RIDE THE WAVES

DATA SCIENTIST PORTFOLIO.

Waves

CONTACT
gghdwl1103@gmail.com

CONTENTS

거대한
데이터의 파도를
즐기는 **서퍼**

데이터 사이언티스트를 꿈꾸는,
홍지원 입니다.

PROFILE

홍지원

1998.11.03
Goyang

Mobile +82)10.9179.5518
E-mail gghdwl1103@gmail.com
GitHub github.com/sweetpersimmon

EDUCATION

2022.08
국민대학교 빅데이터경영통계전공 졸업

2017.02
능곡고등학교 졸업

License

2017.04
MOS Excel Expert

2018.02
컴퓨터 활용능력 2급

2021.12
Adsp(데이터분석 준전문가)

ACTIVITIES

2021.07 - 현재
전국 빅데이터 연합동아리 BOAZ 활동

2021.06 - 2021. 10
LG DX대학 Python 프로그래밍
기초과정 튜터

2021.08
한국장학재단 대학생 재능봉사 캠프

2018.03 - 2020.12
교내 빅데이터 분석 학회 D&A 활동

AWARDS

2021.08
2021 기상청 빅데이터 콘테스트, 우수상

2020.12
2020 빅콘테스트, 혁신아이디어 최우수상

SKILLS

Python

R

MS Office

01 중고거래 자판기 입지 선정 제안

2020 Bigcontest

OVERVIEW

2020 빅콘테스트 혁신아이디어분야 최우수상 수상작
패션 산업을 돕는 비대면 중고거래 자판기 아이디어 제시, MCLP 알고리즘으로
자판기의 최적 입지 선정

PROJECT INFORMATION

Date 2020.08 - 2020. 12

Project in 2020 빅콘테스트

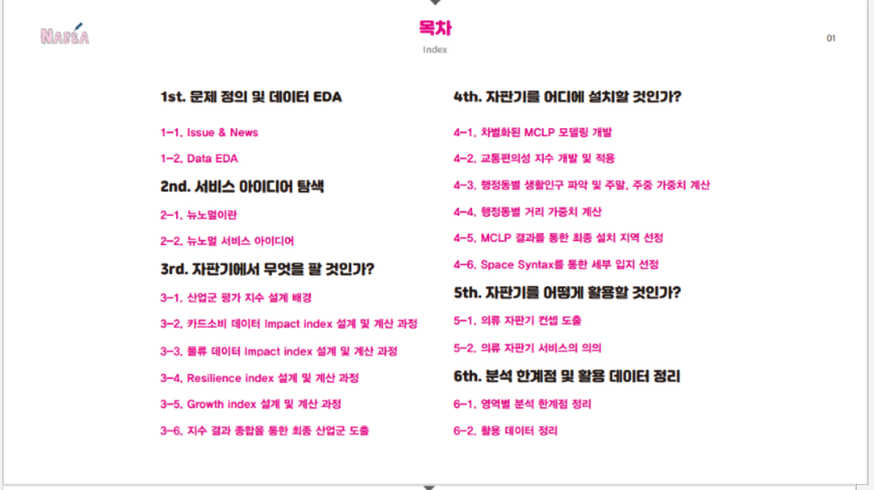
AWARD

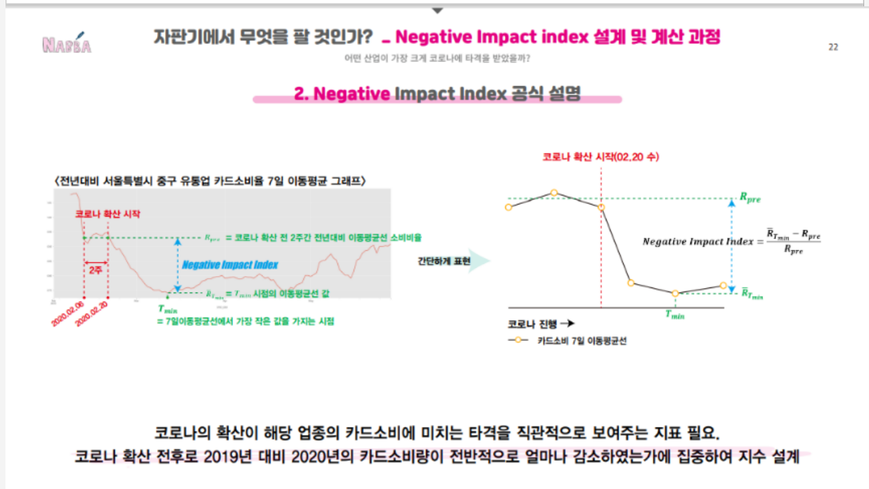
2020 빅콘테스트

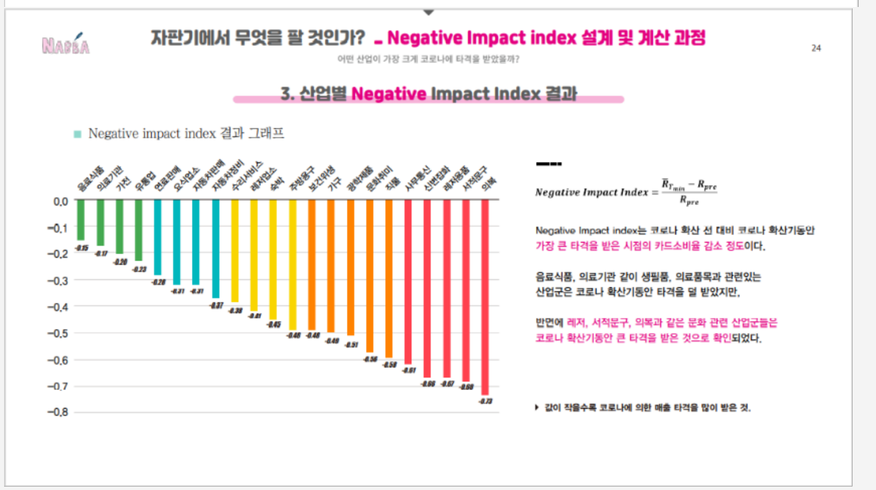
혁신 아이디어 부문
뉴노멀 시대 준비를 위한 아이디어 및 POC 제시

- 산업군을 평가하는 네 개의 인덱스를 개발하여 코로나 시기 가장 타격을 많이 받은 산업으로 패션 산업군을 도출, 이에 기반해 중고거래와 비대면 요소를 충족하는 자판기 아이디어 제시

- 메인 아이템인 의류 중고거래 자판기를 설치할 최적의 입지를 선정하기 위해 기존의 MLCP 모델을 개선하여 실용성 강화







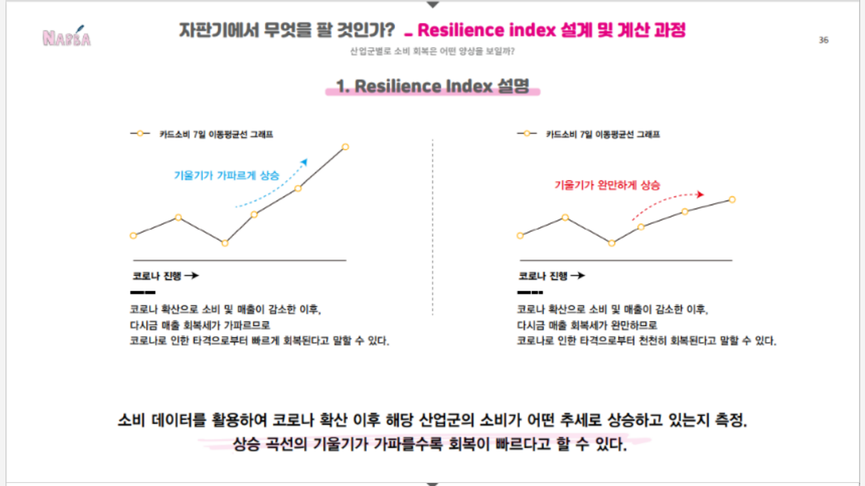
2020'S

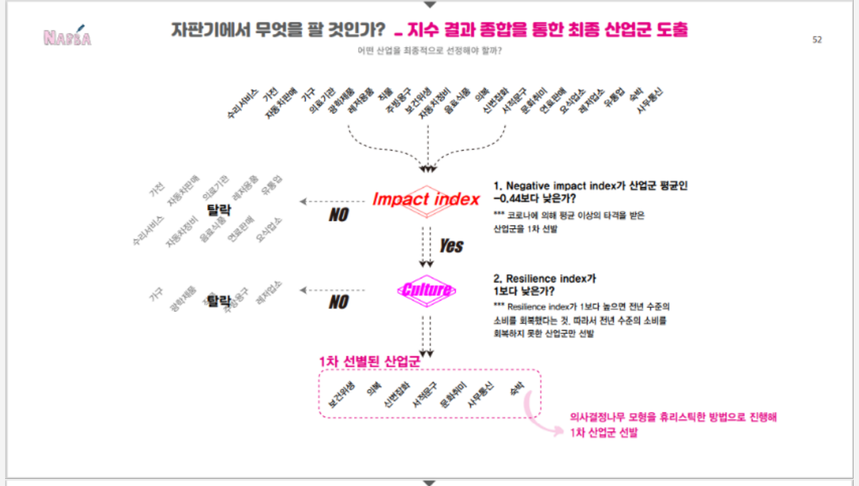
2020 빅콘테스트

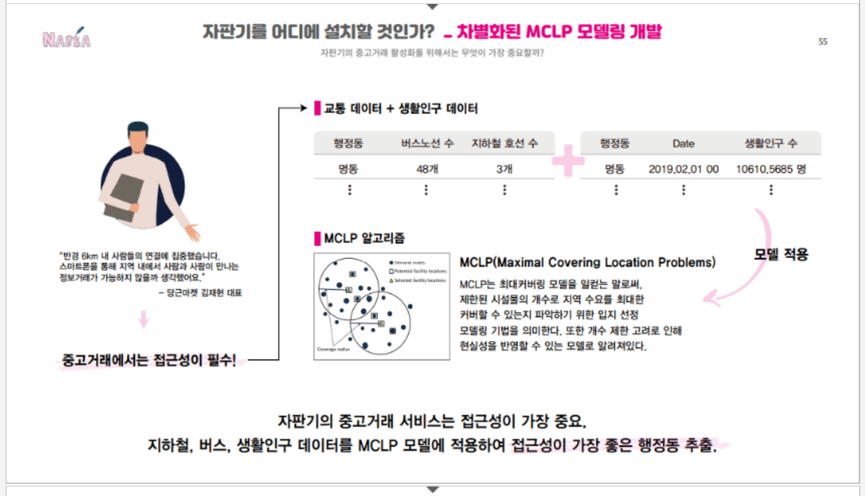
혁신 아이디어 부문
뉴노멀 시대 준비를 위한 아이디어 및 POC 제시

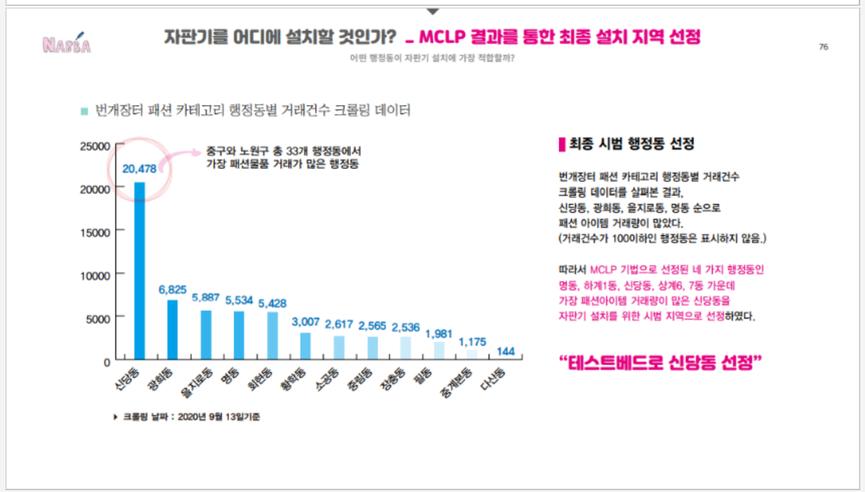
- 산업군을 평가하는 네 개의 인덱스를 개발하여 코로나 시기 가장 타격을 많이 받은 산업으로 패션 산업군을 도출, 이에 기반해 중고거래와 비대면 요소를 충족하는 자판기 아이디어 제시

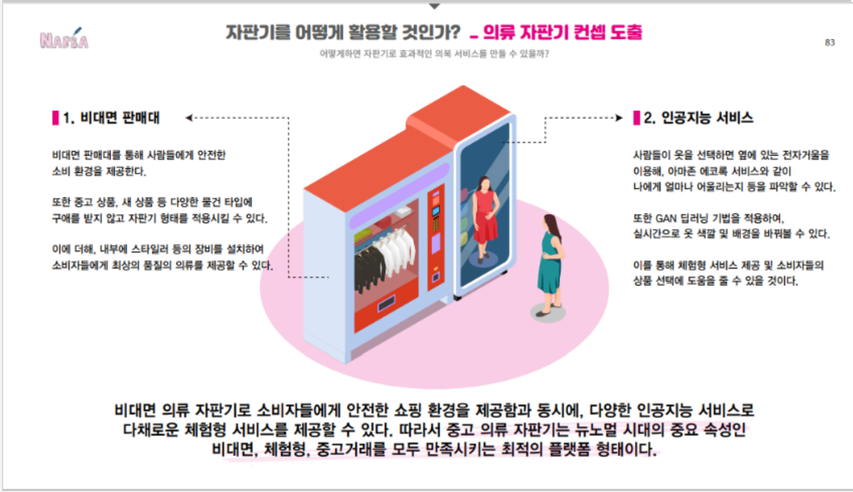
- 메인 아이템인 의류 중고거래 자판기를 설치할 최적의 입지를 선정하기 위해 기존의 MLCP 모델을 개선하여 실용성 강화











02

당근마켓 APP 리뷰 감성분석

Lecture

OVERVIEW

중고거래 플랫폼의 구글 플레이스토어 평점과 리뷰를 토대로 사용자 앱 사용 경험 분석

LSTM 알고리즘으로사용자 리뷰 긍부정 예측 모델 제작하여 새로운 리뷰 관리 모델 제시
리뷰 사용 어휘를 시각화하여 긍정 의견과 부정 의견의 핵심 키워드 탐색

PROJECT INFORMATION

Date 2020.06

Project in 텍스트 분석 수업 중간과제

2020'S

당근마켓 APP 리뷰 감성분석

사용자의 앱 텍스트 리뷰 기반 이용 경험 분석

- 코로나 시기 많은 사람들이 이용하는 중고거래 플랫폼으로 부상한 당근마켓의 리뷰를 크롤링하여 데이터 수집

- LSTM 알고리즘을 이용해 긍/부정 예측 모델을 제작하여 사용자 경험에 기반한 새로운 리뷰 관리 모델 제시

- FastText으로 워드 임베딩 후 T-SNE로 리뷰 속 어휘들을 시각화하여 긍/부정 리뷰 간 주로 사용하는 어휘의 분포 확인

당신 근처의 마켓, 얼마나 편한가요?

사용자 리뷰를 중심으로

국민대 빅데이터 경영통계전공 20172865 홍지원

CONTENTS,

01 분석 내용 요약

사용자 리뷰들은 긍정과 부정의 공통된 특징이 있을 것이다

사용자리뷰의 내용은 앞뒤 맥락이 있다고 가정하여 **순환신경망 LSTM**을 이용한 긍정/부정 감성분석을 실시한 결과, **93%의 정확도**로 리뷰가 잘 분류되었습니다.

긍정, 부정 리뷰에서 주로 사용하는 단어가 있을 것이다

긍정 리뷰와 부정 리뷰에서 사용자들이 **주로 느끼는 장점과 단점**이 무엇인지 **FASTTEXT**로 좌표평면에 단어들을 임베딩하여 군집화 했습니다.

당근마켓의 확연한 장점과 단점

GOOD

안 쓰는 물건을 필요한 곳으로 간편한 우리동네 직거래

나에게 **필요 없던 것이** 다른 사람에게 **중요한 물건으로**, 멀리 나가지 않아도 당일 거래 가능

BAD

제대로 오지 않는 키워드 알림, 아쉬운 지역 설정

필요한 물건의 키워드가 판매글에 올라와도 **알림이 제때 오지 않아** 매물을 놓치는 경우가 있어 오류 수정 시급

바로 옆 동네이지만, 지역 구분이 달라 사고 싶은 매물을 살 수가 없으므로 **지역 범위 확대를 고려**해야

02 분석 목적 - 가설 설정

01 긍정적 리뷰와 부정적 리뷰 간 서로 다른 특징이 있을 것이다

사용자가 만족한 상황이나 불만 사항을 느꼈을 때, 표현하는 문장이 다르다.
긍정적인 리뷰와 부정적인 리뷰는 문장 구성이나 단어 선택 면에서 확연히 다른 특징이 있을 것이다.

이들의 특징을 잘 파악할 수 있다면,
불만족스러운 고객에게 즉각적인 피드백을 줄 수 있다.

02 상황 별 주로 사용하는 특정한 단어가 있을 것이다

긍정인 상황과 부정인 상황에서도 어떤 요인 때문에 긍부정 판단을 내리게 되었는지는 사람마다 차이가 있다.

감정 별 **사용하는 단어들 간의 관계를 파악하여 군집화**하면, 어플의 장점과 단점을 알 수 있어 고객 분석에 도움이 된다.

02 분석 배경

날로 치솟는 당근마켓의 인기

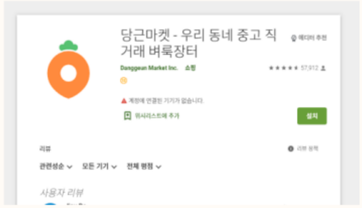
'당신 근처의 마켓'이란 뜻으로, 중고 거래 플랫폼

사용자 위치로부터 3~4km 이내에서 최대 6km 내의 이웃에게만 판매글이 노출되어 직거래를 유도

지역 기반의 우리 동네에서의 직거래 서비스로 선풍적인 인기를 끌고 있다.

자료출처 : 스포츠서울 http://www.sportsseoul.com/news/read/929355?ref=naver
파이낸셜뉴스 https://www.fnnews.com/news/202006250923241182

03 수집 데이터



Google Play Store에서
Selenium으로 데이터 수집

총 2142건의 데이터

별점 순으로 나누기에는 긍정과 부정이 명확하지 않아
직접 긍정(1), 부정(0) 라벨링 진행

2020'S

당근마켓 APP 리뷰 감성분석

사용자의 앱 텍스트 리뷰 기반 이용 경험 분석

- 코로나 시기 많은 사람들이 이용하는 중고거래 플랫폼으로 부상한 당근마켓의 리뷰를 크롤링하여 데이터 수집

- LSTM 알고리즘을 이용해 긍/부정 예측 모델을 제작하여 사용자 경험에 기반한 새로운 리뷰 관리 모델 제시

- FastText으로 워드 임베딩 후 T-SNE로 리뷰 속 어휘들을 시각화하여 긍/부정 리뷰 간 주로 사용하는 어휘의 분포 확인

03 분석1 - LSTM

maxlen=100으로 pad squences

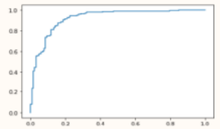
Optimizer=Adam의 최적의 학습률 0.003

Early Stopping의 평가지표 AUC

» LSTM »

정확도 93.33%

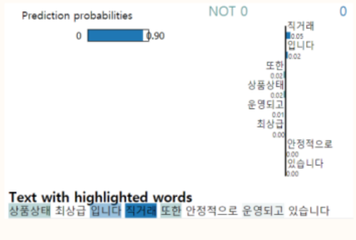
AUC-ROC 92.42%

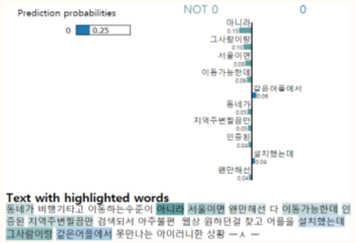


△AUC-ROC 곡선

03 분석1 - LIME

(Stop Words는 동일)
Okt로 토큰화 하는 함수를 만들어
CountVectorize
» DNN 모델로 학습 » 93.5% 정확도





긍정 평가와 부정 평가의 차이가 확연한 것을 알 수 있다.

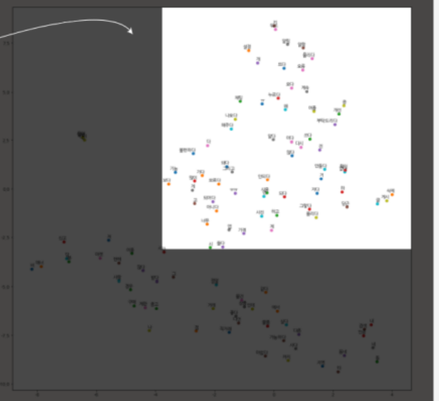
03 분석2 - FASTTEXT

알람, 뜨다, 오류,
키워드, 설정, 모르다,
불안하다

부정 평가

=키워드 알림이 제때 뜨지 않아 불안하거나 오류로 알림이 많이 뜨는 상황이 있음을 추측

군집으로 분리



03 분석2 - FASTTEXT

처음, 이용, 해봤다, 유용하다,
감사하다, 도움, 중고, 당근, 마켓

=중고거래 어플이나 당근마켓을 처음 이용해 본 사용자들이 유용하다고 평가한 내용임을 알 수 있다.

군집으로 분리

04 실행제안

긍정 평가와 부정 평가는 **확연히 구분되는 문장 구성**을 갖추고 있음을 확인

+

긍정 평가와 부정 평가의 **단어들을 군집화**하여 당근 마켓의 **장점**과 **단점**을 확인할 수 있음

=

1 불만족 이용자에게 빠른 응대
긍정리뷰와 부정리뷰의 패턴을 파악할 수 있기 때문에 불만족 상태인 이용자에게 빠른 고객응대를 할 수 있다.

2 다음 업데이트 때 단점을 반영하여 개선
리뷰 중에 많이 언급되는 단점을 인지하고 다음 업데이트 때 단점을 반영하여 문제에 적극 대처하는 모습을 보여줄 수 있다.

3 홍보 시에 장점 강조
이용자들이 많이 언급한 장점을 홍보 시에 강조하면 당근마켓을 아직 이용해보지 않은 사람들도 쉽게 어플에 접근이 가능할 것이다.

감사합니다.

03

뉴스 토픽 분류 AI 경진대회

Dacon

OVERVIEW

한국어 뉴스 헤드라인을 이용하여, 뉴스의 주제를 분류하는 알고리즘 개발
Transformers 기반 언어 모델로 모델링 진행

PROJECT INFORMATION

Date 2021.06 - 2021. 08

Project in Dacon 월간대회

2021'S

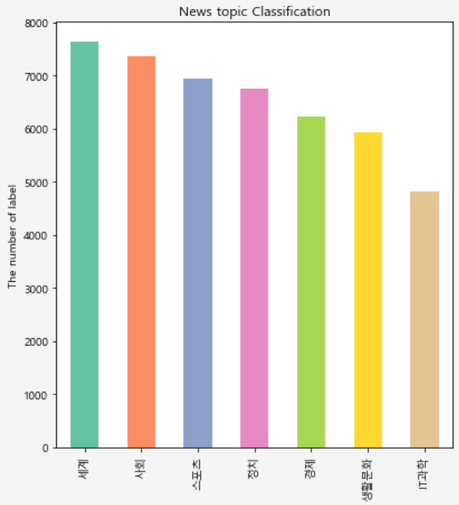
뉴스 토픽 분류 AI 경진대회

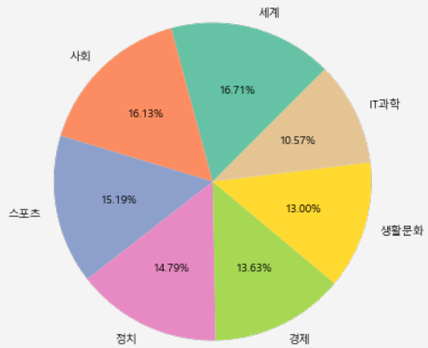
한국어 뉴스 헤드라인으로 뉴스의 주제를 분류하는 알고리즘 개발

- 6개의 뉴스토픽 분류를 위해 트랜스포머 기반 뉴스 기사로 사전 학습한 모델로 예측 진행

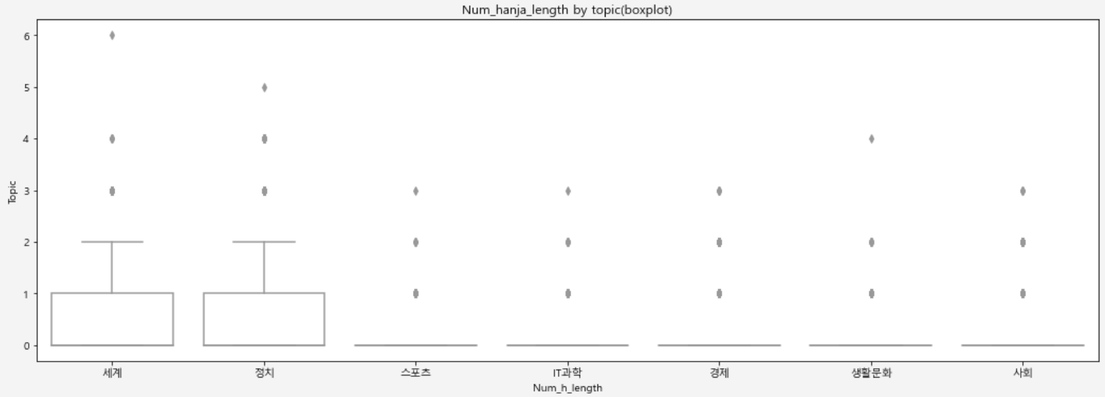
- 뉴스에 주로 등장하는 한자 표기를 한글로 번역하는 전처리를 통해 성능 향상

01. EDA





Train 데이터 내 6개의 뉴스 토픽 (세계, 정치, 스포츠, IT/과학, 경제. 생활문화, 사회)의 분포 확인



의미 함축 등을 위해 한자 사용이 잦은 뉴스 헤드라인의 특성 상 세계, 정치 토픽에서 주로 한자를 많이 사용

2021'S

뉴스 토픽 분류 AI 경진대회

한국어 뉴스 헤드라인으로 뉴스의 주제를 분류하는 알고리즘 개발

- 6개의 뉴스토픽 분류를 위해 트랜스포머 기반 뉴스 기사로 사전 학습한 모델로 예측 진행

- 뉴스에 주로 등장하는 한자 표기를 한글로 번역하는 전처리를 통해 성능 향상

02. 모델링

'뉴스'라는 특정 분야의 데이터에 적합한 사전 학습 모델을 선정,
이들을 앙상블하여 최종 예측

KoBert

기본적으로 뉴스를 기반으로 한 데이터를 사전학습하여, 뉴스 관련 텍스트를 처리하는 데 탁월
뉴스 댓글도 포함한 KcBert에 비해서는 공식적인 언어에 더 강점을 보임

+

KoElectra

뉴스는 기본으로, KoBert보다 학습한 데이터의 양이 광범위하며
모든 입력 내용을 사용하여 학습하기 때문에 더 좋은 성능을 보임

+

Roberta

Bert를 기반으로 더 큰 규모의 파라미터 사용
Bert보다 더 긴 길이의 데이터로 사전학습되어 좋은 성능을 보임

최종 정확도 0.83464(Top 4%) 달성

THANK YOU.

CONTACT
gghdwl1103@gmail.com