

Term Project 제안서

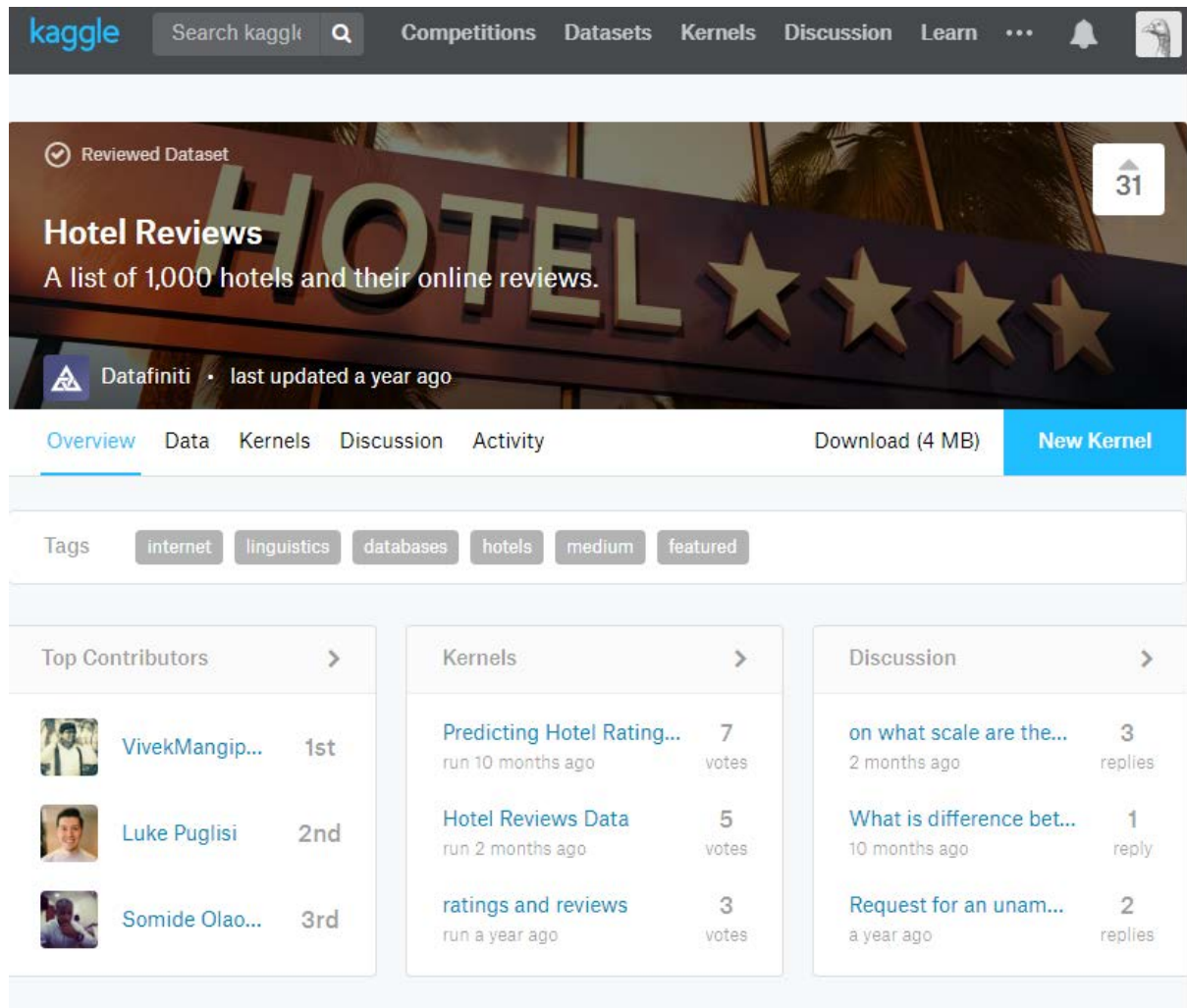
Department	Computer Science
Course title	Advanced Skills in Machine Learning
Instructor	Prof. Sung Soo Lim
Student ID	M2018075
Student Name	정성민
Due date	18.05.08

목 차

1. 데이터셋 (Dataset)	3
2. 프로젝트의 목적	4
3. 관련 알고리즘 (Algorithm).....	5
4. 프로젝트 관련 정보.....	7
4.1 딥러닝 프레임워크.....	7
4.2 Github URL.....	7
4.3 역할 분담.....	7
4.4 프로젝트 마일스톤 (milestone)	7
5. 참고자료	8

1. 데이터셋 (Dataset)

프로젝트를 위한 dataset 은 Kaggle site 의 Datasets channel 을 활용하였습니다.
(URL: <https://www.kaggle.com/datafiniti/hotel-reviews>)






Hotel Reviews
A list of 1,000 hotels and their online reviews.

Datafiniti · last updated a year ago

Overview Data Kernels Discussion Activity Download (4 MB) New Kernel

Tags: internet, linguistics, databases, hotels, medium, featured

Top Contributors	Kernels	Discussion
 Vivek Mangipudi 1st	Predicting Hotel Rating... 7 votes run 10 months ago	on what scale are the... 3 replies 2 months ago
 Luke Puglisi 2nd	Hotel Reviews Data 5 votes run 2 months ago	What is difference bet... 1 reply 10 months ago
 Somide Olao 3rd	ratings and reviews 3 votes run a year ago	Request for an unam... 2 replies a year ago

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

Preview (first 100 rows)		Column Metadata		Column Metrics								
	country	latitude	longitude	name	postalCode	province	reviews.date	reviews.dateAdded	reviews.doRecommend	reviews.id	reviews.rating	reviews.text
pleton	US	45.421611	12.376187	Hotel Russo Palace	30126	GA	2013-09-22T00:00:00Z	2016-10-24T00:00:25Z			4	Pleasant 10 min walk along the sea front to the Water Bus, restaurants etc. Hotel was comfortable breakfast was good - quite a variety. Room aircon didn't work very well. Take mosquito repellent!
pleton	US	45.421611	12.376187	Hotel Russo Palace	30126	GA	2015-04-03T00:00:00Z	2016-10-24T00:00:25Z			5	Really lovely hotel. Stayed on the very top floor and were surprised by a jacuzzi bath we didn't know we were getting! Staff were friendly and helpful and the included breakfast was great! Great location and great value for money. Didn't want to leave!
pleton	US	45.421611	12.376187	Hotel Russo Palace	30126	GA	2014-05-13T00:00:00Z	2016-10-24T00:00:25Z			5	Étt mycket bra hotell. Det som drog ner betyget var att vi fick ett rum under takarna dr det endast var full stöjd i 80 av rummets yta.
pleton	US	45.421611	12.376187	Hotel Russo Palace	30126	GA	2013-10-27T00:00:00Z	2016-10-24T00:00:25Z			5	We stayed here for four nights in October. The hotel staff were welcoming, friendly and helpful. Assisted in booking tickets for the opera. The rooms were clean and comfortable - good shower, light and airy rooms with windows you could open wide. Beds were comfortable. Plenty of choice for breakfast.Spa at hotel nearby which we used while we were there.
pleton	US	45.421611	12.376187	Hotel Russo Palace	30126	GA	2015-03-05T00:00:00Z	2016-10-24T00:00:25Z			5	We stayed here for four nights in October. The hotel staff were welcoming, friendly and helpful. Assisted in booking tickets for the opera. The rooms were clean and comfortable - good shower, light and airy rooms with windows you could open wide. Beds were comfortable. Plenty of choice for breakfast.Spa at hotel nearby which we used while we were there.
pleton	US	45.421611	12.376187	Hotel Russo Palace	30126	GA	2015-04-05T00:00:00Z	2016-10-24T00:00:25Z			5	We loved staying on the island of Lido! You need to take a water taxi from Venice to get there. From the train station, a boat ride takes 45 minutes but has beautiful views along the way. Hotel is an EASY walk from the boat dock. The room was very clean and the breakfast was plentiful. We would definitely recommend this hotel!
pleton	US	45.421611	12.376187	Hotel Russo Palace	30126	GA	2014-06-10T00:00:00Z	2016-10-24T00:00:25Z			4	Lovely view out onto the lagoon. Excellent view. Staff were welcoming and helpful.
pleton	US	45.421611	12.376187	Hotel Russo Palace	30126	GA	2015-05-14T00:00:00Z	2016-10-24T00:00:25Z			4	ottimo soggiorno e ottima sistemazione nei giorni frenetici di inaugurazione della Biennale. Le signore alla reception sono efficientissime e squisite e non sono da meno le ragazze che servono la prima colazione. Da tornarci

[Figure 1] Dataset 예시

2. 프로젝트의 목적

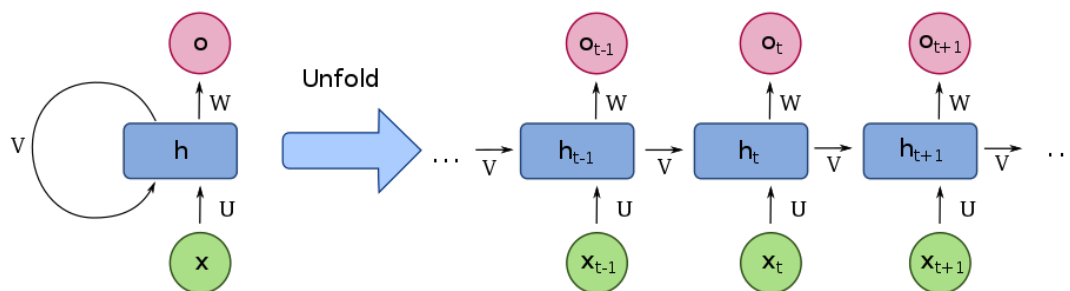
호텔 리뷰에 대한 데이터를 바탕으로 각 호텔별 평점을 예측하는 기계학습 기반의 모델을 만들 것입니다. 기본적으로 머신러닝을 활용하여 예측 모델을 세우기 위해서는 RNN, LSTM 과 같은 모델을 사용합니다. 본 프로젝트는 deep 하게 모델을 설계할 수 있는 LSTM 모델을 사용함으로써 빅데이터 기반의 딥러닝 모델을 학습하고, 높은 학습률 및 정확도를 나타내는 것을 목표로 하고 있습니다.

주어진 데이터셋은 호텔의 이름, 고객 정보, 휴대폰 번호, 지역, 리뷰 등 다양한 정보를 포함하고 있습니다(참고자료 4 참고). 본 프로젝트는 영어로 쓰인 리뷰 중에서, 리뷰의 점수(review.rating)와 리뷰에 해당하는 글(review.text)만을 파싱하여 사용하는 것을 목표로 하고 있습니다. 프로젝트 이후에는 다른 정보를 융합하여 호텔 추천 시스템을 만들 계획에 있습니다.

3. 관련 알고리즘 (Algorithm)

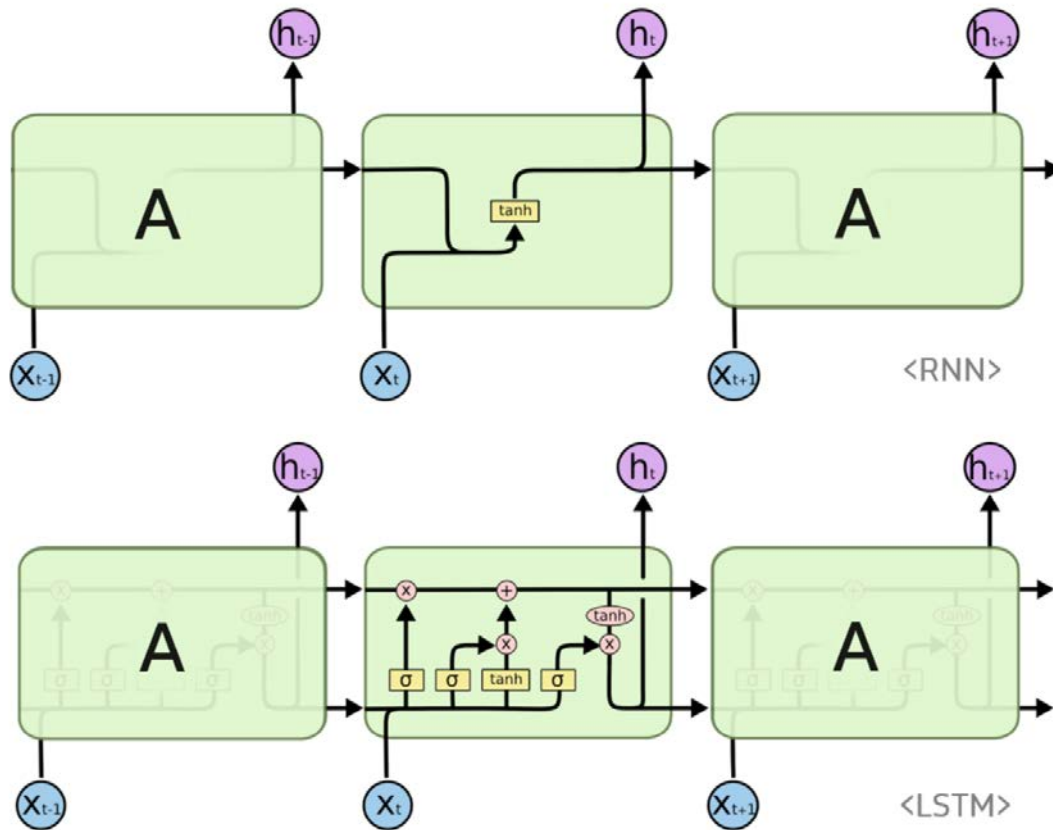
본 프로젝트는 리뷰 데이터(data)와 평점 데이터(label)를 이용하여 LSTM 기반의 예측모델을 만들 것 입니다.

RNN 은 인공 신경망의 한 종류로, 유닛 간의 연결이 순환적 구조를 갖는 특징을 갖고 있습니다. 이러한 구조는 순차적 동적 특징을 모델링 할 수 있도록 신경망 내부에 상태를 저장할 수 있게 해준다. **Feedforward** 신경망과 달리, **recurrent** 인공 신경망은 내부의 메모리를 이용해 시퀀스 형태의 입력을 처리할 수 있습니다. 따라서 **recurrent** 인공 신경망은 필기체 인식이나 음성 인식과 같이 순차적 특징을 가지는 데이터를 처리할 수 있습니다.



[Figure 2] Wikipedia RNN

하지만 RNN 은 관련 정보와 그 정보를 사용하는 지점 사이 거리가 멀 경우 **back propagation** 시 **gradient** 가 점점 줄어 학습 능력이 크게 저하되는 현상을 가집니다. 이를 **vanishing gradient problem** 이라고 합니다. 이 문제를 보완하기 위해 RNN 의 **hidden state** 와 **cell state** 를 추가로 가지고 있는 LSTM 모델을 사용할 계획입니다.



[Figure 3] RNN 와 LSTM 의 차이

LSTM 은 cell state 와 hidden state 를 통해 추가적인 연산을 하게 되는데, 이를 통한 3 가지의 gate 가 생성됩니다. Forget gate, input gate, output gate 인데, 이 gate 들은 이전 값을 얼마나 기억할지, 그리고 현재 값을 얼마나 기억할지 등의 값을 제어하기 위한 목적으로 사용이 됩니다. 본 프로젝트는 간단한 LSTM model 을 설계함으로써 호텔 리뷰와 같은 다대일 (many-to-one) 문제를 해결할 것입니다.

4. 프로젝트 관련 정보

4.1 딥러닝 프레임워크

본 프로젝트는 Tensorflow v1.8 을 사용하여 진행할 것입니다.

4.2 Github URL

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

4.3 역할 분담

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

4.4 프로젝트 마일스톤 (milestone)

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

5. 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>