CS598 Data Mining Capstone Task 1

Exploration of the Dataset

dipakp2@illinois.edu

Overview:

The goal of this task is to explore the Yelp data set to get a sense about what the data looks like and their characteristics. In this document, I am going to analyze part of Yelp's academic dataset and mine this data to discover interesting and useful knowledge.

Task 1.1

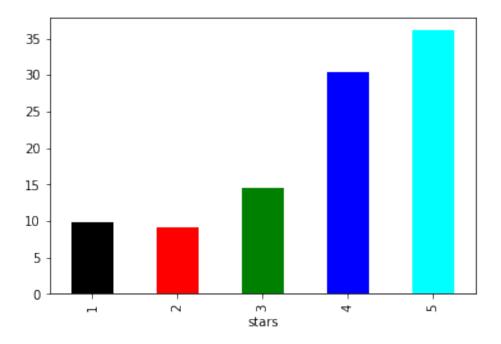
LDA (Latent Dirichlet Allication) – The main idea of this model is to assume that each document is a (different) mixture of topics. The topic is represented as a multinomial probability distribution over words. For yelp dataset, we will use LDA to extract topics from all the review text (positive and negative reviews) and visualize to better understand what everyone have talked about in these reviews.

Step 1. Preprocessing

In order to perform exploratory analysis we first read following Yelp's dataset –

- yelp academic dataset business.json
- yelp academic dataset checkin.json
- yelp academic dataset review.json
- yelp academic dataset tip.json
- yelp academic dataset user.json

Step 2. Distribution Plot



Here we are creating distribution plot of ratings to answer one of the high level goals of our exploratory data analysis. We can see that vast majority of the reviews are positive (rating 5 light blue bar) based on Yelp's rating system on a scale between 1-5.

Step 3. Topic Mining of restaurant reviews

In order to perform topic mining let us take a large random sample of overall dataset and then we perform following preprocessing steps:

- stop word removal
- remove new lines
- remove single quotes
- forming unigram and bigram
- applying spacy's lemmatization
- display wordcloud

At the end, we create a dictionary and corpus for our topic modeling.

Step 4. Topic Model LDA

For this topic model, we are using Gensim's fast and efficient LDA implementation to train a model on the corpus we just created from Yelp dataset and extract topics. For easiness, we decided to use 10 documents per training chunk, and limit the maximum

number of iterations to 100 when inferring the topic distribution. Further, in order to visualize the weight of words we have decided to use word cloud that shows us the exact importance of each topic.



```
open big cheap special side bad little less
```

```
give food always good try great well really service
```

```
order leave

order love

time
restaurant
experience thing
eat
table back
```

```
florist
prefect
prefect
harry five_guys
cashew_chicken
papa
bernaise shibuya
```

```
bernaiseprefect
five_guys
differently
caring
florist harry
shibuyapapa
cashew_chicken
```

shibuya papa five_guys cashew_chicken harry differently florist caring prefect

coffee enjoy
excellent
coolwonderful
perfect buy
atmosphere
customerselection

find people

price place
work drink
never look
much nice

```
papashibuya
prefect
five_guys
cashew_chicken
caring_florist
```

Step 5. LDA Positive Subset

In order to alleviate positive reviews from other reviews we will filter out results that has star ratings of 5 and resulting word cloud is shown below –

```
shibuyadifferently
caring florist
harry papa
prefect
bernaise
five_guys
cashew_chicken
```

```
side bad
open less game
hotel
little
special big
cheap
```

```
always
really
give

place
try
food
great goodservice
```

```
order love
restaurant back
leave eat

time
table experience
```

```
bernaiseshibuya
five_guys
prefectflorist
harry caring
papadifferently
cashew_chicken
```

```
five_guys
cashew_chicken
harry papaflorist
prefect
shibuyabernaise
```

```
caring
shibuya

prefect papa
cashew_chicken
five_guysharry
florist bernaise
```

```
atmosphere
customer
buy cool
excellent
selection, wonderful
enjoy
coffee perfect
```

```
find much
price work
drink
look place
never people
```

```
floristcaring cashew_chicken five_guys harry bernaise prefect papa differently shibuya
```

Step 6. LDA Negative Subset

In order to alleviate negative reviews from other reviews we will filter out results that has star ratings less than 3 and resulting word cloud is shown below –



side

```
Topic 1
               good
    dish
     little taste
least
     restaurant
          <sup>look</sup>nice
meal
        Topic 3
          impersonal
  facialbrutal
     extraction
real_mexicanpint
ırısh
           testicle
         Topic 5
       server food
try
   good
            order
        table
              eat
  service
            place
  drink
         Topic 7
    high
             play
       weekvegetable
  meat
            hear
 big
       usually
          today
   wife
         Topic 9
           mexican
       testicle
  extraction
facial
```

Discussion

Overall after looking at the results that are displayed in word cloud visualization we can say that LDA topic have provided us very good results. Taking glance at positive topics we can say that we were able to compile the results based on positive feelings with keywords such as "love", "wonderful", "caring", "great". Whereas negative topics reveals mix of negative and mix expression with keywords such as "wait", "old", "brutal", "impersonal" etc... These prominent words within the topics make intuitive sense, but some of the topics themselves are reasonable. For instance, positive reviews could include: Five Guys Burgers & Fries are so delicious. The negative reviews could be something like: We had to wait in line for long time had to leave due to slow service of restaurant.

Moreover we can also say that LDA topic has done excellent job in creating word topic cluster for positive and negative reviews. If we look at word cloud topic we can say that topics 0, 1, 3, 4, 5, 7, 8, 9 are based of positive reviews and topics 2 and 6 are based on negative reviews.

References

- https://radimrehurek.com/gensim/similarities/docsim.html
- https://www.datacamp.com/community/tutorials/wordcloud-python
- https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/#4whatdoesldado