NLP와
NLP의 활용

이 한 재 임 혜 원 정 노 아

INDEX

01

NLP

❶ NLP와 활용
❷ 토큰화
❸ 불용어
❹ 인코딩
❺ 패딩

01 NLP와 활용

NLP란?

National Language Processing의 약자로 자연어 처리 혹은 자연 언어 처리
컴퓨터가 인간의 언어를 이해하고 해석하여 처리할 수 있도록 하는 일

NLP 활용

텍스트 분류 - 스팸 메일 분류기, 뉴스 분류기
감성 분석 - 상품, 영화 등의 리뷰 분석
기계 번역 - 번역기

02 토큰화

토큰화란?

토큰은 보통 명사나 형태소와 같이 의미있는 단위로 정의
데이터를 토큰이라 불리는 단위로 나누는 작업

토큰화 예시(명사)

"나는 1학기 시간표를 만들고 있다." → '나' , '학기' , '시간표'

02 토큰화

형태소 분석기

- Okt (Open Korea Text)
- Hannanum
- Kkma
- Mecab
- Komoran

형태소 분석기 함수

종류	공통	추가
Okt	pos (품사 태깅) nouns (명사 추출) morphs (형태소 추출)	normalize (정규화) phrases (구 추출)
Hannanum		sentences (문장 추출)
Kkma		
Mecab		
Komoran		

02 토큰화

형태소 분석기 별 분석 차이 (품사 태깅)

"나는 1학기 시간표를 만들고 있다."				
Okt	Kkma	Mecab	Komoran	Hannanum
('나', 'Noun') ('는', 'Josa') ('1', 'Number') ('학기', 'Noun') ('시간표', 'Noun') ('를', 'Josa') ('만들고', 'Verb') ('있다', 'Adjective') ('.', 'Punctuation')	('나', 'NP') ('는', 'JX') ('1', 'NR') ('학기', 'NNG') ('시간표', 'NNG') ('를', 'JKO') ('만들', 'VV') ('고', 'ECE') ('있', 'VXV') ('다', 'EFN') ('.', 'SF')	('나', 'NP') ('는', 'JX') ('1', 'SN') ('학기', 'NNG') ('시간표', 'NNG') ('를', 'JKO') ('만들', 'VV') ('고', 'EC') ('있', 'VX') ('다', 'EF') ('.', 'SF')	('나', 'NP') ('는', 'JX') ('1', 'SN') ('학기', 'NNG') ('시간표', 'NNG') ('를', 'JKO') ('만들', 'VV') ('고', 'EC') ('있', 'VX') ('다', 'EF') ('.', 'SF')	('나', 'N') ('는', 'J') ('1학기', 'N') ('시간표', 'N') ('를', 'J') ('만들', 'P') ('고', 'E') ('있', 'P') ('다', 'E') ('.', 'S')

03 불용어

불용어 제거

분석에서 큰 의미가 없는 단어(형태소)를 제거하는 과정

불용어 처리 예시

	"인터넷 검색 시 검색 용어로 사용하지 않는 단어. 관사, 전치사, 조사, 접속사 등 검색 색인 단어로 의미가 없는 단어이다. 다만 각 검색 엔진마다 동일하지 않기 때문에 다를 수도 있다."
불용어	" 검색 시 로 등 않는 가 없는 이다 다만 각 마다 하지 않기 때문 에 수도 있다 . ,"
불용어 처리 전	['인터넷', '검색', '시', '검색', '용어', '로', '사용', '하지', '않는', '단어', '.', '관사', ',', '전치사', ',', '조사', ',', '접속사', '등', '검색', '색인', '단어', '로', '의미', '가', '없는', '단어', '이다', '.', '다만', '각', '검색', '엔진', '마다', '동일하지', '않기', '때문', '에', '다를', '수도', '있다', '.']
불용어 처리 후	['인터넷', '용어', '사용', '단어', '관사', '전치사', '조사', '접속사', '색인', '단어', '의미', '단어', '엔진', '동일하지', '다를']

04 인코딩

정수 인코딩

텍스트를 숫자로 바꾸는 방법 중 하나로 빈도수가 높은 순서대로 낮은 숫자부터 단어에 정수를 부여하는 방법

정수 인코딩 예시

"인터넷 검색 시 검색 용어로 사용하지 않는 단어. 관사, 전치사, 조사, 접속사 등
검색 색인 단어로 의미가 없는 단어이다. 다만 각 검색 엔진마다 동일하지 않기 때문에 다를 수도 있다."

↓

['인터넷', '검색', '시', '검색', '용어', '사용', '단어', '관사', '전치사', '조사', '접속사',
'등', '검색', '색인', '단어', '의미', '단어', '다만', '각', '검색', '엔진', '때문']

↓

[[3], [1], [4], [1], [5], [6], [2], [7], [8], [9], [10], [11], [1], [12], [2], [13], [2], [14], [15], [1], [16], [17]]

05 패딩

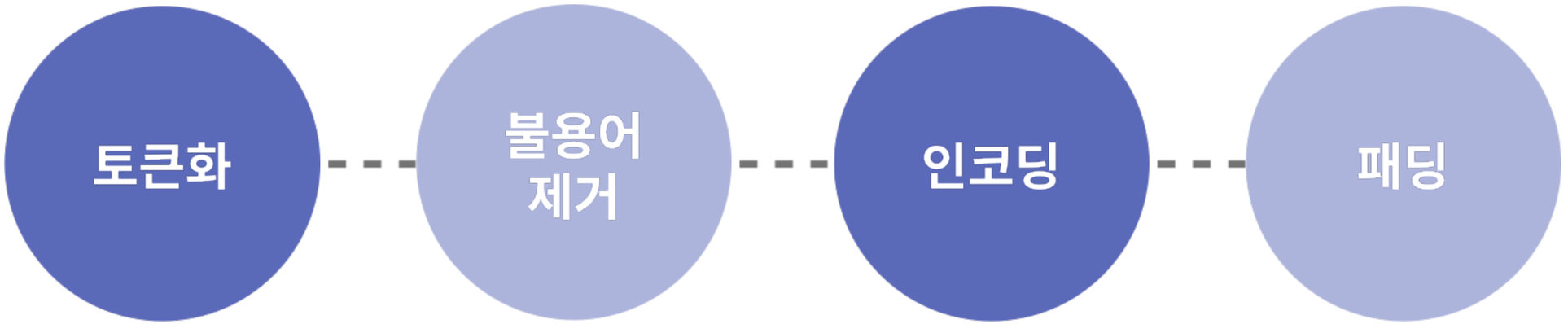
패딩

병렬 연산을 위해 여러 문장들의 길이를 임의로 동일하게 맞추는 작업

패딩 예시

[1,3,4]	[1,3,4,0,0,0,0]
[2,1,5,6,1]	[2,1,5,6,1,0,0]
[2,7,1,2]	[2,7,1,2,0,0,0]
[1,3,8,9,10]	[1,3,8,9,10,0,0]
[2,11,12,13,1]	[2,11,12,13,1,0,0]
[1,2,1,3,14,15,16]	[1,2,1,3,14,15,16]

06 정리



02

모델 활용

1) 프로젝트 및 데이터 소개
2) EDA
3) 텍스트 전처리(Text Preprocessing)
4) 모델링
5) 결론

01

프로젝트

프로젝트 및 데이터 소개

자연어 기반 기후기술분류

국가 연구 개발 과제를
‘기후기술분류체계’에 맞추어 라벨링하는 알고리즘 개발

01

데이터

프로젝트 및 데이터 소개

train data : 174304 obs , 13 var

index	제출년도	사업명	…	과제명	요약문_연구목표	…	label
0	2016	농업기초기반연구		유전정보를 활용한 새로운 해충 분류군 동정기술 개발	○ 새로운 해충분류군의 동정기술 개발 및 유입확산 추적		24
1	2019	이공학학술연구기반구축(R&D)		대장암의 TRAIL 내성 표적 인자 발굴 및 TRAIL 반응 예측 유전자 지도 구축…	최종목표: TRAIL 감수성 표적 유전자를 발굴하고 내성제어 기전을 연구. 발굴된…		0

target 변수
(0~45)

NA 개수
요약문_연구목표 3002
요약문_연구내용 3001
요약문_기대효과 3051
요약문_한글키워드 3028
요약문_영문키워드 3087

➢ 요약문 column에 결측치 존재

01

데이터

프로젝트 및 데이터 소개

test data : 43576 obs , 12 var (label 미포함)

index	제출년도	사업명	…	과제명	요약문_연구목표	…
174304	2016	경제협력권산업육성		R-FSSW 기술 적용 경량 차체 부품 개발 및 품질 평가를 위한 64채널 C-SC…	○ 차체 점용접부의 품질 검사를 위한 64채널 무선 기반 C-Scan 탐촉자 개발₩…	
174305	2018	개인기초연구(과기정통부)(R&D		다입자계를 묘사하는 편미분방정식에 대한 연구	자연계에는 입자의 개수가 아주 큰 다양한 다입자계가 존재한다. 이런 다입자계의 효…	

NA 개수

요약문_연구목표 755
요약문_연구내용 755
요약문_기대효과 761
요약문_한글키워드 760
요약문_영문키워드 772

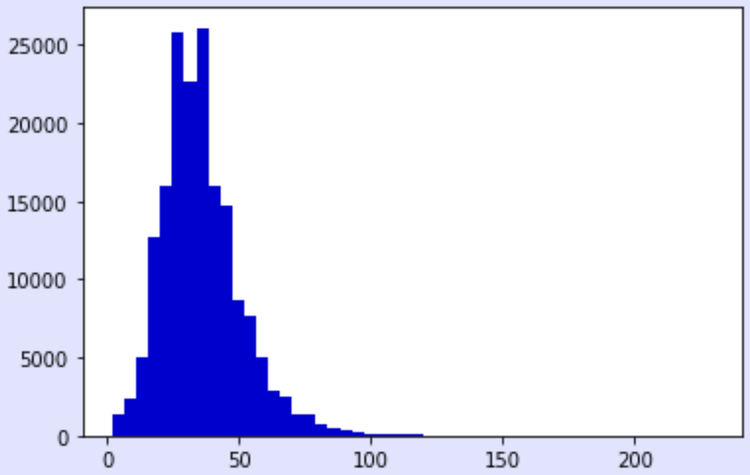
➢ 요약문 column에 결측치 존재

02

과제명

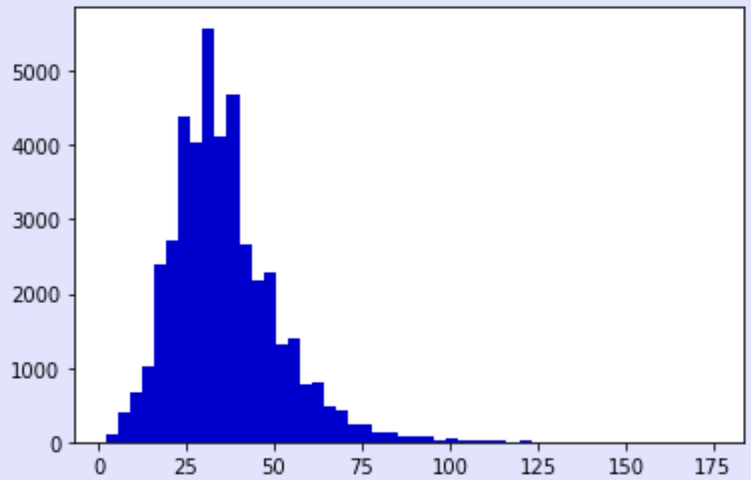
EDA

과제명 길이



train

최대값 : 229
최소값 : 2
평균값 : 35.8425
중간값 : 34.0



test

최대값 : 175
최소값 : 2
평균값 : 35.9448
중간값 : 34.0

02

과제명

EDA

과제명 길이

길이가 짧은 과제명

train

len = 2	len = 3	len = 4
9407 아리	1534 테스트	3839 총괄과제
110338 시예	5879 EMF	4988 양자제어
…	9147 인건비	6625 학사사업
	…	…

test

len = 2	len = 3	len = 4
15542 아리	4725 시설비	12683 양산제품
	9787 인건비	15042 CCRC
	13994 라오닐	15885 본부과제
	…	…

02

과제명

EDA

과제명 중복

train

보안과제정보	716
사회맞춤형산학협력선도대학(LINC+)육성(0.5)	185
해상부유식 LNG 벙커링 시스템 기술개발	98
한국형 e-Navigation 서비스를 위한 핵심기술 연구개발	80
산학협력선도대학(LINC)육성(0.5)	49
...	...

test

보안과제정보	174
사회맞춤형산학협력선도대학(LINC+)육성(0.5)	46
한국형 e-Navigation 서비스를 위한 핵심기술 연구개발	20
해상부유식 LNG 벙커링 시스템 기술개발	15
대학 창의적 자산 실용화 지원(BRIDGE+)사업	14
...	...

02

과제명

EDA

과제명 중복

과제명이 '보안과제정보' 일 때 label

label	count
0 (NAN)	711
10 (바이오에너지)	2
40 (생태 모니터링 복원)	1
13 (연료전지)	1
8 (풍력)	1

➢ 동일한 과제명임에도 label 다름

label이 0(NAN)이 아닌 경우

index	사업명	사업_부처명	내역사업명	과제명	요약문_연구목표	요약문_연구내용	요약문_기대효과	…	label
3396	기후변화대응기술개발(R&D)	과학기술정보통신부	보안과제정보	보안과제정보	보안과제정보	보안과제정보	보안과제정보	…	10
68725	개인기초연구(교육부)(R&D)	교육부	보안과제정보	보안과제정보	보안과제정보	보안과제정보	보안과제정보	…	13
78988	개인기초연구(과기정통부)(R&D)	과학기술정보통신부	보안과제정보	보안과제정보	보안과제정보	보안과제정보	보안과제정보	…	8
99892	개인기초연구(교육부)(R&D)	교육부	보안과제정보	보안과제정보	보안과제정보	보안과제정보	보안과제정보	…	10
103017	생물자원발굴및분류연구(R&D)	환경부	보안과제정보	보안과제정보	보안과제정보	보안과제정보	보안과제정보	…	40

잘못 mapping -> obs 삭제

03

토큰화

텍스트 전처리

한글, 영어, 공백 제외한 모든 문자 제거

Kkma

- 문자의 개수가 클수록 처리속도 느려짐
- 띄어쓰기 오류에 덜 민감

Okt

- 어느정도 띄어쓰기가 되어있는 경우 사용
- norm(정규화) stem(어간추출) 기능

Mecab

- 처리 속도가 가장 빠름

03

불용어

텍스트 전처리

stop_words.csv 에서 불용어

불용어 추가작업

✓ 조사, 접미사, 접속사 모두 제거

ex. '에는', '에서의', '은', '과' … '력', '억', '개', '번째'…

✓ 의미없는 단어 제거

ex. 통한, 되다 …

✓ 올바르게 분류되지 않은 단어 제거

ex. 위한 : 위(명사)+한(동사) -> '위' 제거

✓ 자주 등장하는 단어 제거

ex. 개발, 기술, 연구 …

03

정수 인코딩

텍스트 전처리

['유전 정보 활용 새롭다 해충 분류군 동정',
'대장암 trail…' …]

{ … '나노': 25, '환경': 26, '세포': 27, '생산': 28, '스마트': 29,… }

[359, 43, 6, 103, 653, 8034, 1244]
[610, 5368, 24, 64, 127, 177, 57, 5368, 77, 59, 41, 807, 11, 46] …

STEP 01

토큰화된 말뭉치를 리스트 형태로 반환

STEP 02

단어-숫자(Key-value) 딕셔너리 생성

가장 많이 등장하는 토큰 1로 배정

STEP 03

토큰화된 각 과제명을 숫자 시퀀스 형태로 반환

03

패딩

텍스트 전처리

병렬 연산을 위해 길이를 동일하게 만들어준다

pre : 앞에 0을 추가하여 패딩

[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 359 43 6 103 653 8034 1244]
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 610 5368 24 64 127 177 57 5368 77 59 41 807 11 46] …

post : 뒤에 0을 추가하여 패딩

[359 43 6 103 653 8034 1244 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
[610 5368 24 64 127 177 57 5368 77 59 41 807 11 46 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0] …

LSTM과 같은 recurrent model에서는
pre-padding 성능이 더 높다

04

Parameter

모델링

vocab_size
38912

총 단어 개수
: 단어 - 숫자 쌍
딕셔너리 길이

embedding_dim
32

임베딩 벡터 크기

max_length
40

사용할 텍스트 길이

oov_tok
“<OOV>”

토큰화 되어있지 않은 단어는
<OOV>로 인덱싱

04

모델

모델링

Layer	Output shape	
Embedding	**(None, 40, 32)**	임베딩 레이어 단어를 매핑할 수 있도록 벡터화
Global Average Pooling	**(None, 32)**	layer 평균을 구하여 (vocab_size, embedding_dim) 배열 생성 *overfitting 막는 효과
Dense (128) - relu	**(None, 128)**	활성화 함수 relu - 은닉층 입력이 0을 넘으면 그대로 출력. 0 이하면 0을 출력
Dropout(0.2)		학습시 overfitting 을 막기 위해 추가 신경망의 일부를 사용하지 않는 방법
Dense (46) - softmax	**(None, 46)**	활성화 함수 softmax 다중 클래스 분류문제에서 출력층에 사용

04

모델

모델링

Layer	Output shape	
Embedding	(None, 40, 32)	임베딩 레이어 단어를 매핑할 수 있도록 벡터화
LSTM(64)	(None, 64)	데이터가 길어도 초기 정보가 사라지지 않음 긴 시퀀스를 기억할 수 있음
Dense (128) - relu	(None, 128)	활성화 함수 relu - 은닉층 입력이 0을 넘으면 그대로 출력. 0 이하면 0을 출력
Dropout(0.2)		학습시 overfitting 을 막기 위해 추가 신경망의 일부를 사용하지 않는 방법
Dense (46) - softmax	(None, 46)	활성화 함수 softmax 다중 클래스 분류문제에서 출력층에 사용

04

Compile & Fit

모델링

Compile

✓ loss = ‘sparse_categorical_crossentropy’
다중 클래스 분류 , 출력층 활성화 함수 softmax 인 경우

✓ optimizer = ‘adam’
방향과 스텝사이즈 고려하는 알고리즘

✓ metric = ‘accuracy’
정확성을 지표로 학습

Fit

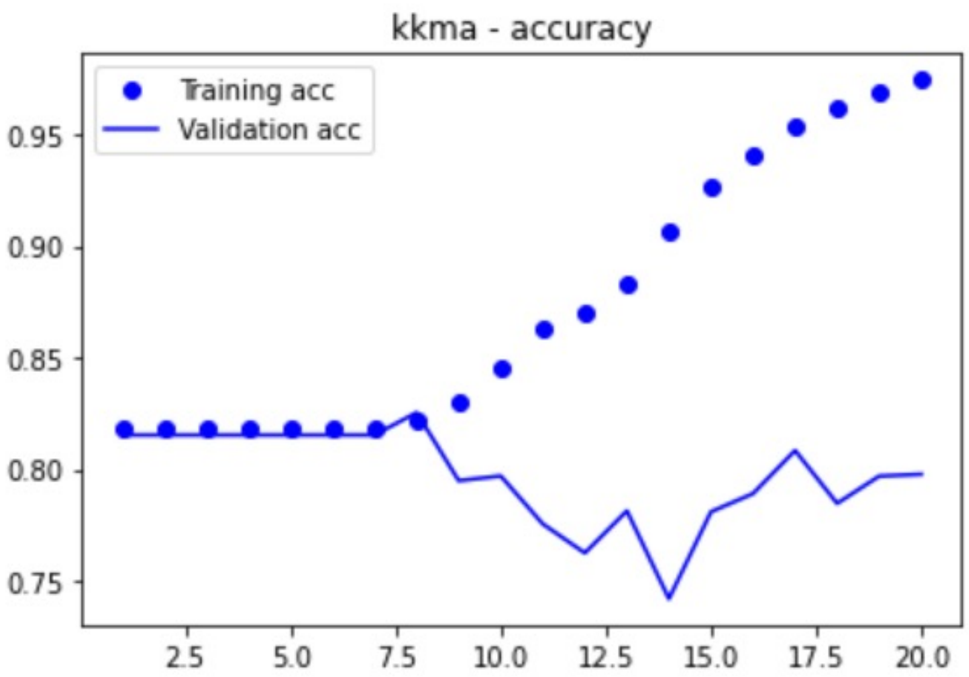
✓ epochs = 20
epoch 만큼 모델 학습
overfitting을 막기 위해 적절하게 튜닝

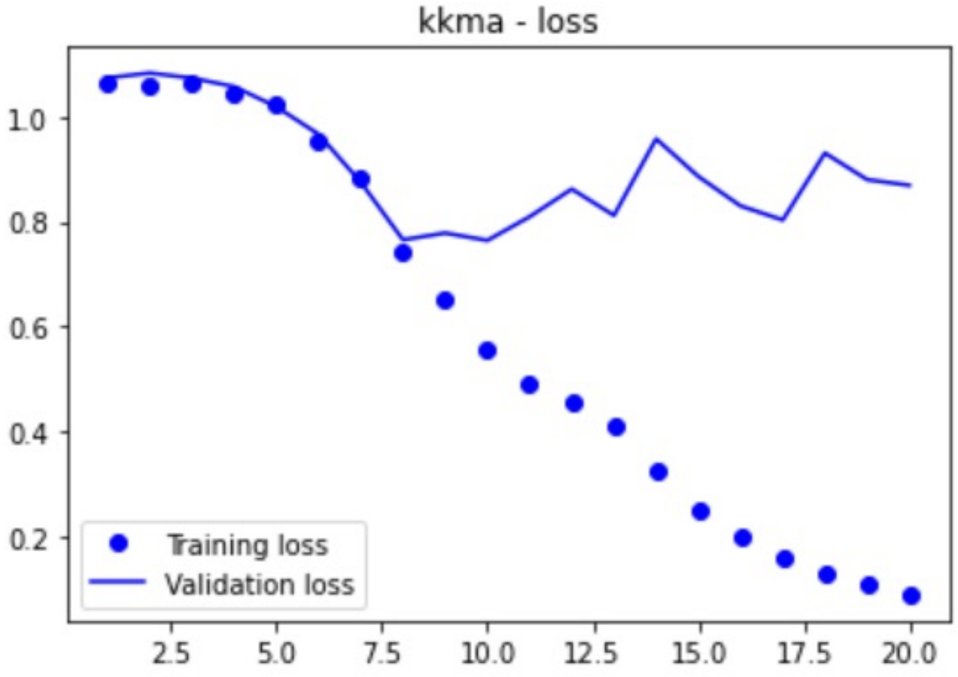
✓ validation split = 0.2
8:2 비율로 나누어 학습

04

Compile & Fit

모델링

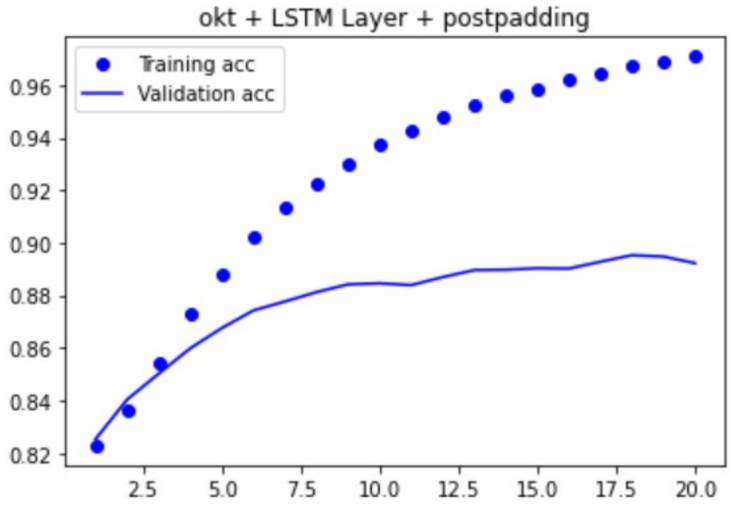


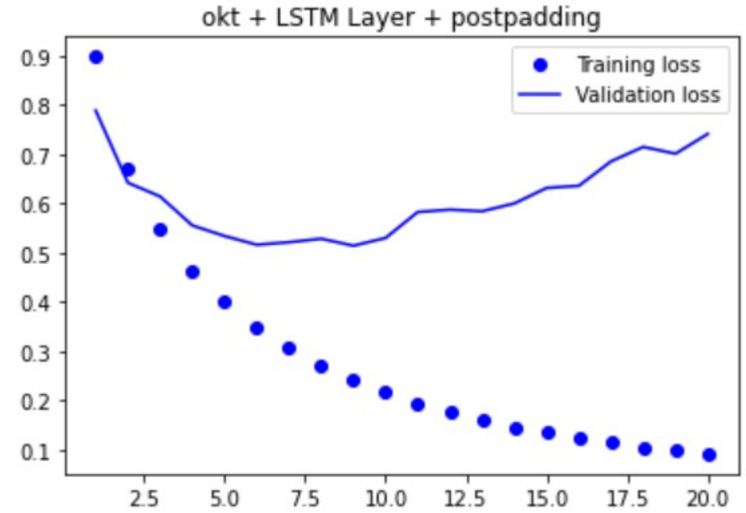


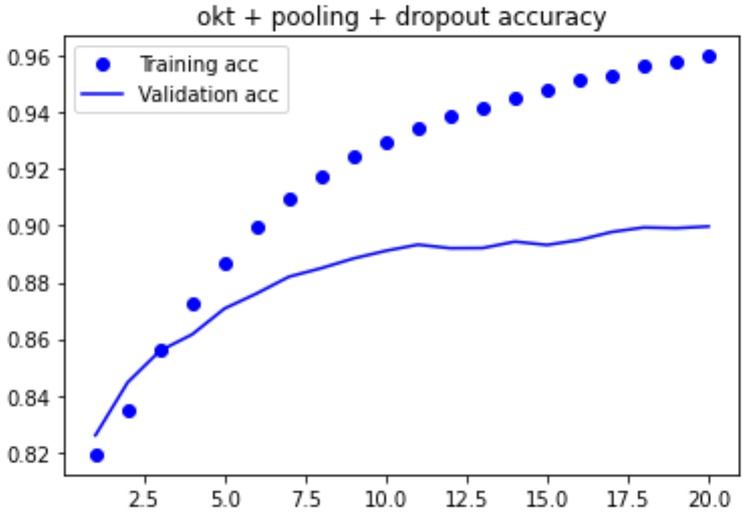
04

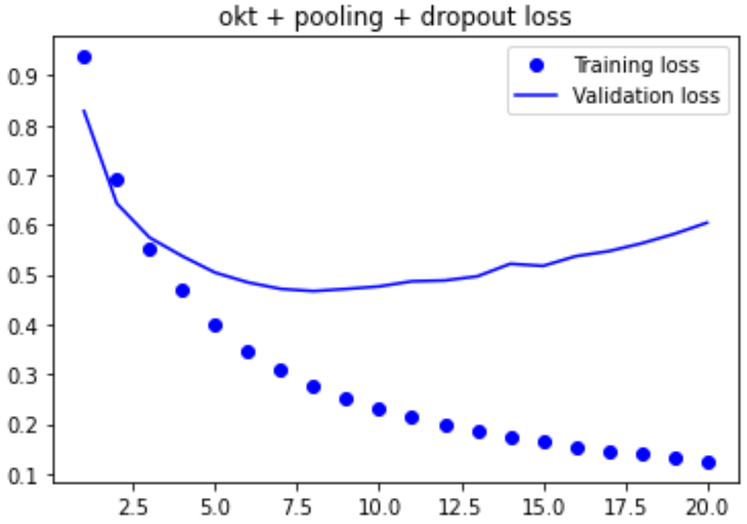
Compile & Fit

모델링





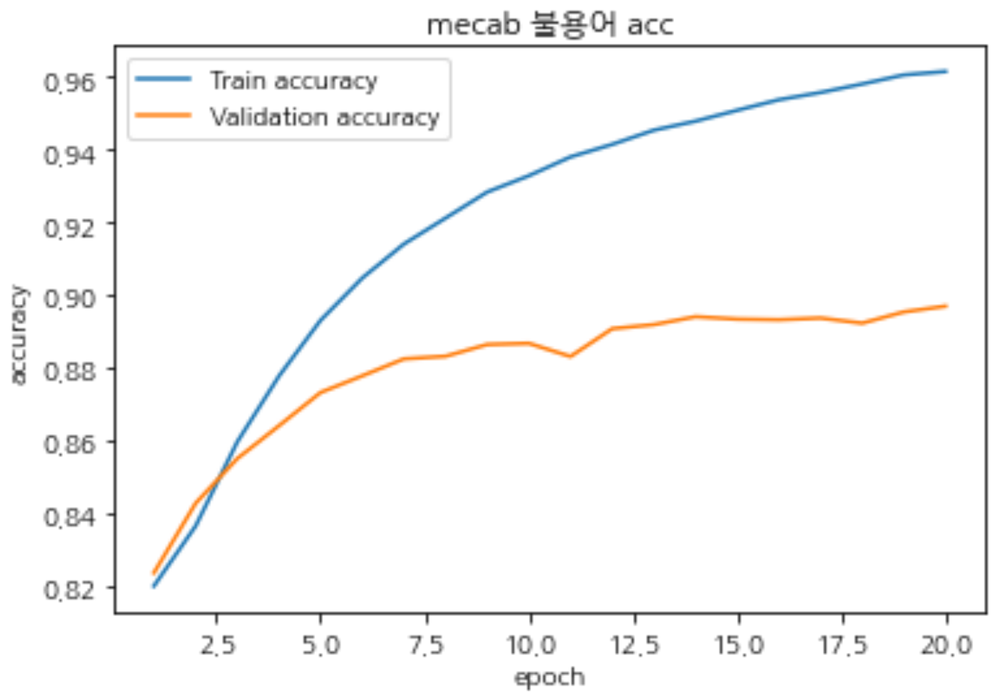


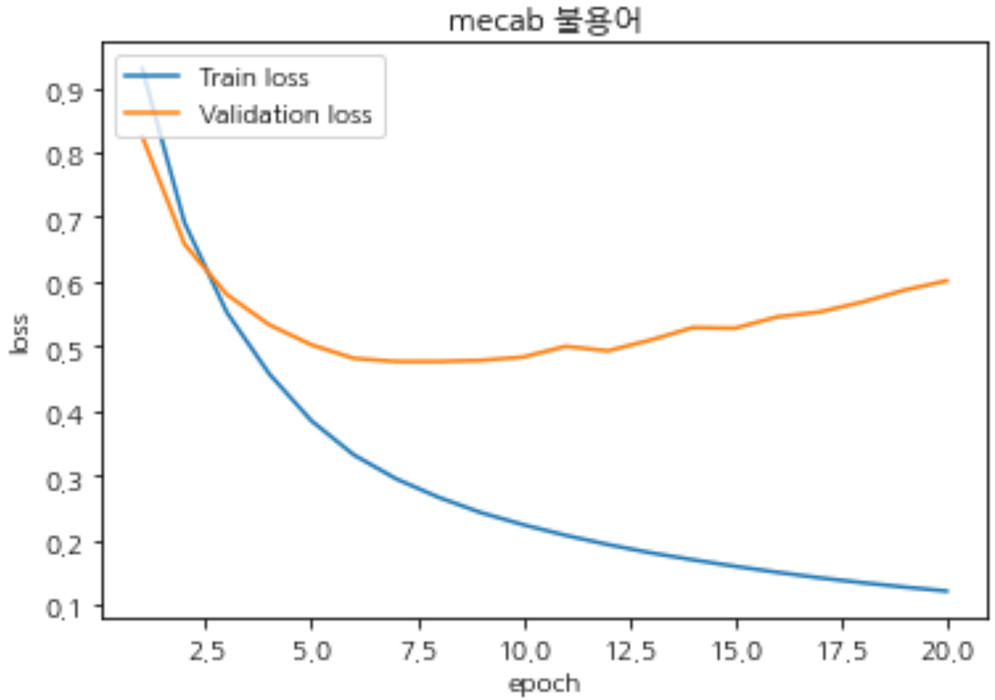


04

Compile & Fit

모델링





05

결론 및 한계점

결론

결론

○ Mecab 속도와 정확성 측면에서 가장 좋음

○ 모델 Global Average Pooling layer

과적합을 막는데 효과적

한계점

○ 정확성이 높지 않음

원인 1) 과제명 변수만을 사용

요약문 텍스트길이가 길어 코드 실행 오래 걸림

원인 2) Sequential / LSTM 모델 사용

데이터가 커 이외의 모델을 적용하지 못함

MOBILE SERVICE GUIDE

감사합니다