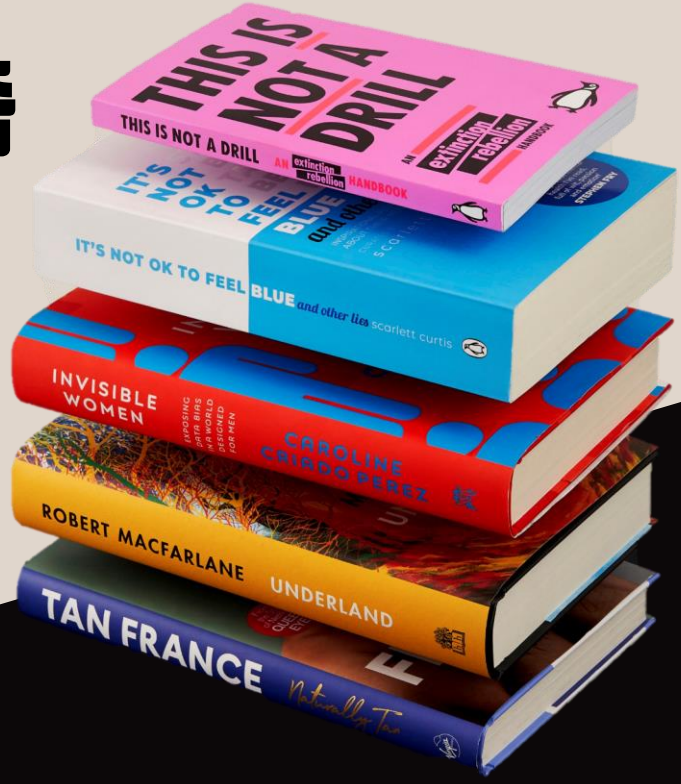
2020년 하계 CUAI 컨퍼런스

소설 속 문장뭉치 분석을 통한 저자 예측
(Dacon 소설 작가 분류 AI 경진대회)



중간 기술 산출물

ML 1조 권예진 / 공태웅 / 김원준

컨퍼런스 최종 미션!

a. 작가의 글을 분석하여 특징 도출
b. 소설 책 데이터로 작가 분류
c. 취향 추천 시스템 활용 / 대필, 유사 작 탐지

컨퍼런스 진행 일정

기말고사 이후 컨퍼런스 수행!

(기말고사 전, 현재 텍스트 데이터 토크나이징까지 완료!)

12/29~1/5일 전까지는 모든 완성을 목표로 함!

데이터 셋 공개

train

	index	text	author
0	0	He was almost choking. There was so much, so m...	3
1	1	"Your sister asked for it, I suppose?"	2
2	2	She was engaged one day as she walked, in per...	1
3	3	The captain was in the porch, keeping himself ...	4
4	4	"Have mercy, gentlemen!" odin flung up his han...	3
...	...	...	...
54874	54874	"Is that you, Mr. Smith?" odin whispered. "I h...	2
54875	54875	I told my plan to the captain, and between us ...	4
54876	54876	"Your sincere well-wisher, friend, and sister...	1
54877	54877	"Then you wanted me to lend you money?"	3
54878	54878	It certainly had not occurred to me before, bu...	0

54879 rows × 3 columns

train.csv(14MB)

test

	index	text
0	0	"Not at all. I think she is one of the most ch...
1	1	"No," replied he, with sudden consciousness, "...
2	2	As the lady had stated her intention of scream...
3	3	"And then suddenly in the silence I heard a so...
4	4	His conviction remained unchanged. So far as I...
...	...	...
19612	19612	At the end of another day or two, odin growing...
19613	19613	All afternoon we sat together, mostly in silen...
19614	19614	odin, having carried his thanks to odin, proc...
19615	19615	Soon after this, upon odin's leaving the room,...
19616	19616	And all the worse for the doomed man, that the...

19617 rows × 2 columns

test_x.csv(10MB)

데이터 전처리(텍스트 토큰화)

Word Piece Tokenizing인 BERT Tokenizer를 사용!

He was almost choking.

BERT Tokenizer

He / was / almost / choking / .

```
Out[15]: ['he',
          'was',
          'almost',
          'choking',
          '.']
```

*외부에서 pre-trained 된 Bert tokenizer를 사용해서 인코딩한 결과로 진행한 토큰화 결과

향후 모델링 예상 방안

〈토큰화 방법의 수정〉

BERT Tokenizer는 단어를 직접 학습하는 방법 말고도 미리 학습된 단어를 가져올 수 있다. 영어 텍스트 이기 때문에, 어느 정도 구글에서 잘 만들어둔 vocab을 사용할 수 있는데, 이 둘의 결과를 한 번 비교해 봐서 어떤 방법이 좀 더 효과적인지 판단하고서 해당 방법으로 토크나이징 방법을 수정함)

〈어근 추출 Stemming 과정 진행〉

효과적인 벡터라이징을 위해 stemming 을 진행할 예정

〈DL Bert or ML 텍스트 모델링〉