18 - 2 인공지능 과제 2

영화 리뷰 긍정/부정 분류하기

조교 오주민
ojm9898@hanyang.ac.kr

과제 개요

- **Naive Bayes Classification 기법을 이용하여 네이버 영화 리뷰 데이터셋 긍정/부정 분류하기**

- **세부사항**
  - 네이버 영화 리뷰 데이터셋의 Raw sentence를 가공하는 방법(Tokenizer 또는 형태소 분석 등)의 선택은 개인 자유.
  - 분류기의 성능을 높이기 위한 최적화 기법 적용가능.
  - 긍정 / 부정 분류 정확도로 점수 산출

데이터셋 예시

- **document(Raw sentence) / 긍정 · 부정 label** (0 = 부정, 1 = 긍정)

```
id  document    label
8112052 어릴때보고 지금다시봐도 재밌어요ㅋㅋ    1
8132799 디자인을 배우는 학생으로, 외국디자이너와 그들이 일군 전통을 통해 발전해가는 문화산업이 부러웠는
4655635 폴리스스토리 시리즈는 1부터 뉴까지 버릴께 하나도 없음.. 최고.   1
9251303 와.. 연기가 진짜 개쩔구나.. 지루할거라고 생각했는데 몰입해서 봤다.. 그래 이런게 진짜 영화지 1
10067386    안개 자욱한 밤하늘에 떠 있는 초승달 같은 영화.  1
2190435 사랑을 해본사람이라면 처음부터 끝까지 웃을수 있는영화   1
9279041 완전 감동입니다 다시봐도 감동   1
7865729 개들의 전쟁2 나오나요? 나오면 1빠로 보고 싶음   1
7477618 굿  1
9250537 바보가 아니라 병 쉰 인듯    1
9730759 내 나이와 같은 영화를 지금 본 나는 감동적이다..하지만 훗날 다시보면대사하나하나 그 감정을완벽하
640794  재밌다  1
9537008 고질라니무 귀엽다능ㅋㅋ 1
4911311 영화의 오페라화라고 해야할 작품. 극단적 평갈림은 어쩔 수 없는 듯.   1
6686673 3도 반전 좋았제 ^^  1
9034036 평점 왜 낮아? 긴장감 스릴감 진짜 최고인데 진짜 전장에서 느끼는 공포를 생생하게 전해준다.    1
979683  네고시에이터랑 소재만 같을 뿐.. 아무런 관련없음..   1
165498  단연 최고   1
8703997 가면 갈수록 더욱 빠져드네요 밀회 화이팅!!   1
9468781 어?생각없이 봤는데 상당한 수작.일본영화 10년내 최고로 마음에 들었다.강렬한 임팩트가 일품.   1
5185638 오랜만에 본 제대로 된 범죄스릴러~   1
10221267    그런 때가 있었다. ('사랑해'도 아니고) 그저 좋아한다는 한 마디 말을 꺼내기도 벅차서 밤 잠 
```

ratings_train.txt (label있음)

ratings_valid.txt (label있음)

결과물 예시

- **ratings_train.txt으로 나이브-베이즈 모델을 학습하여 ratings_test.txt의 label 예측하여 예측 결과를 ratings_result.txt에 저장.**

나눠줄 파일 예시

ratings_test.txt

(label 없음)

```
id  document    label
6270596 굳 ㅋ
9274899 GDNTOPCLASSINTHECLUB
8544678 뭐야 이 평점들은.... 나쁘진 않지만 10점 짜리는 더더욱 아니잖아
6825595 지루하지는 않은데 완전 막장임... 돈주고 보기에는....
6723715 3D만 아니었어도 별 다섯 개 줬을텐데.. 왜 3D로 나와서 제 심기를 불편하게 하죠??
7898805 음악이 주가 된, 최고의 음악영화
6315043 진정한 쓰레기
6097171 마치 미국애니에서 튀어나온듯한 창의력없는 로봇디자인부터가,고개를 젖게한다
```

제출할 파일 예시

ratings_result.txt

(label 있음)

```
id  document    label
6270596 굳 ㅋ 1
9274899 GDNTOPCLASSINTHECLUB 0
8544678 뭐야 이 평점들은.... 나쁘진 않지만 10점 짜리는 더더욱 아니잖아 0
6825595 지루하지는 않은데 완전 막장임... 돈주고 보기에는.... 0
6723715 3D만 아니었어도 별 다섯 개 줬을텐데.. 왜 3D로 나와서 제 심기를 불편하게 하죠?? 0
7898805 음악이 주가 된, 최고의 음악영화 1
6315043 진정한 쓰레기 0
6097171 마치 미국애니에서 튀어나온듯한 창의력없는 로봇디자인부터가,고개를 젖게한다 0
```

과제 조건

- **환경**
  - 프로그래밍 언어 : **Python 3.x**
  - OS : **Ubuntu 16.04 LTS**
  - 데이터 가공의 목적을 제외한 외부 라이브러리 사용 **불가**(내장함수만 사용 가능)

- **제출 사항**
  - 파이썬 파일 : **본인학번_assignment_2.py**
  - 결과물 **: ratings_result.txt**
  - 결과 보고서 : **본인학번_assignment_2.docx**
    - 코드 설명
    - 실험 결과

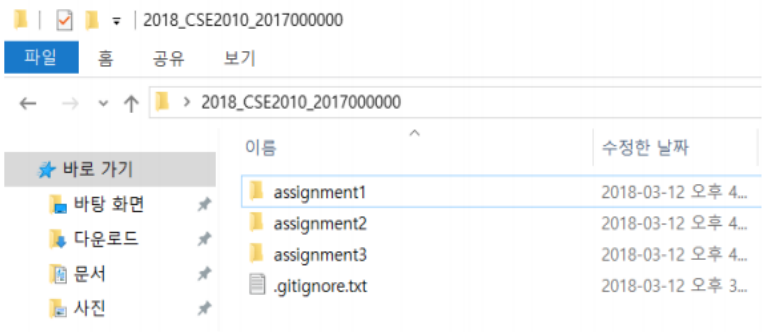
주의 사항

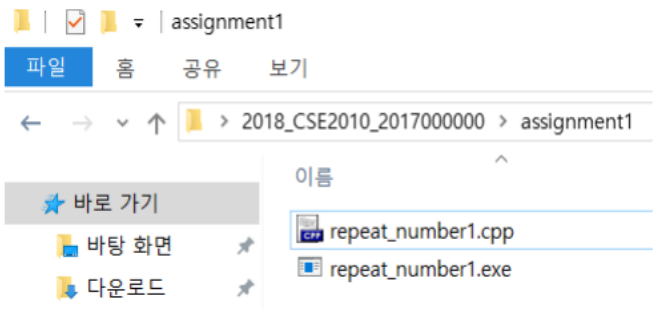
- **파일명** 반드시 준수.
- 코드와 같은 경로에 **ratings_train.txt, ratings_valid.txt, ratings_test.txt, ratings_result.txt** 파일 관리.
- 파일은 **GitLab**에 올려주세요.
- 제출 기한 : **2018.11.23.**
- **추가 제출 기한 없음.**
- 점수 비중 : **코드 70% 보고서 30%**

주의 사항

- **파일은 GitLab에 올릴 것!**
  - **경로** : (GitLab init 경로) – (assignment2) – [파일]
  - **파일명 :** ex) 본인학번_assignment_2.py
  - GitLab 업로드 시, 빈 디렉토리가 존재하지 않도록 할 것!

- 프로젝트는 아래 그림과 같이 관리!
- 소문자 assignment1, assignment2, … 로 폴더를 만들어 과제 제출





Thank you!