빅데이터 응용 프로젝트 계획서**- 알고리즘 문제 추천 -**

2022. 3

목 차

Ⅰ. 프로젝트 목적 3p

1. 프로젝트 개요 3p

Ⅱ. 수집할 데이터 5p

1. 데이터 정보 5p

Ⅲ. 데이터 수집 방법 6p

* 수집 방법 6p

1. 비공식 API 기반 데이터 수집 6p
2. 웹 크롤링 기반 데이터 수집 6p

Ⅳ. 데이터 분석 방법 7p

1. 군집화 7p
2. 필터링 8p

V. 예상 결과 11p

Ⅰ. 프로젝트 목적

1. 프로젝트 개요

□ 개요

▷ 프로젝트명 : 알고리즘 문제 추천

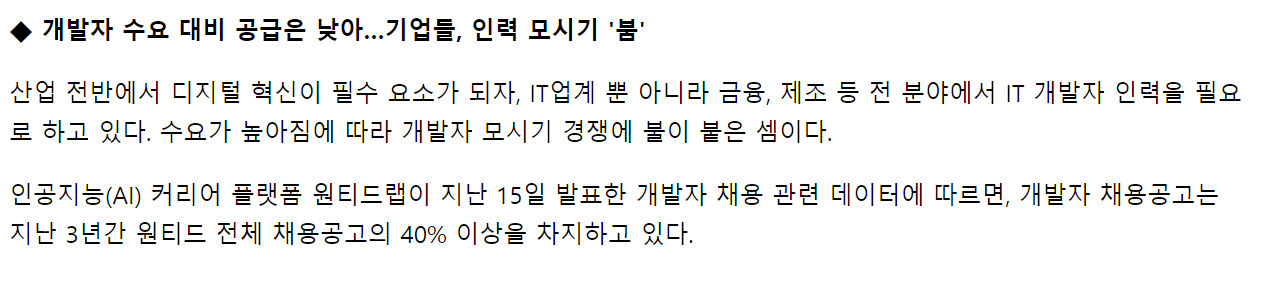
▷ 프로젝트 기간 : 2022. 03 .17 ~ 2022. 06. 02 (총 12 주간)

▷ 프로젝트 도구 : Python(jupyter Notebook)

□ 프로젝트 목적

▷ 프로젝트 필요성

최근 IT업계에 진입해 있거나 진입을 준비하는 사람들이 많아지고 있습니다.  
아래의 사진[[1]](#footnote-1)을 참고한다면 최근 4차산업을 기점으로 기업들이 다수의 양질의 개발자들에 대한 수요가 폭발적으로 증가하고 있지만 이에 대응하는 양질의 공급은 부족한 현상이 지속되고 있습니다.



따라 기업에서 양질의 개발자를 채용하기 위해 각 기업들은 채용과정에서 개발자들이 양질의 개발자임을 확인하기 위해 PS(Problem Solving)능력을 평가합니다.

보통 이러한 PS를 확인하기 위해 기업들은 채용 시에 “코딩 테스트”를 통해 개발자의 역량을 평가합니다.

이에 따라 개발자들은 원하는 기업에 들어가기 위해 각 기업이 원하는 기초적인 역량을 갖추고 있어야 합니다.

따라서 프로그래머들은 PS능력을 기르기 위해 다양한 알고리즘 문제들을 접해보고 풀어 보기 원하고 있으므로 이에 맞게 다양한 알고리즘의 문제를 제공하는 사이트들이 서비스되고 있습니다.

백준, 프로그래머스, 삼성 SW Expert Academy, 등 많은 사이트들이 있고 이러한 사이트에서 다양한 알고리즘 문제를 제공하고 있습니다.

이렇게 다양한 알고리즘을 풀 수 있는 문제를 제공하는 사이트에서 프로그래머들 갖는 공통적인 의문점은 “도대체 많은 문제들 중에서 나는 어떠한 문제들을 풀어야하지?”라는 의문을 갖고 있습니다.

따라서 저희는 이러한 의문에 착안해 국내 최대 문제를 지니고 있고, 다수의 이용자를 사용하고 있는 “백준”를 선점하였고 백준 사이트에서 사용자에게 알고리즘 문제를 추천하는 인공지능을 제공할 것입니다.

Ⅱ. 수집할 데이터

1. 데이터 정보

▷ 데이터 수집 기간

* 2022. 03. 25 ~ ing

▷ 데이터 목록

**- 유저 기본 데이터(Type : json)**

handle(text) : 사용자 명  
solvedCount (int64) : 푼 문제수  
exp(int64) : 경험치량  
tier(int64) : 현재 티어 - 브론즈 5기준 1부터 1씩 증가.  
atingByProblemsSum(int64) : 푼 문제의 난이도 합으로 계산한 레이팅  
ratingByClass(int64) : 클래스로 계산한 레이팅

- 문제 데이터(상세)(Type : json)

ProblemId (int64): 문제 ID  
isSolvable(Boolean) : 체점 가능 여부(true만 사용하고 제거)  
acceptedUserCount(int64) : 맞은 사람 수  
level(int64) : 문제의 tier  
averageTries(number(실수형)) : 평균 시도 횟수  
key : tag ID

**- 유저 상세 데이터(Type : html)**

bojRank(int64) : 백준 등수 (등수)  
solvedProblemCount(int64) : 맞은 문제   
submissionCount(int64) : 제출 횟수 (제출)  
collectCount(int64) : 정답 횟수 (맞았습니다.)  
wrongCount(int64) : 오답 횟수 (틀렸습니다.)  
tiemOverCount(int64) : 시간초과 횟수 (시간초과)  
memoryOverCount(int64) : 메모리초과 횟수 (메모리 추가)  
solvedProblemList : 맞은 문제(text : List)  
unSolvedProblemList : 시도했지만 맞지 못한 문제(text : List)

Ⅲ. 데이터 수집 방법

1. 수집 방법

▷ 데이터 종류: 스크립트 파일(JSON)/ 웹 페이지(HTML,CSS)

▷ 수집 기술: 웹 크롤링 / 웹 스크래핑 / API 사용

▷ 크롤러 구현 방법:

**Python 3.9 / jupyter notebook / Beautiful soup / .csv**

**▷ 비공식 API 기반 데이터 수집 – 유저 목록 및 solved.ac관련 데이터, 문제 정보**

<https://solved.ac/api/v3/user/show> - 랭킹 기반 유저 목록

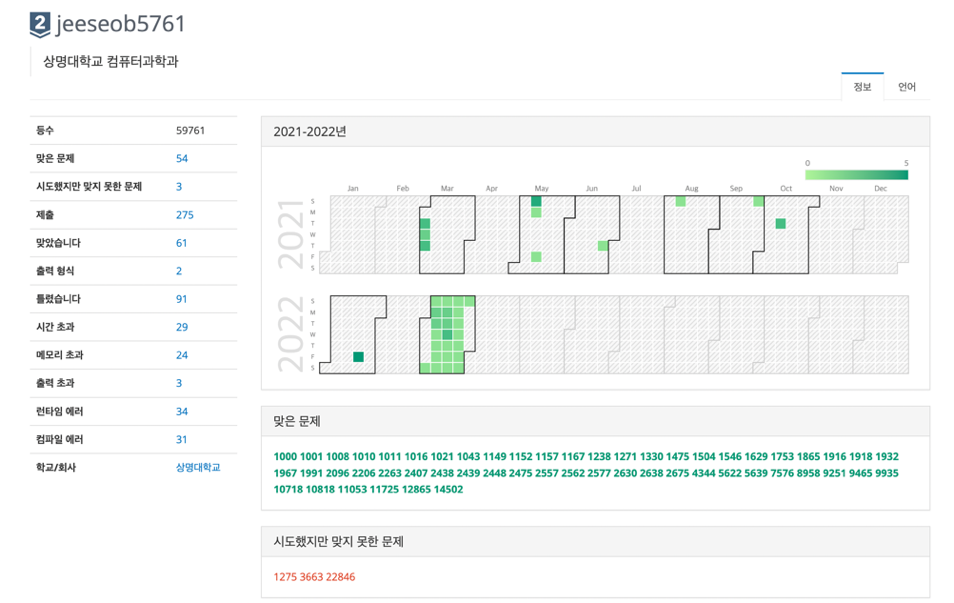
<https://solved.ac/api/v3/problem/lookup> - 문제 목록

**텍스트이(가) 표시된 사진

자동 생성된 설명**

**▷ 웹 크롤링 기반 데이터 수집 – 유저 개개인의 세부 정보,**

<https://www.acmicpc.net/user/UserID> - 유저 세부 정보(유저 목록 기반 크롤링)

****

Ⅳ. 데이터 분석 방법

1. clustering(군집화)

텍스트이(가) 표시된 사진

자동 생성된 설명

**□** Clustering Algorithm에는 여러가지 방법이 있기 때문에, 실제 데이터를 확인후, 데이터의 형태와 알고리즘의 특성을 고려하여 선택할 계획입니다.

▷ 현재 고려 중인 알고리즘은 K-Means, DBSCAN 등 입니다.

▷ **K-Means**

주어진 [데이터](https://ko.wikipedia.org/wiki/%EB%8D%B0%EC%9D%B4%ED%84%B0)를 k개의 [클러스터](https://ko.wikipedia.org/wiki/%ED%81%B4%EB%9F%AC%EC%8A%A4%ED%84%B0)로 묶는 알고리즘으로, 각 클러스터와 거리 차이의 [분산](https://ko.wikipedia.org/wiki/%EB%B6%84%EC%82%B0)을 최소화하는 방식으로 동작한다.

이 알고리즘은 [자율 학습](https://ko.wikipedia.org/wiki/%EC%9E%90%EC%9C%A8_%ED%95%99%EC%8A%B5_(%EA%B8%B0%EA%B3%84_%ED%95%99%EC%8A%B5))의 일종으로, 레이블이 달려 있지 않은 입력 데이터에 레이블을 달아주는 역할을 수행한다.

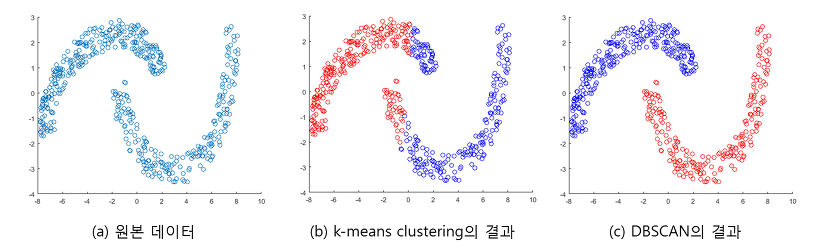
이 알고리즘은 [EM 알고리즘](https://ko.wikipedia.org/wiki/EM_%EC%95%8C%EA%B3%A0%EB%A6%AC%EC%A6%98)을 이용한 클러스터링과 비슷한 구조를 가지고 있다.

▷ **DBSCAN**

밀도 기반 알고리즘의 한 종류이다.

밀도 기반의 클러스터 알고리즘은 "동일한 클래스에 속하는 데이터는 서로 근접하게 분포할 것이다"라는 가정을 기반으로 동작한다.

아래의 그림은 중심 기반 클러스터링 알고리즘 (k-means clustering)과 밀도 기반 클러스터링 알고리즘 (DBSCAN)의 클러스터링 결과의 차이를 시각적으로 표현한 이미지 입니다.



Two moons 데이터셋에 대한 k-means clustering과 DBSCAN의 군집화 결과로 같은 색의 데이터는 같은 클래스에 군집 되었다는 것을 의미한다.

2. filtering()

**□ contents based filtering(콘텐츠 기반 필터링)**

콘텐츠 기반 필터링은 말 그대로 콘텐츠에 대한 분석을 기반으로 추천을 구현하는 방법이다.

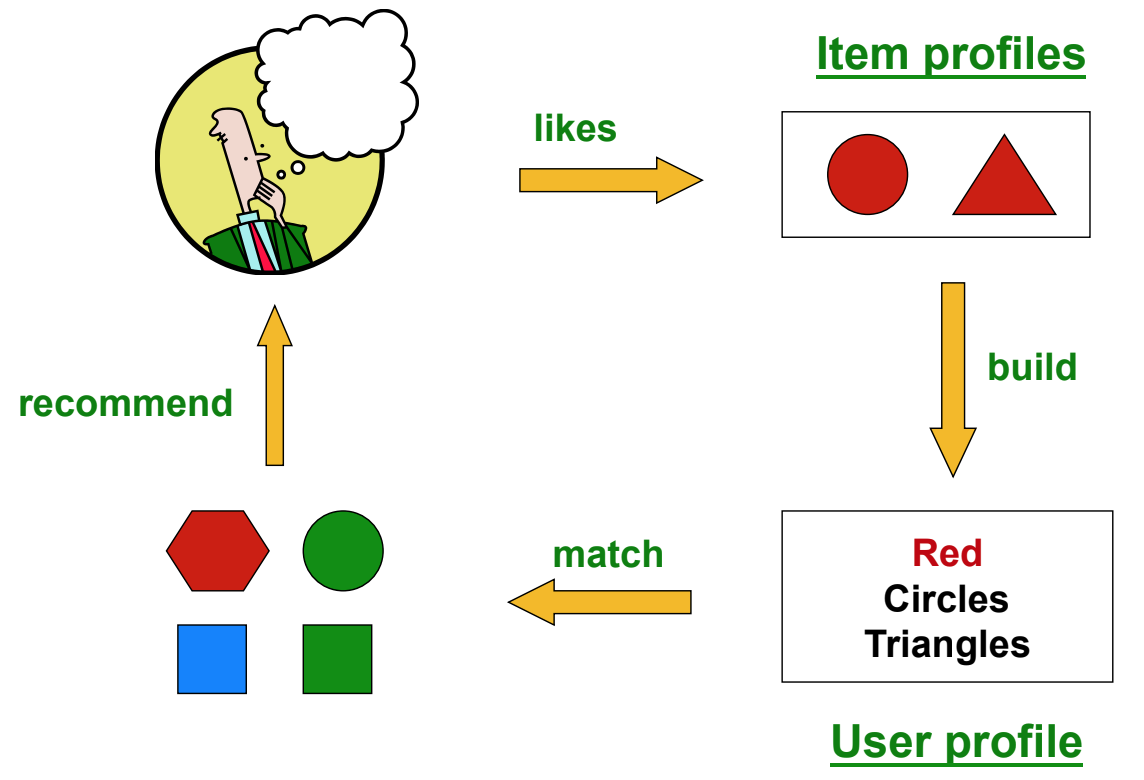
콘텐츠를 분석한 프로파일(item profile)과 사용자의 선호도(user profile)를 추출하고 유사성 분석을 통해 추천을 수행합니다.

가령 슈퍼맨이라는 영화에 대한 사전 분석을 통해 SF, 영웅 등의 특징을 기록하고 슈퍼맨을 본 사람에게 배트맨이라는 비슷한 종류의 영화를 추천해주는 방식입니다.

사용자가 특정 아이템을 선호하는 경우, 그 아이템과 비슷한 콘텐츠를 가진 다른 아이템을 추천해주는 방식입니다.

이 기법은 콘텐츠의 내용을 분석해야 하므로 아이템 분석 알고리즘이 핵심적이며, 이를 위해 군집분석(Clustering analysis), 인공신경망(ANN, Artificial Neural Network), TF-IDF(Term Frequency-Inverse Document Frequency) 등의 기술이 사용된다.

콘텐츠 기반 필터링은 내용 자체를 분석하므로 협업 필터링에서 발생하는 콜드 스타트(Cold-start problems)[[2]](#footnote-2) 문제를 자연스럽게 해결할 수 있습니다.

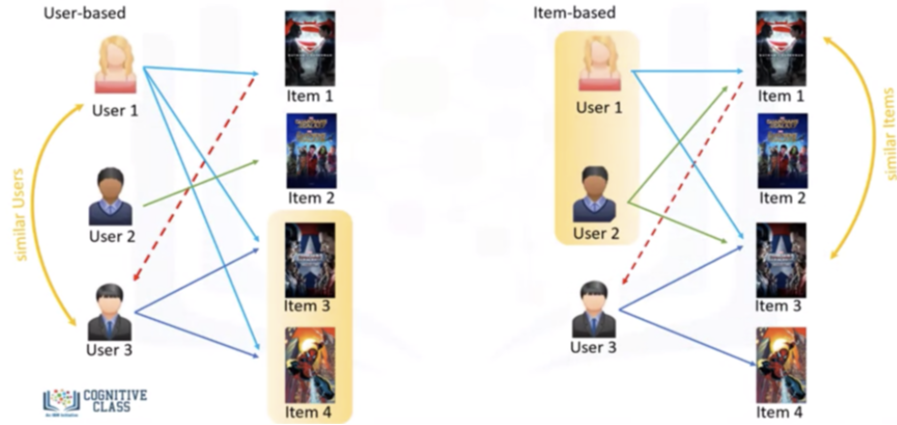
하지만 다양한 형식의 항목을 추천하기 어려운 단점이 있습니다. 예를 들어 음악과 사진, 비디오를 동시에 추천해야 하는 경우, 각각의 항목에서 얻을 수 있는 정보가 다르기 때문에 프로파일을 구성하기 매우 어려워집니다.

**□ collaborative filtering(협업 필터링)**

협업 필터링이란 같은 콘텐츠를 좋아하는 사람은 콘텐츠 취향이 비슷할 것이라는 가정에 기반한 알고리즘으로, 컴퓨터 이용자의 모든 사용 및 소비 기록을 입력시켜 그걸 근거로 불필요한 정보를 필터링해주는 기술입니다.

‘갑’이 A·B 콘텐츠를 좋아하고 ‘을’이 B·C 콘텐츠를 좋아한다면 갑에게는 C를, 을에게는 A를 추천하는 식입니다.

콘텐츠 평가에 참여하는 사람이 많아지고 데이터가 쌓일수록 추천 정확도가 높아지는 머신 러닝 기술이 적용됐습니다.

협업 필터링은 같은 성향이나 특성을 가진 사람은 비슷한 선택을 한다는 가정을 기반으로 추천하는 방식이기 때문에 기본적으로 많은 양의 데이터가 있어야 가능하기 때문에 데이터가 부족한 초기에 사용하기 어렵습니다. 그래서 유저기반 협업 필터링과 아이템기반 협업 필터링으로 나눠 볼 수 있습니다.

* **유저기반 협업 필터링**

비슷한 성향을 보이는 다른사용자들의 기록을 분석하여 추천할 영상을 찾습니다.

* **아이템기반 협업 필터링**

특정 영상A와 유사한 성향을 가진 영상을 모은 뒤. 사용자가 A를 초과했을 때 유사도가 높은 다른 영상을 추천합니다.

▷ 현재 고려 중인 알고리즘은 contents based filtering(콘텐츠 기반 필터링) 입니다.

V. 예상 결과

분류의 A.I에서 성능을 평가하는 기준은 정확도입니다.

즉 정확도는 전체 데이터 개수에서 옳게 분류한 데이터의 개수의 비율로 나타냅니다.

따라서 저희가 사용자에게 알고리즘 문제를 추천하기 위해 아래의 방식을 통해 A.I를 제작할 것입니다.

먼저 비공식 API 기반 데이터들과 웹 크롤링 기반의 데이터들을 전부 수집을 합니다.

저희 팀은 전체 데이터의 100%를 학습을 진행하는 것이 아니라 80%를 가지고 인공지능의 학습 데이터로 사용해 K-Means, DBSCAN의 ML 테크닉을 사용해 진행하고 이를 Content-base-filter를 적용해 A.I모델을 훈련을 하고 이렇게 훈련된 A.I를 검증해 만족스러운 결과를 얻었다면 나머지 20%의 데이터를 검증(validation) 데이터로 사용해 A.I모델을 검증하는 방식에 착안해 저희는 기본적으로는 8대 2 정도로 훈련 데이터 set과 테스트 데이터 set을 분리하여 진행을 할 것이며 진행 상황을 고려해 비율 또는 방법을 변경해 진행할 것입니다.

이러한 작업이 마무리가 되면 만족스러운 알고리즘 문제 추천 A.I가 제공될 것으로 판단됩니다.

1. 참고 사이트 주소: <https://www.inews24.com/view/1386692> [↑](#footnote-ref-1)
2. **콜드 스타트: 앞선 출력 결과를 이용해 동작하게 되므로, 데이터가 없는 시작 상태에서는 제대로 동작하지 않는 시스템** [↑](#footnote-ref-2)