ADBI :: Project Report

Business Recommendations using prediction of review rating

Sai Sameer Tirumalasetti (stiruma)

Jayanth Reddy Vontela (jvontel)

Introduction

Food business is quite tricky and starting a restaurant is a challenging task. Restaurant business owners have to keep up with the customer demands. They should continuously update their restaurant experience according to the trends. Otherwise, they could fall behind.

In this complex ever changing environment, we are in a need of lot of restaurant data to make any logical reasoning for future predictions. So, we make use of available Yelp data.

Data Source for the project

For this project, we used a dataset of Yelp Dataset Challenge which is available online (https://www.yelp.com/dataset_challenge). This dataset has the data of 144,073 businesses and provide useful information such as business profile, review text, user profile, friends and votes. All the data are in json format.

For reviews, the data looks like

```
{
    "review_id":"encrypted review id",
    "user_id":"encrypted user id",
    "business_id":"encrypted business id",
    "stars":star rating, rounded to half-stars,
    "date":"date formatted like 2009-12-19",
    "text":"review text",
    "useful":number of useful votes received,
    "funny":number of funny votes received,
    "cool": number of cool review votes received,
    "type": "review"
}
```

For businesses, the data looks like

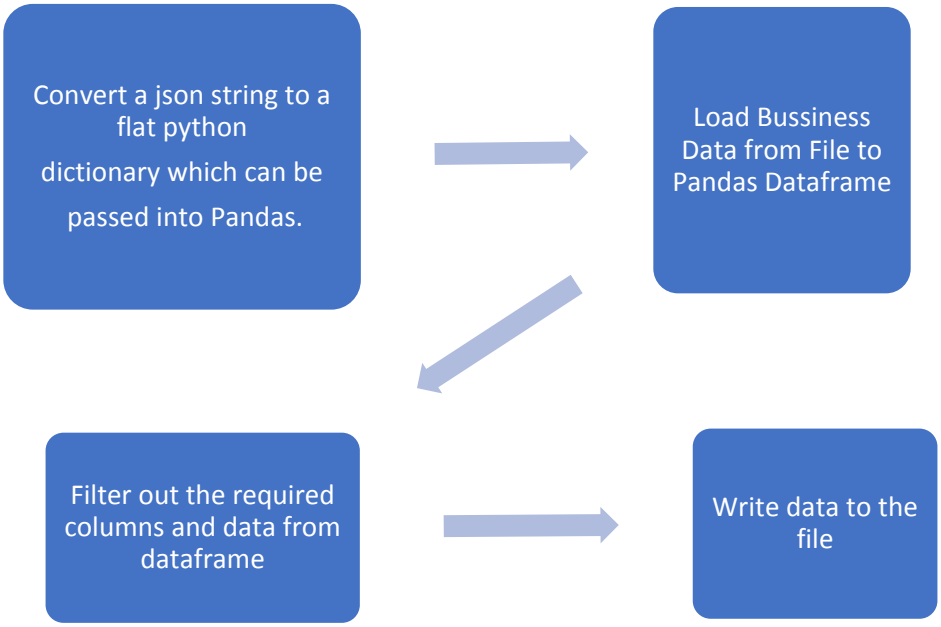
```
{
    "business_id":"encrypted business id",
    "name":"business name",
    "neighborhood":"hood name",
    "address":"full address",
    "city":"city",
    "state":"state -- if applicable --",
    "postal code":"postal code",
    "latitude":latitude,
    "longitude":longitude,
```

```
    "stars":star rating, rounded to half-stars,
    "review_count":number of reviews,
    "is_open":0/1 (closed/open),
    "attributes":["an array of strings: each array element is an attribute"],
    "categories":["an array of strings of business categories"],
    "hours":["an array of strings of business hours"],
    "type": "business"
}
```

Methodology

Data Pre-processing

At first, we trained our model based on review texts and star rankings from the same user. So, we take business dataset and filter data categorically and based on city. A file containing all the reviews and ratings given based on this condition was generated. Then, review texts were cleaned, by removing format, punctuation and extra whitespace. Processed the reviews by tokenizing, removing stop words, lemmatizing and POS tagging and stores it into MongoDB CORPUS collection.. Word stemming was achieved which erased word suffixes to retrieve the root or stem. Stopwords and words which occur in less than 5 documents or more than 50% of documents i.e., words with no information value but appear too common in a language were removed.



Models

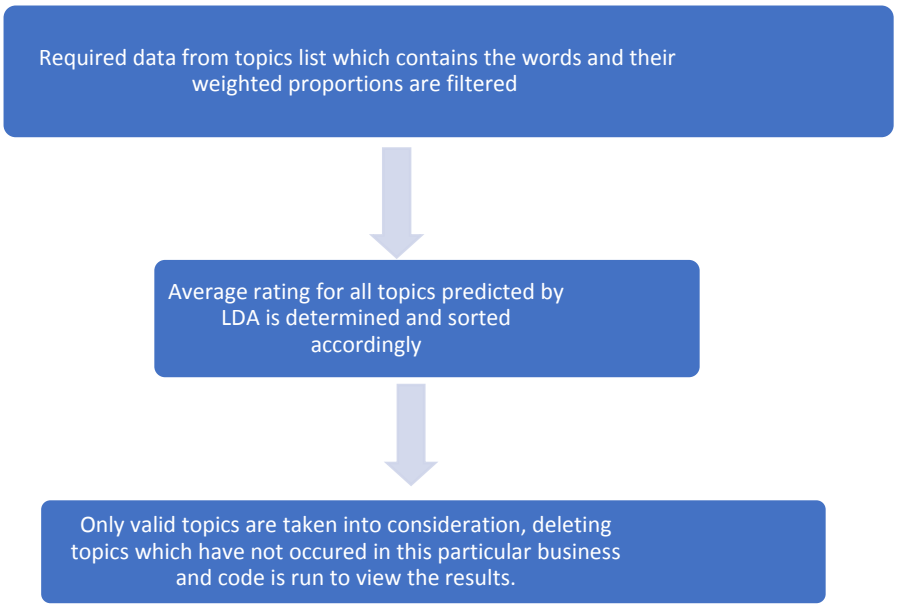
Now we used KLdivergence function to create KLDivergence graph to find out optimal number of topics for LDA. This is done by running the code for 30 times which took a time of 5.1 hours to run. We extracted only

key features of a review, called topics, and used them as training features. To extract topics, we used Latent Dirichlet Allocation(LDA). We use the Latent Dirichlet Allocation (LDA) factor model to approach the unsupervised learning of factors and topics for the Yelp restaurant review data. This model treats the probability distribution of each document over topics as a K-parameter hidden random variable rather than a large set of individual parameters.

Using the LDA function we determine the number of words present in a document and on this document we determine the mixture of topics in the respective document and using each document and topic we generate a multinomial distribution and then output word is generated to fill the topic's word slot.

Example of topics: Kind of Food, Facilities, Service
Example of word: Ingredients used like onion, broccoli, butter, salami.

Now we run display.py to view the result. It's function is as follows:



Tools

- We mainly used Python scripts.
- **Gensim** Python Library: This is a topic modeling tool for documents.
- **NLTK**: We used NLTK package for data cleaning and lexical feature extraction. The built in functions for removing stop words and retrieving unigrams, bigrams were helpful. The NLTK package worked well out of the box, but it was quite slow for POS tagging. We researched this topic for a fair amount of time, and came across the hunpos tagger. This combined with a model specifically meant for web data sped up our tagging process.
- **MongoDB** : Fast joins between tables, helped with metadata and user history features. We go into further detail under the Lessons learnt section how MongoDB was very useful. The highlight was how simple it was to use, and how it worked glitch free.
- **Scipy**: Standard implementations of ML models. Converge on the complete dataset.

Results and Discussion

The topics and gradient that are obtained as the output are as follows:

Topic	Gradient	Topic	Gradient	Topic	Gradient
Place	0.029	Chicken	0.026	Restaurant	0.021
Soup	0.019	Spicy	0.018	Rice	0.018
Roll	0.016	Sushi	0.013	Sauce	0.013
place	0.036	bar	0.030	beer	0.024
service	0.022	time	0.021	night	0.018
drink	0.013	restaurant	0.013	selection	0.012
menu	0.012	Place	0.037	Order	0.022
Burger	0.021	Sandwich	0.012	Don	0.012
Lunch	0.012	Fry	0.011	Service	0.011
People	0.010	Pizza	0.093	Sauce	0.023
Slice	0.015	Place	0.015	Salad	0.014
Tomato	0.014	Cheese	0.013	Crust	0.012
Delivery	0.012	Bread	0.012	Meal	0.017
Restaurant	0.016	Dinner	0.015	Flavour	0.014
Desert	0.014	Menu	0.013	Plate	0.011
Pork	0.011	Meat	0.011	Potato	0.010
Taco	0.031	Place	0.023	Breakfast	0.021
Coffee	0.020	Egg	0.019	Chip	0.019
Brunch	0.018	Time	0.010	Potato	0.009
Day	0.009				

This data can be summarized and given as:

Bakery	20.4%
Bar	18.4%
Taste	16.8%
Service	15.9%
Breakfast	16.4%
Food	12.1%

Questions that are to be answered:

- Review ratings for a restaurant with a business ID: 0TE3fZY8IzTeCaKIPbCsYg

Topic	Rating
Bakery	4.578
Bar	4.7
Taste	4.6

Service	4.16
Breakfast	4.34
Food	4.23

This shows that the particular restaurant needs to improve its service to get more positive feedback.

- Review ratings for a restaurant at a particular location: here given Bloomfield,Pittsburg

Topic	Rating
Bakery	4.6
Bar	4.35023
Taste	4.16
Service	4.14
Breakfast	4.06
Food	4.32094

This shows that for a new restaurant to gain good reviews should serve good breakfast.

References:

- *https://www.kaggle.com/c/yelp-recsys-2013*
- *https://www.yelp.com/html/pdf/YelpDatasetChallengeWinner_ImprovingRestaurants.pdf*
- *http://blog.cigrainger.com/2014/07/lda-number.html*
- *http://stackoverflow.com/questions/20886565/pythonusingmultiprocessingprocesswithamaximumnumberofsimultaneouspro*
- *Shuyan Wang. 2015. Predicting Yelp Review Upvotes by Mining Underlying Topics*