Mining Customer Reviews for Business Opportunities

George Liu

November 17, 2015

Introduction

Yelp is an online service that allows users to review various businesses. Users can write detailed reviews and also rate the businesses using a 5-star scale. Yelp.com is available in 15 languages and has 142 million unique visitors per month. As a result, a huge amount of customer review data is generated on a regular basis. Customer feedback is one of the most important and informative data sources for business decision making. How can we turn the huge amount data generated by Yelp into business insights that are beneficial for the business community? In this project, we try to answer this question by examining data provided by the Yelp Dataset Challenge to garner business insights. The dataset comprises 1.6M reviews by 366K users for 61K businesses in 5 North American and European countries.

In particular, we'll study the ratings and reviews for different business categories and try to answer questions such as:

1. What are the lowest rated categories (such as restaurant, doctor and bars, in terms of stars)?
2. Can Yelp data be used to explore business opportunities - particularly those that can be taken advantage of by enterprises serving these industries? If yes, then:
3. What are the the top pain points in the lowest-rated industries from a consumer's standpoint (based on review contents)?

Methods and Data

We first need to find out the average or median stars (rating scores) of different categories so that comparison is possible. Then based on the result, we can select the lower rated categories and perform text mining on corresponding customer review data for insights. For category rating comparison, as the dataset provided covers all the review info for select cities and the cities were not randomly selected, the dataset represents population for the areas. Therefore, there is no need for statistical inference. In order to perform previously mentioned analysis on the data, "data wrangling" is necessary. The following is the detailed step-by-step actions taken for the analysis.

1. Get the data To get the data, we use "stream_in" function to read in the data, then use "flatten" function to convert the nested JSON structure to regular data frames. Finally, the data frame objects are saved in RDS files for easier later retrival.

2. Clean the data In order to answer the question which categories have the lowest rating, we need to calculate the ratings for all the categories. To do this, it is necessary to transform the data so that all the ratings appear alongside the businesses' corresponding categories. In the dataset provided, we do have business ID, category and corresponding stars available. However, the "categories" column has mixed data - some businesses have multiple categories recorded. To resolve this, we first need to decide on a list of categories to use. Using the category information available on Yelp's homepage (yelp.com), a list of 22 major categories is created. Next, we keep the categoy that is on the list for each business, for records where there are more than 1 major categories, random selection is used to determine the category for the business.

3. Analyze the data Once data wrangling is completed, the data is grouped based on categories, "average stars" and median stars are calculated. An ordering of categories from low to high is given. Since for the vast majority, average stars and median stars rankings agree(except for Professional Services and Education which have small population sizes), we use the average stars ranking to further investigate.

4. Prepare the data In order to understand the "why's" behind different categories' average ranking, we need to pull all the data that belong to certain categories and try to elicit insights using data mining techniques. We have all the review data in the review RDS file produced earlier, however, category information is absent. We then left-join the review data frame with custom made business data frame based on unique "business_id". Once this is done, we have a data frame that includes both review text and category info. This data frame can then be used for the followig text mining activities.

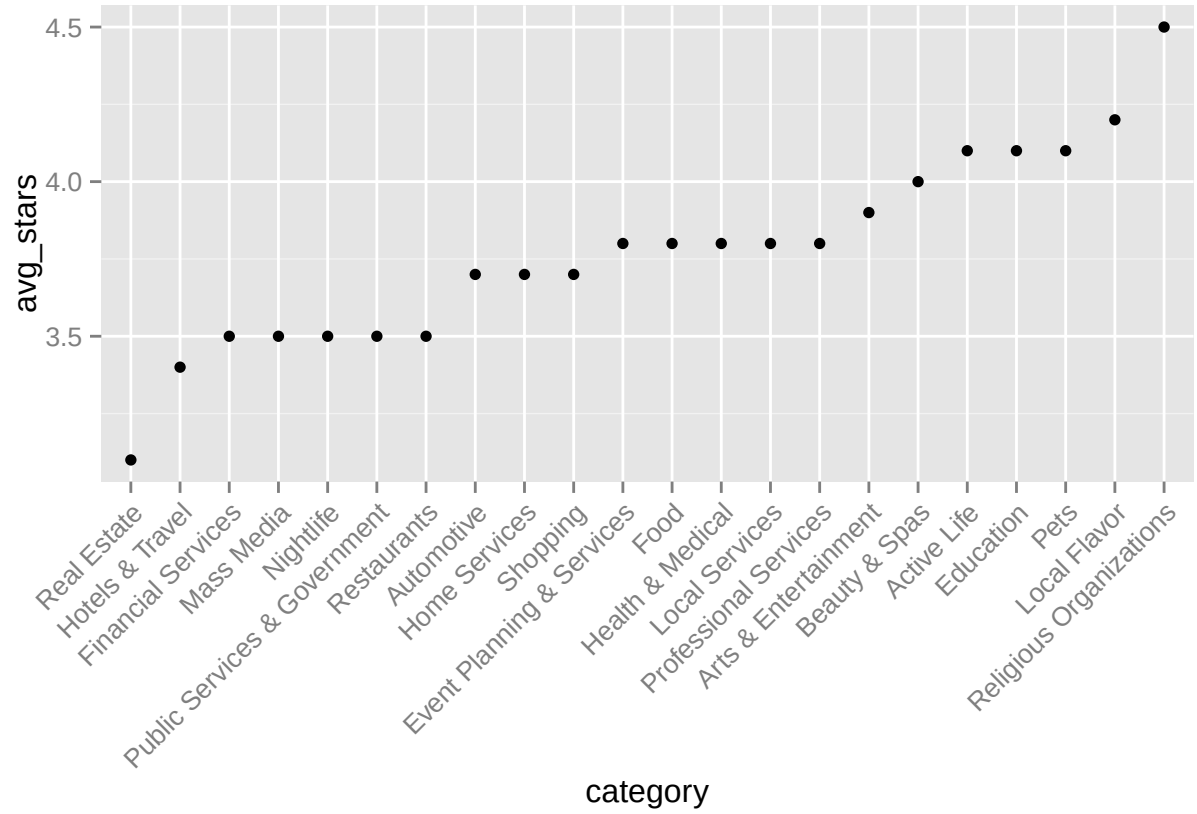
5. Perform text mining Since "Real Estate" category has the lowest ranking, we start by extracting all the review data for it. The review data is then converted into a "corpus" which goes through a series of pre-processing such as stemming, white space stripping etc. Once the text is ready for analysis, we find high frequency terms, plot word cloud and perform term relationship analysis using term correlation and clustering techniques. The correlated words with high frequency terms can then be used to infer hot topics in the review text.

Results

The Lowest Rated Categories

The top six categories with lowest average stars are Real Estate, Hotels & Travel etc. as shown below. In this report we'll focus on the Real Estate category which has an average rating of 3.1 and median rating of 3.0. We now show the ranking table and plot of all categories.

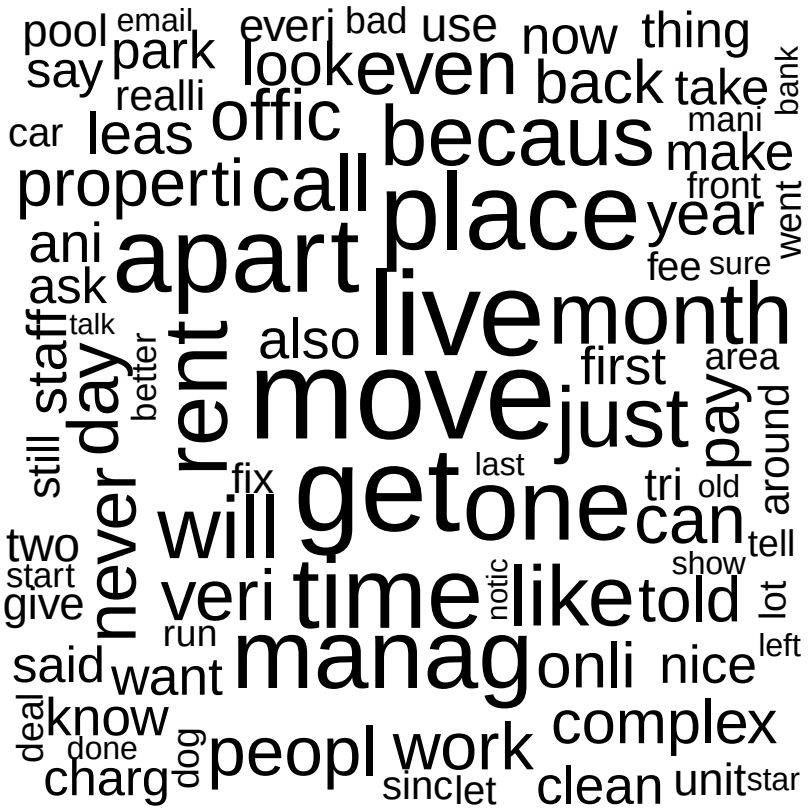
```
##                        category count avg_stars median_stars
## 1                   Real Estate   408       3.1          3.0
## 2                Hotels & Travel  1371       3.4          3.5
## 3             Financial Services   473       3.5          3.5
## 4                     Mass Media    91       3.5          3.5
## 5                      Nightlife  3052       3.5          3.5
## 6  Public Services & Government   395       3.5          3.5
## 7                    Restaurants 19792       3.5          3.5
## 8                     Automotive  2775       3.7          4.0
## 9                  Home Services  2137       3.7          4.0
## 10                      Shopping  7626       3.7          4.0
## 11    Event Planning & Services  1468       3.8          4.0
## 12                         Food  6366       3.8          4.0
## 13             Health & Medical  2920       3.8          4.0
## 14               Local Services  1810       3.8          4.0
## 15        Professional Services   433       3.8          4.5
## 16         Arts & Entertainment  1336       3.9          4.0
## 17                Beauty & Spas  4235       4.0          4.0
## 18                  Active Life  2159       4.1          4.0
## 19                    Education   382       4.1          4.5
## 20                         Pets  1130       4.1          4.0
## 21                 Local Flavor   151       4.2          4.5
## 22      Religious Organizations   157       4.5          4.5
```



Text Mining Results

First, let's look at the most frequent terms and the word cloud.

```
##     move      get     live    place    apart     time      one    manag
##     1203     1120     1112     1001      987      920      907      883
##     rent     will    month     just     like   becaus     call     even
##      799      778      761      751      748      737      695      664
##      day    never    offic     veri      can    peopl properti     work
##      607      605      593      588      562      562      549      544
##     onli     told     year     look      pay  complex     back     leas
##      509      506      496      495      493      479      476      465
##     also    staff      ani     want     nice    first     make      now
##      431      426      422      412      410      404      402      402
##     park     know    clean     take      ask    charg     said      say
##      399      397      388      383      382      367      366      365
##    thing     good
##      365      364
```



Now find relationships between terms with term correlations (only first 6 items shown below):

```
## $get
##   absolutley       adding       agenda     apraisal         bids
##         0.43         0.43         0.43         0.43         0.43
##         ceci    conducted    consigned consignmenth     explicit
##         0.43         0.43         0.43         0.43         0.43
##       failed         gari       goinhg         hosu     housenev
##         0.43         0.43         0.43         0.43         0.43
##   imediatley    interests       leeway      listnig        needd
##         0.43         0.43         0.43         0.43         0.43
##         ored       pushed      relianc       reposn       scrawl
##         0.43         0.43         0.43         0.43         0.43
##         snse          bid          war          mls      brother
##         0.43         0.42         0.41         0.39         0.38
##        buyer   contractor       market      passing        sales
##         0.33         0.33         0.33         0.32         0.32
##          add         hose        title
##         0.31         0.30         0.30
##
## $place
##        able        adam      bodili        emit  especially     extract
##        0.31        0.31        0.31        0.31        0.31        0.31
##         hub      inland  inoperable    jeopardi      occurs    overheat
##        0.31        0.31        0.31        0.31        0.31        0.31
##    reported restrictors     scammer
```

```
##        0.31        0.31        0.31
##
## $rent
## paid
## 0.31
##
## $month
##   per addit
##  0.32  0.31
##
## $like
## sound
##  0.31
##
## $becaus
## made
## 0.31
##
## $call
## later  made
##   0.3   0.3
##
## $even
## though
##   0.39
##
## $day
## notic waltz
##  0.33  0.32
##
## $offic
## office
##   0.31
```

Discussion

In summary, the research does reveal the lowest rated categories on Yelp, also provides data to gather insights about opportunities in the industry, showing there are numerous pain points in the real estate category (mainly apartment renting). These problem areas include:

- biased lease
- staff issue (not responsive, rude)
- price confusion (lack of accuracy)
- environment problems (cigs, squirrel, syringes, neighbourhood, common area, parking/car towing, hygiene)
- hardware problems (floor/carpet issue, baths, electricity, stuffy, hot water, water leak, hvac, door)

These pain points can then be considered by exising players, new entrants or businesses serving the industry to improve service quality, design new services or create business strategies.