기말과제

홍정하

과제의 데이터 파일

- entity_data_utf8.txt
  - 2018년 네이버 NLP Challenge 개체명 인식 데이터
  - 총 1,063,571 라인
  - 라인별 형식: '문장내어절번호\t어절\t개체명\n'

```
1	비토리오	PER
2	양일	DAT
3	만에	-
4	영사관	ORG
5	감호	CVL
6	용퇴,	-
7	항룡	-
8	압력설	-
9	의심만	-
10	가율	-
```

과제 I: 학습 및 테스트 파일 추출[배점: 2점]

- 다음의 작업을 수행하는 py 파일 만들기
  - 작업폴더의 entity_data_utf8.txt를 불러와서
  - 다음과 같이 문장내어절번호를 제거하고
    라인별 형식: '어절\t개체명\n'

비토리오	PER
양일	DAT
만에	-
영사관	ORG
감호	CVL
용퇴,	-
항룡	-
압력설	-
의심만	-
가율	-

  - 첫 1,00,000 라인은 train.txt 파일에, 그 후 63,571 라인은 test.txt 파일로 현재 작업 폴더에 출력하기(정답 출력 파일은 첨부된 '과제 I_출력결과'의 파일 참조)

과제 II: 기계학습기 만들기[배점: 8점]

- 다음을 수행하는 py 파일 만들기
  - 과제 I에서 만든 train.txt 파일을 불러와서
  - nltk.NaiveBayesClassifier.train을 이용하여 정확도가 높은 자질들을 구성하여 [어절별 개체명]을 학습하고
  - nltk.classify.accuracy을 이용하여 test.txt 파일에 대해 정확도를 측정하고, 정확도를 print 함수를 이용하여 화면 출력
- 과제 II 평가 기준
  - 정확도 83% 이상 8점
  - 정확도 82% 이상 7점
  - 정확도 80% 이상 6점
  - 정확도 78% 이상 5점
  - 정확도 76% 이상 4점
  - 그 이하 3점

제출물 및 기한

- py 파일
  - 과제I, 과제II를 해결하는 코드를 하나의 py 수록
- 채점 방식
  - 각 과제 배점 한도 내에서 기능 구현 미진 사항마다 -1점 감점제
- 제출기한: 6월28일(금) 23시59분(지각제출 불허)
- 주의
  - Q&A에 과제 의도 관련 질문만 가능
  - 과제 해결과 관련된 코드 사용법 질문 불허
  - 부정행위 적발 시 F