(itertools)
```

```
## Loading required package: itertools
```

```
## Warning in library(package, lib.loc = lib.loc, character.only = TRUE,
## logical.return = TRUE, : there is no package called 'itertools'
```

Function to read data from JSON File and convert it into dataframe

```
createDF <- function(jsonFile){

  lines <- read_lines(jsonFile, n_max = 100000, progress = FALSE)
  combined <- str_c("[", str_c(lines, collapse = ", "), "]")
  remove(lines)
  df <- fromJSON(combined) %>%flatten() %>%tbl_df()

  df
}
```

Read the business and review json

```
business.df <- createDF('/Users/JohnAntony/Desktop/Main/Applications/R/MachineLearning/Yelp/yelp_dataset_challeng
e_round9/yelp_academic_dataset_business.json')

review.df <- createDF('/Users/JohnAntony/Desktop/Main/Applications/R/MachineLearning/Yelp/yelp_dataset_challenge_
round9/yelp_academic_dataset_review.json')
```

Analysis:By categorizing the popular types of restaurants that are present in a state, we can identify new opportunities for prospective business owners to start new projects.

For eg if the count of particular type of cusine is less in a particular state, it could be identified as potential opportunity to start that type of restaurant in that state or

if people in particular area prefer only one type of cuisine then it better to start the business

First, we filter the business data to get the Restaurant related information. The categories field provide this information but it is usually a list data type which contains all the variables that define the restaurant type like Fast food, burger, etc.,

```
business_flat <- flatten(business.df)
business_table <- as_data_frame(business_flat)
```

To understand what type of restaurant are more common in the dataset we need to unnest the categories list and assign a value to each row

```
business_table %>% mutate(categories = as.character(categories)) %>% select(categories)
```

```
## # A tibble: 100,000 × 1
##                                                                  categories
##                                                                       <chr>
## 1                    c("Tobacco Shops", "Nightlife", "Vape Shops", "Shopping")
## 2  c("Caterers", "Grocery", "Food", "Event Planning & Services", "Party & Even
## 3                         c("Restaurants", "Pizza", "Chicken Wings", "Italian")
## 4  c("Hair Removal", "Beauty & Spas", "Blow Dry/Out Services", "Hair Stylists"
## 5                   c("Hotels & Travel", "Event Planning & Services", "Hotels")
## 6                                          c("Nail Salons", "Beauty & Spas")
## 7                                     c("Baby Gear & Furniture", "Shopping")
## 8                          c("Tex-Mex", "Mexican", "Fast Food", "Restaurants")
## 9                                       c("Local Services", "Self Storage")
## 10                                                 c("Food", "Bakeries")
## # ... with 99,990 more rows
```

```
#Removing unneccessary variables
business_table %>%
  select(-starts_with("hours"), -starts_with("attribute"))
```

```
## # A tibble: 100,000 × 14
##              business_id                   name   neighborhood
## *                  <chr>                  <chr>          <chr>
## 1  0DI8Dt2PJp07XkVvIElIcQ      Innovative Vapors
## 2  LTlCaCGZE14GuaUXUGbamg          Cut and Taste
## 3  EDqCEAGXVGCH4FJXgqtjqg            Pizza Pizza Dufferin Grove
## 4  cnGIivYRLxpF7tBVR_JwWA    Plush Salon and Spa
## 5  cdk-qqJ71q6P7TJTww_DSA            Comfort Inn  Downtown Core
## 6  Q9rsaUiQ-A3NdEAloy0aJA            A Plus Nail
## 7  Cu4_Fheh7IrzGiK-Pc79ig         Boomerang Baby
## 8  GDnbt3isfhd57T1QqU6flg              Taco Bell
## 9  qwAHit4Tuj1zpO7CxVwOMA CubeSmart Self Storage
## 10 Nbr0kbtIrVlEcKIZoXWbSw        Sehne Backwaren
## # ... with 99,990 more rows, and 11 more variables: address <chr>,
## #   city <chr>, state <chr>, postal_code <chr>, latitude <dbl>,
## #   longitude <dbl>, stars <dbl>, review_count <int>, is_open <int>,
## #   categories <list>, type <chr>
```

```
#counting number of restaurants
library(stringr)
business_table %>% select(-starts_with("hours"), -starts_with("attribute")) %>%
  filter(str_detect(categories, "Restaurant"))
```

```
## # A tibble: 33,634 × 14
##               business_id                      name   neighborhood
##                     <chr>                     <chr>          <chr>
## 1  EDqCEAGXVGCH4FJXgqtjqg              Pizza Pizza Dufferin Grove
## 2  GDnbt3isfhd57T1QqU6flg                Taco Bell
## 3  42romV8altAeuZuP2OC1gw       Ohana Hawaiian BBQ
## 4  DNyYOxVAfu0oUcPNL1ljCQ              Chez Lionel
## 5  a1Ba6XeIOP48e64YFD0dMw                  La Prep    Ville-Marie
## 6  826djy6K_9Fp0ptqJ2_Yag    Chipotle Mexican Grill  Downtown Core
## 7  Mi5uhdFB9OJteXPd0_IKfw  Carrabba's Italian Grill
## 8  Uxh0fXFH_QQBivRnIBpdiw              Don Tequila
## 9  YPavuOh2XsnRbLfl0DH2lQ Lo-Lo's Chicken & Waffles
## 10 saWZO6hB4B8P-mIzS1--Xw             Kabob Palace  Spring Valley
## # ... with 33,624 more rows, and 11 more variables: address <chr>,
## #   city <chr>, state <chr>, postal_code <chr>, latitude <dbl>,
## #   longitude <dbl>, stars <dbl>, review_count <int>, is_open <int>,
## #   categories <list>, type <chr>
```

```
# filtering only Business column and count
business_table %>% select(-starts_with("hours"), -starts_with("attribute")) %>%
  filter(str_detect(categories, "Restaurant")) %>%
  mutate(categories = as.character(categories)) %>% select(categories)
```

```
## # A tibble: 33,634 × 1
##                                                                  categories
##                                                                       <chr>
## 1                       c("Restaurants", "Pizza", "Chicken Wings", "Italian")
## 2                         c("Tex-Mex", "Mexican", "Fast Food", "Restaurants")
## 3                                   c("Hawaiian", "Restaurants", "Barbeque")
## 4                                                c("Restaurants", "Cafes")
## 5               c("Sandwiches", "Breakfast & Brunch", "Salad", "Restaurants")
## 6                                   c("Fast Food", "Mexican", "Restaurants")
## 7                                     c("Restaurants", "Italian", "Seafood")
## 8                       c("Restaurants", "Mexican", "American (Traditional)")
## 9                        c("Restaurants", "Waffles", "Southern", "Soul Food")
## 10 c("Persian/Iranian", "Restaurants", "Ethnic Food", "Food", "Greek", "Specia
## # ... with 33,624 more rows
```

```
library(tidyr)
business_table %>% select(-starts_with("hours"), -starts_with("attribute")) %>%
  filter(str_detect(categories, "Restaurant")) %>%
  unnest(categories) %>%
  select(name, categories)
```

```
## # A tibble: 124,057 × 2
##                  name    categories
##                 <chr>         <chr>
## 1         Pizza Pizza   Restaurants
## 2         Pizza Pizza         Pizza
## 3         Pizza Pizza Chicken Wings
## 4         Pizza Pizza       Italian
## 5           Taco Bell       Tex-Mex
## 6           Taco Bell       Mexican
## 7           Taco Bell     Fast Food
## 8           Taco Bell   Restaurants
## 9  Ohana Hawaiian BBQ      Hawaiian
## 10 Ohana Hawaiian BBQ   Restaurants
## # ... with 124,047 more rows
```

```
# to get count of categories in the table
business_table %>% select(-starts_with("hours"), -starts_with("attribute")) %>%
  filter(str_detect(categories, "Restaurant")) %>%
  unnest(categories) %>%
  select(name, categories) %>%
  count(categories)
```

```
## # A tibble: 558 × 2
##             categories     n
##                  <chr> <int>
## 1           Acai Bowls    10
## 2          Accessories     1
## 3          Accountants     1
## 4          Active Life   121
## 5          Acupuncture     1
## 6      Adult Education     2
## 7  Adult Entertainment    10
## 8          Advertising     1
## 9               Afghan    60
## 10             African    87
## # ... with 548 more rows
```

What are the most common restaurant types per state / province?

```
# Getting top counts and getting rid of common tags like "Restautant" and "Food"

cat_table <- business_table %>% select(-starts_with("hours"), -starts_with("attribute")) %>%
  filter(str_detect(categories, "Restaurant")) %>%
  unnest(categories) %>%
  filter(categories != "Restaurants") %>%
  filter(categories != "Food") %>%
  count(state, categories) %>%
  arrange(desc(n))

library(DT)
datatable(cat_table, options = list(pageLength = 25))
```

Show 25 entries Search:

	state	categories	n
1	AZ	Fast Food	1053
2	AZ	Mexican	993
3	AZ	Sandwiches	959
4	AZ	American (Traditional)	932
5	ON	Chinese	910
6	AZ	Nightlife	868
7	AZ	Pizza	850
8	ON	Nightlife	834
9	AZ	Bars	832
10	ON	Bars	800
11	AZ	American (New)	730

	state	categories	n
12	NV	Fast Food	682
13	ON	Italian	667
14	AZ	Burgers	646
15	ON	Canadian (New)	636
16	ON	Sandwiches	626
17	ON	Pizza	611
18	NV	Mexican	595
19	NV	American (Traditional)	587
20	OH	American (Traditional)	571
21	ON	Breakfast & Brunch	570
22	ON	Japanese	565
23	AZ	Italian	556
24	NV	Nightlife	552
25	AZ	Breakfast & Brunch	550

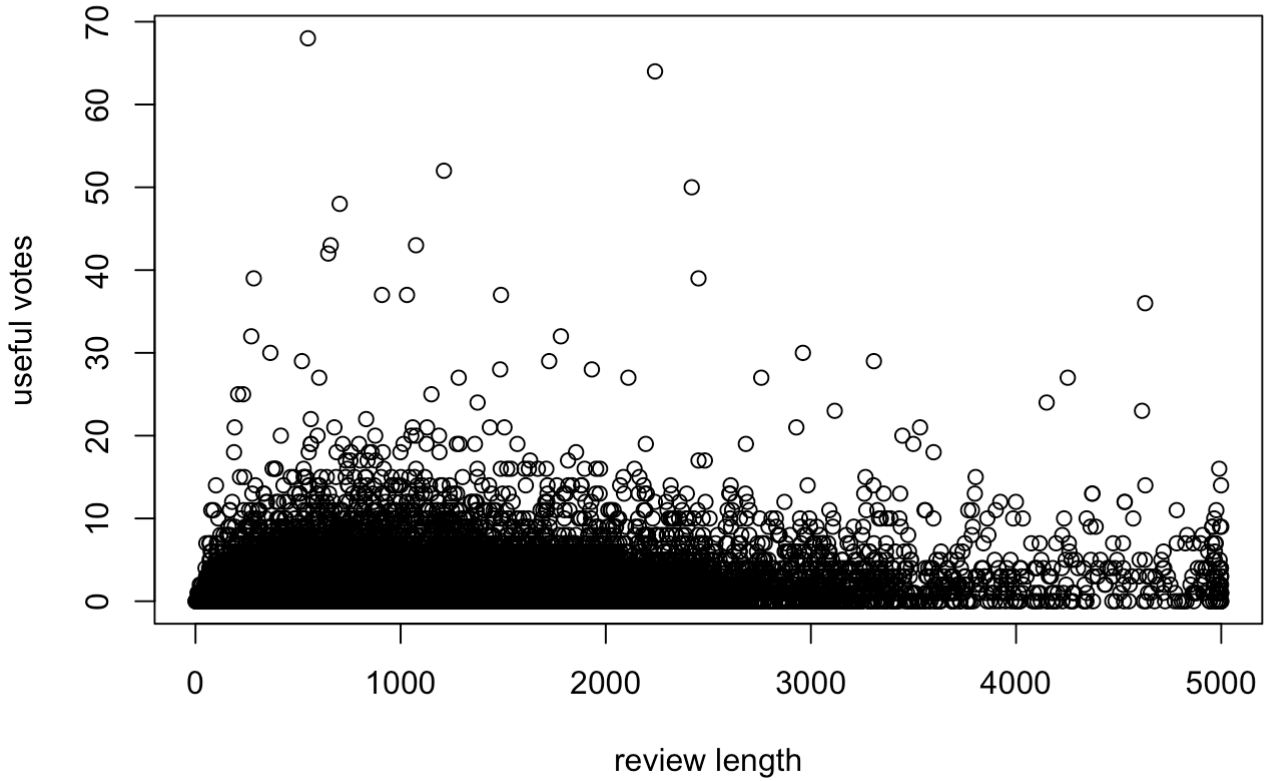
Showing 1 to 25 of 2,828 entries

Previous 1 2 3 4 5 … 114 Next

Analysis on whether lengthy reviews are useful

```
master <- merge(business.df, review.df, by = "business_id")

master$review_length <- nchar(master$text)
x <- master$review_length
y<-master$useful
plot(x,y,xlab="review length",ylab="useful votes")
```



#From the plot we can infer that

there is very little correlation between review length and usefulness of the review. THe more longer the review is the lesser the people interested to read the review.