**아래 내용은 참고 사항으로, 우리가 채워나가야 합니다.**

**1. 문제정의, 서론**

기존의 문제가 뭐였길래?

왜 이 플랫폼을 개발하려 하는가?

에 대한 내용 필요

**2. 문헌조사 요약**

1) 이전 연구

기술 기반 매칭을 위한 지식 그래프를 활용한 연구가 진행

https://arxiv.org/pdf/2109.02554.pdf

구직 추천 시스템을 위한 의미론적 주석 방법을 소개

https://aip.vse.cz/getrevsrc.php?identification=public&mag=aip&raid=377&type=fin&ver=3

우도를 활용하여 테이블 형식의 데이터에 대한 의미론적 개념 주석에 대한 추론이 연구

https://arxiv.org/pdf/2012.08594.pdf

이러한 선행 연구들은 구직 정보의 효율적 제공에 중요한 기여를 하였으나, 대부분의 연구가 일반적인 구직자를 대상으로 하여 개인별 맞춤화에는 한계가 있었음

기존 연구들에서 다루지 않았던 새로운 접근 방식으로, 부산광역시 청년들에게 보다 효과적이고 개인화된 구직 정보를 제공할 수 있을 것으로 기대

2) Notation 및 접근방식

**-Notation**

 Given : 이력서 T는 구직자의 skill들을 포함하고 있습니다. 이력서로부터의 skill 추출에 대한 설명은 섹션 5에서 제공됩니다. 구직자의 skill에 대한 벡터는 로 표현되며, , n은 구직자가 가진 skill의 수를 나타냅니다.

 는 한국표준직업분류(KSCO)에 따른 직업의 집합을 의미합니다.

 는 직업이 가질 수 있는 모든 Skill의 집합이라고 합니다. Skill에 대한 설명은 섹션 4에서 제공됩니다.

 각 직업의 skill은 로 표현됩니다.

****  는 구직자의 skill이 직업의 skill이라는 것을 나타냅니다. 마찬가지로, 는 구직자가 그 직업을임을 나타냅니다.

-접근방식

우도[3]는 주어진 증거에 대한 특정 가설의 가능성을 의미합니다. 본 보고서에서는 우도 개념을 통해 각 직업이 요구하는 Skill을 파악하며, 이를 바탕으로 구직자의 Skill과 가장 잘 매칭되는 직업을 추천하는 방식을 제안합니다. 이를 위해 각 직업의 Skill 분포를 이산확률분포로 표현하고, 이 확률 분포를 이용하여 우도 계산을 진행합니다. 이러한 방식을 효과적으로 적용하기 위해서는 각 직업별 Skill 및 Skill의 확률 데이터가 필요합니다.

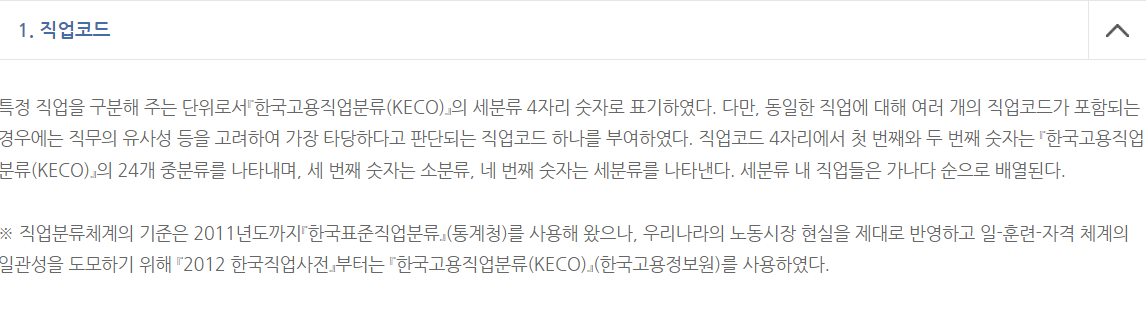
**3. 개발 요구사항(큰 제목만 보고, 세부내용 수정. 큰 제목도 수정 가능하면 우리 방식대로 변경)**

**2.2 데이터셋 구축**

직업과 그에 관련된 Skill 데이터셋의 구축은 초기 단계에서 중요합니다. 현재로서는 직업별 Skill의 표준화된 정의가 부재한 상태입니다. 본 연구에서는 한국표준직업분류(KSCO)를 바탕으로 워크넷 플랫폼에서 제공하는 채용공고 데이터를 활용하였습니다. 해당 데이터 활용의 주된 이유는 채용공고에서 직업과 그에 따른 필요 Skill이 명시되어 있기 때문입니다. 이런 방식으로, 구직자의 현재 Skill과 시장의 요구사항 간의 괴리를 최소화하고, 구직자가 자신의 Skill을 최대한 활용할 수 있는 직업을 찾는 데 도움을 제공합니다. 따라서, 직업 및 직업별 Skill 데이터셋 구축에 이러한 데이터는 필수적입니다.

[그림1] Dataset Framework

**<참고!!>**



https://www.work.go.kr/consltJobCarpa/srch/jobDic/jobDicIntro.do?pageType=jobDicNotes

**2.2.1 채용공고 데이터 수집**

직업정보는 제7차 한국표준직업분류 직종코드(KSCO)를 사용하였고, KSCO를 기준으로 채용공고 사이트인 워크넷에 있는 채용공고 데이터를 수집하였습니다. 채용공고 데이터는 2023년 7월 워크넷에서 대한민국의 모든 지역과 각 직종에 대하여 수집을 하여 총 78227개의 채용공고 데이터를 수집하였습니다.

**2.2.2 Skill추출을 위한 GPT Fine-Tuning**

채용공고에는 매일 수천 건의 데이터가 업데이트되며, 각각의 채용공고는 서로 다른 형식과 구조를 가집니다. 이런 다양성 때문에 일반적인 자연어 처리로는 원하는 Skill을 효과적으로 추출하기 어렵습니다. 이 문제를 해결하기 위해 LLM(Large Language Model)을 사용하여 워크넷의 채용공고로부터 Skill 정보를 추출하였습니다. 본 연구에서는 GPT 3.5 Model을 활용하였습니다. GPT는 Skill 정보를 추출하는 데 유용하나, 정확한 결과를 얻기 위해서는 추가적인 튜닝이 필요합니다. 따라서, GPT-3.5-turbo Model을 Fine-tuning하는 과정을 거쳤습니다. 서울을 제외한 지역에서 각 직종별로 무작위로 선정된 100개의 채용공고를 훈련 예제로 사용하였으며, 이를 통해 원하는 형식의 Skill 정보를 추출하는 모델을 구축하였습니다.

**2.2.3 채용정보에서 Skill 데이터 추출 후 전처리**

Fine-Tuning된 Model을 활용해 서울을 제외한 지역의 채용공고에서 Skill을 추출하였습니다. 그러나, 모든 채용공고에서 원하는 형식에 맞게 Skill이 추출되지 않았습니다. 예컨대, Skill 정보가 제공되지 않은 채용공고에서는 'Skill 정보가 제공되지 않았습니다.'로 표시되었습니다. 또한, 'Java, C, etc..'와 같이 콤마로 나열된 Skill 정보에서는 각 Skill을 개별적으로 분리하고, 행 단위로 저장할 필요성을 느꼈습니다. 이에 따라 추가적인 전처리 과정을 거쳤으며, 이 과정은 하드코딩으로 진행되었습니다. 결과적으로, 각 직업별로 정리된 Skill dataset을 생성하였습니다.

**2.2.4 TF•IDF 가중치 부여 및 확률 부여**

수집된 각 직업별 Skill dataset을 활용하여 사용자의 이력서를 기반으로 직업 추천을 위한 준비 작업이 필요하였습니다. 채용공고에서는 특정 Skill이 빈번하게 등장하는 반면, 다른 Skill은 상대적으로 덜 등장할 수 있습니다. 이러한 Skill들의 중요도나 상대적인 가치를 객관적으로 평가하는 것은 쉽지 않습니다. 이 문제를 해결하기 위해 TF•IDF 기법을 활용하여 Skill들의 중요도를 측정하였습니다. 이를 위해 각 직업의 채용공고를 하나의 문서로 간주하였습니다. 아래와 같은 노테이션을 바탕으로 TF 및 DF를 계산하였습니다.

-  : 직업별 채용공고 문서에서의 Skill 빈도

-  : Skill이 포함된 직업별 채용공고 문서 수

이후, 각 직업별 Skill에 대한 TF\*IDF 값을 계산하여 해당 직업에서의 Skill 중요도를 산출하였습니다.

**2.2.5 Smoothing**

Smoothing은 모델의 일반화 능력을 강화하는 중요한 기법입니다. 특정 직업의 채용공고에서 어떤 Skill이 등장하지 않는다면, 해당 직업이 될 확률이 0이 되는 것을 막기 위한 작업입니다. 이 방법을 통해, 모델은 새로운 데이터나 예상치 못한 입력에 대해 더욱 유연하게 대응할 수 있게 됩니다. Smoothing을 적용함으로써, 특정 직업의 Skill 확률이 0이 되는 상황을 방지하며, 모델이 다양한 상황에서도 안정적으로 동작하게 합니다.

**2.2.6 확률부여**  
앞서 구한 직업별 Skill의 TF•IDF 값은 Skill의 중요성을 파악하는 데 큰 도움을 주지만, 직접적인 확률 계산에는 적합하지 않습니다. 이를 보완하기 위해, 각 직업별 Skill의 확률을 아래의 방식으로 계산하였습니다.

위의 수식은 특정 직업에서 Skill의 중요성을 상대적으로 표현하는 값입니다. 이렇게 계산된 확률을 바탕으로, 각 직업별 Skill의 확률을 추천 알고리즘에 효과적으로 활용할 수 있습니다.

**2.3 추천모델**

이 섹션에서는 구직자의 Skill 집합을 바탕으로 최적의 직업을 추천하는 알고리즘의 구체적인 절차를 설명합니다. 우리의 목표는 구직자가 보유한 Skill과 시장에서 요구하는 Skill 사이의 일치도를 극대화하여, 구직자에게 가장 적합한 직업을 추천하는 것입니다. 아래는 본 연구에서 제안하는 직업 추천 프레임워크의 구조를 나타내는 [그림2]의 설명입니다.

[그림 2] 직업추천 Framework

**2.3.1 사용자의 Skill 추출**

먼저, 구직자의 이력서를 분석하여 보유하고 있는 Skill을 파악합니다. 이 과정에서 구직자는 PDF 형식의 이력서를 시스템에 입력하게 됩니다. PyPDF2 라이브러리를 활용하여, PDF 내의 텍스트 정보만을 추출합니다. 이후, 채용공고에서의 Skill 추출과 동일한 GPT 3.5 turbo Fine-Tuning Model을 사용하여 이력서에서 Skill 정보를 추출합니다.

**2.3.2 가능도 계산**

추출된 Skill을 기반으로 각 직업에 대한 가능도를 계산합니다. 우리는 이미 직업과 직업별 Skill, 그리고 그 Skill의 확률을 이산확률분포로 표현하는 방법을 구축하였습니다. 이 정보를 활용하여 구직자의 Skill 집합과 각 직업의 Skill 집합 사이의 가능도를 아래와 같이 계산합니다.

Log Likelihood(LL)를 사용하여 계산의 효율성을 높힙니다.

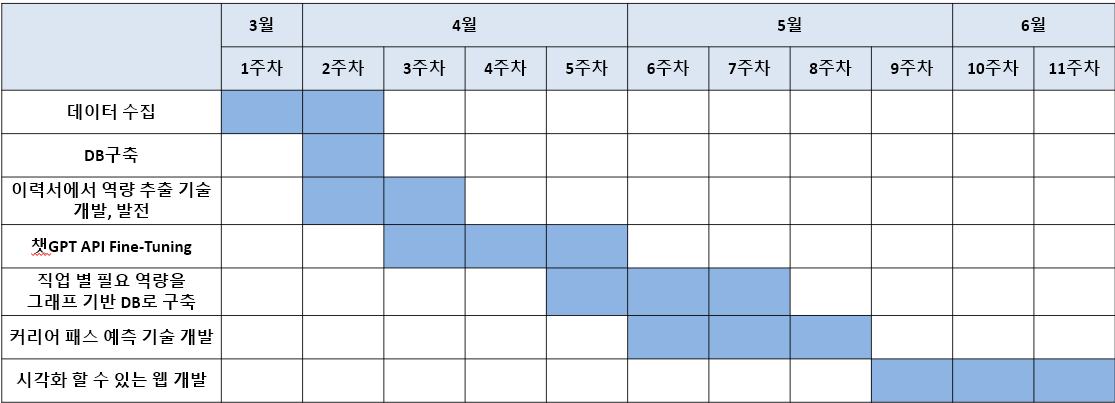
가능도가 가장 높은 직업을 추출하기 위한 식은 다음과 같습니다.

이 알고리즘은 구직자의 Skill 집합과 각 직업의 Skill 집합 사이의 일치도를 최대화하여 구직자에게 최적의 직업을 추천하는 방식으로 설계되었습니다.

**4. 예상 결과물(기존 문제정의서에서 가져옴)**

|  |  |
| --- | --- |
|  |  |
| <기 개발된 직업 추천 앱[1]> | <워라벨을 고려한 직업 추천> |

**5. 개발 계획**

* 1주차 : 데이터 수집
* 2주차 : DB 구축, 이력서에서 역량 추출 기술 분석
* 3주차 : 이력서에서 역량 추출 기술 개발, 발전
* 4주차(중간고사 전주) : 챗GPT API Fine-Tuning
* 5주차(중간고사) : 챗GPT API Fine-Tuning
* 6주차 : 직업 별 필요 역량을 그래프 기반 DB로 구축
* 7주차 : 직업 별 필요 역량을 그래프 기반 DB로 구축, 커리어 패스 예측 기술 개발
* 8주차 : 커리어 패스 예측 기술 개발
* 9주차 : 시각화할수 있는 웹 개발
* 10주차 : 시각화할수 있는 웹 개발
* 11주차 : 시각화할수 있는 웹 개발
* 기말고사 전주 : 발표 준비
* 기말고사 : 발표
* 그 후 : 데이터 볼륨 늘리기, 정확도 향상, 프로그램 최적화, 서비스 추가 [](https://private-user-images.githubusercontent.com/100757595/316289778-28db01e9-8de8-49af-a041-7b10e0f574d7.png?jwt=eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9.eyJpc3MiOiJnaXRodWIuY29tIiwiYXVkIjoicmF3LmdpdGh1YnVzZXJjb250ZW50LmNvbSIsImtleSI6ImtleTUiLCJleHAiOjE3MTE2Mjk0OTEsIm5iZiI6MTcxMTYyOTE5MSwicGF0aCI6Ii8xMDA3NTc1OTUvMzE2Mjg5Nzc4LTI4ZGIwMWU5LThkZTgtNDlhZi1hMDQxLTdiMTBlMGY1NzRkNy5wbmc_WC1BbXotQWxnb3JpdGhtPUFXUzQtSE1BQy1TSEEyNTYmWC1BbXotQ3JlZGVudGlhbD1BS0lBVkNPRFlMU0E1M1BRSzRaQSUyRjIwMjQwMzI4JTJGdXMtZWFzdC0xJTJGczMlMkZhd3M0X3JlcXVlc3QmWC1BbXotRGF0ZT0yMDI0MDMyOFQxMjMzMTFaJlgtQW16LUV4cGlyZXM9MzAwJlgtQW16LVNpZ25hdHVyZT0wYjcxODk5YjJhM2FiYjc5ZTgzOTkwMDg2NDNlZGVmODgxYTEyMDkwNzg3NjIyZDkyNzI4MjdkOWY1YzVkYTZkJlgtQW16LVNpZ25lZEhlYWRlcnM9aG9zdCZhY3Rvcl9pZD0wJmtleV9pZD0wJnJlcG9faWQ9MCJ9.5SlMxc1WrSNfZ176wr-NPArxCOG3DXCYtFTsXoZX5EM)

**+a(성능 테스트 결과)**

테스트를 평가하기 위해 사용한 방법에 대해 설명한 것. 우리도 어떻게 평가할 것인지 알아두면 좋지만, 이는 나중에 파악해도 괜찮을 것이라 생각함.

[그림 12] 지역별 채용공고 개수

실험을 평가하기 위해 채용공고 데이터를 기반으로 평가를 진행하였습니다. 일반적으로 생각될 수 있는 이력서 기반으로의 직업 추천 평가는, 사용자가 받은 추천의 정확성에 대해 확신하기 어려운 문제점이 있습니다. 이러한 문제를 해결하기 위해 동일한 메커니즘(즉, Skill을 추출하여 평가)을 따르는 채용공고 데이터를 활용하여 평가를 진행하기로 결정하였습니다. 이를 위해 각 채용공고를 지역별로 분류하였습니다.

데이터 전처리 및 준비 과정에서는 전체 데이터셋을 두 부분으로 나누었습니다. 첫 번째로, 확률 추정용 데이터셋으로 [그림12]을 참고하여 전체 데이터 중 약 90%에 해당하는 서울을 제외한 지역의 채용공고를 활용하였습니다. 이 데이터셋은 주요 변수의 확률 분포를 추정하는 데 중심적으로 사용되었습니다. 두 번째로, 검증 데이터셋으로는 전체 데이터 중 약 10%인 서울 지역의 채용공고를 사용하였습니다. 이 데이터셋은 초기에 추정한 확률이 얼마나 타당한지, 그리고 해당 확률 추정치가 일관성을 보이는지 검증하는 데 사용되었습니다. 이런 방식의 데이터 분할은 확률 추정의 정확성과 타당성을 확보하기 위한 표준 접근법을 따랐습니다.

[그림 13] 실험결과

추천의 정확한 예측 비율을 측정하기 위하여 Precision@k, 그리고 추천한 직업 중 정확한 측정이 있는 비율을 나타내는 Recall@k를 사용하여 전체적인 정확도에 대하여 평가하였습니다. 실험을 진행한 결과 직업에 대한 추천 정확도는 약 73%인것으로 뛰어난 것을 확인할 수 있습니다. 이를 통하여 구직자에게 가장 적합한 직업을 추천할 수 있다는 것을 알 수 있습니다. 추후 연구에서는 직업에 대한 Skill의 전제관계까지 고려를 진행한다면, 더 높은 성능을 기대할 수 있습니다.