데이터 사이언스 Term Project Report

컴퓨터소프트웨어학부

2016024839 박정민

1. 코드 구성

주어진 training dataset을 pandas dataframe으로 받아서 저장하고, dataframe을 바탕으로 여러 함수를 통해 prediction을 수행하는 Recommender class를 만들었습니다.

```
class Recommender:

    def __init__(self, test):
        self.test = test
```

test : test 파일의 이름입니다.

2. Class 내 method

```
# Read Test Data
def readTestData(self):
    self.testdf = pd.read_table(self.test, sep='\t', header=None, names=['user_id', 'item_id', 'rating', 'time_stamp'])

    self.testdf.drop('time_stamp', axis=1, inplace=True)
    self.testdf.drop('rating', axis=1, inplace=True)
```

readTestData(self)

test file을 읽어 dataframe으로 받아오는 함수입니다. 클래스의 멤버로 test dataframe을 저장합니다. 그리고 불필요한 time_stamp, rating column을 제거합니다.

```python
# Read Train Data
def readTrainData(self, input):
    self.df = pd.read_table(input, sep='\t', header=None, names=['user_id', 'item_id', 'rating', 'time_stamp'])
    self.size = len(self.df)
    self.df.drop('time_stamp', axis=1, inplace =True)

    self.user_movie_rating = self.df.pivot_table('rating', index='user_id', columns='item_id')

    self.user_movie_rating.fillna(0, inplace=True)

    self.movie_user_rating = self.user_movie_rating.values.T

    SVD = TruncatedSVD(n_components=12)
    matrix = SVD.fit_transform(self.movie_user_rating)

    self.corr = np.corrcoef(matrix)
```

readTrainData(self, input)

train dataset을 읽어 dataframe으로 가져옵니다. 가져온 data를 user-rating matrix로 변환하여 각 item간의 상관관계 계수를 나타내는 matrix를 구합니다.

```python
# Predict
def predict(self):
    ratepredict = []

    for tup in self.testdf.values:
        user_id = tup[0]
        item_id = tup[1]

        if item_id not in list(self.user_movie_rating.columns.values):
            ratepredict.append(random.randint(1,6))

        else:
            corr_idx = list(self.user_movie_rating.columns.values).index(item_id)
            rating = self.user_movie_rating.loc[user_id]

            div = 0
            rate_predict = 0
            for idx, rate in enumerate(rating.values):

                if rate != 0:
                    div += 1
                    corr = self.corr[corr_idx][idx]
                    if corr >= 0.5 :
                        rate_predict += rate
                    elif corr > 0 and corr < 0.5:
                        rate_predict += rate * corr
                    elif corr == 0:
                        rate_predict += random.randint(1, 5)
                    elif corr < 0:
                        rate_predict += 5 / rate

            ratepredict.append(rate_predict / div)

    ratepredict = np.array(ratepredict)
    self.testdf['rating'] = ratepredict
```

predict(self)

위 함수에서 구한 상관계수 matrix를 이용하여 특정 user의 특정 item의 rating을 예측하는 함수입니다. 해당 user가 평가한 아이템들과 예측하려는 item간의 상관계수에 따라 rating을 예측하였습니다. 상관계수가 0.5 이상이면 별점을 그대로 반영하였고, 0.5 이하면 점수에다가 상관계수를 곱한값을 더하였습니다. 계수가 0이면 랜덤으로 1~5사이의 값을 더하였습니다. 음수면 반대의 상관관계가 있다는 의미이므로 반비례한 값을 더하였습니다. 더한 횟수로 나누어 평균을 구하였습니다.

```python
# Write Output File
def writeFile(self):
    fileNum = self.test[1]
    fileName = "u" + fileNum + ".base_prediction.txt"

    self.testdf.to_csv(fileName, header=None, index=False, sep='\t')
```

writeFile(self)

예측한 결과를 text파일로 출력하는 함수입니다.

3. 컴파일 환경 및 실행방법

Python 3.92버전을 사용하였고 launch.json파일을 이용하여 training set과 test set을 파일 실행 파라미터로 전달하였습니다.

```jsonc
{
    // Use IntelliSense to learn about possible attributes.
    // Hover to view descriptions of existing attributes.
    // For more information, visit: https://go.microsoft.com/fwlink/?linkid=830387
    "version": "0.2.0",
    "configurations": [
        {
            "name": "Python: Current File",
            "type": "python",
            "request": "launch",
            "program": "${file}",
            "console": "integratedTerminal",
            "args": [
                "u2.base",          //training data
                "u2.test"           //test data
            ]
        }
    ]
}
```

```
PS G:\내 드라이브\데이터사이언스\TermProject> g:; cd 'g:\내 드라이브\데이터사이언스\TermProject'; & 'C:\Users\pmw14\AppData\Local\Programs\Python\Python39\python.exe' 'c:\Users\pmw14\.vscode\extensions\ms-python.python-2021.5.842923320\pythonFiles\lib\python\debugpy\launcher' '58293' '--' 'g:\내 드라이브\데이터사이언스\TermProject\recommender.py' 'u2.base' 'u2.test'
```

파일 실행 스크린샷입니다.