제6회 산학연계 SW프로젝트 최종보고서

팀명	도요새와사람들
프로젝트 수행기간	2021. 07. 01 ~ 2022. 05. 31
프로젝트 주제	뉴스/블로그 스크래핑 데이터활용 딥러닝 자연어처리 기반 미래유망기술 센싱 및 애널리틱 기술개발
지도 교수	정보융합학부 조재희 교수
참여업체 명	㈜티엠넘버스

2022. 06. 10.

광운대학교
KwangWoon University

산학연계SW프로젝트 최종보고서

<table>
<tr><td>팀 명</td><td colspan="4">도요새와 사람들</td></tr>
<tr><td>과제 명</td><td colspan="4">뉴스/블로그 스크래핑 데이터활용 딥러닝 자연어처리 기반 미래유망기술 센싱 및 애널리틱 기술개발</td></tr>
<tr><td>GitHub URL</td><td colspan="4">https://github.com/kimtaeyong98/Technology-Sensing-Evaluation</td></tr>
<tr><td>YouTube URL</td><td colspan="4">https://youtu.be/w4wPPvA-5e8</td></tr>
<tr><td>수행기간</td><td colspan="4">2021년 07월 01일 ~ 2022년 05월 31일</td></tr>
<tr><td>과제비</td><td colspan="4">총 1,800,000 원</td></tr>
<tr><td>지도교수</td><td>성 명</td><td>조재희</td><td>학 부</td><td>정보융합학부</td></tr>
<tr><td rowspan="6">참여학생</td><td>성 명</td><td>학 부</td><td>학 번</td><td>email</td></tr>
<tr><td>박지영</td><td>정보융합학부</td><td>2018204085</td><td>jessy34150@naver.com</td></tr>
<tr><td>오재호</td><td>정보융합학부</td><td>2017204065</td><td>woghsla20@naver.com</td></tr>
<tr><td>김태용</td><td>정보융합학부</td><td>2017204004</td><td>kasamdi5@naver.com</td></tr>
<tr><td>소현수</td><td>정보융합학부</td><td>2017204069</td><td>tgt5248@naver.com</td></tr>
<tr><td>김주한</td><td>정보융합학부</td><td>2017204047</td><td>mae054954@naver.com</td></tr>
<tr><td rowspan="2">참여업체</td><td>회사명</td><td>㈜티엠넘버스</td><td>담당자</td><td>최점기</td></tr>
<tr><td>연락처</td><td>010-5364-9867</td><td>email</td><td>pointkey.choi@trimaran.co.kr</td></tr>
</table>

『산학연계 SW프로젝트』 지원 계획에 따라 최종보고서를 제출합니다.

2022년 06 월 10 일

팀 장 박지영 (인)
팀 원 오재호 (인)
팀 원 김태용 (인)
팀 원 소현수 (인)
팀 원 김주한 (인)

지도교수 (인)

광운대학교 소프트웨어융합대학 귀하

목 차

1. 과제의 개요

가. 배경 및 필요성

기술은 시간이 갈수록 빠르게 발전해가고 있고, 경제 사회적 영향력도 지속해서 커지고 있다. 이런 과정에서 어떤 기술 분야에 먼저 투자할 것인지는 중요시된다. 하지만, 기술의 우선순위 결정이 어떻게 수행될 수 있는지, 특정 아이디어가 정말 실효성 있는 좋은 아이디어인지 판단하는 문제는 쉬운 일이 아니었다.

특히 1980년대 이후에는, 과학기술 투자에 대해 정부나 정책 입안자들은 구체적인 목표를 설정하고 우선순위를 매길 것을 촉구받았다. 투입할 수 있는 자원은 한정되어 있기에 공공영역의 과학기술 연구·개발의 경우 특히나 투자의 효율화가 중요했다. 이를 위해서는 미래에 상대적으로 더 큰 효용가치를 안겨다 줄 유망기술을 발굴하여 투자 해야 했다.

이에 따라 자연스럽게 과학기술 추세 분석과 미래기술 예측을 바탕으로 R&D 사업을 기획하는 방법은 일반적으로 전문가의 지식과 직관적 판단의 영역에 대부분 맡길 수밖에 없었다. 이러한 방법은 실질적으로 전문가의 의견에 전적으로 의존하는 경향성이 두드러지기에 전문가의 편향된 의견이 나타날 수밖에 없다.
또한, 유망기술을 발굴하고 판단하기 위해서는 전문가들이 특허 기술을 직접 검색해서 전부 읽어보고 판단해야 했다. 이는 시간이 오래 걸릴 수밖에 없는 구조이다.

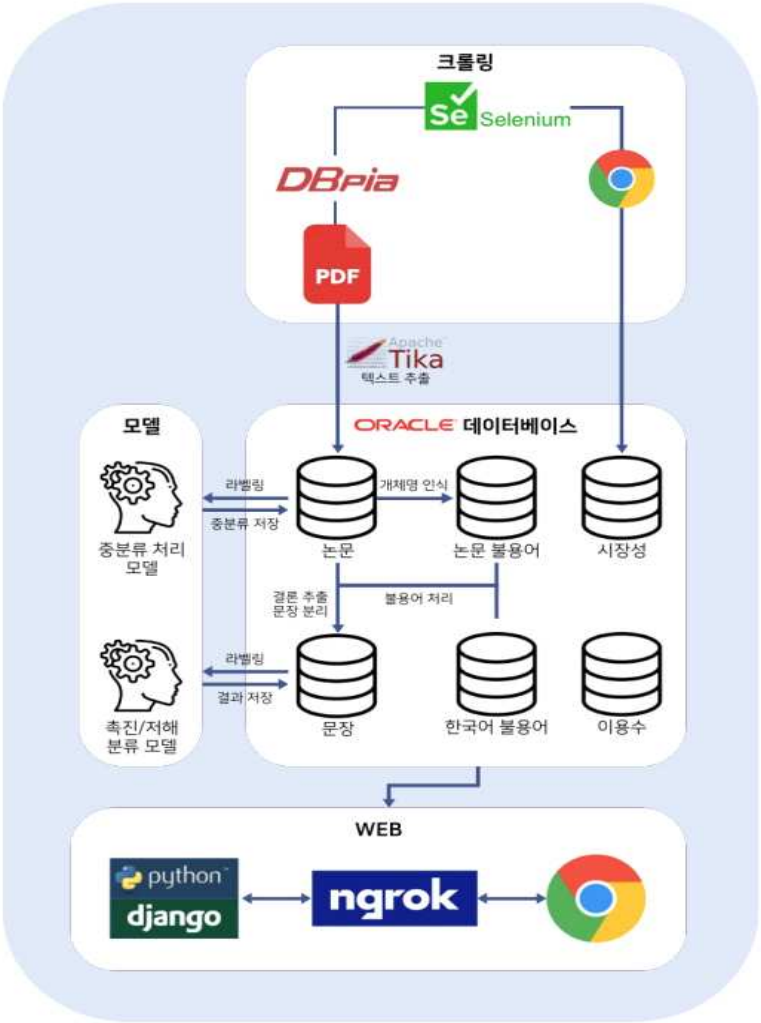
현재, 인공지능(Artificial Intelligence, AI) 및 기계학습(Machine Learning, ML) 및 자연어처리(Natural Language processing) 관련 기술이 눈부시게 발전하고 있고, 큰 주목도 받고 있다. 따라서 더욱 객관성과 확보한 형태의 미래 유망기술 발굴 모델을 만들 수 있는 여건이 완성되었다.

나. 목표

과제의 최종 목표는 객관성과 정합성을 확보한 형태의 미래 유망기술을 발굴하고 이를 심층 분석할 수 있는 웹 서비스를 구현하는 것이다. 이를 위해 기술에 대한 방대한 텍스트 데이터에서 촉진요인 및 저해요인 중립요인을 구분 할 수 있는 인공지능 모델이 필요하다. 인공지능 모델이 거대화 됨에 따라 많은 인공지능이 실시간 결과를 도출하기 어려워지고 있다. 따라서 실시간 웹서비스를 구현하기 위해서는 모델의 경량화가 필수적이다. 즉, 경량화 된 성능좋은 모델을 구축하여 실시간 서비스를 제공하고, 구축한 데이터에 대해 심층분석할 수 있는 대시보드 웹을 구현하는 것이 최종 목표이다.

다. 개발의 내용

개발 해야할 내용은 아래와 같다.



<시스템 아키텍처>

크게 3가지 부분으로 나누어 개발을 진행했다.

첫 번째, 데이터 파트

두 번째, 모델구축 파트

세 번째, 웹서비스 구현 파트

보다 자세한 개발 내용은 2-가 “설계 및 개발의 내용”에서 설명한다.

2. 과제의 내용

가. 설계 및 개발의 내용

1) 개념 설계 (구조 설계)

“1-다 시스템아키텍처”를 보면 크게 3가지 부분으로 설계된 것을 알 수 있다. 중앙에 구조에 어디서든 접근가능한 오라클 클라우드 데이터 베이스를 활용했다.

1. 데이터 파트에서 논문 데이터를 오라클 데이터 베이스에 저장한다.
2. 구출된 데이터로 문서를 주제별로 분류 하는 모델과, 텍스트 데이터의 촉진/중립/저해 요인을 추출할 수 있는 인공지능 모델을 구축한다.
3. 데이터에 대한 인공지능 결과를 다시 데이터 베이스에 저장한다.
4. 인공지능에 대한 결과까지 저장된 데이터를 웹에서 시각화한다.

2) 상세 설계 (기능 설계)

데이터 : 데이터 파트에서 필요한 기능은 5가지 이다.

1. 크롤링 : 도메인을 논문데이터로 지정했기 때문에, 논문 PDF 파일을 다운받는 크롤링 방식이 필요하다. 또한 아이템의 시장성 정보를 인터넷 상에서 크롤링 해야 한다.
2. PDF TO TEXT : PDF는 파일이기 때문에 저장되어 있는 텍스트를 추출하는 방법이 필요하다.
3. 문장분리 : 텍스트로 추출하고 난 후, 대량의 텍스트를 문장 단위로 분리하는 기술이 필요하다.
4. 불용어 구축 : 분석에 필요없는 단어를 정의하는 불용어 테이블이 필요하다
5. 데이터 베이스 저장 : 클라우드 방식으로 원격 저장소에 있는 데이터베이스에 저장할 수 있어야한다.

모델 구축 : 모델 구축 파트에서 필요한 기능은 2가지 이다.

1. 수집된 문서를 주제별로 분류
2. 분리된 문장들을 촉진/저해/중립으로 판단

웹 구현 : 웹 구현 파트에서 필요한 기능은 2가지이다.

1. 구축한 모델을 실시간으로 테스트
2. 구축한 데이터에 대한 모델 결과 대시보드 제작

3) 개발의 내용

데이터 파트

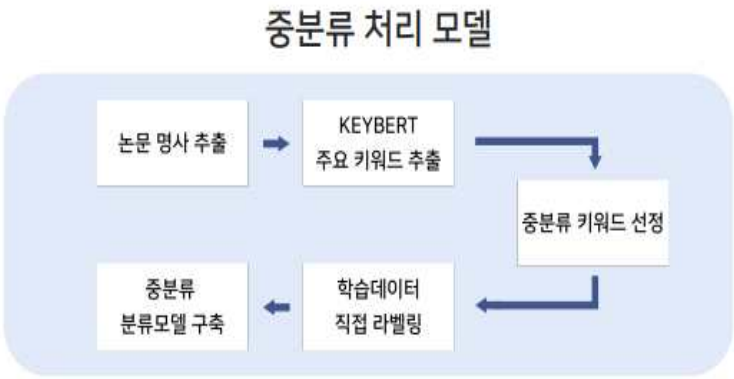
1. 크롤링 : 크롤링을 개발할 때는 selenium을 사용했다. selenium을 사용해 DBPIA 홈페이지에서 최근 3개년 논문PDF를 약 4300건 다운 받았다. 또한 논문에 대한 제목, 이용수, 년도 등을 함께 수집했다. 또한 네이버 기사, 블로그 등에서 시장에서 대한 정보도 함께 수집했다.
2. PDF TO TEXT : 수집한 pdf를 text로 변환하기 위해서 tika라이브러리를 사용하여 논문에서 텍스트를 추출 했고, 추가적으로 regular expression을 사용하여 논문의 결론 부분만 추출하여, 방대한 데이에서 필요한 부분만 남겼다.
3. 문장분리 : 문장단위로 모델에 학습시키기 위해 KoalaNLP를 활용하여 방대한 텍스트를 문장으로 분리했다.
4. 불용어 구축 : 사전에 정리되어 있는 한글 불용어 데이터셋을 데이터베이스에 저장하였고, 추가적으로

논문에서 나타날 수 있는 불용어인 저자명, 기관이름을 개체명인식 기술을 사용해 추출하여 저장했다.
5. 데이터 베이스 저장 : pdf -> 텍스트 -> 결론 추출 ->문장분리 -> 불용어처리 과정을 거친 데이터를 오라클 클라우드 데이터베이스에 저장했다.

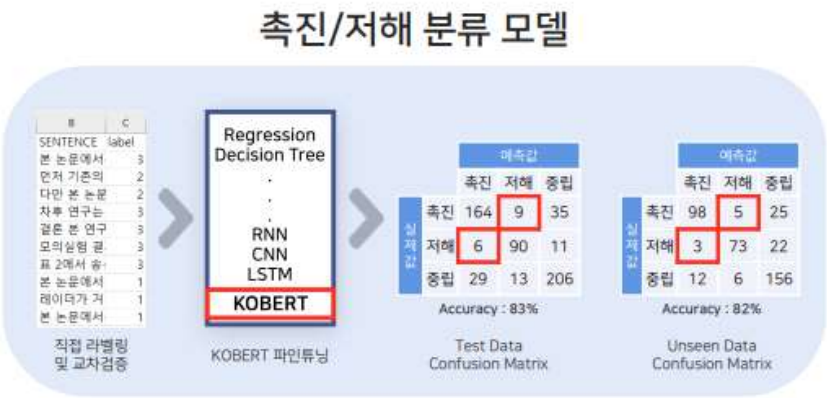
모델 구축 파트

모델 구축은 파이토치를 이용하여 진행했다.

1. 수집된 문서를 주제별로 분류 : 웹에서 대시보드를 구현할 때, 방대한 정보를 정제해서 보여주기 위해 논문별 주제 분리가 필요했다. keybert를 활용하여 각 논문에서 주제를 추출한 다음, 추출된 주제들 중에서 실제 분류로 사용할 키워드를 선정 했다. 그 후 선정한 키워드에 맞춰서 각 논문별 라벨링을 진행했고, 해당 데이터로 논문을 주제별로 분류 하는 모델을 구축했다. 사용한 알고리즘(모델)로는 기본적인 회귀 모델부터, 사전학습된 모델일 kobert, kogpt까지 사용 했고, 가장 성능이 좋았던 kobert를 채택했다.

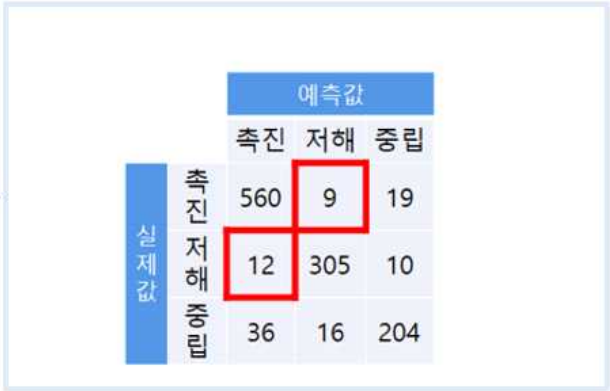


2. 분리된 문장들을 촉진/저해/중립으로 판단 : 분리된 문장들에 대해 지도학습을 수행하기 위해 촉진/저해/중립 3가지로 나누어 교차검증 라벨링을 진행했다. 분류모델과 마찬가지로 기본적인 회귀부터 사전학습된 모델까지 모두 사용해 비교한 결과, kobert 모델이 가장 성능이 좋아 채택했다. kobert 모델은 한국어 wiki로 사전학습된 모델이기 때문에 서술격어조가 많은 논문 도메인과 비슷하기 때문에 성능이 가장 높았던 것으로 유추된다. 또한 하이퍼 파라미터 튜닝을 통해 튜닝을 진행하지 않았을 때 정확도인 83%에서 91% 까지 정확도를 올렸다.



실제값 \ 예측값	촉진	저해	중립
촉진	98	5	25
저해	3	73	22
중립	12	6	156

팜플렛 제출시 모델 정확도(82%)



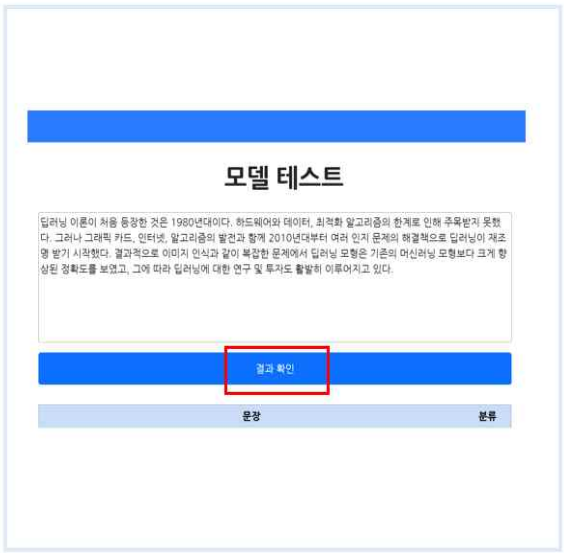
실제값 \ 예측값	촉진	저해	중립
촉진	560	9	19
저해	12	305	10
중립	36	16	204

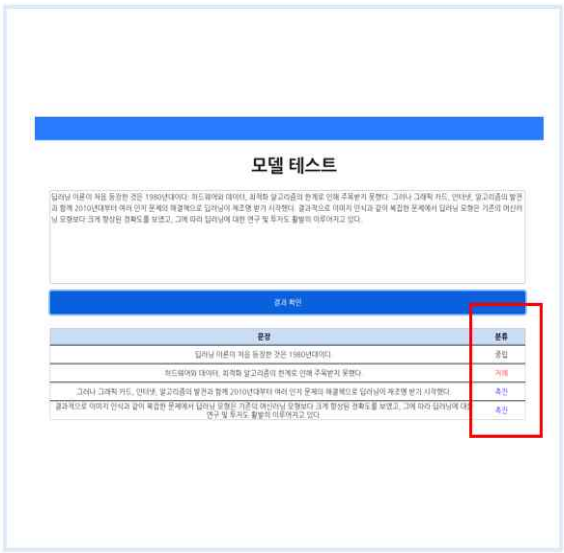
추가 데이터 학습 및 하이퍼 파라미터 튜닝을 통한 91%의 예측 정확도

웹 구현 파트

웹 구현에는 DJango, python, html, css, js, ajax, ngrok, docker 등이 사용되었다.

1. 구축한 모델을 실시간으로 테스트 : 구축한 모델을 웹에 이식하기 위해 모델의 파라미터 부분만 장고 서버에 저장하여 백엔드 로직에서 모델을 구동할 수 있게 했다. 따라서 사용자가 문단을 입력하면, 문장으로 분리 후 촉진 저해 중립을 판단하여, 확인 할 수 있다.



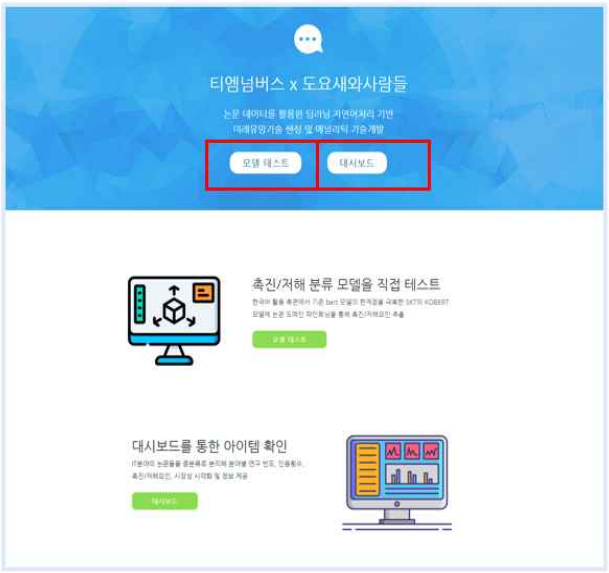


모델 테스트 페이지

문장 입력 후 결과 확인 시 분류 결과 화면

2. 구축한 데이터에 대한 모델 결과 대시보드 제작

구축 한 데이터와 모델에 대한 결과를 시각화 하기 위해 django에서 오라클 클라우드db로 접근해 데이터를 가져와 tag클라우드, 파이차트, 라인차트, 막대그래프 등으로 시각화하고, 문장별 촉진/저해에 대한 확률 값까지 확인 할수 있게 구현했다.



나. 수행 방법 및 추진 과정 (예시 두 개 중 적절한 것으로)

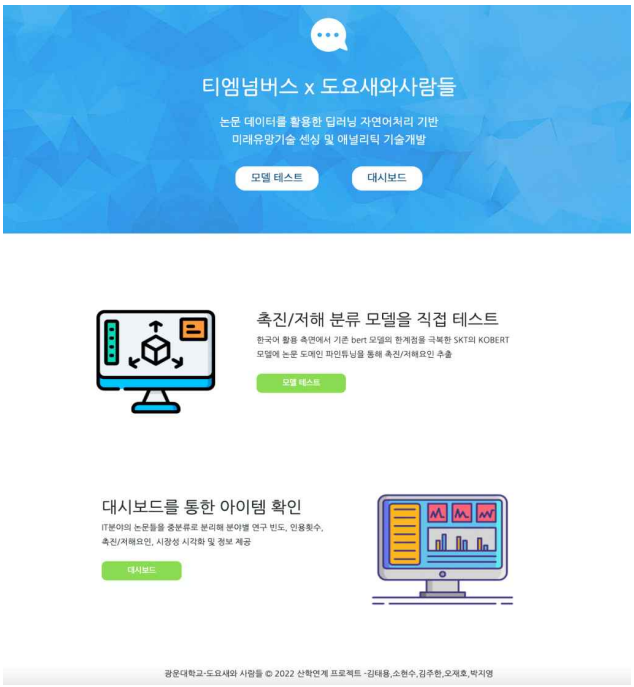
역할 분담

성명	역할
박지영	텍스트 마이닝, 데이터 시각화 및 웹 서비스 프론트엔드 총괄
오재호	텍스트 마이닝, 데이터 시각화 및 웹 서비스 백엔드 총괄
김주한	크롤링, 텍스트 전처리 및 웹 서비스 백엔드 총괄
김태용	크롤링, 텍스트 전처리 및 웹 서비스 백엔드 총괄
소현수	크롤링, 텍스트 전처리 및 웹 서비스 프론트엔드 총괄

수행 일정

일정 (월차)	내용	세부 내용	비고
1	프로젝트 주제 확립	회의를 통해 프로젝트의 정체성 및 주제 확립 업무 흐름 체계(WBS) 구성	
2	논문 텍스트 전처리	정규 표현식, Konlpy 등을 사용하여 전처리 직접 라벨링 및 교차검증	
3	데이터 수집	DBpia에서 논문 데이터 크롤링 크롤러 문제점 해결	
4	논문 텍스트 전처리	정규 표현식, Konlpy 등을 사용하여 전처리 직접 라벨링 및 교차검증	
5	텍스트 마이닝 및 모델 구축	토픽 모델링, 감정분석 등 여러가지 시도 가장 효율이 좋은 KOBERT 모델 사용 파인튜닝을 통해 모델 정확도 판단	
6	웹 구성 및 이식	Django를 사용하여 프로토타입 제작 웹 페이지에 적용이 가능하도록 백엔드 구성 정상적인 입출력이 가능한지 테스트	
7	웹 서비스 디자인 및 서버 연결	사용자 친화적 UI 구상하여 웹 디자인 다양한 시각화 방법 사용 서버를 연결한 웹페이지 서비스 제공	
8	테스트 및 마무리	반복적인 테스트 피드백 및 문제점 해결 발표 준비	

다. 최종 결과물



1. 메인페이지

프로젝트의 주제와 내용을 간략하게 설명하고

사용자가 직접 촉진/저해 분류 모델을 테스트 할 수 있는 페이지와

데이터베이스에 저장된 아이템들을 대시보드를 통해 확인하는 페이지로 나누었다.

2. 모델 테스트

한국어 활용 측면에서 기존 BERT모델의 한계점을 극복한 SKT의 KOBERT모델에

논문 도메인 파인튜닝을 진행한 모델을 사용자가 직접 테스트해 볼 수 있는 페이지로

기존 데이터베이스에 저장되어 있는 아이템 뿐만 아니라

사용자가 원하는 새로운 데이터에서도 촉진/저해 요인을 추출할 수 있다.

중분류명	개수	예시 아이템
네트워크	42	미래 대대급 전술 네트워크 구축을 위한 5G 기반 네트워크 활용방안
통신	41	스마트팩토리 PLC 데이터 수집을 위한 Modbus 통신 설계 및 구현
영상	22	영상의 이진화평면 분해에 기반한 확장된 블록매칭 잡음제거
데이터	20	복잡한 구조의 데이터 중복제거를 위한 효율적인 알고리즘 연구
인공지능	16	스마트시티 고층 광각 CCTV를 활용한 인공지능 재난 정보의 취득 및 대응 연구
융합	15	스마트 의료와 ICT융합보안에 관한 정책적 연구
환경	12	스마트팜 환경 관리를 위한 계측 시스템
사이버 공격	12	MEMS 센서대상 오류주입 공격 및 대응방법
컴퓨팅	10	엣지컴퓨팅기술의 변화와 동향
신경망	10	인공신경망을 활용한 항공산업 연구개발성과 평가 분석 및 예측
콘텐츠	9	5G 시대 콘텐츠의 변화와 과제
사업	9	IoT 클라우드 서비스 플랫폼 사례 연구 심플랫폼 창업과 성장
모형	9	연구장비산업의 비즈니스 생태계 모형 개발 및 응용 연구

3. 대시보드 (1)

분류처리 모델을 통해 분류한 데이터베이스 속 수많은 아이템들을
TagCloud를 통해 시각화하여 상대적으로 많은 관심이 많은 분류를 한 눈에 알아볼 수 있게 했다.
그리고 리스트로 작성하여 분류별 아이템의 개수와 예시 아이템 또한 사용자가 쉽게 인지하도록 했다.

4. 대시보드 (2)

분류별 상세 페이지로 해당 분류의 시장성과 전망을 확인할 수 있다.
그리고 PieChart를 통해 해당 분류가 전체 아이템 중 차지하는 비율을 확인할 수 있고
분류에 속한 아이템(논문)의 작성년도별 개수를 BarChart로 시각화하여 분류의 전망을 유추할 수 있게 했다.
이후 리스트로 아이템들을 나열하여 분류에 해당되는 아이템들을 확인하고 각각의 아이템의

상세페이지로 접근이 가능하도록 하였다.

5. 대시보드 (3)

아이템(논문)별 상세 페이지로 해당 아이템의 촉진/저해 요인 비율을 PieChart로 시각화하였고 아이템의 이용수를 LineChart로 시각화하여 사용자가 아이템의 가치와 전망을 쉽게 파악하도록 했다. 모델이 예측한 촉진요인과 저해요인을 확률과 함께 제공하여 도메인 전문가가 아닌 사람들에게 기술의 가치를 판단할 가이드를 제공하였다.

라. 소프트웨어 저작권 등록

SW등록번호	내용	
2022-023776	명칭	뉴스/블로그 스크래핑 데이터활용 딥러닝 자연어처리 기반 미래유망기술 센싱 및 애널리틱 기술개발
	저작자	광운대학교 산학협력단

마. 예산 집행

예산을 사용하지 않음.

바. 개선 방안

- 딥러닝 모델 구축 시 광범위한 범위의 기술들을 하나의 모델을 통해 평가할 수 없어 6T 기술을 IT 분야로 한정하여, 추후에 6T 기술 중 IT를 제외한 나머지 5가지 기술들에 관한 프로젝트를 진행해야 할 필요가 있다.

- 문장 라벨링 시 도메인 전문성 부족으로 인하여 교차검증을 하였음에도 불구하고 라벨링 결과가 애매한 문장이 존재하였다. 이 딥러닝 모델을 사용하기 위한 라벨링 시 도메인 전문성을 가진 참여자가 라벨링을 해야 한다.
- 프로젝트 결과물인 웹의 UI를 사용자가 직관적으로 볼 수 있도록 재배치해야 할 필요가 있다. 사용자는 각 기술에 대한 비전문가일 가능성이 있기 때문이다.

3. 오픈소스SW 활용 및 기여

가.오픈소스SW 활용

1) 활용한 오픈소스SW 소개

Selenium
DBpia 사이트 크롤링 및 논문 pdf 다운로드 자동화
https://selenium-python.readthedocs.io/#

tika
다운받은 논문 pdf에서 텍스트 추출
https://github.com/chrismattmann/tika-python

KeyBERT
논문 제목에서 주요 key추출, 논문 중분류 구분에 사용
https://github.com/MaartenGr/KeyBERT

KoBERT
한글 위키 기반 사전학습 BERT모델, SK개발
논문 텍스트 데이터 모델링 및 파인튜닝
직접 촉진/저해요인을 라벨링 한 데이터로 파인튜닝
라벨이 없는 문장들에 대해서 촉진/저해 예측
https://github.com/SKTBrain/KoBERT

Django
웹 제작에 사용
https://www.djangoproject.com/

2) 활용 내용

나. 오픈소스SW 기여

Github에 KoBERT FineTuning 코드 및 web, data 공개
https://github.com/kimtaeyong98/Technology-Sensing-Evaluation

4. 과제의 향후 계획

가. 활용 방안

인공지능(Artificial Intelligence, AI) 및 기계학습(Machine Learning, ML) 및 자연어처리(Natural Language processing) 관련 기술로 전문가들의 의사결정을 돕기 위해 유망기술을 발굴하고, 유망기술의 가치를 한눈에 파악할 수 있는 시스템을 만든다면, 유망기술의 발굴과 가치판단에 대한 객관성과 시간을 확보할 수 있다.

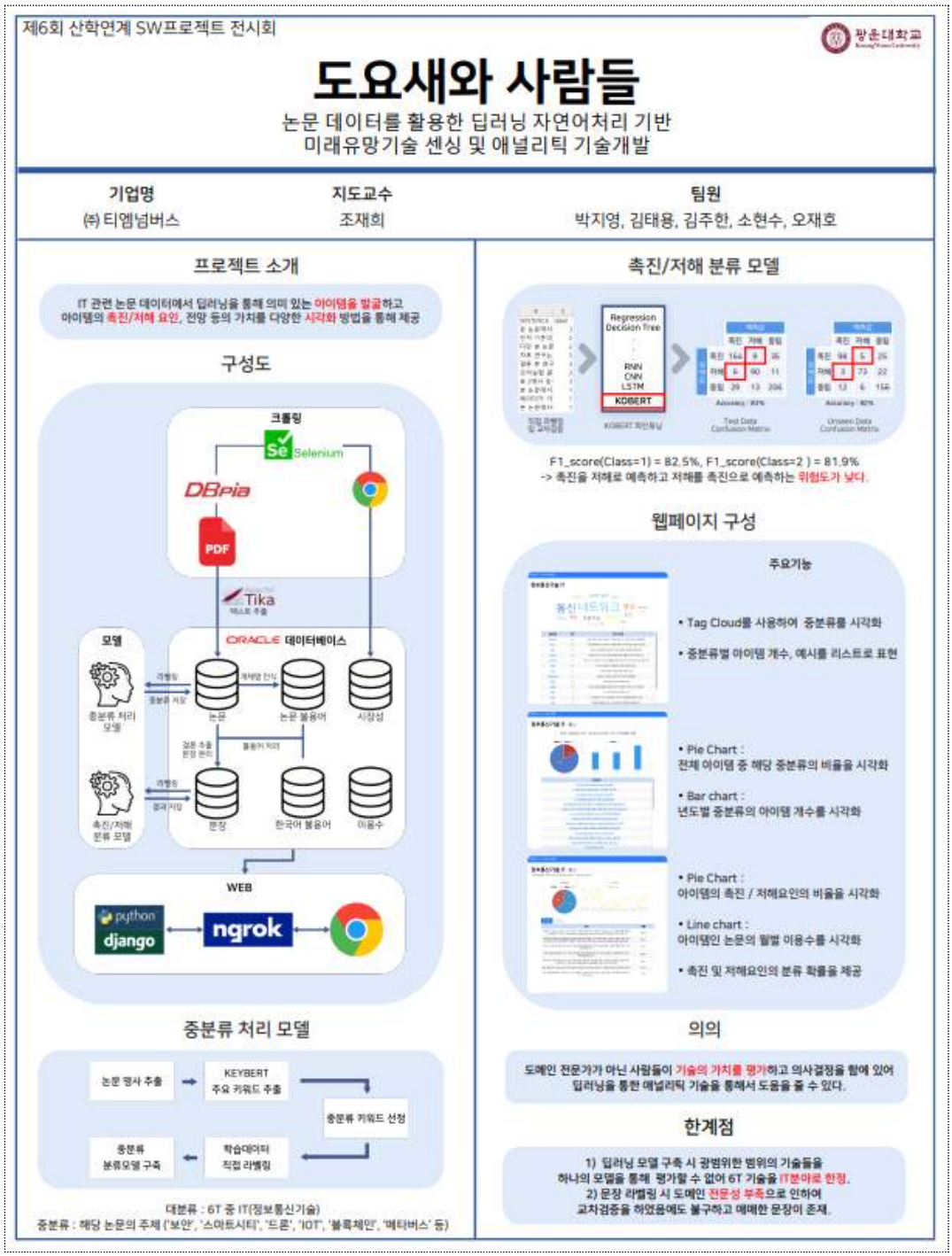
나. 기대 효과

프로젝트의 기대효과는 해당 아이템과 연계된 R&D투자 정보, 논문·특허정보, 기술·시장 동향이나 펀딩 정보, 경제지표, 관련 기업 경쟁정보 등의 다양한 데이터까지 확장될 경우 유망 아이템 분석 시 이에 대한 통찰력을 보다 쉽게 얻을 수 있고, 구조화된 정보를 실시간으로 연계하여 한번에(One-Stop) 탐색할 수 있을 것으로 생각한다.

5. 참고문헌

- [기술가치평가] 2021 기술평가 실무가이드 – 한국산업기술진흥원
  http://www.valuation.or.kr/research_view.do?content_no=675
- Bigkinds
  https://www.bigkinds.or.kr/
- nearbydelta / KoalaNLP
  https://github.com/nearbydelta/koalanlp
- 미래 유망아이템 발굴을 위한 분석플랫폼 연구 - LOD 활용을 중심으로
  https://scienceon.kisti.re.kr/srch/selectPORSrchArticle.do?cn=NPAP12584673&dbt=NPAP
- SKTBrain / KoBERT
  https://github.com/SKTBrain/KoBERT
- 모델 저장하기 & 불러오기
  https://tutorials.pytorch.kr/beginner/saving_loading_models.html

6. 별첨


제6회 산학연계 SW프로젝트 전시회
광운대학교
도요새와 사람들
논문 데이터를 활용한 딥러닝 자연어처리 기반
미래유망기술 센싱 및 애널리틱 기술개발
기업명
㈜ 티엠넘버스
지도교수
조재희
팀원
박지영, 김태용, 김주한, 소현수, 오재호
프로젝트 소개
IT 관련 논문 데이터에서 딥러닝을 통해 의미 있는 아이템을 발굴하고
아이템의 촉진/저해 요인, 전망 등의 가치를 다양한 시각화 방법을 통해 제공
구성도
크롤링
Selenium
DBpia
PDF
Tika
ORACLE 데이터베이스
모델
논문
논문 불용어
사장성
문장
한국어 불용어
이음수
중분류 처리 모델
촉진/저해 분류 모델
WEB
python
django
ngrok
중분류 처리 모델
논문 명사 추출
KEYBERT
주요 키워드 추출
중분류 키워드 선정
학습데이터
직접 라벨링
중분류
분류모델 구축
대분류 : 6T 중 IT(정보통신기술)
중분류 : 해당 논문의 주제 ('보안', '스마트시티', '드론', 'IOT', '블록체인', '메타버스' 등)
촉진/저해 분류 모델
Regression
Decision Tree
RNN
CNN
LSTM
KOBERT
F1_score(Class=1) = 82.5%, F1_score(Class=2) = 81.9%
-> 촉진을 저해로 예측하고 저해를 촉진으로 예측하는 위험도가 낮다.
웹페이지 구성
주요기능
• Tag Cloud를 사용하여 중분류를 시각화
• 중분류별 아이템 개수, 예시를 리스트로 표현
• Pie Chart :
전체 아이템 중 해당 중분류의 비율을 시각화
• Bar chart :
년도별 중분류의 아이템 개수를 시각화
• Pie Chart :
아이템의 촉진 / 저해요인의 비율을 시각화
• Line chart :
아이템인 논문의 월별 이용수를 시각화
• 촉진 및 저해요인의 분류 확률을 제공
의의
도메인 전문가가 아닌 사람들이 기술의 가치를 평가하고 의사결정을 함에 있어
딥러닝을 통한 애널리틱 기술을 통해서 도움을 줄 수 있다.
한계점
1) 딥러닝 모델 구축 시 광범위한 범위의 기술들을
하나의 모델을 통해 평가할 수 없어 6T 기술을 IT분야로 한정.
2) 문장 라벨링 시 도메인 전문성 부족으로 인하여
교차검증을 하였음에도 불구하고 매끄한 문장이 존재.