Artificial Intelligence

Homework Assignment 7.

1. 강의시간에 활용했던 영화평(English text data) 데이터를 사용하여 다음에 해당하는 코드를 작성하시오. 아래 모든 문항에 대한 설명과 충분한 증빙자료(코드, 스크린 샷, 주석 등)가 보고서에 포함되어야 합니다.

(a) 영화평 데이터 파일을 load_files를 사용하여 train data와 test data로 분리하여 loading 하고, 모델 학습을 위해, train data와 test data를 각각 data와 label로 분리하시오. (1점)

(b) 모델의 성능을 향상시키기 위해, text data중에서
를 삭제하고 삭제 전 후 test data의 일부를 출력하고 비교 및 설명하시오. (1점)

- Python의 comment로 처리하여 답을 제출하거나 별도의 파일로 답을 작성하여 zip하여 제출

(c) CountVectorizer를 사용하여 학습 데이터에 대한 BOW를 생성하고 built-in English stopword list를 사용하여 불용어를 제거하시오. (1점)

(d) Naive Bayesian 방식의 Gaussian Distribution을 사용하여 분류 모델을 train 시키고, 이를 이용하여 prediction을 실행하고 정답률을 출력하시오. (2점)

- 정답률은 소수점 둘째자리에서 반올림하여야 함.

2. 강의시간에 활용했던 한글 영화평 데이터(ratings_train.txt, ratings_test.txt)를 사용하여 다음에 해당하는 코드를 작성하시오. 아래 모든 문항에 대한 설명과 충분한 증빙자료(코드, 스크린 샷, 주석 등)가 보고서에 포함되어야 합니다.

(a) 영화평 데이터 파일을 read_csv를 사용하여 train data와 test data로 분리하여 loading 하고, 모델 학습을 위해, train data와 test data를 각각 data와 label로 분리하시오. (1점)

(b) 모델의 성능을 향상시키기 위해, text data중에서 Josa, Eomi, Punctuation, Korean Particle 들을 삭제하고 삭제 전 후 test data의 일부를 출력하고 비교 및 설명하시오. (1점)

- Python의 comment로 처리하여 답을 제출하거나 별도의 파일로 답을 작성하여 zip하여 제출

(c) TfidfVectorizer를 사용하여 학습 데이터에 대한 BOW를 생성하고, BOW의 일부를 sparse matrix 형태와 array 형태로 출력하시오. (1점)

(d) Multinomial Naive Bayesian classification을 사용하여 분류 모델을 train 시키고, 이를 이용하여 prediction을 실행하고 정답률을 출력하시오. (1점)

- 정답률은 소수점 둘째자리에서 반올림하여야 함.

(e) Multinomial distribution smoothing (α값 tuning)을 통하여 prediction 정답률(accuracy) 변화 추이를 확인하고 각 α값에 대한 정답률을 출력하시오. (1점)

- 최적에 근접한 정답률이라고 생각되는 α값과 정답률 pair를 반드시 포함하여야 함

Submitting your assignment :

- Due date: Zip your file and upload it at https://lms.mju.ac.kr/ by 24:00 Monday May 11th, 2022.
- Your homework cover page must be of the form provided by the lms.
- You must zip the homework cover page and your jupyter notebook assignment file(*.ipynb).
- Both of your file names must be of the form "hw7_StudentId_StudentName.ipynb", i.e., hw7_60063539_강예진.ipynb
- You must protect your homework from others. Any form of academic dishonesty will not be tolerated. If you get caught, you will receive -10 points for this homework!
- This assignment is 10 points total and the late penalty is 3 points per day.