**LDA Topic Modeling on Amazon Review**

Anak Agung Ngurah Bagus Trihatmaja, Ayush Bhandari, Anupam Gupta, Xiaomeng Peng

# **1. Introduction**

Latent Dirichlet Allocation (LDA) is a powerful unsupervised learning method to identify topics within the documents and map documents to this topic. In this project, we evaluate LDA on Amazon reviews dataset [1] to identify whether LDA can discover the products category solely based on review. Getting good topics in LDA is challenging, since the quality of LDA depends on the predefined number of topics and hyperparameters such as prior of document topic distribution (𝛂) and prior of topic word distribution (𝛈). In order to find an optimal LDA model for this dataset, sensitivity analysis between 𝛂 and 𝛈 and the effect of pre-processing will be studied. The optimal LDA model will be selected based on different metrics. Metrics such as perplexity and Normalized Pointwise Mutual Information (NPMI) are standard tools to evaluate LDA. Based on our study, NPMI is proved to be a good measurement in determining the number of topics for LDA.

## **1.1. Exploratory Data Analysis**

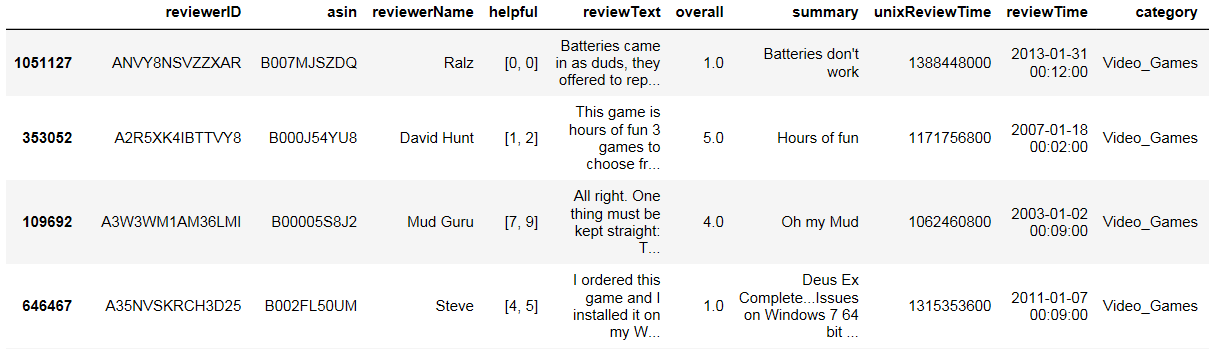
Dataset Overview:

The dataset has total 200,152 reviews collected from Amazon from 1999 to 2013. For each review, it has following attributes:

Table 1: Attribute of reviews

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| reviewId | Row Id |
| asin | Unique identifier for the product |
| reviewerId | Unique identifier for the reviewer |
| Helpful | [Number of users who found the review helpful, Number of users who voted for whether they found the review helpful] |
| overall | Rating between 1 and 5 |
| UnixReviewTime | Time index for the review |
| Time | Timestamp for the review |
| reviewText | Text of the review |
| Category | Category/topic of the product |

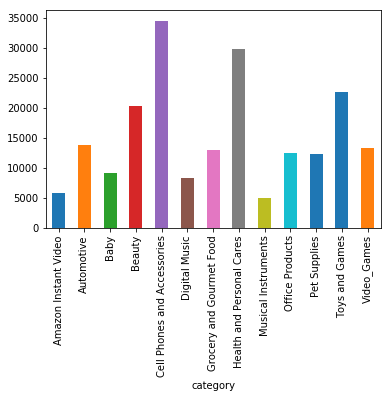
The reviews are illustrated in Figure 1.

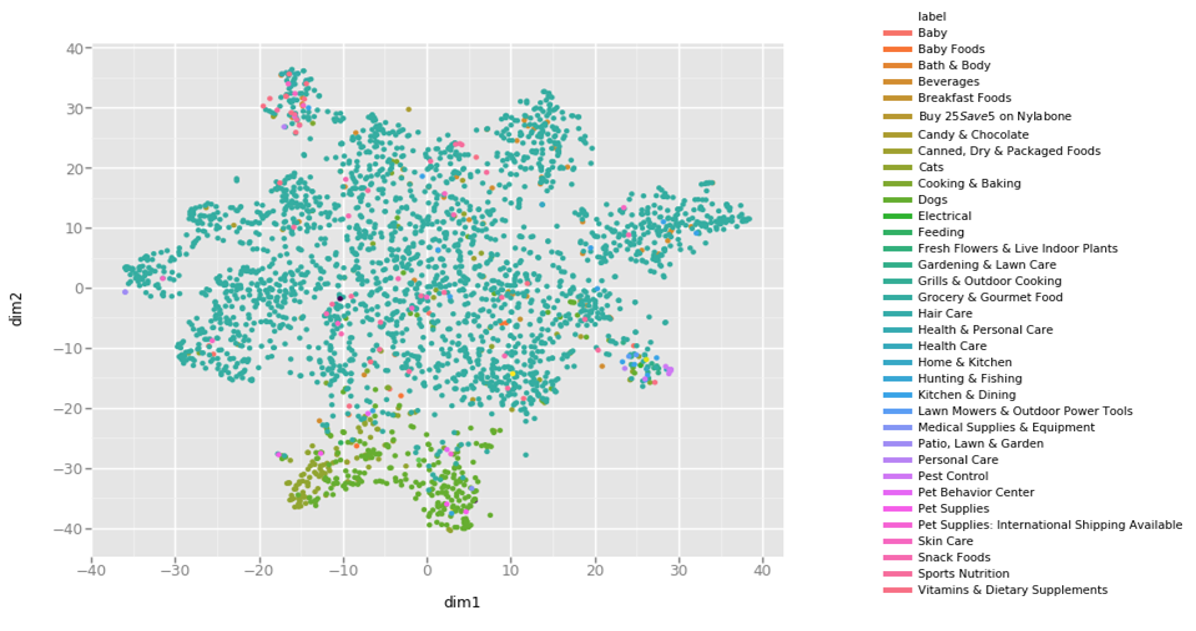


*Figure 1.* Illustration of the dataset

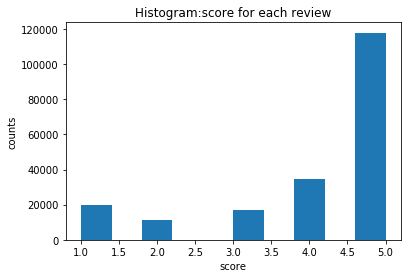
In this dataset, we have 13 categories such as: Video Games, Cell Phones and Accessories, Health and Personal Cares, Toys and Games, Beauty, Automotive, Baby, Office Products, Pet Supplies, Grocery and Gourmet Food, Digital Music, Amazon Instant Video and Musical Instruments. As shown in Figure 2, the number of reviews from different categories are similar, which means the dataset can be considered as balanced. If the dataset is imbalanced, LDA will not efficient separate the categories since majority of the topics will tend to be majority class. For example, if 82% of reviews in one dataset are belongs to one specific category: Grocery and Gourmet Food. The LDA result is shown as Figure 3. Clearly, the clusters are not separable. Thus, with imbalanced dataset, the quality of LDA will not be ensured.

Figure 4 indicates that most of our reviews are positive reviews which has rated the products as 4 or 5 stars. This characteristic will affect the result of topic modeling, we will discuss this part in Section 3.3.

*Figure 2.* Number of reviews in categories



*Figure 3.* T-sne representation of LDA result of imbalanced dataset

*Figure 4.* Number of reviews in each score

# **2. Data Pre-Processing and Hyperparameter Analysis**

In performing LDA, we split our datasets into three categories: development, training and test categories with random seeds of 42. Our development dataset consists of ~ 20,000 reviews. The purpose of development datasets is to speed up our hyperparameter (𝛂 and 𝛈) tuning. The perplexity and NPMI are measured against our test dataset.

Perplexity is a metric used to evaluate language models. It is a measure of how well a probability model predicts a sample. It can be defined as exponentiation of entropy where entropy is the amount of information contained in the variable. The lower the perplexity the better.

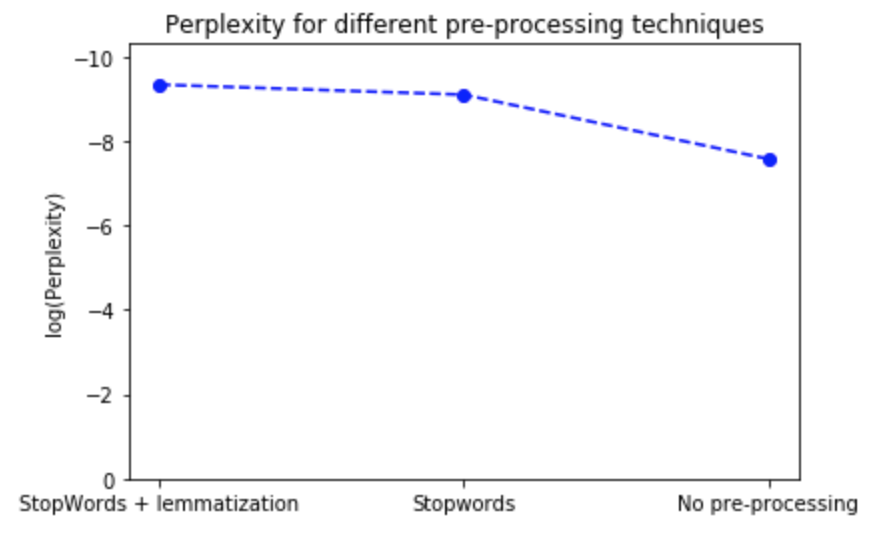
Similarly, normalized pointwise mutual information (NPMI) in topic modelling is a measure of association between words in topics. Unlike PMI, Normalised PMI (NPMI) reduces the bias for PMI towards words of lower frequency. NPMI also provides a standardised range of [−1, 1] for the calculated values [2]. When NPMI is -1 means there is no co-occurrence between the topic words and the text, 0 means co-occurrence at randoms and 1 means complete co-occurrence.

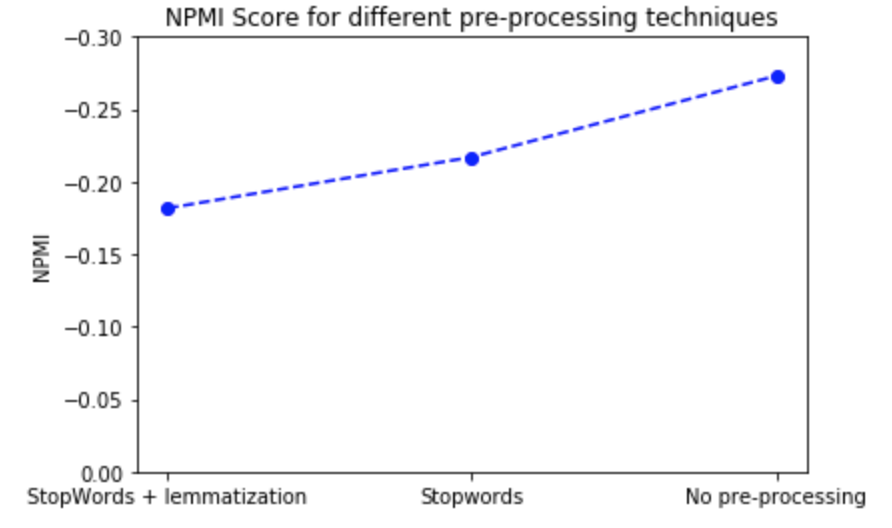
## **2.1. Pre-processing Analysis**

We evaluate the LDA model under three different experimental setting and try to find best pre-processing for LDA model. Three different setting are shown as below:

1. Removal of stopwords + lemmatization
2. Removal of stopwords only
3. No pre-processing

Figure 5 shows LDA with removing stopwords plus lemmatization will have lowest perplexity. As shown in Figure 6, we also get the highest NPMI score for removing stopwords plus lemmatization. Thus, we can infer that removal of stopwords and lemmatization provides us with a better corpus to run topic modelling on to get clear, segregated and meaningful topics.

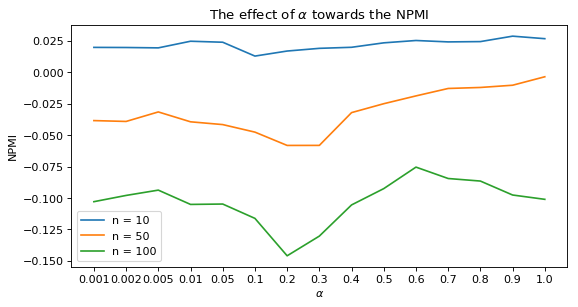
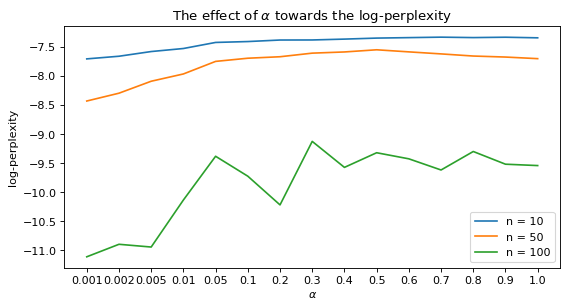
*Figure 5.* Perplexity for different pre-processing 

*Figure 6.* NPMI for different pre-processing 

## **2.2. Sensitivity Analysis: Finding the 𝛂 and 𝛈**

Document topic distribution (𝛂) and prior of topic word distribution (𝛈) play a role in the result of LDA. Most implementation of LDA assumes symmetric distribution. A high 𝛂 on symmetric distribution will make each document is likely to contain a mixture of most of the topics, and not any specific single topic. In the other hand, a low 𝛂 would make a document contains mixture of just a few or single topics more likely. Similarly for the 𝛈, a higher 𝛈 means that each topic would contain a mixture of most of the words, while lower 𝛈 means that the topic would contain a mixture of a few words.

We evaluate the perplexity and the NPMI with various alpha and number of documents 10, 50 and 100 to determine the 𝛂. As we can see in Figure 7, the perplexity keeps on increasing with higher 𝛂, except when the number of topics is at 100. When the topics is 100, there are some fluctuations in our perplexity although we can consider this still increasing. In another graph, we can see the more fluctuate results of NPMI. We see that for 10 topics, the perplexity is lower when 𝛂 is below 0.01, but it drops for 50 and 100 topics. Here we conclude that the sweet range of 𝛂 is [0.001, 0.01].



In finding 𝛈, we use the same procedure as finding 𝛂, that is we evaluate the perplexity and the NPMI with various alpha and number of documents 10, 50 and 100 to determine the 𝛈. Again, in our case, the lowest perplexity starts at 0.001. In the other hand, for some number of topics, we have a good perplexity at 0.01. Here we also conclude that the sweet range of 𝛈 is [0.001, 0.01].

## 

## 

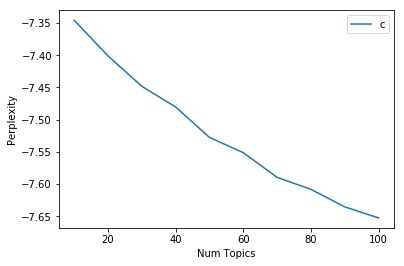
## 

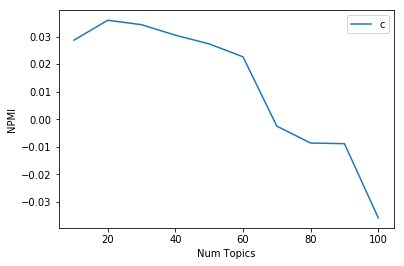
## 

## **2.3. Finding the Optimal Number of Topics**

Defining the optimal number of topics in LDA is important factor to get the most relevant topics. However, there is no hard rule in how many topics we should pick in LDA. The number of topics varies depending on the task we want to perform in the dataset.

We use two metrics in determining the number of topics, which are perplexity and NPMI. Perplexity is often used to get the initial idea on how many topics we need to set. As previously mentioned, the lower the perplexity the better. However, the value of the perplexity does not correlated with human judgement [3] -- we might get a lower perplexity value but the topics might not be relevant according to the human judgement. Thus, we will also use another metric called Normalized Pointwise Mutual Information (NPMI) in determining the number of topics. It is said NPMI has largest correlation with human judgment [4].





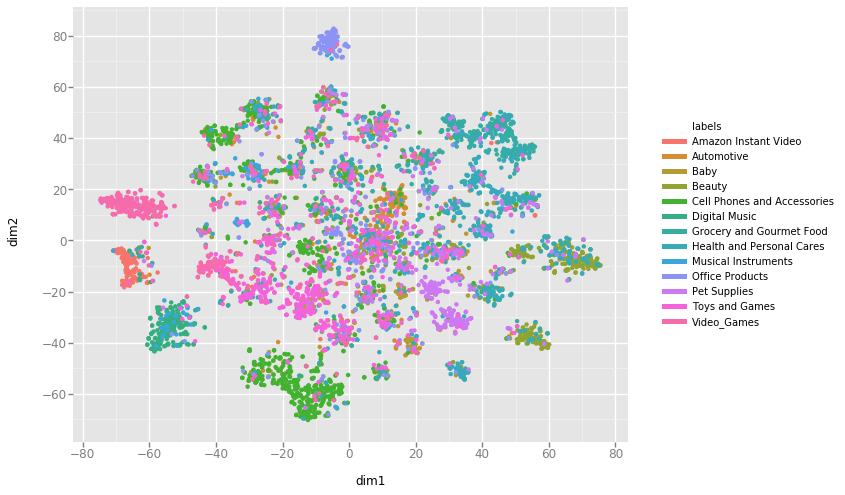
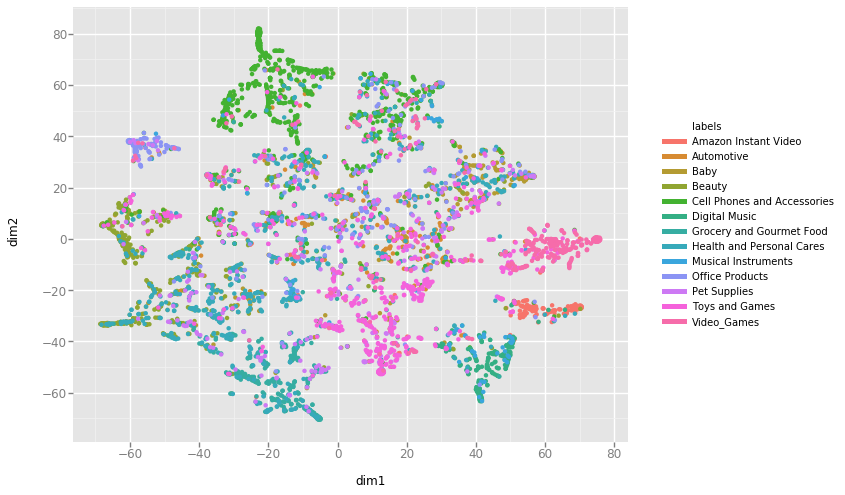
In our case, the perplexity and the NPMI shows different result. On the figure above are the results of assigning different number of topics towards perplexity. We fix the 𝛂 = 0.01 and 𝛈 = 0.01. The result of the perplexity shows that 100, as it is the lowest perplexity, is the best number of topics, while in the other hand, NPMI show the best number of topics is 20 as it shows the highest perplexity. In the result, we will see the difference between assigning 20 topics and 100 topics to our LDA model.

# **3. Result and Discussion**

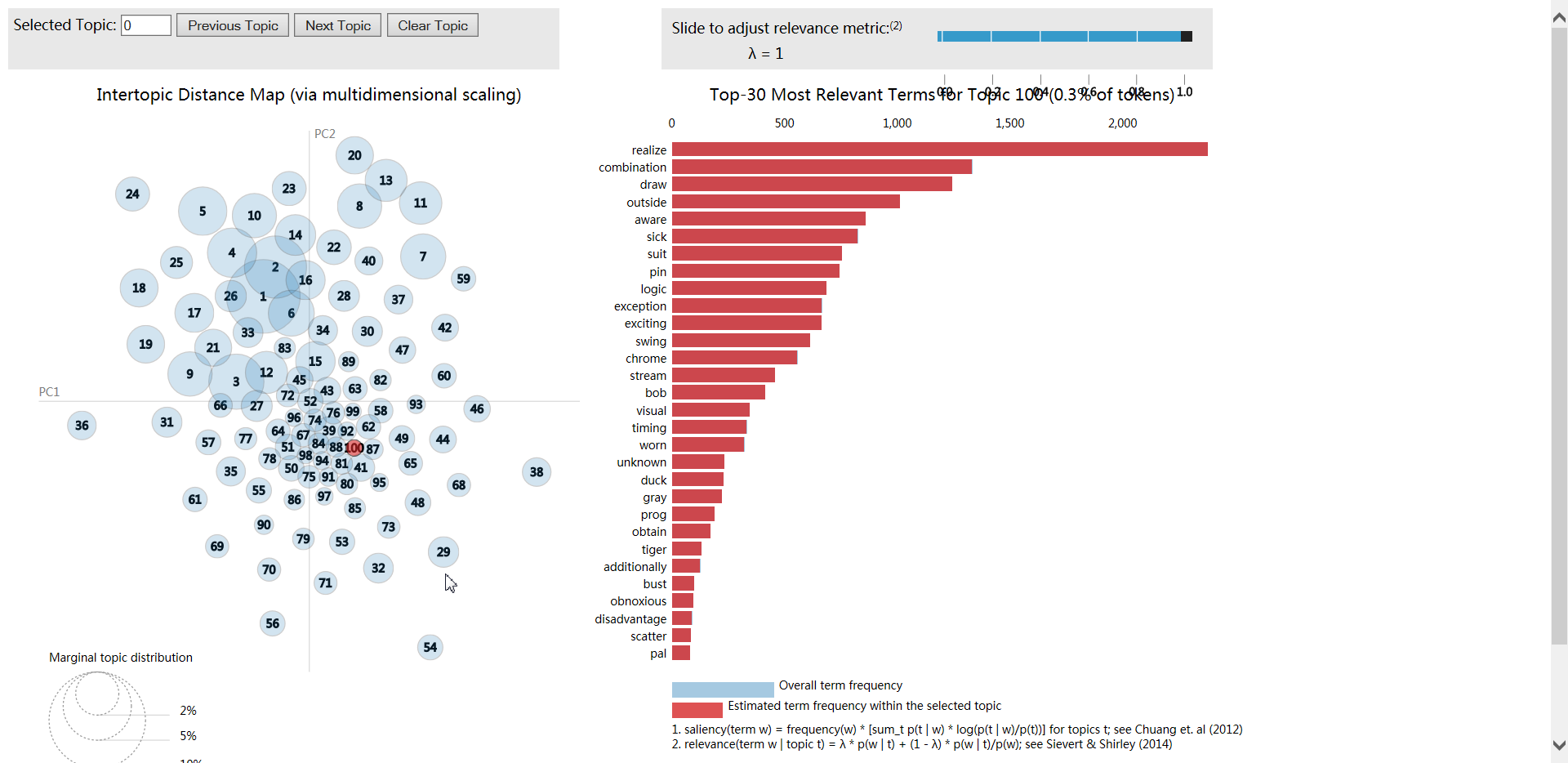
The LDA results on 20 topics shows clearer topic separation compared to the 100 topics. We evaluate our LDA results in steps as follow:

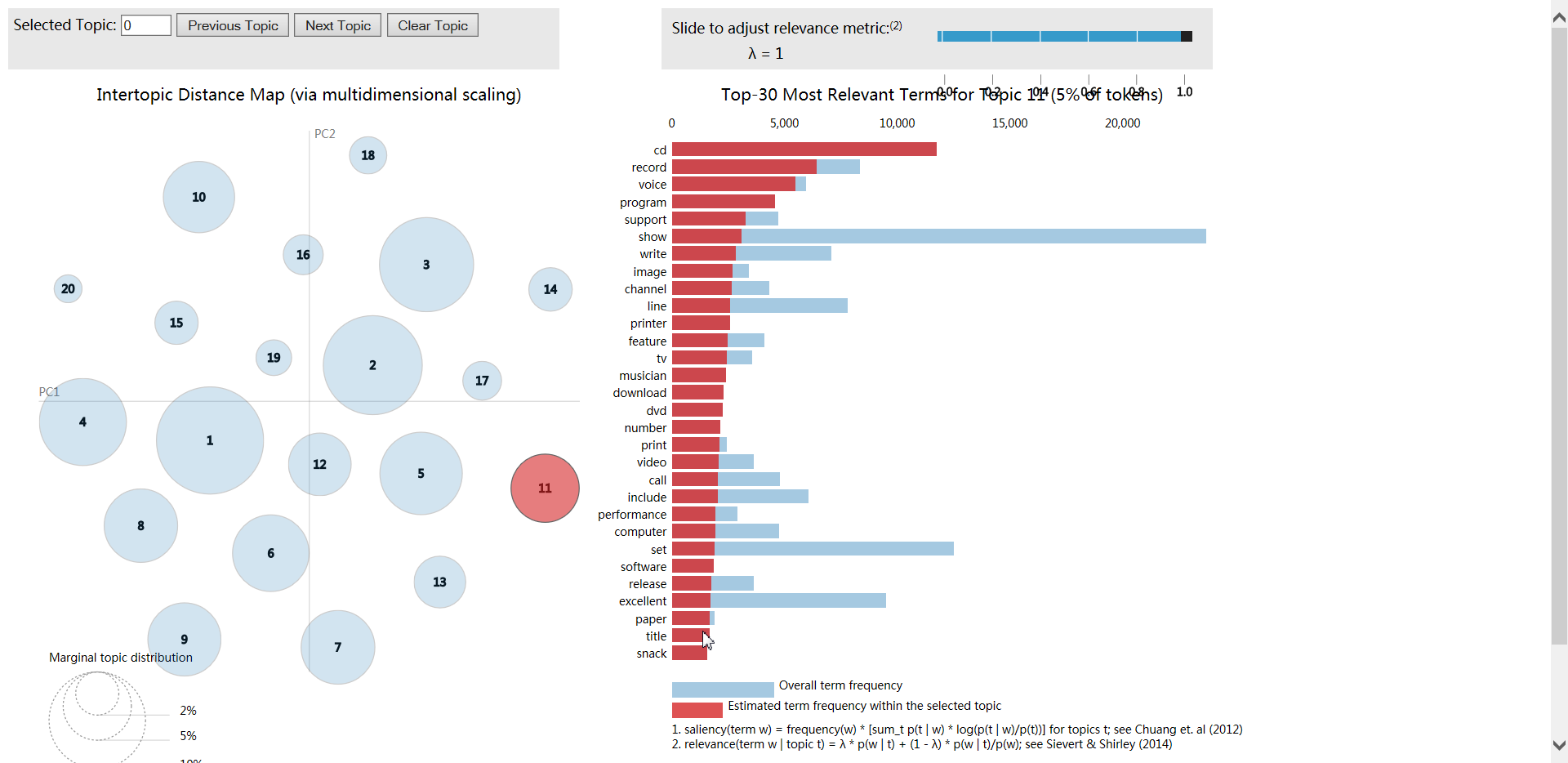
1. We pre-process our data by removing stopwords, lemmatization and converting to bi-gram.
2. We fix 𝛂 = 0.01 and 𝛈 = 0.01.
3. We run LDA in all of our training dataset.
4. We transform our LDA model to our test dataset to get the soft-assignments.
5. We run T-SNE to evaluate the result.
6. We also evaluate the result using pyLDAvis. pyLDAvis is a python library for topic model visualization [5].

## **3.1. LDA Result Visualization**



Our T-SNE result shows that each topics, in LDA with 20 topics, are farther from each other. While in the other hand, the topics in LDA for 100 topics, are closer to each other. To get a clearer view, we will see the result from pyLDAvis. With 20 topics, we have more separable topics, while for 100 topics, we can see some topics get overlapping with the other topics. Better LDA model has more separable topics.





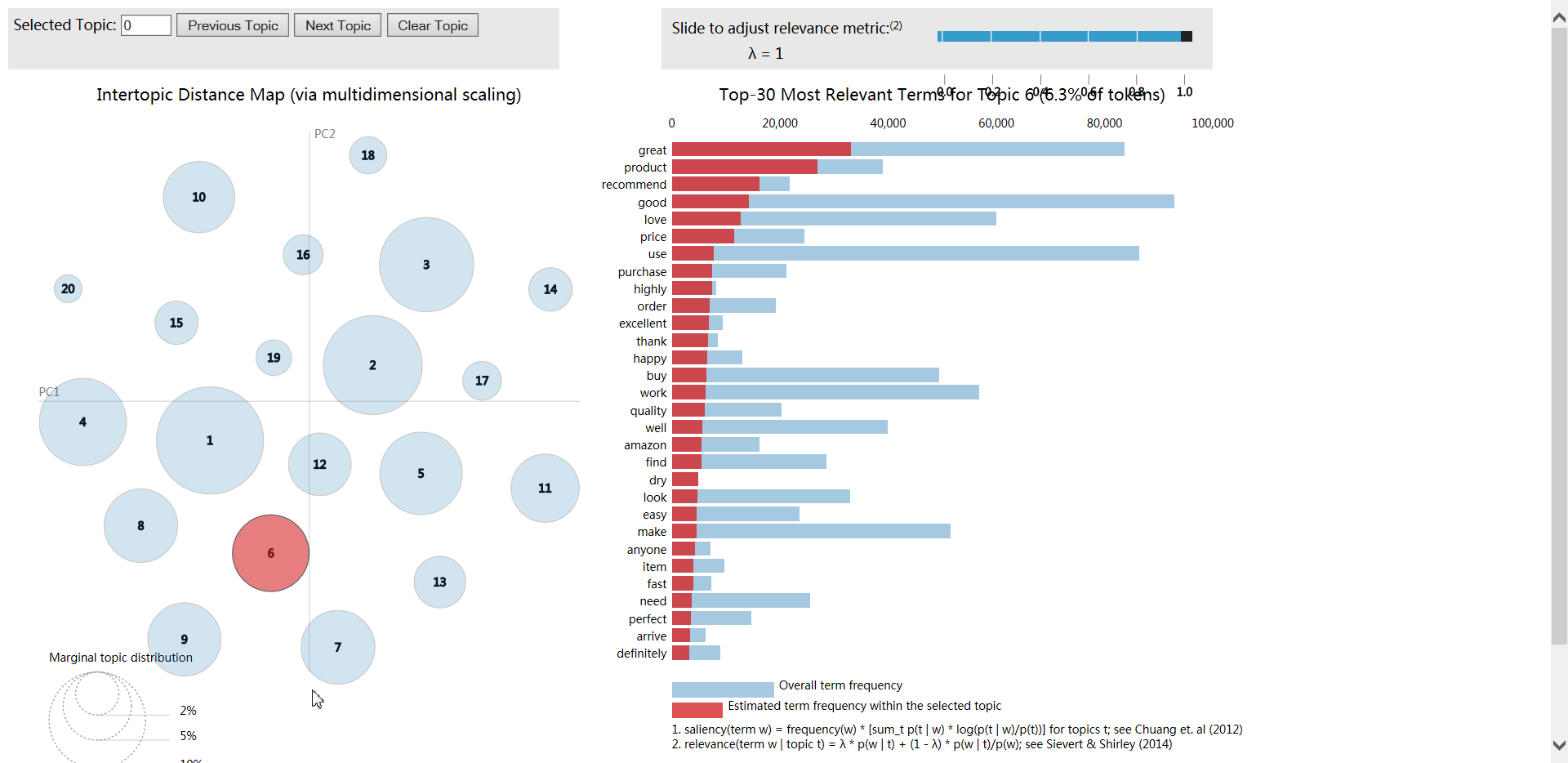
## 

## **3.2. Example of LDA Result**

|  |  |
| --- | --- |
| **Text** | **Assigned Topic:** Cell Phones and Accessories |
| I use this desk headset every day in both my office and my home. I love the way it feels. I love the clarity when I talk into the phone. I love the fact that I can see my screen and carry on a private conversation. I have an iphone 4 and I really like the button in the middle of this headset because it allows me to hang up, or answer the phone from the headset... hold that same button down and it will pull up voice control and I can ask my phone to call someone for me. Click the button lightly and it will pick up on the last thing playing on itunes. This product is easy to use, designed with the consumer in mind and most of all it makes it enjoyable to talk on the phone again. I love the retro feel and look. The fact that I get a little less radiation because I'm holding this headset to my head instead of my phone is just an added bonus. |  |

We use one test from our test dataset and get the soft assignment. From the word cloud, we can see the assigned topic with the largest probability consists of words such as phone, charge, battery, headset. The words in the word cloud represents the keywords assigned to that topic and the size of the word represents the importance (weightage) of the word in that particular topic. From there, we can infer that the topic assigned to this text is about phone and its accessories.

## **3.3. The Effect of Good Reviews**



In our exploratory data analysis step, we showed that the majority of the reviews are positive reviews. Therefore it is inevitable that we have one topics containing only positive words, such as good, great, recommend, love, which are not relevant to a type of product categories.

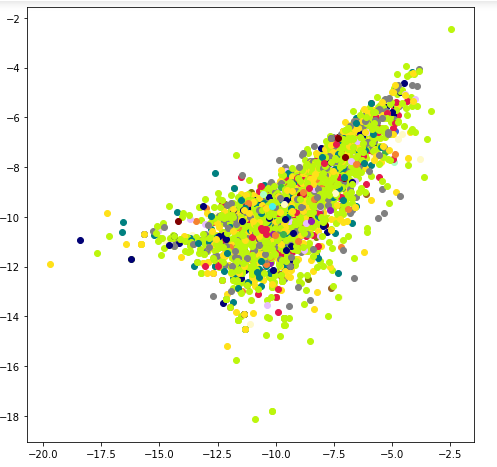
**3.4 Comparison with PCA and k-means**

In order to verify whether another method can yield better results dimensionality reduction aka Principal Component Analysis and k-means clustering was applied.

The procedure used was as follows:

1. Stop words were removed
2. Stemming was performed
3. The reviews were converted into document term matrix format
4. PCA was performed and the dimensions was reduced to 2
5. K-means clustering was performed and the data was clustered into 50 clusters
6. The clustered data was visualized in the form of a 2-D graph

The following was the result



As it can be seen the many reviews overlap with each other resulting in the plot being inconclusive.

# **4. Conclusion and Future Work**

## **4.1. Conclusion**

Based on the experiment, we have shown:

1. Pre-processing affect the quality of LDA
2. NPMI has largest correlation to human topic coherence ratings
3. LDA topics also affected by the abundant of positive words available on most of the reviews
4. PCA on text data is not useful as it results in information loss.

## 

## **4.2 Future Work**

1. Another implementation of LDA called LDA Mallet, which often said give a better result
2. Outside of NPMI, there are another metrics for coherence model such as UMass for intrinsic measure and UCI for extrinsic measure which we can use to determine the best number of topics

# **References**

[1] R. He, J. McAuley. Modeling the visual evolution of fashion trends with one-class collaborative filtering. WWW, 2016.

J. McAuley, C. Targett, J. Shi, A. van den Hengel. Image-based recommendations on styles and substitutes. SIGIR, 2015.

[2] Lau, JH, Newman, D, Baldwin, T. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In: European Chapter of the Association for Computational Linguistics (EACL), 2014, 530.

[3] Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. Reading tea leaves: how humans interpret topic models. In Proceedings of the 22nd International Conference on Neural Information Processing Systems (NIPS'09), 2009, 288-296.

[4] Röder, Michael, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. Proceedings of the eighth ACM international conference on Web search and data mining. ACM, 2015.

[5] Sievert, C & Shirley, K.E. LDAvis: A method for visualizing and interpreting topics. Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces. 63-70.

[6] Řehůřek, Radim & Sojka, Petr. Software Framework for Topic Modelling with Large Corpora. 2010. 45-50. 10.13140/2.1.2393.1847.