**Bi-weekly Report 1**

-Introduction To Brain and Machine Learning-

**Group 2**

**Members and their Role**

Shareen Rai - Making a report

김동환 - Data analysis and preprocessing

서범석 - Reference paper survey

송윤아 - Define the question and select a topic

**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]**

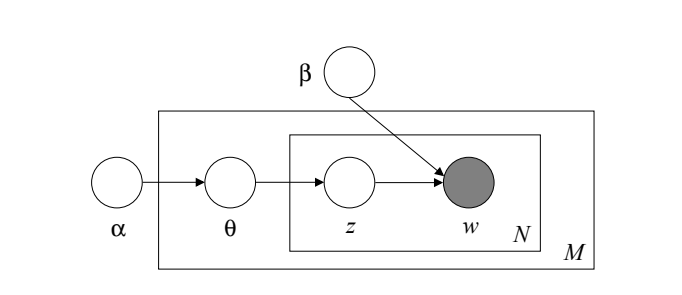
**Topic**

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 [3].



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

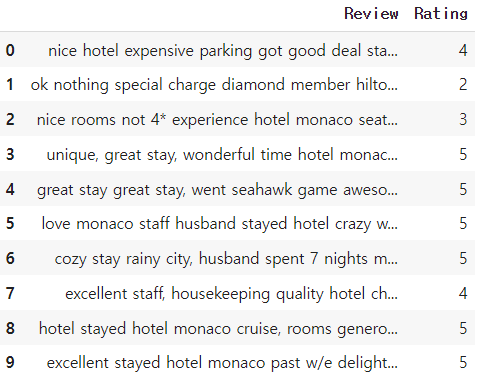
**Data set**

We use data named ‘Tripadvisor Hotel Revisions’ in Kaggle. This data is from a site called ‘TripAdvisor’ that crawled over 20,000 reviews[4]. 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.

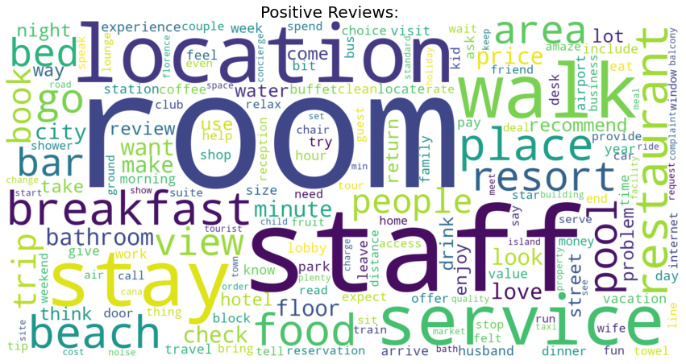
**Data Exploration**

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

전자기기이(가) 표시된 사진

자동 생성된 설명

**Figure 3 : Format of Data Figure 4 : The ratio of positive and negative**

텍스트이(가) 표시된 사진

자동 생성된 설명

**Figure 5 : Word cloud of positive review Figure 6 : Word cloud of negative review**

**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://www.kaggle.com/andrewmvd/trip-advisor-hotel-reviews>