캡스톤디자인(2) 계획서

프로젝트 명

Web Crawling, NLP, Word Cloud를 이용한 기사 통합 요약 시스템

프로젝트 구성원

팀 no.	학번	학과	이름

프로젝트 개요

현재 우리가 살아가고 있는 지식 정보화 시대는 사회 일반 산업 자원으로 활용할 수 있도록 가공된 정보와 지식이 사회 구조나 관습 및 인간의 가치관에 큰 영향을 미치는 시대이다. 이러한 정보의 습득은 주로 검색을 통해 이루어진다. 정보의 양이 거대해짐에 따라 기대하는 정보의 선택적인 습득이 어려워지고 이를 검색하기 위한 키워드를 설정하는 것 역시 사람들에게 또다른 과제로 주어진다.

Word Cloud는 핵심 단어 시각화 표현 기법으로 문서의 문구와 단어를 분석하여 중요도와 사용 빈도를 기반으로 문서의 직관적인 파악을 보조한다. KorLex는 단어의 동의어 집합, 상위어, 그리고 하위어에 대한 정보를 바탕으로 하는 한국어 Wordnet이다. 이러한 Wordnet을 활용하면 기존에 보유하고 있던 정보 또는 단어의 양을 확장할 수 있다. Word2Vec은 단어를 벡터로 변환하는 도구로 단어의 의미와 관련성을 수학적으로 표현하는 기법이다. 이는 Neural Network 기반의 언어 모델로 주변 단어의 문맥을 이용하여 단어를 벡터로 표현한다. 이를 통해 단어 간 유사성을 계산하고 단어 간 관련성을 파악할 수 있다. Word2Vec을 이용하면 주어진 단어와 Wordnet으로 확장한 단어들 간의 벡터 유사성을 기반으로 키워드의 확장이 가능하다. 또한 이를 Word Cloud 생성 전처리 과정에 사용하면 비슷한 단어의 반복을 방지할 수 있고 관련 있는 단어들을 묶을 수 있다.

이 프로젝트에서는 위에서 언급한 기술들을 기반으로 사용자가 뉴스 검색 시 키워드를 입력하면 주간 뉴스 정보를 Word Cloud로 제공해주는 서비스를 개발한다. 사용자는 해당 키워드 단독, 또는 확장된 키워드에 대한 검색의 여부를 선택할 수 있다. 제공된 Word Cloud는 키워드에 해당하는 뉴스들에 대한 내용을 사용자가 한 눈에 요약하여 파악할 수 있게 돕는다. 이는 원하는 뉴스의 내용을 개별적으로 읽고 정리하기 힘든 현대인들에게 정보 습득의 효율을 높일 것을 기대한다.

프로젝트 진행 계획

적용 기술

- Python
- 자연어 처리(Word Cloud, WordNet): KoNLPy, NLTK, WordCloud, Word2Vec, KorLex
- Web Crawling: requests, beautifulsoup4

진행 방법

프로젝트 개요 도출 → 유사 서비스 검토 → 프로젝트 주제 기능 확정 → 소요기술, 소요자원 검토 및 학습→ 구현 → 테스트 → 결과보고서 제출

일정 계획

날짜	일정
1학기	데이터 크롤링 및 전처리 불용어처리
9월 – 10월	전처리에 Word2Vec 도입
11월	KorLex 도입 프로그램 UI 구현
12월	최종 보고서 작성