감성 분석을 위한 토크나이징 및 워드 임베딩 기법 탐색

20203065 소프트웨어학부 박규연

TABLE OF CONTENTS

01

Introduction

주제, 데이터 소개

감성 분석 모델 구축을 위한 토크나이징, 워드 임베딩 기법 탐색

토큰화 기법 탐색

Kkma, Komoran, Okt, Mecab, KoNLTK

Word2Vec, FastText

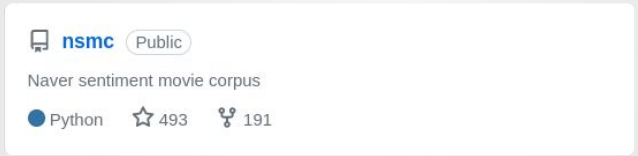
워드 임베딩 기법 탐색

모델 훈련

Logistic Regression, LSTM

NSMC Dataset

Naver sentiment movie corpus (20만 개)

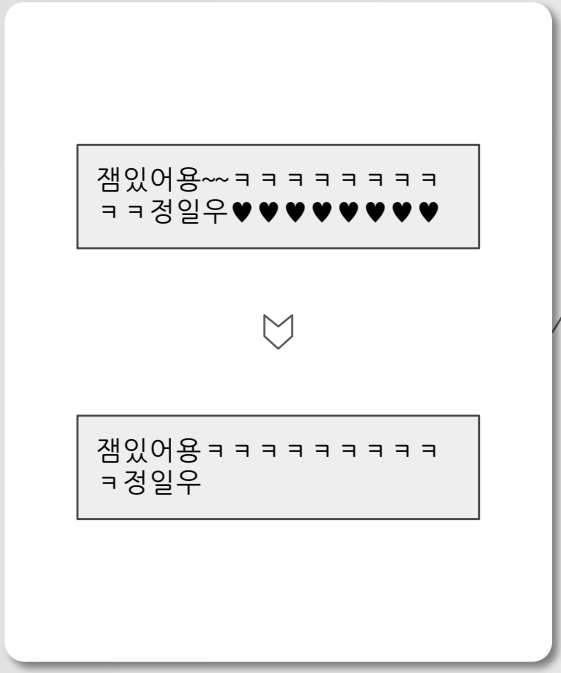


02

Preprocessing

데이터 전처리 및 토큰화

Preprocessing and Tokenizing



정규 표현식을 이용한 특수문자 제거

`pattern=r'[^a-zA-Z0-9가-힣 \n]'`

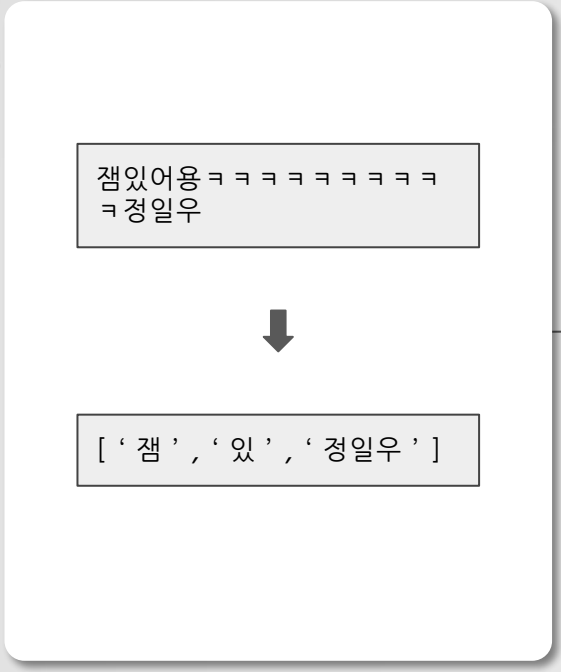
명사 토큰화

`tagger.nouns(text)`

불용어 제거

`if token not in stopwords`

Preprocessing and Tokenizing



정규 표현식을 이용한 특수문자 제거

```
pattern=r'[^a-zA-Z0-9가-힣 \n]'
```

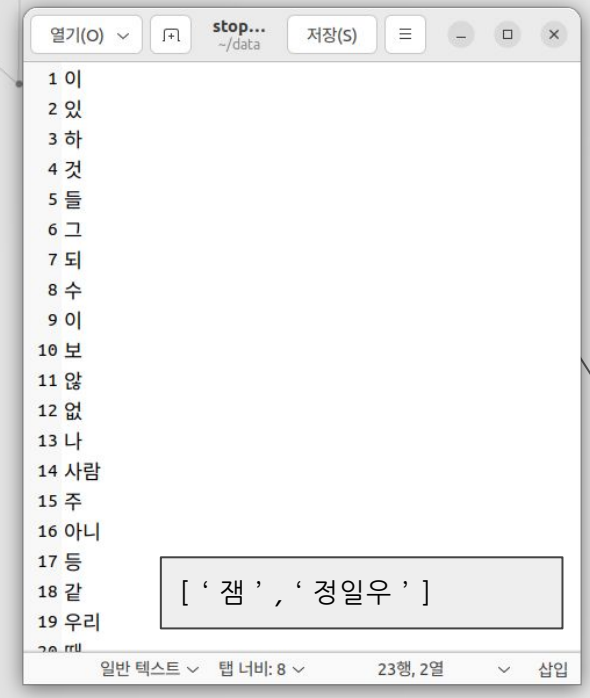
명사 토큰화

```
tagger.nouns(text)
```

불용어 제거

```
if token not in stopwords
```

Preprocessing and Tokenizing



정규 표현식을 이용한 특수문자 제거

```
pattern=r'[^a-zA-Z0-9가-힣 \n]'
```

명사 토큰화

```
tagger.nouns(text)
```

불용어 제거

```
if token not in stopwords
```

Preprocessed Dataframe

	id	document	label	tokens
0	8112052	어릴때보고 지금다시봐도 재밌어요ㅋㅋ	1	['보고']
1	8132799	디자인을 배우는 학생으로, 외국디자이너와 그들이 일군 전통을 통해 발전해가는 문화산...	1	['디자인', '학생', '외국', '디자이너', '일군', '전통', '통해', ...
2	4655635	폴리스스토리 시리즈는 1부터 뉴까지 버릴께 하나도 없음.. 최고.	1	['폴리스스토리', '시리즈', '부터', '뉴', '최고']
3	9251303	와.. 연기가 진짜 개쩔구나.. 지루할거라고 생각했는데 몰입해서 봤다.. 그래 이런...	1	['연기', '진짜', '몰입', '다그', '진짜', '영화']
4	10067386	안개 자욱한 밤하늘에 떠 있는 초승달 같은 영화.	1	['안개', '밤하늘', '초승달', '영화']
5	2190435	사랑을 해본사람이라면 처음부터 끝까지 웃을수 있는영화	1	['사랑', '라면', '처음', '끝', '영화']
6	9279041	완전 감동입니다 다시봐도 감동	1	['완전', '감동', '감동']
7	7865729	개들의 전쟁2 나오나요? 나오면 1빠로 보고 싶음	1	['전쟁', '빠']

03

Tokenizer Comparison

형태소 분석기 비교

Accuracy Comparison

	‘내 나이와 같은 영화를 지금 본 나는 감동적이다..하지만 훗날 다시보면대사하나하나 그 감정을완벽하게 이해할것만 같다… ’
Kkma	['내', '나이', '영화', '나', '감동', '훗날', '보면대', '사', '하나', '하나', '감정', '완벽', '이해', '것']
Komoran	['나이', '영화', '감동', '하지', '훗날', '다시', '보면', '대사', '하나하나', '감정', '을', '완벽', '이해', '것']
Okt	['내', '나이', '영화', '나', '감동적', '훗날', '대사', '대사하나하나', '하나하나', '감정', '완벽', '이해']
Mecab	['내', '나이', '영화', '지금', '나', '감동', '훗날', '다시', '사하나', '그', '감정', '이해']
KoNLTK	['나', '나이', '영화', '나', '감동적', '훗날', '다시보면대사하나', '감정을완벽', '이해할것']

Kkma, Komoran, Okt

띄어쓰기가 잘 되어있는 리뷰에서 괜찮은 성능을 보임임

Accuracy Comparison

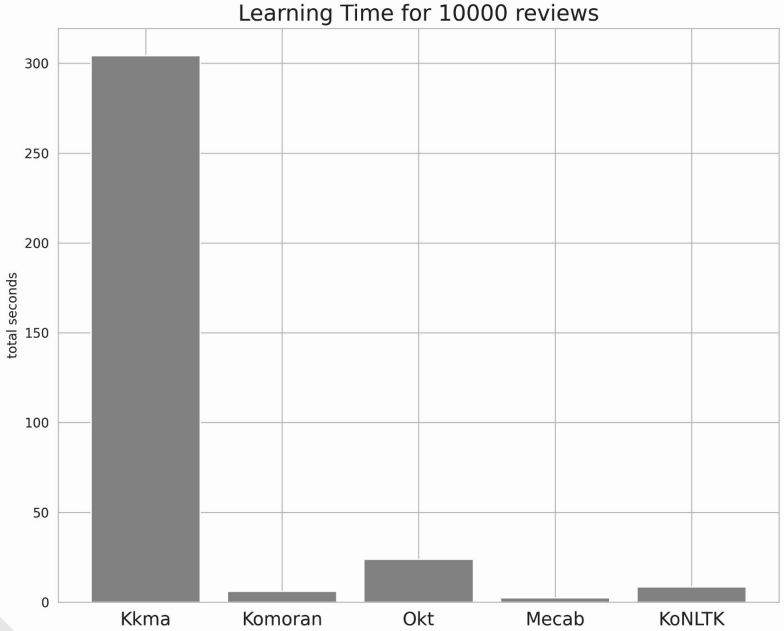
	'이거어렸을때되게재밌게봄ㅋㅋ이정재 이범수ㅋㅋㅋㅋ연기쩜'
Kkma	['이거', '때', '이정재', '이범수', '연기']
Komoran	[]
Okt	['이거', '때', '봄', '이정재', '이', '이범수', '범수', '연기']
Mecab	['거', '때', '봄', '이정재', '이범수', '연기', '쩜']
KoNLTK	['이거어렸을때되게재밌게봄', '이정', '이범수', '연기쩜']

Kkma가 띄어쓰기가 잘 안 되어 있는 리뷰에서도 유의미한 토큰들을 잘 뽑아냄
Komoran은 띄어쓰기가 잘 안되어 있는 리뷰에서는 성능이 떨어짐

Speed Comparison

Speed

Mecab이 속도 측면에서 매우 우수
Kkma는 속도 측면에서 매우 떨어짐



Tokenizer Comparison

	Kkma	Komoran	Okt	Mecab	KoNLTK
Accuracy	○	X	○	△	△
Speed	X	○	△	○	○

정확도 측면에서 가장 우수한 **Okt**
속도 측면에서 가장 우수한 **Mecab**
을 **선정**하여 이후 실험 진행

04

Word Embedding

Word2Vec와 FastText 비교

Most Similar Top 5

“ 배우 ”

	Word2Vec	FastText
Okt	‘ 연기자 ’ , ‘ 영화배우 ’ , ‘ 김혜수 ’ ‘ 양동근 ’ , ‘ 아역배우 ’	‘ 배우진 ’ , ‘ 단역배우 ’ , ‘ 재연배우 ’ ‘ 영화배우 ’ , ‘ 유명배우 ’
Mecab	‘ 연기자 ’ , ‘ 조연 ’ , ‘ 차승원 ’ ‘ 신인 ’ , ‘ 손예진 ’	‘ 파배우 ’ , ‘ 남배우 ’ , ‘ 명배우 ’ ‘ 영화배우 ’ , ‘ 여배우 ’

Most Similar Top 5

positive=[“ 감동 ” , “ 마지막 ”]

	Word2Vec	FastText
Okt	‘ 핑 ’ , ‘ 콧 ’ , ‘ 피아노 ’ ‘ 글썽 ’ , ’ 주룩주룩 ’	‘ 마지막모습 ’ , ’ 마지막방송 ’ , ’ 핑 ’ ‘ 글썽 ’ , ’ 방울 ’
Mecab	‘ 콧물 ’ , ’ 방울 ’ , ’ 흘 ’ ‘ 물결 ’ , ’ 하이라이트 ’	‘ 흘 ’ , ’ 콧물 ’ , ’ 방울 ’ ‘ 감동이 ’ , ’ 핑 ’

Most Similar Top 5

positive=[“ 황정민 ” , “ 최우식 ”]
negative=[“ 별로 ”]

	Word2Vec	FastText
Okt	‘ 김영호 ’ , ’ 경력 ’ , ’ 김갑수 ’ ‘ 백성현 ’ , ’ 이성민 ’	‘ 상우 ’ , ’ 백성현 ’ , ’ 손현주 ’ ‘ 김영호 ’ , ’ 정려원 ’
Mecab	‘ 차승원 ’ , ’ 디카프리오 ’ , ’ 엄정화 ’ ‘ 손예진 ’ , ’ 임창정 ’	‘ 김혜수 ’ , ’ 차승원 ’ , ’ 안성기 ’ ‘ 엄정화 ’ , ’ 신인 ’

05

Keyword Analysis

워드카운트 / 워드클라우드 / 임베딩벡터 시각화

Word Count

	count.most_common(8)
Okt	[('영화', 57883), ('정말', 11748), ('진짜', 10643), ('평점', 8394), ('최고', 7805), ('연기', 7047), ('스토리', 6734), ('이영화', 6516)]
Mecab	[('영화', 75236), ('연기', 8901), ('최고', 8586), ('평점', 8238), ('스토리', 7020), ('드라마', 6703), ('감동', 6309), ('배우', 5643)]

*Okt 사용

Word Cloud

Positive

Negative

Word Cloud

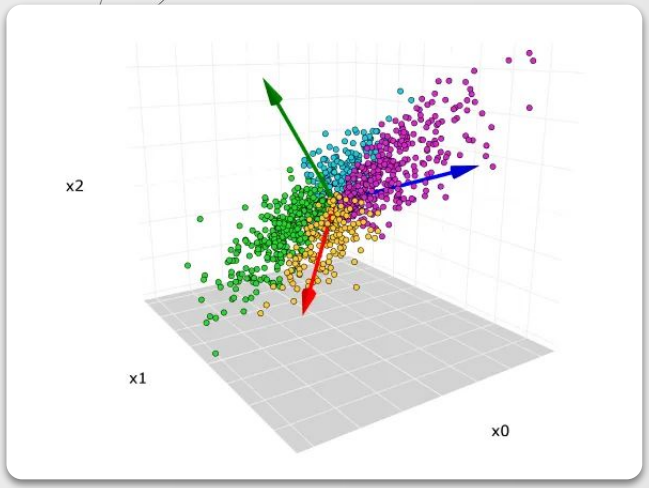
Positive

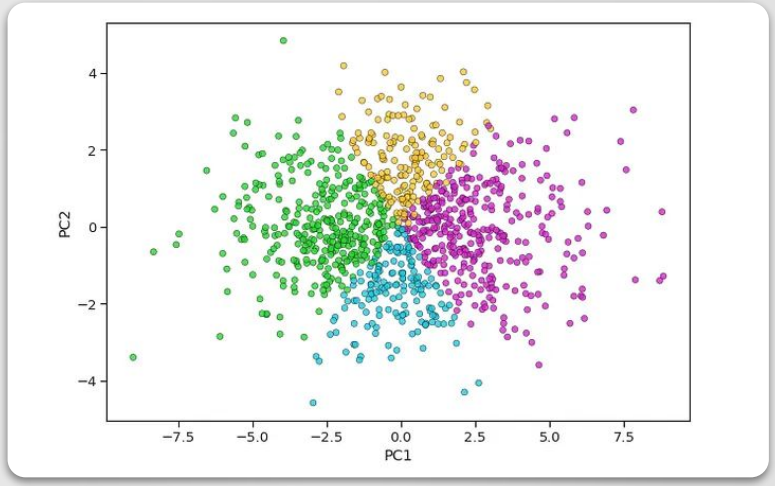
Negative

지배적인 단어,
키워드 분석에 의미 없는 단어 제외

PCA(주성분 분석)

고차원의 데이터를 저차원의 데이터로 환원시키는 기법



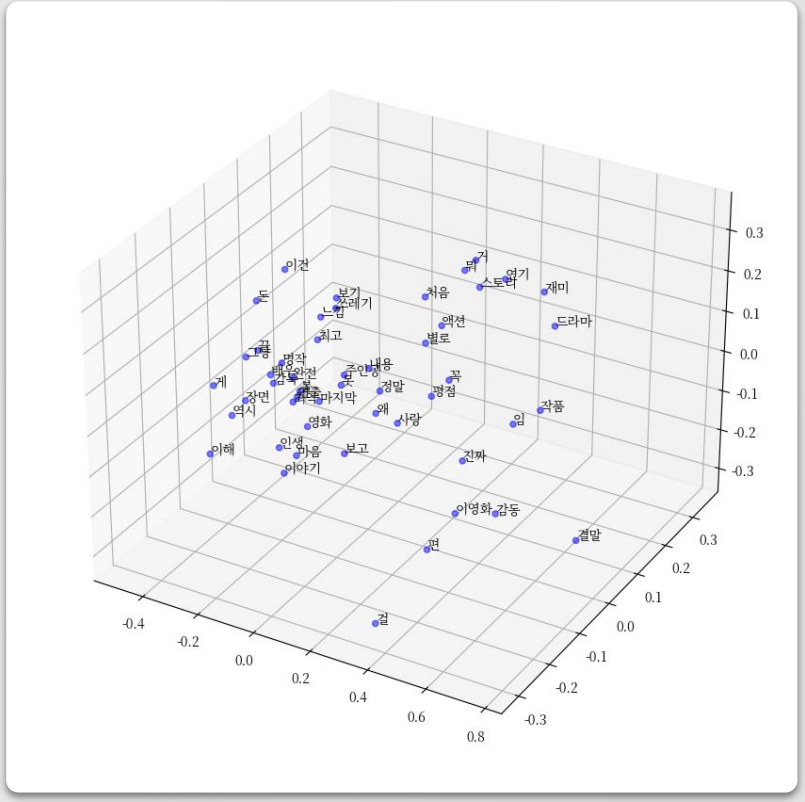


https://towardsdatascience.com/principal-component-analysis-pca-explained-visually-with-zero-math-1cbf392b9e7d

3D Visualization

Most Similar Top 100

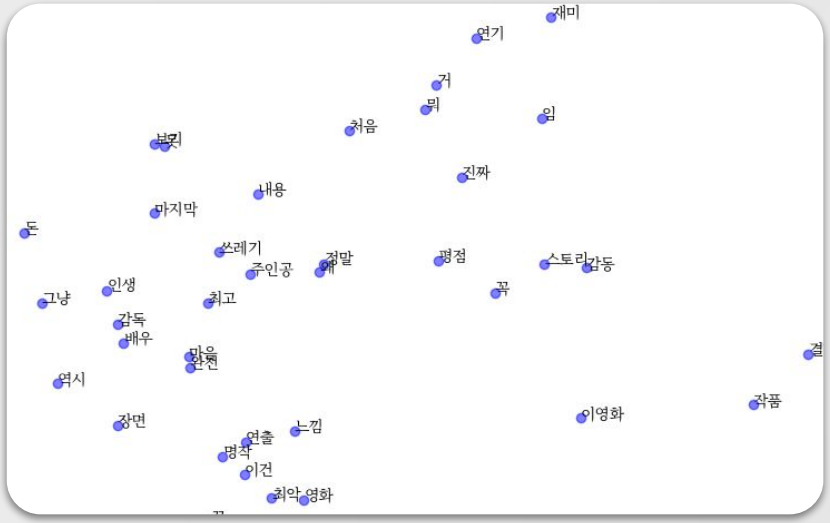
positive=[“ 황정민 ” , “ 마동석 ”]



2D Visualization

Most Similar Top 100

positive=[“ 황정민 ” , “ 마동석 ”]

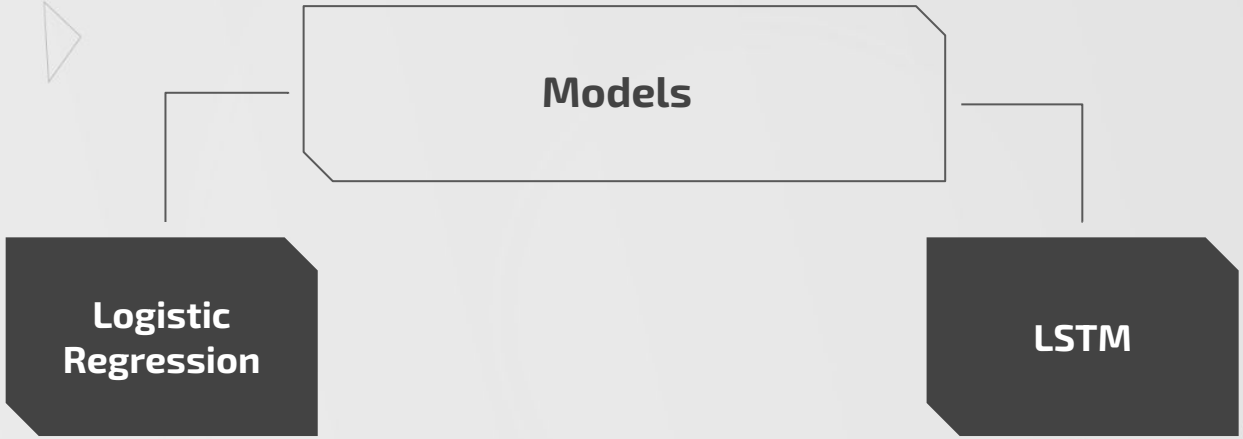


06

Classification Model

감성 분류 모델 훈련 및 비교

Modeling



Embedding

Word2Vec
(CBOW)

Word2Vec
(Skip-gram)

FastText

```
[[1, 1, 1, 1, 1, 1, 1, 1, 1, 1],
 [1, 2, 1, 3, 1, 1, 1, 1, 1, 1],
                ...
 [0, 1, 2, 1, 3, 2, 1, 3, 2, 1]]
```



```
[1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
```

단어 벡터들을 평균내서 문장 벡터 생성

Logistic Regression

	Word2Vec (CBOW)	Word2Vec (Skip-gram)	FastText
Okt	69.61%	72.00%	69.45%
Mecab	69.48%	71.00%	69.04%

LSTM

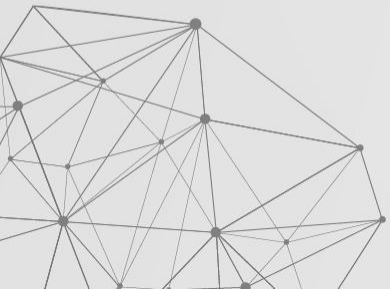
	Word2Vec (CBOW)	Word2Vec (Skip-gram)	FastText
Okt	63.28%	72.67%	70.45%
Mecab	69.58%	71.08%	70.00%

Conclusion

Okt — Tokenizer

Word2Vec (Skip-gram) — Word Embedding

LSTM — Classification Model



Process

NSMC Data
Naver Sentiment Movie Corpus

Okt
Tokenizer

Mecab
Tokenizer

Word2Vec
Word Embedding

FastText
Word Embedding

Word2Vec
Word Embedding

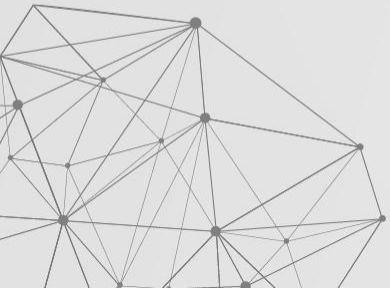
FastText
Word Embedding

LR
LSTM

LR
LSTM

LR
LSTM

LR
LSTM



References

- ChatGPT
- Word2vec 및 fastText 임베딩 모델의 성능 비교, 2020, 강현석 외 1명, 디지털콘텐츠학회논문지
- 감성 분류를 위한 워드 임베딩 성능 비교, 2021, 윤혜진 외 2명, ACK 2021 학술발표대회 논문집

https://github.com/noooey/Exploring-for-Sentiment-Analysis

THANKS

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**.

Please keep this slide for attribution.