Term Project 제안서

|  |  |
| --- | --- |
| Department | Computer Science |
| Course title | Advanced Skills in Machine Learning |
| Instructor | Prof. Sung Soo Lim |
| Student ID | M2018075 |
| Student Name | 정성민 |
| Due date | 18.05.08 |

**목 차**

[1. 데이터셋 (Dataset) 3](#_Toc514344856)

[2. 프로젝트의 목적 4](#_Toc514344857)

[3. 관련 알고리즘 (Algorithm) 5](#_Toc514344858)

[4. 프로젝트 관련 정보 7](#_Toc514344859)

[4.1 딥러닝 프레임워크 7](#_Toc514344860)

[4.2 Github URL 7](#_Toc514344861)

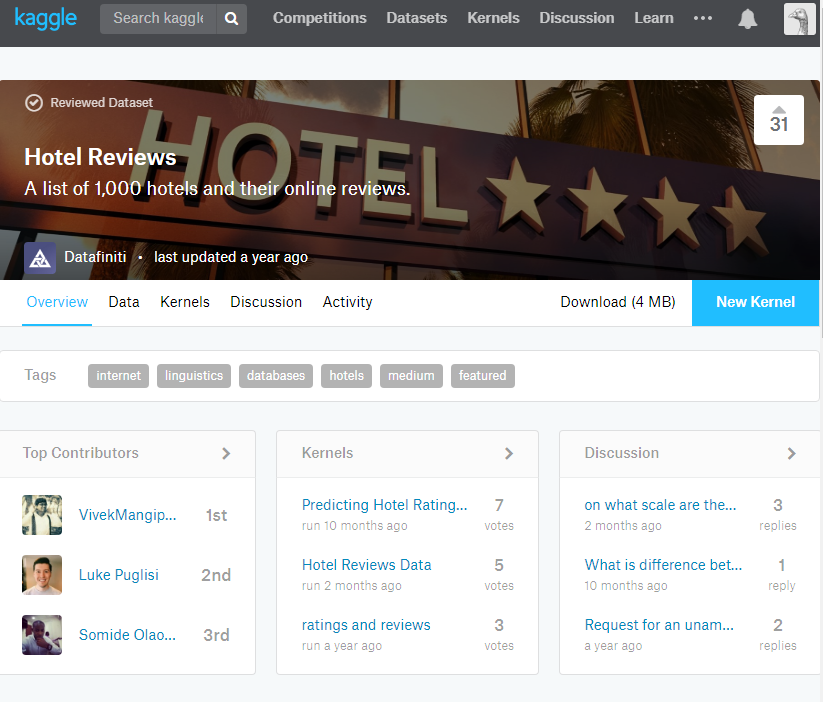
[4.3 역할 분담 7](#_Toc514344862)

[4.4 프로젝트 마일스톤 (milestone) 7](#_Toc514344863)

[5. 참고자료 8](#_Toc514344864)

# 데이터셋 (Dataset)

프로젝트를 위한 dataset은 Kaggle site의 Datasets channel을 활용하였습니다. (URL: https://www.kaggle.com/datafiniti/hotel-reviews)



해당 데이터셋은 Datafiniti의 비즈니스 데이터베이스에서 제공하는 1000개의 호텔 및 리뷰 목록입니다. 데이터셋에는 호텔 위치, 이름, 평점, 리뷰 데이터, 제목, 사용자 이름 등이 포함되어 있습니다. 데이터셋의 크기는 약 36000개입니다.



**[Figure 1] Dataset 예시**

# 프로젝트의 목적

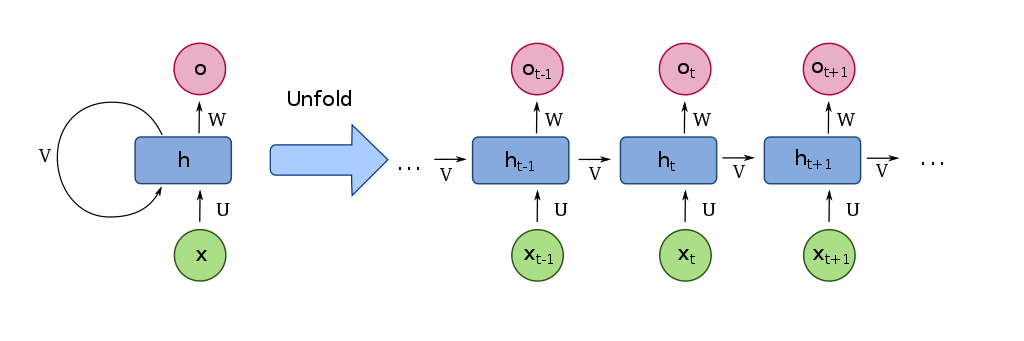
호텔 리뷰에 대한 데이터를 바탕으로 각 호텔별 평점을 예측하는 기계학습 기반의 모델을 만들 것입니다. 기본적으로 머신러닝을 활용하여 예측 모델을 세우기 위해서는 RNN, LSTM과 같은 모델을 사용합니다. 본 프로젝트는 deep하게 모델을 설계할 수 있는 LSTM 모델을 사용함으로써 빅데이터 기반의 딥러닝 모델을 학습하고, 높은 학습률 및 정확도를 나타내는 것을 목표로 하고 있습니다.

주어진 데이터셋은 호텔의 이름, 고객 정보, 휴대폰 번호, 지역, 리뷰 등 다양한 정보를 포함하고 있습니다(참고자료4 참고). 본 프로젝트는 영어로 쓰인 리뷰 중에서, 리뷰의 점수(review.rating)와 리뷰에 해당하는 글(review.text)만을 파싱하여 사용하는 것을 목표하고 있습니다. 프로젝트 이후에는 다른 정보를 융합하여 호텔 추천 시스템을 만들 계획에 있습니다.

# 관련 알고리즘 (Algorithm)

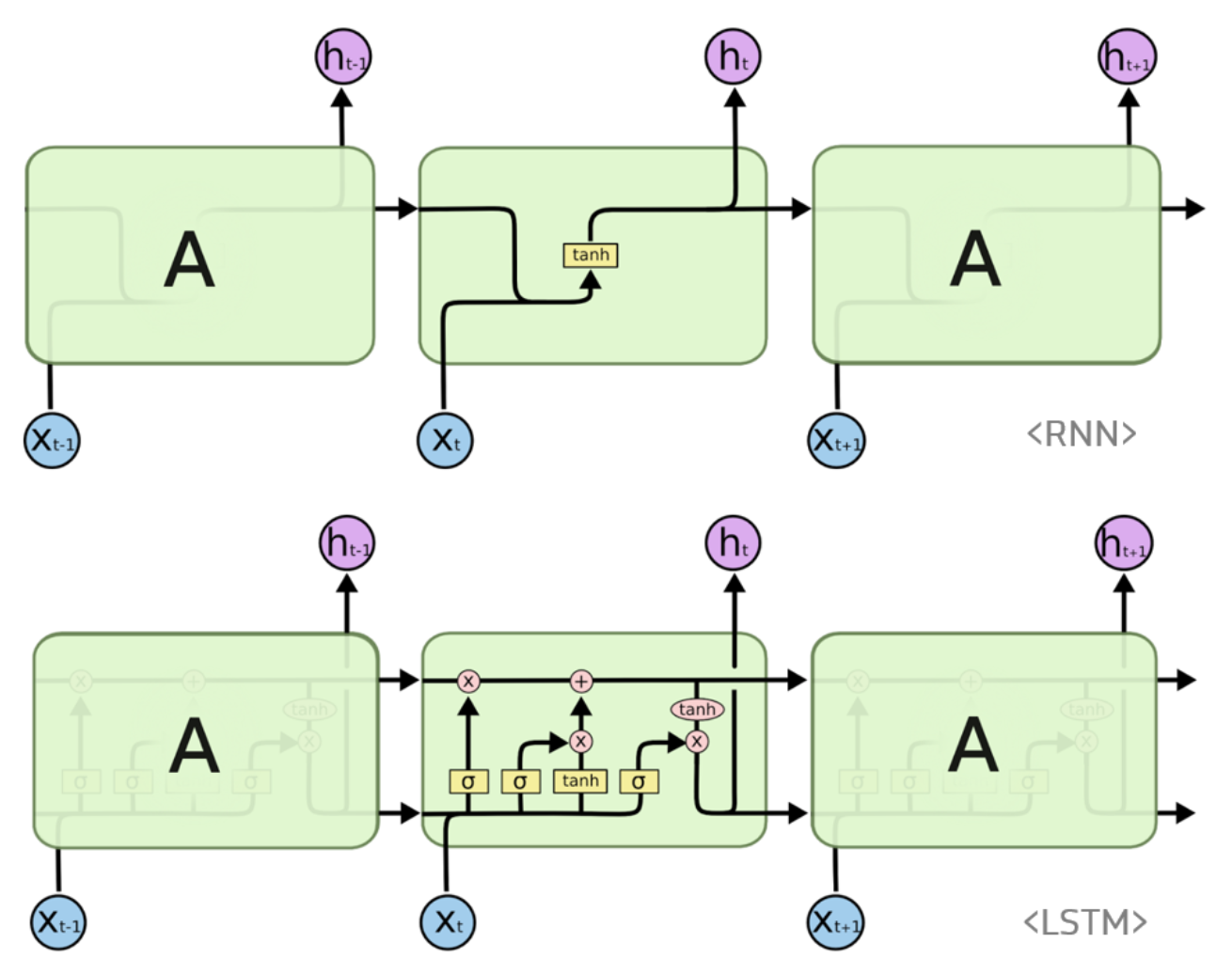
본 프로젝트는 리뷰 데이터(data)와 평점 데이터(label)를 이용하여 LSTM 기반의 예측모델을 만들 것 입니다.

RNN은 [인공 신경망](https://ko.wikipedia.org/wiki/%EC%9D%B8%EA%B3%B5%EC%8B%A0%EA%B2%BD%EB%A7%9D)의 한 종류로, 유닛 간의 연결이 [순환](https://ko.wikipedia.org/wiki/%EC%88%9C%ED%99%98_(%EA%B7%B8%EB%9E%98%ED%94%84_%EC%9D%B4%EB%A1%A0))적 구조를 갖는 특징을 갖고 있습니다. 이러한 구조는 순차적 동적 특징을 모델링 할 수 있도록 신경망 내부에 상태를 저장할 수 있게 해준다. [Feedforward 신경망](https://ko.wikipedia.org/w/index.php?title=%EC%A0%84%EB%B0%A9_%EC%A0%84%EB%8B%AC_%EC%8B%A0%EA%B2%BD%EB%A7%9D&action=edit&redlink=1)과 달리, recurrent 인공 신경망은 내부의 [메모리](https://ko.wikipedia.org/wiki/%EB%A9%94%EB%AA%A8%EB%A6%AC)를 이용해 [시퀀스](https://ko.wikipedia.org/wiki/%EC%8B%9C%ED%80%80%EC%8A%A4) 형태의 입력을 처리할 수 있습니다. 따라서 recurrent 인공 신경망은 필기체 인식이나 [음성 인식](https://ko.wikipedia.org/wiki/%EC%9D%8C%EC%84%B1_%EC%9D%B8%EC%8B%9D)과 같이 순차적 특징을 가지는 데이터를 처리할 수 있습니다.



**[Figure 2] Wikipedia RNN**

하지만 RNN은 관련 정보와 그 정보를 사용하는 지점 사이 거리가 멀 경우 back propagation시 gradient가 점점 줄어 학습 능력이 크게 저하되는 현상을 가집니다. 이를 vanishing gradient problem이라고 합니다. 이 문제를 보완하기 위해 RNN의 hidden state와 cell state를 추가로 가지고 있는 LSTM 모델을 사용할 계획입니다.



**[Figure 3] RNN와 LSTM의 차이**

LSTM은 cell state와 hidden state를 통해 추가적인 연산을 하게 되는데, 이를 통한 3가지의 gate가 생성됩니다. Forget gate, input gate, output gate인데, 이 gate들은 이전 값을 얼마나 기억할지, 그리고 현재 값을 얼마나 기억할지 등의 값을 제어하기 위한 목적으로 사용이 됩니다. 본 프로젝트는 간단한 LSTM model을 설계함으로써 호텔 리뷰와 같은 다대일(many-to-one) 문제를 해결할 것입니다.

# 프로젝트 관련 정보

## 딥러닝 프레임워크

본 프로젝트는 Tensorflow v1.8을 사용하여 진행할 것입니다.

## Github URL

<https://github.com/bonopi07/2018-1_advML_project>

## 역할 분담

* 데이터 수집 및 가공, 방법론 설계: 정성민
* Tensorflow 기반의 학습 모델 구현: 정성민
* 학습 모델 하이퍼 파라미터 튜닝 및 정확도 검증: 정성민

## 프로젝트 마일스톤 (milestone)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 5월  2주차  (5/7-5/13) | 5월  3주차  (5/14-5/20) | 5월  4주차  (5/21-5/27) | 5월  5주차  (5/28-5/31) | 6월  1주차  (6/1-6/3) | 6월  2주차  (6/4-6/10) | 6월  3주차  (6/11-6/17) |
| 프로젝트  준비 |  |  |  |  |  |  |  |
| 데이터 가공 |  |  |  |  |  |  |  |
| 학습 모델  설계 |  |  |  |  |  |  |  |
| 모델 튜닝 및 정확도 검증 |  |  |  |  |  |  |  |

# 참고자료

1. 모두를 위한 머신러닝/딥러닝 강의. <https://hunkim.github.io/ml/>
2. github blog. <https://ratsgo.github.io/natural%20language%20processing/2017/03/09/rnnlstm/>
3. Kaggle dataset URL. <https://www.kaggle.com/datafiniti/hotel-reviews>
4. <https://datafiniti-api.readme.io/docs/business-data-schema>