텍스트분석 스터디
4조

김기호 이가영 이승준

텍스트 분석 기법 스터디

⇩

데이콘 프로젝트 진행

발표 순서

1 텍스트 분석 순서

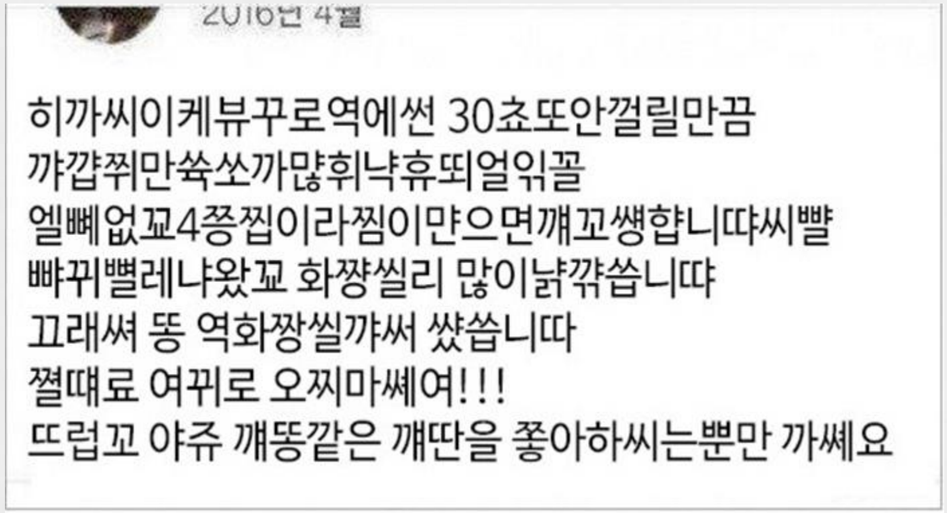
2 프로젝트 소개

3 데이터 전처리 및 EDA

4 모델링 및 소감

① 토큰화(Tokenization) - 벡터화 - 모델링

자연어처리(NLP)에서 토큰화 (Tokenization)이란 데이터를 처리하기 위해서
최소한의 의미를 기반으로 토큰(Token)을 뽑는 것을 말한다.



KoNLPy

한국어 처리 패키지

① 토큰화(Tokenization) - 벡터화 - 모델링

KoNLPy 패키지 내의 형태소 분석기

Hannanum	KAIST 말뭉치를 이용해 생성된 사전	일부 케이스 분석 품질 저하 ex) 띄어쓰기 없는 문장
Kkma	세종 말뭉치를 이용해 생성된 사전	정확한 품사 분류, 긴 분석 시간
Mecab	일본어용 형태소 분석기를 한국어로 수정	속도가 제일 빠르다 새로운 단어 추가 가능
Komoran	Java로 쓰여진 오픈소스 한글 형태소 분석기	빠른 속도와 보통의 분석 품질 자바가 설치된 환경이라면 어디서든 사용가능
Okt	오픈 소스 한국어 분석기, 과거 Twitter 형태소 분석기	단어들을 정규화, 오타 수정 기능이 있음 비형식어, 신조어 등을 상대적으로 잘 찾아냄.

① 토큰화(Tokenization) - 벡터화 - 모델링

<형태소 분석기 예시>

mecab.morphs → ['대한민국', '헌법', '유구', '한', '역사', '와', '전통', '에', '빛나', '는', '우리', '대한', '국민', '은', '3', '·', '1', '운동', '으로']

kkma.morphs → ['대한민국', '헌법', '유구', '하', 'ㄴ', '역사', '와', '전통', '에', '빛나', '는', '우리', '대하', 'ㄴ', '국민', '은', '3', '·', '1', '운동', '으로']

okt.morphs → ['대한민국', '헌법', '유구', '한', '역사', '와', '전통', '에', '빛나는', '우리', '대', '한', '국민', '은', '3', '·', '1', '운동', '으로']

① 토큰화 - 벡터화(Vectorize) - 모델링

Frequency Based Embedding

TF-IDF (Term Frequency - inverse Document Frequency)

(TF:단어빈도) * (IDF:역문서빈도)

TF = 특정 문서에서 단어가 나타난 수 / 특정 문서에 있는 전체 단어의 개수
IDF = log(말뭉치에서의 전체 문서의 수 / 말뭉치에서 해당 단어가 나타난 문서의 수)

TF-IDF는 단어의 빈도와 역 문서 빈도를 사용하여 문서 단어 행렬(DTM) 내의 각 단어들마다 중요한 정도를 가중치로 주는 방법입니다. 우선 문서 단어 행렬(DTM)을 만든 후, TF-IDF 가중치를 부여한다.

문서의 핵심어를 추출하거나, 검색 엔진에서 검색 결과의 순위를 결정하거나,
문서들 사이의 비슷한 정도를 구하는 등의 용도로 사용할 수 있다.

	과일이	길고	노란	먹고	바나나	사과	싶은	저는	좋아요
문서1	0	0	0	1	0	1	1	0	0
문서2	0	0	0	1	1	0	1	0	0
문서3	0	1	1	0	2	0	0	0	0
문서4	1	0	0	0	0	0	0	1	1

<문서 단어 행렬, DTM>

① 토큰화 - 벡터화(Vectorize) - 모델링

Frequency Based Embedding

Word2vec - Word to Vector

단어를 벡터로 변환

유사도를 계산하지 못하고, 차원이 복잡하다는 One-hot encoding의 단점을 보완
'비슷한 위치에서 등장하는 단어들은 비슷한 의미를 가진다' 라는 가정 하에 만들어진 표현 방법

Ex)
One-hot encoding
단어 - [0,0,0,0,1,0,…,0]
Word2vec
단어 - [0.2,0.3,0.5,0,7,0.2,…,0.2]

단어를 표현하기 위해 사용자가 설정한 차원을 가지는 벡터가 되면서 각 차원은 실수형의 값을 가진다.

① 토큰화- 벡터화 - 모델링(modeling)

Bert

pre-training이 가능한 모델

MLM - 입력 문장에서 15%의 단어를 가리고 가려진 단어를 맞추는 방법으로 학습

NSP - 두 문장이 주어졌을 때 문맥상 첫번째 문장의 다음에 두번째 문장이 올 수 있는지를 예측

-주어진 질문에 적합하게 대답하기
-번역기
-문장 주제 찾기 또는 분류하기

한글화

KoBERT
- Korean BERT

위키피디아나 뉴스 등에서 수집한 수백만 개의 한국어 문장으로 이루어진 대규모말뭉치(corpus)를 학습

KcBert
- Korean comments BERT

네이버 댓글과 대댓글에 나타나는 구어체 특징, 신조어, 오탈자 등을 반영한 모델

등등...

① 토큰화- 벡터화 - 모델링(modeling)

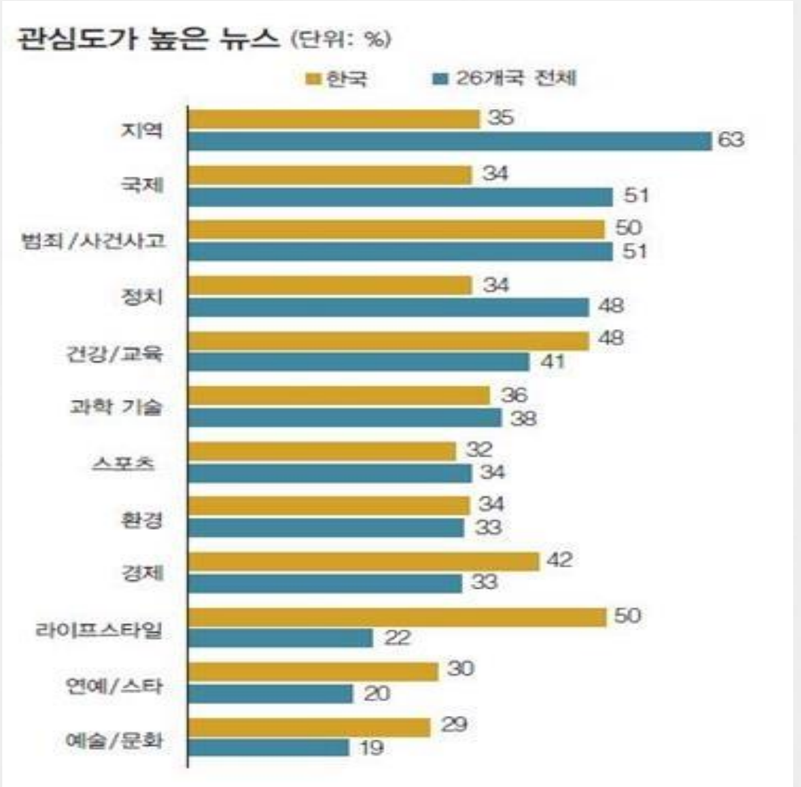
LGBM - Light Gradient Boosting Machine	Logistic Regresion - 로지스틱 회귀
GBM(Gradient Boosting Machine)(틀린부분에 가중치를 더하면서 진행하는 알고리즘)	선형 회귀를 먼저 실행한 후, 그 결과값에 로지스틱 함수를 사용하여 값을 분류
메모리를 적게 차지, 속도가 빠름, GPU 활용 가능	쉬운 정규화, 편리한 해석 빠른 학습, 예측 속도
과적합의 우려가 다른 Tree 알고리즘 대비 높은 편	훨씬 더 좋고 더 복잡한 예측을 생성할 수 있는 모델 有 비선형 문제를 해결하는 데 사용할 수 없음 과적합에 취약

② 프로젝트 소개 - 데이터 소개

뉴스 토픽 분류 AI 경진대회

<분석 목적>

연합뉴스 헤드라인을 이용하여 뉴스의 주제를 분류하는 알고리즘을 개발하고자 함.
개발한 알고리즘을 이용하여 원하는 주제의 뉴스를 쉽게 찾고자 함.



② 프로젝트 소개 - 데이터 소개

-Topic_dict

- 실제 뉴스 토픽을 나타낸 데이터
- 뉴스 토픽 인덱스 값

Topic	Topic_idx
IT과학	0
경제	1
사회	2
생활문화	3
세계	4
스포츠	5
정치	6

-Train<총 45654행 X 3열>

- 총 45654개 뉴스 헤드라인으로 구성된 데이터

-Test<총 9131행 X 2열>

- 총 9131개 뉴스 헤드라인으로 구성된 데이터

Index	Title	Topic_idx
0	인천→핀란드 항공기 결항… 휴가철 여행객 분통	4
1	실리콘밸리 넘어서겠다…구글 15조원 들여 美전역 거점화	4
…	…	…
45652	답변하는 배기동 국립중앙박물관장	2
45653	2020 한국인터넷기자상 시상식 내달 1일 개최… 특별상 김성후	2

③ 함수 생성 - 토큰화 - 벡터화

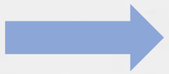
<데이터 전처리를 위한 함수 생성>

1. 필요 없는 단어 삭제 및 교체

인천→핀란드 항공기 결항…

미국 산업생산 한달만에 0.1%↑…제조업 회복 기대

LG 방출→호주야구 응시생 장진용 야구 끈 놓지 않아



인천에서 핀란드 항공기 결항…

미국 산업생산 한달만에 0.1% 증가… 제조업 회복 기대

LG 방출에서 호주야구 응시생 장진용 야구 끈 놓지 않아

③ 함수 생성 - 토큰화 - 벡터화

<데이터 전처리를 위한 함수 생성>

2. 한자 한글로 교체

실리콘밸리 넘어서겠다…구글 15조원 들여 美전역 거점화

NYT 클린턴 측근韓기업 특수관계 조명…공과 사 맞물려

日 오키나와서 열린 강제징용 노동자 추도식

실리콘밸리 넘어서겠다…구글 15조원 들여 미국전역 거점화

NYT 클린턴 측근한국기업 특수관계 조명…공과 사 맞물려

일본 오키나와서 열린 강제징용 노동자 추도식

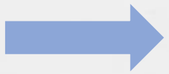
③ 함수 생성 - 토큰화 - 벡터화

<데이터 전처리를 위한 함수 생성>

3. 특수문자 제거

신중국70년 ①차이나 미라클…최빈국서 G2 경제대국 부상

현행 헌법과 다른 점은 ②지방자치 그리고 경제민주화 개념 강화



신중국70년 차이나 미라클 최빈국서 G2 경제대국 부상

현행 헌법과 다른 점은 지방자치 그리고 경제민주화 개념 강화

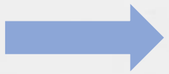
③ 함수 생성 - 토큰화 - 벡터화

<okt 패키지를 통한 토큰화>

인천에서 핀란드 항공기 결항 휴가철 여행객 분통

실리콘밸리 넘어서겠다 구글 15조원 들여 미국전역 거점화

이란 외무 긴장완화 해결책은 미국이 경제전쟁 멈추는 것



인천 핀란드 항공기 결항 휴가 철 여행객 분통

실리콘밸리 넘어서다 구글 15조원 들이다 미국 전역 거점 화

이란 외무 긴장 완화 해결 책 미국 경제 전쟁 멈추다 것

③ 함수 생성 - 토큰화 - 벡터화

<토큰화 결과>

okt.morphs → ['실리콘밸리', '넘어서다', '구글', '15조원', '들이다', '미국', '전역', '거점', '화']

okt.nouns → ['실리콘밸리', '구글', '미국', '전역', '거점', '화']

okt.pos → [('실리콘밸리', 'Noun'), ('넘어서다', 'Verb'), ('구글', 'Noun'), ('15조원', 'Number'), ('들이다', 'Verb'), ('미국', 'Noun'), ('전역', 'Noun'), ('거점', 'Noun'), ('화', 'Noun')]

③ 함수 생성 - 토큰화 - 벡터화

<tf-idf를 이용한 벡터화>

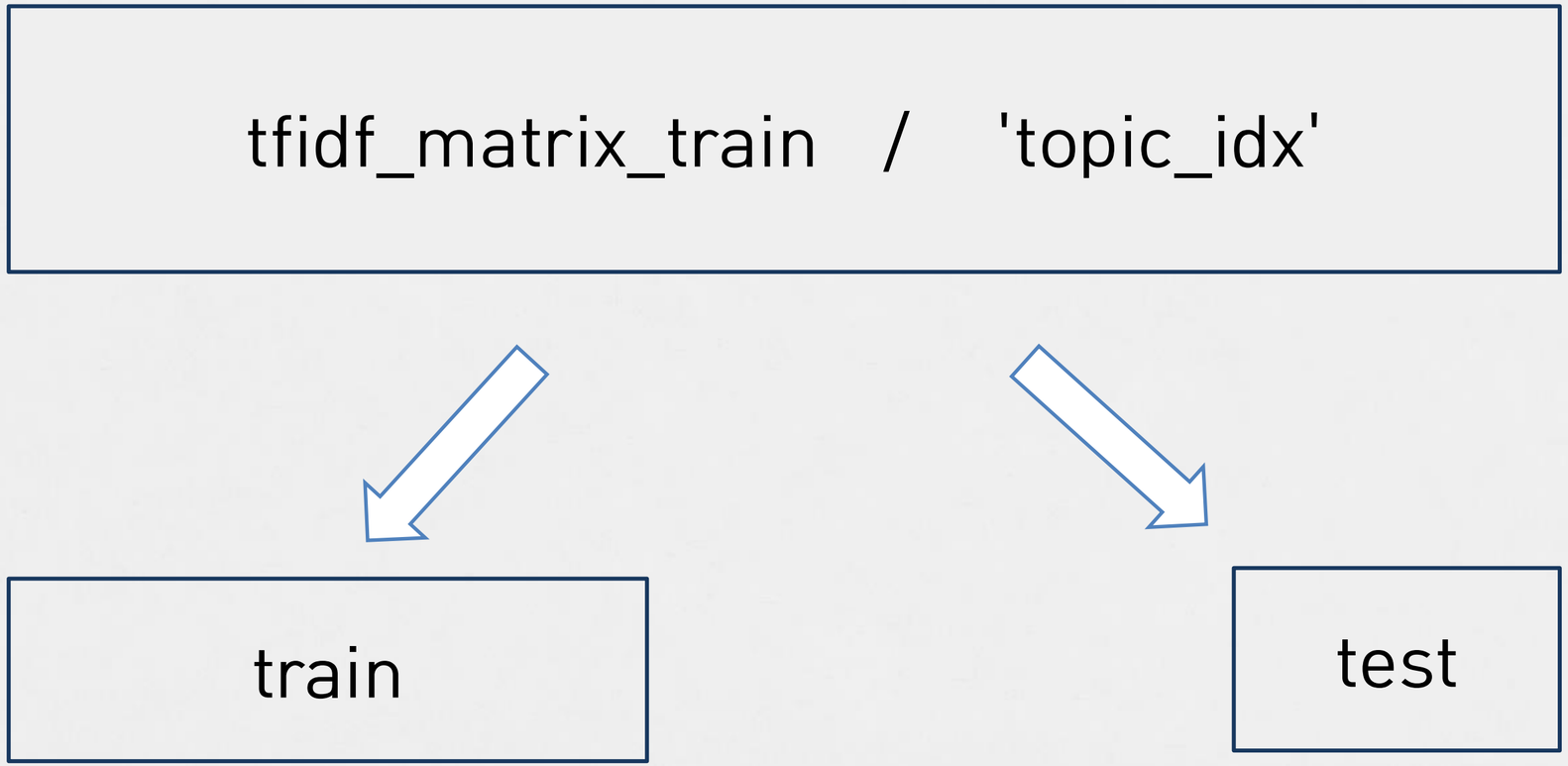
train['title']

(45654,)

tfidf_matrix_train

(45654, 30316)

④ train,test set 분리 - LGBM - Logistic Regresion



④ train,test set 분리 - LGBM - Logistic Regresion

1. 파라미터 튜닝

- max_depth=5
- min_data_in_leaf=20
- num_iterations=1500
- num_leaves=30
- random_state=42

2. 성능(accuracy)

0.815

④ train,test set 분리 - LGBM - Logistic Regresion

1. 파라미터 튜닝

- C = 6
- max_iter = 500
- penalty = ‘l2’
- multi_class = ‘ovr’

2. 성능(accuracy)

0.845

④ 최종 모델 선택

모델 종류	LGBM	Logistic Regresion
데이콘 결과	public점수 : 0.773 private점수 : 0.759	public점수 : 0.811 private점수 : 0.791
순위	2	1

RANDOM FOREST REGRESSOR 선택

④ 소감

이승준	스터디를 시작하기 전에는 자연어를 어떻게 모델에 적용할 수 있을 지 궁금했는데, 벡터화를 통해 가능했다. 형태소 분리, 명사 추출 등 이미 만들어진 패키지가 많다는 것을 알았다.
이가영	스터디를 통해 자연어처리의 전반적인 부분을 알 수 있어서 좋았다. 프로젝트를 진행하면서 자연어 처리과정에 다양한 방법이 존재한다는 것을 새롭게 알게 되었다.
김기호	스터디를 시작하기 전에는 자연어 처리 하면 한국어나 영어에 대한 방법만 떠올랐었는데 스터디를 하며 중국어, 일본어 등 다양한 언어들도 분석 할 수 있다는 것을 알았고, 프로젝트를 진행하면서 자연어를 처리해주는 다양한 패키지들의 장단점 및 사용방법을 알 수 있었다.

감사합니다