**모델정의서**

**Team 청약 알려드림**

**유상범, 강진모, 강이삭, 박기범, 유다은**

**<목차>**

1. **프로젝트 개요**
   1. **프로젝트 명**
   2. **목적**
2. **데이터 설명**
   1. **데이터 출처**
   2. **전처리 과정**
3. **모델 설계**
   1. **모델 유형**
   2. **입력 및 출력**
   3. **하이퍼파라미터**
4. **프로젝트 개요**

* **프로젝트 명 : 상담 이력 기반 가상(Virtual Agent) 서비스**
* **목적 : 복잡한 청약 제도와 정보 접근성 부족 문제를 해결하기 위한 AI 기반 상담 서비스**

**제공**

* 상담 이력 기반 가상 상담(Virtual Agent) 서비스는 복잡한 청약 제도와 정보 접근성 부족으로 인해 발생하는 문제를 해결하기 위해 설계되었다. 본 프로젝트의 주요 목적은 AI 기반 상담 챗봇을 활용하여 개인별 맞춤형 정보를 제공하고, 입력 오류를 방지함으로써 청약 부적격 사례를 감소시키는 것이다. 이를 위해 청약 기초 정보 제공, 개인별 자격 조회 및 실시간 정보 업데이트와 같은 핵심 기능을 구현하며, 사용자 친화적인 상담 환경을 조성하고 24시간 접근성을 보장한다. 최종적으로 본 서비스는 청약 대상자의 주거 안정과 경제적 자립을 지원하고, 청약 제도에 대한 이해도를 높여 보다 효율적인 청약 프로세스를 돕는 것을 목표로 한다.

1. **데이터 설명**

* **데이터 출처 :**
  + **청약 자격 및 공고 데이터**
* 각 개인의 청약 자격 여부를 계산하기 위해 필요한 조건 데이터를 포함한다.
* 예: 나이, 무주택 기간, 부양가족 수, 소득 요건 등.
  + **청약 일정 및 경쟁률 데이터**
* 실시간 청약 일정과 지역별 경쟁률 데이터를 포함하여 사용자에게 최신 정보를 제공한다.
* 크롤링을 통해 수집.
  + **정책 및 규제 데이터**
* 청약 관련 정책 변경 사항을 정책 업데이트 알림에 활용된다.
  + **사용자 입력 데이터**
* 사용자가 제공하는 개인 정보와 선호 조건으로, 개인화 서비스에 사용된다.
* **전처리 과정**

대부분의 데이터는 크롤링(Crawling)을 통해 수집하였다. 모집 공고문의 경우 PDF 파일 형태로 제공되기 때문에, **LlamaCloud의 Llama Parse**를 활용하여 텍스트로 변환하였다. 변환된 각 모집 공고문은 JSON 파일 형식으로 저장되었으며, 문서명은 Title 필드에, 내용은 Content 필드에 담았다.

텍스트화 과정에서 목차 부분이 # 기호로 구분된 것을 확인하여 **MarkdownSplitter**를 사용해 각 문단을 나누는 추가 전처리를 수행하였다. 또한, 원본 파일을 여러 단위로 청킹(Chunking)하여 별개의 문서로 활용할 수 있도록 준비하였다.

* FAQ 데이터의 경우에도 크롤링을 통해 수집하였으며, 질문(question)과 답변(answer)을 구분하여 JSON 파일 형태로 저장하였다.
* 이러한 전처리 과정을 통해 데이터를 구조화하고, 효율적으로 모델에 활용할 수 있는 형태로 가공하였습니다.

1. **모델 설계**

**1.1 모델 유형**

* **RAG(Retrieval-Augmented Generation) 기반 챗봇 모델**
  + ChromaDB를 활용한 벡터 데이터베이스 구축
  + OpenAI GPT-4 기반 답변 생성
  + FAQ 매칭 시스템 통합
  + 다국어 임베딩 모델(distiluse-base-multilingual-cased-v2) 활용

**1.2 입력 및 출력**

**입력 데이터**

1. **문서 데이터**
   * JSON 형식의 청약 관련 문서
   * CSV 형식의 테이블 데이터
   * FAQ 크롤링 데이터
   * PDF 파싱 데이터
2. **메타데이터 구조**
   * 문서 제목
   * 섹션 타입 (subscription, schedule, notice, general)
   * 아파트명
   * 공고월
   * 문서 코드
   * 추가 속성 (가격 정보, 평면도 정보 등)
3. **사용자 쿼리**
   * 자연어 형태의 질문
   * 전처리된 키워드

**출력 데이터**

1. **응답 구조**
   * 참고 문서 정보
   * 핵심 답변
   * 상세 정보
   * 주의사항
2. **응답 포맷**
   * YYYY.MM.DD 형식의 날짜
   * 표 형식의 수치 데이터
   * 출처가 포함된 텍스트

**1.3 하이퍼파라미터**

### 1. 임베딩 및 검색 관련

* **청크 크기**: 500 (max\_chunk\_size)
* **청크 오버랩**: 15 (overlap)
* **검색 결과 수**: 10 (n\_results)
* **ChromaDB 처리 단위**: 150 (chunk\_size)
* **유사도 임계값**: 0.5 (similarity threshold)

### 2. GPT-4 생성 파라미터

* **Temperature**: 0.4
* **Maximum Tokens**: 1,000
* **Model**: gpt-4

### 3. 데이터 처리 설정

* **문서 타입 분류 기준**
  + 청약 관련: ['청약', '신청', '당첨자', '모집']
  + 일정 관련: ['일정', '계획', '날짜', '기간', '예정']
  + 공지 관련: ['유의', '주의', '안내', '참고', '확인']

### 4. FAQ 매칭 파라미터

* **키워드 기반 매칭**
* **다중 키워드 매칭 시 교집합 크기 기준 최적 답변 선정**