딥러닝 기반 논문 질의응답 챗봇 개발 계획서

1. 프로젝트 배경 및 선행 연구

배경

논문에 대한 질의응답 시스템은 연구자들이 방대한 양의 문서에서 필요한 정보를 빠르게 찾아내는 데 도움을 줄 수 있습니다. LangChain을 사용하면 이러한 시스템을 쉽게 구축할 수 있으며, 특히 NLP와 딥러닝 모델 통합이 용이해집니다.

선행 연구

- 파인튜닝 모델 받아서 나만의 로컬 LLM 호스팅하기 + RAG
  - 🔥성능이 놀라워요🔥 무료로 한국어🇰🇷 파인튜닝 모델 받아서 나만의 로컬 LLM 호스팅 하기(#LangServe) + #RAG 까지!! (youtube.com)

2. 프로젝트 목표 및 개요

프로젝트 목표

LangChain을 활용하여 논문 내용을 기반으로 질의응답을 수행할 수 있는 챗봇 시스템을 개발하고, 연구자들이 논문 정보를 효과적으로 검색하고 활용할 수 있도록 하는 것입니다.

개요

이 프로젝트는 다음 단계를 포함합니다:

1. 데이터 수집 및 전처리
2. LangChain 모델 및 데이터 통합
3. 챗봇 시스템 개발
4. 평가 및 성능 개선
5. 결과 문서화 및 배포

3. 기술 스택

프로그래밍 언어

- **Python**: NLP 및 딥러닝 모델 개발

프레임워크 및 라이브러리

- **LangChain**: NLP 기능 통합 및 모델 연결
- **OpenAI API**: 사전 훈련된 NLP 모델 (GPT 등)
- **TensorFlow**: 딥러닝 모델 학습 및 개발

개발 도구

- **Jupyter Notebook**: 데이터 분석 및 모델 실험 환경

4. 프로그램 구조

구조 개요

프로그램은 데이터 수집 및 전처리, LangChain 모델 설정 및 통합, 챗봇 개발, 평가 및 개선의 네 가지 주요 모듈로 나뉩니다.

모듈 설명

1. **데이터 수집 및 전처리 모듈**
    - 논문 데이터 수집: Google Scholar, PubMed, arXiv 등에서 논문 다운로드
    - 텍스트 전처리: 텍스트 정제, 토큰화, 정규화
2. **LangChain 모델 및 데이터 통합 모듈**
    - **LangChain 설정**: LangChain 라이브러리를 사용하여 언어 모델과 데이터 통합
    - **문서 검색**: LangChain을 활용한 문서 검색 기능 구현
    - **질의응답 시스템**: 질문에 대한 응답을 생성하기 위한 LangChain 및 NLP 모델 활용
3. **챗봇 시스템 개발 모듈**
    - **API 통합**: LangChain 및 딥러닝 모델을 챗봇 API와 통합
4. **평가 및 개선 모듈**
    - **모델 평가**: 챗봇의 정확성, 응답 적합도 평가
    - **성능 개선**: 사용자 피드백을 기반으로 모델 및 시스템 개선

5. 데이터 수집 방법 및 가능 여부

데이터 수집 방법

- **논문 데이터베이스**: Google Scholar, PubMed, arXiv 등에서 논문 다운로드
- **API 활용**: 논문 메타데이터와 텍스트 추출을 위한 API 사용

데이터 가능 여부

- 공개된 논문 데이터베이스에서 데이터 수집 가능
- LangChain을 활용하여 논문 데이터와 통합하여 활용 가능

6. 기본적인 데이터 분석

데이터 전처리

- **텍스트 정제**: 특수문자 제거, 불용어 제거
- **토큰화**: 문장 및 단어 단위로 분리
- **정규화**: 소문자 변환, 어간 추출

기본 분석 내용

- **문서의 주요 키워드 추출**: TF-IDF, LDA 등
- **질문-응답 쌍 분석**: 질문과 답변의 유사성 분석

7. 예상 결과물

모델 및 시스템

- **딥러닝 기반 질문-응답 모델**: GPT 기반의 모델
- **웹 기반 챗봇**: LangChain을 통합한 시스템

문서 및 보고서

- **프로젝트 보고서**: 프로젝트 개요, 방법론, 결과 및 결론 포함
- **기술 문서**: 시스템 설계, 코드 설명, 사용 방법
- **튜토리얼**: 사용자 가이드 및 챗봇 사용법

8. 일정

1. **일차**: 데이터 수집 및 전처리

2. **일차**: LangChain 모델 설정 및 파인튜닝
3. **일차**: 챗봇 개발 및 통합
4. **일차**: 모델 평가 및 성능 개선
5. **일차**: 결과 문서화 및 발표