캡스톤디자인 II 계획서

<table>
<tr><td rowspan="2">제 목</td><td>국문</td><td colspan="4">키워드 기반 경제 콘텐츠 큐레이션</td></tr>
<tr><td>영문</td><td colspan="4">Keyword-based economic contents curation</td></tr>
<tr><td>프로젝트 목표
(500자
내외)</td><td colspan="5">현재 재테크에 대한 관심이 뜨거워지는 가운데, 이에 대한 여러 정보를 얻기 위해 커뮤니티 게시글이나, 지인들로부터 정보를 얻던 과거와 달리 현재는 뉴스, 신문 등 관련 기사를 통해 보다 더 전문적인 재테크 정보를 얻으려는 모습이 나타나고 있다. 이에 대하여, 바쁜 현대사회에서 모든 경제 뉴스를 전부 보는 것은 어려운 일이기에, 용이하게 재테크 관련 정보를 얻기 위해서 인기 키워드로 연결되어있는, 재테크에 관한 콘텐츠들을 요약하여 보여주는 것을 목표로 한다.</td></tr>
<tr><td>프로젝트
내용</td><td colspan="5">이 프로젝트의 궁극적인 목표는, 경제 관련 인기 키워드를 콘텐츠 큐레이팅 모델을 통해 큐레이션 하는 것이다. 인기 키워드를 중심으로 연관된 재테크 관련 콘텐츠들을 크롤링한 후, 학습을 위한 전처리 과정을 거친다. 그리고, 전처리 된 데이터를 NLP 모델에 학습시켜, 모델을 통해 요약한다. 이와 관련해, 부동산 핀테크 기업인 루센트블록에서 캡스톤 자문하고 있다.</td></tr>
<tr><td>기대효과
(500자
이내)
(응용분야 및
활용범위)</td><td colspan="5">일별 인기 경제 키워드를 기반으로 주요 내용들을 요약, 설명, 추천글 등의 형태로 사용자에게 친화적으로 전달해준다. 이로써 일일이 찾아보는 수고를 덜어주어 시간을 절약하고 특히 경제 관련 기사들(투자, 부동산, 금융 등)과 같이 다양한 분야의 정보를 얻을 때 도움을 줄 수 있다.</td></tr>
<tr><td>중심어(국문)</td><td>데이터 크롤링</td><td>텍스트 전처리</td><td>자연어 처리</td><td>문서 요약</td></tr>
<tr><td>Keywords
(english)</td><td>Data Crawling</td><td>Text
Preprocessing</td><td>NLP</td><td>Text
Summarization</td></tr>
<tr><td>멘토</td><td>소속</td><td colspan="2">루센트 블록</td><td>이름</td><td>진준호</td></tr>
<tr><td rowspan="4">팀
구성원</td><td>학년/
반</td><td>·학 번</td><td>이 름</td><td colspan="2">연락처(전화번호/이메일)</td></tr>
<tr><td>4학년</td><td>20181622</td><td>송재민</td><td colspan="2">010-5118-4832/20181622@edu.hanbat.ac.kr</td></tr>
<tr><td>4학년</td><td>20181617</td><td>박재필</td><td colspan="2">010-2936-8255/20181617@edu.hanbat.ac.kr</td></tr>
<tr><td>4학년</td><td>20181630</td><td>이용경</td><td colspan="2">010-5784-0928/20181630@edu.hanbat.ac.kr</td></tr>
<tr><td colspan="6">컴퓨터공학과의 캡스톤디자인 관리규정과 모든 지시사항을 준수하면서 본 캡스톤디자인을 성실히 수행하고자 아래와 같이 계획서를 제출합니다.
2023년 6월 19일
책 임 자 : 송 재 민 (인)
지도교수 : 장 한 얼 (인)</td></tr>
</table>

캡스톤디자인 계획서

1. 캡스톤디자인의 배경 및 필요성

1990년대 이전에는 금리가 연 10~20%인 예금 상품이 적지 않았다. 그러므로, 은행에 저축만 해두어도 높은 금리 때문에 재산 형성이 가능했다. 그러나, 요즘은 과거에 비해 금리가 낮아졌고, 집값은 급등하였으며, 물가도 매년 오르기 시작하면서 인플레이션이 발생하고 있다. 그래서 자신의 근로 소득만으로는 재산을 형성하기에 어려움이 발생하였고, 재테크의 방식으로 저축보다는 투자를 선호하는 사람들이 증가하였다. 이렇게 자신의 소득과 수입을 늘리는 방법을 찾는 재테크에 대한 관심이 높아지면서, 정보의 중요성이 심화하였는데, 오늘날에 이르러 무더기로 쏟아져 나오는 뉴스와 같은 정보 들을 일일이 다 확인하고 정리하기에는 너무 많은 시간이 소요되기 때문에, 간단함과 편리함을 추구하는 사회적 요구에 따라, 존재하는 많은 정보 중에서 현재 인기 키워드와 그에 관련된 부동산 등의 재테크 관련 정보들이 무엇이 있는지 요약해주는 큐레이션 된 콘텐츠의 필요성이 대두된다.

2. 캡스톤디자인 목표 및 비전

캡스톤디자인1를 진행하면서 크롤링을 통한 뉴스 기사 추출과 뉴스 기사 간의 관련도 있는 데이터 추출과 요약까지의 기본적인 토대는 갖춰졌다. 하지만 NLP 모델에 사용한 학습 데이터의 부족과 아직 만족할만한 요약문의 퀄리티, 크롤링의 소요 시간, 데이터 간의 관련도 높은 데이터 추출 등 해결해야 할 보완점 및 문제점이 존재한다.

첫 번째 목표는 문서 요약 및 문장 작성에 사용할 NLP 모델을 작성하는 것이고, 두 번째 목표는 크롤링의 소요 시간 및 크롤링 풀 확장을 위한 데이터 수집에 질과 양을 높여 활용성을 높이는 것이다. 마지막으로 키워드들 간의 관련도 높은 데이터를 추출하여 결과값으로 나올 요약문이나 작성될 문장의 퀄리티를 높여 완성도를 높이는 것이다.

추후 이 프로젝트가 완성된다면 바쁜 현대인들을 위해 정보 검색의 시간을 줄이고, 정확한 정보를 전달하기 위하여 다양한 뉴스 기사를 수집하고, 수집된 기사를 자연어 처리 모델을 활용하여 경제 관련 정보를 효율적으로 탐색할 수 있어 주식, 부동산 등과 같은 사용자의 재테크 활동에 도움을 줄 수 있다.

3. 캡스톤디자인 내용

1) 캡스톤 요구사항

기능적 요구사항

1. 네이버 뉴스 크롤링
 - 경제 관련 키워드들의 뉴스 기사를 공신력 있는 상위 20개의 언론사를 선정하여 해당 언론사의 기사들을 수집하여 텍스트 전처리 후, 제목, 언론사, url, 기사 본문 형태로 크롤링

2. 크롤링 된 데이터들 중 키워드와 관련 데이터만 추출
 - 크롤링한 데이터들을 k-means 군집 알고리즘을 통해 군집을 형성하고 형성된 군집 중 키워드와 관련된 군집만 추출함.

3. NLP 모델을 활용하여 내용 요약
 - 한국어 자연어 처리 모델 중 하나인 Ko-BERT 모델 기반으로 여러 가지 기사들을 요약하고, 키워드와 관련 있는 내용을 추출함

4. 요약된 내용을 기사 형식으로
 - 모델을 통해 요약된 내용을 사용자가 보기 쉬운 형태로 제공함으로써 경제 관련 정보를 얻는데 도움을 준다.

5. 클라우드 환경 이용
- 클라우드 환경의 장점인 유연성을 이용해, 가장 시간이 오래 걸리는 작업인 Crawling에 대한 부담을 감소시킨다.

6. Django 웹 프레임 워크 사용
- 풀스택 프레임워크인 Django를 사용해, 사용자에게 웹 형식으로 정보를 제공한다.

비기능적 요구사항

、 성능 요구사항
- 기사 작성을 위해 필요한 기사를 크롤링하는데 걸리는 시간
- 최근 정보를 빠르게 전달해야 하므로 빠른 기사 작성 필요

、인터페이스 요구사항
- 인기 경제 키워드를 통해 키워드 기반 콘텐츠 출력
- 이미 입력된 키워드 관련 데이터 확인 및 업데이트

、유지보수 요구사항
- 한번 입력한 키워드 관련 데이터 DB에 저장 유지
- 저장된 데이터의 최신 데이터 업데이트 지속

、품질 요구사항
- 한 번에 많은 양의 크롤링 시 해당 사이트에서 접근을 방지할 수 있음. 이를 예방하고자 time 라이브러리의 sleep() 사용

2) 추진일정

수행 목표	수행기간(월)			
	6월	7월	8월	9월
크롤링 모델 보완				
DB 관리				
Pre-processing				
Post-processing				
UI 구성 및 웹 배포				

4. 캡스톤디자인 추진전략 및 방법

1) 캡스톤디자인 목표 달성을 위한 추진전략, 수행방법 및 추진절차

(1) 캡스톤디자인에 대한 이해

프로그램의 구현을 위해서 데이터 분석에 가장 많이 사용되는 언어인 Python을 기반으로 프로젝트를 진행하였다. 콘텐츠 큐레이션을 하기 위해서는 자연어 생성 모델을 사용해야 하는데 모델의 빠른 최적화를 위하여 학교에서 제공하는 GPU 서버를 사용하며, 사용자에게 제공되는 데이터는 객관적인 데이터를 제공해야 하기에 공신력 있는 언론사를 20개를 채택하여 기사 데이터를 가져오기로 하였다.

(2) 캡스톤디자인 경험

캡스톤디자인을 진행하기 위하여 선행된 경험으로 3학년 과목 중 마이크로프로세서와 모바일 컴퓨팅과 응용, 인공지능, 데이터베이스 과목에서의 프로젝트를 기반으로 파이썬 라이브러리인 Requests, Selenium, BeautifulSoup 등을 활용하여 다양한 형식의 웹사이트를 크롤링 하였고, DB의 기본적인 틀이나 저장 방식 등을 활용해 DB를 관리하였다.

(3) 검증된 멘토 활용

루센트블록의 CTO와의 미팅을 통해 현재 진행 상황을 피드백을 받으며 프로젝트의 방향성을 확립한다.

(4) 프로젝트 관리체계 수립

역할 분담을 크게 프론트엔드 및 크롤링 1명, 백엔드 및 DB 관리 1명, 인공지능 모델 구현 담당(자연어처리) 1명으로 구성했다. 제일 먼저 크롤링은 다양한 언론사의 웹 환경을 조사하고, 이를 토대로 자신이 원하는 경제 키워드에 대한 결과를 수집한다. 이에 맞춰 크롤링된 데이터들 중 관련도 있는 데이터만 남기고 자연어 처리는 다양한 기사를 요약할 수 있는 모델을 이용하여 환경을 조성한 후, 모델의 파라미터를 조정하여 더욱더 의미 있는 텍스트를 생성하여 고객에게 제공하는 것이 프로젝트의 목표이다.

2) 캡스톤디자인 목표 달성을 위한 팀 구성 체계 및 역할

팀원	역할
송재민(팀장)	백엔드, DB 관리, Pre-processing
이용경	프론트엔드, Crawling
박재필	Post-processing, NLP 모델 작성

5. 캡스톤디자인 결과의 활용방안

경제 관련 키워드를 사용하여 뉴스들을 검색하고, 검색된 기사들을 크롤링하여 다량의 뉴스 기사를 수집한다. 그리고 이를 요약함에 있어서 여러 검증을 거치며 다중 키워드와 최대로 관련된 기사를 수집하기에, 사용자에게 친화적이며 다중문서를 요약하여 출력하기에 사용자로 하여금 시간 절약에 큰 힘을 실어줄 수 있다. 특히 경제 관련 기사를 위주로 기사를 가져오기에 투자, 부동산, 금융 관련 등 경제 관련 다양한 분야에서 도움을 줄 수 있다

6. 참고문헌

1. Yang Zhang, Yunqing Xia, Yi Liu, and Wenmin Wang. 2015. Clustering Sentences with Density Peaks for Multi-document Summarization. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1262–1267, Denver, Colorado. Association for Computational Linguistics.

캡스톤디자인 II 계획발표 채점표

팀 구성원	학년/반	학 번	이 름
제 목			

항목	점수				
	1	2	3	4	5
1. 프로젝트 주제의 필요성이나 중요성이 적절히 서술되었는가?					
2. 국내외 동향(문제 제기), 주요 기능(특징 포함) 및 범위가 적절히 서술되었는가?					
3. 기대효과(사회적, 기술적, 경제적 파급효과)가 적절히 서술되었는가?					
4. 추진 전략과 수행방법이 적절한가?					
5. 팀 구성과 역할 분담이 적절히 이루어졌는가?					
합계					

*수정 및 개선 의견

2013년 월 일

심사위원 : (인)

※ 채점은 각 영역별 5점 만점을 기준으로 채점함.(상 5, 중 3, 하 1)

※ 계획서와 발표내용을 참고하여 채점표에 따라 평가함.