3차과제

정보처리및 자연어처리

홍정하

과제 작성 및 제출 시 주의사항

- 제출기한: 6월22일(화) 23시59분(지각제출 불허)
- 개요
  - 데이터파일 entity_data_utf8.txt을 대상으로
  - NaiveBayesClassifier를 이용하여 accuracy가 높은 분류 모형 만들기
- 배점: 총 20점(평가 I + 평가 II + 평가 III)
- 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, konlpy 4개의 모듈만 사용 가능)

과제의 데이터 파일

- entity_data_utf8.txt
  - 문자코드: UTF-8
  - 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	-
```

	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 6점)

- 기계학습 절차를 충분히 숙지하고 적절한 코드를 사용하고 있는가?
  - 데이터파일의 어절단위 또는 문장단위 90%분량을 train set으로, 나머지 10%를 test set으로 분할(development set은 고려하지 말 것)
  - 기계학습에 필요한 일련의 절차를 문장으로 기술하고, 사용한 코드 및 출력결과(전체 또는 일부)를 적절히 제시
- 채점기준
  - 절차의 적절성 위반 하나당 -1점
  - 코드/출력결과의 적절성 위반 하나당 -1점

평가 II: feature extraction(배점: 8점)

- 다양한 측면에서 feature extraction을 수행 및 검토하고 있는가?
  - 다음 1~4 항목 중 3가지 항목을 선택하여 accuracy를 측정하여 제시
    1. n-gram: 어절/음절/자음모음
    2. 한국어형태소분석
    3. 한국어자음모음분리/합성
    4. 위치/길이: 어절/음절/자음모음
  - 주의: 1의 항목을 여러 측면에서 제시 가능, 예를 들어 어절에 대한 tri-gram, bi-gram, 또는 어절과 음절의 bi-gram을 제시 가능. 그러나 동일 항목 선택으로 간주됨.
- 채점기준
  - 3가지 항목에 대한 feature extraction 준수 위반 하나당 -2점
  - 코드/출력결과의 적절성 위반 하나당 -1점

평가 III: 정확도(배점: 6점)

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

- 채점 기준: accuracy 기준에 따라 평가
  - 85% 이상 6점
  - 83% 이상 5점
  - 81% 이상 4점
  - 79% 이상 3점
  - 77% 이상 2점
  - 그 미만 1점