Project #2

M1505.001600 정보모델링기법과 응용

2019년도 봄학기

제출기한: **2019년 5월 20일 23:55까지**

1. (100점) 주어진 질의어와 관련이 높은 순서대로 문서들을 나열하는 검색 엔진 모듈을 python whoosh 라이브러리를 사용하여 구현하시오.
   - 사용 데이터: WebAP 데이터셋
     - document.json : 9665개 문서 파일
     - query.json : 80개 질의어 파일
     - relevance.json : 각 질의어의 실제 연관 문서가 명시된 정답 파일
   - 작성 모듈
     - QueryResult.py: 텍스트 형태의 질의어를 입력 받아 whoosh 질의어 객체로 변환 후 검색 결과를 반환.
     - CustomScoring.py: 문서들을 질의어와 관련 높은 순서대로 나열할 때 사용하는 문서 채점 함수. 사용 가능한 기본 정보로는
       ✓ 문서 내 단어 빈도(TF)
       ✓ 역문서 빈도(IDF)
       ✓ 전체 데이터셋 내 단어 빈도
       ✓ 문서 개수
       ✓ 문서 길이(단어 개수)
       ✓ 전체 데이터셋 내 단어 개수
       ✓ 문서 당 평균 단어 개수
       등이 있으며 제공되는 정보 이외의 정보를 추출 가능하다면 추가적으로 사용 가능.
   - 평가 방법
     - 80개 질의어 중 임의로 정해진 15개의 test 질의어에 대한 검색 성능 평가
     - 평가 지표로는 BPREF를 사용

$$\text{BPREF} = \frac{1}{R}\sum_{d_r}\left(1 - \frac{N_{d_r}}{R}\right)$$

       - d_r은 연관 문서
       - N_{d_r}은 d_r보다 높게 랭크된 비연관 문서의 수
     - 점수 산정: 100점 중 성능 평가 점수 80점(15개 test 질의어의 평균 BPREF * 80), 성능 개선 방법에 대한 근거와 정당화 내용 20점
   - 주의 사항
     - 제공되는 질의어에 test 질의어까지 포함된 프로젝트 상황 특성을 활용한 질의어-문서 매칭 금지(직접적인 질의어-문서 매칭금지)
     - 문서 분석은 허용, 질의어 분석은 금지

2. (80점) 노래의 가사를 이용한 장르를 분류, 군집화하는 모델을 각각 구현하시오

2-1. (60점) 노래 가사를 이용한 장르 분류

- Country, Jazz, Pop, R&B, Rock 총 5개의 category
- 각 category마다 약 800~900개의 가사 데이터 제공
- Lyrics_train.csv 에 training을 위한 모든 데이터 존재. 여기서 Train/Validation set을 나눠 Cross-Validation에 활용하면 됨.
- 파일은 Song(노래제목), lyrics(가사), genre(장르)로 구성됨
- 주어진 데이터 외에 추가로 크롤링하여 활용해도 되지만 장르와 직접적 연관성이 존재하는 데이터의 경우는 금지 (가수명 등)
- 테스트는 각 category별 150~200개의 노래에 대해 진행

2-1-1. (30점) Naïve Bayes Classifier

- 점수 = 테스트 셋에서의 장르 분류 정확도(Accuracy)

2-1-2. (30점) SVM

- 점수 = 테스트 셋에서의 장르 분류 정확도(Accuracy)

- 각 모델에 대해 학습시킨 모델을 **pickle 파일로 저장하여** 코드와 함께 제출
- 제공된 데이터를 전처리 과정을 거쳐 일부만 training에 사용했다면, 전처리 과정이 포함된 코드도 제출하고, 보고서 상에 명확하게 과정을 명시 (크롤링하여 데이터 추가한 경우 추가 데이터도 같이 제출)

2-2. (20점) 노래 군집화

- K-means Clustering
- 분류와 같은 데이터 사용
  보고서에 군집화 결과에 대한 분석 포함 필수
- 평가 지표로는 V-measure 사용
  V-measure = homogeneity와 completeness의 조화 평균
  - Homogeneity: 각 군집이 단일 클래스의 데이터만 가지는 정도

$$\mathrm{h} = 1 - \frac{H[C|K]}{H[C]}$$

  - Completeness: 같은 클래스의 값이 하나의 군집으로 모여있는 정도

$$\mathrm{c} = 1 - \frac{H[K|C]}{H[K]}$$

  - V-measure

$$\mathrm{v} = \frac{2hc}{h + c}$$

  $H[C]$: 클래스 엔트로피, 여러 클래스에 분산되어 있을수록 큰 값
  $H[C|K]$: 군집화 종료 후 클래스 엔트로피
  $H[K]$: 군집 엔트로피, 여러 군집에 분산되어 있을수록 큰 값
  $H[K|C]$: 클래스 별로 분류한 후의 군집 엔트로피
- 분류에서 test 데이터로 사용한 데이터까지 모두 포함하여 측정한 V-measure * 20으로 점수 부여

- 보고서 점수 10점
- 발표자료 및 발표 점수 10점

eTL에 업로드된 예시 코드를 응용하시오. 발표자료는 PPT나 PDF의 파일 형태로, 보고서는 Word나 PDF로, python 코드 결과물과 함께 압축하여 eTL 사이트에 제출하시오.