캡스톤디자인 I 계획서

(표지)

제 목	국문	키워드 기반 콘텐츠 큐레이션
	영문	Keyword-based contents curation
프로젝트 목표 (500자 내외)	현재 재테크에 대한 관심이 뜨거워지는 가운데, 이에 대한 여러 정보들을 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.	
프로젝트 내용	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, 전처리된 데이터를 NLP 모델에 학습시켜, 모델을 통해 요약한다. 이와 관련해, 부동산 핀테크 기업인 루센트 블록에게 캡스톤 자문하고 있다.	

중심어(국문)	데이터 크롤링	텍스트 전처리	자연어 처리	문서 요약
Keywords (english)	Data Crawling	Text Preprocessing	NLP	Text Summarization

멘토	소속	루센트블록	이름	진준호

팀 구성원	학년/반	학 번	이 름	연락처(전화번호/이메일)
	4학년	20181622	송재민	010-5118-4832/20181622@edu.hanbat.ac.kr
	4학년	20181617	박재필	010-2936-8255/20181617@edu.hanbat.ac.kr
	4학년	20181630	이용경	010-5784-0928/20181630@edu.hanbat.ac.kr

컴퓨터공학과의 캡스톤디자인 관리규정과 모든 지시사항을 준수하면서 본 캡스톤디자인을 성실히 수행하고자 아래와 같이 계획서를 제출합니다.

2023년 3월 3일

책 임 자 : 송 재 민 (인)

지도교수 : 장 한 얼 (인)

(내용)

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

바쁜 현대인들을 위해 정보 검색의 시간을 줄이고, 정확한 정보를 전달하기 위하여 다양한 뉴스 기사를 수집하고, 수집된 기사를 자연어 처리 모델을 활용하여 부동산 시장 예측에 도움을 줄 수 있다.

3. 캡스톤디자인 내용

기능적 요구사항

1. 인기 키워드 추출
 - 실시간 기사 내용을 제공하기 위하여 구글 인기 검색어에서 키워드를 채택하여 관련된 내용을 검색한다.

2. 네이버, 다음 특정 언론사들의 뉴스 크롤링
 - 공신력 있는 상위 20개의 언론사의 기사를 선정하여 해당 언론사의 기사들을 수집하여 텍스트 전처리 후, 제목, 언론사, url, 기사 본문형태로 csv파일로 저장.

3. NLP 모델을 활용하여 내용 요약
 - 한국어 자연어 처리 모델 중 하나인 Ko-BERT 모델 기반으로 여러가지 기사들을 요약하고, 부동산 관련있는 내용을 추출함.

4. 요약된 내용을 기사형식으로
 - 모델을 통해 요약된 내용을 루센트 블록의 포스팅 형식에 맞추어 사용자에게 정보를 제공한다.

비기능적 요구사항

、 성능 요구사항
- 크롤링 소요 시간
  기사 작성을 위해 필요한 기사를 크롤을 하는데 소요되는 시간
- 기사 작성 완료 시간
  최근 정보를 빠르게 전달해야 하기 때문에 빠른 기사 작성 필요

、인터페이스 요구사항
- 큐레이션 콘텐츠 Interface
  당일 인기 키워드 및 키워드를 통한 뉴스 기사를 출력

ㆍ유지보수 요구사항

실시간 키워드 갱신

매일 실시간 키워드를 갱신하여 새로운 기사를 작성하기 때문에 지속적인 업데이트 및 유지 관리 필요

ㆍ품질 요구사항

- 크롤러 차단 방지

한번에 많은 양의 크롤링 시 해당 사이트에서 접근을 방지할 수 있다.
이를 예방하고자 time 라이브러리의 sleep() 사용

4. 캡스톤디자인 추진전략 및 방법

(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 실시간성과 객관적인 데이터를 제공해야 하기에 구글의 인기검색어 순위를 통해 검색어를 정하고, 가장 공신력있는 언론사를 20개를 채택하여 기사 데이터를 가져오기로 하였다.

(2) 캡스톤디자인 경험

캡스톤디자인을 진행하기 위하여 선행된 경험으로 3학년 과목 중 마이크로프로세서와 모바일 컴퓨팅과 응용, 인공지능 과목에서의 프로젝트를 기반으로 파이썬 라이브러리인 Requests, Selenium, BeautifulSoup 등을 활용하여 다양한 형식의 웹사이트를 크롤링 하였고, 이를 csv 파일로 저장하며, Pytorch를 사용하여 사전학습된 모델로 요약을 하여 사용자에게 텍스트 형식으로 제공될 것이다.

(3) 검증된 멘토 활용

(4) 프로젝트 관리체계 수립

역할 분담을 크게 데이터 수집 및 전처리(크롤링)를 2명, 인공지능 모델 구현 담당(자연어처리) 1명으로 구성했다. 제일 먼저 크롤링은 다양한 언론사의 웹 환경을 조사하고, 이를 토대로 자신이 원하는 검색어에 대한 결과를 수집한다. 이에 맞춰 자연어 처리는 다양한 기사를 요약할 수 있는 모델을 이용하여 환경을 조성한 후, 모델의 파라미터를 조정하여 더욱 더 의미있는 텍스트를 생성하여 고객에게 제공하는 것이 프로젝트의 목표이다.

팀원	역할
송재민 (팀장)	네이버 뉴스 콘텐츠 크롤링을 통한 자료 수집 및 데이터 전처리
이용경	다음 뉴스 콘텐츠 크롤링을 통한 자료 수집 및 데이터 전처리
박재필	NLP 모델링 기반 콘텐츠 큐레이팅 모델 개발

+ 프로젝트 과정에서 추가적으로 작업이 필요할시, 역할분담을 추가할 예정임