

11조 기말 프로젝트 보고서 

* 호텔 리뷰 데이터를 활용한 자연어 처리 및 모델-

|  |  |
| --- | --- |
| **과목명** | **딥러닝 분석** |
| **교수명** | **정원일 교수님** |
| **학 과** | **정보통계 보험수리학과** |
| **팀 원** | **조성찬(20110726)**  **황병호(20170845)**  **김시헌(20180728)** |
| **제출일** | **2023. 06. 11.** |

**목 차**

[**1. 개요 3**](#_heading=h.txwlcr1tho2b)

[**2. 데이터 탐색 / 데이터 전처리 4**](#_heading=h.k4yk74p83p9i)

[**2.1 데이터 확인 및 시각화 4**](#_heading=h.gjdgxs)

[**2.2 이상값 정제 및 결측값 처리 6**](#_heading=h.30j0zll)

[**3. 모델 생성 8**](#_heading=h.1fob9te)

[**3.1**](#_heading=h.3znysh7) **자연어 처리와 RNN 8**

[**3.2**](#_heading=h.2et92p0) **다양한 RNN 적용 방법 [10](#_heading=h.2et92p0)**

[**3.3**](#_heading=h.tyjcwt) **모델 생성 및 적합 [11](#_heading=h.tyjcwt)**

[**3.4**](#_heading=h.3dy6vkm) **모델 평가 15**

[**3.5 성능비교 18**](#_heading=h.2s8eyo1)

[**4. 호텔 리뷰 단어 분석**](#_heading=h.cybpobw7nzpx)

**5. 추가 분석**

**6. 결론** [**20**](#_heading=h.cybpobw7nzpx)

# 1. 개요

최근 텍스트 데이터를 이용한 다양한 자연어 처리 서비스가 개발되면서, 인터넷을 통한 소비자 리뷰의 중요성이 더욱 부각되고 있다. 호텔 업계에서도 소비자 의견을 체계적으로 분석하고 경영 전략에 효과적으로 반영하는 것이 핵심 전략으로 떠오르고 있다.

본 보고서는 호텔 리뷰 데이터를 활용하여 자연어 처리 및 모델 비교에 대한 연구를 다루고 있다. 약 2만 건의 테스트 데이터와 14만 건의 데이터를 사용하여 분석을 수행하였다. 주요 목표는 다양한 통계적 방법을 활용하여 텍스트 데이터를 분석하고, 모델 간의 성능 차이를 비교한다.

자연어 처리 과정에서는 텍스트 데이터의 전처리, 특징 추출, 모델 학습 순서로 진행한다. 1점부터 5점까지의 평가 점수를 가지고 있는 텍스트 데이터를 바탕으로 자연어 처리를 진행한다. 전처리 과정에서는 정규화, 토큰화, 불용어 처리, 표제어 추출 등을 통해 텍스트 데이터를 정제하였으며, TF-IDF와 같은 특징 추출 방법을 사용하여 데이터를 벡터 형태로 변환한다.

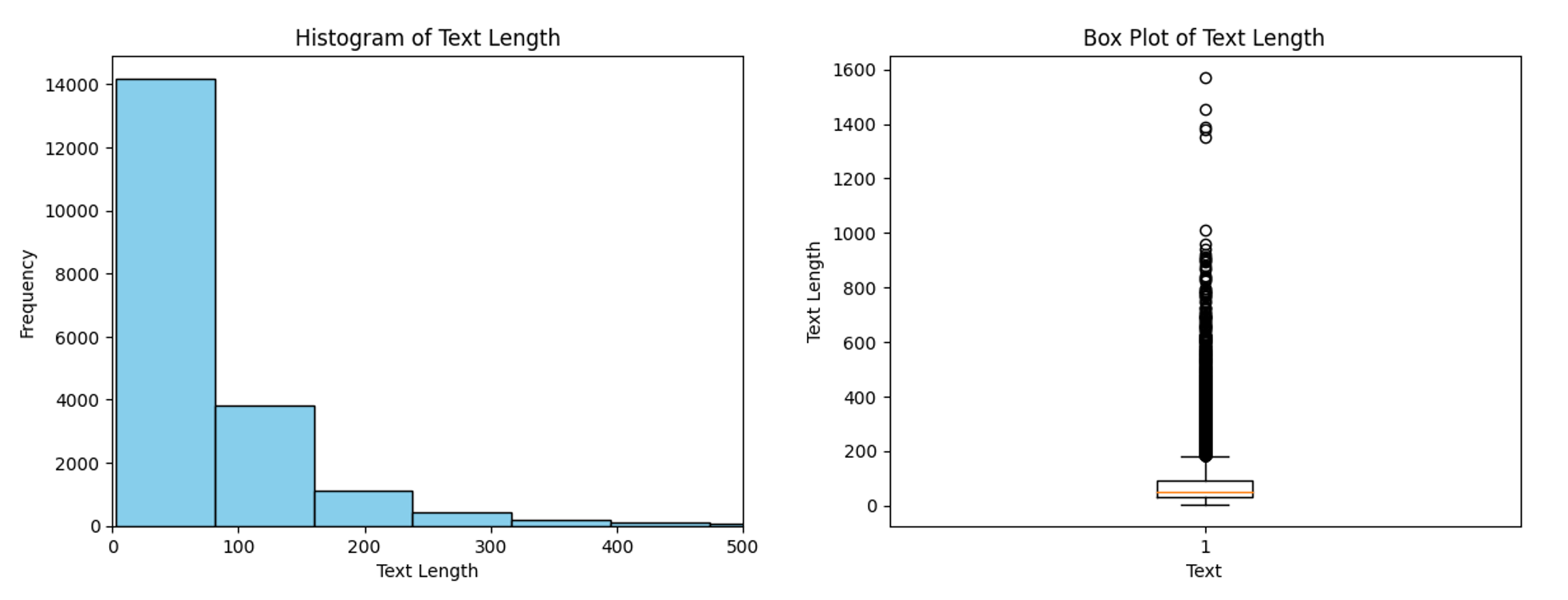
그 후, 딥러닝 모델인 순환신경망의 한 종류인LSTM를 활용하여 모델을 학습시킨 후 성능을 비교하였다. 뿐만 아니라 단어의 빈도수를 시각화 하고, 이를 활용해 성능 향상을 기대할 수 있도록 데이터를 가공하였다. 비지도 학습 방법 중 하나인 K-means 클러스터링을 이용하여 리뷰 데이터를 클러스터로 그룹화하고, 이를 시각화 하여 클러스터간 특징들을 추출한다.

# 2. 데이터 전처리

# 데이터 확인

데이터 셋의 분포를 확인하여 간단하게 EDA를 진행한다. 모델 학습에 사용할 텍스트 데이터가 어떤 형태로 이루어져 있는지 확인하기 위해 각 텍스트가 가지고 있는 단어 개수를 파악한다.

약 2만개의 텍스트 데이터 중에 대부분의 데이터는 100개 이하의 단어로 이루어져 있다. 일부 텍스트의 경우 500개 이상의 단어로 구성되어 있지만 그 빈도수는 상대적으로 작아, 이후 처리 과정에서 제외시킨다. (그림 1)

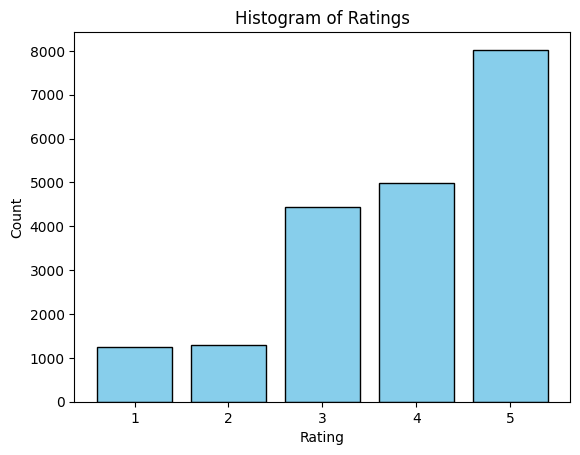


**<그림 1 단어 길이 별 텍스트 데이터 셋 분포>**

데이터 셋의 반응 변수(response variable)로 사용하는 평가 점수는 1점부터 5점까지 총 5개의 클래스에 대한 정수로 구성이 되어 있다. 5점에 가까울수록 좋은 평가를 남긴 리뷰이며, 1점에 가까울수록 나쁜 평가를 한 리뷰를 뜻한다.

대체로 긍정적인 평가를 한 리뷰가 가장 많으며 1점과 2점을 평가한 점수가 전체 데이터 셋에 12% 정도 수준만 차지하고 있어 다른 점수를 메긴 리뷰에 비해 상대적으로 적은 것을 확인할 수 있다. (그림 2)

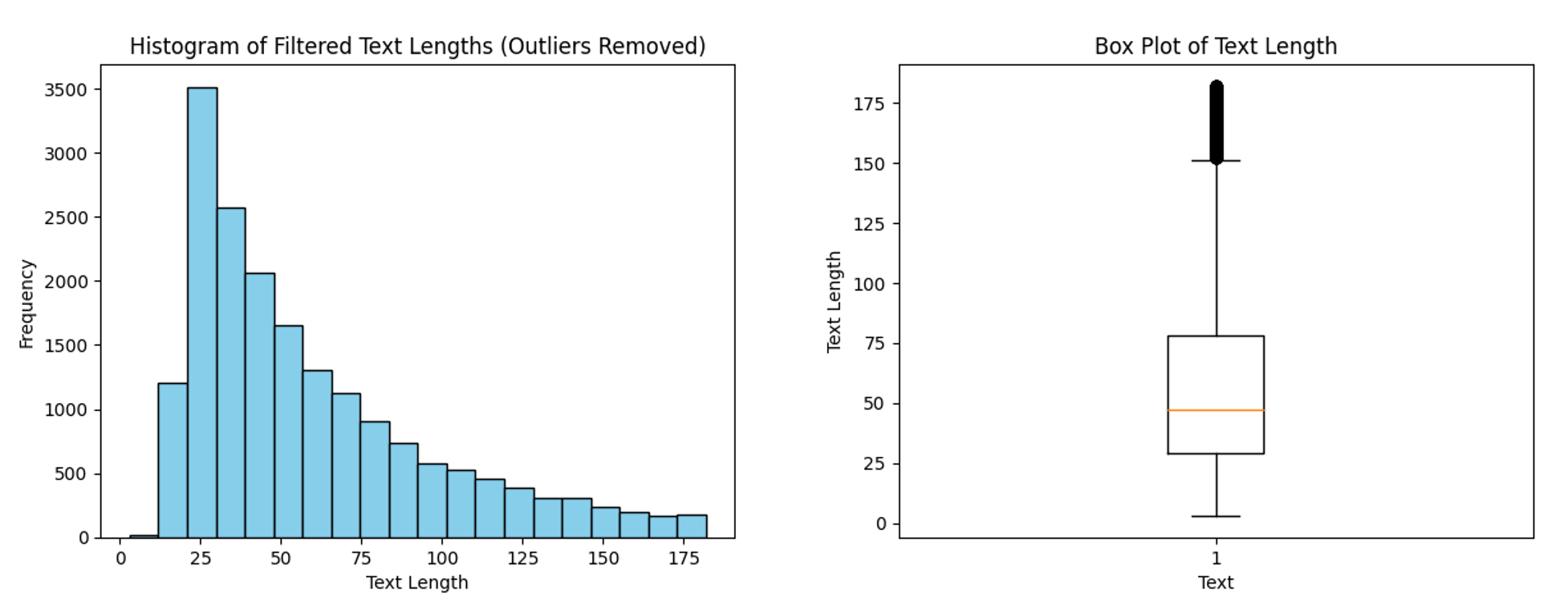
이후 텍스트 데이터를 일부 가공하고, 기계가 이해할 수 있는 숫자 형태의 데이터로 변환하기 위한 과정을 진행한다. 이 과정은 자연어 처리(Natural Language processing, 이하 NLP)라 말하며, 텍스트 데이터가 가지고 있는 복잡한 구조를 단순한 벡터 형태로 가공하는 과정도 포함되어 있다.



**<그림 2 평가 점수 별 데이터 셋 분포>**

# 텍스트 데이터 전처리(Natural Language Processing, NLP)

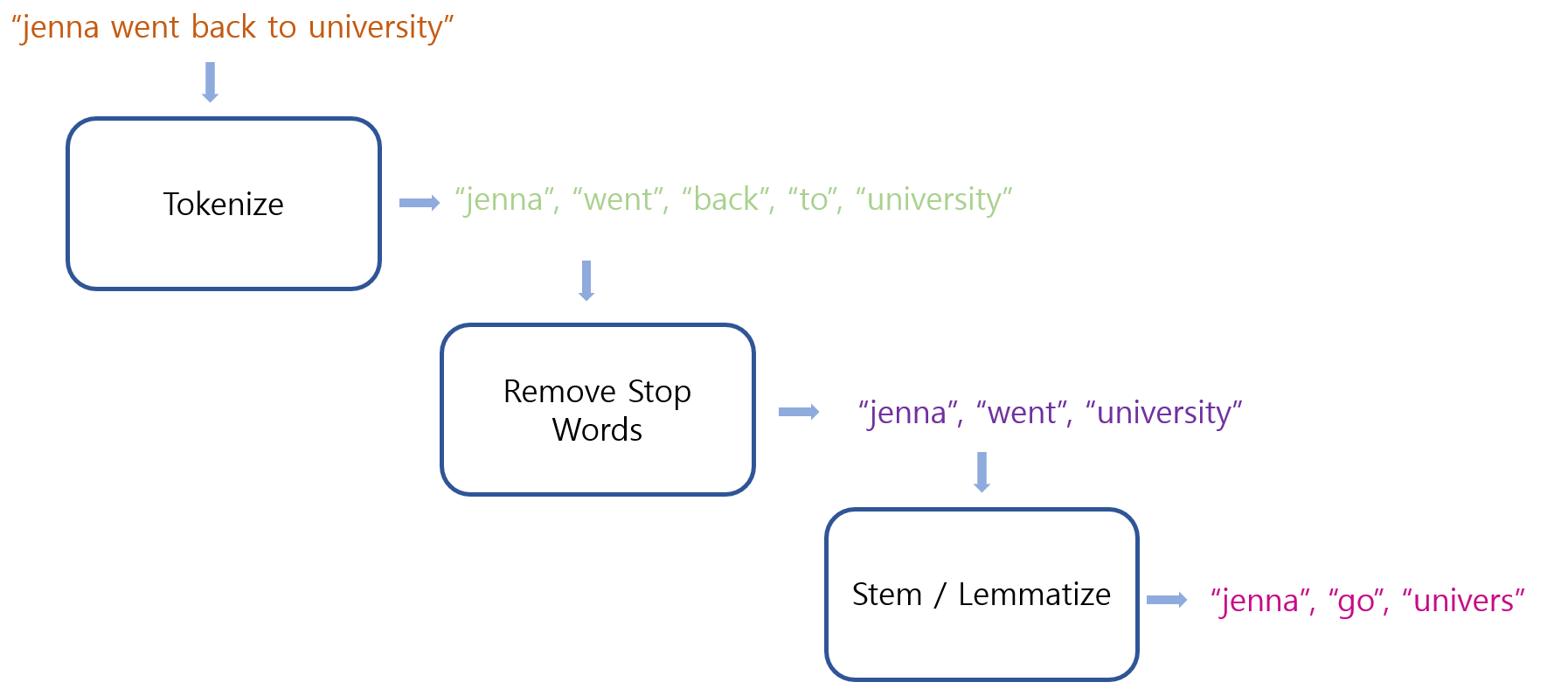
NLP 과정에 앞서서 앞서 확인한 데이터의 이상치 제거를 수행한다. 이전 과정에서 확인했던 텍스트 별 문장의 길이를 기준으로 하여, 특별히 긴 단어를 포함하고 있는 리뷰를 제외시킨다. 박스 플랏의 사분위수와IQR을 사용하여 182개 이상의 길이의 데이터를 제외하여, 약 1.8만개의 데이터를 사용한다. (그림 3)



**<그림 3 이상치 제거 후 단어 길이 별 텍스트 분포>**

이후 NLP 과정은 아래와 같은 순서로 진행이 된다. (그림4)

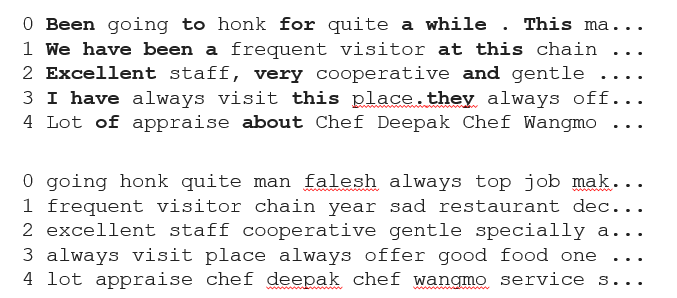
* **토큰화**
* **불용어 제거 및 표제어 추출**
* **시퀀스 생성 및 패딩**

****

**<그림 4 토큰화, 불용어 제거, 표제어 추출 과정>**

텍스트 전처리 과정 중 가장 기본이 되는 토큰화를 통해 텍스트 데이터를 단어로 분할하여 나눈다. 이 과정에서 정규표현식을 사용해서 이번 분석에서 제외할 특수 문자나 마침표 등을 제외하고, 나머지 단어들을 분할하여 리스트 형태로 저장한다.

이후 문장에서 중요하지 않으며, 자주 등장하는 단어인 'a', 'the', 'is'와 같은 관사나 접속사는 문장의 의미를 해칠 수 있기 때문에 제거한다. nltk 라이브러리의 Stopwords를 사용하여 쉽게 제거할 수 있으며, 이후 마찬가지로 nltk 라이브러리를 사용하여 표제어를 추출한다. (그림 5)



**<그림 5 자연어 처리 전후 비교, 처리 전(상), 처리 후(하)>**

이후 모든 텍스트 데이터의 단어를 사용해 단어 사전을 만들고, 이를 사용해 시퀀스 생성한다. 시퀀스는 단어 사전에서 인덱스를 추출해 텍스트 데이터를 정수 형태로 가공된 데이터이다. One-hot Encoding에 비해 단어의 등장 순서를 유지시킨다는 장점이 있다.

시퀀스를 생성함과 동시에, 가장 긴 문장을 기준으로 데이터의 길이를 맞추는 패딩 과정을 진행하여, NLP과정을 마친다. 이외에도 TF-IDF이나 Word2Vec 와 같은 다양한 NLP 모델이 있지만, 초기 모델 생성 과정에서는 제외시켰으며, Tensorflow가 제공하는 Embedding 레이어를 사용했다.

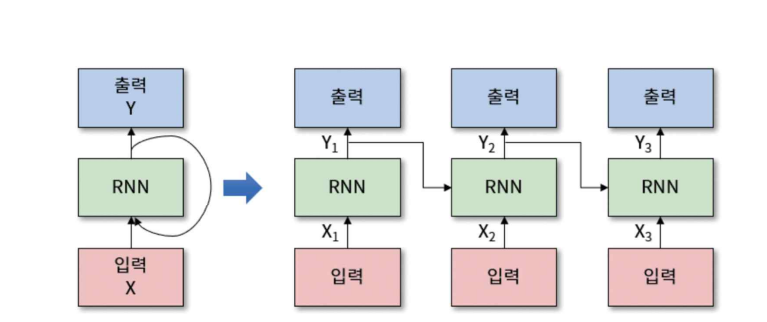
# 

# 3. 모델 생성

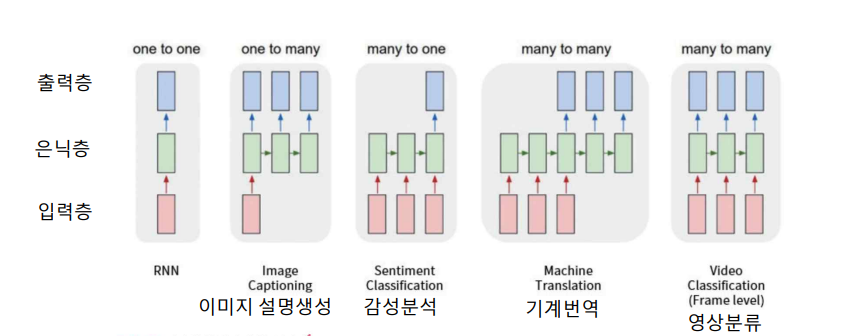
# 3.1 자연어 처리와 RNN

호텔 리뷰의 분석을 위해서는 자연어 처리 과정이 필수적이다. 자연어 처리는 컴퓨터가 인간의 언어를 이해하고 조작할 수 있도록 해주는 인공지능의 분야이다. 또한, 구조화되지 않은 텍스트 기반 데이터로부터 유용한 정보를 얻기 위해 사용되며, 사용자가 추출된 정보에 액세스하여 해당 데이터에 대해 새로운 이해를 생성할 수 있게 해준다.

이런 자연어 처리는 순환 신경망인 RNN을 통해 처리할 수 있다. 자연어 처리는 입력의 순서가 중요한데, RNN은 기존 신경망과 달리 이전 출력값이 현재 결과에 미치는 ‘되먹임’ 구조를 가지기 때문에 용이하다. 또한, 입력과 출력의 길이 제한이 없기 때문에 다양한 형태의 네트워크를 만들 수 있는 강점이 있다.



**<순환 신경망 구조 풀이>**



**<다양한 형태의 순환 신경망>**

# 3.2 다양한 RNN 적용 방법

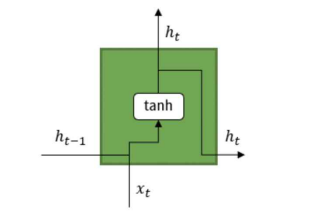
RNN을 이용하여 분석을 진행하기 위해, 다음과 같은 레이어를 이용하였다.

1. **SimpleRNN**
2. **GRU**
3. **LSTM**

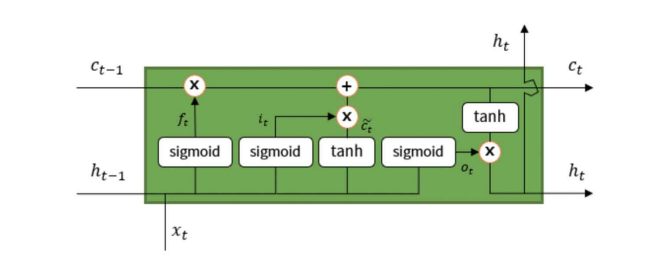
simpleRNN은 가장 간단한 형태의 RNN 레이어로, 변화하는 입력을 받기 때문에 위 그림에서는 각 단계에서 입력이 변할 때의 흐름을 보여준다. 보통 활성화 함수로는 tanh나 ReLU를 사용한다. simpleRNN은 간단하지만, 장기 의존성 문제가 존재하며, 입력 데이터와 출력 데이터의 사이의 길이가 멀어질수록 연관 관계가 적어진다는 단점이 있다. 현재 예측값을 얻기 위해 과거의 정보에 의존하면서도, 과거 시점이 너무 멀어지면 문제를 해결하기 어렵다는 것이다.

1. LSTM

위의 simpleRNN의 단점을 해결하기 위한 구조로 LSTM을 고안했다. LSTM은 RNN의 특별한 종류로, 긴 의존 기간을 필요로 하는 학습을 수행할 능력을 가지고 있다. 단순한 neural network layer 한 층 대신, 4개의 레이어가 특별한 방식으로 서로 정보를 주고 받도록 되어있다. 따라서, simpleRNN에 비해 복잡한 구조를 가지고 있으며, 출력 외에 LSTM 셀 사이에만 공유되는 셀 상태(cell state)를 가지고 있는 레이어이다. 밑의 그림에서 보면 SimpleRNN 셀에서 타임 스텝의 방향으로 ht가 전달되고 있는 반면, LSTM 셀에서는 셀 상태인 ct가 평행선을 그리며 함께 전달되는 모습을 볼 수 있다. 이처럼 타임 스텝을 가로지르며 셀 상태가 보존되기 때문에 장기 의존성 문제를 해결할 수 있다는 것이 LSTM의 핵심 아이디어이다.



<셀로 나타낸 simpleRNN의 흐름>



<셀로 나타낸 LSTM의 계산 흐름>

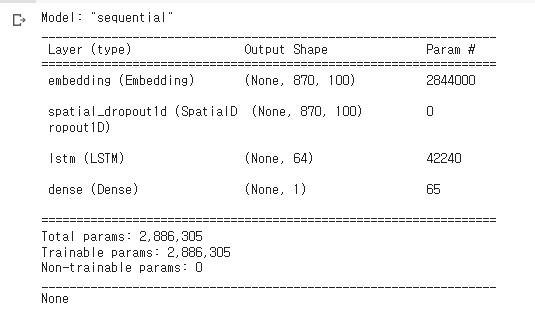
다음에 나와있는 식은 그림을 구성하는 주요 수식으로, 마지막의 두 줄은 셀 상태인 ct와 출력인 ht를 최종적으로 결정하는 부분이다.

3.3 모델 생성 및 적합

이제 본격적으로 데이터에 RNN 신경망을 이용하여 모델을 적합한다. RNN 신경망 중에서, SimpleRNN의 성능을 보완하고, 많은 데이터를 분석할 것으로 예상되어 LSTM을 사용하여 분석을 진행하기로 결정했다.

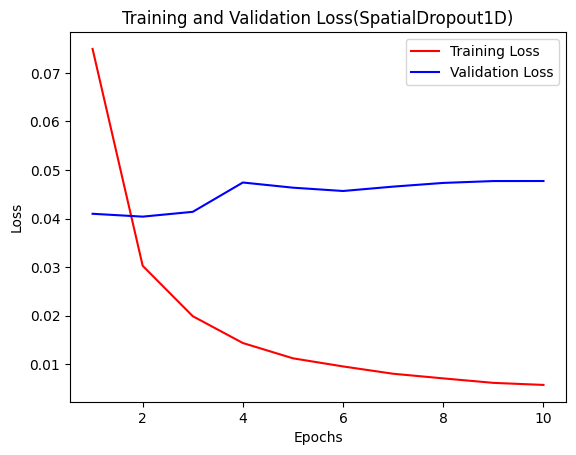
LSTM 모델을 구축하기 위해 먼저 임베딩 차원을 100으로 설정한다. 임베딩은 단어를 고정된 길이의 밀집 벡터로 표현하는 과정으로, 단어 간의 의미적 유사성을 고려하여 단어를 표현한다. 단어 인덱스의 길이에 1을 더하여 어휘 사전의 크기를 구하고, 이를 모델의 첫 번째 층인 Embedding 층에서 사용한다. Sequential 모델을 생성하여, 정수로 인코딩된 단어를 입력받아 임베딩 벡터로 변환하는 Embedding 층을 추가한다. 또한 SpatialDropout1D층을 추가하여 과적합을 방지한다.

다음엔 메인인 LSTM 층을 추가한다. 유닛의 수는 64로 설정하였다. Dropout 층으로 과적합을 줄이고, 선형 활성화 함수를 사용하였고, 하나의 출력 뉴런을 가지도록 설정하였다. 모델을 컴파일 할 때는 mse 함수를 사용하고 Adam optimizer를 사용하여 모델을 최적화 하였다.



다음과 같이 모델 훈련을 했다. LSTM 모델을 사용하여 훈련하였고, 훈련과정에서의 손실값 변화를 시각화 하였다. 에포크 수는 10번으로 설정하였고, 전체 데이터 셋을 32번 반복 훈련하게 설정하였다.

또한 그래프를 통해 훈련 손실 그래프와 검증 손실 그래프를 시각화 할 수 있게끔 하였다.

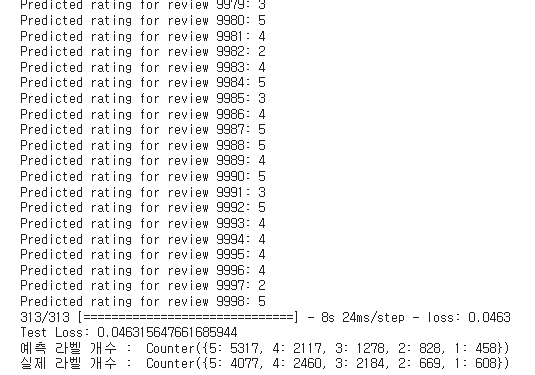


다음의 시각화 자료는 training loss가 줄고 있음을 보여주고 있다. validation loss 값은 줄어들고 있지는 않지만, 과적합 방지를 위한 드롭아웃을 추가해서 많이 늘어나지는 않음을 보여주고 있다.

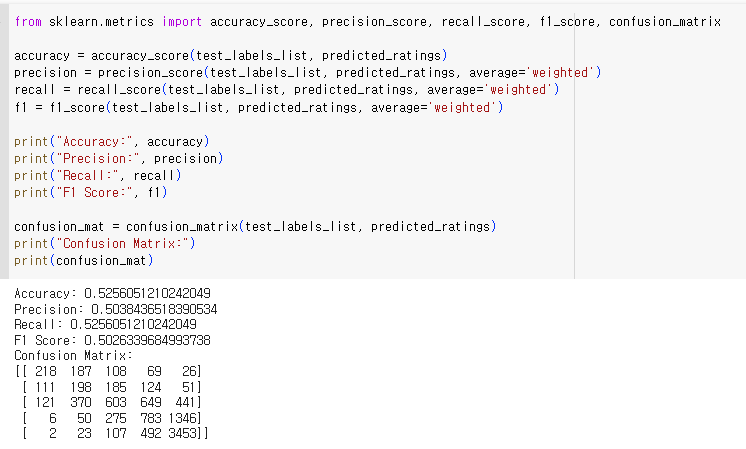
비교를 위해 드롭아웃 부분을 제외하고 훈련을 시켜본 결과 다음과 같은 그래프가 나왔다.

1. 모델 평가

다음으로 모델 평가를 진행하였다. 테스트 데이터에 대한 예측 결과를 실수값으로 반환한다. 그리고 convert\_data 함수는 predicted\_ratings 리스트의 각 요소를 받아서 해당 값을 정수로 코드에 나와있는 기준대로 변환한다. loss를 통해 테스트 데이터에 대한 모델의 손실 값을 평가하고, 예측 라벨 개수와 실제 라벨 개수를 출력한다.



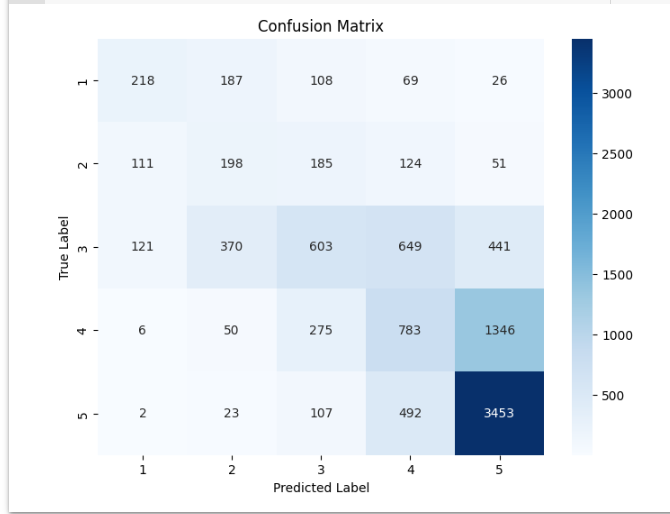
다음과 같은 결과가 출력되었고, 테스트 손실값은 0.04 정도가 나왔다. 예측과 실제 라벨의 개수를 비교해 볼 수 있다.



다음은 모델의 정확도 확인을 해봤다. sklearn.metrics 모듈을 사용하여 평가 지표를 계산하고, 혼동행렬을 출력했다. 정확도와 정밀도, 재현율, F1 스코어를 계산하였다. 그 결과 정확도는 약 52.56%가 나왔다.



다음은 혼동 행렬 가시화 하였다. 계산된 혼동행렬을 데이터 프레임으로 변환하고, index와 column 매게변수를 사용하여 데이터프레임의 인덱스와 컬럼을 클래스 라벨로 지정한다. 그리고 히트맵 시각화를 진행하였다.



다음과 같이 히트맵이 시각화 되었다. raiting 값이 5점일 때 5점으로 예측한 값이 가장 많이 나오는 결과를 보여주고 있다

# 4. 모델 평가

주어진 분석을 수행하기 위해 호텔의 리뷰인 텍스트 데이터를 이용하여 분석하는 방법과 rating값을 이용하여 분석하는 방법 2가지를 제시한다.

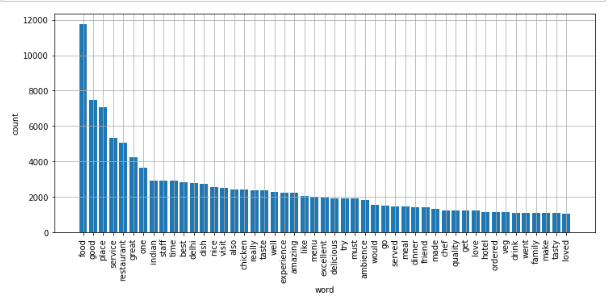
# 4.1 텍스트 데이터 자체를 이용한 분석

어떤 단어가 좋은 호텔 리뷰에 많이 쓰이고 어떤 단어가 나쁜 호텔 리뷰에 많이 쓰이는지 분석하기 위해 rating값이 1,2점은 부정적인 평가로, 4,5점은 긍정적인 평가로 기준을 세운다. rating값이 3점인 리뷰는 중립적인 평가로 간주하고 제외한 후 분석을 수행한다.

4.1.1 단어 빈도 분석 및 Word Cloud

TF-IDF를 활용하여 좋은 리뷰와 나쁜 리뷰 각각에서 단어의 중요도를 계산하고 TF-IDF점수가 높은 단어를 확인하여 차이를 분석하고자 한다. 또한 각 리뷰에서 워드 클라우드를 생성하여 차이를 시각적으로 확인한다.

-좋은 리뷰에서 단어 빈도 수 상위 50개의 단어 시각화



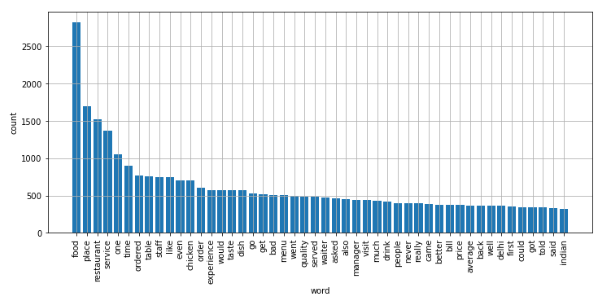
<그림 좋은 리뷰에서의 상위 50개 단어 히스토 그램>

좋은 호텔 리뷰에서 주로 "food", "good", "place", "service", "great" 등과 같은 단어들이 등장함을 알 수 있다. 구체적으로는, 각각 11771, 7488, 7063, 5354, 4235 횟수만큼 나타났다. 이를 통해 음식, 좋은 서비스, 멋진 장소 등이 좋은 리뷰에서 강조되는 것을 알 수 있다.



<그림 좋은 리뷰에서의 상위 50개 단어 워드 클라우드>

-나쁜 리뷰에서 단어 빈도 수 상위 50개의 단어 시각화



<그림 나쁜 리뷰에서의 상위 50개 단어 히스토그램>

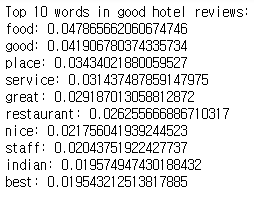
나쁜 호텔 리뷰에서 주로 "food", "bad", "place", "restaurant", "service" 등과 같은 단어들이 등장함을 알 수 있다. 구체적으로는, 각각 2825, 509, 1693, 1525, 1365 횟수만큼 나타났다. 이를 통해 음식, 나쁜 서비스, 나쁜 장소 등이 나쁜 리뷰에서 강조되는 것을 알 수 있다.



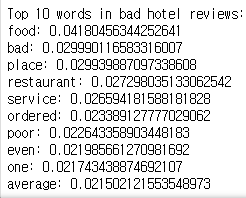
<그림 나쁜 리뷰에서의 상위 50개 단어 워드 클라우드>

4.1.2 TF-IDF를 활용한 단어 중요도 계산

단어 빈도 수와 워드 클라우드 시각화만을 통한 해석은 객관화가 부족하다고 생각하여 객관적인 지표인 TF-IDF를 활용하였다. TF-IDF는 단어의 빈도를 고려하여 계산되고 단어의 상대적인 빈도를 파악할 수 있다. 높은 점수를 가질수록 해당 문서에서 자주 나오는 것으로 식별할 수 있다. 따라서 좋은 호텔 리뷰와 나쁜 호텔 리뷰에서 각각 TF-IDF 점수 상위 10개의 단어들을 확인하였다. 결과는 4.1.1의 방법과 매우 유사하게 나옴을 확인할 수 있다.



<그림 좋은 호텔 리뷰에서 TF-IDF 점수가 높은 상위 10개 단어>



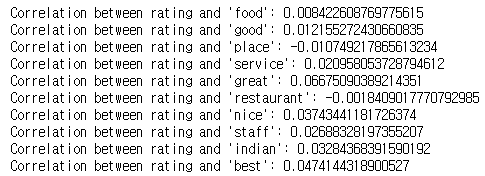
<그림 나쁜 호텔 리뷰에서 TF-IDF 점수가 높은 상위 10개 단어>

# 4.2 Rating값을 활용한 분석

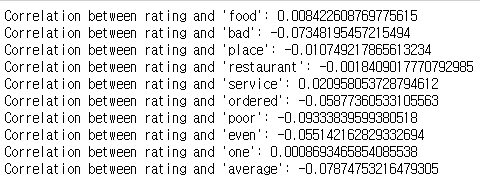
4.1은 텍스트 데이터 자체를 이용해서 분석한 반면 4.2는 rating값을 활용하여 분석하고자 한다.

호텔 리뷰의 rating값과 특정 단어들의 등장 빈도 혹은 중요도를 비교하여, 특정 단어들이 좋은 리뷰 혹은 나쁜 리뷰에서 더 자주 등장하는지 또한 rating값과 어떤 상관 관계를 가지는지 등을 분석한다.

방법은 TF-IDF를 활용하여 리뷰 텍스트를 벡터화한 후 벡터화 된 리뷰와 rating값을 결합한 데이터 프레임을 생성하여 4.1에서 도출한 각 리뷰의 상위 10개 특정 단어들과 rating값의 상관 관계를 계산한다.



<그림 좋은 호텔 리뷰 상위10 특정 단어의 등장 빈도 혹은 중요도와 rating 값의 상관 관계>



<그림 나쁜 호텔 리뷰 상위10 특정 단어의 등장 빈도 혹은 중요도와 rating 값의 상관 관계 계산>

[결론]

: 단어 빈도 수 및 WordCloud 분석뿐만 아니라 이외의 분석을 통해서도 좋은 호텔 리뷰에서는 음식, 좋은 서비스, 멋진 장소 등이 강조되는 경향을 보이는 반면, 나쁜 호텔 리뷰에서는 음식, 나쁜 서비스, 나쁜 장소 등이 강조되는 경향을 보인다. 이를 통해 음식, 서비스, 장소, 레스토랑 등과 같은 요소들이 호텔 평가에 많은 영향을 미치는 것으로 해석할 수 있다.

팀원 별 역할

공통 : 보고서 작성

조성찬 :

황병호 :

김시헌 :