KNOW 직업추천 대회
User-based CF로 분석

김지선

지난 내용

- 12월 말~ 2월까지 KNOW 직업추천 대회 참여
- 자체 구현한 추천 방식 시도 (완료)
- 낮은 성능에 CF 알고리즘 도입 후 분석 중 (진행중)

용어 정리

1. 재직자

- 설문조사에 답한 사람
- train 데이터의 한 행
- 직업 아는 상태

- 2. 구직자
- 직업을 추천 받는 사람
- Test 데이터의 한 행
- 재직자와 같은 설문에 응답했으나 직업을 모름

User-based CF 알고리즘으로 분석

- CF 알고리즘을 직업추천 데이터에 적용

  - Test User 중에 Train User설문 응답이 가장 비슷한 유저의 직업을 Test User에게 추천 하는 방식 (by Cosine Similiarity)
  - 응답 결과가 비슷하다 == User의 특성이 비슷하다

http://www.edujin.co.kr/news/articleView.html?idxno=32192

CF 적용 프로세스

- 1~5: 임베딩 프로세스
- 6~7: CF 프로세스

1. 텍스트 데이터와 숫자 데이터를 분리(학습 데이터)
2. 전체 텍스트 데이터에 대하여 토큰화 해서 단어 모델 생성
3. 텍스트 데이터에 각각에 대해 모델 돌리기
   - 각각의 행,열 값에 대한 토큰화
   - 토큰화된 값을 모델에 돌리기

4. 임베딩 완료
5. 임베딩한 텍스트 데이터를 숫자 데이터에 붙히기
   (=설문 응답 결과를 모두 벡터화)
6. 구직자와 재직자의 설문 응답의 상호 유사도 행렬 생성
7. 구직자와 가장 비슷하게 응답한 재직자의 직업을 추천(구직자에게)

지금 진도 유사도 행렬 생성완료(6)

- 유사도 행렬 기반 예측 하면 끝남

	구직자의 응답 (idx=0)	구직자의 응답 (idx=1)	구직자의 응답 (idx=2)
재직자의 응답 (idx=0)	0.32(A)	0.67	0.54
재직자의 응답 (idx=1)	0.55	0.87	0.23
재직자의 응답 (idx=2)	0.74	0.32	0.89(B)

- A-구직자 0과 재직자 0의 설문 응답 유사도
- B-구직자 2에게 재직자 2의 직업 추천 !!

  ⇨구직자 idx=2와 재직자 idx=2의 응답이 가장 비슷하기 때문

- idx=user 식별
- 구직자-test data
- 재직자-train data

재직자와 구직자의 유사도 행렬

df

	0	1	2	3	4	5	6	7	8	9	...	9476	9477	9478
0	0.989769	0.989816	0.976401	0.994628	0.986329	0.981328	0.969389	0.994168	0.988	0.986297	...	0.50472	0.5913	0.490905
1	0.9897	0.990248	0.981862	0.99699	0.985012	0.976548	0.96643	0.996568	0.988766	0.985015	...	0.504609	0.591735	0.490359
2	0.990785	0.989133	0.982164	0.997176	0.986233	0.977241	0.968157	0.996843	0.989907	0.986231	...	0.505593	0.59206	0.491416
3	0.926455	0.991522	0.962739	0.968672	0.913065	0.905469	0.8745	0.965747	0.92703	0.913039	...	0.463384	0.568345	0.445657
4	0.99099	0.988498	0.984703	0.998133	0.985882	0.974904	0.967262	0.997859	0.990539	0.985882	...	0.506003	0.59238	0.491653
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
9481	0.576219	0.600533	0.571513	0.58703	0.574508	0.581071	0.560837	0.586166	0.574704	0.575175	...	0.974149	0.99204	0.970651
9482	0.575238	0.571528	0.57835	0.58184	0.571281	0.559485	0.559289	0.582325	0.577249	0.571967	...	0.990114	0.99895	0.986531
9483	0.499186	0.476732	0.473269	0.487903	0.502144	0.502476	0.502533	0.489026	0.498278	0.502865	...	0.997827	0.989377	0.998278
9484	0.619232	0.6664	0.632267	0.643967	0.613484	0.617989	0.590737	0.642182	0.618534	0.614102	...	0.944696	0.977686	0.938458
9485	0.588113	0.581158	0.600035	0.597659	0.581974	0.562454	0.567675	0.598356	0.591731	0.582636	...	0.98521	0.99551	0.980699

9486 rows × 9486 columns

앞으로 할 내용

- 자체 구현한(결정 트리)라 불리는 추천 방식에 히트텐을 적용해서 MRR로 다시 성능을 측정해볼 예정
  - 히트텐: 상위 10개를 추천하는 방식

- 늦어도 다음주까지는 끝내서 연구실 방문 예정

이 프로젝트의 방향성

⇨ 논문 (개인적인 바람)

⇨ 기술 보고서

- 데이콘 데이터 X (일부),
  워크넷 한국직업정보 설문조사 데이터 공개(전체)
  바꿔서 정리할 예정