**Bi-weekly Report 2**

-Introduction To Brain and Machine Learning-

**Group 2**

**Members and their Role**

Shareen Rai -  Making a report

김동환 - Data analysis, preprocessing and model constructing

서범석 - Reference paper survey and model learning

송윤아 - Define the question and select a topic and summarize the result

**Problem definition**

Due to the pandemic, the vacation in the form of "Staycation" has increased worldwide [1], hotels are now not only just accommodations but also being served as a travel destination. As such, we expect consumers to consider a variety of factors in the process of deciding on a hotel. We thought that hotels needed the means to examine what factors satisfy consumers, and also consumers needed the means to help select the rational hotel.

In general, when people choose a hotel, they refer to the hotel review. However, as the number of review-set is way too large, and every reviewer is interested in each different factor. So the evaluation of those reviews is mandatory for every individual who wants to get information from them. To solve this problem, we use machine learning to make this easier for those who have to select the hotel to spend memorable days.

텍스트, 사람, 스크린샷, 실내이(가) 표시된 사진

자동 생성된 설명

**Fig 1 : Article about “Staycation” [1]**

**Topic**

Using machine learning, will examine in detail the factors that hotel consumers think positively and negatively.

**Model**

To solve this problem, we use the LDA (Latent Dirichlet Allocation) model for the topic. Because LDA technique does not have to know in advance what the topics will look like. And we can explore topic formation and resulting document clusters by tuning the LDA parameters to fit different dataset shapes [2].

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 in a corpus D [3]:

1. Choose N ∼ Poisson(ξ).

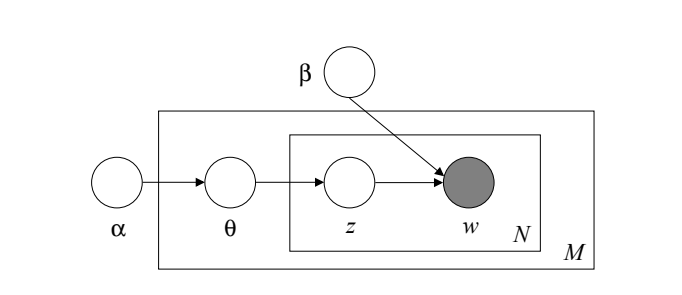
2. Choose θ ∼ Dir(α).

3. For each of the N words :

(a) Choose a topic

(b) Choose a word from , a multinomial probability conditioned on the topic .

‘sklearn.decomposition.LatentDirichletAllocation’ module of Scikit-learn is a very good tool for implementing the LDA model as Python [4]. This module receives the number of topics and Term Frequency-Inverse Document Frequency (TF-IDF) of a document and finds the topic distribution of a given document and the frequency value of words used in the topic.



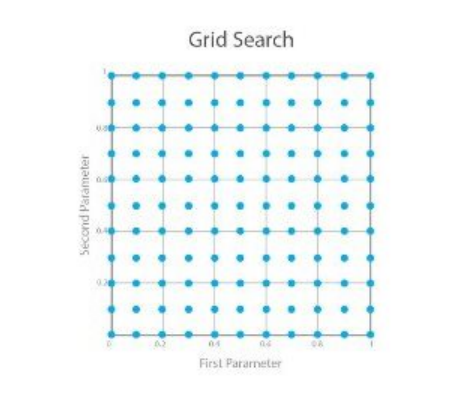
**Fig 2 : Graphical model representation of LDA [3]**

**Setting parameter values**

Several search methods such as random search and grid search or genetic algorithm can be used to find the optimal parameter value during model learning. Among them, grid search was used to find the most optimal result value in this project.

Grid search is a traditional method of hyperparameters optimization, which simply makes a complete search over a given subset of the hyperparameters space of the training algorithm [5]. This process takes a long time, but it's the surest way to find the optimal parameter values.

‘sklearn.model\_selection.GridSearchCV’ module of Scikit-learn is a very good tool for implementing search methods as Python [6]. This module receives various parameters as param\_grid and compares the result values of substituting them with stratified k-fold cross-validation to create optimal parameters and models made of those parameters.



**Fig 3 : An illustration of a grid search space [5]**

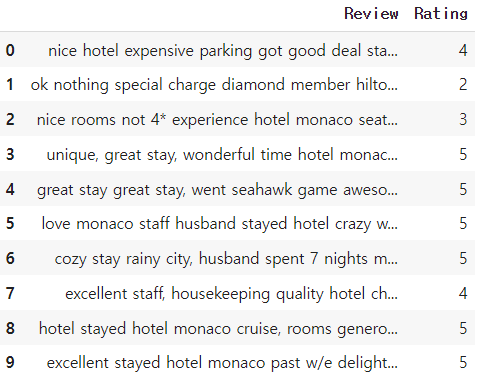
**Data set**

We use data named ‘Tripadvisor Hotel Revisions’ in Kaggle. This data is from a site called ‘TripAdvisor’ that crawled over 20,000 reviews [7]. And we classified them as positive/negative reactions using the number of stars and reviews in the data set.

To see the use of specific words for each data, we created a word cloud and checked. At this time, ‘hotel’, ’everything’, ’anything’, ‘nothing’, ’thing’, ‘need’, ‘stay’, ‘say’, ’go’, ‘day’, ‘night’, ‘time’ were additionally excluded, in terms of selecting main factor that affects the evaluation.

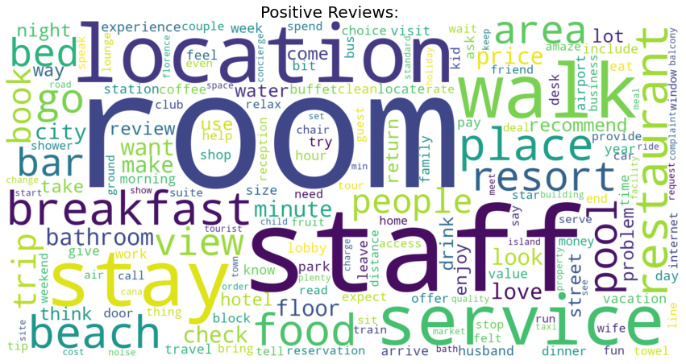
**Data Exploration**

* **Followed Pictures are the output of our preprocessing units**

전자기기이(가) 표시된 사진

자동 생성된 설명

**Fig 4 : Format of Data Fig 5 : The ratio of positive and negative**

텍스트이(가) 표시된 사진

자동 생성된 설명

**Fig 6 : Word cloud of positive review Fig 7 : Word cloud of negative review**

**Training Model**

Step1. Import the module that we’ll use.

텍스트이(가) 표시된 사진

자동 생성된 설명

Step2. Make the ‘LDA\_model’ function.

텍스트이(가) 표시된 사진

자동 생성된 설명

Step3. Find the best parameter in pos\_lem(data) and ned\_lem(data).

텍스트이(가) 표시된 사진

자동 생성된 설명

텍스트이(가) 표시된 사진

자동 생성된 설명

Step4. Make ‘display\_topics’ function.

텍스트이(가) 표시된 사진

자동 생성된 설명

Step5. Print the result for each positive and negative.

텍스트이(가) 표시된 사진

자동 생성된 설명









**Result**

We printed out the words that had the most influence from 1st to 20th. It was confirmed that both positive and negative ranked first and second. The biggest factors that people feel positive and negative are room and staff. It was found that the result was similar to the word cloud result output from data expansion. For the hotel to receive a good review from the users, it will have to be equipped with rooms and staff. To get a positive response, it is necessary to pay attention to location and service. As services and food are at the top of the factors that affect negative reactions, we think it is necessary to consider those factors to reduce negative reactions.

|  |  |
| --- | --- |
|  |  |
| Result1. Positive Result | Result2. Negative Result |



**Fig 8. Result**

**Reference**

[1] <https://www.forbes.com/advisor/travel-rewards/2021-is-the-year-to-book-a-holiday-staycation-heres-why/>

[2] Nguyen, Eric. “Text Mining and Network Analysis of Digital Libraries in R.” (2014).

[3] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *the Journal of machine Learning research* 3 (2003): 993-1022.

[4] <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html>

[5] Liashchynskyi, Petro, and Pavlo Liashchynskyi. "Grid search, random search, genetic algorithm: a big comparison for NAS." *arXiv preprint arXiv:1912.06059* (2019).

[6] <https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html>

[7] <https://www.kaggle.com/andrewmvd/trip-advisor-hotel-reviews>