Latent Dirichlet Allocation and Predatory Pricing Online Data

Xiaotian Hu and Shanshan Ding

February 24, 2021

Abstract

In this paper, we study Latent Dirichlet Allocation (LDA; Blei et al., 2012) for topic modeling of Amazon unfair pricing data during Covid-19. A topic model is designed to capture topics relating to words in text document or corpus. LDA is a generative probabilistic model with helping to collect topics from discrete data, like text & corpora. It is also known as a three-level hierarchical Bayesian model, where each item of the collection is modeled as a finite mixture over an underlying set of topics. For each topic, it is modeled as an infinite mixture on an underlying set of basic topic probabilities in turn. We conduct analysis of unfair pricing data by sellers from Amazon during the Covid-19 period using LDA. Specifically, we perform topic modeling and generate topics under Amazon product description. Our goal is to capture information and topics on what kind of surgical masks and products are in unfair pricing during Covid-19. Finally, we conclude that N95 is the most unfairly priced product under the topic modeling. By generating graphical illustrations with the Python pyLDAvis package, we are able to summarize and provide more detailed information based on Predatory Pricing Online model.

1. Introduction

For a collection of documents, if an article talks about sports, it may discuss basketball, archery, swimming, etc. Then, the frequency of words related to basketball should be higher in the part of basketball, such as 'basket','net', etc. Similarly, the frequency of words related to archery should be higher in the part of archery, such as 'pull', 'Bow', 'arrow', etc. In the part of swimming, words related to swimming should appear more frequently, such as 'water', 'swim' and so on. Therefore, a document should contain multiple topics, and each of the topics should account for a different proportion. In addition, each topic should consist of many words, and the proportion of each word in each topic is likely different.

Topic modeling uses a mathematical framework to reflect the feature of a document. A topic model automatically analyzes each document under specific corpus, counts the words in the document, and determines which topics are contained in the current document or corpus and the proportions of the different topics according to statistical information. Latent Dirichlet allocation (LDA) is a popular topic modeling method and is a technique that we can use to generate topics and summarize documents.

LDA was proposed by Blei, Ng, and Jordan in 2012 and published in Journal of Machine Learning Research. The paper describes the model and method in detain and discusses its broad applications in document modeling, text classification, and collaborative filtering. The paper also compares the LDA model with a mixture of unigrams model and the probabilistic LSI model to demonstrate advantages of the LDA model. This is our main reference paper for mathematical foundations. In addition, we also refer to the paper of 'Analysis of Variational Bayesian Latent Dirichlet Allocation: Weaker Sparsity than MAP' by Nakajima et al. (2014) for theoretical results and properties.

On the other hand, the paper 'Online Learning for Latent Dirichlet Allocation' by Hoffman, Blei and Bach (2010) proposes an online LDA model that can handily analyze massive document collections, including documents arriving in a stream. It has been shown to produce good parameter estimates dramatically faster than batch algorithms on large datasets. Kapadia (2019) offers a good guidance to perform topic modeling based on Python and is very helpful for us to implement the methods in different scenarios and parameter settings. Moreover, akashii(2019) demonstrates the implementation and mechanism of LDA, where two methods including variational inference and collapsed Gibbs sampling were used for model fitting, providing a useful connection between technical background and implementation.

Finally, we conduct the analysis of predatory pricing online data from Amazon during the Covid-19 period using LDA models to generate topics and make conclusions. We firstly perform data cleaning and processing and then use the Workcloud package in python to illustrate word distribution and importance based on the processed data. We then train LDA models and generate topics, and further produce topic plots with the pyLDAvis package. Based on the topic and graphical results from the fitted models, we find that the disposable N95 mask is the most unfairly priced product under the topic modeling.

We organize the paper as follows. We introduce notation & terminology and describe the LDA methods in Section 2. We introduce the Amazon unfair pricing data and data processing in Section 3.1. We prepare LDA model training in Section 3.2, and present results and conclusions in Section 3.3. Finally, Section 4 provides discussions and future improvement of the work.

2. LDA topic modeling

This section provides a review of LDA topic modeling by Blei, Ng, and Jordan (2012).

2.1 Notation & Terminology

We will use the language of text collections throughout this paper instead of the entities such as "words", "documents", and "corpora". It will help to make inference and capture abstract notions when we introduce variables in the model. In order to help with understanding, we define the following terms: 1). the fundamental unit of discrete data is a word, defined as an item from 1 to the vocabulary of V which is noted as 1,...,V. We represent words using unit-basis vectors which have a single component equal to 1 and all other components equal to 0. Thus, we need to set up a way to denote the components. The vth word in the vocabulary is represented by a V-vector w such that $w^v = 1$ and $w^u = 0$ for u = v; 2). We denote the document under a sequence of N words as W = $(w_1, w_2, ..., w_n)$, where w_n is the nth word; 3). When M documents are collected, it forms a corpus where we denote as $D = (W_1, W_2, ..., W_M)$.

2.2 Latent Dirichlet allocation

Latent Dirichlet allocation (LDA) is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words.

LDA assumes the following generative process for each document w in a corpus D: 1. Choose $N \sim Poisson(\xi)$; 2. Choose $\theta \sim Dir(\alpha)$; 3. For each of the N words w_n: (a) Choose a topic $z_n \sim Multinomial(\theta)$; (b) Choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

In the model, some simplifying assumptions are made. First, the dimensionality k of the Dirichlet distribution is set as known and fixed. Second, the word probabilities are parameterized by a k x V matrix β where $\beta_{ij} = p(w^j = 1|z^i = 1)$, which we treat as a fixed quantity. The quantity is to be estimated.

A k-dimensional Dirichlet random variable θ can take values in the $(k-1)$-simplex (a k-vector θ lies in the (k-1)-simplex if $\theta_i \geq 0, \sum_{i=1}^{k} \theta_i = 1$), and has the following probability density on this simplex:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k})\alpha_i}{\prod_{i=1}^{K} \Gamma(\alpha_i)} {\theta_I}^{\alpha_1 - 1} ... {\theta_k}^{\alpha_k - 1},$$

where the parameter α is a k-vector with components $\alpha_i > 0$, and where $\Gamma(x)$ is the Gamma function. The Dirichlet is a convenient distribution on the simplex — it is in the exponential family, has finite dimensional sufficient statistics, and is conjugate to the multinomial distribution.

Given the parameters α and β, the joint distribution of a topic mixture θ, a set of N topics z, and a set of N words w is given by:

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n, \beta)$$

,

where $p(z_n|\theta)$ is simply θ_i for the unique i so that ${z_n}^i = 1$. We get the marginal distribution of a document by Integrating over θ and summing over z:

$$p(w|\alpha, \beta) = \int p(\theta|\alpha)(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta)p(w_n|z_n, \beta))d\theta$$

.

Finally, we are able to get the probability of a corpus by taking the product of the marginal probabilities in single documents:

$$p(D|\alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha)(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn}, \beta))d\theta_d.$$

The LDA representation has three levels: corpus level, document level and word level. The parameters α and β are corpus level parameters, which are proposed to be generated once before processing through a corpus. The variables θ_d are document-level variables, sampled once per document. At last, in each document, the variables z_{dn} and w_{dn} are word level variables, sampled once for each word.

3. LDA for Predatory Pricing Online Datas

3.1 Data Cleaning & Preparing Data for LDA Analysis.

The data are collected from a blog (Paul, 2020) that investigates unfair pricing practices due to the Covid-19 pandemic, called the predatory pricing online data. The data are extracted from amazon original product database. The predatory pricing online data contain variables including details about product

reviews, product discounts, the time of listing of products, and product descriptions.

Covid-19 is an infectious disease that has caused global pandemic and coronavirus outbreak. When we fight coronavirus disease, some sellers such as mask sellers had utilized this emergent situation to gain benefits much more than usual sales. Consequently, the supplies of masks in many countries have suffered shortage. For masks sold on Amazon and some other e-commerce companies, unfair pricing has become a severe problem to customers as the prices of some masks and pandemic related products has increased to unbelievable levels.

The data that we study focus on products on surgical masks from amazon especially for products with unfair pricing. The data are extracted from Amazon on March 9th, 2020 and obtained based on datahut's platform. The sample size of the dataset is 532 (products), meaning that we have 532 collected products. There are 12 variables in the dataset including 'Product Name', 'Asin', 'Product url', 'Brand Name', 'Image url', 'MRP', 'Sale Price', 'Discount Percentage', 'Product Description', 'Date First Available', 'Number of Reviews', and 'Seller Name'. In order to perform LDA for the dataset, we need to select the key variables. LDA is a technique that can be used to generate topics from corpus. We target on some variables that has enough information about the products to generate topics from the data. After variable selection, we finalize two variables to be used in the data analysis, which are 'Product Name' and 'Product Description'.

After the pre-processing, our next step is to remove the effect of the missing value in the data. Due to missing at random, we delete the rows where the missing values are located, and then unify the cases of the text. Also, removing punctuation is in great importance to make our data more clear for LDA analysis. The aim for the above step is to make description of the product more amenable for the data analytics.

3.2 LDA Model Training

After cleaning and preparing, we next verify whether the processing of the product description works well. We import the Wordcloud package (wordcloud 1.8.1, https://pypi.org/project/wordcloud/) to get a visual representation of the most common words first. This is a package in python, which using to summarize important contents(or words) in corpus or documents. It can help to check whether the data are suitable for the analysis.

Figure 1 is a plot generated from word cloud. It indicates that the text and corpus works well and the description is mainly about face masks. It also generates some content around the main topic of face masks. Next, we are able to train data to get preparation for plotting the most frequent words. Figure 2 is a generated plot of ten most frequent words. Also, this is a further examination

Figure 1: Wordcloud Figure.

of the picture of the word cloud plot.

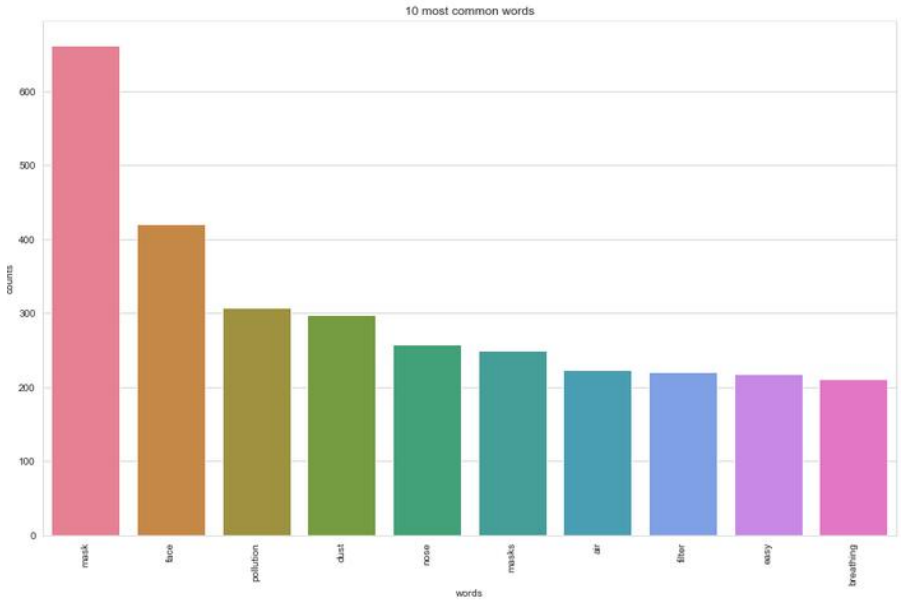
The most ten common words here are: 'mask', 'face', 'pollution', 'dust', 'nose', 'masks', 'air', 'filter', 'easy', and 'breathing'. We can see that these are the words most commonly used to describe face masks. Although there are some repetition between 'mask' and 'masks', it still reflects a functional topic around face masks. Based on the relevant topics, we can generate some topics by relevant circle of the topic.

Figure 3 stands for an auto-generated topic that summarizes the relevant topic under a topic relevancy circle. In this figure, we can generate some topics' range and develop a main idea or topic from the given words instead of topic number. We will describe more under the pyLDAvis output. pyLDAvis is a python package, which is a port of the R package by Sievert and Shirley (2015). pyLDAvis is designed to help users to interpret topic models which have been performed for text data. This package generates information and results from LDA topic models to form web-based visualization plots.

Figure 4 is a LDA visualization plot. This plot helps interactively understanding individual topics and relationships between topics. In order to understand the output, there are some points to be demonstrated:

First, LDA has it's own format. We need to organize the text format into the format required for LDA model. It has specific requirements for the format. A list whose length is equal to the number of the documents. Each element of documents is an integer with two rows. Each column of documents[[i]] (i.e., document i) represents a word occurring in the document. $documents[[i]][1, j]$ is a 0-indexed word identifier for the jth word in document i, which should be an index − 1 into vocab. $documents[[i]][2, j]$ is an integer specifying the number

Figure 2: 10 Frequent Words Figure.

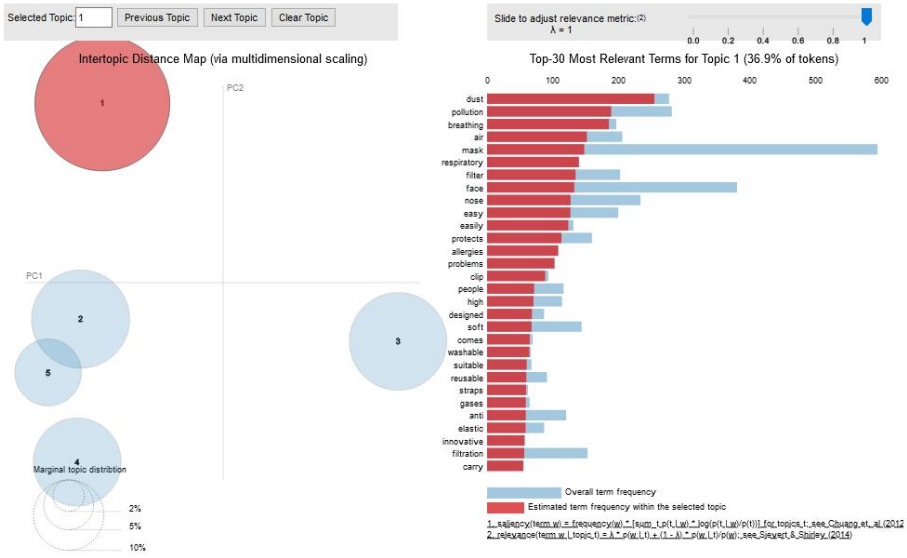


it generated ten frequent words

Figure 3: Summarized Topic Figure.

```
Topics found via LDA:

Topic #0:
gloves free latex n95 quality comfort nose anti protection used

Topic #1:
dust pollution breathing air mask respiratory filter face nose easy

Topic #2:
mask face pollution n95 disposable protection comfortable anti respirator ply

Topic #3:
masks surgical mask face covers medical particulate provides nose physical

Topic #4:
face masks germs day efficiency 99 dental wearer filtration easy
```

Figure 4: PyLDAvis Plot Figure.



The figure is generated by PyLDAvis Package under Python

of times that words appear in the document. The first line is the ID of words, and the second line is the frequency of words in documents, in order to generate a desired structure.

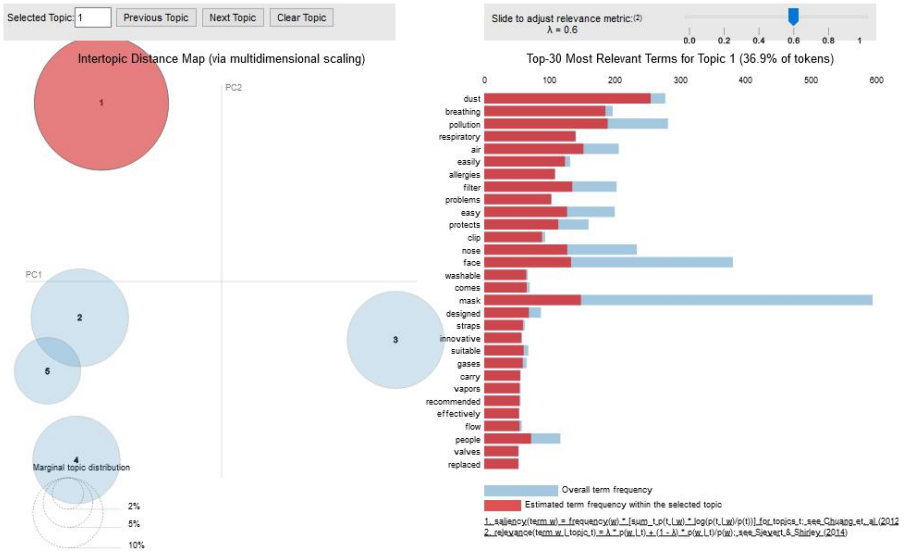
Subsequently, we have two prior parameters α and β to be specified in the model. For α, we can vary the size. The result of increasing α will lead to each document closer to the same topic. It describes that the role of $p(w_i|topic_i)$ is small, so that the role of $p(d_i|topic_i)$ is large. For , the result of increasing is that the role of $p(di|topici)$ is reduced, and the role of $p(w_i|topic_i)$ is increased, indicating that each topic is more concentrated on a few words. In another words, it means that every word is transferred to a topic as much as possible. Depending on the size of the text and the number of iterations, the time will vary, ranging from dozens of minutes to one or two minutes.

Furthermore, when a web-based interactive topic model visual display is completed, λ will be a parameter that controls whether a topic is a specific topic or general topic. pyLDAvis provides the following relationship:

$$relevance(term_w|topic_t) = \lambda * p(w|t) + (1 - \lambda) * p(w|t)/p(w).$$

The relevance of a word topic is adjusted by the parameter λ. If λ is close to 1, then the words that appear more frequently under the topic are more related

Figure 5: Topic 1 Figure.



to the topic; if λ is closer to 0, then the more special and exclusive words under the topic are more related to the topic. Thus, readers can change the relevance of words and topics by adjusting the size of λ. As for the best value of λ, readers can validate over different sizes.

After running topic modeling, we can describe the frequency of each topic. pyLDAvis uses the size of a circle to represent this number, so the size of the bubble indicates the frequency of the theme. Then, pyLDAvis uses multi-dimensional scale analysis to extract the principal components as dimensions, and distributes the topics to these two dimensions. The distance between the topics indicates the proximity between the topics.

Moreover, the plot only summarizes 30 most relevant terms in each topic and it will not create any specific topic. However, it will analyze the relevancy of each word and put relevant words together under one topic. We can make conclusion for the topic using those 30 words.

3.3 Results

Before we conclude from the plot, we need to decide what is the λ value we use to analyze the plots. When topics are too general, it stands for low value of λ. In this case, we cannot observe the features under a great position. Thus, we prefer a higher λ value. This will develop the description features much better. Moreover, the product description has already had a lot of breath words for

Figure 6: Topic 2 Figure.

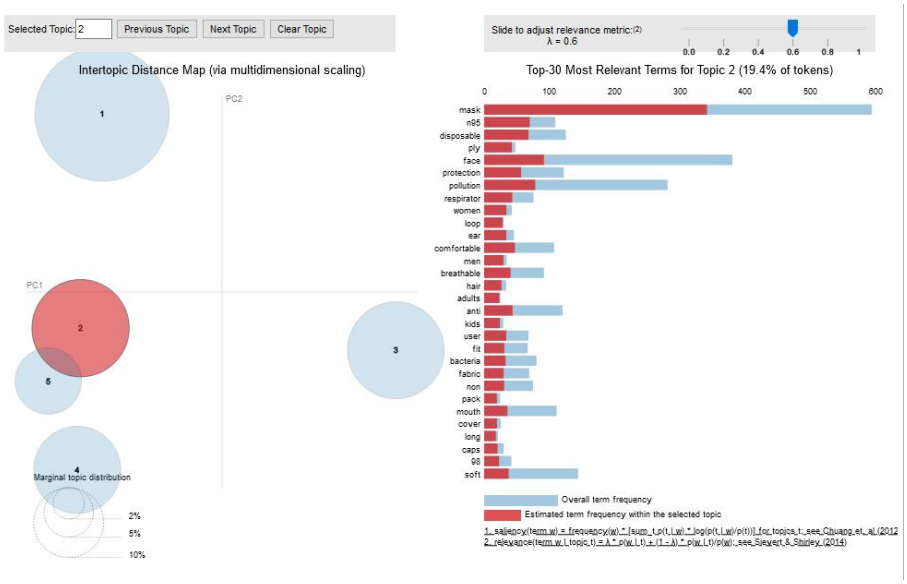


Figure 7: Topic 3 Figure.

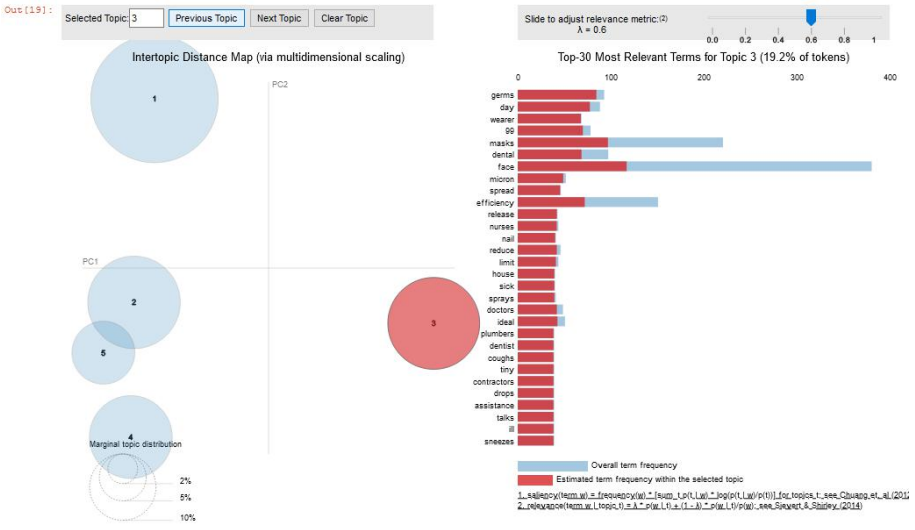


Figure 8: Topic 4 Figure.

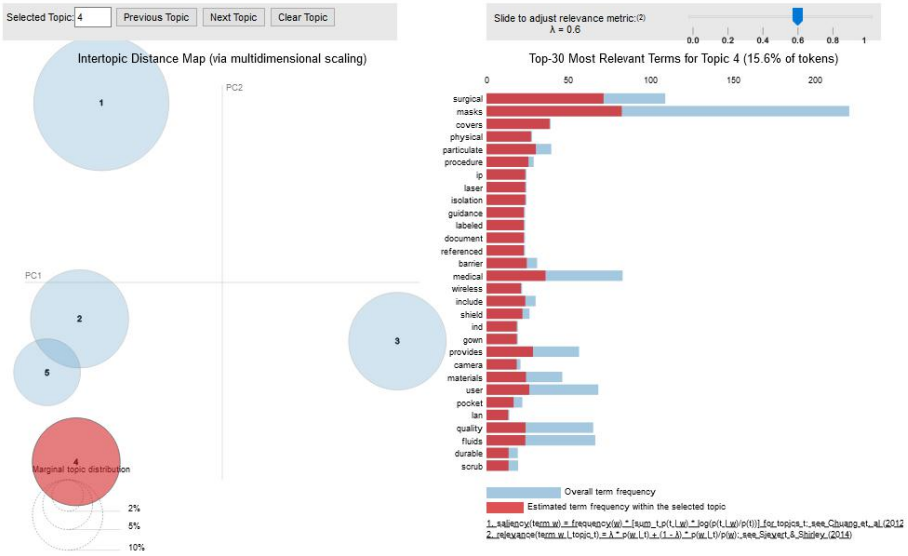
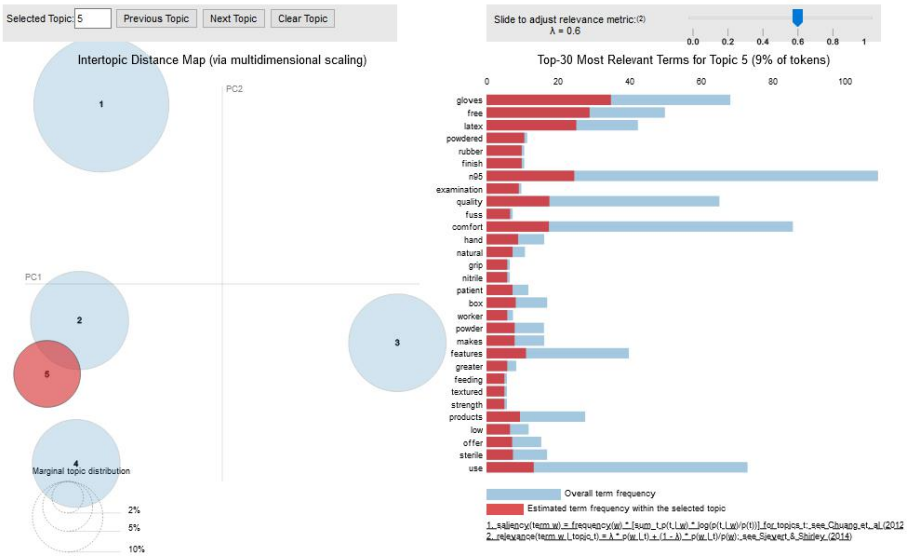


Figure 9: Topic 5 Figure.



surgical product, so features for the product show more importance. Relatively, when λ is 0.6, it shows most of features of the data, and it will not suffer the problem that topics are too narrow to precise terms using in surgical product. N95 is a significant term we focus on. When λ is 0.6, N95 could show in the second place in topic 2. The reason why we focus on topic 2 is that topic 1 is too general. It generally summarizes an idea about surgical products. For topic 2, we can see the topic is about the disposable N95 which is a specific mask type. Thus, the term of N95 could be a detailed feature of interest. It stands for a very specific type under face masks. In addition, N95 masks are frequently used in preventing Coronavirus and has the best protection. By analyzing this indicator, we select a relatively fitted λ value which is 0.6.

Figure 5 represents topic 1 under the pyLDAvis plot. We can see that topic 1 indicates the function of masks using. It offers some general terms using in surgical products including 'dust', 'breathing', 'pollution', etc.

Figure 6 represents topic 2 under the pyLDAvis plot. It indicates that topic 2 is about mean features of surgical masks where N95 and disposable are main features. It offers some functional terms and types of masks including 'N95', 'disposable', 'ply', 'protection' and so on. We could conclude that disposable N95 mask is the largest unfair pricing product.

Figure 7 represents topic 3 under the pyLDAvis plot. We can see that topic 3 shows where the masks are from and some specific terms in medical treatment.

Figure 8 represents topic 4 under the pyLDAvis plot. This plot shows that topic 4 summarizes some possible types of surgical masks.

Figure 9 represents topic 5 under the pyLDAvis plot. We can describe that topic 5 is mainly about the material of the masks. It also has some relevancy with topic 2, since the relevancy circles between the two topics have certain overlaps.

Based on the fact that N95 is the mask type with the best effect on preventing coronavirus disease, the results shown above are meaningful as we can see that the disposable N95 is summarized by topic 2 in a prominent place and topic 2 is the most meaningful/important topic summarizing features and unfair pricing products, particularly masks. These results are consistent with the market common observations. We thus conclude that based on the LDA topic models we fit, disposable N95 seems to represent the most unfair pricing product under the predatory pricing online data.

4. Discussions & Further Improvements

The graphs and results we generated in the previous section can be used to summarize what kind of masks is in unfair pricing. When we focus on concluding the mask type, the topic model suggests that N95 is the most unfair pricing product. LDA can feasibly analyze topics from corpus text data as we conduct the unfair pricing model. Some further research development might be needed for the modeling produce to provide more detailed or meaningful topics. For instance, topic models based on RNN (give full name and some references) techniques might help enhance the performance of LDA by increasing short term memory around some local sentences and documents. In that direction, we might combine RNN with LDA topic models. Such techniques have demonstrated efficiency in the field of document analysis (give references) because RNN can help to grasp local structures of the document, while topic models are more focusing on global structures.

References

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Editor: John Lafferty. *Latent Dirichlet Allocation*. Journal of Machine Learning Research (2003) 3, 993-1022. Retrieved from: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

[2] Shashank Kapadia *How to get started with topic modeling using LDA in Python*. Topic Modeling in Python: Latent Dirichlet Allocation (LDA) Apr 14, 2019. Retrieved from: https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0

[3] Adji B. Dieng, Chong Wang, Jianfeng Gao, John Paisley. *TOPICRNN*, a Recurrent Neural Network With Long-Range Semantic Dependency. Retrieved from: https://openreview.net/pdf?id=rJbbOLcex

[4] Predatory Pricing Online Data from Amazon. *Dataset In Datahut*. Retrieved from: https://data.world/data-hut/predatory-pricing-data-from-amazon

[5] Tony Paul *COVID-19 And Predatory Pricing Online*. Mar 18, 2020 Amazon Scarping, Big Data, E-Commerce, Pricing Strategy, Uncategorized. Retrieved from: https://blog.datahut.co/covid-19-and-predatory-pricing-online/

[6] Shinichi Nakajima, Issei Sato, Masashi Sugiyama, Hiroko Kobayashi and Kazuho Watanabe. *Latent Dirichlet Allocation*. Analysis of Variational Bayesian Latent Dirichlet Allocation: Weaker Sparsity than MAP Retrieved from: https://papers.nips.cc/paper/2014/file/5487315b1286f907165907aa8fc96619-Paper.pdf

[7] Matthew D. Hoffman, David M. Blei and Francis Bach. *Online Learning for Latent Dirichlet Allocation*. (2010). In advances in neural information processing systems, 856-864.

[8] wordcloud 1.8.1 Released: Nov 11, 2020 *wordcloud package*. Retrieved from:https://pypi.org/project/wordcloud/

[9] Ben Mabey. *pyLDAvis package*, Copyright 2015. This is a port of the R package by Carson Sievert and Kenny Shirley. Retrieved from: https://pyldavis.readthedocs.io/en/latest/readme.html

[10] akashii99, Git stats *Topic Modelling* Implementation of Blei, Andrew Ng, Jordan paper of LDA, Apr 16, 2019. Retrieved from:https://github.com/akashii99/Topic-Modelling-with-Latent-Dirichlet-Allocation