Homework Assignment Hw 8

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 : 60182196

이 름 : 이동혁 (서명)

1.

(a).

```
from sklearn.datasets import load_files
import numpy as np

reviews_train = load_files("/Users/leedonghyeok/Downloads/aclImdb/train")
reviews_test = load_files("/Users/leedonghyeok/Downloads/aclImdb/test")

text_train, y_train = reviews_train.data, reviews_train.target
text_test, y_test = reviews_test.data, reviews_test.target

print("done")
```

```
done
```

(b).

```
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("done")

# 전(일부) : doesn't hurt either.<br /><br />Stargate SG1
# 후(일부) : doesn't hurt either.  Stargate SG1
```

```
done
```

(c).

```
from sklearn.feature_extraction.text import CountVectorizer

vect = CountVectorizer(stop_words = 'english').fit(text_train)

X_test = vect.transform(text_test)

print("done")
```

```
done
```

(d).

```
from sklearn.ensemble import RandomForestClassifier
from sklearn import metrics, model_selection
from time import time

start = time()

clf = RandomForestClassifier(n_estimators = 100)
clf.fit(X_test, y_test)

scores = model_selection.cross_val_score(clf, X_test, y_test, cv = 10)

end = time()

print("Execution time(seconds)", str(round((end-start), 2)))
print("각 validation 정답률 = ", scores)
print("평균 정답률 :", "%.2f"%scores.mean())
```

```
Execution time(seconds) 407.48
각 validation 정답률 =  [0.8728 0.8584 0.8532 0.8768 0.8516 0.8592 0.8588 0.8568 0.8372 0.856 ]
평균 정답률 : 0.86
```

(e).

```
from sklearn.ensemble import RandomForestClassifier
from sklearn import metrics, model_selection
from time import time

start = time()

clf = RandomForestClassifier(n_estimators = 10)
clf.fit(X_test, y_test)

scores = model_selection.cross_val_score(clf, X_test, y_test, cv = 10)

end = time()

print("Execution time(seconds)", str(round((end-start), 2)))
print("각 validation 정답률 = ", scores)
print("평균 정답률 :", "%.2f"%scores.mean())
```

```
Execution time(seconds) 40.56
각 validation 정답률 =  [0.782  0.776  0.7852 0.7784 0.7716 0.7636 0.7816 0.7752 0.7612 0.7644]
평균 정답률 : 0.77
```

```
from sklearn.ensemble import RandomForestClassifier
from sklearn import metrics, model_selection
from time import time

start = time()

clf = RandomForestClassifier(n_estimators = 3)
clf.fit(X_test, y_test)

scores = model_selection.cross_val_score(clf, X_test, y_test, cv = 3)

end = time()

print("Execution time(seconds)", str(round((end-start), 2)))
print("각 validation 정답률 = ", scores)
print("평균 정답률 :", "%.2f"%scores.mean())
```

```
Execution time(seconds) 3.52
각 validation 정답률 =  [0.70266379 0.69914797 0.7023881 ]
평균 정답률 : 0.70
```

Classifier을 늘릴수록 더 높은 정답률이 나왔지만 시간은 더 걸렸다.