스마트테크놀로지AIR – HW5_Naive Bayes

컴퓨터과학과

2018147563 주우진

1. 코드 설명

In[1] ~ In[5]: App에 관한 문장들을 전처리하는 과정이다. 모든 문자를 소문자로 바꾸고 각종 문장 기호들을 공백으로 대체하여 출력하였고 세 글자 이하인 단어들은 제외하되 공백을 기준으로 단어들을 구분하여 각 단어를 토큰으로 만들었다. 그 후, 모든 토큰에 대해 등장한 횟수에 1씩 더하고 등장할 확률과 확률에 로그를 취한 값을 출력하였다. 등장 횟수에 1을 더하는 이유는 이후 테스트할 데이터에서 확률에 로그를 취하기 위해 단어의 등장 횟수가 0보다 크게 되게 하기 위함이다.

In[6] ~ In[8]: App의 경우와 마찬가지로 Others에 대해 데이터를 전처리하고 각 단어별로 등장 확률과 확률에 로그를 취한 값을 출력하였다.

In[9] ~ In[11]: 테스트할 데이터에 대해 전처리를 하고 토큰별로 특정 토큰이 등장했을 때 해당 토큰이 등장했을 때 전체 문장이 App에 관련되었을 확률과 Others에 관련되었을 확률을 각각 리스트로 만들어 저장하고 출력하였다.

In[12] ~ In[13]: 각 문장에 대해 모든 토큰의 로그값을 더하여 특정 문장이 App에 관련되었을 확률과 Others에 관련되었을 확률의 로그값을 저장하고 출력하였다.

In[14]: 두 로그값을 비교하여 App과 Others 중 어느 것에 더 관련되었을지 추정한 값을 저장하고 출력하였다.

2. 분석

Naïve Bayes 기법을 이용하기 위해 모든 문장을 전처리하여 단어들을 토큰으로 분리하였고 각 토큰들이 App과 관련되었을 확률과 Others와 관련되었을 확률을 각각 구한다. 그 후 한 문장 안의 모든 토큰들의 확률을 합산하여 한 문장 전체가 App과 관련되었을 확률과 Others와 관련되었을 확률을 비교하여 그 확률이 높은 쪽으로 추정한다. 테스트 데이터의 상위 10개는 App과 관련된 문장이었고, 하위 10개는 Others와 관련된 문장이었는데, 추정 결과, 상위 10개 문장에 대해 결과가 일치하였고, 하위 10개 중 1개의 문장을 제외한 나머지 문장에 대해 결과가 일치하였다. 전체적으로는 95%의 정확도를 보였다.