정보처리 및 자연언어처리

과제3 : NaiveBayesClassifier를 이용하여 accuracy 높은 모델 만들기

경제학/빅데이터사이언스 전공 20180590 이윤주

0.Data 정제

-entity\_data\_utf8.txt 파일의 정보를 전체/한국어만 어휘 별로 (어절, 개체명)만 뽑아 각각 labels, kor\_word로 저장한다.

 텍스트이(가) 표시된 사진

자동 생성된 설명

텍스트이(가) 표시된 사진

자동 생성된 설명

-이 후 각 방법(ngram, 자소분리 등등)을 위한 feature expression을 한 후,

-훈련셋과 테스트셋으로 나누어 훈련셋으로 모델을 훈련시킨다.

-테스트셋으로 정확도를 평가한다.

1.ngram을 이용한 모델

텍스트이(가) 표시된 사진

자동 생성된 설명

-라벨링 되어있는 배열을 90%는 훈련셋, 10%는 테스트셋으로 지정한다.

-유니그램, 바이그램, 트라이그램에 훈련셋을 넣어 학습시킨다.

-테스트셋을 넣고 정확도를 평가한다.

2.한글 자소 분리를 이용한 모델

텍스트이(가) 표시된 사진

자동 생성된 설명

-한글 자음/모음을 분리하는 함수를 작성 후 kor\_word에서 가장 많이 쓰이는 종성을 common\_suffix에 저장한다.

텍스트이(가) 표시된 사진

자동 생성된 설명 텍스트이(가) 표시된 사진

자동 생성된 설명

-feature extraction 진행

테이블이(가) 표시된 사진

자동 생성된 설명

-훈련셋 90%, 테스트셋 10%로 나누어 naivebayesclassifier로 훈련시킨 후 정확성을 체크한다.

3. 끝글자를 추출하여 비교

텍스트이(가) 표시된 사진

자동 생성된 설명

-마지막 글자를 feature로 삼아서 동일하게 진행.

3)-2. Feature 두 개를 합쳐서 시도

텍스트이(가) 표시된 사진

자동 생성된 설명 텍스트이(가) 표시된 사진

자동 생성된 설명

텍스트이(가) 표시된 사진

자동 생성된 설명

-2보다는 향상됐지만 3보다는 낮으므로 단일 feature가 낫다.

3)-3. 마지막 두 글자를 추출하여 비교

텍스트이(가) 표시된 사진

자동 생성된 설명

-가장 높은 정확도를 보인다.

4.(시도)문장 별 형태소 분석을 이용한 모델

텍스트이(가) 표시된 사진

자동 생성된 설명

-어휘 별로 끊어져 있는 것을 문장으로 만들어 꼬꼬마 형태소 분석기로 분석하고,

(아웃라인에 출력된 것은 꼬꼬마 형태소 분석기가 분석한 형태소)

가장 많이 쓰인 형태소와 일치하는 것을 feature로 두어 훈련을 시키려고 계획하였습니다.

그러나 코딩 과정에서 해결되지 않는 점이 있어 정확도를 산출하지 못했습니다.

5.마치며

-의외로 끝 두 글자로 학습한 것이 가장 높은 정확도를 보였습니다.

더 다양하고 높은 정확도의 feature extraction이 존재할 것인데 아직 시도해보지 못해서 아쉽습니다. 그리고 자연어처리의 가장 중요한 부분이 데이터 정제임을 깨닫게 되는 과제였습니다. 텍스트를 모델링하게 쉽게 변환하는 과정이 가장 힘들었기 때문입니다.