Homework Assignment hw8

보고서 및 논문 윤리 서약

1. 나는 보고서 및 논문의 내용을 조작하지 않겠습니다.
2. 나는 다른 사람의 보고서 및 논문의 내용을 내 것처럼 무단으로 복사하지 않겠습니다.
3. 나는 다른 사람의 보고서 및 논문의 내용을 참고하거나 인용할 시 참고 및 인용 형식을 갖추고 출처를 반드시 밝히겠습니다.
4. 나는 보고서 및 논문을 대신하여 작성하도록 청탁하지도 청탁받지도 않겠습니다.

나는 보고서 및 논문 작성 시 위법 행위를 하지 않고, 명지인으로서 또한 공학인으로서 나의 양심과 명예를 지킬 것을 약속합니다.

학　　과 : 융합소프트웨어학부 데이터테크놀로지전공
과　　목 : 인공지능
담당교수 : 전종훈
강좌 번호: 6019
학　　번 : 60201901
이　　름 :　　권성중　　(서명)

a)

loading 후 data와 label을 분리했습니다.

```
from sklearn.datasets import load_files
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.metrics import accuracy_score
import numpy as np
from sklearn.ensemble import RandomForestClassifier
from sklearn import metrics

# a) load_files 이용해서 train,test로 분리해서 loading 후 data와 label 분리
reviews_train = load_files("/Users/wnd180/Downloads/aclImdb/train")
text_train,y_train = reviews_train.data, reviews_train.target

reviews_test = load_files("/Users/wnd180/Downloads/aclImdb/test")
text_test, y_test = reviews_test.data, reviews_test.target
```

b)

```
# b) br 삭제 후 일부 출력하기
print("테스트 데이터 문서의 수 : ", len(text_test))
print("클래스별 샘플 수 (테스트 데이터) : ", np.bincount(y_test))

print("삭제 전")
print(text_train[0:2])
text_train = [doc.replace(b"<br />",b" ")for doc in text_train]
text_test = [doc.replace(b"<br />",b" ")for doc in text_test]

print("삭제 후")
print(text_train[0:2])
# 삭제전
# [b"Zero Day leads you to think, even re-think why two boys/young men would
do what they did - commit mutual suicide via slaughtering their classmates. It
captures what must be beyond a bizarre mode of being for two humans who have
decided to withdraw from common civility in order to define their own/mutual
world via coupled destruction.<br /><br />It is not a perfect movie but given
what money/time the filmmaker and actors had - it is a remarkable product. In
terms of explaining the motives and actions of the two young suicide/murderers
it is better than 'Elephant' - in terms of being a film that gets under our
'rationalistic' skin it is a far, far better film than almost anything you are
likely to see. <br /><br />Flawed but honest with a terrible honesty.",
b'Words can\'t describe how bad this movie is. I can\'t explain it by writing
only. You have too see it for yourself to get at grip of how horrible a movie
really can be. Not that I recommend you to do that. There are so many
clich\xc3\xa9s, mistakes (and all other negative things you can imagine) here
that will just make you cry. To start with the technical first, there are a
LOT of mistakes regarding the airplane. I won\'t list them here, but just
```

```
mention the coloring of the plane. They didn\'t even manage to show an
airliner in the colors of a fictional airline, but instead used a 747 painted
in the original Boeing livery. Very bad. The plot is stupid and has been done
many times before, only much, much better. There are so many ridiculous
moments here that i lost count of it really early. Also, I was on the bad
guys\' side all the time in the movie, because the good guys were so stupid.
"Executive Decision" should without a doubt be you\'re choice over this one,
even the "Turbulence"-movies are better. In fact, every other movie in the
world is better than this one.']
# 삭제 후
# [b"Zero Day leads you to think, even re-think why two boys/young men would
do what they did - commit mutual suicide via slaughtering their classmates. It
captures what must be beyond a bizarre mode of being for two humans who have
decided to withdraw from common civility in order to define their own/mutual
world via coupled destruction.  It is not a perfect movie but given what
money/time the filmmaker and actors had - it is a remarkable product. In terms
of explaining the motives and actions of the two young suicide/murderers it is
better than 'Elephant' - in terms of being a film that gets under our
'rationalistic' skin it is a far, far better film than almost anything you are
likely to see.   Flawed but honest with a terrible honesty.", b'Words can\'t
describe how bad this movie is. I can\'t explain it by writing only. You have
too see it for yourself to get at grip of how horrible a movie really can be.
Not that I recommend you to do that. There are so many clich\xc3\xa9s,
mistakes (and all other negative things you can imagine) here that will just
make you cry. To start with the technical first, there are a LOT of mistakes
regarding the airplane. I won\'t list them here, but just mention the coloring
of the plane. They didn\'t even manage to show an airliner in the colors of a
fictional airline, but instead used a 747 painted in the original Boeing
livery. Very bad. The plot is stupid and has been done many times before, only
much, much better. There are so many ridiculous moments here that i lost count
of it really early. Also, I was on the bad guys\' side all the time in the
movie, because the good guys were so stupid. "Executive Decision" should
without a doubt be you\'re choice over this one, even the "Turbulence"-movies
are better. In fact, every other movie in the world is better than this
one.'])
```

Br이 사라진 것이 확인이 가능하다.

c) 불용어 제거하기

```
# c) countvectorizer를 사용해 BOW 생성과 동시에 english stop word list를 이용해
불용어 제거
vect = CountVectorizer(stop_words="english").fit(text_train)
X_train = vect.transform(text_train)
X_test = vect.transform(text_test)
```

d) Random Forest 방식을 사용하여 분류 모델을 train 시키고, 이를 이용하여 prediction을 실행하고 코드의 실행 시간과 정답률을 출력하시오.

Base classifier는 n_estimators = 100이다.

```
# d) RandomForest 방식을 사용하여 분류 모델을 train 시키고,이를 이용하여 prediction 을 실행하고 코드의 실행시간과 정답률을 출력하시오
import time
start = time.time()
# base classifier -> n_estimators= 100
clf = RandomForestClassifier(n_estimators=100)
clf.fit(X_train,y_train)
predict = clf.predict(X_test)
ac_score = metrics.accuracy_score(y_test, predict)
print("수행 시간 :",round(time.time()-start,1))
print("정답률 : ",round(ac_score,1))
```

e) 모델 향상을 위해 n_estimator를 늘렸다.

```
clf = RandomForestClassifier(n_estimators=500)
clf.fit(X_train,y_train)
predict = clf.predict(X_test)
ac_score = metrics.accuracy_score(y_test, predict)
print("정답률 : ",round(ac_score,1))
```