KDT AI-빅데이터 양성과정

대구 뉴스 플랫폼

(대구 뉴스를 기반으로 한 뉴스플랫폼)

TEAM 대구타임즈

허유나 최성진 이승수 박윤미

Contents

Step 01 프로젝트 개요

1. 프로젝트 주제 선정배경
2. 프로젝트 구현내용
3. 개발환경
4. 프로젝트 구조
5. 기대효과

Step 02 팀구성 및 역할

1. 구성원별로 프로젝트를 진행하면서 주도적으로 참여한 부분을 중심으로 작성

Step 03 수행절차 및 방법

1. 프로젝트 사전 기획
2. 프로젝트 수행
3. 프로젝트 완료 과정

Step 04 수행결과

1. 탐색적 분석 및 전처리
2. 모델 개요
3. 모델 선정 및 분석
4. 모델 평가 및 개선
5. 시연 동영상

Step 05 자체 평가 의견

1. 한계점 및 개선사항

KDT 빅데이터-AI 양성과정

프로젝트 개요

1. 프로젝트 주제 선정배경
2. 프로젝트 구현내용
3. 개발환경
4. 프로젝트 구조
5. 기대효과

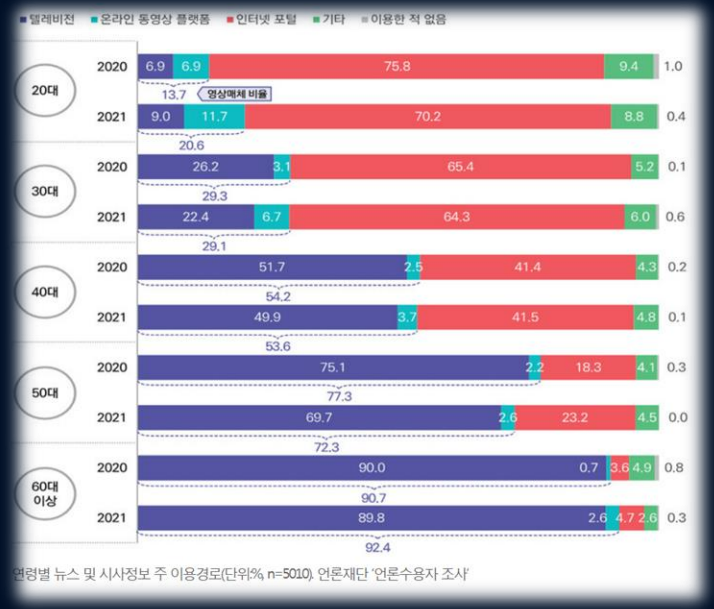
프로젝트 개요

프로젝트 주제 선정 배경

한겨레

한국언론진흥재단 '2021 언론수용자 조사' "뉴스, 포털·유튜브로 본다." 는 비율 더 높아졌다

지난해 포털과 온라인 동영상 플랫폼을 통해 뉴스를 접하는 비율이 더욱 높아진 것으로 조사됐다. 유튜브 등 동영상 플랫폼에 대해선 이제 "뉴스 매체로서 공고한 위치를 점하기 시작했다"는 평가가 나왔다.



최근 텍스트를 통해 뉴스기사를 접하기보다 유튜브와 같은 **온라인 동영상 플랫폼을 통해 뉴스를 접하는 비율이 높아**지고 있음

코로나 19 이후 **온라인 동영상 플랫폼 이용률**이 지난해 **24.4%**에서 **26.7%**로 증가함

관심있는 뉴스를 쉽게 접할 수 있는 **동영상 플랫폼의 장점을 반영한 뉴스 플랫폼 서비스 기획**

프로젝트 개요

프로젝트 구현 내용

요약분석

감성분석

주제분류 모델

신뢰도 분석

뉴스 기사 간결 요약 :
기사의 핵심 내용을 파악할 수 있는 기능

기사 감정 톤 분석 : 긍부정 식별 기능

주요 주제 식별 :
주요 토픽 모델링 구분

신뢰도 측정: 논문의 가중치 점수

기사검색 (유사도 분석으로 유사 기사 추천)

기사 감정분석 (키워드 긍부정 판단)

토픽 모델링 (연관키워드)

기사 신뢰성(요인별 가치점수 부여)

프로젝트 개요

개발환경

개발환경

o/s : window

언어

Python

라이브러리

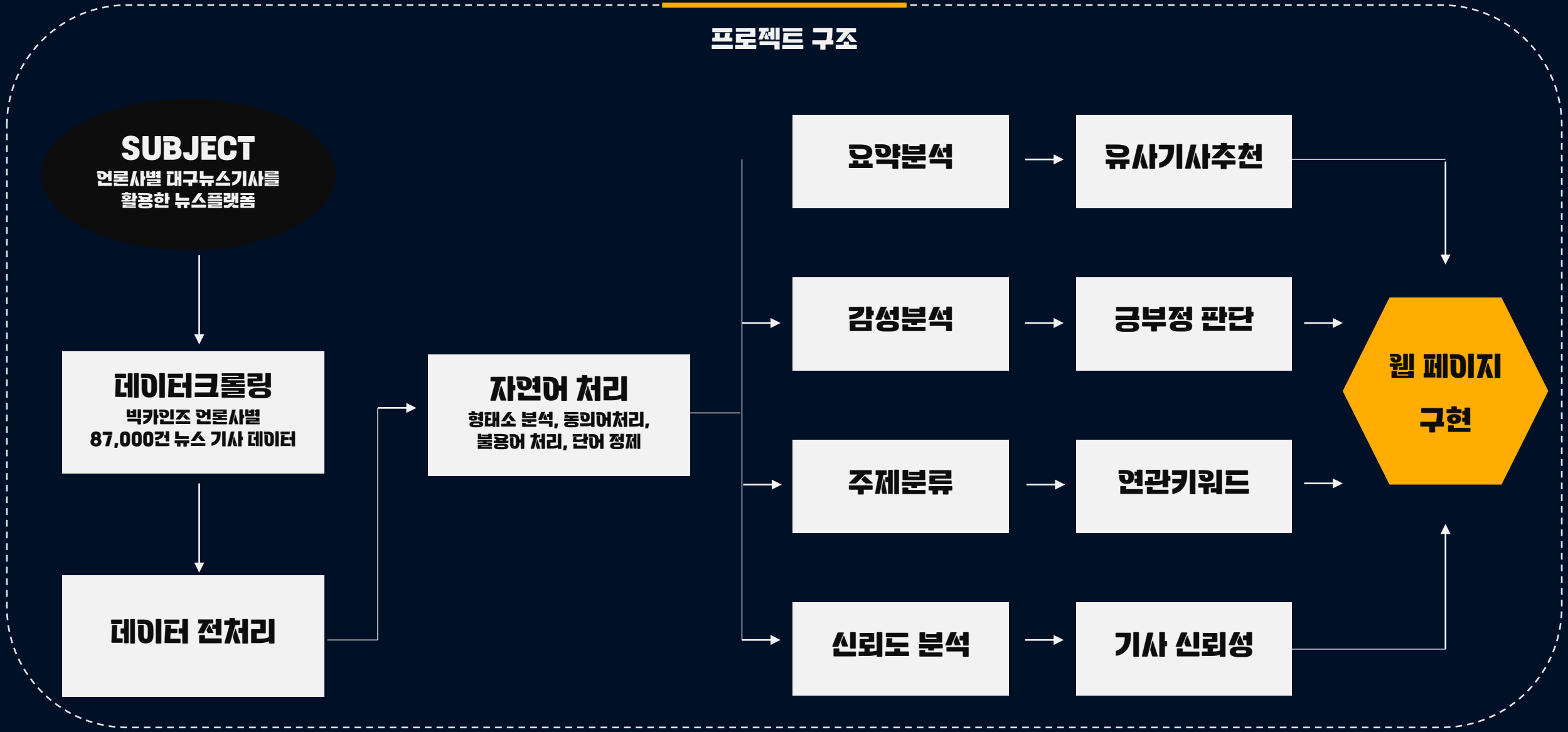
데이터 전처리 : Pandas, Numpy, os, glob

데이터 크롤링 : BeautifulSoup, requests, time, random

텍스트마이닝 : KoBERT, KoBART, CNN, LSTM, LDA, Scikit-Learn

프로젝트 개요

프로젝트 구조



프로젝트 개요

기대효과

오락적 요소

Daegu Times
뉴스플랫폼

실용적 요소

주요 토픽 모델링 구분

관심 기사 혹은 유사기사의 신뢰도 파악

기간설정 후 핫 키워드를 통해 핵심 주제를 파악함에 있어 편리성 제공

감성분석을 통해 기사의 긍부정 판단 정보 제공

키워드 검색시 유사 기사 추천을 통한 데이터 정보의 다양성 제공

긴 본문내용을 핵심 문장으로 요약

KDT 빅데이터-AI 양성과정

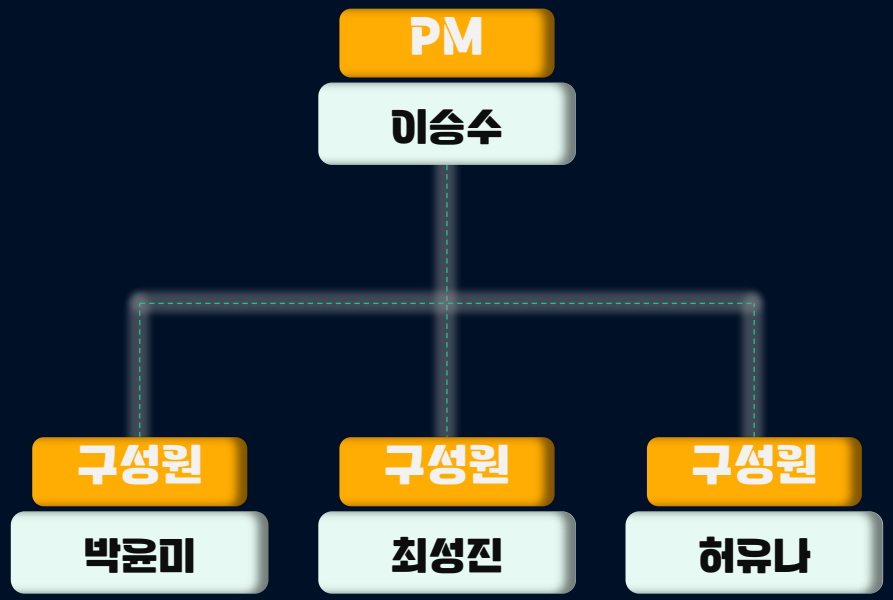
팀구성 및 역할

1. 구성원별로 프로젝트를 진행하면서 주도적으로 참여한 부분을 중심으로 작성

팀구성 및 역할

담당자 및 업무분장

프로젝트 조직도



인력별 역할

구분	담당자	역할
프로젝트 관리	이승수	일정관리, 문서작성, 개발보조
데이터수집	전원	뉴스기사 데이터 크롤링
분석/설계	전원	데이터 전처리 EDA
개발 (AI Model)	전원	모델링(요약모델, 감성분석, 토픽모델링, 신뢰도분석)
개발 (Client-Side)	전원	클라이언트사이드 웹 구축 및 웹 디자인
개발 (Server-Side)	전원	서버사이드 DB구축 및 모델 적용
테스트	전원	모델 테스트, 디버깅, 유지보수

KDT 빅데이터-AI 양성과정

수행절차 및 방법

1. 프로젝트 사전 기획
2. 프로젝트 수행
3. 프로젝트 완료 과정

수행절차 및 방법

프로젝트 사전 기획

	분석 기획	데이터 준비	데이터 분석	시스템 구현	평가 및 전개
주요 Task	1. 프로젝트 정의 2. 계획 수립 3. Kick-off 회의 4. 프로젝트 세부 일정 계획수립	1. 필요 데이터 정의 2. 데이터 수집 3. 데이터 전처리 4. 자연어 처리	1. 탐색적 분석 2. 모델링	1. 플랫폼 업로드	1. 모델 발전 계획 수립 2. 프로젝트 평가/보고
주요 산출물	1. 요구사항 정의서 2. 프로젝트 계획서	1. 뉴스기사 데이터셋			1. 최종보고서

수행절차 및 방법

프로젝트 수행

Step 01
데이터 수집

- 빅카인즈

Step 02
데이터 전처리

- 형태소 분석
- 동의어 처리
- 불용어 처리
- 단어 정제

Step 03
모델 개발 및 학습

- KoBART
- CNN
- LSTM
- LDA

Step 04
서비스 개발 및 배포

- Flask
- HTML
- CSS

수행절차 및 방법

프로젝트 완료 과정

01

~ 11/10

기획

빅데이터 수집

02

~ 11/17

데이터 준비

데이터전처리

자연어처리

03

~ 11/24

모델 구성 및 학습

텍스트 마이닝

모델링

04

~ 12/1

플랫폼 업로드

웹페이지 기능 구현

서버 구축

05

~ 12/8

평가 및 전개

한계점 및 개선사항

06

~ 12/15

발표

KDT 빅데이터-AI 양성과정

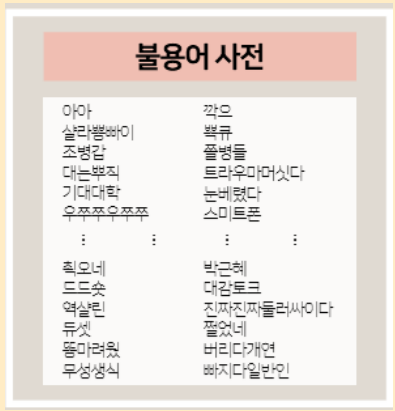
수행결과

1. 탐색적 분석 및 전처리
2. 모델 개요
3. 모델 선정 및 분석
4. 모델 평가
5. 시연 동영상

수행결과

탐색적 분석 및 전처리

자연어 처리



특수문자, 동의어, 불용어 처리

[~, %, .] 제외하고 모든 특수문자 제거

언론사명 삭제

Before

서울신문은 최근 '학폭위 10년, 지금 우리 학교는' 기획
기사를 통해 현행 학폭위 제도의 문제점과 실태를 고스란히
보도했다. 보도 이후 조희연 서울시교육감은 서울신문이
제기한 학폭위의 문제에 대해 적극 공감한다며, 함께 대안을
찾아보자는 의견을 보내왔다. 그는 초등학교 저학년생을 학폭
제도에서 제외하고 경미한 사안은 학교생활기록부 기재를 아예
하지 않는 등 교육적 회복이 가능한 방안을 제시했다.
전국시도교육감협의회장인 조 교육감은 이런 방안들을 전국
교육감 합의를 거쳐 법 개정으로 이끌겠다고 강조했다.
인터뷰는 지난 10일 서울시교육청에서 대면으로 진행했다.
다음은 조 교육감과의 일문일답. -학폭 제도를 도입한 지
10년을 맞았다. 현장에서는 학폭 제도가 과연 우리 교실을
정말 행복하게 만들었는지에 대한 의문이 가득한 상황이다.
"서울신문의 보도를 학폭 업무 담당자들과 인상 깊게 살펴봤다.

자연어 처리 전 기사 내용

After

은 최근 학폭위 10년 지금 우리 학교는 기획 기사를 통해
현행 학폭위 제도의 문제점과 실태를 고스란히 보도했다.
보도 이후 조희연 서울시교육감은 이 제기한 학폭위의 문제에
대해 적극 공감한다며 함께 대안을 찾아보자는 의견을 보내왔다.
그는 초등학교 저학년생을 학폭 제도에서 제외하고 경미한
사안은 학교생활기록부 기재를 아예 하지 않는 등 교육적
회복이 가능한 방안을 제시했다. 전국시도교육감협의회장인
조 교육감은 이런 방안들을 전국 교육감 합의를 거쳐 법 개정으로
이끌겠다고 강조했다. 인터뷰는 지난 10일 서울시교육청에서
대면으로 진행했다. 다음은 조 교육감과의 일문일답. 학폭 제도를
도입한 지 10년을 맞았다. 현장에서는 학폭 제도가 과연 우리
교실을 정말 행복하게 만들었는지에 대한 의문이 가득한 상황이다.
의 보도를 학폭 업무 담당자들과 인상 깊게 살펴봤다.

자연어 처리 후 기사내용

모델 성능을 높이기 위한 함수 실행

Model 01

요약분석 모델

수행결과

모델 개요 - 요약모델

원문내용

'삼보모터스 삼보문화재단 왼쪽 이재하 회장 은 대구 문화예술발전을 위해 연 3천만원씩 총 3억원을 대구문화예술진흥원에 기부약정한다. 대구문화예술진흥원 제공
삼보모터스 삼보문화재단 회장 이재하 은 대구 문화예술발전을 위해 연 3천만원씩 총 3억원을 대구문화예술진흥원 이하 문예진흥원 에 기부 약정한다.삼보모터스는 지역을 대표하는 기업 중 하나로 2015년 삼보문화재단을 설립해 지역의 예술인을 지원 육성하고 전통문화 계승과 발전을 위해 사회공헌과 나눔을 실천하는 기업이다. 청년 시절 미술 교사였던 이재하 회장은 6일 출범한 대구메세나협의회 의 회장으로 추대돼 지역 메세나 활성화와 문화예술 발전을 위해 앞장서고 있다.삼보모터스 삼보문화재단은 기부 약정과 더불어 삼보미술상 제정 의사를 밝혔다. 대구문예진흥원은 미술 분야에서 창작활동에 전념하고 예술적 성과를 인정받은 훌륭한 지역의 원로 작가 1명과 예술적 잠재력과 발전 가능성이 큰 청년 작가 2명을 선정해 삼보미술상을 시상할 계획이다. 삼보미술상에 선정되면 상금과 함께 이듬해에 대구문화예술회관에서 기념전시를 개최할 수 있다.'

요약내용

'삼보 모터스는 지역을 대표하는 기업 중 하나로 2015년 삼보문화재단을 설립해 지역의 예술인을 지원 육성하고 전통문화 계승과 발전을 위해 사회 공헌과 나눔을 실천하는 기업이다. 대구 문예 진흥원은 미술 분야에서 창작활동에 전념하고 예술적 성과를 인정받은 훌륭한 지역의 원로 작가 1명과 예술적 잠재력과 발전 가능성이 큰 청년 작가 2명을 선정해 삼보 미술상을 시상할 계획이다.'

Sumy Library 사용하여,

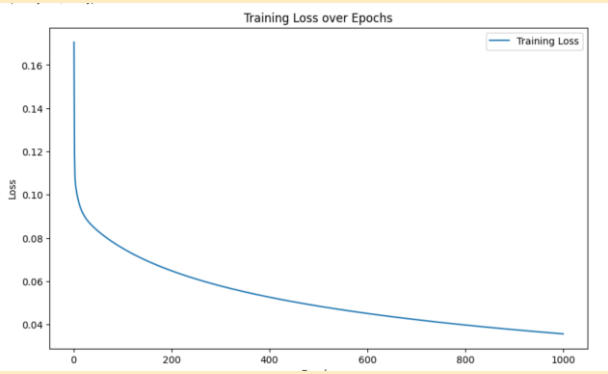
긴 본문 기사 내용을 2~3개의 문장으로 요약

원문내용을 Feature로, 요약내용을 Target으로 지정한 요약분석모델 학습

수행결과

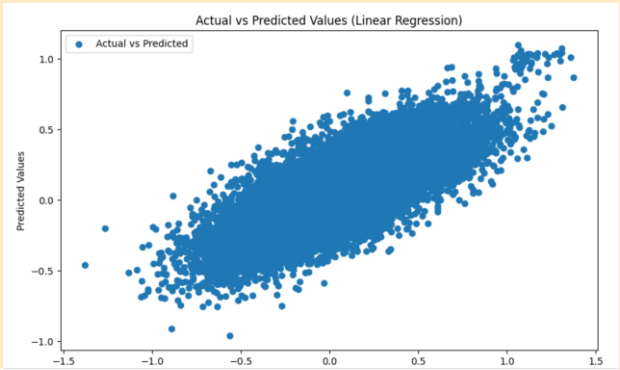
모델 개요 - 요약모델

선형회귀(Loss)



Epoch가 진행될수록 Train loss가 감소함에 따라 모델의 훈련 데이터에 대한 예측 능력이 향상되고 있음을 확인

선형회귀(MSE)



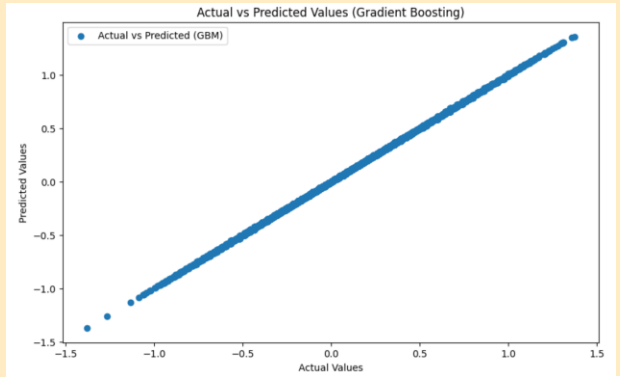
Mean Squared Error: 1.1857052666097712e-11

평균 제곱오차(MSE)가 0에 가까움

모델이 데이터를 잘 적합한 것으로 판단

GBM

랜덤포레스트



Mean Squared Error: 9.373886260502044e-06

GBM과 랜덤포레스트도 평균제곱오차(MSE)가 큰 차이는 나지 않지만 선형회귀 모델보다 높게 나타남

KoBERT 로 임베딩 하였기 때문에 모델 비교 의미 X

But. 모델 예측결과가 임베딩 된 수치 데이터로 표기

수행결과

모델 개요 - 요약모델

KoBERT 임베딩

단어	1	2	3	4
love	0x6C	0x6F	0x76	0x65
live	0x6C	0x69	0x76	0x65
like	0x6C	0x69	0x6B	0x65

KoBERT를 사용한 임베딩이란 ?

사람이 쓰는 자연어를 컴퓨터가 보고 인식할 수 있게

숫자형태인 벡터화(vector) 시킨 결과

KoBART

Hugging Face Transformers 라이브러리를

사용하여 KoBART의 사전 교육된 모델을 사용

사전학습 된 AutoTokenizer(벡터화)

AutoModelForSeq2Seq(예시:번역기)

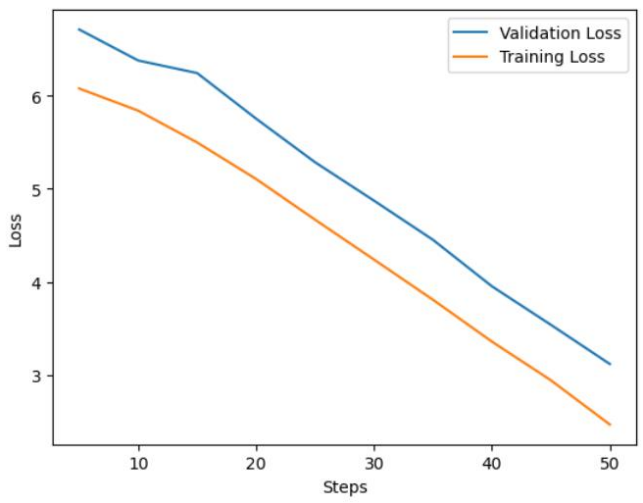
특정 도메인, 즉 위에서 수집한 대구 뉴스

데이터에 맞춰 미세 조정된 학습 모델 생성

수행결과

모델 선정 및 분석 - 요약모델

KoBART 모델 선정



오차율은 Epoch 초기부터 낮은 지표를 보임

Epoch = 50 기준, Loss = 3

예측 결과 예시

```
'(대구=연합뉴스) 박세진 기자 = 대구시는 김선조 신임 행정부시장이 오는 4일 취임한다고 3일 밝혔다. 김 부시장은 취임식을 생략하고 홍준표 대구시장 주재 간부회의에 참석하는 것으로 일정을 시작한다. \n그는 부산 출신으로 서울대 철학과를 졸업했고 37회 행정고시에 합격해 1994년 4월 공직에 입문했다. \n환경부, 울산시 안전행정국장, 울산 중·동구 부구청장, 행정자치부 지역발전과장, 울산시·부산시 기획조정실장, 행정안전부 균형발전지원관 등을 거쳤다. 대구에서는 1995년 환경부 대구지방환경청 소속으로 1년간 근무한 바 있다. \n김 부시장은 "대구가 대한민국 3대 도시의 영광을 되찾도록 행정 업무 경험과 전문성을 살리겠다"고 말했다.\n
```

```
Generated Summary: 3일 대구시는 4일 취임하는 김선조 신임 행정부시장이 취임식을 생략하고 홍준표 대구시장 주재 간부회의에 참석하는 것으로 일정을 시작한다고 밝혔다.
```

Model 02

감성분석 모델

수행결과

모델 개요 - 감성분석모델

감성사전

	word	polarity
0	ㅡㅡ	-1
1	ㅠㅠ	-1
2	ㅠ_ㅠ	-1
3	ㅠ	-1
4	ㅜㅡ	-1
...	...	...
14849	(^-^)	1
14850	(;_;)	-1
14851	(-_-)	-1
14852	(-;	1
14853	-_-^	-1

산출 결과

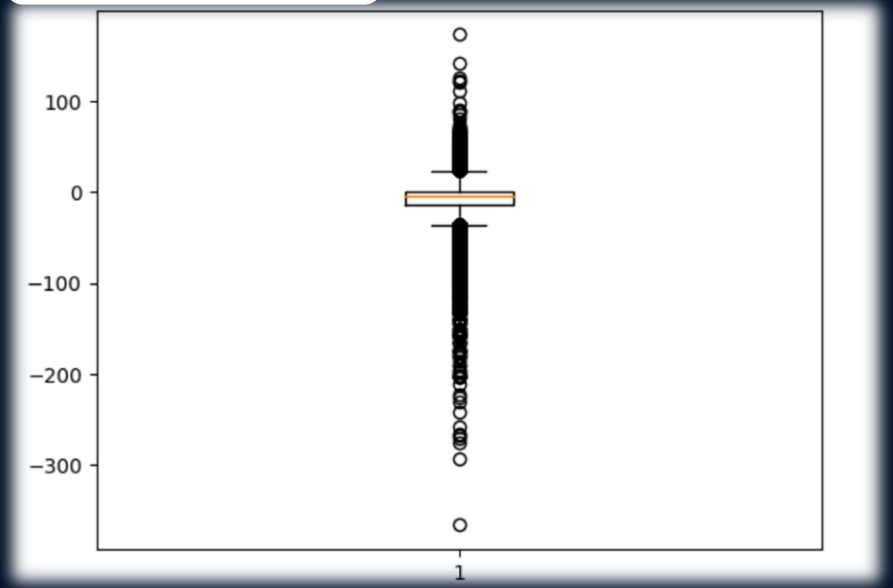
	명사	감성점수
0	['이대성', '허웅', '제압', '성현', '점슛', '분전', '허웅', '생...	6
1	['진박', '감별', '사가', '쥐락펴락', '나경원', '비판', '응수', ...	-3
2	['양금', '국민', '의원', '대구', '북구', '대구', '북구', '주민...	9
3	['김상훈', '국민', '의원', '대구', '서구', '대구', '서구', '주...	14
4	['류성걸', '국민', '의원', '대구', '류성걸', '국민', '의원', '...	12
...	...	...
39940	['우상혁', '부다페스트', '세계', '육상', '선수권', '남자', '높이뛰...	-7
39941	['중소', '벤처기업', '실장', '전보', '중소기업', '정책', '실장',...	15
39942	['유튜버', '김용호', '유튜브', '채널', '강용석', '나이트', '라이브...	-8
39943	['윤희', '경찰청장', '충북', '충주시', '중앙', '경찰', '학교', ...	11
39944	['날씨', '기록', '지난', '서울', '경복궁', '관광객', '양산', '...	-10

군산대학교 감성사전을 이용하여 감성점수 산출

수행결과

모델 개요 - 감성분석모델

감성점수 분포도



긍정/중립/부정 으로 분류하기 위해

Boxplot 을 이용해서 감성점수 분포도 확인

1사분위수(25%) : -14.0 점

3사분위수(75%) : 1.0 점

1사분위수, 3사분위수를 기준으로 나눠서

긍정 : 1.0 점 이상

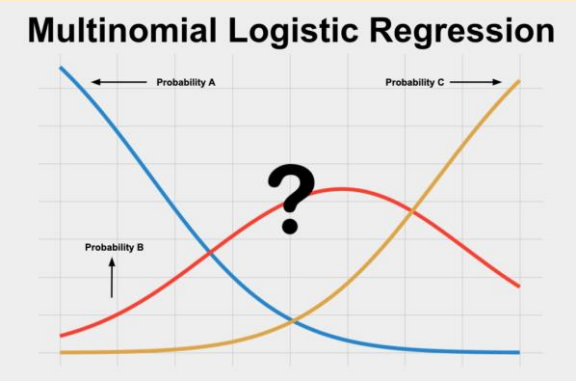
부정 : -14.0 점 이하

중립 : -14.0 과 1.0 점 사이

수행결과

모델 선정 및 분석 - 감성분석모델

모델 선정 배경



앞서 사용한 KoBART 모델을 사용해도 되지만

transformer 모델이라 무겁고 어려움

따라서 쉬운 모델이 적합하다고 생각함

분석 방법

비교적 정형화된 뉴스데이터이기 때문에

군산대학교 감성사전을 이용해 점수를 산출한 후

긍정, 부정, 중립으로 분류해도 괜찮다고 판단

결과적으로 산출한 **감성점수**를 **타겟**, 본문의 **단어**를 **특성**으로 받는 모델 구현

수행결과

모델 선정 및 분석 - 감성분석모델

다항로지스틱 회귀모델

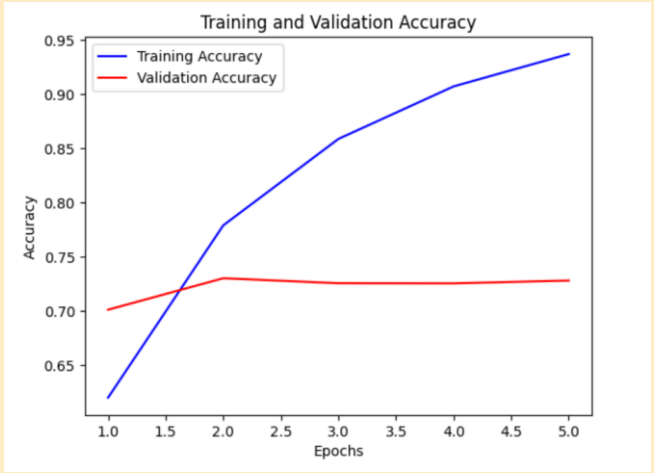
```
lr = LogisticRegression(multi_class='multinomial', solver='newton-cg')
lr.fit(train_X, train_y)
lr.score(test_X, test_y)
```

```
0.8237576667918388
```

	precision	recall	f1-score	support
-1	0.85	0.81	0.83	2002
0	0.81	0.84	0.82	3786
1	0.83	0.81	0.82	2201
accuracy			0.82	7989
macro avg	0.83	0.82	0.82	7989
weighted avg	0.82	0.82	0.82	7989

⇒ 성능 평가 결과가 0.8 정도로 좋은 성능을 보임

LSTM



accuracy : 0.9 / val_accuracy : 0.7

파라미터 수정 후 val_accuracy 는

큰 변화없고 정확도가 떨어짐

현재까지는 LSTM 모델 성능이 과적합을 보임으로 다항 로지스틱 회귀 모델이 적합한 것으로 판단

수행결과

모델 선정 및 분석 - 감성분석모델

새로운 텍스트

'"엄마, 강아지는 어떻게 만들어?"\n\n"강아지들을 철창에 가둬놓고
발정제 주사를 잔뜩 맞혀. 강제로 여러 차례 교배를 시키지.
임신한지 60일 정도가 지나면, 엄마 강아지의 배를 갈라 아기 강아지들을 꺼낸단다.
"\n\n빨래를 개던 엄마가 아들의 질문에 친절히 답해준다.
이를 듣는 아이 표정이 점점 심각해진다.\n\n/사진=위액트 유튜브 영상 '
사지않을개' 캠페인, '펫숍 강아지는 어디서 올까?' 화면 캡쳐\n/
사진=위액트 유튜브 영상 \'사지않을개\' 캠페인, \'펫숍 강아지는 어디서 올까?\'
화면 캡쳐\n"엄마 강아지의 배는 다시 꿰매주면 돼. 이미 여러번 그렇게 했거든.
태어난 새끼들은 죽지 않을만큼 굶겨. 그럼 아주 조그맣게 만들 수 있어.
정말 예쁘지? 강아지가 너무 크면 잘 팔리지 않거든."\n/사진=위액트 유튜브 영상
'사지않을개' 캠페인, '펫숍 강아지는 어디서 올까?'
화면 캡쳐\n/사진=위액트 유튜브 영상 \'사지않을개\' 캠페인, \'펫숍 강아지는
어디서 올까?\' 화면 캡쳐\n엄마는 웃으며 말을 이어간다.\n\n"그래서 새끼들 중에
제일 작은 강아지는 다시 철창에 가두어서 임신을 시키고, 나머지는 가게에 내다 파는
거란다."\n\n이를 다 들은 아이 얼굴이 잔뜩 시무룩해진다.
집안에 있던 작고 하얀 반려견, 몰티즈를 새삼스레 바라본다.

기사내용을 보면 내용이 부정적이라는 것을 확인할 수 있음

예측 결과 예시

```
sentiment = '긍정' if predicted_label == 1 \
    else ('부정' if predicted_label == -1 else '중립')
print(f'예측 감성: {sentiment}')
```

```
예측 감성: 부정
```

기사 원문을 토대로 예측한 감성 분석 결과를 보면 "긍정", "중립", "부정" 중 "부정" 으로 정확히 예측한 것을 확인할 수 있음

Model 03

주제분류모델

수행결과

모델 개요 - 주제분류모델

토픽모델링

```
nouns_df = pd.read_csv('../datas/nouns_df.csv')
nouns_df['본문']
```

```
0        팀 이대성  122 117로 팀 허웅 제압 전성현 3점슛 9개 분전허웅  생애 첫 ...
1        제2의 진박감별사가 당 쥐락펴락  나경원 비판에 응수   국민의힘 장제원 의원  나...
2        양금희 국민의힘 의원 대구 북구갑 .                        대구...
3        김상훈 국민의힘 의원 대구 서구 .                        대구 ...
4        류성걸 국민의힘 의원 대구 동구갑 .                        류성...
                              ...
39940    우상혁이 20일 2023 부다페스트 세계육상선수권 남자 높이뛰기 예선에서 바를 넘고...
39941    중소벤처기업부  실장급 전보  중소기업정책실장 이대희  소상공인정책실장 원영준  국...
39942    유튜버 김용호씨. 유튜브 채널       강용석 나이트 라이브  캡처연예  출신 유튜...
39943    윤희근 경찰청장이 18일 충북 충주시 중앙경찰학교에서 열린 신임경찰 제312기 졸업...
```

뉴스 데이터의 본문 열을 이용하여 토픽 모델링 실행

벡터화

CounterVectorizer과 TF-IDF로

벡터화 후 LDA 모델을 통해 토픽 모델링 실행

```
from sklearn.decomposition import LatentDirichletAllocation

lda = LatentDirichletAllocation(n_components=7, random_state=42)
lda.fit(dtm)
```

```
LatentDirichletAllocation
LatentDirichletAllocation(n_components=7, random_state=42)
```

수행결과

모델 개요 - 주제분류모델

토픽 분류

```
topicList = []
for n, i in enumerate(lda.components_): # 인덱스, 값
    idx = np.argsort(i)[::-1][:7] # 낮은 순
    topic = cv.get_feature_names_out()[idx] # 0번 데이터는 3번째 가중치가 낮았다는 뜻
    print(f'Topic {n+1} : {topic}') # 1번 토픽 많은 순서, 2번 많은 순서
    topicList.append(topic)
```

```
Topic 1 : ['아파트' '주택' '가구' '은행' '가격' '하락' '증가']
Topic 2 : ['의원' '국민' '대통령' '대표' '민주당' '후보' '시장']
Topic 3 : ['사업' '산업' '기업' '지원' '교육' '공항' '추진']
Topic 4 : ['경찰' '혐의' '사건' '마약' '수사' '사람' '범죄']
Topic 5 : ['병원' '안전' '의료' '발생' '피해' '지원' '환자']
Topic 6 : ['기온' '예상' '도로' '날씨' '오전' '남부' '최고']
Topic 7 : ['문화' '행사' '공연' '축제' '한국' '예술' '작품']
```

- 1 : 부동산
- 2 : 정치 이슈
- 3 : 정책
- 4 : 사건 사고 범죄
- 5 : 건강 및 안전
- 6 : 날씨 및 기상
- 7 : 문화와 예술
- 8 : 기타

부동산, 정치이슈, 정책, 사건사고 범죄,
건강 및 안전, 날씨 및 기상, 문화와 예술,
기타 **7개의 토픽으로 분류**

분류된 토픽을 보고 **stop_words** 에
지우고 싶은 단어를 추가하여 반복 실행

Stop_words

```
# 지우고 싶은 단어가 있을 경우
stop_words = ['대구', '경북', '전국', '대구시', '서울', '지역', '오늘',
              '무단', '배포', '금지', '저작권', '내일', '광주',
              '강원','우리','위해','지금','지난','지점','대한','지난해',
              '제주', '오후', '아침', '경기', '부산', '코로나', '제공', '대해']
```

수행결과

모델 개요 - 주제분류모델

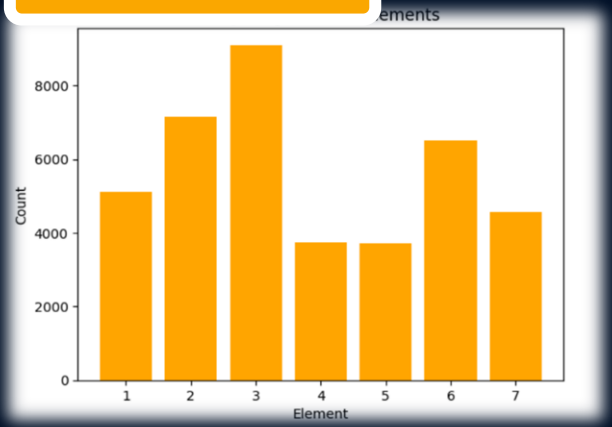
주제 분류 함수

```
    total_sum = []
    for j in range(len(topicList)):
        total_sum.append(dtm_df.loc[:, dtm_df.columns.isin(topicList[j])].sum(axis=1).values[0])

    max_value = max(total_sum)
    max_index = total_sum.index(max_value) + 1
else:
    max_value = 0
    max_index = 8

max_value_list.append(max_value)
max_index_list.append(max_index)
```

```
  0%|          | 13/39945 [00:00<05:50, 113.91it/s]
100%|██████████| 39945/39945 [02:57<00:00, 224.87it/s]
```

각 본문 열에 대해 **주제를 분류하는 함수** 생성

빈도수 그래프



각 주제 별 **빈도수**에 대한 그래프

CounterVector

0.0s

	본문	Topic
0	팀 이대성 122 117로 팀 허웅 제압 전성현 3점슛 9개 분전허웅 생애 첫 ...	7
1	제2의 진박감별사가 당 쥐락펴락 나경원 비판에 응수 국민의힘 장제원 의원 나...	2
2	양금희 국민의힘 의원 대구 북구갑 . 대구...	2
3	김상훈 국민의힘 의원 대구 서구 . 대구 ...	2
4	류성걸 국민의힘 의원 대구 동구갑 . 류성...	2
...	...	...
39940	우상혁이 20일 2023 부다페스트 세계육상선수권 남자 높이뛰기 예선에서 바를 넘고...	7
39941	중소벤처기업부 실장급 전보 중소기업정책실장 이대희 소상공인정책실장 원영준 국...	1
39942	유튜버 김용호씨. 유튜브 채널 강용석 나이트 라이브 캡처연예 출신 유튜...	4
39943	윤희근 경찰청장이 18일 충북 충주시 중앙경찰학교에서 열린 신임경찰 제312기 졸업...	4
39944	무더운 날씨를 기록한 지난 14일 서울 경복궁을 찾은 관광객들이 양산으로 햇볕을 피...	6

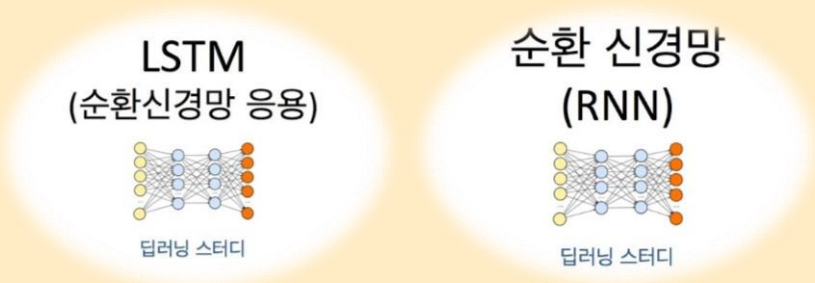
토픽 열을 Target 데이터로 하여 딥러닝 실행

TF-IDF는 토픽 분류 성능이 좋지 않아 **CounterVectorizer 사용**

수행결과

모델 선정 이유 - 주제분류모델

Sequency



텍스트의 복잡한 의미 구조를 분석하기 용이

특징과 패턴을 효과적으로 분석

문맥의 의존성이 있는 텍스트 분석에 자주 사용

Non_Sequency

1차원 시퀀스의 로컬 패턴과 구조를 캡쳐해서 분석

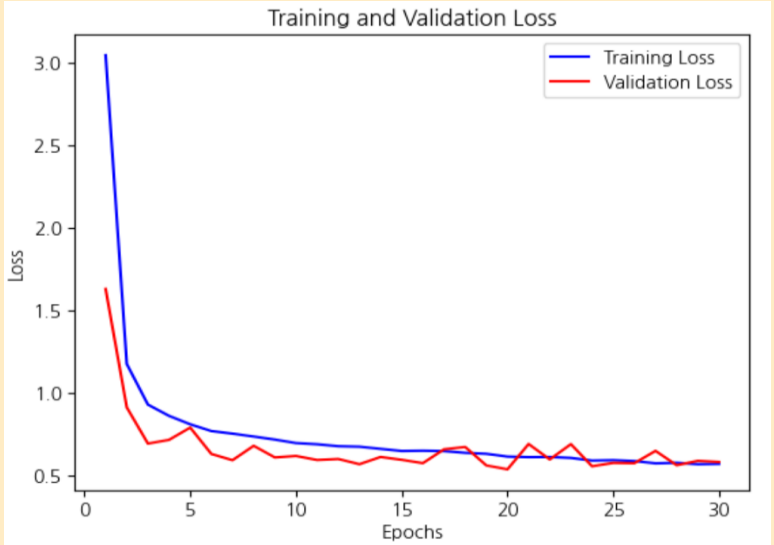
문장안에 단어, 성분의 특징, 문장의 지역 정보 위주

텍스트 분석에 자주 사용

수행결과

모델 선정 및 분석 - 주제분류모델

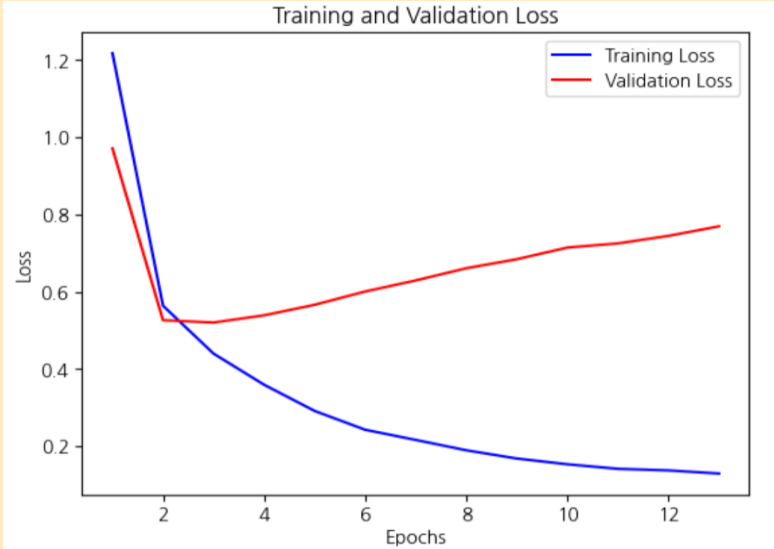
LSTM



accuracy: 0.8269

val_accuracy: 0.8346

CNN



accuracy: 0.9754

val_accuracy: 0.8400

LSTM은 과적합 없이 모델이 잘 생성된 것을 볼 수 있는 반면 CNN은 과적합이 존재

따라서 CNN 모델이 더 가볍지만 오차율과 과적합으로 인해 LSTM 모델을 선택

수행결과

모델 선정 및 분석 - 주제분류모델

LSTM 모델 선정

```
text = '''호흡기 약하신 분들은 불필요한 외출을 자제하시는 게 좋겠습니다.

먼지는 찬 바람이 불며 밤부터 점차 해소되겠는데요.

날이 추워지겠습니다.

현재는 서울의 기온 13.1도, 대구 20.4도, 부산이 18.9도 보이고 있고요.

그 밖의 지역도 안동 17도, 창원 18도, 포항이 20도 안팎으로 어제만큼 공

하지만 밤사이 북서쪽에서 찬 공기가 남하하며 기온이 큰 폭으로 떨어지겠

내일 서울 아침 기온 영하 4도, 주말인 모레는 영하 6도까지 곤두박질하겠

찬 바람까지 불어 체감하는 추위가 심하겠습니다.
'''
```

- 1 : 부동산
- 2 : 정치 이슈
- 3 : 정책
- 4 : 사건 사고 범죄
- 5 : 건강 및 안전
- 6 : 날씨 및 기상
- 7 : 문화와 예술
- 8 : 기타

텍스트를 확인하면 **날씨 및 기상 토픽**임을 확인

예측 결과 예시

```
predicted_category = model.predict([text_to_predict])

# Print the predicted category
print(f'Predicted Category: {predicted_category[0]}')
```

```
Predicted Category: 6
```

6:날씨 및 기상 카테고리로 본문의 토픽을 **정확히 예측**한 것을 볼 수 있음

Model 04

신뢰도분석모델

수행결과

모델 개요 - 신뢰도분석모델

논문 : 뉴스 기사 신뢰도 측정방안
(뉴스 트러스트 사례를 중심으로)

계량 요인	점수 부여 방식
기자 명	• 기자 명 DB 필드를 통해 추출 • DB 필드에는 없으나 본문에 기자 명이 있을 경우 언론사별 패턴 분석을 통해 기계적으로 추출 • 추출된 기자 명들 중 인터넷뉴스팀 등 기자가 작성하지 않은 특수 케이스 처리 • 실명 기자 명과 이메일 있을 경우 1점, 실명 기자 명만 있으면 0.8점, 기자 명이 아예 없으면 -1점, 비실명 기자 명만 있으면 0점
기사의 길이	• 길이에 따라 0~1점 사이 가점 부여(길이가 긴 것에 대한 보상이 있어야 함) • 평균 이상인 기사에 대해 단계별 가점(표준편차 따른 단계 구분), 분류(category) 및 신문·방송 등 유형에 따른 적용 필요 • if (content_length <mean) then 0, else (content_length <mean + 0.5SD) then 0.165, else (content_length <mean + SD) then 0.33, else (content_length <mean + 1.5SD) then 0.495, else (content_length <mean + 2SD) then 0.66, else (content_length <mean + 2.5SD) then 0.835, else (content_length >mean + 2.5SD) then 1
인용문의 수	• 0~15개까지 균등하게 점증 가점 • 15개 이상은 모두 1점
제목의 길이	• 명확한 기준을 세우기 애매하기 때문에 지나치게 긴 제목에 대해서만 감점 • 제목의 길이가 45자를 넘으면 감점
제목의 물음표, 느낌표 수	• 따옴표는 문제 삼지 않으며, 물음표, 느낌표를 사용했을 경우 사용 여부에 따라 감점 • 물음표 느낌표 1개인 경우 -0.5, 2개인 경우 -1점
수치 인용 수	• 분류 및 유형 별로 숫자가 많으면 가점, 한글을 빼고 무조건 숫자만 추출 • 평균 이하는 0점, 0~0.5SD=0.33, 0.5SD~1SD=0.66, 1SD 이상 = 1 • 기사 분류 및 매체 유형에 따라 평균 및 표준편차 적용
이미지의 수	• 3개를 기점으로 전후 차등하여 점수 부여 • 0개 = 0점, 1개 = 0.33, 2개 = 0.66, 3개 = 1점, 4개 = 0.66점, 5개 = 0.33점 6개 이상 = 0점
평균 문장의 길이	• 문장의 길이가 평균적으로 지나치게 길 경우 감점 • 평균 + 1SD 이상인 경우 -1점, 나머지는 0점 • 분류 및 매체 유형에 따라 평균 및 표준편차 적용
제목에 사용된 부사 수	• 부사 수가 많을 경우 감점(제목에 사용된 경우로 개수별로 감점) • 0에서 1까지는 0점, 2개는 -0.5점, 3개 이상은 -1점
문장당 평균 부사 수	• 형용사, 접속사는 제외하고 부사가 지나치게 많은 것에만 감점(분류 및 매체 유형에 따라 적용) • 평균값 + 2SD보다 많은 경우 -1점 적용
기사 본문 중 인용문의 비중	• 전체 기사 내에서 인용문이 차지하는 비중이 지나치게 높을 경우 감점 • 전체 기사에서 인용문의 비중이 0.5~0.8까지는 -0.5, 0.8이상은 -1점

가치점수

- 독이성 : 기자 명 + 기사 길이 + 인용문 수 + 제목 길이 + 제목 물음표·느낌표 + 수치 인용 수 + 이미지 수 + 평균 문장 길이 + 제목 부사 수 + 문장 평균 부사
- 투명성 : 기자 명 + 기사 길이 + 인용문 수 + 제목 물음표·느낌표+수치 인용 수 + 이미지 수 + 인용문 비중
- 사실성 : 기자 명 + 기사 길이 + 인용문 수 + 제목 물음표·느낌표 + 수치 인용 수 + 이미지 수 + 제목 부사 수 + 문장 평균 부사 수 + 인용문 비중
- 유용성 : 기자 명 + 기사 길이 + 인용문 수 + 수치 인용 수 + 이미지 수
- 균형성 : 기자 명 + 기사 길이 + 인용문 수 + 제목 물음표·느낌표 + 제목 부사 수 + 문장 평균 부사 수 + 인용문 비중
- 다양성 : 기자 명 + 기사 길이 + 인용문 수 + 제목 길이 + 수치 인용수 + 이미지 수 + 평균 문장 길이 + 인용문 비중
- 독창성 : 기자 명 + 기사 길이 + 인용문 수 + 제목 물음표·느낌표 + 수치 인용수 + 이미지 수
- 중요성 : 기자 명 + 기사 길이 + 인용문 수 + 수치 인용 수 + 이미지 수
- 심층성 : 기자 명 + 기사 길이 + 인용문 수 + 수치 인용 수 + 이미지 수 + 인용문 비중
- 선정상 : 기자 명 + 제목 길이 + 제목 물음표·느낌표 + 제목 부사 수 + 문장 평균 부사 수 + 인용문 비중
- 총합

본문, 내용을 통해 점수 환산

점수를 타겟으로 선정

본문, 내용을 특성으로 선정

딥러닝 모델 적합

참고 논문 : https://www.kpf.or.kr/front/research/selfDetail.do?seq=591492 75페이지~

수행결과

모델 선정 및 분석 – 신뢰도분석모델

LSTM 모델정보

```
Model: "model"
__________________________________________________________________________________________________
 Layer (type)                Output Shape              Param #   Connected to
==================================================================================================
 input_1 (InputLayer)        [(None, 100)]             0         []

 input_2 (InputLayer)        [(None, 100)]             0         []

 embedding (Embedding)       (None, 100, 64)           64000     ['input_1[0][0]',
                                                                  'input_2[0][0]']

 lstm (LSTM)                 (None, 100, 64)           33024     ['embedding[0][0]']

 lstm_2 (LSTM)               (None, 100, 64)           33024     ['embedding[1][0]']

 lstm_1 (LSTM)               (None, 64)                33024     ['lstm[0][0]']

 lstm_3 (LSTM)               (None, 64)                33024     ['lstm_2[0][0]']

 concatenate (Concatenate)   (None, 128)               0         ['lstm_1[0][0]',
                                                                  'lstm_3[0][0]']

 dropout (Dropout)           (None, 128)               0         ['concatenate[0][0]']

 dense (Dense)               (None, 64)                8256      ['dropout[0][0]']

 dense_1 (Dense)             (None, 1)                 65        ['dense[0][0]']

==================================================================================================
Total params: 204417 (798.50 KB)
Trainable params: 204417 (798.50 KB)
Non-trainable params: 0 (0.00 Byte)
```

본문, 내용 2가지를 Input으로 받고

점수를 산출하는 선형 딥러닝 모델 생성

모델학습 결과

```
Epoch 1/150
243/243 [==============================] - 66s 255ms/step - loss: 7717019.5000 - mse: 7717019.5000 - val_loss: 2
Epoch 2/150
C:\Users\LG\anaconda3\envs\daegu\lib\site-packages\keras\src\engine\training.py:3000: UserWarning: You are savin
g instead the native Keras format, e.g. `model.save('my_model.keras')`.
  saving_api.save_model(
243/243 [==============================] - 74s 306ms/step - loss: nan - mse: nan - val_loss: nan - val_mse: nan
Epoch 3/150
243/243 [==============================] - 90s 369ms/step - loss: nan - mse: nan - val_loss: nan - val_mse: nan
Epoch 4/150
243/243 [==============================] - 75s 307ms/step - loss: nan - mse: nan - val_loss: nan - val_mse: nan
Epoch 5/150
243/243 [==============================] - 103s 425ms/step - loss: nan - mse: nan - val_loss: nan - val_mse: nan
Epoch 6/150
240/243 [============================>.] - ETA: 0s - loss: nan - mse: nan
```

첫 학습 이후에는 모델 평가가 불가능하게

손실함수가 커짐에 따라 모델이 적합하지 않음

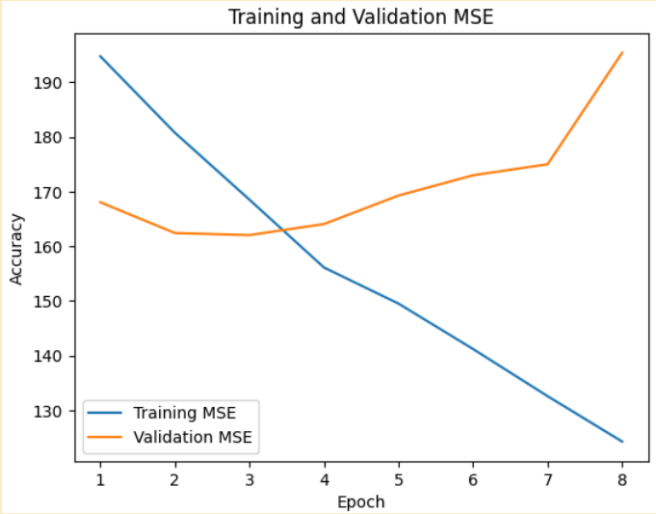
수행결과

모델 선정 및 분석 - 신뢰도분석모델

CNN 모델정보

```
Model: "model"
_________________________________________________________________________________________________
 Layer (type)                Output Shape              Param #   Connected to
=================================================================================================
 input_1 (InputLayer)        [(None, 100)]             0         []

 input_2 (InputLayer)        [(None, 100)]             0         []

 embedding (Embedding)       (None, 100, 64)           64000     ['input_1[0][0]',
                                                                  'input_2[0][0]']

 conv1d (Conv1D)             (None, 97, 64)            16448     ['embedding[0][0]']

 conv1d_1 (Conv1D)           (None, 97, 64)            16448     ['embedding[1][0]']

 layer_normalization (Layer  (None, 97, 64)            128       ['conv1d[0][0]']
 Normalization)

 layer_normalization_1 (Lay  (None, 97, 64)            128       ['conv1d_1[0][0]']
 erNormalization)

 global_max_pooling1d (Glob  (None, 64)                0         ['layer_normalization[0][0]']
 alMaxPooling1D)

 global_max_pooling1d_1 (Gl  (None, 64)                0         ['layer_normalization_1[0][0]'
 obalMaxPooling1D)                                                ]

 concatenate (Concatenate)   (None, 128)               0         ['global_max_pooling1d[0][0]',
                                                                  'global_max_pooling1d_1[0][0]
                                                                  ']

 dropout (Dropout)           (None, 128)               0         ['concatenate[0][0]']

 dense (Dense)               (None, 64)                8256      ['dropout[0][0]']

 dropout_1 (Dropout)         (None, 64)                0         ['dense[0][0]']

 dense_1 (Dense)             (None, 32)                2080      ['dropout_1[0][0]']

 dense_2 (Dense)             (None, 1)                 33        ['dense_1[0][0]']

=================================================================================================
Total params: 107521 (420.00 KB)
Trainable params: 107521 (420.00 KB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________________________________________
```

모델학습 결과




```
Epoch 1/150
243/243 [==============================] - 19s 69ms/step - loss: 302.4314 - mse: 299.9179 - val_loss: 185.5934 - val_mse: 183.1004
Epoch 2/150
C:\Users\LG\anaconda3\envs\daegu\lib\site-packages\keras\src\engine\training.py:3000: UserWarning: You are saving your model as an HDF5 file v
g instead the native Keras format, e.g. `model.save('my_model.keras')`.
  saving_api.save_model(
243/243 [==============================] - 20s 81ms/step - loss: 197.2349 - mse: 194.7615 - val_loss: 170.5132 - val_mse: 168.0573
Epoch 3/150
243/243 [==============================] - 19s 77ms/step - loss: 183.2180 - mse: 180.7799 - val_loss: 164.8550 - val_mse: 162.4315
Epoch 4/150
243/243 [==============================] - 15s 64ms/step - loss: 170.9430 - mse: 168.5366 - val_loss: 164.4411 - val_mse: 162.0497
Epoch 5/150
243/243 [==============================] - 13s 53ms/step - loss: 158.5068 - mse: 156.1279 - val_loss: 166.4391 - val_mse: 164.0714
Epoch 6/150
243/243 [==============================] - 13s 52ms/step - loss: 151.8781 - mse: 149.5224 - val_loss: 171.6351 - val_mse: 169.2853
Epoch 7/150
243/243 [==============================] - 13s 55ms/step - loss: 143.5515 - mse: 141.2132 - val_loss: 175.3264 - val_mse: 172.9973
Epoch 8/150
243/243 [==============================] - 15s 62ms/step - loss: 134.9118 - mse: 132.5888 - val_loss: 177.3195 - val_mse: 175.0060
Epoch 9/150
242/243 [============================>.] - ETA: 0s - loss: 126.5952 - mse: 124.2899Restoring model weights from the end of the best epoch: 4.
243/243 [==============================] - 17s 68ms/step - loss: 126.5891 - mse: 124.2837 - val_loss: 197.7047 - val_mse: 195.4037
Epoch 9: early stopping
```

데이터 값들의 특성을 중심으로 손실함수가 크지 않으므로 CNN모델이 적합한 것으로 판단

다만 모델학습이 반복 될 경우 과적합 되므로 과적합되지 않은 선에서 모델 학습 정지

새로운 데이터에서도 괜찮은 예측 결과를 가져옴

수행결과

모델 선정 및 분석 - 신뢰도분석모델

CNN 모델 선정

```
Model: "model"
_________________________________________________________________________________________________
 Layer (type)                Output Shape              Param #   Connected to
=================================================================================================
 input_1 (InputLayer)        [(None, 100)]             0         []

 input_2 (InputLayer)        [(None, 100)]             0         []

 embedding (Embedding)       (None, 100, 64)           64000     ['input_1[0][0]',
                                                                  'input_2[0][0]']

 conv1d (Conv1D)             (None, 97, 64)            16448     ['embedding[0][0]']

 conv1d_1 (Conv1D)           (None, 97, 64)            16448     ['embedding[1][0]']

 layer_normalization (Layer  (None, 97, 64)            128       ['conv1d[0][0]']
 Normalization)

 layer_normalization_1 (Lay  (None, 97, 64)            128       ['conv1d_1[0][0]']
 erNormalization)

 global_max_pooling1d (Glob  (None, 64)                0         ['layer_normalization[0][0]']
 alMaxPooling1D)

 global_max_pooling1d_1 (Gl  (None, 64)                0         ['layer_normalization_1[0][0]'
 obalMaxPooling1D)                                                ]

 concatenate (Concatenate)   (None, 128)               0         ['global_max_pooling1d[0][0]',
                                                                  'global_max_pooling1d_1[0][0]
                                                                  ']

 dropout (Dropout)           (None, 128)               0         ['concatenate[0][0]']

 dense (Dense)               (None, 64)                8256      ['dropout[0][0]']

 dropout_1 (Dropout)         (None, 64)                0         ['dense[0][0]']

 dense_1 (Dense)             (None, 32)                2080      ['dropout_1[0][0]']

 dense_2 (Dense)             (None, 1)                 33        ['dense_1[0][0]']

=================================================================================================
Total params: 107521 (420.00 KB)
Trainable params: 107521 (420.00 KB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________________________________________
```

CNN 모델이 LSTM 모델보다 해석 용이

하이퍼파라미터 조정 후 최적 모델 선정

예측 결과 예시

title

'"서울부터 대구까지 232㎞ 거리"…22만개 팔린 '이 소파''

content

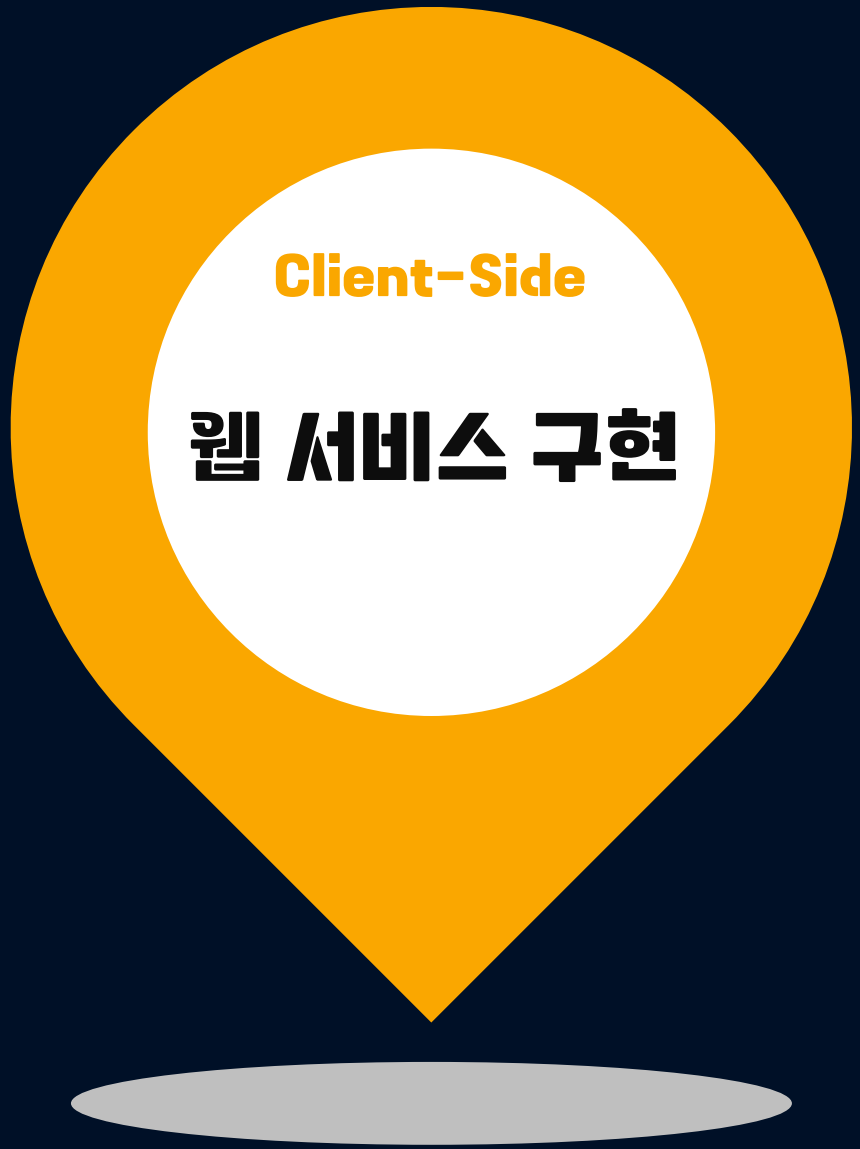
"신세계까사 까사미아 캄포 소파패브릭·착석감, 모듈기능이 강점[서울=뉴시스] 신세계까사 까사미아 캄포 소파. (사진=신세계까사 제공) 2023.12.06. photo@newsis.com[서울=뉴시스] 배민욱 기자 = 신세계까사의 베스트셀러 소파 '캄포'가 누적 판매 22만개 돌파했다. 6일 신세계까사에 따르면 캄포는 신세계까사가 신세계그룹 편입 이후 대표 브랜드 '까사미아'의 상품 경쟁력 강화 일환으로 선보인 소파다. 2019년 출시 당시 가죽 소파 선호도가 높던 국내 시장에서 10분에 1개씩 팔리며 패브릭 소파 열풍을 주도했다. 판매된 캄포 소파 제품을 일렬로 나열하면 232㎞다. 서울에서 대구까지 갈 수 있는 거리다.캄포는 4년째 베스트셀러의 자리를 지키고 있다. 캄포 소파의 인기 배경에는 기능성 패브릭, 편안한 착석감 , 모듈 기능의 장점이 고르게 작용된 것으로 보인다. 신세계까사는 캄포 누적 판매 22만개 돌파를 기념해 '어메이징 캄포' 프로모션을 내년 1월 14일까지 진행한다. 지난달 30일 이전까지 신세계까사의 온라인 쇼핑 플랫폼 '굳닷컴'과 기타 온라인몰, 까사미아 오프라인 매장에서 캄포를 구매한 고객은 '굳닷컴'에서 캄포 구매 인증 시 쇼핑 지원금 1만 포인트를 받는다. 굳닷컴이나 까사미아 오프라인 매장에서 캄포를 구매한 고객이 매장에서 직원을 통해 구매 인증 후 까사미아 제품을 구매하면 금액별로 신세계상품권 2만·4만·6만원권도 선물한다.이벤트 기간 내 굳닷컴에서 캄포 클래식 또는 슬림 소파 신규 구매 고객은 기본 모듈 커버 1세트와 굳포인트 5만점 페이백 등의 혜택을 제공 받는다. 17일까지 매장에서 웨딩·입주 클럽 가입 후 캄포 플러스 제품을 300만원 이상 구매하면 신세계상품권 10만원권이 제공된다. SNS(사회관계망서비스) 공유 이벤트도 있다. 캄포와 함께 한 일상의 순간을 개인 SNS에 공유하면 추첨을 통해 까사미아 상품권 50만원, 굳포인트 3만점, 스타벅스 기프티콘 등이 증정된다. 자세한 내용은 까사미아 공식 인스타그램에서 확인할 수 있다. "

```
pred_
1/1 [==============================] - 1s 508ms/step
63
```

예측 점수 결과 : 0~100점 사이로 집계

사용자들이 기사의 신뢰성을 보고 믿을만한지 판단 가능

예시 예측 결과 63점이 나온 것으로 확인


Client-Side
웹 서비스 구현

웹 서비스 구현

웹서비스 - 웹 구상도

Template 폴더 생성 후 HTML 파일을 넣어 웹 서버를 구축

Static 폴더에 CSS, JS 파일을 넣어 웹 서버에 렌더링

JavaScript 활용한 웹 페이지 애니메이션 구현

HTML과 CSS를 활용하여 웹 서비스 기능 구현


Daegu
Times
뉴스 핫 트렌드
뉴스에 대해 얼마나 알고 계시나요?
시작하기

KDT 빅데이터-AI 양성과정

자체평가

1. 한계점 및 개선사항

자체평가

자체 평가 및 의견 – 한계점 및 개선사항

NER Tagging

Tagging을 활용한 키워드(인물, 기관)
현재 행보 추적 기능 구현 추가 예정

모델 퀄리티

모델링 과정에서 다양한 벡터화 방법과
모델학습기법을 사용했어야 했지만,
실증랩실 사용 제한과 시간적인 문제로
모델생성의 한계

기사의 퀄리티 차이

신문사마다 기사의 양과 퀄리티
차이로 모델링 전 전처리 과정의
어려움

대한민국 전 지역 상용화

대구 지역에서 벗어나
전국의 뉴스데이터로 확대 가능

Daegu Times

Thank You