CSC522 Automated Learning and Data Analysis, FALL 2015

Providing Business Insights to Restaurants from Yelp User Reviews

Jigar Sharda, Ronak Patel and Shrey Sanghavi

North Carolina State University, Raleigh, North Carolina, USA.

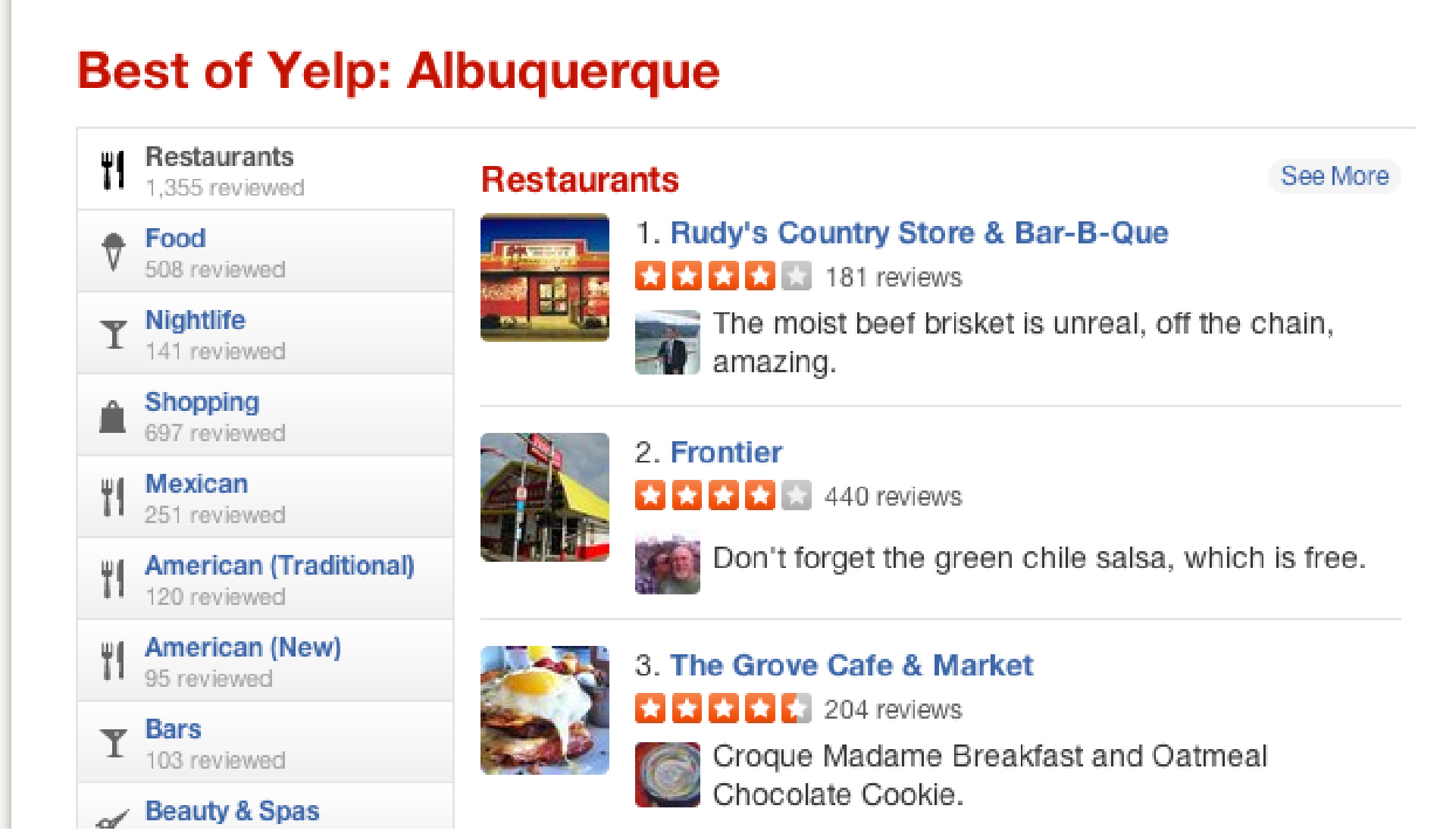
Scan to view our code on GitHub

Abstract

Yelp reviews provide useful information and several opinions to users and help them choose the best businesses in their region. Currently, if a prospective entrepreneur wants to understand the market scenario, the user needs and expectations, he has to read all the reviews in the region to get an idea about the demand. This strategy is impractical, nor feasible or accurate.

Introduction

The reviews provided in the dataset do not explicitly list the features provided in the available ratings. We try to solve this problem using Topic Modeling in order to get hold of the inherent features. Once we have these features, we intend to find the relative importance of each of the features per restaurant.



Best of Yelp: Albuquerque

Restaurants 1,355 reviewed
Food 508 reviewed
Nightlife 141 reviewed
Shopping 697 reviewed
Mexican 251 reviewed
American (Traditional) 120 reviewed
American (New) 95 reviewed
Bars 103 reviewed
Beauty & Spas

Restaurants See More

1. Rudy's Country Store & Bar-B-Que
181 reviews
The moist beef brisket is unreal, off the chain, amazing.

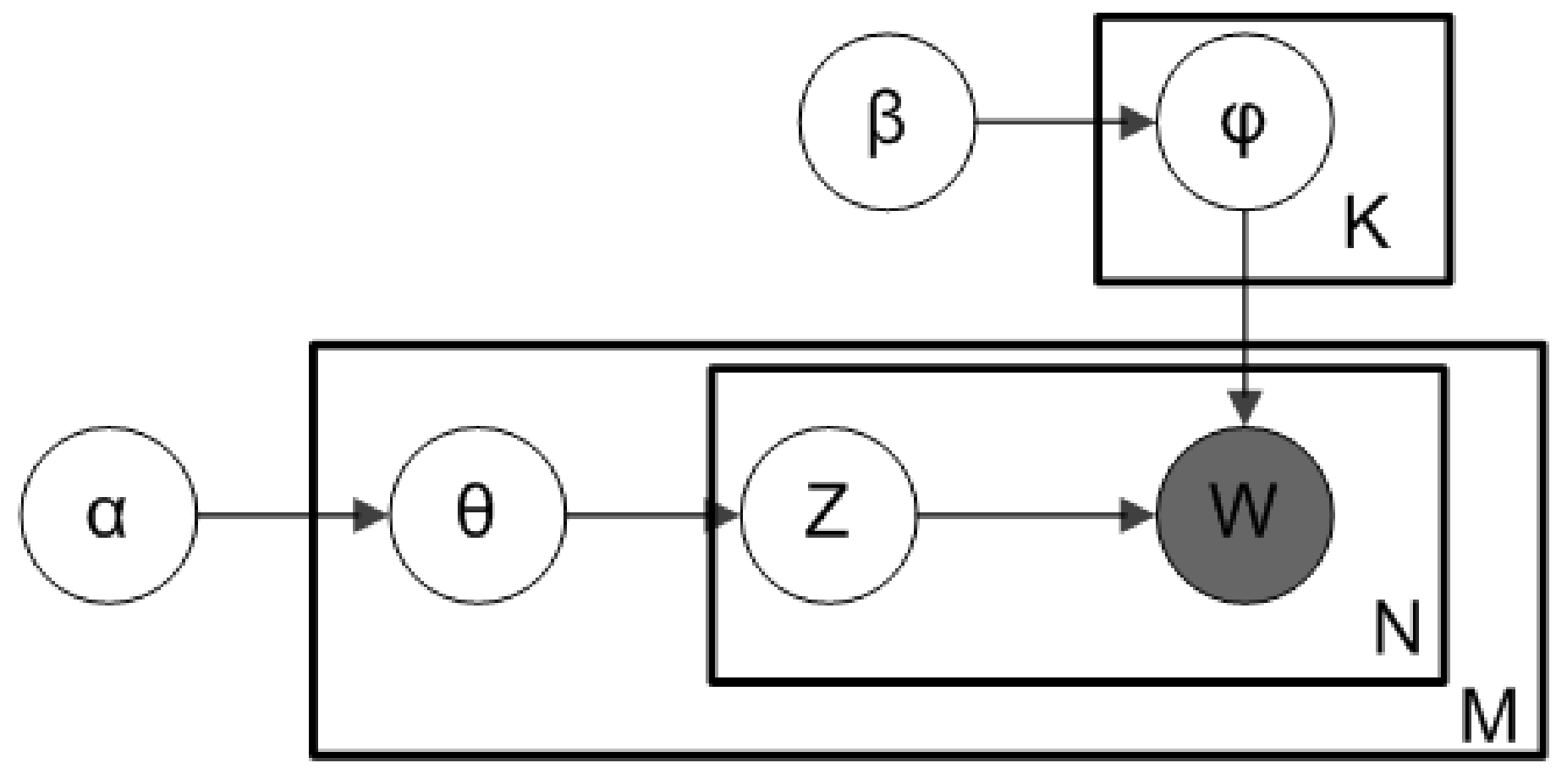
2. Frontier
440 reviews
Don't forget the green chile salsa, which is free.

3. The Grove Cafe & Market
204 reviews
Croque Madame Breakfast and Oatmeal Chocolate Cookie.

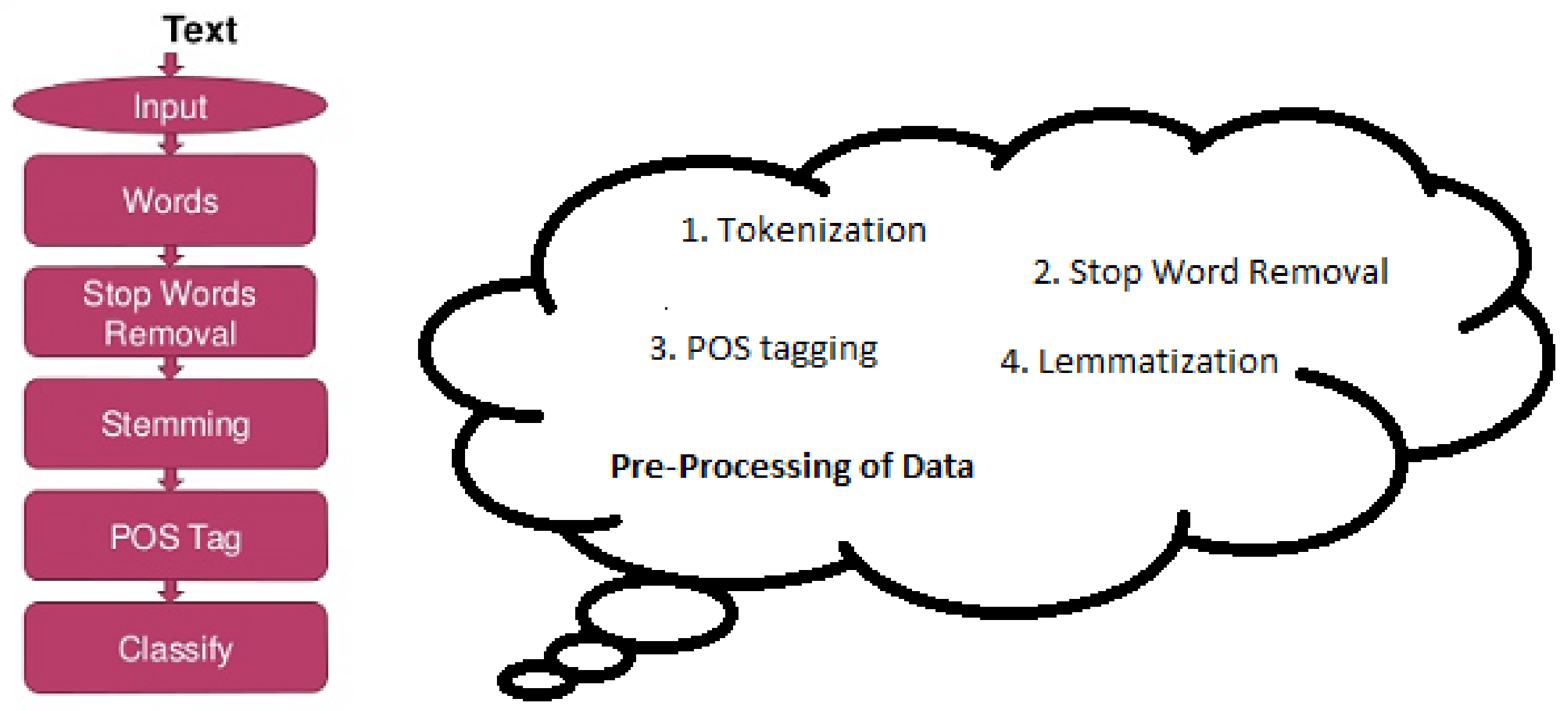
Method

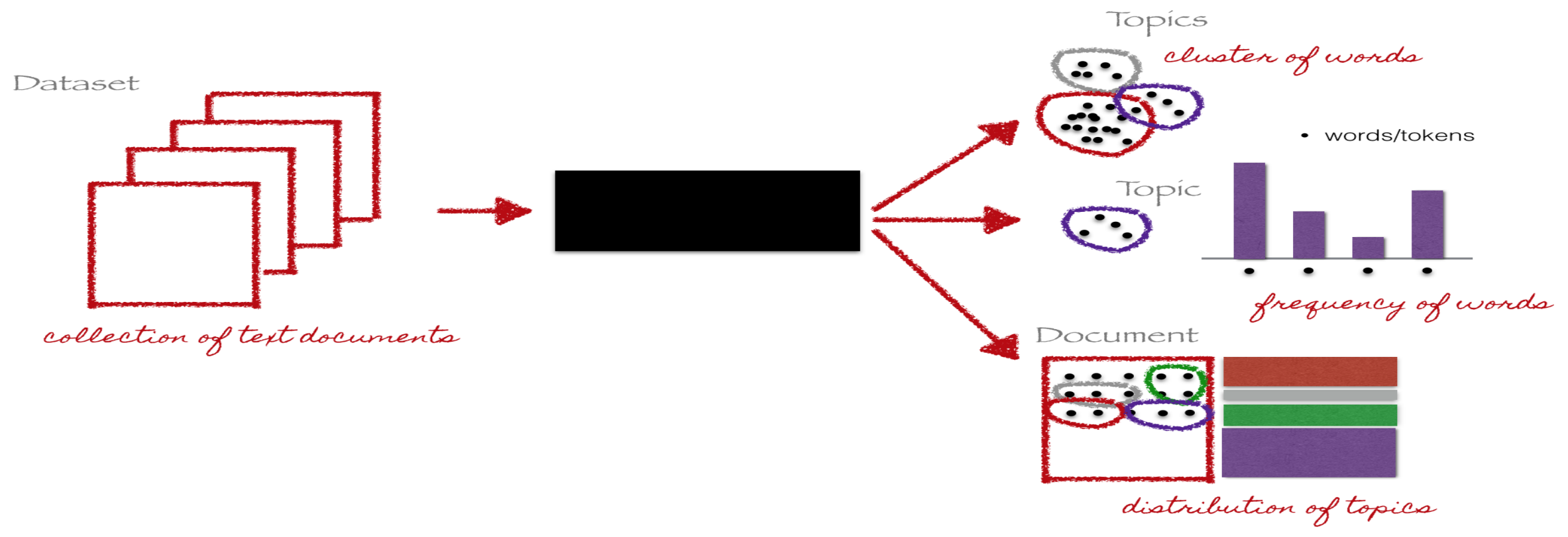
We loaded necessary data from json files into MongoDB and used Python as our language of implementation.

Algorithm: Latent Dirichlet Allocation



Approach

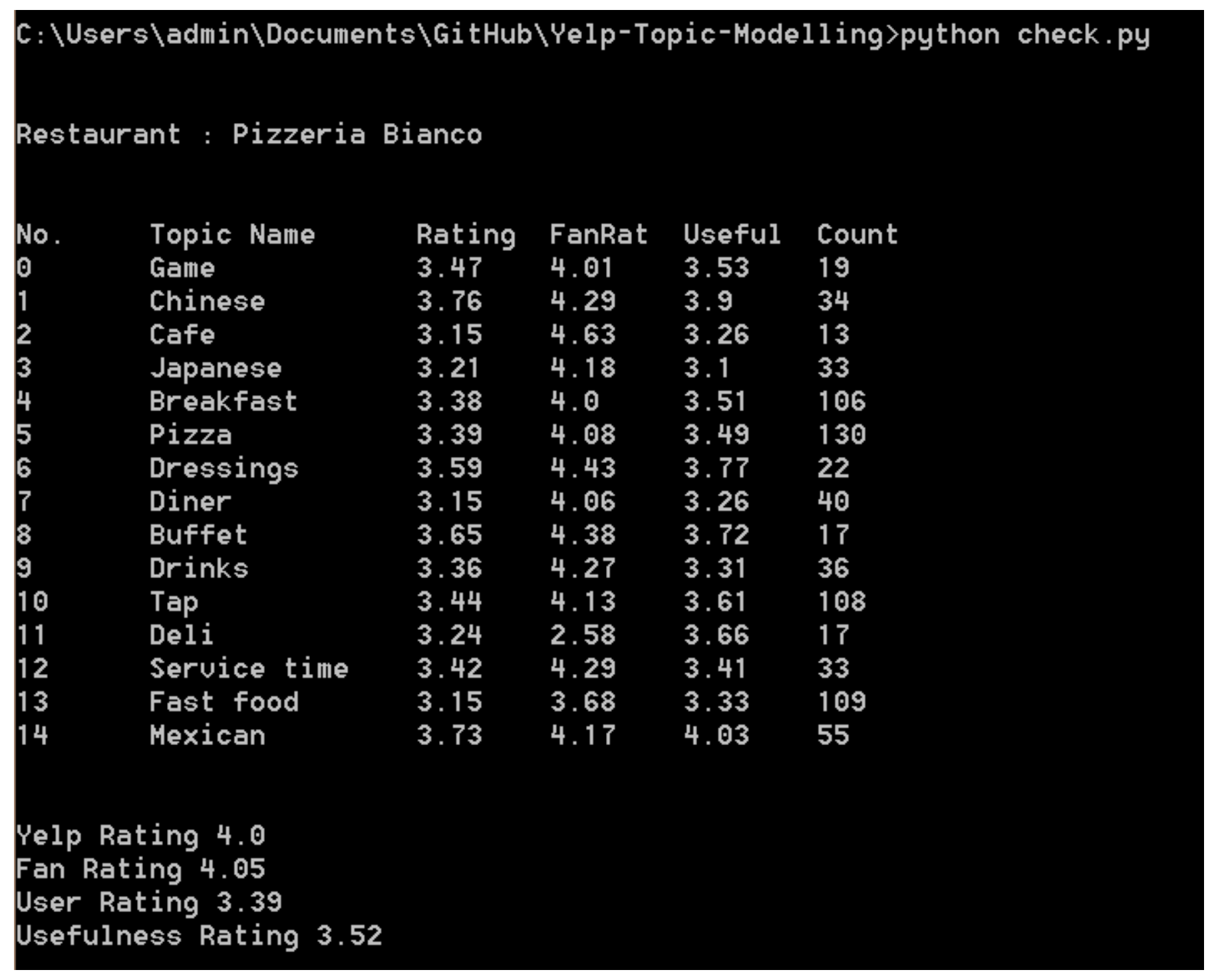




Fast
Deli
Diner
Chinese
Service
Tap
Dressings
Japanese
Buffet
Breakfast
Game
time
Pizza
Drinks
Mexican
Café
food

Results

We identified various areas for restaurants to focus on, based on user reviews. We found out a relationship between original Yelp ratings and ratings based upon User's fans. We found out a relationship between the usefulness of a User's review and the aggregate of all users ratings. In order to decide the number of topics for LDA we used log perplexity.

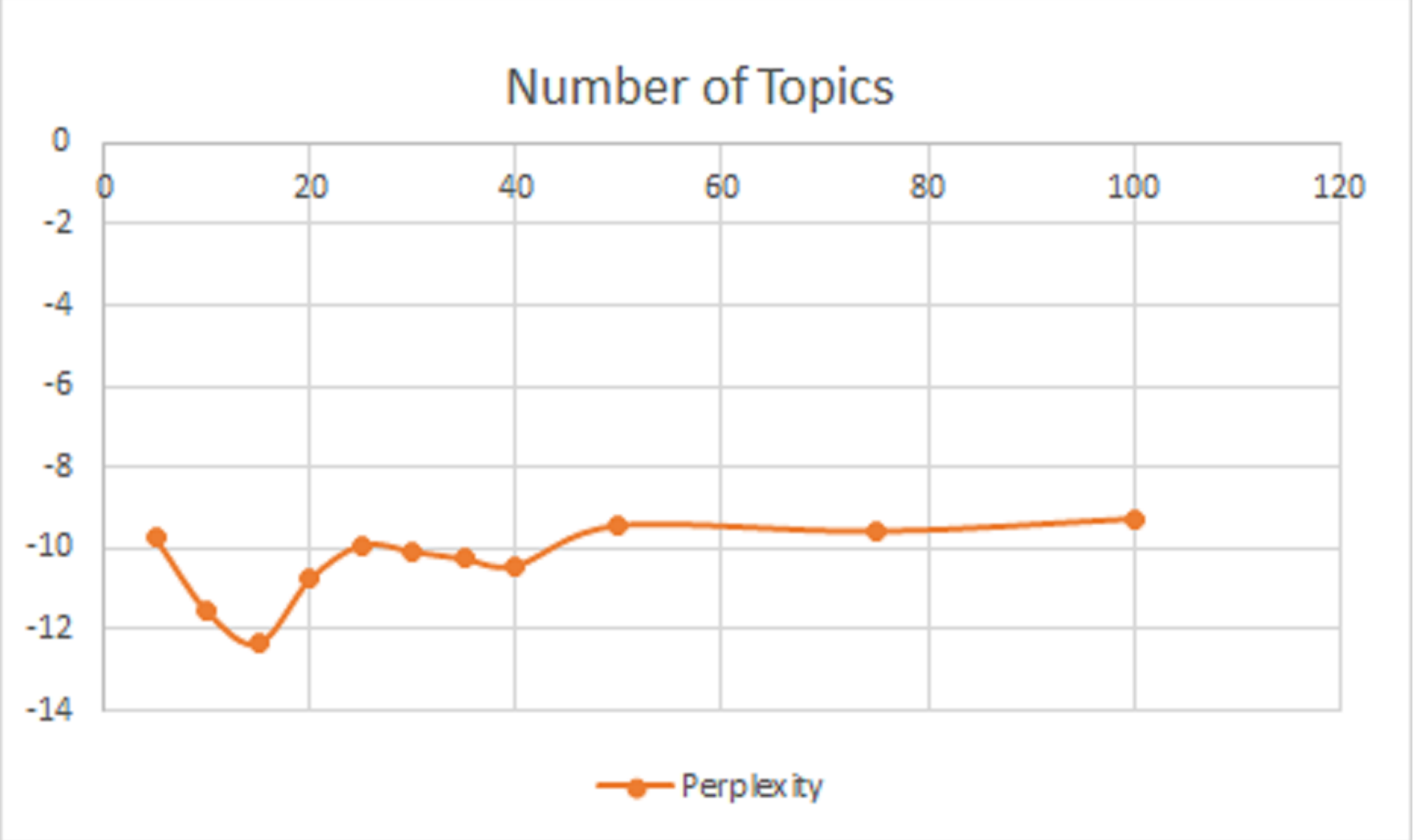


```
C:\Users\admin\Documents\GitHub\Yelp-Topic-Modelling>python check.py


Restaurant : Pizzeria Bianco


No.     Topic Name      Rating  FanRat  Useful  Count
0       Game            3.47    4.01    3.53    19
1       Chinese         3.76    4.29    3.9     34
2       Cafe            3.15    4.63    3.26    13
3       Japanese        3.21    4.18    3.1     33
4       Breakfast       3.38    4.0     3.51    106
5       Pizza           3.39    4.08    3.49    130
6       Dressings       3.59    4.43    3.77    22
7       Diner           3.15    4.06    3.26    40
8       Buffet          3.65    4.38    3.72    17
9       Drinks          3.36    4.27    3.31    36
10      Tap             3.44    4.13    3.61    108
11      Deli            3.24    2.58    3.66    17
12      Service time    3.42    4.29    3.41    33
13      Fast food       3.15    3.68    3.33    109
14      Mexican         3.73    4.17    4.03    55


Yelp Rating 4.0
Fan Rating 4.05
User Rating 3.39
Usefulness Rating 3.52
```



Future Challenges

Yelp dataset corpus is pretty huge and the possibilities for exploration are endless. Following are some tasks which can be taken as a continuation of our work :

1.Sentence level text analysis

2.Multi-aspect topic analysis of User reviews

3.Finding clusters of users to predict the trend setters

Citations

1. James Huang, Stephanie Rogers, Eunkwang Joo.2013. Improving Restaurants by Extracting Subtopics from Yelp Reviews. In Yelp Dataset Challenge Winner
2. Gensim Library. Available at: https://radimrehurek.com/gensim/
3. Ivan Titov and Ryan McDonald. 2008. Modeling online reviews with multigrain topic models. In Proceedings of the 17th international conference on World Wide Web (WWW '08).