**Milestone 2**

*Anak Agung Ngurah Bagus Trihatmaja, Anupam Gupta, Ayush Bandari, Xiaomeng Peng*

## **Abstract**

From the exploratory data analysis done in milestone 1, we want to, using totally unsupervised method, discover the product categories based on the review. To achieve this, we will run LDA on our review dataset. Therefore, in this milestone 2, we tuned hyperparameters of LDA mode for Amazon Fine Food Reviews [1]. The purpose of tuning the hyperparameters is to find the best model for LDA. We also quickly evaluate our model by using TSNE to see if there is any structure.

## **Sensitivity Studies**

In this sensitivity analysis, we try to change the hyperparameters of LDA model and determine modeling parameters.

There are three parameters:

1. Topic number N: Number of topics in this reviews
2. Alpha: Dirichlet distributions for (per document) topic
3. Beta: Dirichlet distributions for (per word) topic

The trained LDA represents the topic-distributions for the documents of the training set, and can therefore be used to compute the likelihood of unseen documents, i.e. test dataset. The performance of LDA in unseen documents decides the quality of LDA.

The metric we used to evaluate the model is perplexity. Perplexity is a measurement of how well a probability distribution or probability model predicts a sample. Low perplexity indicates the probability distribution is good at predicting the sample.

For this study, we run pre-processing for our dataset which includes stopwords removal, stemming and converting the text into bag of words model. We will divide the dataset into 60% training and 40% for test dataset. We run LDA against our training dataset and then measure the perplexity in our test dataset.

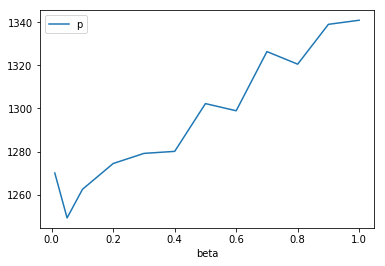


Figure 1. Perplexity with beta values [0.01, 1], alpha = 0.01, and 30 topics

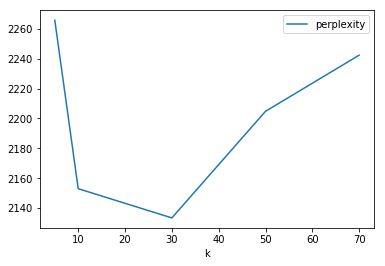


Figure 2. Perplexity with beta values 0.01, alpha = 0.01, and number of topics [5,70]

From Figure 1, Dirichlet hyperparameter *β* affects perplexity, when topic word prior probability increase, in general perplexity increases. The best value of *β* is 0.05. A high *β* value means that each topic is likely to contain a mixture of most of the words, and not any word specifically, while a low *β* value means that a topic may contain a mixture of just a few of the words. In addition, a high *β* value will lead to topics being more similar in terms of what words they contain. From our simulation result, lower *β* is better. Then each topic is determined by few number of words.

In Figure 2, we pick our number of topics by evaluate the perplexity in function of number of topics. There, we see that the lowest perplexity obtained by assigning 30 as the number of topics.

## **TSNE Visualization of Current Model**

We also run t-SNE to see the visualization of our model in 2D dimension. Each color represents the topic assigned from the LDA for our test dataset. The hyperparameters of the LDA is *β* = 0.05, k = 30. From the plot we can see a little bit of structure, however it seems still random as we see many points also overlapping with other groups.

In this visualization, we randomly sample the data into 1000 dataset, to make the computation runs faster. However we also worried that our approach fails to make use the information that the undisplayed data points provide about the underlying manifold.

## **Future Work**

1. Perplexity might not be the best measure to evaluate the topic model as it doesn’t consider the context and semantic association between words. Therefore for the next step, we will also incorporate NPMI (Normalized Pointwise Mutual Information) which we can run through a NLP library called *gensim* [2].
2. As our dataset doesn’t have categories of the product, we need to get the categories by scrapping the HTML of the product page on Amazon. We also have ratings and usefulness of a review.

[1] http://snap.stanford.edu/data/web-FineFoods.html

[2] https://radimrehurek.com/gensim/models/coherencemodel.html