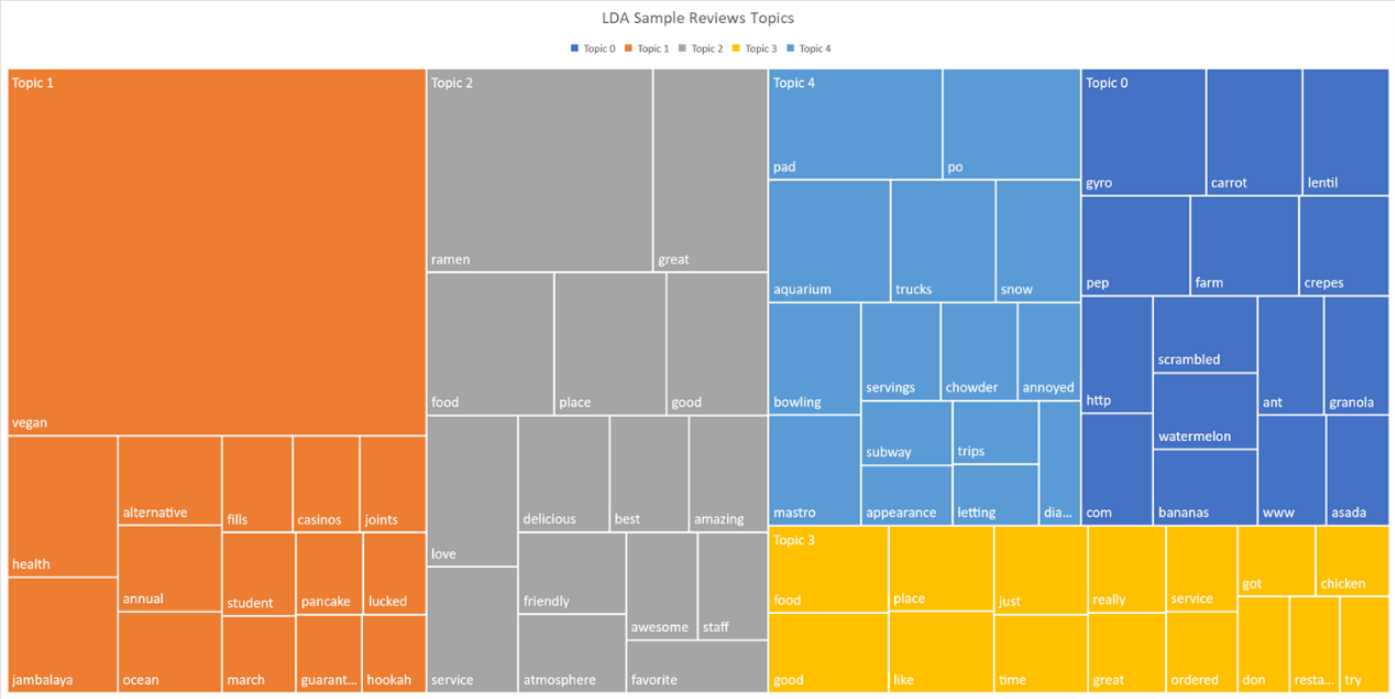
Usage of Topic model

In my data exploration project, I mainly used two types of topic model provided in python gensim package, which are LDA (Latent Dirichlet Allocation), LSI (Latent Semantic Analysis). I will talk about the difference between these two models in later Chapter.

According to gensim's guideline, I used TF-IDF model to vectorize sample corpus, and use Sparse2Corpus to convert sparse vectors to corpus for lda model. Then I used LdaModel with number of topics as 5 to generate topics.

Task 1: What are the topic people talk about on Yelp?

I sampled 100000 reviews from all reviews from given dataset, and generated a TreeMap for the 5 topics generated by LDA model.



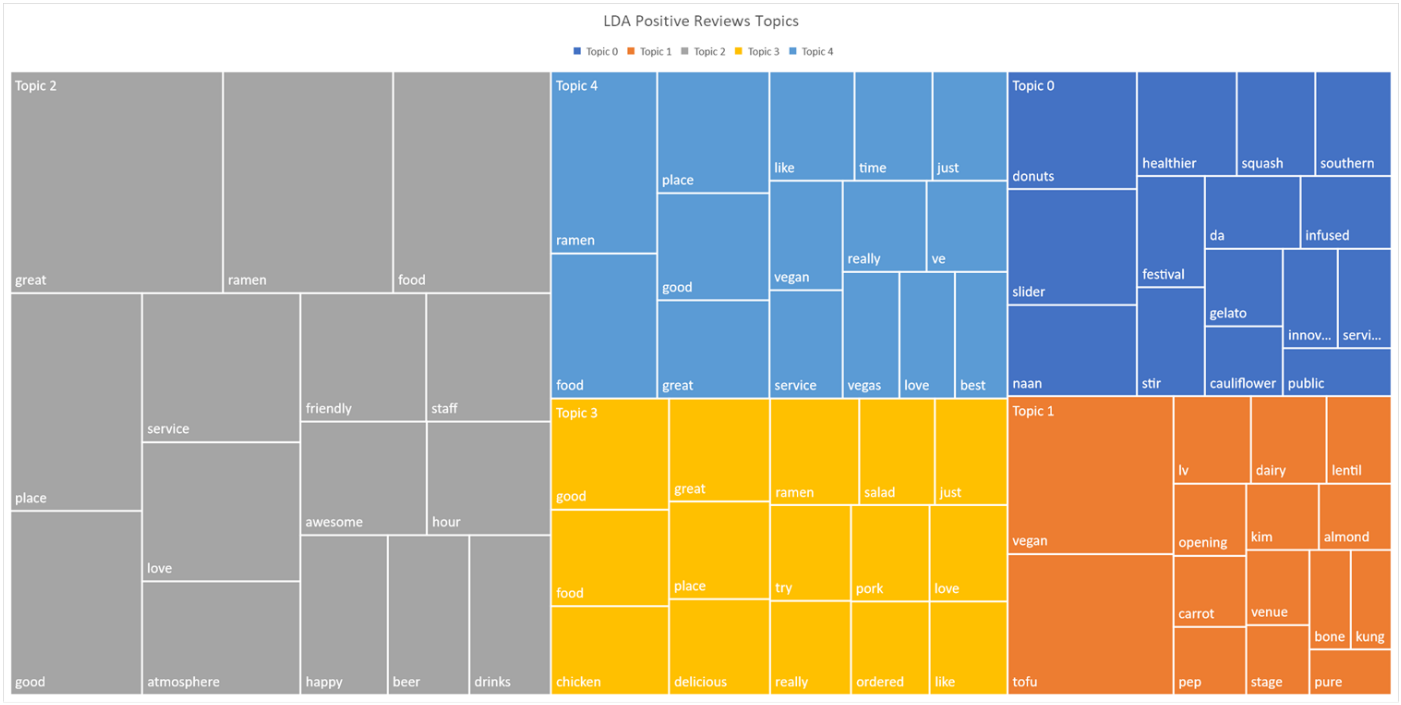
Picture 1 100000 Sample Reviews Topics from LDA model.

In this treemap, different color represents different topics, and the area of each word rectangle represents weight of a particular word in this topic.

Task 2: What are the topics of positive and negative reviews?

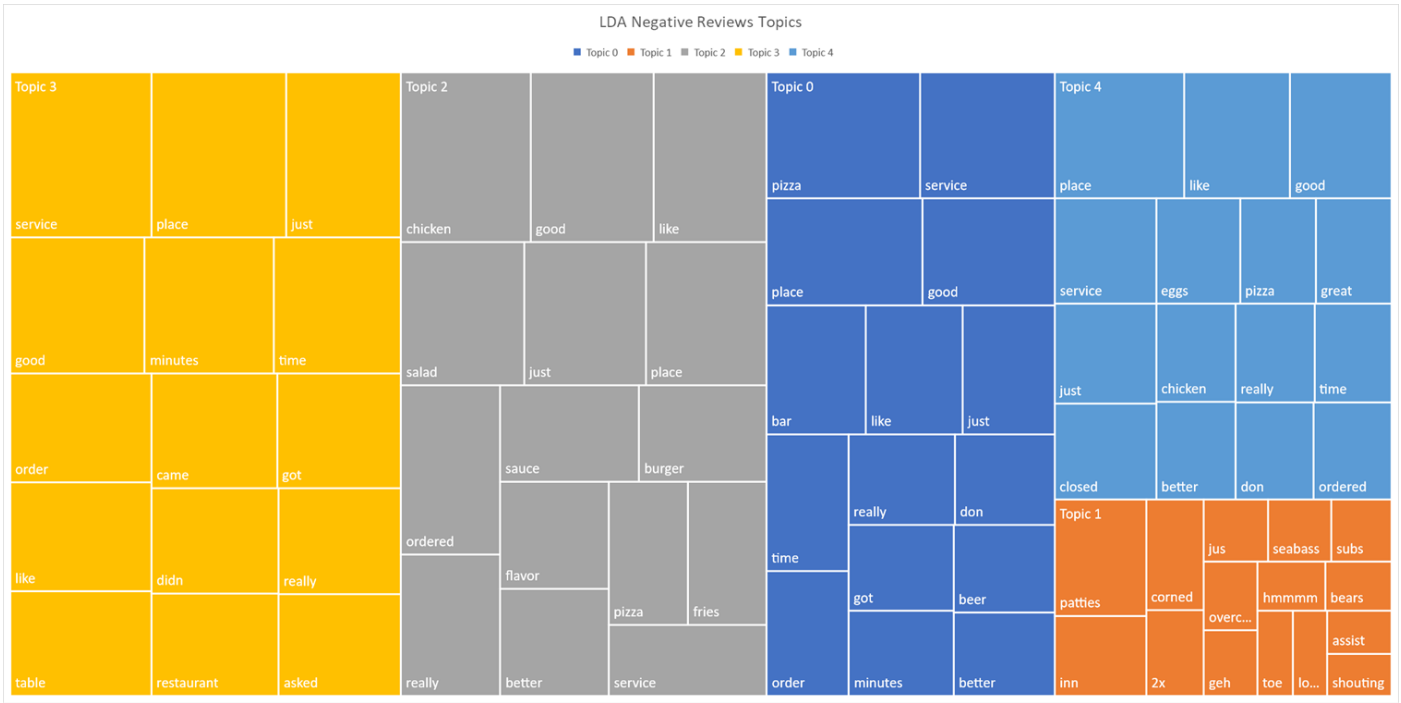
I separated all reviews according to their ratings. I consider reviews with ratings above 3 as positive ratings, while reviews with ratings below 3 as negative ratings. Reviews with ratings exactly 3 is ignored. LDA model is used in this task.

For positive reviews, the 5 topics are visualized as following:



Picture 2: Positive Review Topics from LDA model

For negative reviews, the 5 topics are visualized as following:



Picture 3 Negative Review Topics from LDA model

Comparing these two models, we can identify that they have similarities in every topic when describing what they are commenting about. However, there are some keywords in negative reviews identical, like “didn” and “don”. These are frequently used negative words in reviews. In fact, “not” and “no” is removed when vectorizing with built-in English stopwords.

Task 3: Any difference with Completely Different Cuisines?

I chose two completely different cuisines to compare their review topics, which are vegan and barbeque. I used LDA in this task.

For vegan, 5 topics are visualized as following:

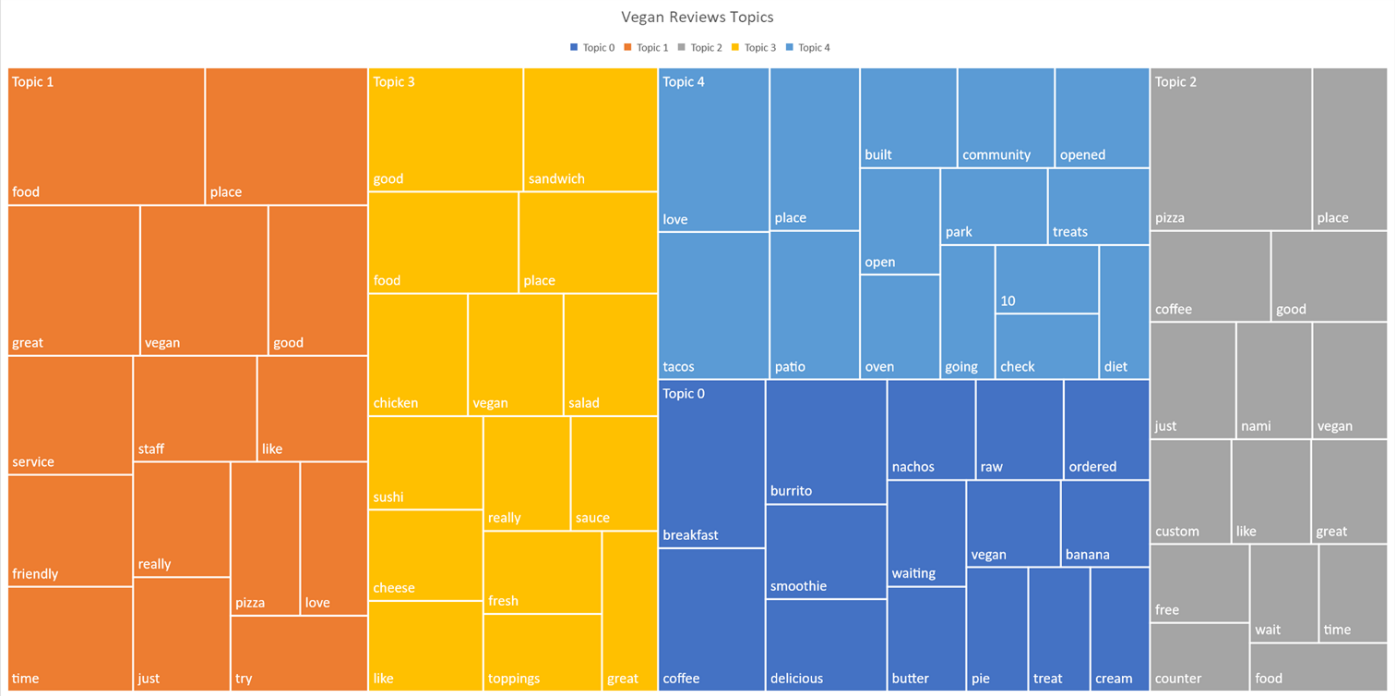


Figure 4 Vegan Cuisine Reviews Topics from LDA Model

For barbeque, 5 topics are visualized as following:

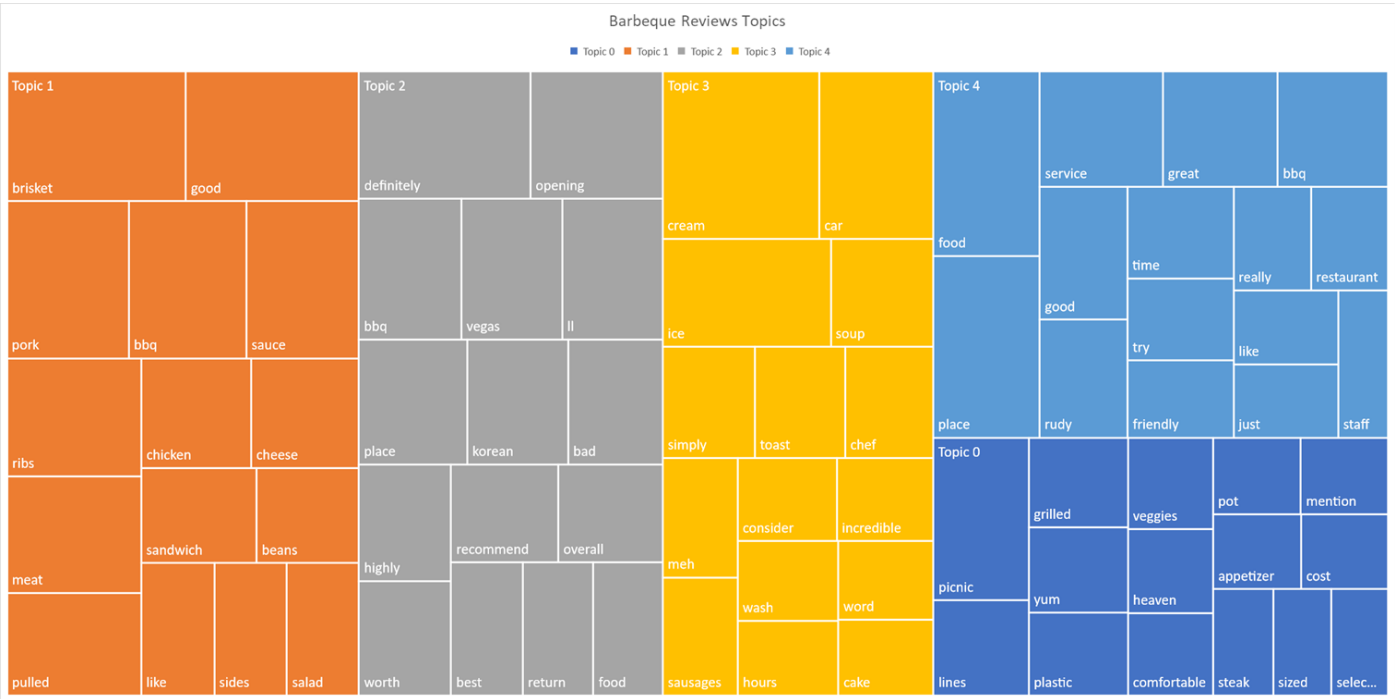


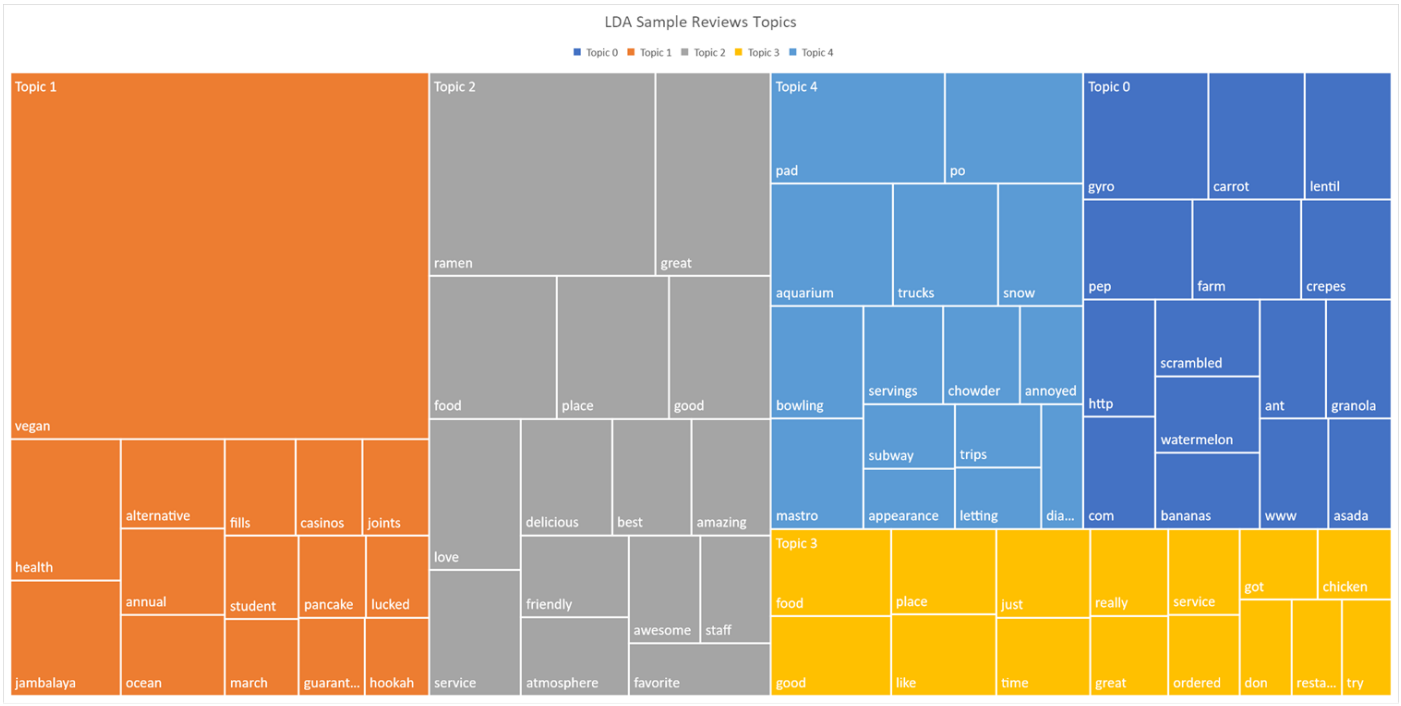
Figure 5 Barbeque Cuisine Reviews Topics from LDA Model

From my personal perspective, they have a little similarity, that vegan topics also included “non-vegan” words like “chicken” and “cheese” (for some strict vegetarians). But barbeque topics include way much more keywords related

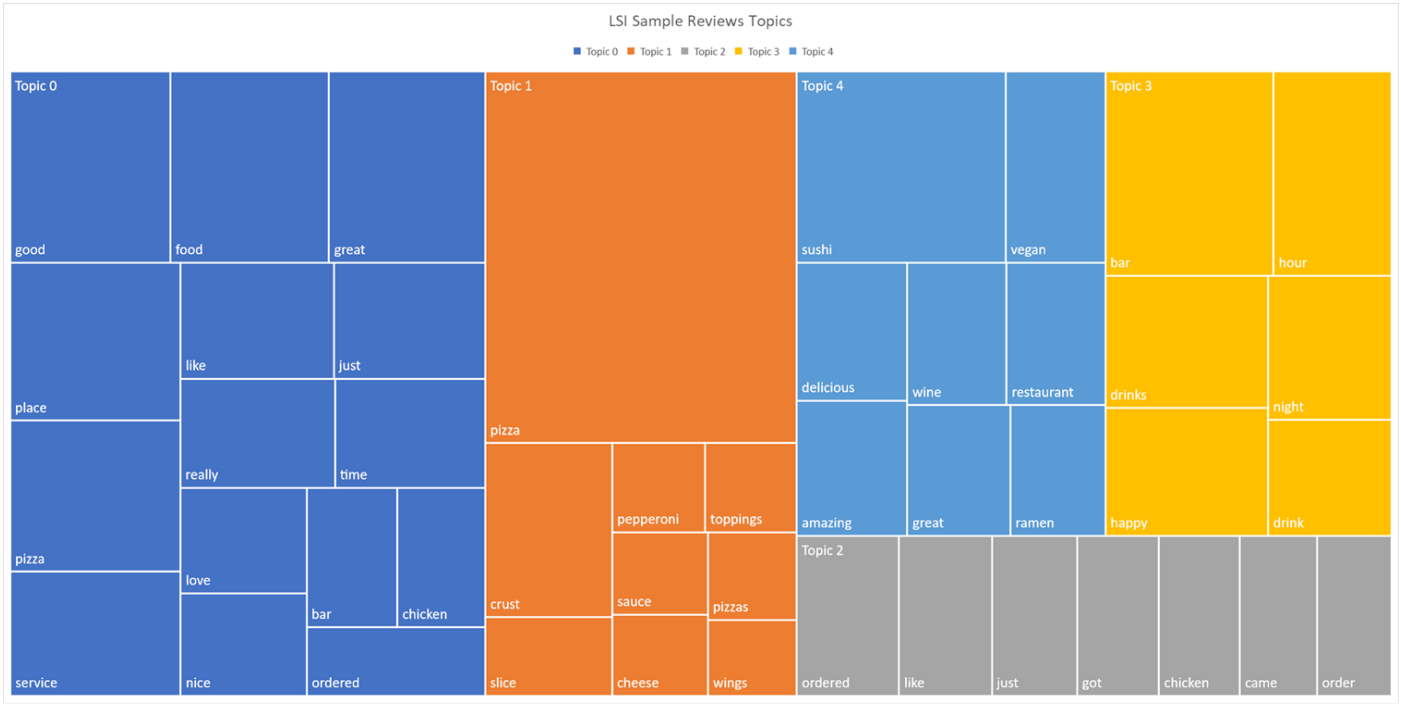
to meat, like “pork”, “ribs”, “sausages”.

Task 4: Different Models

I compared result between LDA model and LSI model. I chose 15 topic words for every model. Since LSI topics included negative weights which indicates that the word has negative contribution to the topic, while TreeMap allows only positive numbers, so the count of keywords in each topic might be different.
The topics for LDA model are visualized as following:



The topics from LSI model are visualized as following:



From the result we can see that topics in LSI model are more exclusive compared with LDA model topics. Like topic 1

in LSI model, there are only dishes described in the reviews, but no judgement positively contributes to this topic. This is untrue because all reviews are supposed to express a positive judgement toward all dishes. So, compared with LDA model, this is less accuracy. From performance perspective, LSI is more efficient.
Thus, LSI is more efficient compared with LDA, but less accurate.

Conclusion

In conclusion, after mining all kinds of topics in different subsets of these reviews, we can see that both LDA and LSI are very helpful topic modeling algorithms. There are several minor problems including the coverage of stop words. The preset topic number also affects generation of topics.