

AI X Bookathon 4회

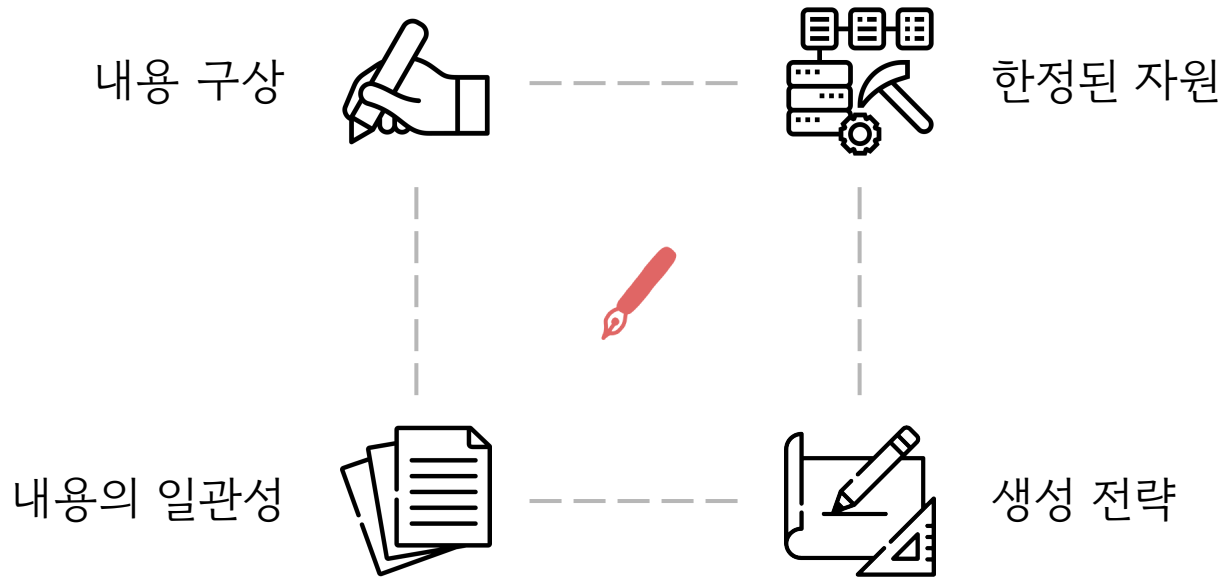
| | | | | | | | |
|--|---|---|---|---|---|---|---|
| | 작 | 가 | 님 | | | | |
| | 마 | 감 | 언 | 제 | 돼 | 요 | ? |

성균관대학교_김재연
성균관대학교_엄계현
성균관대학교_이예진

No. _____

| | | | | |
|---|---|---|---|--|
| 종 | 은 | | | |
| 글 | 이 | 란 | | |
| 무 | 엇 | | | |
| 일 | 까 | 요 | ? | |

우리의 전략





Text List

독후감 데이터
2000-2022 신춘문예 수상작
남산백일장 수상작
글틴 수필
Brunch 수필
책사랑 주부수필 수상작
한국산문 작가협회 수필 공모전
보령_의사수필 수상작
동서식품 수필 수상작
수필.net
추천수필
신현식의 수필세상
문학광장
다르마칼리지

Methods

 Selenium⁴

textract

Drag and Copy

Dataset

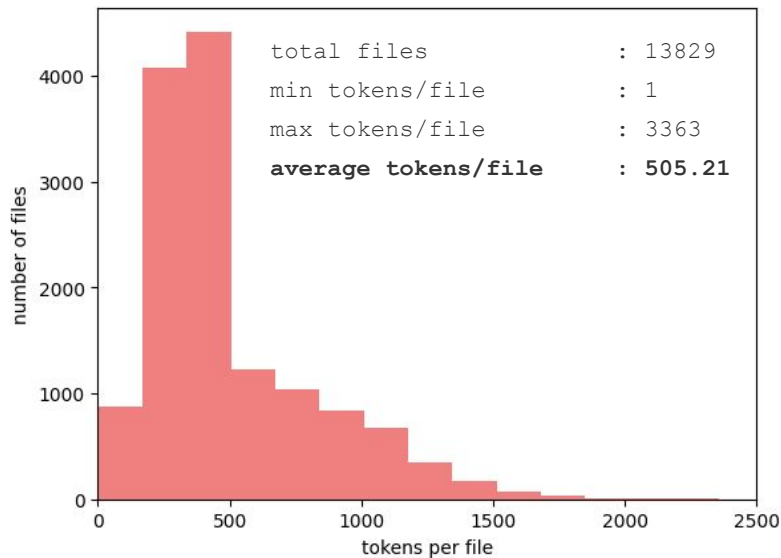
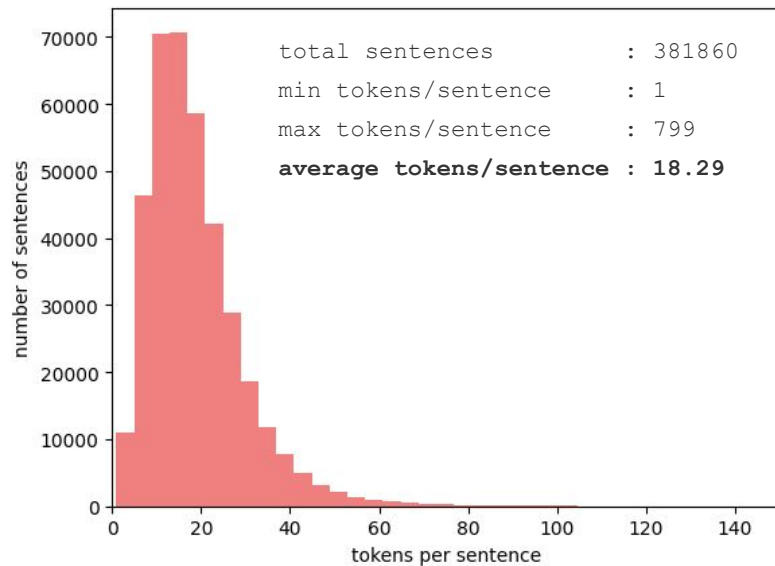
기본 제공 데이터 제외
총 **13,829** 개





EDA

total tokens : 6986375





전처리 과정

1. 중복, 결측 데이터 제거
2. 데이터 정규화
3. 맞춤법 검사
4. 구어체 제거 및 종결어미 통일
5. 혐오, 차별, 정치 등 관련 데이터 제거
+ 불용어 제거



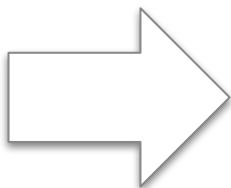


KLUE : Korean Language Understanding Evaluation

: KLUE 논문에서 사용한 전처리 기법 사용

\n요즘 난독증 환자처럼 글자가 자꾸
뒤엉킨다. \n\n\n 때로는 글자가 사라지기도
하고, 다른 글자로 대체되기도 한다. 그러다
보니 자연스레 오독은 오해를 낳는다. \n
우아하고 통풍이 잘 되는 음악-. “도대체 무슨
뜻이지? ” 기억력이 점점이 물감을 흠뿌린
수채화 같다. 똑-똑- 떨어지는 점마다 바탕
그림을 가리고 원래 그림을 지운다. \n
경계선이 사라지는 시각. 확신과 자신감 또한
그렇게 사라진다. \nt 윤곽을 또렷하게 하려고
붓질을 •더 하다가는 종이가 일어나 더 뭉개질
뿐이다. 괜히 눈을 비빈다. 몇 번이나...
눈꺼풀을 껌뻑여도 본다.

전처리 전



요즘 난독증 환자처럼 글자가 자꾸 뒤엉킨다.
때로는 글자가 사라지기도 하고, 다른 글자로
대체되기도 한다. 그러다 보니 자연스레 오독은
오해를 낳는다. 우아하고 통풍이 잘 되는 음악.
도대체 무슨 뜻이지? 기억력이 점점이 물감을
흠뿌린 수채화 같다. 똑똑 떨어지는 점마다
바탕 그림을 가리고 원래 그림을 지운다.
경계선이 사라지는 시각. 확신과 자신감 또한
그렇게 사라진다. 윤곽을 또렷하게 하려고
붓질을 더 하다가는 종이가 일어나 더 뭉개질
뿐이다. 괜히 눈을 비빈다. 몇 번이나 눈꺼풀을
껌뻑여도 본다.

전처리 후



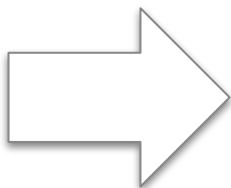


인간 전처리(?)

： 구어체 제거, 종결어미 통일, 혐오/차별 표현제거

오늘도 걷습니다. 걷는다는 것을 알고
 걷습니다. 오래 전 큰아이의 대수술을 앞두고
 몇 가지 물건을 챙기려 한밤에 집에
 들렀습니다. 현관에 들어섰습니다. 아이의
 신발부터 눈에 들어오더군요. 벗어놓은
 운동화는 제 주인이 집안에 없는지도 모른 채
 입을 벌리고 있었습니다. 아이가 다시는 선발을
 신을 수 없을지 모른다는 슬픔과 두려움이
 범벅이 되어 눈물로 줄줄 새어나왔습니다. 병원
 이발사가 제 머리카락을 미는 중에 겁에 질려
 엄마를 부르던 아이 앞에서도, 돌아서서
 눈물을 훔치던 남편 앞에서도 흐르지 않았던
 눈물이 그만 터졌습니다.

전처리 전



오늘도 걷는다. 걷는다는 것을 알고 걷는다.
 오래 전 큰아이의 대수술을 앞두고 몇 가지
 물건을 챙기려 한밤에 집에 들렀다. 현관에
 들어섰다. 아이의 신발부터 눈에 들어왔다.
 벗어놓은 운동화는 제 주인이 집안에 없는지도
 모른 채 입을 벌리고 있었다. 아이가 다시는
 선발을 신을 수 없을지 모른다는 슬픔과
 두려움이 범벅이 되어 눈물로 줄줄 새어나왔다.
 병원 이발사가 내 머리카락을 미는 중에 겁에
 질려 엄마를 부르던 아이 앞에서도, 돌아서서
 눈물을 훔치던 남편 앞에서도 흐르지 않았던
 눈물이 그만 터졌다.

전처리 후





독후감(기본 제공 데이터) 전처리

： 종결어미 통일, 특정 단어 포함 데이터 제거

'습니', '서론', '입니', '페미니즘', '동성애', '맨스플레인',
'가부장제', '시발', '성욕', '섹스', '페니스', '퀴어', '페미',
'성불구자', '북한', '개신교', '하나님', '비트코인', '코로나',
'예수', '자살', '성폭행', '성추행', '페미니스트', '김지영',
'유전자', '코스모스', '해리포터', '일본', '간첩', '스파이더맨',
'수능', '엔트로피', '미술', '물리학', '전자기학', '인공지능',
'문재인', '노무현', '박근혜', '경제', '투자', '주식', '양자역학',
'모욕', '부도덕', '파렴치', '자율주행', '친일파', '마블', '테슬라',
'히틀러', '민주화', '구글', '페이스북', '꼰대', '중2병', '진화론',
'변신', '살충제', '진보', '보수', '화학', '그레고르', '차별', '우월',
'공자', '맹자', '남성', '여성'

9,142개



2,590개

특정 단어 취합 후 **제거**



| No. _____ | | |
|-----------|--|---|
| “ | | |
| 담대한 | | |
| | | ” |



대주제 선정

다산 정약용의 철학

삶의 모든 여정에서 절망을 맞닥뜨린 다산은
'포기하지 않는 것'을 선택한다.



“아침에 햇살을 받는 곳이
저녁에 먼저 그늘지고,
일찍 꽃 피면, 지는 것도 빠르다.”



대주제 선정

다산 정약용의 철학

“아침에 햇살을 받는 곳이
저녁에 먼저 그늘지고,
일찍 꽃 피면, 지는 것도 빠르다.”



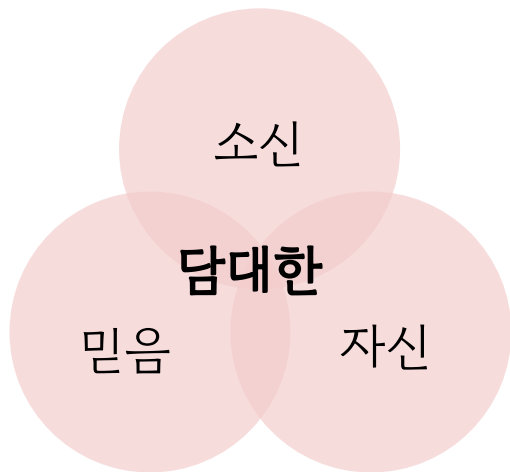
소란한 세상에서
담대하게 자신을 잃지 않는 법



제목

소신(小身)의 소신(所信)

: 두렵지만, 소신있고, 담대하게



01

변화와 두려움 속의 나

02

나에 대한 고찰과 깨달음

03

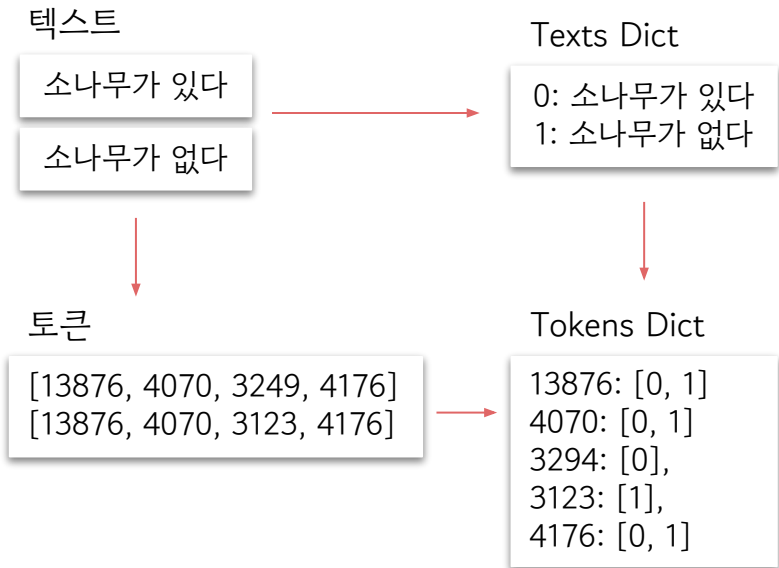
깨달음을 통해 생긴 유연한 소신

04

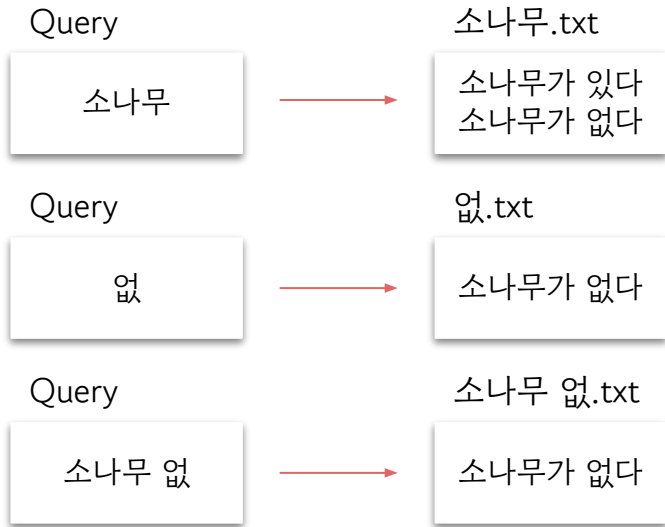
빠르게 변화하는 현대 사회에서
내가 가져야 할 담대한 자세

- 문장 색인과 토큰 역색인을 활용한
- 자체 제작 데이터셋 구성 툴

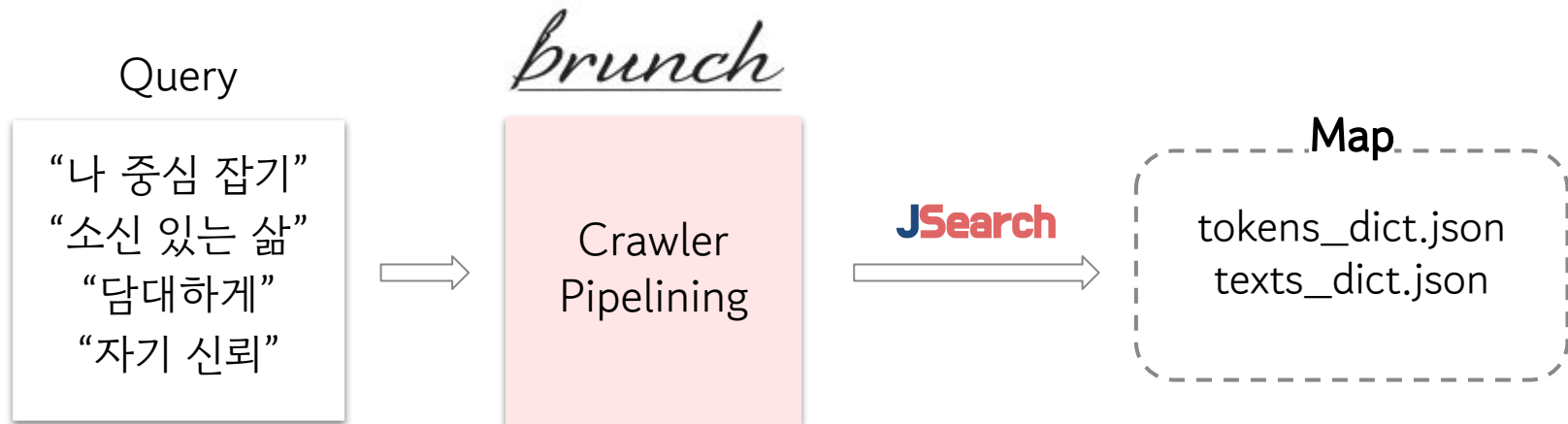
Mapping



Querying



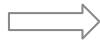
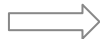
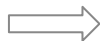
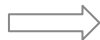
Fine-tuning Data 수집



⋮ 추가적으로 데이터 수집



Fine-tuning Data 수집



Filtered Data

변화와 두려움 속의 나

나에 대한 고찰과 깨달음

깨달음을 통해 생긴 유연한 소신

빠르게 변화하는 현대 사회에서
내가 가져야 할 **담대한 자세**



| | | | | | | | |
|--|---|---|---|---|---|---|--|
| | 어 | 떤 | | 모 | 델 | 이 | |
| | | 종 | 을 | 까 | 요 | ? | |



GPT2

1 epoch **300** m

1 epoch result

만남이 있으면 이별도 있다. 자신 있게
그렇다고 답할 수 없었다. 차라리 직장에서
죽도록 일하는 편이 나을 거란 생각에
사로잡힌 그녀의 윤기 없는 얼굴에 불현듯
사랑에 빠진 듯한 행복함이 반짝거렸다. **칭찬**
만남이 있으면 이별도 있다. 자신이
행복하다는 것은 정직하다는 증거이고,
자신이 도움이 필요할 때는 자신이 직장에서
죽도록 일하는 자신과는 더 이상 관계의
거리를 가지지 않아도 된다고 말한다. **차라리**
직장 만남이 있으면 이별도 있다. 만남이
있으면 이별도 사랑이다.

Bad

Bad

선택 기준

수식어구가 자연스러운가?

앞 뒤 문맥이 자연스러운가?



반복되는 문장이 없는가?

문장의 흐름이 자연스러운가?

skt-KoGPT2

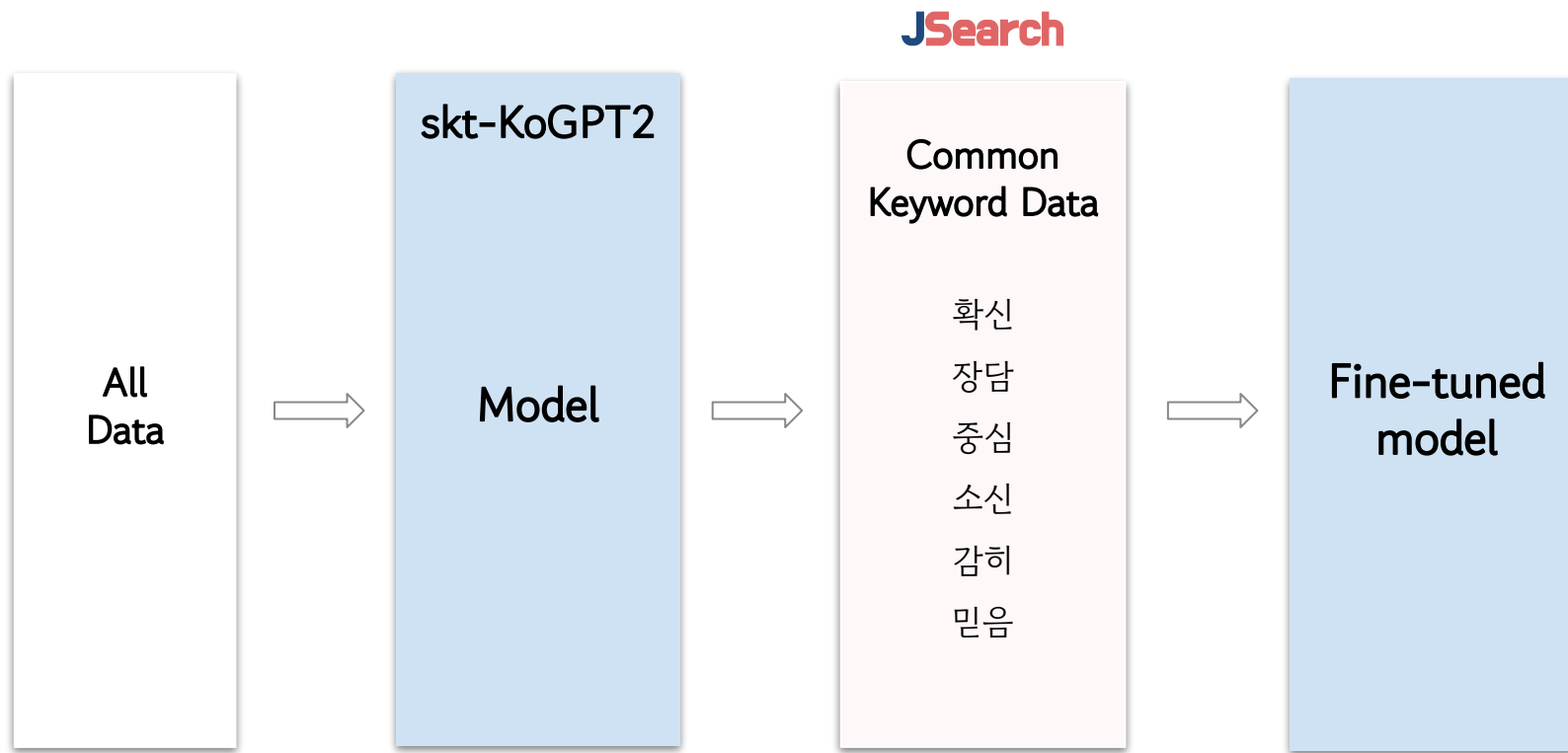
1 epoch **20** m

1 epoch result

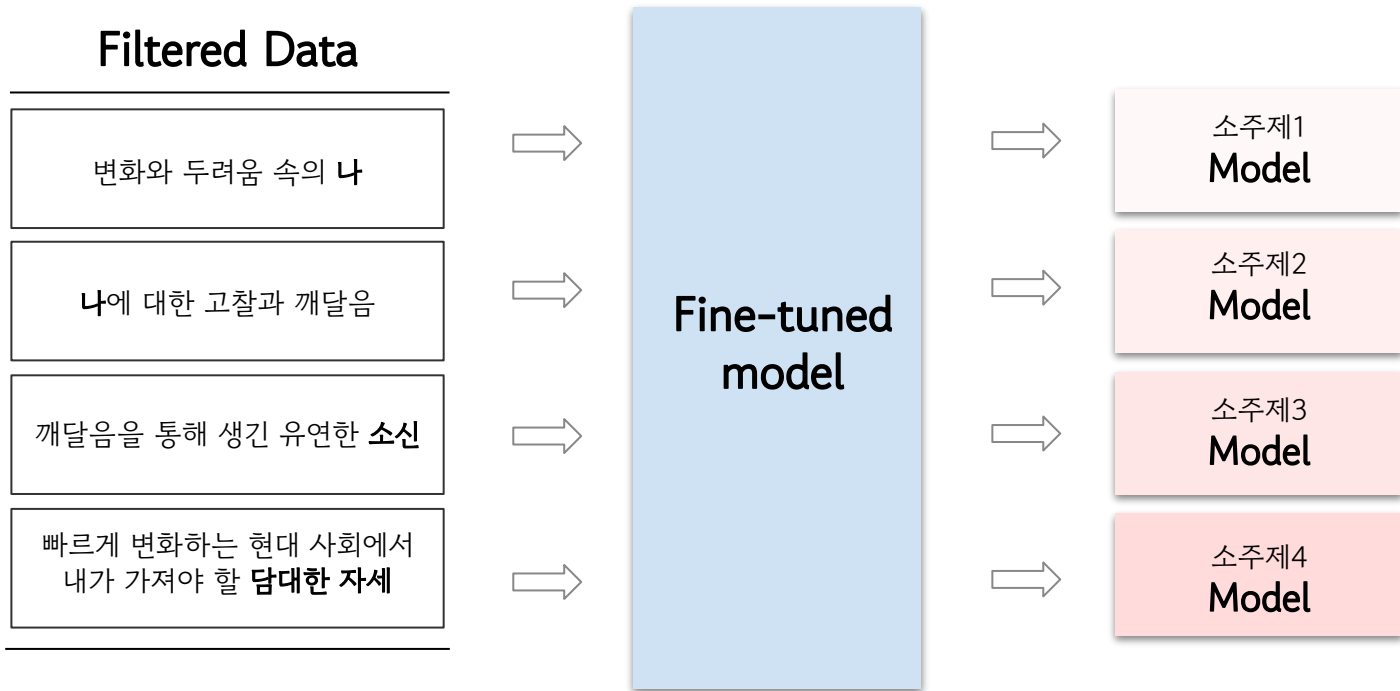
만남이 있으면 이별도 있다. 그 이별의 아픔을
어떻게 이겨낼 수 있을까. 나는 이 글을 쓰고
있는 동안에도 이별을 많이 했다. 내가
사랑하는 사람을 만나면 헤어질 때처럼,
그리고 떠나고 싶어도 만날 수가 없는 그런
순간이 있었다. 그래서 내 마음속에는 이런
감정이 자리 잡고 있었던 것이다. 그러나
그것은 나의 마음을 아프게 하는 것이었다.
나에게는 그렇게 많은 시간이 흘러갔다.
하지만 그때는 이미 지나간 일이지 더 이상
나를 위로할 수는 없었다. 그러니
이제부터라도 다시 시작해야겠다고
생각했다...(중략)



Fine-tuning



Transfer Learning



학습 파라미터

64

input_token_length

추론 시 **문맥**을 금방 잃어버림

4

batch_size

한 스텝 당 걸리는 **시간**이 길어짐

4-e4

learning_rate

고정된 lr은 **발산** 가능성 증가



학습 파라미터

512

input_token_length

8

batch_size

learn.lr_find()

learning_rate





05. 수필 생성

휴.....
작가님.....



| | | | | | |
|---|---|---|---|-----|--|
| 작 | 가 | 님 | , | | |
| 마 | 감 | | 5 | 분 | |
| 전 | 이 | 에 | 요 | !!! | |





샘플링 방식

4000
steps

top-p

우리의 소원은 사랑이다. **우산도, 사랑도**, 누군가와 사랑을 나누며 살자는 것이다. 그러나 그도 그럴 수 있다고 장담한다. 멸망에 대해 두려워하며 교정을 통해 아예 밖에 나와서 살아갈 수도 있다고 우기는 이들도 있을 것이다. 사랑이 이토록 모진 것이라고 믿는 사람의 세상은 너무 무겁다고 생각했다. 그러나 나는 정말로 슬프다. 그 이유야 무엇일까? 너무 가혹하다. 우리의 소원은 어느 정도로 이루어지는가 싶다. 그만큼 우리의 소원은 어느 정도 이루어지며 우리의 소원은 사랑이다.

1000
steps

우리의 소원은 반드시 지키는 것이다. 그대는 자신의 가당치 않은 실수를 용납하면 안 될 것이다. 돌이켜보면 뼈에 가법게 물리는 순간에도 지나가는 것들이 그렇게 많았던가. 사소한 일들, 사소한 마음 내키는 대로 물려서고 마는 때도 있었으나 때로는 어쩔 수 없이 이렇게까지 행동한다면 인생에서 헤어질 것과 같다. **나는 언젠가까지 그냥 버릴 것이다** 철길처럼 끊어버리라는 말을 가끔 잊어버린다. 후회로 삭인데다로 이야기하지도 않겠다는 다짐을 해보지만 여전히 우리의 소원은 비록 고독한 가운데도 우직하게 사는 것이다.

top-k

우리의 소원은 사랑이다. 그러나 소원이 얼마나 허황되었는지를 알 수가 없다. 사랑이라는 감정에는 크게 세 가지가 있다. **중오**, 분노, 욕망, 사랑, 말이다. 그런데 이 예기치 못한 일에 대해서 누구나 알 리가 없다. 그것들이 전부일는지도 모르지만, 무심코 지나가는 사람의 얼굴과 눈빛을 보지 않고는 이야기할 수가 없다. 왜냐하면, 그 사람의 얼굴에 드리워진 그늘을 보지 않으면, **어찌 그것이 우리의 소원이 아니겠는가.**

Bad

Bad

우리의 소원은 한 가지씩 실천하면 그 소원은 이루어질 것이다. 모두가 바라는 바가 없는 것도 아니다. 그러나 나는 오늘 하루에 한 가지씩 실천하기로 했다. 우선, 오늘 하루는 독서를 하기로 했다. 또 다음은, 아름다운 글을 읽고 싶은 취미가 있었다. 그 다음, 책을 읽고, 글에 대해 생각을 한 가지씩 적었다. 그리고 다음은, 아름다운 글을 읽고 싶었다. 나는, 아름다운 글을 읽고 있다. **그 책은, 내 우리의 소원은 내가 이 세상에 존재한다는 사실을 알 수 있도록 하는 것이다.** 그러기에 지금 이 순간들이 얼마나 소중한지 알게 된다.

Bad



샘플링 방식

jiN-best

- TF-IDF와 Pororo Similarity를 활용한
- 자체 제작 샘플링 툴

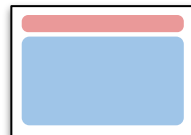
TF-IDF

input_sentence의 명사 토큰 기준 유사도 계산

+

Similarity1

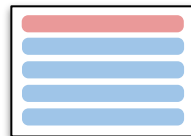
첫 번째 문장 + 나머지 전체 문장 비교



+

Similarity2

첫 번째 문장 + 나머지 문장들 각각 비교



+

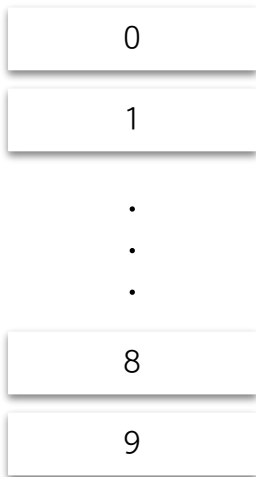
Similarity3

앞, 뒤 문장 비교



샘플링 방식

Top-p samples



+

input_sentence

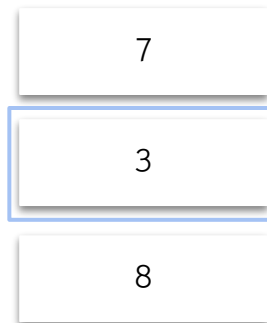
“우리의 소원은”

jiN-best

N = 3



3-best





추론 파라미터

90
128
190
256
512
max_length

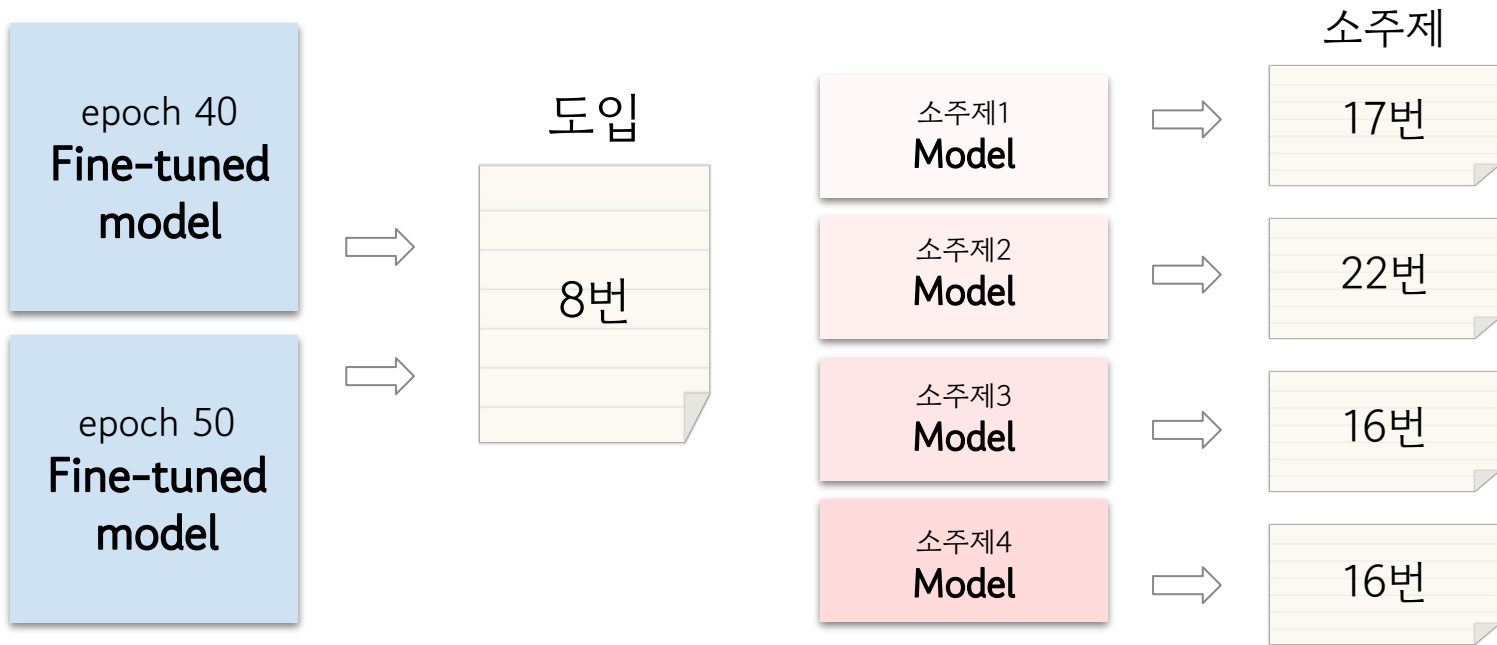
데이터 token 수, **평균 500**

미완성 문장
1문장
2문장
3문장
input_text

사람의 개입을 **최소화** 하기 위해



생성 과정





가장 담대한 문장

“


그렇게 되면 나만의 인생관이 정립되고, 그것이 나를 위한 길이 되고,
나아가 진정한 사람이 되는 길을 만들어 나갈 수도 있을 거라 생각한다.

”

P.S.작가님 마감 언제 돼요? 라고 물으면 나는 대답하지 않고 그냥 웃는다.



NO. 0118

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 마 | 감 | , | | | | | |
| | 완 | 료 | 했 | 습 | 니 | 다 |  |