Trang Nguyen

Remesh Take-home Assignment

April 13, 2021

Background

Having open-ended responses data sometimes is a challenge for use to get the thorough ideas of what features our users are looking for when they use social media platforms. However, some responses would have similar contents even though the words used in the answers were different. Hence, applying topic models to the text dataset we have could be the solution to this problem since it is highly valuable to businesses to know what people are discussing and understand their demands and opinions. Topic models provide a simple way to analyze large volumes of unlabeled text. It clusters the words that often occur together and uses contextual clues to connect words with similar meaning and gives us comprehensive "topics." With the new topics we have, we then can validate the responses using the answers from participants preferences associated with the open-ended responses text.

Data Preprocessing

Before applying the topic models to the responses text, we have to preprocess the data. The responses text is removed any words that are non-English. Any distracting characters, punctuations in the sentences are removed as well. We then put every response text into lists of words and create bigram models to make sure we do not miss any meaningful words. Bigrams are two words that are frequently occurring together in the documents (or the responses text in our case), such as back yard, really like, and really appreciate. It is the probability of word w_1 occurring after the word w_2 which is also known as $P(w_2|w_1)$. We also lemmatize words and keeep the words in noun, adjective verb, and adverb forms.

Methods

One of the popular algorithms for topic modeling is Latent Dirichlet Allocation (LDA) [1]. LDA considers each document consists of a combination of topics, and each topic consists of a combination of words in a certain proportion. For instance,

- Document 1: Topic1 = 0.33, Topic2 = 0.33, Topic3 = 0.33
- Topic1: Psychology = 0.25, Computer Science = 0.36, Music = 0.39

LDA is a type of Bayesian Inference Model. It assumes that the topics are generated before documents, and suggested topics that could have generated a corpus of documents. The algorithm also uses collapsed Gibbs Sampling. The algorithm will go through each document and randomly assign each word in the document to one of k topics that is assigned before the algorithm. Then for each document d, go through each word in d and compute:

- $p(topic\ t \mid document\ d)$ which is the proportion of words in document d that are assigned to topic t.

- $p(word\ w|\ topic\ t)$ which is the proportion of assignments to topic t over all documents that come from this word w.

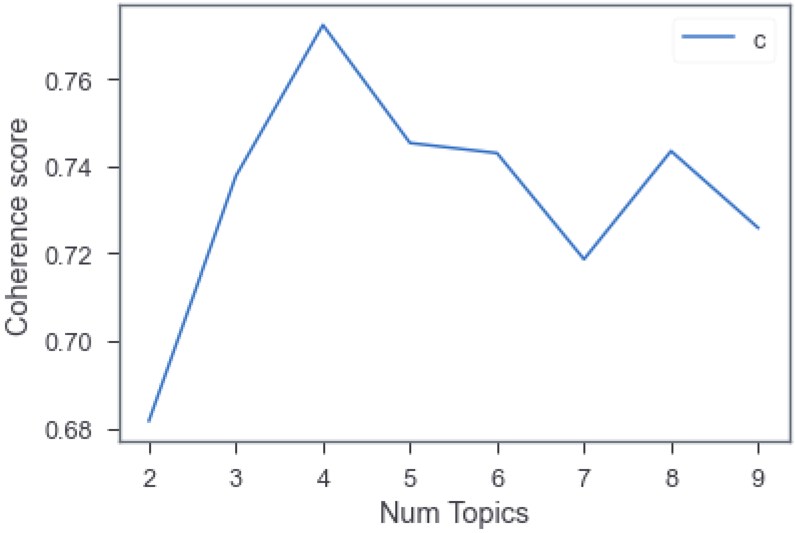
In this scope of this project, LDA Mallet Model [2] was used. Mallet is an open-source toolkit developed by Andrew McCullum which provides us the Mallet Topic Modelling toolkit which includes efficient and sampling-based implementation of LDA as well as Hierarchical LDA. It also has a very fast and highly scalable implementation of Gibbs sampling, efficiency methods for hyperparameter optimization and to suggest new topics for new documents given trained models.

Coherence score which is to decide interpretability and the quality of learned topics was used.

$$Coherence\ Score\ = \sum\nolimits_{i<j} score(w_i, w_j)$$

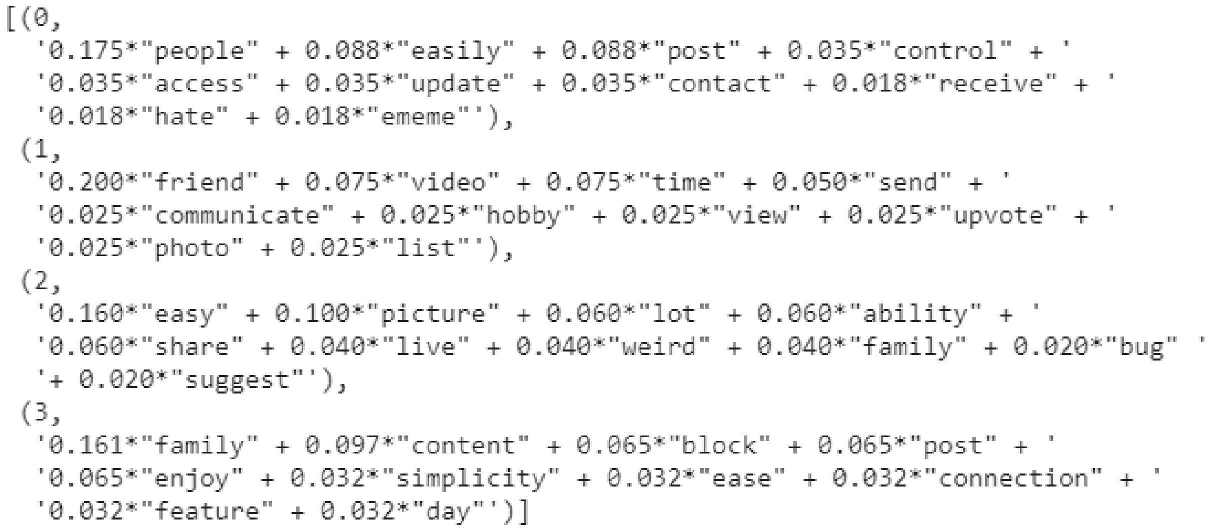
where w_i, w_j are the top words of the topics.

To choose the k topics, we run LDA Mallet model with k goes from 2 to 10 to find the optimal number of topics that would give us the most optimal coherence score. According to Figure 1, at $k = 4$ or the number of topics should be 4, we could get the highest coherence score.



(Figure 1)

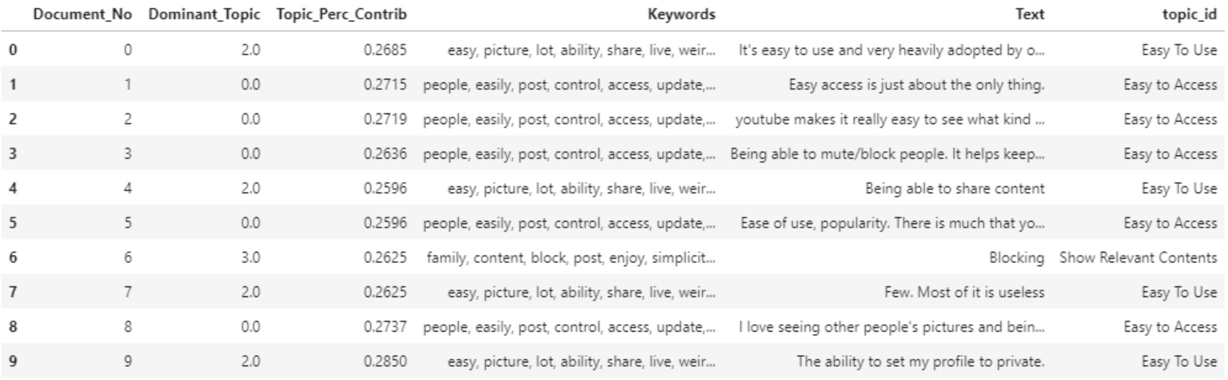
We then extracted the learned topics from the LDA Mallet model with 4 pre-defined number of topics (Figure 2).



```
[(0,
  '0.175*"people" + 0.088*"easily" + 0.088*"post" + 0.035*"control" + '
  '0.035*"access" + 0.035*"update" + 0.035*"contact" + 0.018*"receive" + '
  '0.018*"hate" + 0.018*"ememe"'),
 (1,
  '0.200*"friend" + 0.075*"video" + 0.075*"time" + 0.050*"send" + '
  '0.025*"communicate" + 0.025*"hobby" + 0.025*"view" + 0.025*"upvote" + '
  '0.025*"photo" + 0.025*"list"'),
 (2,
  '0.160*"easy" + 0.100*"picture" + 0.060*"lot" + 0.060*"ability" + '
  '0.060*"share" + 0.040*"live" + 0.040*"weird" + 0.040*"family" + 0.020*"bug" '
  '+ 0.020*"suggest"'),
 (3,
  '0.161*"family" + 0.097*"content" + 0.065*"block" + 0.065*"post" + '
  '0.065*"enjoy" + 0.032*"simplicity" + 0.032*"ease" + 0.032*"connection" + '
  '0.032*"feature" + 0.032*"day"')]
```

(Figure 2)

We then defined the name of our learn topic and find the dominant topic in each response like in the example from Figure 3. Our four defined learned topics from the responses about their favorite features are ‘Easy to Use’, ‘Easy to Access’, ‘Show Relevant Contents’, and ‘Ease of Connectivity.’

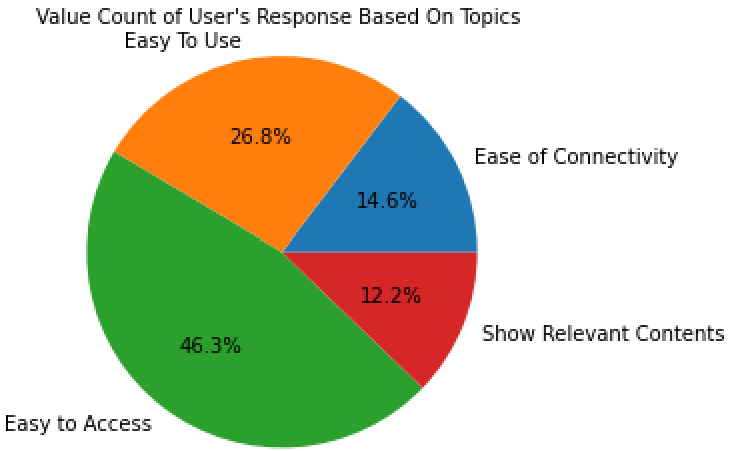


	Document_No	Dominant_Topic	Topic_Perc_Contrib	Keywords	Text	topic_id
0	0	2.0	0.2685	easy, picture, lot, ability, share, live, weir...	It's easy to use and very heavily adopted by o...	Easy To Use
1	1	0.0	0.2715	people, easily, post, control, access, update,...	Easy access is just about the only thing.	Easy to Access
2	2	0.0	0.2719	people, easily, post, control, access, update,...	youtube makes it really easy to see what kind ...	Easy to Access
3	3	0.0	0.2636	people, easily, post, control, access, update,...	Being able to mute/block people. It helps keep...	Easy to Access
4	4	2.0	0.2596	easy, picture, lot, ability, share, live, weir...	Being able to share content	Easy To Use
5	5	0.0	0.2596	people, easily, post, control, access, update,...	Ease of use, popularity. There is much that yo...	Easy to Access
6	6	3.0	0.2625	family, content, block, post, enjoy, simplicit...	Blocking	Show Relevant Contents
7	7	2.0	0.2625	easy, picture, lot, ability, share, live, weir...	Few. Most of it is useless	Easy To Use
8	8	0.0	0.2737	people, easily, post, control, access, update,...	I love seeing other people's pictures and bein...	Easy to Access
9	9	2.0	0.2850	easy, picture, lot, ability, share, live, weir...	The ability to set my profile to private.	Easy To Use

(Figure 3)

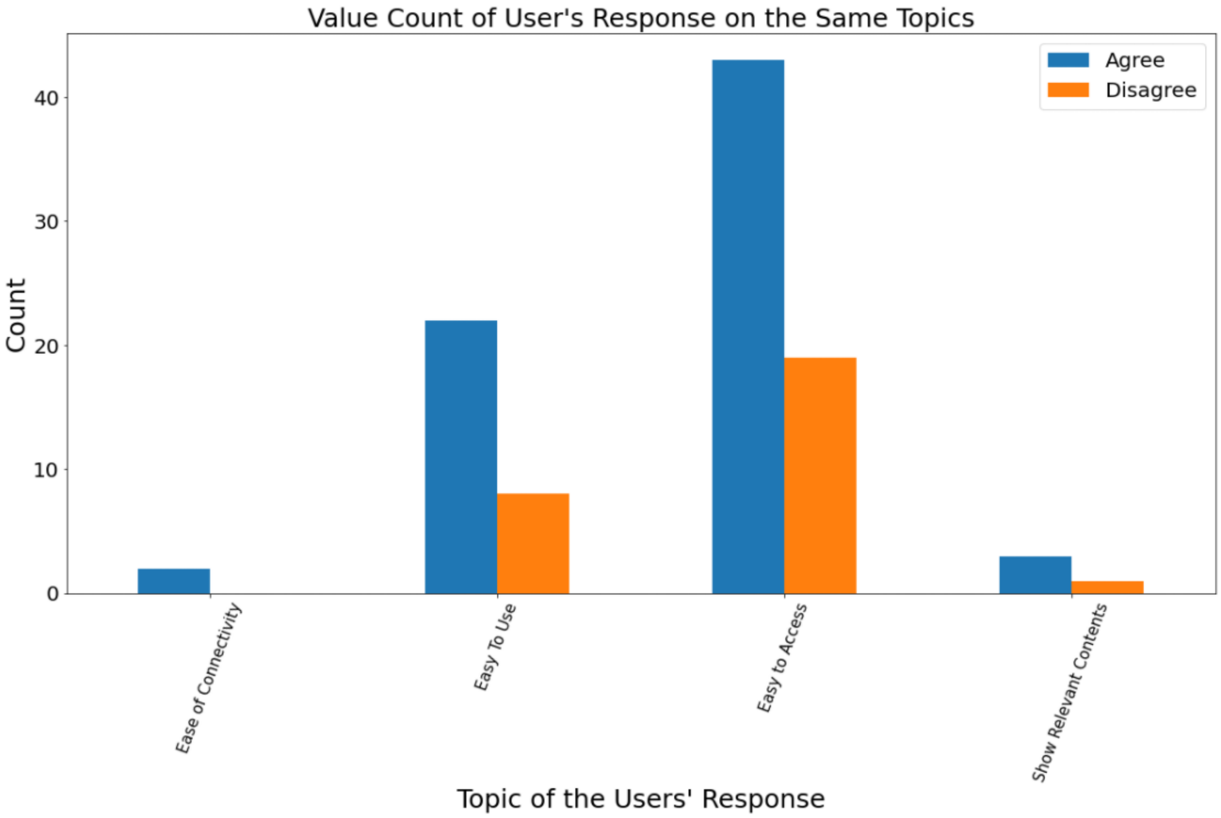
Analysis

After topic modeling, we now can proceed with the analysis of the given data.



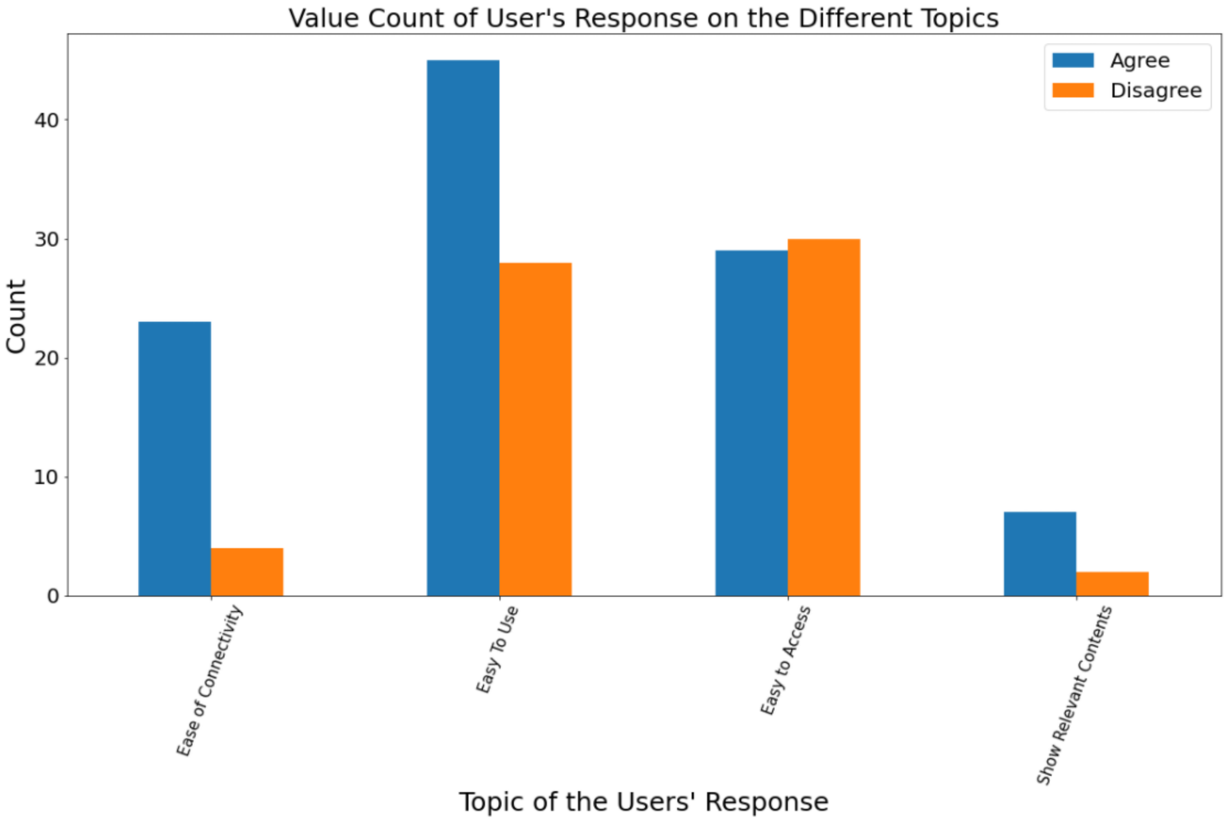
(Figure 4)

Figure 4 shows that 46.3% of the text responses about the favorite feature on social media platform are related to "Easy to Access" topic while "Show Relevant Contents" has lowest number of favorable. However, since these answers are based on the questions of what features do the user like about the social media platforms that they use, all of these four topics should be taken into consideration to produce a new product.



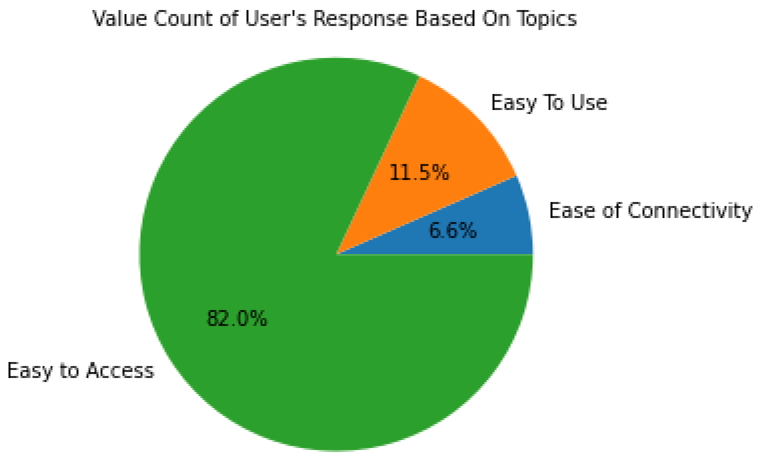
(Figure 5)

We then take further analysis to learn more about the data. Figure 5 shows the frequency of responses from respondents if they are consistent with the previous answers or not when they are given the response from the same topic. As we can see, all the responses are pretty consistent with the text responses. On the other hand, when the respondents are given responses with different topic from their answer in text responses, the results are quite interesting (Figure 6). There is a barely different in opinion with "Easy to Access." However, there are a lot of agreements in the rest of the topics. This would give us a better idea of what feature we should prioritize to produce first.



(Figure 6)

Then we work around with dataset which has binary choices. From Figure 7, we can see that if the respondents that wrote easy to access text responses are consistent with their answers even though they are given different kind of responses. This might give us ideas that 'Easy to Access' is one of the most important features when using social media platforms and we can set the 'Show Relevant Contents' to lower priority since the number of responses related to this topic are quite low throughout our analysis.



(Figure 7)

Drawbacks and Future Improvement

LDA is unsupervised learning model; hence, the way we defined the name for learned topic might be bias and subjective. LDA works better with data with larger sample size; however, with the trained model, we can set the baseline for the future when we could collect more data. We could also make use of the topics we have for classification problems later if needed.

In the text responses, there are some answers which do not give a clear answer of which feature they like. For example, “Few. Most of them are useless,” LDA still couldn’t define this well. However, since our goal is to find favorable feature to develop our own, these answers will not add in any insight to our needs anyway; hence, it is bearable for us to have these responses misclassified.

Conclusion

From all the above analysis, feature that helps users to access easily without any problems to their social media platforms should be the main priority to develop out new feature at Remesh. After that, we can consider developing the feature that can help users to use social media platform in an easier way. Feature related to showing relevant contents should be at low priority based on the responses from the respondents.

References

[1] https://www.seas.harvard.edu/courses/cs281/papers/blei-ng-jordan-2003.pdf

[2] http://mallet.cs.umass.edu/topics.php