Analysis of relation between Stars rating vs. most relevant Attributes and n-grams (words) in Reviews filtered by State and Business Category.

Jose Maria Vilaverde

November 22th 2015

Introduction

In order to get top rate as business in a city is analyzed where have to be established my business, which services I need to offer, what positive review words I need to be associated to my business and what negative review words I need to avoid

To do that analysis is used as input Business Category and State. As output: Top Rate, Location-Neighborhood, Services, Top-5 positive words to promote, Top-5 negative words to avoid.

For example, If I want to open a business for category "dentist" in Arizona, I need to know where is the best place to open the business, services that I need to offer to my customers, like "credit card accepted", and identify most relevant positive review words that I need to get from my customers and negative review words for my business category in Arizona.

Methods and Data

Steps:

1.ETL -> 1.Get_Data.R

1.1.Download data from https://d396qusza40orc.cloudfront.net/dsscapstone/dataset/yelp_dataset_challenge_academic_dataset.zip and unzip it.

1.2.Extract json information from files "business","checkin","tip","review" (Excluded "user", not relevant for my analysis)

1.3.Store it into RDS file. Review is divided into files with 100.000 lines/file in order to make affordable the calculation for my computer. For review processing are obtained 16 files.

1.4.Identify business per state.

1.5.Filter files per business_id per State, create RDS files "checkin","tip","review" per State.

2.Process -> 2.Proces Data.R

Obtain auxiliar dataset

2.1.Use as input State and Business Category.

List of States:

[1] "AZ, BW, CA, EDH, ELN, FIF, HAM, IL, KHL, MA, MLN, MN, NC, NTH, NV, NW, ON, OR, PA, QC, RP, SC, SCB, WA, WI, XGL"

List of Categories:

[1] "(Other), Accessories, Active Life, American (New), American (Traditional), Apartments, Arts & Crafts, Arts & Entertainment, Asian Fusion, Auto Repair, Automotive, Bakeries, Barbeque, Bars, Beauty & Spas, Beer, Wine & Spirits, Books, Mags, Music & Video, Breakfast & Brunch, Burgers, Cafes, Caterers, Chicken Wings, Chinese, Coffee & Tea, Convenience Stores, Cosmetics & Beauty Supply, Day Spas, Delis, Dentists, Department Stores, Desserts, Diners, Doctors, Drugstores, Dry Cleaning & Laundry, Electronics, Event Planning & Services, Fashion, Fast Food, Fitness & Instruction, Flowers & Gifts, Food, Food Trucks, Furniture Stores, Greek, Grocery, Gyms, Hair Removal, Hair Salons, Health & Medical, Home & Garden, Home Decor, Home Services, Hotels, Hotels & Travel, Ice Cream & Frozen Yogurt, Indian, Italian, Japanese, Jewelry, Latin American, Local Services, Lounges, Massage, Mediterranean, Men's Clothing, Mexican, NA's, Nail Salons, Nightlife, Oil Change Stations, Parks, Pet Boarding/Pet Sitting, Pet Groomers, Pet Services, Pet Stores, Pets, Pizza, Professional Services, Public Services & Government, Pubs, Real Estate, Restaurants, Salad, Sandwiches, Seafood, Shoe Stores, Shopping, Southern, Specialty Food, Sporting Goods, Sports Bars, Steakhouses, Sushi Bars, Tex-Mex, Tires, Venues & Event Spaces, Veterinarians, Wine Bars, Women's Clothing"

2.2.Intermediate Output to support analysis and define model

2.3.Relation Attributes-Stars

2.4.Count words.

Get output

2.5.Attributes to have, top-5 positive words (>=4 stars comments), top-5 negatives words (<3 stars comments).

2.6.Map of top rated business.

2.7.Random forest model fitted for a Business Category and State and analysis to verify that all those attributes and words can predict star rates.

Results

Used as State **“NC”** and **“Food”** as category.

List of top attributes and Top-5 positive and negative words for 1-gram

List filtered, % of positive and negative over total must be over 2.5%, difference between positive and negative must be over 0.15 stars to be relevant.

Attribute	positives	pos_avg	negatives	neg_avg	diff_avg	pos%	neg%
attributes.Parking.street	72	4.07	600	3.71	0.36	10.71	89.29
attributes.Accepts Credit Cards	639	3.76	33	3.50	0.26	95.09	4.91
attributes.Price Range	647	3.76	25	3.58	0.18	96.28	3.72
attributes.Parking.garage	61	3.90	611	3.73	0.17	9.08	90.92
attributes.Good For.breakfast	20	3.90	652	3.74	0.16	2.98	97.02
attributes.Caters	76	3.88	596	3.73	0.15	11.31	88.69

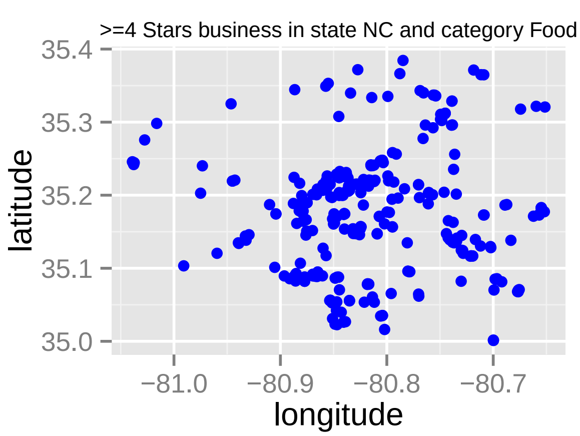
	word	accumulate
1	good	38.00
2	range	23.00
3	food	20.00
4	place	18.00
5	service	15.00

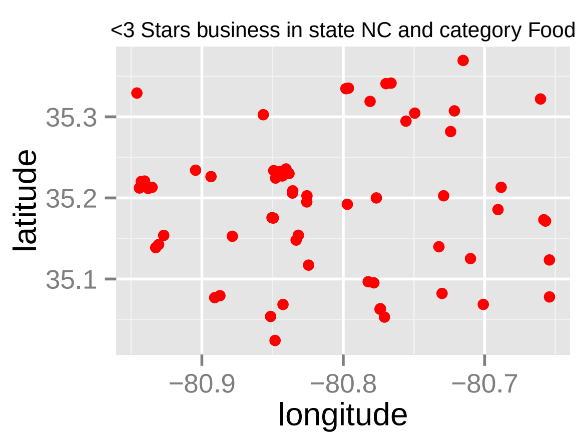
Table 1: Top-5 positive words

	word	accumulate
1	food	21.00
2	time	14.00
3	good	12.00
4	place	12.00
5	no	11.00

Table 2: Top-5 negative words

Map of Business Stars Rates





Random Forest Model

Formula

stars ~ attributes.Parking.validated + attributes.Ambience.hipster + attributes.Parking.street + attributes.Accepts.Credit.Cards + attributes.Music.live + attributes.Price.Range + attributes.Parking.garage + attributes.Good.For.breakfast + attributes.Takes.Reservations + attributes.Caters + attributes.Good.For.Groups + attributes.Has.TV + attributes.Wheelchair.Accessible + attributes.Ambience.casual + attributes.Good.for.Kids + attributes.Parking.lot + attributes.Good.For.dinner + latitude + longitude

Confusion Matrix Table

	1.5	2	2.5	3	3.5	4	4.5	5
1.5	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0
2.5	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	1	0
3.5	0	0	1	0	2	2	0	0
4	1	2	4	16	25	27	18	6
4.5	0	0	3	2	4	7	8	2
5	0	0	0	0	0	0	0	0

Table 3: Confussion Matrix

Confusion Matrix Overall

Accuracy	Kappa
0.2824427	0.0320704

Confusion Matrix by Class

```
##             Sensitivity Specificity Pos Pred Value Neg Pred Value
## Class: 1.5  0.00000000   1.0000000            NaN      0.9923664
## Class: 2    0.00000000   1.0000000            NaN      0.9847328
## Class: 2.5  0.00000000   1.0000000            NaN      0.9389313
## Class: 3    0.00000000   0.9911504      0.0000000      0.8615385
## Class: 3.5  0.06451613   0.9700000      0.4000000      0.7698413
## Class: 4    0.75000000   0.2421053      0.2727273      0.7187500
## Class: 4.5  0.29629630   0.8269231      0.3076923      0.8190476
## Class: 5    0.00000000   1.0000000            NaN      0.9389313
##              Prevalence Detection Rate Detection Prevalence
## Class: 1.5 0.007633588     0.00000000          0.000000000
## Class: 2   0.015267176     0.00000000          0.000000000
## Class: 2.5 0.061068702     0.00000000          0.000000000
## Class: 3   0.137404580     0.00000000          0.007633588
## Class: 3.5 0.236641221     0.01526718          0.038167939
## Class: 4   0.274809160     0.20610687          0.755725191
## Class: 4.5 0.206106870     0.06106870          0.198473282
## Class: 5   0.061068702     0.00000000          0.000000000
##            Balanced Accuracy
## Class: 1.5         0.5000000
## Class: 2           0.5000000
## Class: 2.5         0.5000000
## Class: 3           0.4955752
```

```
## Class: 3.5         0.5172581
## Class: 4           0.4960526
## Class: 4.5         0.5616097
## Class: 5           0.5000000
```

Discussion

Attributes Analysis

At initial analysis are detected attributes that have an impact over $\pm 0.15 stars rate$. Positive % (pos% = $\frac{positive reviews}{total reviews}$) and negative % (neg% = $\frac{negative reviews}{total reviews}$) must be >2.5% to avoid to interprete exceptions as true influencers of stars rate.

With this conditions, are identied some attibutes with an influence between (0.36, 0.15). This attributes and latitude and longitude are the factors for the model.

Top-5 positive and negative 1-gram

Top-5 positive and negative 1-gram analysis appears as no useful in this case, because some 1-gram detected as top positive appears also as most frequent into top negative 1-gram, due to this, these are not included into the model. For a small-medium set of reviews is not useful this kind of analysis, is recomended to try use it over greater data set of review or include 2-grams/3-grams to evaluate if are relevant is that scenario.

Map of business

Map of business shows interesting information, comparing ≥ 4 stars (best rated business) vs. <3 stars (worse rated business) is easy to see the difference between both maps. Moreover, appears an accumulation of best rated business that indicates areas where clients are attracted due to good reviews.

Random Forest Model

Random Forest Model reveals that hypotesis about influence of attributes alone are not useful to predict rates based on attributes and location. Confussion Matrix Table reveals a poor accuracy into prediction, less than 30%, and appears totaly inefficient to classificate 1.5, 2, 2.5, 3 and 5 stars, and low accuracy for 3.5, 4 and 4.5.

Implication for the question

Based on this analysis, in one hand, use of attributes as stars rate predictor and Top-5 positive and negative 1-gram is inefficient, can't be used to recomend to business owner attributes to have or words that have to make arise in reviews. Maybe is possible to recommend some attributes, but is not confirmed their direct influence in stars rate. In the other hand, Map of business can be useful to stablish location of business in a specific category.