



RIDE THE WAVES

DATA SCIENTIST PORTFOLIO.

Waves

CONTACT

gghdwl1103@gmail.com

RIDE THE WAVES

CONTENTS

거대한 데이터의 파도를 즐기는 서퍼

00

이력서

003

- Profile
- Education
- License
- Activities
- Awards
- Skills

02

당근마켓 APP 리뷰 감성분석

010

- 중고거래 플랫폼의 구글 플레이스토어 평점과 리뷰를 토대로 사용자 경험 분석
- LSTM 알고리즘을 이용한 사용자 리뷰 긍부정 예측
- 긍부정 의견의 워드 임베딩으로 핵심키워드 탐색

01

중고거래 자판기 입지 선정 제안 004

- 2020 빅콘테스트 혁신아이디어분야 최우수상 수상
- 패션 산업을 돕는 비대면 중고거래 자판기 아이디어 제시
- MLCP 알고리즘으로 자판기의 최적 입지 선정

03

뉴스 토픽 분류 AI 경진대회 040

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

데이터 사이언티스트를 꿈꾸는, 홍지원입니다.

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 연립동아리 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
빅콘테스트혁신 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

뉴노멀 시대 준비를 위한 아이디어 및 POC 제시

팀명 : 코로나나빠
팀장 : 송재원(rearsilre@naver.com)
팀원 : 홍지원(ghdw1103@naver.com)
: 김민기(xowhdtjd@naver.com)
: 김효은(heun7410@naver.com)
: 이종욱(ollehw@naver.com)

서비스 아이디어 탐색 - 뉴노멀 서비스 아이디어
뉴노멀 시대에 주목받는 서비스로는 무엇이 있을까?

최근 서비스 트렌드를 통해 뉴노멀시대에 가장 알맞은 서비스는 체험형, 중고거래, 비대면 형태임을 파악. 따라서 이에 맞는 서비스 형태로 POC를 발굴하는 것이 중요.

체험형 오프라인 매장, 중고거래 서비스, 비대면 서비스

과정 중심, 융합적 접근, 소유에서 공유로, 지속가능성, 창발 (작은 변화가 큰 변화를 이끈다), 비대면(안전)

자판기에서 무엇을 팔 것인가? - Negative Impact index 설계 및 계산 과정
어떤 산업이 가장 크게 코로나에 타격을 받았을까?

2. Negative Impact Index 공식 설명

〈전년대비 서울특별시 중구 유통업 카드소비를 7일 이동평균 그래프〉

코로나 확산 시작 (2020.02.20)

간단하게 표현

코로나 진행 →

카드소비 7일 이동평균선

코로나의 확산이 해당 업종의 카드소비에 미치는 타격을 직관적으로 보여주는 지표 필요. 코로나 확산 전후로 2019년 대비 2020년의 카드소비량이 전반적으로 얼마나 감소하였는가에 집중하여 지수 설계

목차
Index

1st. 문제 정의 및 데이터 EDA

1-1. Issue & News
1-2. Data EDA

2nd. 서비스 아이디어 탐색

2-1. 뉴노멀이란
2-2. 뉴노멀 서비스 아이디어

3rd. 자판기에서 무엇을 팔 것인가?

3-1. 산업군 평가 지수 설계 배경
3-2. 카드소비 데이터 Impact index 설계 및 계산 과정
3-3. 물류 데이터 Impact index 설계 및 계산 과정
3-4. Resilience index 설계 및 계산 과정
3-5. Growth index 설계 및 계산 과정
3-6. 지수 결과 종합을 통한 최종 산업군 도출

4th. 자판기를 어디에 설치할 것인가?

4-1. 차별화된 MCLP 모델링 개발
4-2. 교통편의성 지수 개발 및 적용
4-3. 행정동별 생활인구 파악 및 주말, 주중 가중치 계산
4-4. 행정동별 거리 가중치 계산
4-5. MCLP 결과를 통한 최종 설치 지역 선정
4-6. Space Syntax를 통한 세부 입지 선정

5th. 자판기를 어떻게 활용할 것인가?

5-1. 의류 자판기 컨셉 도출
5-2. 의류 자판기 서비스의 의의

6th. 분석 한계점 및 활용 데이터 정리

6-1. 영역별 분석 한계점 정리
6-2. 활용 데이터 정리

서비스 아이디어 탐색 - 뉴노멀 서비스 아이디어
뉴노멀 시대에 주목받는 서비스로는 무엇이 있을까?

VENDING MACHINES

중고거래

자판기는 주거지역, 지하철역 주변 등 사람들의 생활권 반경 내로 밀집하게 접근할 수 있는 매개체이다. 따라서 이런 접근성의 증가는 중고거래의 매력성을 증가시킨다.

또한 중고거래의 낮은 단가를 고려해 보았을 때, 화려한 매장 대신 친숙한 자판기가 더 알맞은 플랫폼 형태이다.

체험형

최근 오프라인 체험형 매장의 증가 추세에 따라 자판기를 소형 랩토퍼 컨셉으로 꾸민다면 소비자들에게 큰 호응을 받을 수 있을 것이라 생각.

또한 자판기에 AR, 홀로그램과 같은 부가 장치를 추가하여 소비자들에게 신선한 체험을 제공할 수 있다.

자판기는 중고거래, 오프라인 체험형 서비스, 비대면 서비스를 모두 만족시킬 수 있는 서비스 형태. 따라서 자판기를 중심으로 뉴노멀 서비스 아이디어를 제안

자판기에서 무엇을 팔 것인가? - Negative Impact index 설계 및 계산 과정
어떤 산업이 가장 크게 코로나에 타격을 받았을까?

3. 산업별 Negative Impact Index 결과

■ Negative impact index 결과 그래프

$$\text{Negative Impact Index} = \frac{\bar{R}_{T_{min}} - R_{pre}}{R_{pre}}$$

Negative Impact index는 코로나 확산 전 대비 코로나 확산기동안 가장 큰 타격을 받은 시점의 카드소비를 감소 정도이다.

음료식품, 의류기관 같이 생활품, 의료품목과 관련있는 산업군은 코로나 확산기동안 타격을 덜 받았지만,

반면에 레저, 서적문구, 의복과 같은 문화 관련 산업군들은 코로나 확산기동안 큰 타격을 받은 것으로 확인되었다.

▶ 값이 작을수록 코로나에 의한 매출 타격을 많이 받은 것.

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

자판기에서 무엇을 팔 것인가? - Resilience index 설계 및 계산 과정

산업군별로 소비 회복은 어떤 양상을 보일까?

36

1. Resilience Index 설명

— 코디소비 7월 이동평균선 그래프



코로나 진행 →

코로나 확산으로 소비 및 매출이 감소한 이후, 다시금 매출 회복세가 가파르므로 코로나로 인한 타격으로부터 빠르게 회복된다고 말할 수 있다.

— 코디소비 7월 이동평균선 그래프



코로나 진행 →

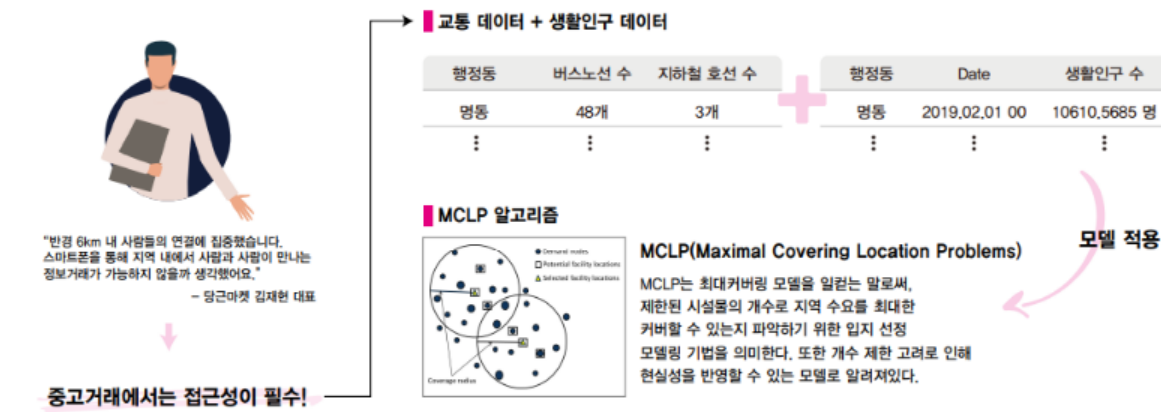
코로나 확산으로 소비 및 매출이 감소한 이후, 다시금 매출 회복세가 완만하므로 코로나로 인한 타격으로부터 천천히 회복된다고 말할 수 있다.

소비 데이터를 활용하여 코로나 확산 이후 해당 산업군의 소비가 어떤 추세로 상승하고 있는지 측정.
상승 곡선의 기울기가 가파를수록 회복이 빠르다고 할 수 있다.

자판기를 어디에 설치할 것인가? - 차별화된 MCLP 모델링 개발

자판기의 중고거래 활성화를 위해서는 무엇이 가장 중요할까?

55



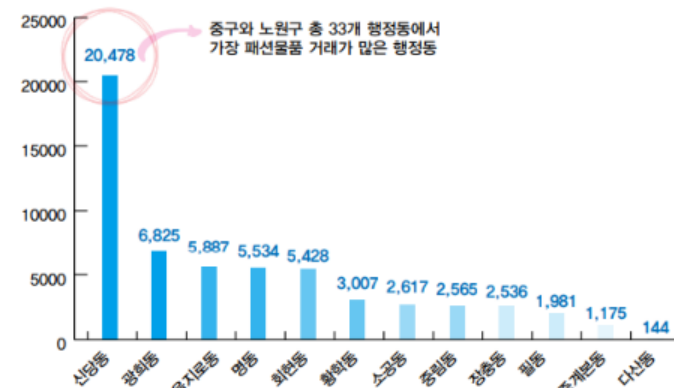
자판기의 중고거래 서비스는 접근성이 가장 중요.
지하철, 버스, 생활인구 데이터를 MCLP 모델에 적용하여 접근성이 가장 좋은 행정동 추출.

자판기를 어디에 설치할 것인가? - MCLP 결과를 통한 최종 설치 지역 선정

어떤 행정동이 자판기 설치에 가장 적합할까?

76

■ 번개장터 패션 카테고리 행정동별 거래건수 크롤링 데이터



▶ 크롤링 날짜 : 2020년 9월 13일 기준

최종 시범 행정동 선정

번개장터 패션 카테고리 행정동별 거래건수 크롤링 데이터를 살펴본 결과, 신당동, 광희동, 을지로동, 명동 순으로 패션 아이템 거래량이 많았다. (거래건수가 1000이 아닌 행정동은 표시하지 않음.)

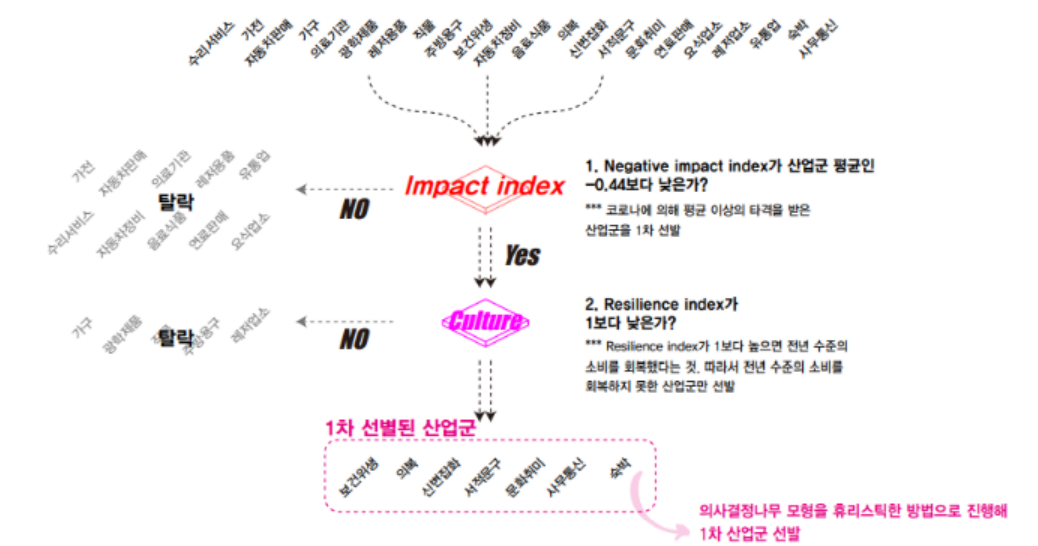
따라서 MCLP 기법으로 선정된 네 가지 행정동인 명동, 하계1동, 신당동, 상계6, 7동 가운데 가장 패션이템 거래량이 많은 신당동을 자판기 설치를 위한 시범 지역으로 선정하였다.

"테스트베드로 신당동 선정"

자판기에서 무엇을 팔 것인가? - 지수 결과 종합을 통한 최종 산업군 도출

어떤 산업을 최종적으로 선정해야 할까?

52



자판기를 어디에 설치할 것인가? - 차별화된 MCLP 모델링 개발

자판기의 중고거래 활성화를 위해서는 무엇이 가장 중요할까?

58

MCLP 모델링 적용 과정



1. 목적함수 정의

본 과제에서는 교통 데이터와 생활인구 데이터를 이용해 목적 함수 정의



2. 첫 번째 행정동 추출

MCLP 계산을 위해 정의한 목적함수를 최대화 하는 하나의 동을 추출



3. 두 번째 행정동 추출

불한 동과 인접동의 영향력을 줄여준 뒤 같은 알고리즘으로 추가적인 행정동 추출

모델링 적용 과정은 교통데이터와 생활인구 데이터로 정의한 목적함수를 최대화 시키는 행정동을 추출.
이후 먼저 추출된 행정동을 제외하고 이 과정을 2~4회 반복하면서 최종 접근성이 우수한 여러개의 행정동 추출.

자판기를 어떻게 활용할 것인가? - 의류 자판기 컨셉 도출

어떻게하면 자판기로 효과적인 의류 서비스를 만들 수 있을까?

83

1. 비대면 판매대

비대면 판매대를 통해 사람들에게 안전한 소비 환경을 제공한다.

또한 중고 상품, 새 상품 등 다양한 물건 타입에 구매를 받지 않고 자판기 형태를 적용시킬 수 있다.

이에 더해, 내부에 스타일러 등의 장비를 설치하여 소비자들에게 최상의 품질의 의류를 제공할 수 있다.

2. 인공지능 서비스

사람들이 옷을 선택하면 옆에 있는 전자거울을 이용해, 아마존 에코룩 서비스와 같이 나에게 얼마나 어울리는지 등을 파악할 수 있다.

또한 GAN 딥러닝 기법을 적용하여, 실시간으로 옷 색깔 및 배경을 바꿔볼 수 있다.

이를 통해 체험형 서비스 제공 및 소비자들의 상품 선택에 도움을 줄 수 있을 것이다.

비대면 의류 자판기로 소비자들에게 안전한 쇼핑 환경을 제공함과 동시에, 다양한 인공지능 서비스로 다채로운 체험형 서비스를 제공할 수 있다. 따라서 중고 의류 자판기는 뉴노멀 시대의 중요 속성인 비대면, 체험형, 중고거래를 모두 만족시키는 최적의 플랫폼 형태이다.

02

당근마켓 APP 리뷰 감성분석

Lecture

OVERVIEW

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

LSTM 알고리즘으로 사용자 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

01 분석 내용 요약

사용자 리뷰들은 긍정과 부정의
공통된 특징이 있을 것이다

사용자리뷰의 내용은 앞뒤 맥락이 있다고 가정하여
순환신경망 LSTM을 이용한 긍정/부정 감성분석을
실시한 결과, 93%의 정확도로 리뷰가 잘 분류되었습니다.

긍정, 부정 리뷰에서
주로 사용하는 단어가 있을 것이다

긍정 리뷰와 부정 리뷰에서 사용자들이 주로 느끼는
장점과 단점이 무엇인지 FASTTEXT로 좌표평면에
단어들을 임베딩하여 군집화 했습니다.



당근마켓의 확연한 장점과 단점

안 쓰는 물건을 필요한 곳으로
간편한 우리동네 직거래
나에게 필요 없던 것이 다른 사람에게 중요한 물건으로,
멀리 나가지 않아도 당일 거래 가능

GOOD

제대로 오지 않는 키워드 알림,
아쉬운 지역 설정

BAD

필요한 물건의 키워드가 판매글에 올라와도 알림이 제때 오
지 않아 매물을 놓치는 경우가 있어 오류 수정 시급
바로 옆 동네이지만, 지역 구분이 달라 사고 싶은 매물을 살
수가 없으므로 지역 범위 확대를 고려해야

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 군집화하면,
어플의 장점과 단점을 알 수 있어 고객 분석에 도움이 된다.

CONTENTS,

01

개요

분석 내용 요약

02

서론

당근마켓이 대체 뭐길래?
분석의 목적, 배경, 필요성

03

본론

1 플레이 스토어 데이터 수집
2 데이터 분석 :
LSTM, FASTTEXT
3 결과 해석

04

결론

본 분석의 실행제한

02 분석 배경

날로 치솟는
당근마켓의 인기

'당신 근처의 마켓'이란 뜻으로, 중고 거래 플랫폼

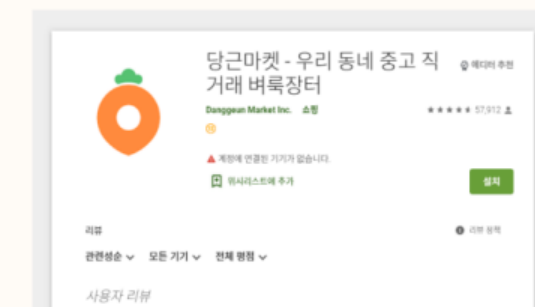
사용자 위치로부터 3~4km 이내에서 최대 6km 내의
이웃에게만 판매글이 노출되어 직거래를 유도

지역 기반의 우리 동네에서의 직거래 서비스로
선봉적인 인기를 끌고 있다.



자료 출처: 스포츠서울 <http://www.sportsseoul.com/news/read/929355?ref=naver>
파이낸셜 뉴스 <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

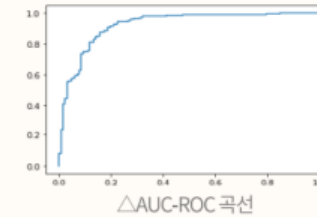
maxlen=100으로 pad sequences

Optimizer=Adam의 최적의 학습률 0.003

Early Stopping의 평가지표 AUC

» LSTM »

정확도 93.33%
AUC-ROC 92.42%

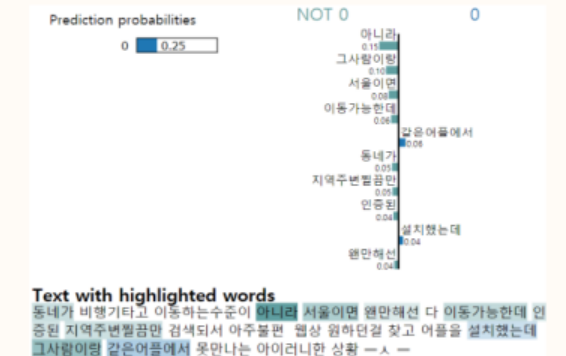
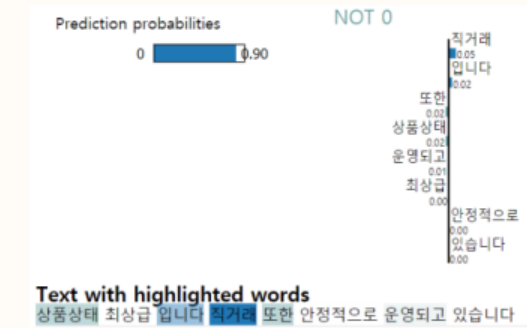


03 분석1 - LIME

(Stop Words는 동일)
Okt로 토큰화 하는 함수를 만들어
CountVectorize

» DNN 모델로 학습

» 93.5% 정확도

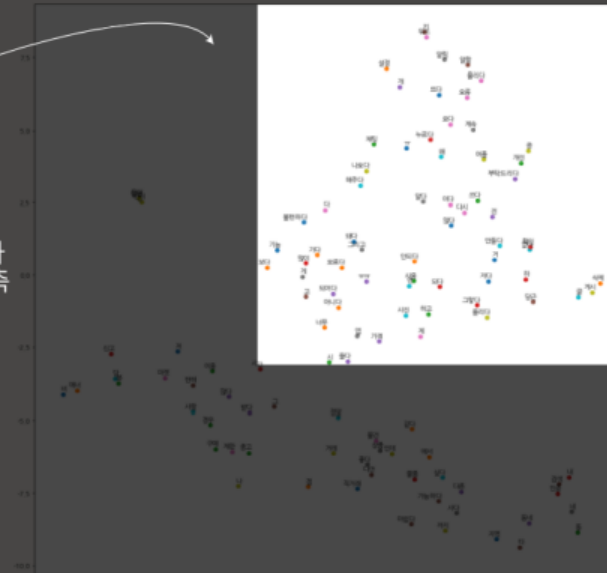


긍정 평가와 부정 평가의 차이가 확연한 것을 알 수 있다.

03 분석2 - FASTTEXT

알람, 뜨다, 오류,
키워드, 설정, 모르다,
불안하다

부정 평가
= 키워드 알람이 제때 뜨지 않아 불안하거나
오류로 알람이 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 평가
한 내용임을 알 수 있다.
군집으로 분리



04 실행제안

긍정 평가와 부정 평가는 확연히 구분되는
문장 구성을 갖추고 있음을 확인

+

=

긍정 평가와 부정 평가의 단어들을
군집화하여 당근마켓의 장점과 단점을
확인할 수 있음

1 불만족 이용자에게 빠른 응대

긍정 리뷰와 부정 리뷰의 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

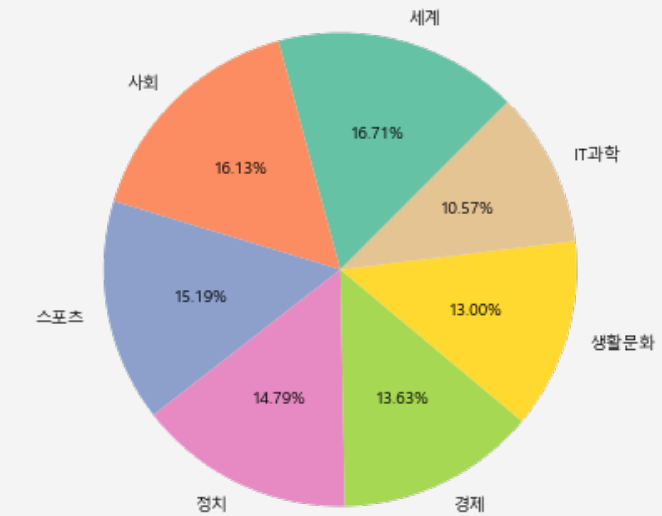
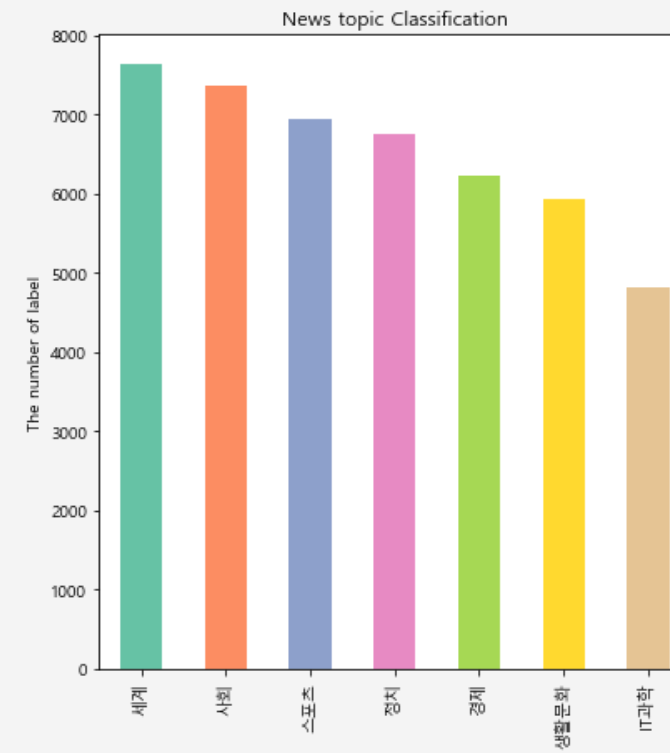
뉴스 토픽 분류 AI 경진대회

한국어 뉴스 헤드라인으로
뉴스의 주제를 분류하는
알고리즘 개발

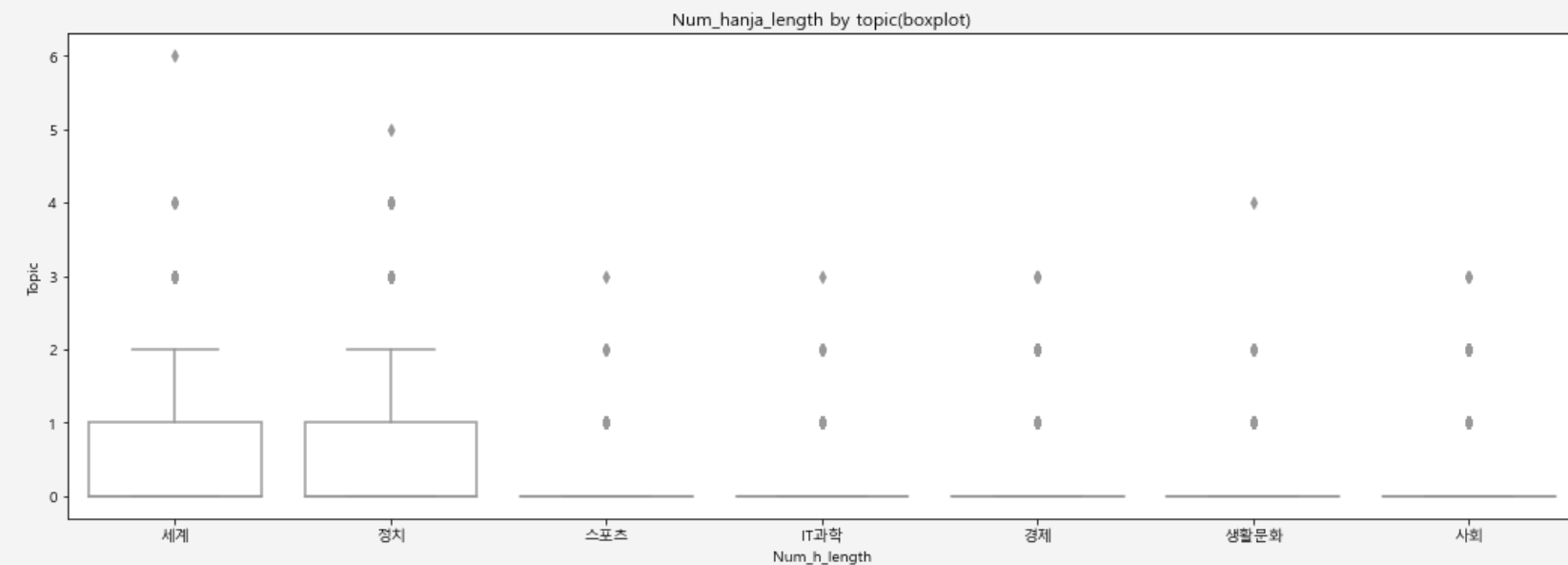
- 6개의 뉴스토픽 분류를 위해 트랜스포머
기반 뉴스 기사로 사전 학습한 모델로 예측
진행

- 뉴스에 주로 등장하는 한자 표기를 한글
로 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 한글
로 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

RIDE THE WAVES