Extracting hidden topic from yelp restaurant reviews using LDA

Lu Yang

ABSTRACT

In this paper, I discover hidden topics from yelp restaurant reviews by using Latent Dirichlet Allocation algorithm. Hidden topics are the core ideas of a large amount of reviews. I extract them from yelp dataset in order to help restaurants know what the customers really care about and improve restaurants effectively. The open dataset I used is from the Yelp Dataset Challenge and has 10765368 customer reviews. I adopted the LDA algorithm to consolidate reviews into many unobserved topics and use POS tag filter and frequency filter to improve the model. Besides, I use the LDA model to extract topics for a new review. Since an unsupervised method cannot apply precision and recall, I use data visualization to draw the bar charts. Comparing these three models, I think that the LDA model within POS tag filter works best. It can extract the topics whose words fall into different categories. However, from these results, I found that the LDA model can only complete the task of data exploration since it only extracts words which describe the topics. We still need to do more to get specific topics for these reviews.

INTRODUCTION

1. Dataset

The dataset I used is Yelp dataset. It is a subset of Yelp's businesses, reviews and user data. This dataset contains 7 csv files that converted from JSON files. For this project, I use yelp_business.csv and yelp_review.csv. In total, there are 5200000 user reviews, information on 174000 businesses.

2. Why Reviews important

Yelp reviews are the feedbacks from customers who have been to the restaurant. They pointed out the advantages, disadvantages and specialties of the restaurants. Besides, these reviews also have a huge impact on whether other customers will come to this restaurant. Overall, these reviews are very valuable and contain what customer really need.

3. Why hidden topic

Reviews are high-dimensional data which is difficult to extract features. For human, reading a lot of reviews is a time-consuming work. Due to the above reasons, we need to reduce the dimensions of the reviews, which means, to extract hidden topic from the text. Therefore, we can extract key information from a large section of reviews accurately, thus we can help to reduce the workload of restaurant owners and improve restaurants.

4. Why using LDA

There are plenty of method to factor models for discrete data, such as latent semantic indexing (LSI), probabilistic latent semantic indexing (PLSI), laplacian probabilistic latent semantic indexing (LapPLSI) and latent dirichlet allocation(LDA).
LSI: Analyzing relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms. It assumes that words that are close in meaning will occur in similar pieces of text (the distributional hypothesis).
PLSI: The analysis of two-mode and co-occurrence data. In effect, one can derive a low-dimensional representation of the observed variables in terms of their affinity to certain hidden variables.
LapPLSI: An algorithm which models the document space in a more discriminatory manner using nearest neighbors.

LDA: It is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar.
The reason that I choose LDA : 1) We need unsupervised learning to extract topics from yelp reviews. 2) We can customize the number of topics rather than a large set of individual parameters.
LDA assumes documents are produced from a mixture of topics. Those topics then generate words based on their probability distribution. It is a matrix factorization technique and make uses of sampling techniques and iterative method in order to improve the matrices.

5. Implementation

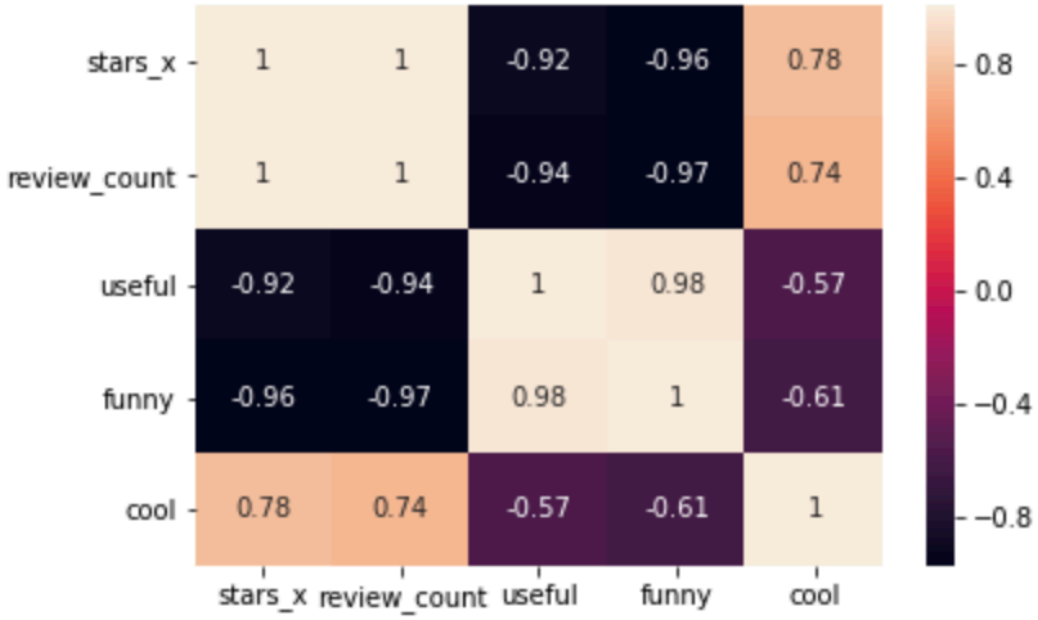
1) Loading the data from yelp database
    a. Loading yelp_business and yelp_reviews
    b. Merging two dataset and select useful columns
    c. Spliting the dataset into 3 part
2) Exploring the data
    a. Explore the relationship among 'stars', 'useful', 'funny', 'cool' columns
    b. Explore different stars distribution among different categories
3) Clean the data
    a. Sample the data
    b. Remove the punctuations, stopwords and normalize the corpus.
4) Build the orginal LDA model and data visualization using df1 dataset
5) Using the trained LDA model to get hidden topic from a new review
6) Using Frequency Filter to improve the model and data visualization
7) Using Part of Speech Tag Filter to improve the model and data visualization

CODE WITH DOCUMENTATION

Jupyter notebook on Github:
https://github.com/lucylucyyangyang/Final-Project-CODE/blob/master/ResearchPaper.ipynb

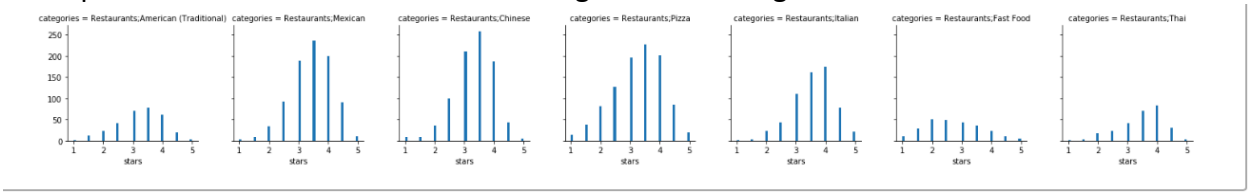
RESULT

1. Exploring the data
    a. Explore the relationship among 'stars', 'useful', 'funny', 'cool' columns



As we can see, 'stars' is strongly correlated with 'cool' and has a negative correlation with 'useful' and 'funny'. There is a strong correlation between 'useful' and 'funny'.

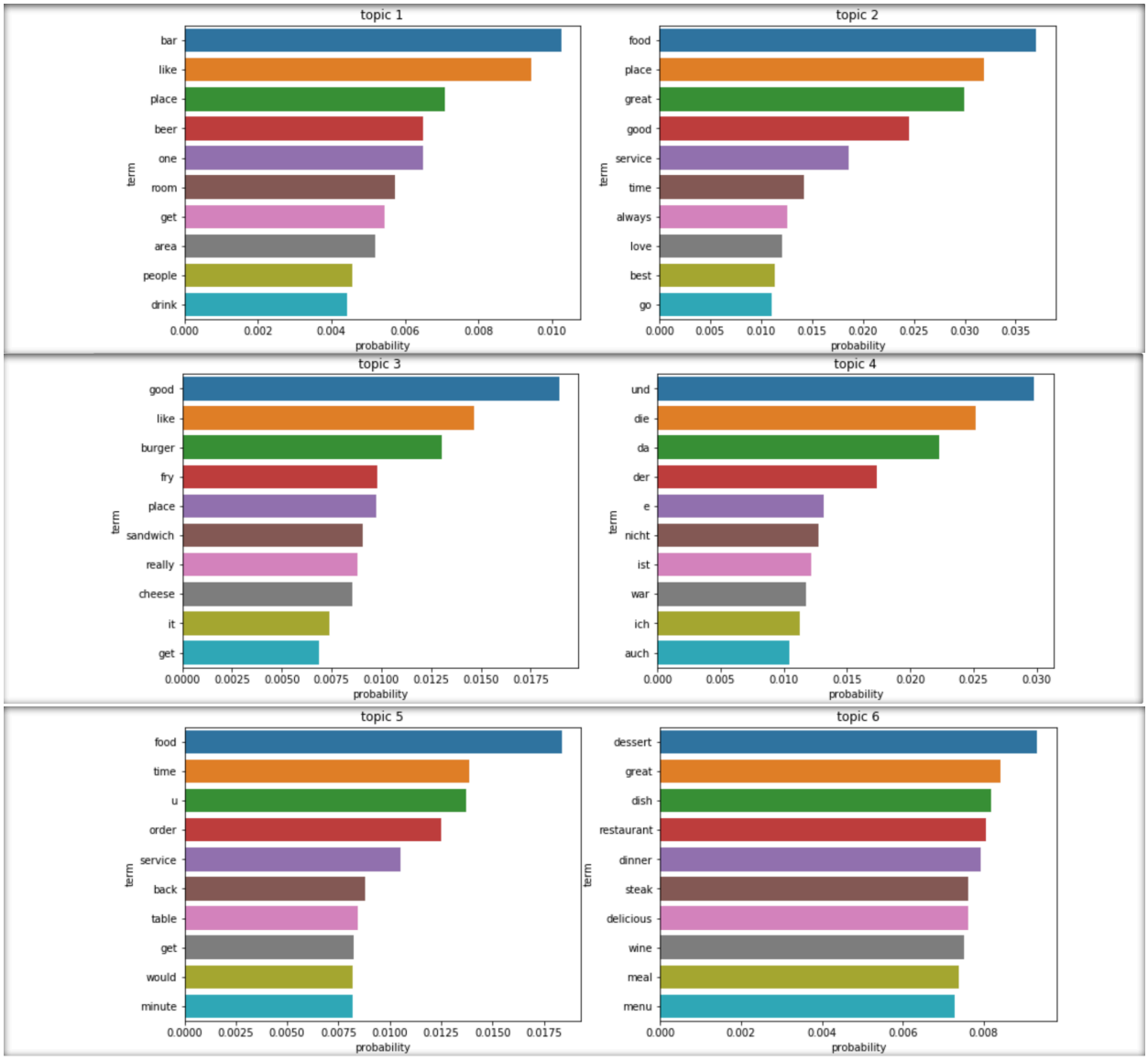
    b. Explore different stars distribution among different categories

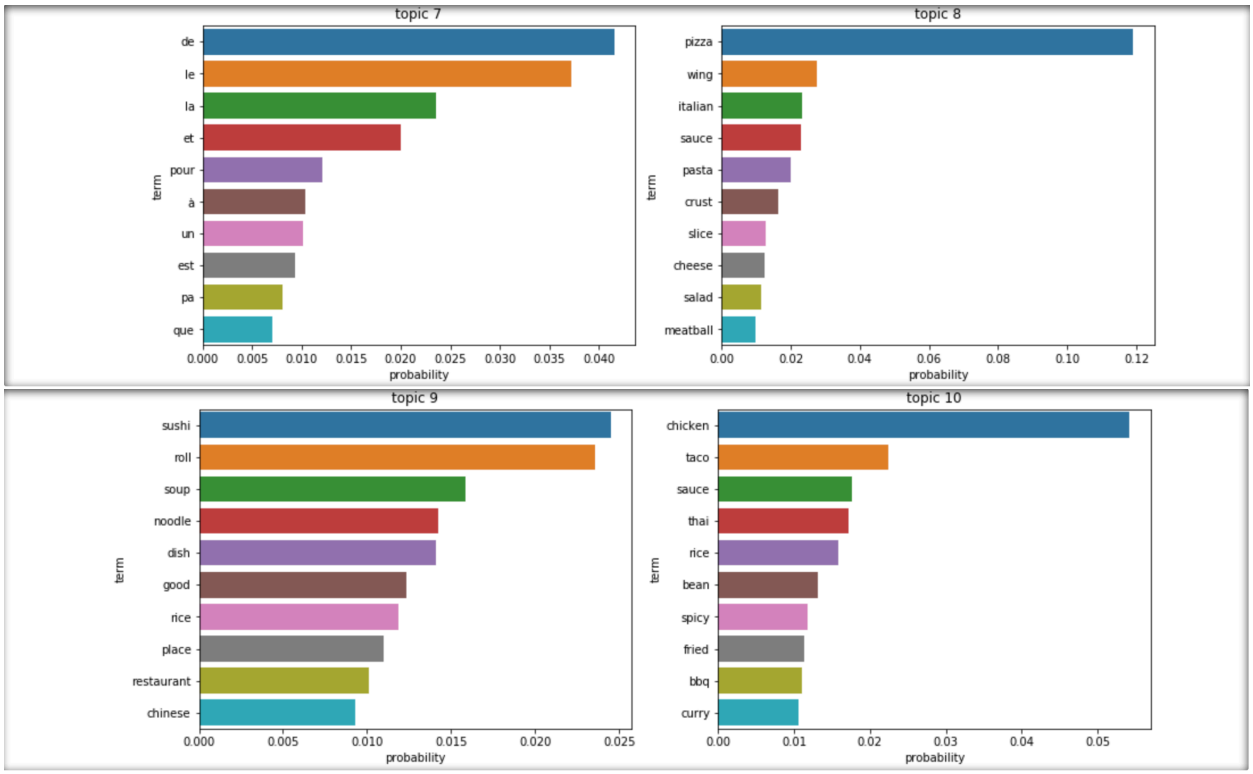


As we can see, fast food is a right-skewed distribution which means the majority of fast food restaurants don't have high ratings. Others are left skewed distributions which means most of them are highly rated.

2. LDA result
    a. Original LDA model

For this model, I extracted Top 10 topics and describe each topic with the five most relevant words


topic 1
bar
like
place
beer
one
room
get
area
people
drink
term
0.000
0.002
0.004
0.006
0.008
0.010
probability
topic 2
food
place
great
good
service
time
always
love
best
go
term
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
probability
topic 3
good
like
burger
fry
place
sandwich
really
cheese
it
get
term
0.0000
0.0025
0.0050
0.0075
0.0100
0.0125
0.0150
0.0175
probability
topic 4
und
die
da
der
e
nicht
ist
war
ich
auch
term
0.000
0.005
0.010
0.015
0.020
0.025
0.030
probability
topic 5
food
time
u
order
service
back
table
get
would
minute
term
0.0000
0.0025
0.0050
0.0075
0.0100
0.0125
0.0150
0.0175
probability
topic 6
dessert
great
dish
restaurant
dinner
steak
delicious
wine
meal
menu
term
0.000
0.002
0.004
0.006
0.008
probability



As we can see:
Topic 1: It is about the bars. Customers care about their beer, place and rooms.
Topic 2: It discussed the place for dating and partying. Customer pay more attention on service and place.
Topic 3: It should be about fast food. Burger, sandwich and cheese are the focus points.
Topic 4: They are some German words.
Topic 5: It should be about restaurant wait time.
Topic 6: It is about French restaurant. Customer care about their wine, dessert and steak.
Topic 7: They are some French word.
Topic 8: It is about Italian restaurant. People care about pasta, meatball and pizza.
Topic 9: It may be about the ramen restaurant. Customers may care about their noodles, sushi and soup.
Topic 10: It may be about some Mexican Restaurant. Foods are the key point.

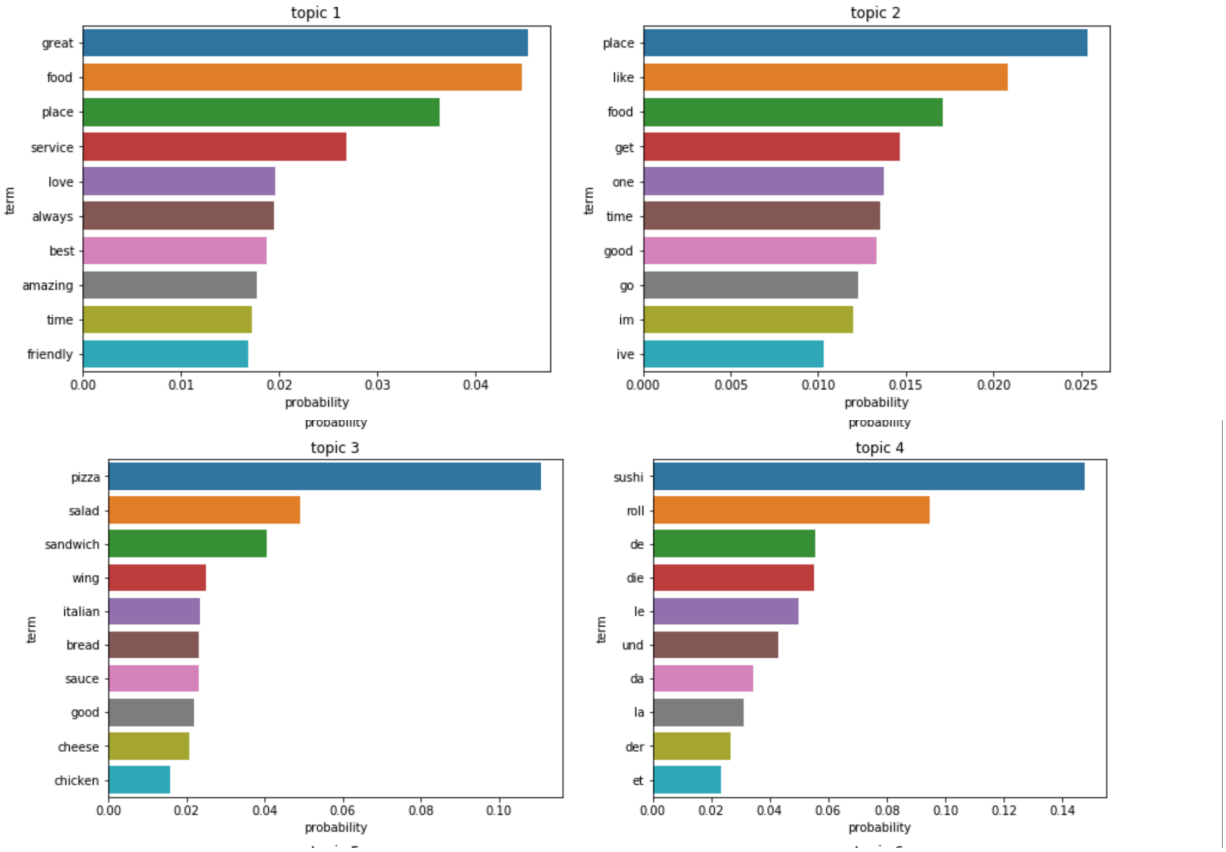
We can see this LDA model extract several words to describe these 10 topics. However, we can't accurately know what the topics are through this model. In addition, this model extracts some French and German words.

b. Using original LDA model to extract topic for a new review

```
[(3, '0.148*"sushi"'),
 (9, '0.040*"chicken"'),
 (5, '0.014*"steak"'),
 (4, '0.073*"taco"'),
 (7, '0.064*"burger"'),
 (6, '0.032*"good"'),
 (8, '0.023*"u"'),
 (2, '0.111*"pizza"'),
 (0, '0.045*"great"'),
 (1, '0.025*"place"')]
```

As we can see, this review is discussing the food in the restaurant. Overall, the customer thinks this restaurant is good.

c. Using frequency filter to improve the LDA model

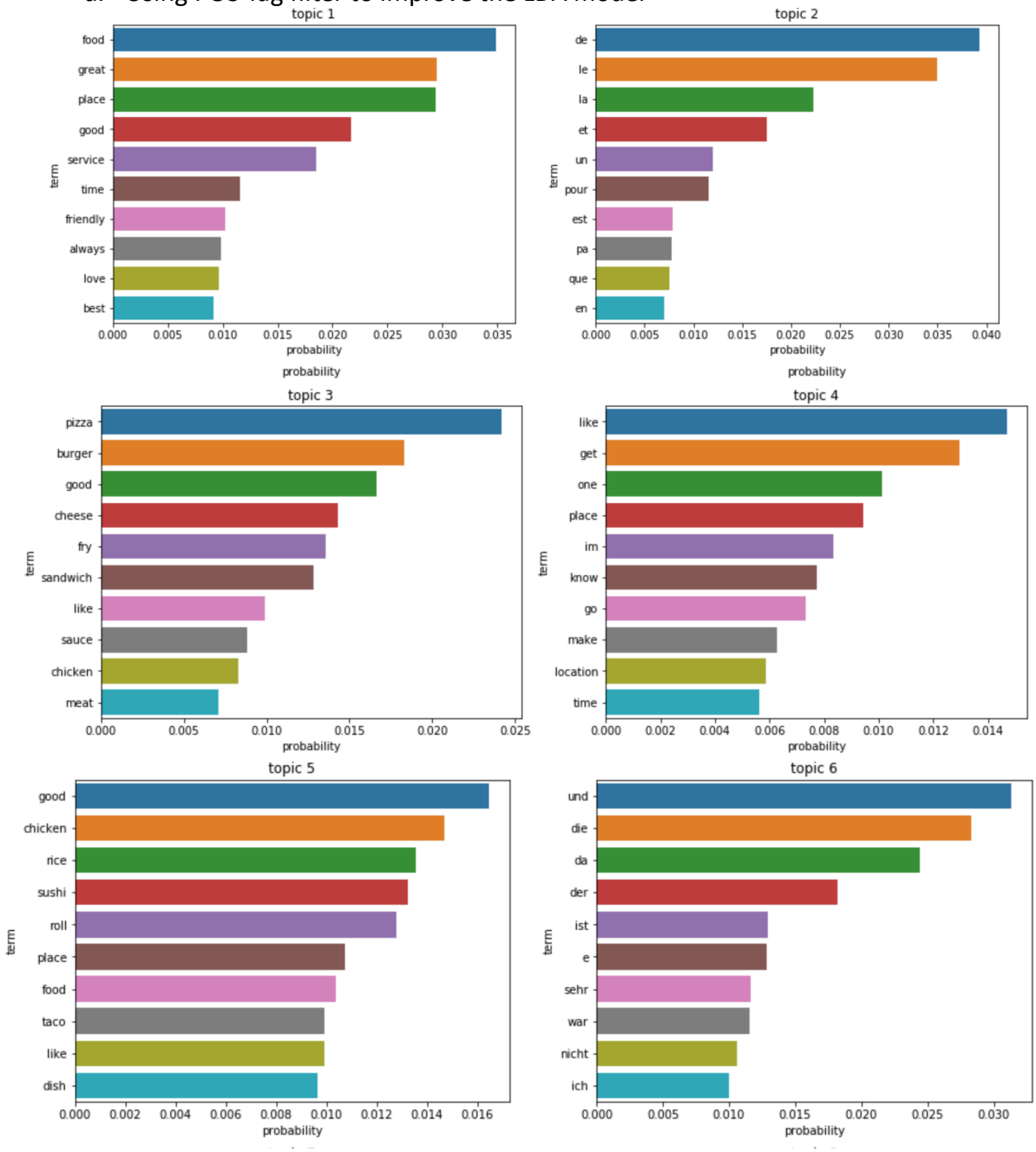


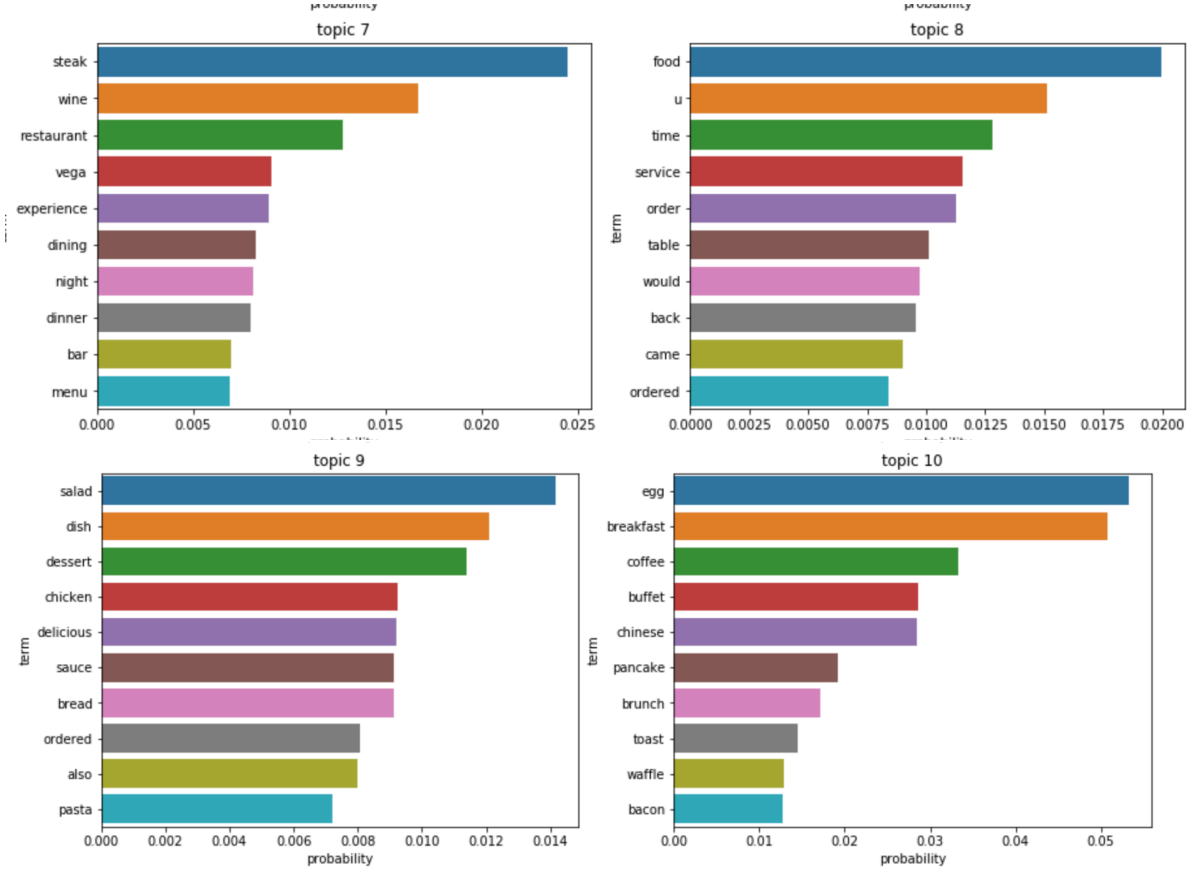
Topic 1: It is a place for dating and partying. Customer pay more attention on service and place.
Topic 2: It is about service and time spent in the restaurant.
Topic 3: It may be about the Italian restaurant. Customers may care about their traditional food.
Topic 4: They are some French words. They are saying something about Asian foods.
Topic 5: It may be about a Mexican restaurant. Customers pay more attention on taco, chip and burrito.
Topic 6: It is about the restaurant which is suitable for dinner. Customer care about their steak, dessert, salad and menu.
Topic 7: It should be about the bars. The drink such as the beer, the service and the place are the focus points.

Topic 8: It should be about the burgers. The food is the focus point.
Topic 9: It is about restaurant wait time.
Topic 10: It may be about Japan restaurants. Customer care about their food most.

I think this LDA model works better than the first model since it uses frequency filters to get the most commonly used words. These words describe the topics of different categories of food. We can also find an interesting phenomenon. Many French people may love to try Asian foods.

d. Using POS Tag filter to improve the LDA model





Topic 1: It is a place for dating and partying. Customer pay more attention on service and place.
Topic 2: They are some French words.
Topic 3: It may be about the fast food restaurant. Customers may care about their burger and sandwich.
Topic 4: It shows that customer also care about the location of the restaurant.
Topic 5: It may be about some Mexican Restaurant. Foods are the key point.
Topic 6: They are some German words
Topic 7: It is about the restaurant which is suitable for dinner or the bar. Customer care about their steak, wine and menu.
Topic 8: It is about restaurant wait time.
Topic 9: It is about ordered food.
Topic 10: It is about breakfast.

I think this LDA model works best among these 3 models. It uses POS tag filters to get the most useful words. These words describe the characteristics of these restaurants themselves such as location, wait time and so on.

DISCUSSION

1. The LDA model

The LDA model can extract several words to describe these topics. However, we can't accurately see what these topics are through this model. We still need some classification technology to help us find the exact topics. We can draw a conclusion that the LDA model is more suitable for data exploration. It really helpful for extracting words to describe the topics, but it cannot help us to get the topics themselves. Besides, we can find that there are many people in Germany and France who love to use yelp too.

2. Improvement of LDA

The results of topic models are completely affected by the features present in the corpus. During the process of LDA, the corpus converted into document term matrix. If we can reduce the dimensionality of the matrix, the results of topic modeling will be better. The common methods to reduce the dimensionality are POS tag filter and frequency filter.

a. Using frequency filter to improve LDA model

Frequency filter is arranging every term according to its frequency. Terms with higher frequencies are more likely to appear in the results as compared ones with low frequency. I think it is really helpful for me to improve the model. The most of 10 extracted topics are clear and useful. The similarity of these topics is not high.

b. Using POS tag filter to improve LDA model

In corpus linguistics, part-of-speech tagging is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech based on both its definition and its context. The POS tag filter helps us to select most useful words from a large amount of corpus. These topics are all unique and valuable because they describe the different characteristics of the restaurants.

3. What customer really need

From these 3 models, we can know that customers pay most attention to the service and place. They discussed a lot about whether this place is good for date or party. Besides, they care most about the food. In addition, location is also a focus point.

REFERENCE

1. Stack overflow (2017). Understanding LDA / topic modelling — too much topic overlap from https://stackoverflow.com/questions/46326173/understanding-lda-topic-modelling-too-much-topic-overlap
2. Stack overflow (2016). Reading file using relative path in python project from https://stackoverflow.com/questions/40416072/reading-file-using-relative-path-in-python-project/40416154
3. Stack overflow (2013). How to predict the topic of a new query using a trained LDA model using gensim? from https://stackoverflow.com/questions/16262016/how-to-predict-the-topic-of-a-new-query-using-a-trained-lda-model-using-gensim
4. Brandon Rose (2018). Document Clustering with Python from http://brandonrose.org/clustering
5. Shivam Bansal (2016). Beginners Guide to Topic Modeling in Python

from https://www.analyticsvidhya.com/blog/2016/08/beginners-guide-to-topic-modeling-in-python/
6. James Huang (2013). Improving Restaurants by Extracting Subtopics from Yelp Reviews From https://www.yelp.com/html/pdf/YelpDatasetChallengeWinner_ImprovingRestaurants.pdf
7. StackExchange (2015). Calculating precision and recall for LDA from https://stats.stackexchange.com/questions/185983/calculating-precision-and-recall-for-lda