4차 6기 세미프로젝트

|  |
| --- |
| RNN을 이용한 네이버 영화 리뷰 분류 |

**2020년 5월 13일**

**딥러닝 기반 AI 엔지니어링 (B)**

영화 말해 Yes or No !

고민정

문다영

김태완

박나은

신하선

**목 차**

**1. 프로젝트 개요**  **1**

1.1 프로젝트 기획 배경 및 목표 1

1.2 구성원 및 역할 2

1.3 프로젝트 추진 일정 3

**2. 프로젝트 결과**  **6**

2.1 데이터 수집 7

2.2 데이터 분석 7

2.3 데이터 분석 결과 7

**3. 기대 효과**  **8**

3.1 향후 개선 사항 8

3.2 기대 효과 8

**4. 개발 후기**  **9**

**1. 프로젝트 개요**

* 1. 프로젝트 기획 배경 및 목표

모든 제품이나 캠페인은 기획과 제작 전에 반드시 시장분석이 선행됩니다. 이 때 업무 효율을 높이기 위해 사원이나 인턴이 각종 커뮤니티 반응, 인터넷 뉴스 기사 본문, 블로그, 카페의 수많은 글을 통해 시장 분위기를 파악하여 하나의 자료로 요약하는 ‘Fact Book’을 제작하는 과정이 필요합니다. 하지만 이는 사람이 하나하나 판별해야 하기 때문에 방대한 양의 글을 모두 분석할 수 없으며 긍정과 부정 어조의 기준을 정하기 어렵기 때문에 많은 시간이 걸리며 데이터 누락이 생길 수 있다는 한계점이 있습니다.

한편 영화 리뷰의 경우, 리뷰와 ‘별점’이 함께 제시되기 때문에, 해당 글의 어조를 유추할 수 있다는 특성을 가집니다. 따라서 네이버 영화의 수십만 건에 달하는 리뷰와 별점을 스크래핑하여 별점에 따라 라벨을 붙이고 학습시킬 수 있습니다. 즉 어조 분석에 필요한 많은 양의 데이터를 얻을 수 있기 때문에 영화 리뷰의 어조 분석으로 시작하고자 합니다.

리뷰와 같은 언어 데이터는 ‘Sequence Data’ 라고 합니다. 시간의 흐름에 따라 발생하는 연속적인 데이터를 뜻합니다. 주식데이터, 영화, 음악, 자연어 처리 등등이 이 Sequence Data 에 포함됩니다. 하지만 기존 Nueral Network 으로는 이런 Sequence Data를 처리할 수 없습니다. 그래서 새로운 network Model을 만들게 되고 이것이 RNN ( Recurrent Neural Network ) 입니다.

RNN은 이전 상태의 정보를 메모리 형태로 저장하는 형식으로 앞에서 얻은 정보가 뒤에서 얻은 데이터와 연관관계를 가지는 데이터를 다룰 때 강력한 효과적인 장점을 가지고 있습니다. 인간의 언어는 앞 뒤 문맥을 가지고 있기 때문에 RNN 적용하기 매우 적합합니다. 하지만 RNN은 관련 정보와 그 정보를 사용하는 지점 사이 거리가 멀 경우 역전파시 그래디언트가 점차 줄어 학습능력이 크게 저하되는 것으로 알려져 있습니다. 이를 vanishing gradient problem이라고 합니다. 이를 보완하기 위해 LSTM 모델을 사용합니다. LSTM 모델은 아주 세심한 선형 연산을 거치기 때문에 정보가 큰 변함 없이 계속적으로 다음 단계에 전달되게 합니다.

따라서 이 RNN과 LSTM을 이용하여 네이버 영화 리뷰 글을 분석하고 해당 영화에 대한 리뷰 글이 긍정적인지 부정적인지 알아내고자 합니다.

1.2 구성원 및 역할

|  |  |  |  |
| --- | --- | --- | --- |
| 이름 | 전공 | 역할 | 구현 부분 |
| 고민정 | 경영학부/  컴퓨터공학과 | 팀장 | RNN 모델링, 데이터전처리 |
| 김태완 | 게임소프트웨어학과 | 팀원 | 리뷰 데이터 크롤링, 전처리 |
| 문다영 | 도시행정학과  /경영학부 | 팀원 | 데이터 크롤링 및 모델링 |
| 박나은 | 광고홍보학과 | 팀원 | 데이터 전처리 및 모델 성능 개선 |
| 신하선 | 교육심리학과  /생명환경공학 | 팀원 | 데이터 전처리 및 모델 성능 개선 |

1.3 프로젝트 추진 일정

|  |  |  |  |
| --- | --- | --- | --- |
| 구분 | 기간 | 활동 | 비고 |
| 사전  기획 | 4.30 | 프로젝트 기획 및 팀 구성 |  |
| 4.30 | PJT주제 선정, 팀(PM/팀원) 구성 | 3~5 인/팀 |
| PJT  수행  /  완료 | 4.30 | 프로젝트 수행 |  |
| 4.30 | 프로젝트 설계 |  |
| 5.01 | 구현 및 테스트 |  |
| 5.04 | 팀별 최종 발표 (구축 완료 보고) | 최우수 한 팀 선발 |

**2. 프로젝트 개발 결과**

2.1 데이터 수집

네이버 영화 리뷰 평점(<https://movie.naver.com/movie/point/af/list.nhn>) 사이트에서 Python 의 beautifulsoup 과 pandas를 사용하여 평점, 리뷰(글)의 묶음 형식으로 스크래핑 했습니다. 총 10,000개 리뷰로 구성된 데이터를 사용하였습니다. 영화 리뷰에 대한 텍스트와 해당 리뷰의 평점이 7이상일때 긍정인 경우로써 1을 부여합니다. 반대로 평점이 4이하일 때 부정인 경우로써 0으로 표시한 레이블로 구성합니다.

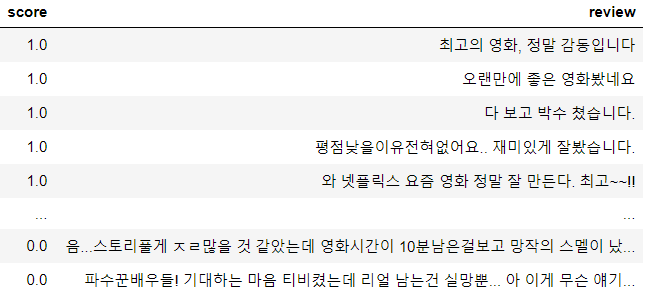


그림 1 해당 리뷰와 1(긍정)과 0(부정)으로 구성된 레이블

2.2 데이터 분석

2.2.1 네이버 영화 리뷰 데이터에 대한 이해와 전처리

학습데이터에 한글과 공백을 제외하고 모두 제거하는 정규 표현식을 수행했습니다.

토큰화 과정에서 불용어를 제거했습니다. 불용어는 정의하기 나름인데, 한국어의 조사, 접속사 등의 보편적인 불용어를 사용할 수도 있겠지만 여기서는 '의', '가', '이', '는', '은', '들', '잘', '좀', '과', '도', '을', '를', '으로', '한', '하다', '에' 를 사용했습니다.

한국어 데이터는 토큰화(tokenization)를 할 때 형태소 분석기를 사용합니다. 여기서는 토큰화를 위한 형태소 분석기로써 KoNLPy의 Kkma를 사용했습니다. 이러한 형태소 분석기를 사용함으로써 일정 수준의 정규화를 수행해주는데, 예를 들어 '이런'이 '이렇다'로 변환되었고 '만드는'이 '만들다'로 변환되었습니다. 훈련 데이터 뿐만 아니라, 테스트 데이터에 대해서도 동일하게 토큰화를 해줍니다.

등장 빈도가 threshold 값인 3회 미만. 즉, 2회 이하인 단어들은 정수 인코딩 과정에서 배제시키고 등장 빈도수가 2이하인 단어들의 수를 제외한 단어의 개수를 단어 집합의 최대 크기로 제한하겠습니다.

그 다음으로 서로 다른 길이의 샘플들의 길이를 동일하게 맞춰주는 패딩 작업을 진행하였습니다. 전체 샘플 중 길이가 100 이하인 샘플의 비율: 95.17% 임을 확인했습니다. 모든 샘플의 길이를 100으로 맞췄습니다.

2.3 데이터 분석 결과

LSTM으로 네이버 영화 리뷰 감성 분류하고자 합니다. 최적의 학습 파라미터를 찾기 위해 그리드서치(GridSearchCV)를 수행했습니다. 그 결과 최적의 파라미터가 <표 1> 과 같이 정해졌습니다..

|  |  |
| --- | --- |
| embedding\_size | 150 |
| LSTM\_node | 128 |
| batch\_size | 10 |
| embedding\_size | 150 |
| epochs | 5 |
| optimizer | adam |

표 1 그리드 서치를 이용한 최적의 학습 파라미터

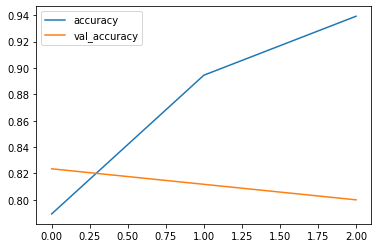


그림 2 RNN 모델의 accuracy : 0.8000

최종적으로 최적의 학습 파라미터를 가지고 테스트 데이터를 예측(Predict)한 결과, <그림 2> 의 결과처럼 80.00% 의 정확률(accuracy) 결과값을 가지게 되었습니다.

**3. 기대 효과**

3.1 향후 개선 사항

‘RNN을 이용한 네이버 영화 리뷰 분류’ 프로젝트는 크게 data의 수(크기), 모델링 분야에 개선 사항이 있습니다.

우선 data의 수(크기) 관련해서는 전체 100,000 데이터를 사용했습니다. 네이버 리뷰 사이트에서 그 이상의 리뷰 글을 제공해주지 않았습니다. 하지만 전체적인 정확도를 올리기 위해서는 다른 영화 리뷰 사이트나(예: 다음 영화 사이트) 개별적인 영화를 직접 검색해서 해당하는 블로그나 여러 SNS 등의 리뷰 글을 수집할 필요성이 보여집니다. 더욱 많이 데이터의 수를 늘려 학습한다면 정확도가 개선될 것으로 기대됩니다.

모델링 관련해서는 GridSearchCV 를 시행할 때 다양한 파라미터로 시행해야 합니다. 현재 이 프로젝트에서는 각 임베딩 노드, LSTM 노드, optimizer, epoch 요소에 2개씩 입력하여 그리드서치를 시행했습니다. 최적의 학습 파라미터를 알기 위해서는 각 요소에 더 많은 파라미터를 삽입하여 다양한 경우에서 더욱 최적의 결과를 얻을 수 있어야 합니다. 이러한 시도는 모델이 더 나은 정확도를 얻을 것으로 예상됩니다.

3.2 기대 효과

이번 프로젝트를 통해 완성된 모델은 단순히 영화 리뷰를 분석하는 데서 그치는 것이 아니라 다양한 분야에서의 활용이 가능합니다. 기존에 사람이 판별해야 했던 소비자들의 반응을 효율적으로 수치화 할 수 있으며, 사람이 보기엔 한계가 있는 많은 양의 데이터를 빠르게 처리하는 것 역시 가능해질 것입니다. 향후 데이터 기반 시장분석이 필요한 마케팅 솔루션에 중요한 근거를 찾는 데에 활용되는 것을 기대합니다.

**4. 개발 후기**

|  |
| --- |
|  |

|  |  |
| --- | --- |
| 성명 | 후기 |
| 고민정 | 시계열 data와 문맥은 완전히 다른 개념으로 생각이 되어서 자연어 처리를 시계열 data를 다루는 RNN을 이용해 작동한다는 이야기를 들었을 땐 이해가 가지 않았지만, 수업을 듣고 프로젝트를 진행하면서 자연어 처리가 어떤 방식으로 진행되는 지 감을 잡을 수 있었습니다. |
| 문다영 | 형태소 분석기도 다양하고, parameter에 따라 나올 수 있는 모델도 다양해서 최고의 성능을 보여주는 모델을 찾기까지 어려움이 많았습니다. 하지만 팀원들이 합심해서 좋은 결과를 얻을 수 있어 기뻤습니다. |
| 김태완 | 여러가지 이미지 관련 모듈과 기능, 다루는 법을 연습할 수 있었음과 RNN 관련 개념을 익히는데 도움을 준 팀원들에게 감사합니다. |
| 박나은 | 웹에서 데이터를 가져오는 스크래핑을 복습할 수 있었고, 좋은 팀원들이 끌어준 덕분에 RNN 프로젝트를 무사히 마칠 수 있어 다행이었습니다. |
| 신하선 | RNN과 LSTM을 이론으로만 학습했을 때는 막막했는데 실제 네이버 영화 댓글을 이용하여 실습 했을 |