텍스트 마이닝과 데이터 마이닝

Part 09. 추천 시스템

정 정 민

programmers

Chapter 25. 휴대폰 추천 시스템 실습

1. 데이터 설명
2. 컨텐츠 기반 필터링 적용
3. 협업 필터링 적용

데이터 설명

핸드폰 추천 데이터

- 이번 실습에서 사용할 Kaggle 데이터 (링크)
  - 다운로드 받아주세요!

- 2022년 미국에서 가장 인기 있었던 휴대폰 관련 데이터
- 총 33종의 휴대폰 & brand, model, OS, 성능, 가격 등 정보 존재
  - 총 13개의 세부 정보 종류

- 사용자 평가 데이터
- 총 99명의 사용자에게 10개의 휴대폰이 제공되어, 구매 가능성(1~10)을 표시하도록 요청
  - 총 990개 데이터 포인트
- 이들의 나이, 성별, 직업 정보도 포함

컨텐츠 기반 필터링 적용

강의 실습 코드 : 링크

프로파일 구성

- 13개의 세부 정부 중, 분석에 사용할 특성을 선택
- 도메인 지식(?)과 구매 경험을 되살려 아래와 같은 특성을 프로파일로 구성
  - 브랜드, 모델, OS, 메모리, RAM, 가격
- Embedding 계획
  - 브랜드, 모델, OS : 텍스트 → TF-IDF 점수 적용
  - 메모리, RAM, 가격 : 숫자 → MinMax Scaling 값 활용

```
# 텍스트 데이터 tf-idf 처리
tfidf_vectorizer = TfidfVectorizer()
tfidf_matrix = tfidf_vectorizer.fit_transform(data['brand',
                                                   'model',
                                                   'operating system'])
# 수치형 데이터 min-max scailing 처리
scaler = MinMaxScaler()
scaled_features = scaler.fit_transform(data['internal memory',
                                            'RAM', 'price'])
# 최종 통합 프로파일 생성
combined_features = np.hstack((tfidf_matrix.toarray(),
scaled_features))
```

유사도 생성

- 생성한 핸드폰 프로파일끼리 유사도 생성
- Cosine similarity 활용

- 두 핸드폰 쌍의 유사도를 미리 구해둠 > Matrix 형태로 출력
  - 정사각 행렬 (num_cellphone x num_cellphone)
  - 대각 행렬은 자기 자신과의 유사도로 모두 1

```
# 코사인 유사도를 사용하여 항목 간 유사도 계산
cosine_sim = cosine_similarity(combined_features, combined_features)

# 행과 열의 수 : 사용 핸드폰 수
[[1.         0.50113071 0.59591399 ... 0.00248105 0.         0.24454722]
 [0.50113071 1.         0.83605019 ... 0.00457353 0.         0.03285111]
 [0.59591399 0.83605019 1.         ... 0.00457353 0.         0.03285111]
  ...,
 [0.00248105 0.00457353 0.00457353 ... 1.         0.46320185 0.40357208]
 [0.         0.         0.         ... 0.46320185 1.         0.48431097]
 [0.24454722 0.03285111 0.03285111 ... 0.40357208 0.48431097 1.        ]]
```

추천 결과 확인

- 목적 : 특정 핸드폰과 비슷한 추천
- 아래와 같은 과정으로 진행

1. 비교 대상이 되는 핸드폰을 선택
2. 유사도 행렬에서 다른 핸드폰과의 유사도를 가져옴
3. 유사도를 기반으로 상위 N개의 핸드폰을 선택
4. 선택된 핸드폰의 정보를 출력

선택한 휴대폰 :

	brand	model	operating system	internal memory	RAM	price
4	Apple	iPhone 13 Pro Max	iOS	256	6	1199

추천된 비슷한 휴대폰 :

	brand	model	operating system	internal memory	RAM	price
3	Apple	iPhone 13 Pro	iOS	256	6	999
2	Apple	iPhone 13	iOS	128	4	699
1	Apple	iPhone 13 Mini	iOS	128	4	699
28	Sony	Xperia Pro	Android	512	12	1998
0	Apple	iPhone SE (2022)	iOS	128	4	429

협업 필터링 적용

강의 실습 코드 : 링크

사용자 – 핸드폰 상호 작용 데이터

- 사용자가 핸드폰에 대해 rating을 진행한 파일을 사용 : “cellphones ratings.csv”

- 전체 33개의 핸드폰 중 10개를 제시 → 23개는 정보가 없음
- 이 부분은 0으로 채워넣기!

- 행에는 사용자 정보, 열에는 핸드폰 정보로 하는 matrix 생성
  - DF.pivot_table : 기존 DF에서 필요한 정보를 요약, 추출해 새로운 DF을 생성 (user_item_matrix)
  - 같은 사용자 ID를 기준으로 rating column의 값을 집계

	user_id	cellphone_id	rating
0	0	30	1
1	0	5	3
2	0	10	9
3	0	9	3
4	0	23	2
...	...	...	...
985	258	31	5
986	258	17	8
987	258	23	9
988	258	27	8
989	258	24	6

990 rows × 3 columns

cellphone_id	0	1	2	3	4	5	6	7	8	9	...	23	24	25	26	27	28	29	30	31	32
user_id																					
0	0.0	0.0	0.0	10.0	0.0	3.0	0.0	0.0	2.0	3.0	...	2.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
1	0.0	0.0	0.0	10.0	7.0	0.0	0.0	8.0	8.0	0.0	...	0.0	0.0	0.0	0.0	0.0	8.0	0.0	0.0	7.0	6.0
6	0.0	2.0	0.0	0.0	0.0	0.0	0.0	9.0	7.0	0.0	...	8.0	0.0	0.0	0.0	4.0	0.0	0.0	0.0	6.0	10.0
8	5.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	8.0	0.0	8.0	0.0	6.0	0.0	0.0	9.0
10	0.0	0.0	0.0	9.0	9.0	3.0	7.0	0.0	0.0	0.0	...	9.0	0.0	0.0	0.0	8.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
254	0.0	5.0	0.0	0.0	0.0	10.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	18.0	0.0	0.0	0.0	5.0	0.0	8.0
255	0.0	10.0	0.0	0.0	10.0	0.0	10.0	0.0	10.0	10.0	...	0.0	0.0	10.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
256	0.0	0.0	7.0	0.0	5.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	8.0	0.0	0.0	0.0	0.0	0.0	7.0	9.0
257	7.0	0.0	6.0	0.0	8.0	0.0	0.0	0.0	5.0	0.0	...	0.0	0.0	8.0	0.0	0.0	0.0	7.0	5.0	0.0	0.0
258	0.0	0.0	0.0	0.0	0.0	0.0	7.0	0.0	0.0	0.0	...	9.0	6.0	0.0	7.0	8.0	0.0	6.0	0.0	5.0	0.0

99 rows × 33 columns

사용자 기반 협업 필터링

- 사용자 기반으로 추천을 진행하려면 '이웃'을 생성해야 함
- 따라서 사용자 간 유사도를 구해야 함
  - 사용자 수 x 사용자 수 크기의 행렬 생성

- Cosine Similarity 유사도 활용

```
# 사용자 유사도 메트릭스 생성
user_similarity = cosine_similarity(user_item_matrix)


# 유사도 행렬을 DataFrame으로 변환
user_similarity = pd.DataFrame(user_similarity,
                               index=user_item_matrix.index,
                               columns=user_item_matrix.index)
```

아이템 기반 협업 필터링

- 반대로 아이템들 사이의 Cosine 유사도를 계산
- 핸드폰 수(33) x 핸드폰 수(33) 크기의 메트릭스 생성

```
# 아이템 - 사용자 메트릭스 생성
item_user_matrix = user_item_matrix.T

# 아이템 유사도 메트릭스 생성
item_similarity = cosine_similarity(item_user_matrix)

# 유사도 행렬을 DataFrame으로 변환
item_similarity_df = pd.DataFrame(item_similarity,
                                  index=item_user_matrix.index,
                                  columns=item_user_matrix.index)
```

E.O.D