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

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

Input

Category, State

Intermediate Output to support analysis and define model

Relation Attributes-Stars

Count words: Per states: total reviews analyzed, total ocurrences, frecuence per review

Output

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

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

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