

## 캡스톤디자인(1) 계획서

### 프로젝트 명

Web Crawling, NLP, Word Cloud를 이용한 기사 통합 검색 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. 이와 같은 정보를 습득할 때 사람들은 주로 검색을 통해 습득하고 있다. 하지만 정보의 양이 워낙 방대하기 때문에 사람들은 원하는 정보를 검색하여도 무엇부터 확인해야하는지 혼란스럽거나 원하는 정보를 검색하는 키워드 자체를 선택하는데 어려움을 느낄 수 있으며 이에 검색하여 정보를 습득하기까지의 시간이 길어질 수 있다.

Word cloud는 핵심 단어 시각화 표현 기법으로 문서의 문구와 단어를 분석하여 중요도나 사용빈도를 직관적으로 파악할 수 있게 돕는다. 따라서 Word Cloud를 활용하면 사용자가 문서의 주요 키워드나 중요도를 한눈에 보고 이해할 수 있다. 또한 KorLex는 단어의 동의어 집합, 상위어, 그리고 하위어에 대한 정보를 바탕으로 하는 한국어 Wordnet이다.

이번 프로젝트에서는 위 두가지를 활용하여 사용자가 원하는 정보를 포함한 최신 뉴스를 검색하기 위해 키워드를 입력하면 해당하는 뉴스 매체들의 주간 뉴스 정보를 Word Cloud로 제공해주는 서비스를 개발한다. 사용자는 해당 키워드 단독, 동의어, 상위어, 하위어에 대한 검색의 여부를 선택할 수 있으며 제공된 Word Cloud의 키워드를 클릭하면 해당 키워드가 포함되어있는 뉴스들을 볼 수 있다. 이를 통해 키워드 검색 시 해당 키워드에 대한 정보를 한 눈에 파악 한 뒤 세부 내용을 확인할 수 있어 정보 습득의 효율을 높일 것을 기대한다.

### 프로젝트 진행 계획

#### 적용 기술

- Python
- 자연어 처리(Word Cloud, WordNet): KoNLPy, NLTK, wordcloud, pytagcloud, 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
3월	1주차: 프로젝트 개요 도출 2주차: 유사 서비스 검토 3주차: 기능 확정
4월	3월 4주차 ~ 4월 1주차: 소요 기술, 소요 자원 검토 4월 2주차 ~ 4월 4주차: 소요 기술 학습
5월	5월 1주차 ~ 5월 4주차: 구현
6월~여름방학	구현 진행: 기능 구현 80% 완료 목표