1p

안녕하세요, AI 부트캠프 1기 김세진입니다. 프로젝트 발표를 시작하겠습니다.

2p

제가 프로젝트에서 사용한 데이터는 이 두 가지 데이터입니다. 첫번째는 2004년부터 2005년까지의 BBC 뉴스를 요약한 데이터이고, 두번째는 2020년 뉴욕타임즈 기사의 제목과 기사 내용을 요약한 데이터입니다. 두 데이터들 모두 캐글 사이트에 제공되어 있습니다. 이 두가지 데이터에 자연어 처리와 KNN이라는 알고리즘을 이용하여 내용 기반 추천 시스템을 만들어보는 프로젝트를 진행했습니다.

3p

최근 많은 기업들에서 추천 서비스를 많이 도입하고 있습니다. 특히 네이버 Clova에서는 개인화추천 시스템을 전담하는 팀이 따로 있을 정도로 추천 알고리즘을 중요하게 생각하고 개발하고 있습니다. 그리고 최근 제가 공모전을 준비하면서 기획했던 서비스에서 내용 기반 추천 시스템을 도입하려고 했었는데, 그당시에는 제가 어떤 기술을 써야 할지 잘 모르겠어서 추천 서비스를 넣지 않았던 경험이 있습니다. 그래서 추천 서비스에 대해 공부해 놓으면 앞으로 제가 취업을 하게 될 때도 유용하게 사용할 것 같고, 한번쯤은 만들어보고 싶은 시스템이기 때문에 이걸 주제로 택하게 되었습니다. 원래는 인터넷 커뮤니티에 있는 글들을 모아 분석해보고 싶었는데, 데이터를 모으는 데 시간이 오래 걸리고 사용하는 언어가 비속어가 포함된 글이 많기 때문에 최대한 문법적으로 올바르고 정형화된 글을 찾아 분석해야겠다고 생각했습니다. 그래서 bbc와 뉴욕타임즈 같은 기사 데이터를 이용하여 프로젝트를 진행했습니다.

4p

이 프로젝트의 목표는 특정 기사와 비슷한 내용을 가지고 있는 기사를 찾는 것이고, nlp와 knn알고리즘을 이용하여 해결할 수 있습니다. Nlp는 사람들이 사용하는 자연스러운 언어를 컴퓨터가 이해할 수 있게 처리하는 과정이고, knn알고리즘은 분류 문제에서 어떤 데이터가 주어지면 그 주변의 데이터를 살펴본 뒤 가장 가까운 k개의 이웃의 정보로 그 데이터가 어떠한 데이터인지 예측하는 방법론입니다.

5p

기사 데이터를 분석하기 전 전처리하는 과정입니다. 먼저 필요 없는 데이터와 중복 데이터를 정리하고, 텍스트에 있는 특수문자와 대문자 등을 정리하고, 의미가 없는 단어인 불용어를 정리해 준 다음에 문법적 요소를 제거하여 의미 단위로 나타낸 단어를 뽑아냅니다.

6p

전처리를 하고 나서 기사에서 가장 많이 나온 단어가 무엇인지 시각화해 보았습니다. Bbc 기사에서는 크게 핵심이 되는 단어를 찾기는 어려웠습니다. 데이터의 개수가 그리 많지 않아서 그렇다고 생각합니다.

7p

뉴욕타임즈 기사의 단어를 시각화한 결과입니다. New, coronavirus, trump, president와 같은 단어들이 높은 빈도를 보이는데요, 이는 2020년에 신종 코로나 바이러스가 유행하기 시작했고 미국 대통령 선거가 있었던 때이기 때문에 사회적으로 그러한 관심이 높아져서 나온 결과라고 추측할 수 있습니다.

8p

knn알고리즘을 이용하여 특정 기사와 내용이 비슷한 5개의 기사들을 뽑아 보았습니다. Bbc 기사에서 100번째 기사와 유사한 기사들을 뽑아 보니 은행 금리와 관련되었다는 공통점이 있었습니다.

9p

뉴욕타임즈에도 알고리즘을 적용해 보니, 10000번째 기사와 유사한 기사들은 주로 아프가니스탄 지역과 관련된 내용이라는 공통점이 있었습니다.

10p

분석을 하다 보니 뉴욕타임즈의 2020년 기사들에 코로나바이러스와 관련된 내용이 많다는 것을 알았고, 이러한 사실을 바탕으로 기사를 분류하는 모델을 만들 수 있겠다는 생각이 들었습니다. 그래서 특정 단어의 유무를 판단하는 랜덤 포레스트 분류 모델을 만들어보았습니다.

11p

분류 모델을 만들기 위해 뉴욕 타임즈 기사 데이터를 다시 전처리했는데요, 기사 내용에 ‘coronavirus’ 라는 단어가 포함되어 있으면 1을 표시하고, 포함되어 있지 않으면 0을 표시하여 라벨을 만들어 주는 작업을 했습니다. 분류 모델이 기사들 중에서 1로 표시된 기사만을 찾아낼 수 있도록 하는 것이 목표입니다.

12p

랜덤포레스트 모델과 비교하기 위해 베이스라인을 만들어 보았습니다. 로지스틱 회귀 모델을 이용했는데, 이 모델에서는 타겟변수가 0과 1로 나누어진 데이터를 사용할 수 있고 선형 모델에 가까운 모델이라 간단하기 때문에 베이스라인으로 사용하기 적합하다고 생각했습니다. 베이스라인 모델의 정확도는 0.86으로 나타났습니다.

13p

분석 모델의 구조입니다. 앞에서의 베이스라인과 달리 랜덤포레스트 분류기를 사용했으며, 분류기의 파라미터를 조정한 후에 모델을 학습시켰습니다. 이 모델의 정확도는 0.93으로 베이스라인 모델보다 정확도가 높게 나온 것을 볼 수 있습니다.

14p

프로젝트 결과는 여기까지입니다. 하지만 아쉬운 점이 조금 남습니다. 처음 프로젝트를 계획할 때는 knn 알고리즘이 아닌 다른 방법을 추가로 공부해서 내용 기반 추천 시스템을 만들어보고 싶었는데, 제가 사용한 뉴욕타임즈 기사 데이터가 너무 커서 모델을 한 번 돌릴 때 거의 서너시간 정도가 소요되었기 때문에 다양한 시도를 하기가 어려웠습니다. 데이터를 임의로 줄일까도 생각했지만 데이터를 줄일 만한 기준이 마땅히 없었기 때문에 그대로 사용한 것이 문제였다고 생각합니다. 그리고 뉴욕타임즈 데이터는 1900년대부터 2000년대까지 매년 방대한 데이터가 제공되었기 때문에 다양한 년도의 데이터를 분석하면서 시대별로 미국의 관심사에 대해 트렌드 분석을 해보고 싶었는데, 마찬가지로 데이터가 너무 커서 그렇게 하지 못했던 점이 아쉽습니다. 다음에 기회가 된다면 트렌드 분석까지 해보고 싶고, 추천 시스템과 관련된 알고리즘을 더 많이 공부해보고 싶습니다. 발표를 마치겠습니다. 감사합니다.