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 City. As output: Top Rate, 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 have to give 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:

ETL -> 1.Get_Data.R

- Download data from https://d396qusza40orc.cloudfront.net/dsscapstone/dataset/yelp_dataset_challenge_academic_dataset.zip and unzip it.
- Extract json information from files "business","checkin","tip","review" (Excluded "user", not relevant for my analysis)
- 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 I obtained 16 files.
- Identify business per state.
- Filter files per business_id per State, create RDS files "checkin","tip","review" per State.

Process -> 2.Proces Data.R

- 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"

- Intermediate Output to support analysis and define model
- Relation Attributes-Stars
- Count words.

Output

- Attributes to have, top-5 positive words (>=4 stars comments), top-5 negatives words (<3 stars comments).
- Map of top rated business.
- Random forest model fitted for a Business Category and State and analysis to verify that all those attributes and words have a correlation with star rates.

Results - Describe what you found through your analysis of the data.

- Use 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

---

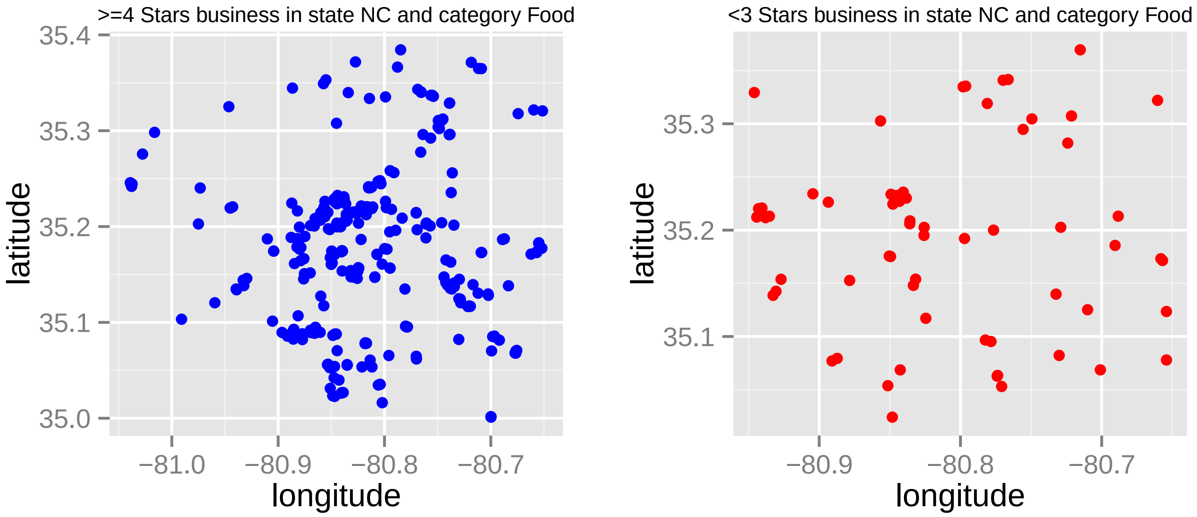
	word	accumulate
1	good	38.00
2	range	23.00
3	food	20.00
4	place	18.00
5	service	15.00

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:



Model:

```
## 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
## <environment: 0x000000000abe1f58>
## ntree      OOB      1       2       3       4       5       6       7       8
##    10:  75.12%100.00%100.00%100.00% 91.53% 83.84% 36.75% 87.36% 96.15%
##    20:  75.12%100.00%100.00%100.00% 93.33% 88.00% 27.97% 92.05%100.00%
##    30:  76.50%100.00%100.00%100.00% 98.33% 91.00% 28.81% 90.91%100.00%
##    40:  74.88%100.00%100.00%100.00% 98.33% 85.00% 28.81% 89.77%100.00%
##    50:  75.58%100.00%100.00%100.00% 98.33% 87.00% 31.36% 87.50%100.00%
## ntree      OOB      1      2      3      4      5      6      7      8
##    10:  74.71%100.00%100.00% 81.48% 88.14% 68.00% 70.34% 68.97% 88.46%
##    20:  76.04%100.00%100.00% 85.71% 90.00% 69.00% 70.34% 70.45% 92.31%
##    30:  76.04%100.00%100.00% 85.71% 86.67% 74.00% 70.34% 69.32% 84.62%
##    40:  76.73%100.00%100.00% 85.71% 83.33% 72.00% 72.03% 73.86% 88.46%
##    50:  76.04%100.00%100.00% 89.29% 85.00% 75.00% 66.10% 73.86% 84.62%
## ntree      OOB      1      2      3      4      5      6      7      8
##    10:  80.23%100.00%100.00% 89.29% 86.67% 79.00% 72.41% 81.40% 80.77%
##    20:  77.88%100.00%100.00% 85.71% 78.33% 76.00% 72.03% 77.27% 92.31%
##    30:  78.11%100.00%100.00% 89.29% 90.00% 78.00% 67.80% 73.86% 88.46%
##    40:  77.65%100.00%100.00% 85.71% 86.67% 79.00% 66.95% 73.86% 92.31%
##    50:  75.81%100.00%100.00% 82.14% 85.00% 76.00% 66.10% 71.59% 92.31%
## ntree      OOB      1      2      3      4      5      6      7      8
##    10:  77.44%100.00%100.00%100.00%100.00% 83.67% 52.14% 71.26%100.00%
##    20:  75.69%100.00%100.00%100.00%100.00% 85.86% 37.61% 79.55%100.00%
##    30:  78.70%100.00%100.00%100.00%100.00% 93.94% 40.17% 81.82%100.00%
##    40:  74.54%100.00%100.00%100.00%100.00% 88.89% 35.90% 72.73%100.00%
##    50:  75.23%100.00%100.00%100.00%100.00% 89.90% 38.46% 71.59%100.00%
## ntree      OOB      1      2      3      4      5      6      7      8
##    10:  80.09%100.00%100.00% 96.30% 89.47% 80.41% 68.42% 77.01% 92.59%
##    20:  76.85%100.00%100.00% 86.21% 93.22% 72.73% 64.10% 75.00% 96.30%
##    30:  74.07%100.00% 88.89% 89.66% 88.14% 71.72% 57.26% 75.00% 96.30%
##    40:  74.54%100.00% 88.89% 86.21% 88.14% 76.77% 56.41% 73.86% 96.30%
##    50:  75.69%100.00% 88.89% 86.21% 91.53% 76.77% 58.12% 75.00% 96.30%
## ntree      OOB      1      2      3      4      5      6      7      8
##    10:  81.54%100.00% 88.89% 82.14% 93.22% 81.82% 73.50% 77.65% 96.30%
##    20:  77.31%100.00%100.00% 79.31% 91.53% 73.74% 70.09% 71.59% 96.30%
##    30:  78.47%100.00% 88.89% 79.31% 94.92% 74.75% 68.38% 77.27% 96.30%
##    40:  78.94%100.00%100.00% 79.31% 94.92% 75.76% 65.81% 80.68% 96.30%
##    50:  78.47%100.00% 88.89% 79.31% 89.83% 79.80% 66.67% 76.14%100.00%
## ntree      OOB      1      2      3      4      5      6      7      8
##    10:  76.67%100.00%100.00%100.00% 96.61% 90.91% 47.46% 67.82%100.00%
##    20:  77.19%100.00%100.00%100.00%100.00% 89.90% 36.44% 84.09%100.00%
##    30:  75.12%100.00%100.00%100.00%100.00% 89.90% 26.27% 87.50%100.00%
##    40:  74.88%100.00%100.00%100.00%100.00% 90.91% 22.88% 89.77%100.00%
##    50:  74.65%100.00%100.00%100.00%100.00% 88.89% 25.42% 87.50%100.00%
## ntree      OOB      1      2      3      4      5      6      7      8
##    10:  78.84%100.00%100.00% 86.21% 84.48% 72.16% 74.36% 78.41% 92.59%
##    20:  77.88%100.00%100.00% 79.31% 88.14% 74.75% 73.73% 71.59% 92.59%
##    30:  76.73%100.00%100.00% 86.21% 88.14% 68.69% 69.49% 77.27% 88.89%
##    40:  79.26%100.00%100.00% 93.10% 88.14% 71.72% 73.73% 78.41% 88.89%
##    50:  76.50%100.00%100.00% 86.21% 93.22% 71.72% 68.64% 71.59% 85.19%
## ntree      OOB      1      2      3      4      5      6      7      8
##    10:  75.64%100.00%100.00% 79.31% 82.76% 69.07% 70.34% 75.00% 92.59%
##    20:  75.81%100.00%100.00% 89.66% 86.44% 61.62% 73.73% 75.00% 88.89%
##    30:  75.12%100.00%100.00% 89.66% 86.44% 65.66% 72.88% 68.18% 88.89%
##    40:  73.73%100.00%100.00% 89.66% 83.05% 63.64% 72.03% 68.18% 85.19%
##    50:  72.81%100.00%100.00% 86.21% 86.44% 61.62% 70.34% 67.05% 85.19%
## ntree      OOB      1      2      3      4      5      6      7      8
##    10:  74.76%100.00%100.00% 96.43% 96.55% 87.50% 27.83% 90.91% 96.15%
##    20:  75.64%100.00%100.00%100.00% 94.92% 85.86% 29.91% 93.18%100.00%
##    30:  72.39%100.00%100.00%100.00% 98.31% 81.82% 24.79% 86.36%100.00%
##    40:  73.78%100.00%100.00%100.00% 98.31% 86.87% 23.93% 88.64%100.00%
##    50:  71.69%100.00%100.00%100.00% 98.31% 90.91% 17.09% 82.95%100.00%
## ntree      OOB      1      2      3      4      5      6      7      8
##    10:  77.78%100.00% 88.89% 93.10% 82.46% 72.63% 76.72% 70.11% 92.31%
##    20:  77.26%100.00% 88.89% 89.66% 89.83% 74.75% 67.52% 71.59%100.00%
##    30:  77.03%100.00% 77.78% 93.10% 91.53% 70.71% 67.52% 73.86%100.00%
##    40:  74.01%100.00% 88.89% 89.66% 91.53% 66.67% 64.10% 69.32% 96.15%
##    50:  75.41%100.00% 77.78% 89.66% 88.14% 76.77% 61.54% 71.59% 96.15%
## ntree      OOB      1      2      3      4      5      6      7      8
##    10:  78.22%100.00% 88.89% 86.21% 86.44% 75.51% 67.83% 79.31% 96.15%
##    20:  77.03%100.00% 88.89% 93.10% 93.22% 73.74% 61.54% 78.41% 92.31%
##    30:  77.73%100.00% 88.89% 89.66% 89.83% 76.77% 64.10% 77.27% 96.15%
##    40:  76.80%100.00% 77.78% 93.10% 89.83% 76.77% 62.39% 75.00% 96.15%
##    50:  78.19%100.00% 77.78% 89.66% 93.22% 80.81% 61.54% 77.27% 96.15%
## ntree      OOB      1      2      3      4      5      6      7      8
##    10:  76.46%100.00%100.00% 96.55% 88.14% 88.54% 37.29% 92.05% 96.00%
##    20:  75.98%100.00%100.00%100.00% 96.61% 75.76% 39.83% 92.05%100.00%
##    30:  75.75%100.00%100.00%100.00% 96.61% 83.84% 30.51% 95.45% 96.15%
##    40:  75.06%100.00%100.00%100.00% 98.31% 78.79% 31.36% 95.45% 96.15%
##    50:  74.36%100.00%100.00%100.00% 98.31% 79.80% 27.97% 95.45% 96.15%
## ntree      OOB      1      2      3      4      5      6      7      8
##    10:  76.22%100.00% 90.00%100.00% 89.83% 68.75% 63.25% 76.14% 96.15%
##    20:  75.98%100.00% 90.00% 96.55% 84.75% 71.72% 62.71% 77.27% 96.15%
##    30:  76.44%100.00% 90.00%100.00% 89.83% 73.74% 62.71% 75.00% 88.46%
##    40:  75.29%100.00% 90.00% 96.55% 88.14% 71.72% 64.41% 70.45% 92.31%
##    50:  77.83%100.00% 90.00% 93.10% 89.83% 71.72% 70.34% 73.86% 96.15%
## ntree      OOB      1      2      3      4      5      6      7      8
##    10:  80.00%100.00% 70.00% 89.66% 87.93% 76.77% 73.28% 81.82% 88.46%
##    20:  77.83%100.00% 80.00% 96.55% 89.83% 73.74% 63.56% 80.68% 96.15%
##    30:  76.91%100.00% 90.00% 96.55% 88.14% 69.70% 65.25% 78.41% 96.15%
##    40:  75.06%100.00% 90.00% 96.55% 88.14% 66.67% 65.25% 72.73% 96.15%
##    50:  75.98%100.00% 90.00% 93.10% 86.44% 70.71% 66.10% 75.00% 92.31%
## ntree      OOB      1      2      3      4      5      6      7      8
##    10:  75.37%100.00%100.00%100.00%100.00% 87.70% 38.36% 75.45%100.00%
##    20:  75.19%100.00%100.00%100.00%100.00% 86.29% 27.21% 90.91%100.00%
##    30:  74.31%100.00%100.00%100.00% 98.65% 88.71% 21.77% 91.82%100.00%
##    40:  75.79%100.00%100.00%100.00%100.00% 92.74% 25.17% 89.09%100.00%
##    50:  75.05%100.00%100.00%100.00%100.00% 92.74% 23.13% 88.18%100.00%
```

```
## Random Forest
##
## 541 samples
##  20 predictor
##   8 classes: '1.5', '2', '2.5', '3', '3.5', '4', '4.5', '5'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 434, 432, 434, 431, 433
## Resampling results across tuning parameters:
##
##   mtry  Accuracy   Kappa        Accuracy SD  Kappa SD
##    2    0.2662353  0.008980148  0.01622787   0.02384463
##   12    0.2256409  0.016980744  0.04285306   0.05150857
##   22    0.2456916  0.044872905  0.04059436   0.05205530
##
```

```
## Accuracy was used to select the optimal model using  the largest value.
## The final value used for the model was mtry = 2.
```

Discussion - Explain how you interpret the results of your analysis and what the implications are for your question/problem.