

데이터 전처리

(1) 기업 분석 보고서 (Domain Name : Company_Report)

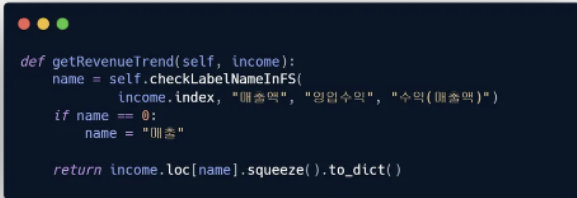
- 기업별 데이터 추출

- 사용 라이브러리: `dart_fss`, `BeautifulSoup`, `requests`, `pandas`

- 사용 API: `Dart OpenAPI`

- 작업 흐름

1. `dart_fss`로 기업 데이터 추출을 진행한다.
2. 위 라이브러리로 추출이 안 될 경우, OpenAPI로 직접 호출한다. - 기업별 ②재무제표
 - 2-1. 원문 데이터를 받아 BeautifulSoup으로 xml 태그 파싱한다.
3. 범용적 사용을 위해 기업별 사용 용어를 Rule-based로 처리한다.



```
def getRevenueTrend(self, income):  
    name = self.checkLabelNameInFS(  
        income.index, "매출액", "영업수익", "수익(매출액)"  
    )  
    if name == 0:  
        name = "매출"  
  
    return income.loc[name].squeeze().to_dict()
```

- pdf 데이터 추출 - 기업별 ①사업내용

- 사용 라이브러리: `pdfplumber`

- 전처리 이유: robots.txt를 위반하지 않으면서 '사업내용' 데이터를 수집하기 위함

- 작업 흐름

1. 수집한 기업별 사업보고서(pdf)를 불러온다.
2. 원문 500여 페이지 중 '사업내용'의 부분의 위치를 찾는다.
3. 해당 페이지의 정보를 추출하여 html 태그 형태로 json 파일로 저장한다.

- **사업 내용 요약 - 기업별 ①사업내용**

- **사용 모델:** gpt-4o-mini
- **사용 프롬프팅 기술 및 패턴:**

```
사용자 입력 메시지의 내용은 한국기업의 사업내용이다. 너는 기업을 전문적으로 분석하는 유능한 분석가이다.
모든 <조건>에 맞춰, <구조>과 같은 구조로 한국기업의 사업내용을 요약하라.

<조건>
1. 개조식으로 작성할 것.
  (예시: [BEFORE] 회사는 지속적인 기술 및 서비스에 대한 투자를 통해 핵심 사업의 경쟁력을 강화하고 있습니다. ->
  [AFTER] 지속적인 기술 및 서비스에 대한 투자를 통해 핵심 사업의 경쟁력을 강화)
2. 1500 token 내로 작성을 마무리할 것.
3. 첫 문단은 취업준비생들을 위해 요청한 사업내용 요약에 대한 전반적인 총평과 기업 공략 포인트에 대해서 정리해서
  기재할 것.
4. 첫 문단 이후에 요청한 사업내용 요약을 기재할 것.

<구조>
1. 글을 HTML 형식으로 요약할것
2. 목록은 <ul>과 <li> 태그로 표현할 것.
3. 태그 사이에 띄어쓰기(' \n') 없이 한 줄로 표현할 것.
예시:
<p>첫 문단(전반적인 총평과 기업
```

Cot, Role-Prompting, Format Insutrctiuon을 적용한 프롬프트

- **CoT** - 단계적으로 요약 구조화
- **Role-Prompting** - "분석가"라는 관점을 설정
- **Format Instruction** - 개조식 작성과 HTML 태그 구조

(2) AI 모의 면접 (Domain Name : AI-Interview)

전처리 목적

- 목표로 하는 최종 결과물
 1. 질문 생성 모델
 2. 면접 채점 모델
- 위 목표를 이루기 위해 원천 데이터를 보고 많은 논의를 진행했고 다음과 같은 결론이 나옴
 - 직무에 대한 정보가 존재하여 직무 별로 질문들을 나누어 처리해서 모델링
 - 질문에 대한 의도(5개로 결정)를 부착하고 각 의도마다 1개의 질문으로 총 5개의 질문-답변을 하나의 면접 세션으로 구성
 - 이렇게 구성하는 이유는, 면접을 진행할 때 한 사람에 대해 여러 질문들을 하게 되는데, 이전 답변을 고려하여 다음 질문이 나와야 자연스러울 것이라고 생각했기 때문
 - 면접 채점 결과 제공

원천 데이터는 아래와 같이 하나의 json 파일에 하나의 질문-답변을 포함하는 구조로 되어 있음

ckmk_d_ard_f_e_101874	JSON 원본 파일
ckmk_d_ard_f_e_101875	JSON 원본 파일
ckmk_d_ard_f_e_101877	JSON 원본 파일
ckmk_d_ard_f_e_101878	JSON 원본 파일
ckmk_d_ard_f_e_101879	JSON 원본 파일
ckmk_d_ard_f_e_101881	JSON 원본 파일
ckmk_d_ard_f_e_101882	JSON 원본 파일
ckmk_d_ard_f_e_102313	JSON 원본 파일
ckmk_d_ard_f_e_102314	JSON 원본 파일
ckmk_d_ard_f_e_102316	JSON 원본 파일
ckmk_d_ard_f_e_102318	JSON 원본 파일
ckmk_d_ard_f_e_102319	JSON 원본 파일

```
{
  "version": "1.0",
  "dataset": {
    "info": {
      "date": "20230116",
      "occupation": "ARD",
      "channel": "MOCK",
      "place": "ONLINE",
      "gender": "FEMALE",
      "ageRange": "35-44",
      "experience": "EXPERIENCED"
    },
    "question": {
      "raw": {
        "text": "디자이너로서 앞으로의 목표에 관해서 설명해 주세요",
        "wordCount": 6
      },
      "emotion": [],
      "intent": []
    },
    "answer": {
      "raw": {
        "text": "저는 디자이너로서 어 많은 경험들을 쌓아서 그것들을 토대로 제 개인 브랜드를 런칭하는 게 목표입니다. 이 회사에서 쌓은",
        "wordCount": 63
      },
      "emotion": [],
      "intent": [
        {
          "text": "",
          "expression": "",
          "category": "attitude"
        }
      ],
      "summary": {
        "text": "저는 디자이너로서 경험들을 토대로 제 개인 브랜드를 런칭하는 게 목표입니다. 그 브랜드를 육성있게 크게 키우는 것이 0",
        "wordCount": 28
      }
    }
  },
  "rawDataInfo": {
    "question": {
      "fileFormat": "wav",
      "fileSize": 241358,
```

전처리 과정

- 질문 생성 모델 학습 데이터

1. Rule Base 기반 의도 부착

```
OVERCOME_KEYWORD = ['상사', '대처', '어떻게 해결', '예기치', '문제', '예상', '다르', '노하우', '극복', '동료',  
                    '갈등', '타 부서와', '다른 부서와', '분쟁', '위기', '산사태', '코드 리뷰', '전문가 팀원',  
                    '전문가인 팀원', '위기관리', '위기', '축박']  
  
ADAPTABILITY_KEYWORD = ['변화', '새로운 환경', '적응', '부서에 배치', '다른 부서에', '갑자기']  
  
COWORKING_KEYWORD = ['어떤 사람', '조직 내', '어떤 역할', '어떠한 역할', '포지션', '무슨 역할', '다른 경우', '사고성', '협업',  
                    '비아이티적 동료']  
  
PROJECT_KEYWORD = ['프로젝트', '책임', '팀워크', '리더십', '일정 관리', '목표 설정', '협업 과정', '성과 도출', '동아리']  
  
SELF_DEVELOPMENT_KEYWORD = ['학습', '성장', '자기 주도', '새로운 기술', '트렌드 파악', '자격증', '아이씨티 전문성',  
                             '교육 참여', '개인 목표', '스킬 향상', '공부', '스터디', '개발 능력 향상']  
  
SKILL_KEYWORD = ['새로운 언어', '프로그래밍 언어', '개발 도구', '시스템 설계', '코드 최적화', '알고리즘', '데이터베이스',  
                 '성능 향상', '기술 스택', '소프트웨어 개발', '테스트 자동화', '애자일 방법론', '클라우드 컴퓨팅', '버전 관리',  
                 '독커', '도커', '프레임워크', '보안', '커널 분석', '소프트웨어', '오에스', '리팩토링', '테스트 코드', '디엔에스',  
                 '에이치티티피', '에이치티티피에스', '페어 프로그래밍', '좋은 개발 문화', '설계 프로그램', '아이씨티 역량',  
                 '아이피 주소', '메서드', '자바스크립트']
```

- 위와 같이 Rule Base로 총 5개의 의도를 면접 질문들에 부착
- 실제 질문과 그 의도가 정말 일치하는지, 정성 평가(랜덤 샘플링 하여 사람이 직접 확인)와 OpenAI API(GPT)를 통해 검증 후, Rule Base로 의도를 부착해도 문제가 없겠다는 결론을 내림

2. 이전 답변과 다음 질문과의 유사도 계산

- 최종 데이터에 similarity가 존재하는데, 이 값이 이전 답변과 다음 질문(이전 답변 기준 다음 질문)의 코사인 유사도를 계산한 값임
- 유사도를 계산하는 이유는 앞서 말했듯이, 이전 답변과 다음 질문이 어느 정도 관련이 있어야 자연스러운 면접이 될 것 같기 때문
- 최종 데이터를 구성할 때는 코사인 유사도가 높은 n개 중 랜덤으로 다음 질문을 선택했음

3. 직무 별 데이터 처리

- 전처리를 진행하면서 직무 별로 면접 질문이 차이가 없고 거의 모든 질문들이 일반적인 질문인 것을 발견
- 직무별로 데이터를 나누지 않고 통합해서 처리하고, 면접은 인성 면접으로 결정

[전처리 된 질문 생성 모델 학습 용 데이터]

```
[
  {
    "question": "협업을 할 때 사교성이 좋은 편인지 궁금합니다 그리고 또 사교성을 키우기 위해 어떤 노력을 하셨나요",
    "answer": "저의 큰 단점이라고 할 수 있는데요. 저는 굉장히 성격이 내성적이고 처음 본 사람하고는 낯가림이 심하기 때문에 사실 그",
    "rule_based_intent": "협업 능력"
  },
  {
    "question": "지원자분 본인의 장점과 단점에 대해서 설명해 주시면 좋겠습니다 그리고 장점을 최대한 특화시키고 단점을 극복하려고",
    "answer": "저는 책임감이 굉장히 강합니다. 제가 책임감이 강해지기까지는 저희 단점이 한몫 했습니다. 저는 부탁과 거절을 잘 못하",
    "rule_based_intent": "대처 능력",
    "similarity": 0.6498361825942993
  },
  {
    "question": "지금까지 살아오면서 새로운 환경에서 전혀 몰랐던 업무를 맡아서 한 경험이 있으실까요 만약 있다면 소개해 주세요",
    "answer": "지금까지 살아오면서 새로운 환경에서 몰랐던 일을 맡아서 한 경험을 말씀드려 보겠습니다. 저는 이전에 회사에서 취업했",
    "rule_based_intent": "적응력",
    "similarity": 0.6491960287094116
  },
  {
    "question": "과거에 일을 쭉 하셨잖아요 지금까지 과거에서 지금까지 어떤 프로젝트를 많이 해보셨는지 한번 이 자리에서 말해보시겠",
    "answer": "저는 경력직으로 지원한 만큼 여러 가지 프로젝트를 수행해 왔습니다. 때로는 팀원으로 때로는 팀장으로서 여러 프로젝트를",
    "rule_based_intent": "프로젝트 경험",
    "similarity": 0.6007181406021118
  },
  {
    "question": "면접자분은 삼 년 내에 본인의 주요 성장 계획이나 목표가 별도로 있으십니까",
    "answer": "삼 년 내에 제 주 성장 계획과 목표를 말씀드리겠습니다. 첫째로 회사 내에서 일 년 안에 제 모든 직무를 파악하고 싶습",
    "rule_based_intent": "자기 개발",
    "similarity": 0.45644432306289673
  }
]
```

- 면접 채점 모델 학습 데이터

1. OpenAI API 기반 채점

```
response = openai.chat.completions.create(
    model="gpt-4o-mini",
    messages=[
        {"role": "system", "content": "너는 면접 대상자의 답변을 채점하는 유용한 채점 담당자야."},
        {"role": "User",
         "content":
            f""" [Question]{question}
                [Intent]{intent}
                [Answer]{answer}
                [Note]
                1. [Answer]는 면접관이 면접 대상자의 [Intent]를 파악하기 위한 [Question]에 대한 면접 대상자의 답변이야.
                2. 면접자가 면접관의 질문에 대해 얼마나 잘 대답했는지를 1~100점으로 채점하고, 답변에 대한 feedback을 제공해줘.
                3. 답변에 아쉬운 점이 존재한다면 점수를 낮게 주었으면 해.
                3. output은 "score:~점<s>feedback:답변에 대한 피드백"처럼 각각을 "<s>"로 분리하여 생성해줘.
            """
        }
    ]
)
```

- 앞서 구성 된 세션 데이터와 위와 같은 프롬프트를 제공하여 OpenAI API 기반으로 면접 채점 데이터 생성
- 실제로 평가된 점수와 그 피드백이 적절한지 정성 평가를 통해 확인하고, OpenAI API로 면접 채점 데이터를 생성해도 되겠다는 결론을 내림

[전처리 된 면접 채점 모델 학습 용 데이터]

```
{
  "question": "지금까지 협업을 하면서 가장 어려웠던 것과 어려움을 어떻게 해결했는지 구체적으로 해결 과정에 대해 말씀해 주시기 바랍니다.",
  "answer": "지금까지 협업을 하면서 가장 어려웠던 것과 어려움을 어떻게 해결했는지 함께 말해 달라고 이렇게 질문을 해 주셨는데요. 지금까지 합",
  "summary": "협업을 할 때 가장 어려운 점은 의견을 좁혀가는 점들이었습니다. 의견을 조율하고 의견을 좁혀 나갈 때 어려움을 느꼈던 적이 많았습",
  "rule_based_intent": "협업 능력",
  "score": "75점",
  "feedback": "답변이 협업의 어려움을 잘 설명하고 상대방의 의견을 경청하는 방식을 제시했지만, 구체적인 사례나 경험을 드러내지 않아 신뢰성이",
  "alternative_answer": "제가 참여한 프로젝트에서 서로 다른 의견을 가진 팀원들이 있었습니다. 이때, 먼저 팀원들의 의견을 경청하고 이해하는",
},
{
  "question": "과거 직면한 어려운 문제를 해결하기 위해 어떤 아이디어를 제한하셨고 또 무엇을 하셨는지 과정을 말씀해 주실 수 있으신지 궁금합니다.",
  "answer": "과거 어려운 문제를 해결하기 위해 저는 어떤 아이디어를 제한했고 어떻게 해결했는지 과정을 말해보겠습니다. 학과 과정 중 안전 캡션",
  "summary": "과거 안전 캡션 디자인 과목에서 안전 어플리케이션 제작 프로젝트에 참가했던 경험이 있습니다. 팀을 만들어 아이디어를 내고 제작",
  "rule_based_intent": "대처 능력",
  "similarity": "0.733162522315979",
  "score": "70점",
  "feedback": "면접 대상자는 과거의 팀 프로젝트 경험을 통해 문제 해결 과정을 설명했으나, 보다 구체적인 어려움의 설명이 부족했습니다. 어려움",
  "alternative_answer": "저는 대학 시절 팀 프로젝트에서 안전 어플리케이션을 개발하는 과정에 참가했습니다. 프로젝트 초반, 팀원 간의 의견 차",
},
{
  "question": "지원자님께서 지금까지 살아오시면서 새로운 환경에서 전혀 몰랐던 새로운 일을 맡아서 그러한 업무를 수행한 경험이 있으신지 궁금합니다.",
  "answer": "새로운 환경에서 이제 완전 모르는 방향에서 이제 맨 땅에 헤딩이라고 하는데 그런 부분에 대해서 한 경험에 대해서 말씀드리겠습니다.",
  "summary": "맨 땅에 헤딩이라고 하는데 경험에 대해서 말씀드리겠습니다. 군 생활을 오래 했는데 부대 이동을 많이 하게 됩니다. 새로운 환경에서",
  "rule_based_intent": "적응력",
  "similarity": "0.47714120149612427",
  "score": "70점",
  "feedback": "지원자의 경험에 대한 설명이 있었지만, 구체적인 사례나 해당 업무에서의 성과에 대한 언급이 부족하여 전반적으로 적응력에 대한",
  "alternative_answer": "저는 군 복무 중 여러 부대로 전이되는 경험을 했습니다. 각 부대마다 새로운 환경과 업무를 맡게 되었고, 처음에는 많은",
},
{
  "question": "본인이 생각하기에 좋은 개발 문화라는 것은 어떤 것인지 설명 부탁드립니다.",
  "answer": "좋은 개발 문화라고 하면 겪어본 적이 있습니다. 바로 존중 배려 이런 건 것 같습니다. 철학이나 어떻게 보면 숫자로 계산할 수 없는",
  "summary": "좋은 개발 문화는 존중 배려, 존중, 배려와 같은 존중 배려가 있다고 생각합니다. 그것을 정량화하는 작업은 필요하지만, 서로 배려하",
  "rule_based_intent": "기술적 역량",
  "similarity": "0.3522321879863739",
  "score": "75점",
  "feedback": ""
}
```

2차 전처리

위 과정으로 만들어진 데이터로 모델을 학습하고 테스트한 결과, 이전 답변을 고려하지 않고 질문을 생성하는 문제점이 발견됨

문제 원인은 원천 데이터 자체에 이전 답변을 고려하여 다음 질문을 한 데이터가 없기 때문이라고 판단

이 문제를 해결하기 위해 면접 채점 데이터와 같이 OpenAI API를 기반으로 데이터를 생성
이 과정에서 초기에 설정했던 질문 의도와 면접의 첫 번째 질문이 달라짐

- 데이터 생성

면접 채점 데이터를 생성할 때와 마찬가지로 적절한 프롬프트와 첫 번째 질문을 같이 제공하여 데이터 생성

```
response = openai.chat.completions.create(
    model="gpt-4o-mini",
    messages=[
        {"role": "system", "content": f"너는 1:1 면접을 진행하는 면접관이야."},
        {"role": "user",
         "content":
            f"""질문 의도: ['자기 분석', '대처 능력', '소통 능력', '프로젝트 경험', '자기 개발']
            이전 질문: "{beforeQuestion}"
            이전 답변: "{beforeAnswer}"

            [Note]
            1. 질문 의도는 면접관이 질문해야 할 질문 리스트들이야. 이중 너는 {intent}에 관한 질문을 생성할 거야.
            2. 이전 답변은 면접자가 이전 질문에 답변한 내용이야.
            3. 질문 생성시 면접자의 이전 답변을 참고하여 만들어줘.
            4. 단, 이전 질문과 유사하지 않고, 면접자의 "{intent}"을 파악할 수 있도록 면접 질문을 생성하고, 그 질문에 대한 예상 답변까지도 생성
            5. {context}
            6. {answerText}
            7. 질문, 예상 답변, 피드백은 각각 3문장으로 생성해줘.
            8. output은 "question:너가 생성한 질문<s>answer:질문에 대한 예상 답변<s>score:~점<s>feedback:답변에 대한 피드백"처럼 각각
            """""
        ]
    )
```

너무 좋은 답변들만 생성하지 않도록 적절한 비율 설정

```
if percent == 20 :
    answerText = '답변에 대한 채점 점수가 50점 미만으로 나오도록 조금 잘못된 답변을 생성해봐.'
elif percent == 30:
    answerText = '답변에 대한 채점 점수가 65점 정도로 나오도록 조금 애매한 답변을 생성해봐.'
else:
    answerText = ''
```


[전처리 된 최종 데이터]

```
{
  "question": "다른 사람이 생각하는 지원자님의 장 단점은 뭐라고 생각하시나요 이유도 함께 말씀 부탁드립니다.",
  "answer": "저를 보는 장점이라 함은 기록이 없다 그런 얘기는 좀 들었던 것 같습니다. 크게 중요하지 않는데 라는 얘기를 하더라도 주변에서. 그래서 글썄이 이게 제가 공감할 하는",
  "intent": "자기 분석",
  "score": "65점",
  "feedback": "지원자께서는 장점과 단점을 모두 언급하며 자기 분석을 시도하셨습니다. 하지만 장점에 대한 구체적인 설명이 부족하고, 주변의 평가를 직접적으로 수용하지 않고 의문",
},
{
  "question": "어려운 상황이나 예상치 못한 문제에 직면했을 때, 지원자님은 어떻게 대처하시나요? 구체적인 예시를 들어주실 수 있나요?",
  "answer": "저는 어려운 상황에 처했을 때, 먼저 상황을 냉정하게 분석하려고 합니다. 예를 들어, 전에 팀 프로젝트 중 마감 직전 갑자기 큰 문제가 발생했을 때, 팀원들과 즉각 회의",
  "intent": "대처 능력",
  "score": "85점",
  "feedback": "구체적인 예시를 들어 대처 방법을 설명한 점이 좋습니다. 다만, 문제의 성격이나 그 과정에서 느낀 감정에 대해서도 조금 더 자세히 언급하면 좋을 것 같습니다. 그러",
},
{
  "question": "이전에 팀원들과의 소통이 원활하지 않아 어려움을 겪었던 경험이 있나요? 그 상황에서 어떻게 대처했는지 말씀해 주실 수 있나요?",
  "answer": "저는 팀원들과의 소통이 잘 이루어지지 않았던 상황이 있었지만, 그때는 그냥 개인적으로 문제를 해결하려고 했습니다. 결국 팀원들이 각자 다른 방향으로 작업을 하게 되",
  "intent": "소통 능력",
  "score": "45점",
  "feedback": "답변에서 소통 문제를 해결하기 위한 구체적인 행동이나 방법이 부족합니다. 팀원들과의 꾸준한 소통을 통해 문제를 예방하거나 해결했던 경험을 공유하면 더 좋을 것",
},
{
  "question": "최근에 참여했던 프로젝트에 대해 설명해 주실 수 있나요? 그 프로젝트의 목표와 진행 과정을 간단히 말씀해 주세요.",
  "answer": "제가 최근에 참여한 프로젝트는 교육 플랫폼을 개발이었습니다. 이 프로젝트의 목표는 사용자 친화적인 인터페이스를 통해 학습자들이 더욱 쉽게 자료를 접근할 수 있도록 하",
  "intent": "프로젝트 경험",
  "score": "85점",
  "feedback": "프로젝트의 목표와 진행 과정을 명확하게 설명해주셨습니다. 하지만, 구체적인 문제 해결 과정이나 팀원들과의 협력에 대한 언급이 부족해 좀 더 강화될 필요가 있습니다",
},
{
  "question": "최근에 자기 개발을 위해 어떤 노력을 해왔는지 구체적으로 말씀해 주실 수 있나요?",
  "answer": "최근에는 여러 온라인 코스를 통해 새로운 기술들을 배우고 있습니다. 특히, 데이터 분석에 관심이 있어 관련된 강의를 수강하고 있으며, 이를 통해 직무에서 필요한 기술",
  "intent": "자기 개발",
  "score": "68점",
  "feedback": "전반적으로 자기 개발에 대한 노력과 방향성이 잘 드러나긴 했지만, 한정적인 시간이 주어진 상황에서 조금 더 구체적인 계획과 성과를 제시했다면 더욱 긍정적인 인상을",
}
```

최종적으로 이 데이터로 질문 생성 모델, 면접 채점 모델 학습을 진행

Data 비율

Train Data: 약 9천개(세션)

Test Data: 약 1천개(세션)

데이터가 적은 이유는 모델 학습을 위한 자원 제한 때문에 더 많은 데이터를 사용하여 학습 할 수가 없음