Explore the Information Delivered by Food Reviews

Qiaowei Jiang
MS in Analytics / 1851 N SCOTT ST
MS in Analytics / Arlington, VA, 22209
qj16@georgetown.edu

Abstract

In this paper I explore the information delivered by Amazon Fine Food Reviews (2013). The main purpose of this paper is to identify the main food topics and emotional tendency presented in each single topic and the reviews as a whole. I propose to apply LDA model and sentiment analysis using ***gensim*** and ***Textblob*** approaches. LDA visualization and ***WordCloud*** will be presented in the related part. These visualizations have important representational properties that will help in further analysis.

1 Introduction

Nowadays, online shopping has gradually become a part of our everyday life, primarily those of us who are busiest. For online retailers, there are no constraints to inventory or space management and they can sell as many different products as they want. While brick and mortar stores can keep only a limited number of products due to the finite space they have available.

But for customers, online shopping has its flaws. People cannot verify the quality of the product in person, the only referential information, except the product information provided by the retailers, are reviews published by other customers. So they need to summarize true information other customers want to delivered from all reviews.

Thus I want to apply Natural Language Processing (NLP) techniques, mainly Topic Modeling and Sentiment Analysis to explore the information delivered by the whole reviews dataset.

2 Methodology

To explore the information delivered by Amazon Fine Food Reviews, I mainly follow the following step via Python:

- Pre-processing original data
- LDA analysis
- Sentiment analysis

The detailed process will be stated next.

2.1 Data

The data I used is Amazon Fine Food Reviews published by Stanford Network Analysis Project (2013). This dataset consists of reviews of fine foods from amazon. The data span a period of more than 10 years, including all ~500,000 reviews up to October 2012. Reviews include product and user information, ratings, and a plain text review. It also includes reviews from all other Amazon categories.

This dataset includes:

- Reviews from Oct 1999 - Oct 2012
- 568,454 reviews
- 256,059 users
- 74,258 products
- 260 users with > 50 reviews

just
can
find
great
dont
also
tried
chips
love
make
well
little
best
one
like
now
water
cup
buy
price
teatry
use
time
flavor
drink
product
get
food
taste
coffee
will
chocolate
much
really
better
good

Figure 1: WordCloud of all reviews

2.2 Methods

2.2.1 Pre-processing Data

Before analyzing, we should pre-process~~ing~~ data, like removing the punctuations and stopwords. Thus I use histogram to plot the most frequent terms in overviews first. From Figure 2, we can see stopwords like 'the', 'I' and 'and' tend to be most frequen~~cy~~, which will undoubtedly impact the analysis results.

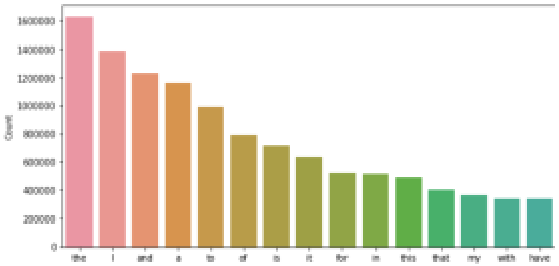


Figure 2: Histogram of most frequent terms

Thus I remove unwanted characters, numbers and symbols, short words (length < 3) from the text at the beginning. And Figure 3 shows the result of this process.

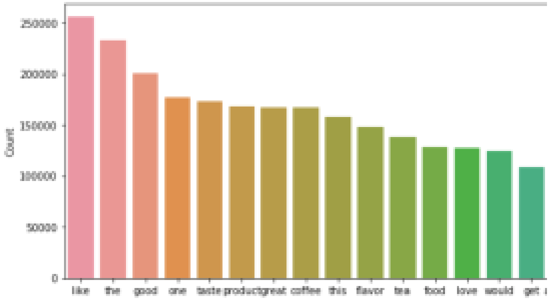


Figure 3: Histogram of most frequent terms after basic pre-processing

Then I use ***Spacy*** to further clean the data. English words have different forms with same etyma, but for text analysis, which will disperse the distribution, leading to lower frequent of the etyma. Thus I apply the ***lemmatization*** function (words in third person are changed to first person and verbs in past and future tenses are changed into present) in ***Spacy*** to process the reviews. And Figure 4 shows the results.

From Figure 4, we can roughly draw some conclusions. Firstly, 'good' ranks top, indicating that most customers give positive feedback~~s~~ on Amazon Fine Food. And 'flavor 'ranks third, meaning customers focused on the flavor of foods most. Also, things 'coffee' and 'tea' ~~also~~ rank high, the most frequent food category in this dataset tends to be liquid food.

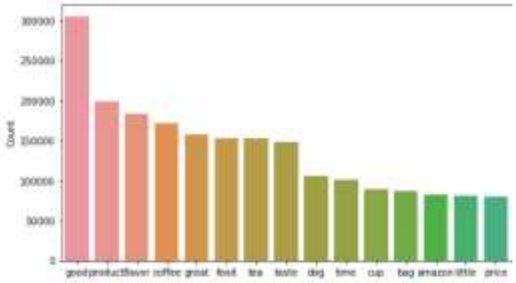


Figure 4: Histogram of most frequent terms after pre-processing

2.2.2 Building LDA Model

After pre-processing data, I apply Latent Dirichlet Allocation (LDA) of topic modeling to analysis. Topic modeling is a type of statistical modeling for discovering the abstract "topics" that occur in a collection of documents. LDA is an example of topic model and is used to classify text in a document to a particular topic.

I expect LDA can show the main topics of the whole Amazon Fine Food Reviews dataset, helping us understand main modes of customer reviews.

2.2.3 Sentiment Analysis

In this part, I apply ***Textblob*** library to do sentiment analysis, mainly focused on polarity and subjectivity. Since these two indicators can show whether the product is worth to buy and to what extent the reviews are reliable.

3 Results

Overall, the LDA model is consistent with Figure 4, since their algorithms are both based on the frequency of terms.

The sentiment analysis indicates that, most feedbacks of foods in this database are positive, encouraging other customers to buy. And the objectivity of these reviews is high enough to serve as references for potential customers. And the specific discussions are stated in the next part.

4 Discussions

4.1 LDA Analysis

In this part, I create the object for LDA model using ***gensim*** library, and apply the reviews after pre-processing as dictionary. I also use ***pyLDAvis*** and ***WordCloud*** to show LDA visualization.

Firstly, I need to determine the specific number of topics in ***gensim.models.ldamodel.LdaModel***. After trying several numbers, I find six is the best choice. This can be explained by Figure 5. The six circles representing six different topics don't overlap in the Intertopic Distance Map, showing these six topics along with their keywords are unique and unduplicated.

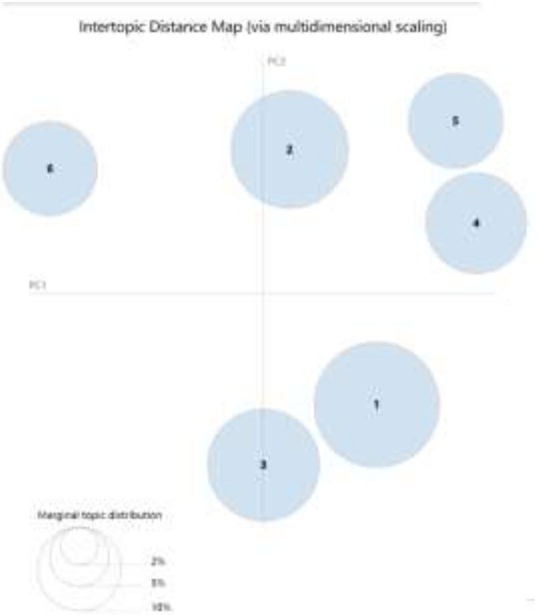


Figure 5: Part 1 of LDA visualization

And Figure 6 is one example of the rest part of LDA visualization (when choosing the fifth topic). The bar are the Top-3 Most Salient Terms in this topic, the blue bar is the overall term frequency of one term, while the red bar is the estimated term frequency of this term within the selected topic.

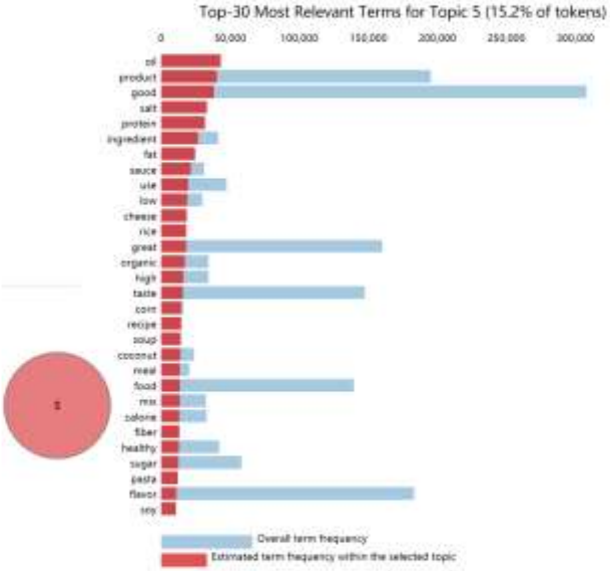


Figure 6: Part 2 of LDA visualization

To view and analyze the six topics easier, I use ***WordCloud***. And the results are showed in Figure 8. We can draw some conclusions from the ***WordCloud***:

- From Topic #0, 'great' and 'good' are most outstanding, showing most customers give positive feedbacks on Amazon Fine Food, which is consistent with Figure 4.
- From Topic #3, 'taste' and 'flavor' are the words with largest percentage. Showing flavor of foods are the key focus of customers, which is also consistent with Figure 4.
- Topic #1, #2, #4 and #5 indicate the main four food categories of reviews: coffee, tea, dog food, protein and oil. And snacks like cookies and chips also appear in the ***WordCloud***, which does not violate our subjective cognition.
- Another point is that, in these four topics ***WordCloud***, 'good' also stands out. This is not weird, as unsatisfactory products will be quickly eliminated in any competitive market, leaving products with good quality. So fine foods in this dataset always with good quality and receive customers praise.

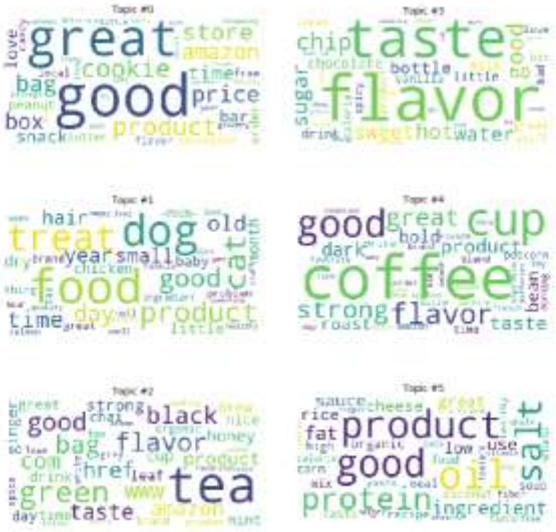


Figure 7: *WordCloud* of six main topics

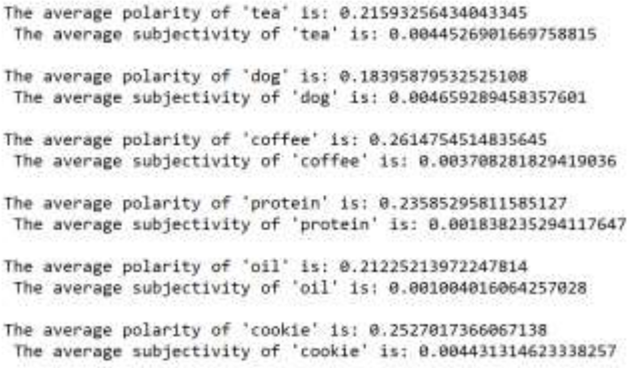
4.2 Sentiment Analysis

In this part, I apply ***TextBlob.sentiment*** to do sentiment analysis. And the results of this function follow these two rules:

- The polarity score is a float within the range [-1.0, 1.0], which -1.0 indicates negative, and 1.0 indicates positive.
- The subjectivity is a float within the range [0.0, 1.0], while 0.0 indicates objective, and 1.0 indicates subjective.

Figure 9 shows the polarity and subjectivity results of six most frequent foods based on LDA analysis. The results are approaching for these six different foods.

The average polarity is fluctuating ~~between~~ 0.2, showing reviews are positive generally. Since there exist unavoidable negative reviews, it is reasonable that the polarity score are not very high. And the average subjectivity is approaching 0, showing the objectivity level of these reviews is very high. Thus customers do not need to worry about the reliability of these reviews. They can refer to existing reviews to make a wise choice.


```
The average polarity of 'tea' is: 0.21593256434043345
 The average subjectivity of 'tea' is: 0.0044526901669758815

The average polarity of 'dog' is: 0.18395879532525108
 The average subjectivity of 'dog' is: 0.004659289458357601

The average polarity of 'coffee' is: 0.2614754514835645
 The average subjectivity of 'coffee' is: 0.003708281829419036

The average polarity of 'protein' is: 0.23585295811585127
 The average subjectivity of 'protein' is: 0.001838235294117647

The average polarity of 'oil' is: 0.21225213972247814
 The average subjectivity of 'oil' is: 0.001004016064257028

The average polarity of 'cookie' is: 0.2527017366067138
 The average subjectivity of 'cookie' is: 0.004431314623338257
```

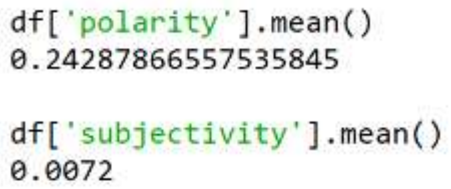
Figure 8: Sentiment analysis for most frequent foods in main topics

Figure 10 is the sentiment analysis results of reviews as a whole, and the results come very close to the results in Figure 9. It is reasonable since these six types of foods are the most frequent of the whole dataset. And this Figure indicates the positivity and objectivity of these reviews once again.

Thus sentiment analysis helps to reveal the emotional tendency of the food reviews, showing

Amazon Fine Foods enjoy great popularity and receive positive reviews generally, which is a good signal for customers who tend to purchase these products.

Also, for potential customers who are still worry about the objectivity of online product reviews, this exploration is a good encouragement. It shows that overall, buyers will post the objective reviews online, which will serve as reliable references for others.



```
df['polarity'].mean()
0.24287866557535845

df['subjectivity'].mean()
0.0072
```

Figure 9: Sentiment analysis for all reviews

5 Limitations

I apply ***TextBlob.sentiment*** to do sentiment analysis in this paper. Though the final results are reliable, this function may not work very well sometimes and the results it returns may have some biases.

A more accurate way ~~is~~ to create new classifiers ~~by~~ my own, like using ***NaiveBayesClassifier***, training and testing the classifier until its accuracy reaches the requirement. But it is a machine learning process that beyond my current level

References

J. McAuley and J. Leskovec. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. WWW, 2013.