텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

위의 코드는 실습에서 다루었던 유사도가 높은 10개의 영화를 추천해주는 함수의 위쪽 일부분으로 함수를 변형한 부분입니다.

우선 내림차순으로 정렬한 이후에 “cos\_sim = cos\_sim[0:11]” 이 코드로 상위 11개의 유사도를 가진 영화들을 뽑아내었다는 건 이해했습니다. 이어서 Return 아래에 주석처리 된 부분의 역할이 자기 자신이 포함되었다면 이를 제거하고 상위 10개의 영화를 추출해서 반환하는 것으로 이해했습니다.

그런데 아래의 사진에서 볼 수 있듯 애초에 cos\_matrix가 각 영화가 서로에 대한 유사도를 갖는 ndarray이고 각 영화는 자기 자신에 대해 유사도가 1임을 확인할 수 있었습니다.

텍스트, 스크린샷, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

그렇다면

cos\_sim.sort(key=lambda x :x[1], reverse=True)

이 코드를 통해 내림차순 했을 때 자기 자신에 대한 유사도는 1이므로 제일 상단에 자기 자신에 해당하는 인덱스가 오고 유사도는 1이어야 할 것이라 생각했습니다.

하지만 (return cos\_sim을 사용해서) 주어진 title에 대해 상위 11개의 유사도를 가진 영화를 인덱스와 유사도의 튜플로 반환하도록 했을 때아래와 같은 결과를 얻었습니다.

텍스트, 스크린샷, 폰트, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

indices["The City of Lost Children"]

의 결과로 28이 출력되기 때문에 유사도 상위 11개의 데이터를 출력했을 때 첫 번째로 28이 온 것은 예상한 것과 일치하는 듯 보입니다.

텍스트, 스크린샷, 소프트웨어, 멀티미디어 소프트웨어이(가) 표시된 사진

자동 생성된 설명

다른 영화 타이틀로 결과를 출력해본 바는 위와 같습니다. 제일 상단의 영화가 자기 자신인 것은 항상 동일하였습니다.

질문1. 유사도가 1이 아니고 0.99999 이런 식으로 표현된 것이 이해가 되지 않습니다.

(chatGPT에 활용한 결과 코사인 유사도 값은 부동소수점 연산의 한계로 인해 정확한 1이나 -1이 아닌 값이 될 수 있다고 하여 이는 이해했습니다. 하지만 그렇다면 왜 cos\_matrix만 출력했던 부분에서는(위에서 2번째 사진) 정확하게 1의 값이 출력되었는지 궁금합니다)

질문2. 제가 예상했던, 자기 자신에 대해 유사도가 1이기 때문에 내림차순으로 정렬하면 자기 자신에 해당하는 인덱스와 유사도의 튜플이 제일 상위에 올라와야 한다는 것이 필연적인 것이 아니라 우연의 일치였는지도 궁금합니다. 우연의 일치가 아니라면

if idx in sim\_movie\_idx:

#         sim\_movie\_idx.remove(idx)

#     sim\_movie\_idx = sim\_movie\_idx[0:10]

# #     return sim\_movie\_idx

이 부분을 사용할 필요 없이 cos\_sim = cos\_sim[1:11]로 해당 작품을 제외할 수 있지 않을까 생각이 듭니다.

질문3. 상위 추천 영화의 대부분이 모두 1에 가깝게 수렴하는 것은 장르가 완전히 동일한 영화들에 대해서 코사인 유사도가 1에 가깝게 수렴하는 것인가요?