[LG U+ Why Not SW Camp 5기]
커리어 설계 AI 에이전트 - CH 팀

모델 정의서

프로젝트명: 커리어 설계 AI 에이전트

팀명: Challenge is Happiness

작성자: 김민지

목차

Ⅰ. 시스템 구성

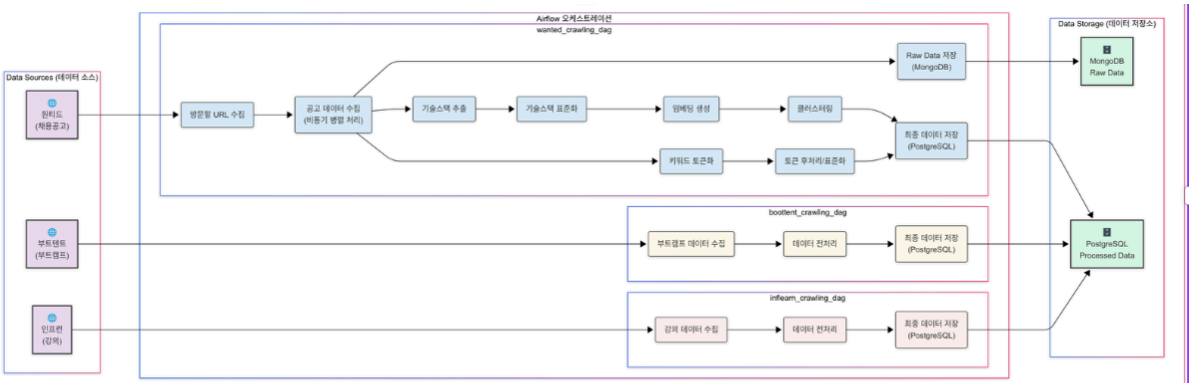
1. 개요

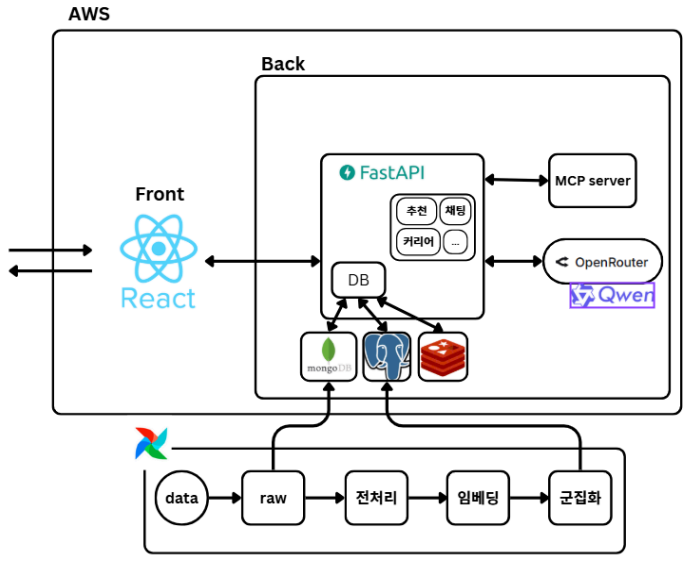
본 프로젝트는 사용자의 이력 정보를 기반으로 채용 공고를 추천하고, 직무 트렌드를 반영한 커리어 설계를 지원하는 AI 에이전트 'JOB자'를 개발하는 것을 목표로 한다.
사용자는 웹 기반 프론트엔드 화면에서 추천 결과를 확인하고, 추천된 공고를 기반으로 맞춤형 커리어 로드맵을 설계받을 수 있으며, 이 모든 흐름은 백엔드 서버 및 데이터 파이프라인이 함께 통합되어 구성된다.

2. 구조

(1) 아키텍처 및 전체 흐름

시스템은 크게 프론트엔드, 백엔드 API 서버, DB 스토리지, Airflow 기반 데이터 파이프라인, LLM API 서버로 구성되어 있으며, AWS 환경에서 운영된다.





전체 흐름은 다음과 같다.

① 데이터 수집 → ② 전처리 및 키워드 추출 → ③ 임베딩 → ④ 클러스터링 → ⑤ DB 저장 → ⑥ 사용자 입력 → ⑦ 유사도 계산 → ⑧ LLM 기반 설명 및 추천 제공

Ⅱ. 모델

1. 임베딩 모델 (Text embedding)

(1) 모델 설명 및 선정 이유

a. 역할
공고와 이력서의 텍스트를 벡터로 변환하여 유사도 기반 추천이 가능하도록 임베딩한다.

b. 사용 모델
intfloat/multilingual-e5-large
- 다국어 텍스트 표현에 특화된 SentenceTransformer 계열 모델
- 한국어 포함 다양한 언어의 의미를 효과적으로 임베딩
- 사전 학습된 모델
- (Query-Answer) 구조로 학습되어 검색 및 추천 시스템에 최적화

c. 구성
- encoder : 입력 문장을 고차원 의미 공간상의 벡터로 변환

d. 주요 알고리즘
- Mean Pooling 기반 Sentence Embedding
- Cosine Similarity를 통한 벡터 간 유사도 계산
- Triplet loss 기반 학습으로 의미적 거리 유지
- 처리된 벡터는 pgvector에 저장

e. 특징 및 선정 이유
- 다국어 지원 (한국어 포함)
- 커리어 정보와 채용 공고 간 의미 기반 유사도 추출에 적합

(2) 입력 텍스트 구조 및 임베딩 처리 방식

- 채용 공고: (주요업무*2) + (자격요건*2) + 우대사항 + 기술스택
- 사용자 정보: 기본 정보 + 희망 직무 + 경력 + 보유 스킬 + 경험 + 자격증

위 항목들을 각각 하나의 텍스트로 합친 후, 임베딩 모델에 입력하여 1024차원 벡터로 변환한다.

(3) 유사도 계산 방식 및 기준

공고 벡터와 사용자 벡터 간 **Cosine Similarity**를 계산하여 유사도를 측정한다.
유사도 점수는 0~1 사이의 실수로, 값이 1에 가까울수록 의미적으로 더 유사하다고 판단한다.
유사도 점수가 높은 순으로 공고를 정렬하여 상위 N개를 추천 후보로 선정한다.
이때, 다음과 같은 보정 로직을 적용하여 실제 추천 점수를 조정하였다.

- 지원자 유형 보정
- 입력 정보 밀도 보정
- 직무명 정합성 보정

(4) 평가 지표

임베딩 모델의 성능을 비교하기 위해 다음과 같은 지표를 사용하였다:

① 항목별 평균 유사도 (주요업무, 자격요건, 우대사항, 전체)

- 정의: 각각의 항목에 대해 사용자와 채용 공고 간의 코사인 유사도를 계산한 평균값
- 목적: 실제 추천 대상이 되는 텍스트 필드에서 모델이 의미 기반 유사도를 잘 반영하는지 평가

- 설명: 주요업무, 자격요건, 우대사항 등은 공고 내에서도 추천의 핵심 필드이며, 이들의 유사도가 높을수록 의미 기반 매칭 품질이 좋다고 판단

② Top-N 평균 유사도 (Top-20 기준)

- 정의: 사용자 기준으로 상위 N개 공고를 추천했을 때의 평균 유사도
- 목적: 실제 추천 리스트에서 얼마나 높은 유사도의 공고들이 선정되었는지를 판단
- 설명: 사용자가 실제로 추천받게 되는 리스트의 품질을 측정하며, 직관적 성능 평가에 적합

③ 유사도 분산

- 정의: 추천된 공고 간 유사도의 분산(Variance).
- 목적: 추천 결과의 일관성을 확인
- 설명: 분산이 낮을수록 유사도 기준이 안정적으로 적용된 것으로 판단. 즉, 고평가 편차가 적고 추천 품질이 일정

④ Recall (재현율)

- 정의: Top-N 추천 리스트 내에 실제로 적합한(Positive) 공고가 포함된 비율
- 목적: 추천 결과의 정확도를 정량적으로 측정
- 설명: 유저의 특성과 맞는 공고를 얼마나 잘 찾아냈는지 판단하는 핵심 지표로 사용 (※ 내부 기준으로 '정답 공고셋' 정의하여 Recall 계산)

(5) 성능 검증 및 모델 선택

세 가지 임베딩 모델(BAAI/BGE, E5-large, Text3S)을 대상으로 주요 지표 기반 성능 비교를 수행하였다. 주요 비교 항목은 본문 유사도 평균, Top-20 유사도 분산, Recall이며, 전체 텍스트 구성 항목(주요업무, 자격요건, 우대사항)별 유사도 또한 정량적으로 측정하였다.

모델명	주요업무	자격요건	우대사항	본문 전체	유사도 분산	Recall
BAAI/bge	44.8%	45.9%	46.2%	61.1%	0.0105%	76.0%
E5-large	**82.0%**	**82.6%**	**83.0%**	**90.3%**	**0.0023%**	**80.0%**
Text3S	26.5%	26.2%	26.8%	53.7%	0.0416%	68.0%

결과 분석 및 선택 이유

- E5-large 모델은 모든 평가 항목에서 압도적으로 높은 점수를 기록하였다.
- 본문 전체 평균 유사도 90.3%, 유사도 분산 0.0023%, Recall 80.0% 로, 정확도와 일관성 모두에서 가장 우수한 성능을 보였다.
- 특히, Query-Answer 학습 구조 덕분에 의미 기반 추천이 잘 작동했으며, 한국어 텍스트 표현력도 뛰어났다.

이에 따라, JOB자 추천 시스템의 임베딩 모델로 intfloat/multilingual-e5-large 를 최종 채택하였다.

2. 클러스터링 모델 (Clustering)

(1) 모델 설명 및 선정 이유

a. 역할
공고 벡터를 클러스터링하여 기존 직무 분류의 한계를 보완하고, 직무 간 의미 기반 유사성을 반영한 재분류 체계를 형성한다.

b. 사용 모델
KMeans
- 단순하고 빠르며 대용량 데이터에서도 효율적인 대표적인 군집화 알고리즘
- 벡터 간 거리 기반으로 동작

c. 주요 알고리즘
- Euclidean Distance 기반 KMeans
- 입력 벡터: SentenceTransformer 기반 임베딩 벡터
- 클러스터 수(K): 내부 품질 지표 및 직무 분포를 고려하여 수동 설정
- 군집 라벨을 기반으로 직무를 재매핑하고 DB에 저장

d. 특징 및 선정 이유
- 데이터 수 10K 미만
- 카테고리 수 파악
- 라벨 없음 (비지도 학습)
- 예측 목적 존재 (직무 분류)
- 벡터 기반 데이터 사용

(2) 벡터 기반 군집화 목적 및 활용 방안

기존 채용 공고에 포함된 직무명은 작성 주체에 따라 표현 방식이 상이하거나, 동일한 업무라도 서로 다른 명칭으로 분류되는 경우가 많다. 이에 따라 공고 임베딩 벡터를 기반으로 의미적으로 유사한 직무들을 하나의 군집으로 묶고, 이를 기준으로 직무 재정의 및 표준화된 직무 체계를 구축하고자 한다.

a. 활용 방안
- 직무 리매핑
- 직무 트렌드 분석
- 사용자 - 직무 간 갭 분석
- 대시보드 시각화

(3) 클러스터 수 설정 및 리매핑 방식

Elbow Method의 분석 지표에 따르면 적정 군집 수는 K=20 수준으로 추정되었으나, 직무 세분화 및 의미 기반 추천의 필요성으로 인해 의도적으로 **K=200**으로 과적합을 시켜 군집 수를 크게 설정하였다. 이는 다양한 직무 간 미세한 의미 차이를 반영하고, 이후 리매핑을 통해 통합한다.

- n_cluster = 200

(4) 품질 평가 지표

군집화 품질을 정량적으로 평가하기 위해 다음 세 가지 대표적인 지표를 활용하였다.

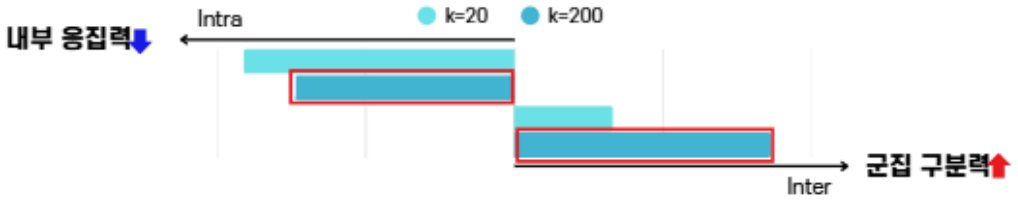
a. Silhouette Score
: 각 데이터가 자신의 클러스터와 얼마나 잘 맞는지를 측정하는 지표

b. Calinski-Harabasz Index
: 클러스터 간 거리 대비 클러스터 내 응집도를 나타내는 비율 기반 지표

c. Davies-Bouldin Index
: 클러스터 간 유사도를 기반으로 군집 간 중복도를 평가

(5) 성능 검증 및 모델 선택

본 프로젝트에서는 기존 직무 분류 체계(K=20)의 한계를 극복하고, 실제 추천 정확도를 향상시키기 위해 K=200 클러스터링 후 리매핑 방식을 도입하였다.



a. 검증 방식
   - K=20 군집과 K=200 리매핑 군집에 대해 정량 지표(Intra/Inter Distance) 비교
   - 공고와 유저 간의 직무 매칭 정확도, 추천 품질 향상 여부를 중심으로 분석
   - 정성적 평가: 동일 직무 내 공고 텍스트 유사도 분석, 직무 분포 시각화 등

b. 모델 선택 사유:

   K=200 리매핑 군집은 다음과 같은 정량적 성능 향상을 보였다.

   - 평균 Intra-cluster Distance: 0.0910 → 0.0736 (↓ 응집력 향상)
   - 평균 Inter-cluster Distance: 0.0331 → 0.0864 (↑ 분리도 향상)

기존 군집 수(K=20)에서는 의미적으로 섞이거나 맞지 않는 직무가 많았고, 직무 간 유사성 기반 분류에는 부족하였다. 반면, K=200 + 의미 기반 리매핑을 통해 세분화된 직무 그룹 형성 및 추천 신뢰도 향상이 가능해졌다.

따라서, 최종적으로는 K=200 클러스터링 + 리매핑 방식이 품질과 실용성 면에서 모두 우수하다고 판단되어 최종 선택하였다.

3. LLM (Large Language Model)

(1) LLM API 설명

a. 사용 API
   - **llama-4-maverick**

b. 역할
   - 임베딩 기반 Top-30 공고 후보군에서 최종적으로 적합한 Top-5 공고를 재랭킹하고, 각 공고에 대해 추천 사유를 자연어로 설명하는 역할을 수행한다.
   - 유사도 기반 추천의 정량적 한계를 보완하고, 사용자 맞춤형 커리어 인사이트를 제공하는 핵심 컴포넌트다.

c. 기술적 구성
   - 입력: 사용자 이력서 요약 + 상위 30개 공고 요약
   - 출력:
     - (1) 지원 적합 여부 (추천)
     - (2) 추천 사유에 대한 자연어 요약 문장
     - (3) 부족한 역량 또는 성장 방향 제안

- 구조
  *Prompt 기반 LLM 호출 구조
  *Prompt 튜닝을 통해 다음을 수행
    - 지원자-공고 간 정합성 판단
    - 자연어 기반 추천 사유 설명
    - 불충분한 조건이 있는 경우, 이를 지적하고 보완 방안을 제시

d. 주요 알고리즘
- Prompt-only 기반 구조이나, 상위 30개 공고의 전문을 DB에서 추출해 Prompt에 포함시킴으로써 RAG-like 구조로 작동
- Few-shot Prompt Engineering
- Instruction-following 기반 자연어 추론 (Natural Language Inference, NLI)

(2) API 특징 및 선정 이유

a. 후보군 모델
- 유료 : GPT-4o-mini, GPT-4.1-mini, GPT-4.1-nano, gemini-1.5-flash, maverick
- 무료 : qwen, maverick

b. 모델 평가 기준
- *사용자 스킬/경력 반영 수준*
  : 공고 요구사항과 사용자의 경력, 스킬셋 매칭을 문장에서 얼마나 반영하는지
- *제외 조건 고려 여부*
  : 보충역/경력/지역 제한 등과 같은 명시적 제외 조건을 자연스럽게 고려하는지
- *모델 일관성 및 안정성*
  : 동일한 입력에 대해 일관된 품질의 답변을 생성하는지 여부
- *모델 추론력*
  : 직접 명시되지 않은 항목에 대해서도 LLM이 스스로 추론하여 문장을 생성하는 능력
- *시간 및 비용*
  : 평균 응답 시간 및 호출 비용 등을 고려하여 실제 서비스 가능성을 평가

c. 선정 이유

이러한 평가 항목을 기반으로 다양한 모델을 비교하였으며, 모든 항목에서 우수한 성능을 보인 **llama-4-maverick (유료)** 모델을 최종 선택하였다. 모델 간 성능 차이는 크지 않았지만, 일관성과 Reasoning 측면의 품질 차이가 존재하였고, 시간(평균 4~5초) 및 비용(1회 호출당 약 2.3원)도 실 서비스에서 수용 가능한 수준으로 판단되었다.

(3) 활용 방안 및 통합 흐름

a. API 사용 흐름

1. 사용자 이력서 요약 및 전처리
2. 유사도 기반 Top-30 공고 벡터 추출
3. 각 공고의 주요 정보 요약 (주요업무, 자격요건 등)
4. 사용자 요약 + 공고 요약 → 프롬프트 구성
5. LLM 호출 (llama-4-maverick via OpenRouter API)
6. 응답 파싱 → 추천 여부, 추천 사유, 성장 제언
7. 적합한 상위 5개 공고를 최종 추천 결과로 제공

b. 실행 환경 및 요구사항

- API 플랫폼 : OpenRouter
- 사용 모델 : llama-4-maverick
- 호출 방식 : Prompt / RAG
- 평균 응답 시간 : 약 6초
- 호출 환경 : FastAPI 기반 백엔드 서버에서 API 호출
- 사전 처리 필요 : 공고 요약, 프롬프트 압축, 중복 제거 등

III. 결론 및 향후 계획

1. 전체 구조 요약 및 추천 흐름

본 프로젝트는 사용자 이력 정보를 기반으로 채용 공고를 추천하고, 직무 트렌드를 반영한 커리어 설계를 지원하는 AI 에이전트 “JOB자”의 MVP 시스템 구축을 목표로 하였다.

시스템은 다음과 같은 통합 아키텍처로 구성되어 있다:

- 데이터 파이프라인 (Airflow): 수집 → 전처리 → 임베딩 → 클러스터링 → DB 저장 자동화
- 백엔드 (FastAPI): 유저 입력을 기반으로 유사도 계산 및 커리어 로드맵 제공
- 프론트엔드 (React): 사용자 입력 및 결과 시각화 UI 제공
- 저장소: PostgreSQL / MongoDB / Redis

이 과정을 통해 사용자는 단순히 공고 추천을 받는 것에 그치지 않고, 어떤 역량이 부족한지, 왜 이 공고가 나에게 적합한지, 어떻게 성장해야 하는지에 대한 설명과 방향성까지 제공받을 수 있다.

2. 보완점 및 향후 개선 방향

구분	보완점	향후 개선 방향
데이터 수집	**특정 산업군/직무군 위주로 편중**	**공고 수집 범위 확대 (예: 공공기관, 스타트업 등)**
임베딩 모델	**사전학습 모델 사용으로 도메인 적합성 제한**	**사용자 및 산업 특화 파인튜닝 진행 검토**
클러스터링	**직무 리매핑 후 수작업이 일부 필요**	**Active Learning 기반 군집 해석 자동화 도입 고려**
LLM 활용	**Prompt-only 기반 설명 생성**	**Retrieval 기반 RAG 구조 도입 및 성능 개선**
정량 평가	**사용자 실 사용 평가 미비**	**사용자 피드백 기반 리콜, NDCG 등 평가 구조 추가**