기말과제

정보처리 및 자연어처리

홍정하

과제 작성 및 제출 시 주의사항

- 제출기한: 6월24일(금) 23시59분(지각제출 불허)
- 개요
  - 데이터파일 train.txt, test.txt을 대상으로
  - NaiveBayesClassifier를 이용하여 accuracy가 높은 분류 모형 만들기
- 배점: 총 15점(평가 I + 평가 II)
- 제출 형식: 보고서 형식 (hwp, doc, pdf)
  - 과제를 수행한 일련의 코드 및 출력결과를 캡쳐하여 보고서에 수록(충분히 식별 가능한 크기로 삽입할 것)
  - 과제에서 요구하는 사항에 대해 문장으로 기술
- 강의에서 다루지 않은 코드/기능 사용 가능(단, 모듈은 re, os, nltk, Hangul.py 5개의 모듈만 사용 가능)

과제의 데이터 파일

- train.txt, test.txt
  - 문자코드: UTF-8
  - 2018년 네이버 NLP Challenge 개체명 인식 데이터
  - train.txt(638,139 라인), test.txt(425,432 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      -
```

- 태그: 비개체명

	개체명 범주	태그	정의
1	PERSON	PER	실존, 가상 등 인물명에 해당 하는 것
2	FIELD	FLD	학문 분야 및 이론, 법칙, 기술 등
3	ARTIFACTS_WORKS	AFW	인공물로 사람에 의해 창조된 대상물
4	ORGANIZATION	ORG	기관 및 단체와 회의/회담을 모두 포함
5	LOCATION	LOC	지역명칭과 행정구역 명칭 등
6	CIVILIZATION	CVL	문명 및 문화에 관련된 용어
7	DATE	DAT	날짜
8	TIME	TIM	시간
9	NUMBER	NUM	숫자
10	EVENT	EVT	특정 사건 및 사고 명칭과 행사 등
11	ANIMAL	ANM	동물
12	PLANT	PLT	식물
13	MATERIAL	MAT	금속, 암석, 화학물질 등
14	TERM	TRM	의학 용어, IT관련 용어 등 일반 용어를 총칭

평가 I: 기계학습 절차 적절성(배점 10점)

- 기계학습 절차를 충분히 숙지하고 적절한 코드를 사용하고 있는가?
  - train.txt을 train set으로, test.txt을 test set으로 하여(development set은 고려하지 말 것)
  - 기계학습에 필요한 일련의 절차를 문장으로 기술하고, 사용한 코드 및 출력결과(전체 또는 일부)를 적절히 제시
  - 이미 train set과 test set으로 구분이 되어 있으므로 shuffling 불필요.
- 정확도가 가장 높은 최적의 모형을 선택하기 위한 일련의 과정과 이유를 제시하시오.
  - 한국어 형태소분석 사용 금지
  - 그 밖의 다양한 방식의 feature extraction을 수행하여 어떻게 정확도가 가장 높은 최적의 모형을 선택하게 되었는지를 과정과 이유를 제시하시오.
- 채점기준
  - 절차의 적절성 위반 하나당 -1점
  - 코드/출력결과의 적절성 위반 하나당 -1점

평가 II: 정확도(배점: 5점)

- 정확도가 높은가?
  - 다양한 feature extraction을 통해 선택한 feature들을 통해 최선의 정확도를 test set으로 제시

- 채점 기준: accuracy 기준에 따라 평가
  - 87% 이상 5점
  - 86% 이상 4점
  - 84% 이상 3점
  - 81% 이상 2점
  - 그 미만 1점