[프로젝트] P02 영화추천시스템

영화추천시스템 (총 100 점)

MovieLens 100k 데이터를 다양한 방법으로 분석해보세요.
MovieLens 100k 데이터는 다음의 주소에서 다운로드 받을 수 있습니다.
https://grouplens.org/datasets/movielens/100k/

하나의 ipynb 파일에 다음 Task 들을 모두 수행하여 제출하세요.
입력파일경로는 각각 "./ua.base", "./ua.test", "./u.item"으로 설정하길 바랍니다.
파일명은 "p02_이름_학번.ipynb"로 하여 제출하세요.

- Task 1. (10 점) 데이터 준비하기
    - Task 1-1. (5 점) 파일을 다운로드 받고 ua.data 파일(학습데이터)과 ua.test 파일(검증데이터)의 내용을 불러와서 Tensor 데이터 생성하기
    - Task 1-2. (5 점) u.item 파일로부터 영화 id 와 title 불러오기

- Task 2. (20 점) Latent Factor 모델을 이용하여 학습하기
    - Task 2-1. (3 점) P, Q, bias_user, bias_item 등 파라미터 초기화하기
    - Task 2-2. (7 점) regularization 과 bias 적용하여 가설, 비용 설정하기
    - Task 2-3. (5 점) torch.optim 을 사용하여 학습하기
    - Task 2-4. (5 점) 학습데이터와 검증데이터에 대해서 각각 RMSE 값을 구하여 출력하기 (training RMSE, test RMSE)

- Task 3. (10 점) 13 번 User 에게 추천하기 (knn search)
    - Task 3-1. (5 점) 13 번 user 의 예상 별점이 가장 높은 영화 top 20 개를 찾아서 id 및 영화이름 출력하기
    - Task 3-2. (5 점) Latent Matrix P 와 Q 를 이용하여 13 번 user 와 cosine similarity 가 가장 유사한 영화 top 20 개를 찾아서 id 및 영화이름 출력하기

- Task 4. (30 점) 영화 클러스터링하기 (k-means clustering)
    - Task 4-1. (15 점) 다음을 만족하는 k-means clustering 알고리즘 구현하기

        - 각 영화가 속한 cluster 를 정할 때, cosine similarity 를 기준으로 정하기
    - Task 4-2. (10 점) k=1, ..., 40 까지 바꿔가면서 cost 값을 계산하고 이를 matplotlib 을 활용하여 그래프로 그리기
    - Task 4-3. (5 점) 가장 적절해보이는 k 선택하기

- Task 5. (30 점) 차원 축소 및 시각화 (PCA)
    - Task 5-1. (5 점) P 행렬와 Q 행렬을 합쳐 Z 행렬 만들기
    - Task 5-2. (10 점)Z 행렬에서 PCA 수행하여 2 차원 데이터로 줄인 Zp 만들기
        - 참고: 외부 library 를 사용해도 무방함
    - Task 5-3. matplotlib 을 활용하여 Zp 의 scatter plot 그리기
        - Task 5-3-1. (5 점) P 행렬과 Q 행렬의 점들을 서로 다른 색으로 그리기
        - Task 5-3-2. (5 점) Task 3 의 결과 점들을 다른 색으로 그려 강조하기
        - Task 5-3-3. (5 점) Task 4 에서 구한 cluster 들을 각기 다른 색으로 그리기