KNOW 기반 직업 추천 알고리즘

2조 : 구현서, 김태환, 임지인

대회설명

- 대회명: KNOW기반 직업 추천 알고리즘 경진대회
- 링크
- 대회 기간
  - 1차 2022년 2월 2일까지 코드 제출
- 대회설명 및 목적
  - KNOW(한국직업정보)에서 여러 직업군의 재직자들을 상대로 한 관한 직무관련 설문 데이터를 통하여, 설문지 작성 정보에 따라 맞는 직업군을 찾아주는 모델을 만들어 본다
  - 또한, 직업과 연관이 높은 설문지 문항 분석 및 영향변수 발굴

배경

- KNOW(한국직업정보) 재직자 조사는 한국고용정보원이 청소년과 성인의 진로 및 경력설계, 진로상담, 구인, 구직 등에 도움을 주기 위해서 2001년부터 개발, 운영하고 있는 조사이다.
- KNOW(한국직업정보)는 다양한 직업에 종사하고 있는 재직자에 대하여 직무관련 조사를 수행하고 있다. (2017년 : 일반업무활동, 2018년 : 업무환경 및 흥미, 2019년 : 지식 및 성격, 2020년 : 업무수행능력 및 가치관)
- 본KNOW(한국직업정보) 데이터를 기반으로 직업추천 모델을 만들어보고 직업과 연관성 높은 직무능력을 탐색 발굴하고자 한다.

목적

- KNOW(한국직업정보) 설문 데이터셋을 활용한 직업 추천 알고리즘 개발
- 직업과 연관이 높은 설문지 문항 분석 및 영향변수 발굴

예상되는 어려움

- 각 연도별 설문지 주제 및 질문이 상이함
- 데이터 결측치 처리 방법 찾기 : 설문조사이기에 임의로 결측치를 채워 넣기 어려움
  - CF 기법으로 빈 값을 채워 넣을 계획
- 해당없음에 대한 처리여부 결정 애매

	idx	cq1	cq2	cq3	cq4	cq5	cq6	cq7	cq8	cq9	...	bq37	bq37_1	bq38	bq38_1	bq38_2	bq39	bq40	bq41_1	bq41_2	bq41_3
0	9486	4	1	3	1	1	3	2	3	1	...	1		2		6		35.0			2000
1	9487	5	4	5	5	5	5	3	3	3	...	4	커뮤니케이션 디자인 전공	1	1		1	65.0	2700	2200	
2	9488	3	3	3	4	3	3	4	4	2	...	4	국문학과	2		6		60.0			3200
3	9489	4	4	4	4	2	3	3	4	4	...	3	극작학과	2		6		20.0			2000
4	9490	5	2	3	2	4	4	4	4	3	...	2	실업	1	4		2	35.0	3000	1500	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
9064	18564	5	4	2	2	2	4	4	3	3	...	4	호텔조리학	2.0		5		72.0			8000
9065	18565	2	2	5	5	5	4	4	4	4	...	5	식품공학	1.0	1		1	40.0	3000	2100	
9066	18566	4	4	3	2	2	3	4	4	4	...	4	환경공학	1.0	2		2	40.0	1800	1800	
9067	18567	4	2	5	5	5	5	3	4	3	...	5	식품공학	1.0	1		1	52.0	3500	2800	
9068	18568	4	2	5	4	5	5	4	5	4	...	4	농학	1.0	1		1	40.0	2500	2500	

9069 rows × 140 columns

예상되는 어려움

- y 값인 knowcode의 직업을 어떤 기준으로 대, 중, 소분류를 할 것인지
- 150개의 질문 중 어떤 질문을 사용할 것인지
    - 설문지 문항에서 필요한 부분만 빼내는 작업 (Feature Engineering)
- 질문의 답변이 5~6가지이기에, 그대로 할 것인지 아니면 두 가지로 나누어서 진행할 것인지

변수명	변수설명	변수값	변수값 설명
knowcode	연계 KNOW 직업코드	11102	행정부고위공무원
		11201	기업 고위임원
		12101	정부행정 관리자
		12201	경영지원 관리자
		12301	마케팅·광고·홍보 관리자(부서장)
		12401	금융관리자
		12402	보험관리자
		13101	연구 관리자
		13201	유치원원장 및 원감
		13202	초등학교교장 및 교감
		13203	중고등학교교장 및 교감
		13204	대학교총장 및 대학학장
		13305	경찰·소방·교도 관리자
		13401	보건·의료 관리자
		13501	사회복지 관리자
		13601	예술·디자인·방송 관리자

예상되는 어려움

- 150개의 질문 중 어떤 질문을 사용할 것인지
  - 설문지 문항에서 필요한 부분만 빼내는 작업 (Feature Engineering)
- 질문의 답변이 5~6가지이기에, 그대로 할 것인지 아니면 두 가지로 나누어서 진행할 것인지

가. 귀하의 업무를 수행하려면 **【빌딩 및 건축】** 관련지식은 얼마나 중요합니까?

중요하지 않다	약 간 중요하다	중요하다	아 주 중요하다	아주 많이 중요하다
①	②	③	④	⑤

기존 방법론 조사 리스트

- World Happiness Report up to 2020
- 머신러닝 알고리즘 분석 및 비교를 통한 Big-5 기반 성격 분석 연구, 김용준 저

문제 해결

- 추천 알고리즘 성능 향상을 위한 정확한 직업 분류
  - 531개의 직업 대/중분류

데이터 설명

- 기본
  - Column: 2017년부터 2020년까지의 직업 관련 설문조사의 각 문항
  - Row: 각 재직자분들의 설문 조사에 대한 답변
- 차이점
  - 2017년: 일반업무활동 설문
  - 2018년: 업무환경 및 흥미 설문
  - 2019년: 지식 및 성격 설문
  - 2020년: 업무수행능력 및 가치관 설문
  - 연도별로 직업 분류 코드가 다르다
  - 설문 질문중에 bq로 시작하지 않는 문항은 연도마다 질문이 다르다
- 공통점
  - 연도별로 직업 분류 코드가 다르지만 연계코드로 모든 연도마다 통일되게 코드가 분류되어 있는것도 있다
  - 설문 질문중에 bq 로 시작하는 문항은 모든 연도에 공통된다

데이터 전처리

- 설문지 문항에서 필요한 부분만 빼내는 작업 (Feature Engineering)
- 결측이 있는 설문대답 처리
- 연도별 마다 맞게끔 전처리
- 직업별로 대, 중, 소 분류로 묶어준다

모델

- K-means clustering
- KNN
- Random Forest
- XGBoost

모델 학습 방식

- 우선 1차로 모든 직업군에 대해 맞추기는 어려우니 대분류 부터 맞추는 작업을 해본다
- 연도별로 설문 문항이 다르니 연도별로 따로 학습할지 아니면 모든 연도별로 공통된 Feature들을 찾아 한번에 학습할지 고민해본다

프로젝트를 수행할 환경

- Jupyter Notebook : 개별 작업
- Notion : 현재 진행상황 정리 및 보고
- Google Workspace : 공동 코딩 작업

기대 효과

- 구직자에게 정확도 높은 직업 추천 시스템 제공