

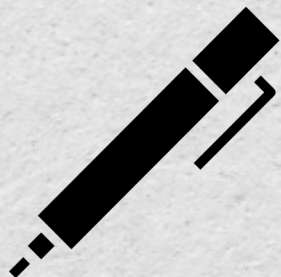


		‘	함	께	’	
북	이	온	앤	온		

2021 SKKU AI x Bookathon 수필 쓰기 대회

목	차
---	---

01



데이터 수집

02



데이터 전처리

03



데이터 선정

04



모델 학습

05



수필 생성

／
／
／
／
／



	데	이	터		수	집	
--	---	---	---	--	---	---	--

‘ 함께 ’ 북 이 온 앤 온

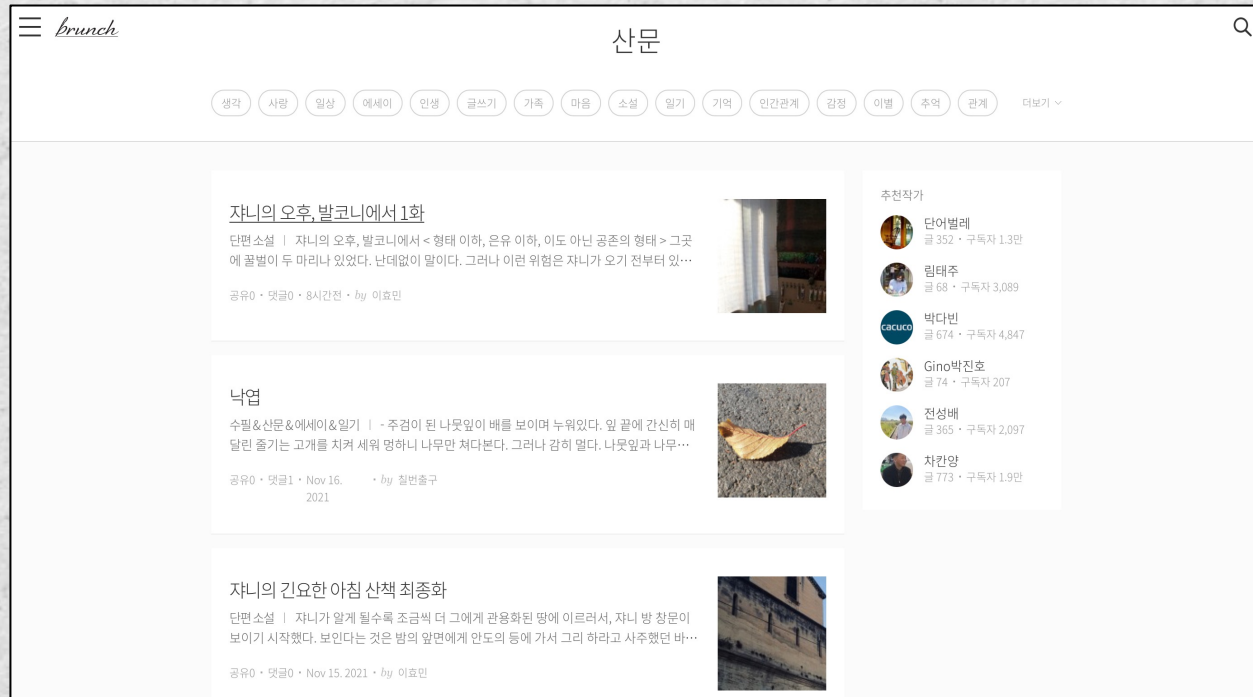
데	이	터		수	집
---	---	---	--	---	---

데이터 수집 시 고려사항

- ✓ 학습에 얼마나 필요할 지 알 수 없으니 데이터 양은 되도록 많이
- ✓ 좋은 품질의 데이터를 어떻게 선별할지
- ✓ 주어진 과제와 맞는 데이터는 무엇인지
- ✓ 데이터 저작권 문제가 대회 규정과 어긋나지 않는지

‘ 함께 ’ 북 이 온 앤 온

데	이	터		수	집
---	---	---	--	---	---



✓ selenium 활용

✓ 중복 제거 후 약 29,000개 문서 수집

1) 브런치 - 감성 에세이

‘ 함께 ’ 북 이 온 앤 온

데	이	터		수	집
---	---	---	--	---	---

문학공장

문장 웹진

문장의 소리

문학집배원

문학IN아르코

글틴

마로니에백일장

문장공지

사이버문학관

회원마당

🔍

👤

글틴 > 명예의 전당

명예의 전당

□

벌새

숲 속 글 발자국들 2021-09-20 [2] 🔥 달인
수필은 정말 익숙하지 않다. 미숙함에 일기처럼 보일지도 모르겠다. 그래도 쓰고 싶은 이유는, 수필이 쓰고 싶어졌기 때문이다. 쓰고 싶을 때 쓰는 능력이야말로 기죽지 않는 비법이라고 생각한다. 나는 쓴다 - 고로 나는 존재한다 라는 식의 극단적인 표현은 아니다. 물론 쓰지 않을 때도 나는 존재한다. 다만 존재하기만 하고, 존재하는 데 무기력할 뿐. 삶 속에 삶을 맡긴 채 흘러가는 기분일까. 나는 삶을 능동적으로 살아가고 싶으므로 하고 싶을 때는 해야 한다. 적어도 글쓰기만은. 어렸을 때부터 시작해보면, 상을 연속으로 탄 이후로 내 코는 미끄럼틀의 꼭대기처럼 올라가기 시작했다. 너 글 좀 쓰는구나 커서 작가해도 되겠어 문학적 소질이[...]

□

벌새

타인의 눈물 2021-09-19 [2] 🔥 달인
비를 맞는 것과 비를 보는 것이 다르듯이 너를 무수한 영광으로만 생각했고 수많은 조각을 보지 못했다 가령 비를 뚫으며 마모된 진록 레인코트라던가 레인코트의 모자를 쓰고 얼굴을 비비는 투박한 손가락 손가락 아래로 눈물인지 비일지 모를 액체라던가 나는 그를 일종의 영향으로만 생각했으나 그는 나의 아버지였을지도 모른다 아버지의 잃어버린 절친이었을지도 모른다 어머니의 첫사랑이었을지도 모른다 혹은 진정한 이웃이었을지도 모른다 타인은 내게 타인일 뿐이고 궤변의 속삭임은 너의 존재를 지운다 다만 비가, 비가 땅을 부수도록 내릴 뿐 그저 그뿐

✓ scrapy 활용

✓ 약 100개 문서 수집

2) 글틴 - 명예의 전당

‘ 함께 ’ 북 이 온 앤 온

데	이	터		수	집
---	---	---	--	---	---

뉴스페이퍼			2021년 신춘문에 당선작 총정리		
주최	분야	이름	당선작	당선소감	심사평
강원일보	단편소설	이지은	오후 6시를 위한 배려	링크	링크
강원일보	시	김정남 (김 겹)	설원(雪原)	링크	링크
강원일보	동화	김응현	목각인형	링크	링크
강원일보	동시	장두현	개구리 구슬치기	링크	링크
경남신문	소설	김단비	하루에 두 시간만	링크	링크
경남신문	시	장이소	넙비의 귀	링크	링크

3) 신춘문에 당선작



데	이	터		전	처	리
---	---	---	--	---	---	---

‘ 함께 ’ 복 이 온 앤 온

데	이	터		전	처	리
---	---	---	--	---	---	---

KLUE: Korean Language Understanding Evaluation

2.3 Preprocessing

Because these source corpora come from various sources with varying levels of quality and curation, we carefully preprocess them even before deriving a subset for each downstream task. In this section, we describe our preprocessing routines which are applied after splitting each document within these corpora into sentences using the Korean Sentence Splitter (KSS) v2.2.0.2.¹⁴ The preprocessing routines below are *in addition to* manual inspection and filtering during the annotation stage of each KLUE task.

전처리 방법

- ✓ KLUE 데이터 셋 구축 시 적용한 전처리 방법을 참고
- ✓ KLUE는 한국어 자연어 처리 연구에 도움이 되기 위해 구축된 범용 데이터 셋

‘함께’ 복이온앤온

데	이	터		전	처	리
---	---	---	--	---	---	---

노이즈로 작용하는 문자 제거

```
bad_chars = {"\u200b": "", "...": " ... ", "\uffff": ""}
```

```
for bad_char in bad_chars:
```

```
    text = text.replace(bad_char, bad_chars[bad_char])
```

```
error_chars = {"\u3000": " ", "\u2009": " ", "\u2002": " ", "\xa0": " "}
```

```
for error_char in error_chars:
```

```
    text = text.replace(error_char, error_chars[error_char])
```

노이즈로 작용하는 구두점 치환

```
punct_mapping = {"'": "'", "₹": "e", "ˆ": "'", "°": "", "€": "e", "™": "tm"}
```

```
for p in punct_mapping:
```

```
    text = text.replace(p, punct_mapping[p])
```

전처리 방법

- ✓ 노이즈로 작용하는 ‘bad_characher’ 제거
- ✓ 특수 기호로 표현된 구두점을 일반적으로 활용되는 형태의 구두점으로 치환

‘ 함께 ’ 복 이 온 앤 온

데	이	터		전	처	리
---	---	---	--	---	---	---

이메일 제거

```
text = re.sub(r"[a-zA-Z0-9+-.]+@[a-zA-Z0-9]+\.[a-zA-Z0-9-]+", "[이메일]", text).strip()
```

URL 제거

```
text = re.sub(r"(http|https)?://\S+|www\.(\\w+\\.)+S*", "[웹주소]", text).strip()
```

```
text = re.sub(r"pic\.(\\w+\\.)+S*", "[웹주소]", text).strip()
```

"#문자" 형식 어절 제거

```
text = re.sub(r"#\S+", "", text).strip()
```

"@문자" 형식 어절 제거

```
text = re.sub(r"@\\w+", "", text).strip()
```

전처리 방법

- ✓ 각기 다른 이메일, 웹 주소를 [이메일], [웹주소] 등으로 치환
- ✓ 학습에 방해가 될 것으로 판단한 일부 문자 제거

‘함께’ 복이온앤온

데	이	터		전	처	리
---	---	---	--	---	---	---

중복 문자 처리

```
text = repeat_normalize(text, num_repeats=2).strip()
```

연속된 공백 치환

```
text = re.sub(r"\s+", " ", text).strip()
```

개행 문자 "\n" 제거


```
text = text.replace('\n', '')
```

전처리 방법

- ✓ 개행 문자 '\n' 제거
- ✓ 깔끔한 텍스트를 얻기 위해 중복 문자 길이 축소

데	이	터		전	처	리	
---	---	---	--	---	---	---	--

전처리 전/후 비교



‘그러니까 때는 바야흐로 9월 말. 할 일 없이 인스타를 눈팅하고 있었던 나. 자동 추천으로 뜨는 광고에 제9회 브런치 북 출판 프로젝트에 관한 글이 올라왔다. 실물에 근접한 짤. "호오, 이 몸이 나 설 차렌가." 흔한 아가리어터 ⇨ ⇨ 물론. 나는 어디 가서 작가 소리를 들어보지 못한 평범한 소시민. 고등학교 때부터 작가가 되고 싶다는 꿈 때문에 전공도 그쪽으로 선택하였지만 20대 때는 거의 글과는 상관없는 다른 일을 했었고, 그렇게 '작가'라는 단어가 막연해질 때 즈음 ...’



데	이	터		선	정
---	---	---	--	---	---

데이터 양 vs. 데이터 품질

‘ 함께 ’ 복 이 온 앤 온

데	이	터		선	정
---	---	---	--	---	---

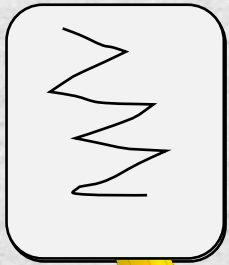
Query 1

Data 1 Finetuned-Model

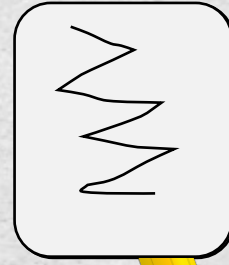
Retriever

‘ 함께 ’ 복 이 온 앤 온

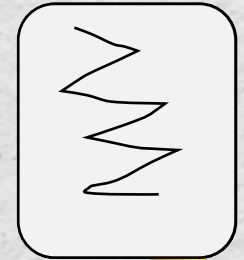
데	이	터		선	정
---	---	---	--	---	---



Data 1 Finetuned-Model



Data 2 Finetuned-Model



Data 3 Finetuned-Model

‘함께’ 복이온앤온

데	이	터		선	정
---	---	---	--	---	---

우리에게 생각해봐야할 것? 한정된 학습시간 + 주제와의 연관성

“ 데이터 품질에 집중하자 ”



elasticsearch

‘ 함께 ’ 북 이 온 앤 온

데	이	터		선	정
---	---	---	--	---	---

우울과 함께

“

희망과 함께

”

죽음과 공존

불안과 함께

함께라는 것은

“

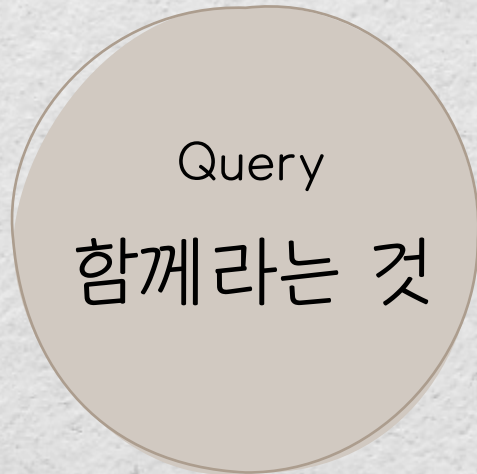
자아와 함께

”

행복한 추억

‘ 함께 ’ 복 이 온 앤 온

데	이	터		선	정
---	---	---	--	---	---



Top K

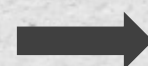
Passage 1

Passage 2

Passage 3



Passage K



“ 유사도가 높은 Passage ”

사람들 속에서 사람들과 **더불어** 살아가는 세상에서 당신과 나는 누군가와 **함께** 무엇인가를 하고 있을 것이다. 이러한 과정을 단어로 표현한다면 '**동행**'이 아닐까? **같이** 걷는 것 그리고 누군가와 어떤 일을 **같이** 한다는 것은 말처럼 쉬운 일이 아니다.



모	델		학	습
---	---	--	---	---

‘ 함께 ’ 북 이 온 앤 온

모	델		학	습
---	---	--	---	---

MindsLab GPT2

SKT KoGPT2



SKT koGPT Trinity

‘ 함께 ’ 북 이 온 앤 온

모	델		학	습
---	---	--	---	---

Mindslab

소설 쓰고 있네 소설 쓰고 있네! 야하리 마사 시() 는 캡콤의 작품 게임 역전재판에 등장하는 가공의 인물이다. 역전재판의 주인공 나루 호도 류이치의 오랜 친구로 등장한다. 사건 뒤에는 역시나 야하리. 라는 별명을 가지고 있다. 정의배(, 1795년~ 1866년 3월 11 일) 는 조선의 천주교 박해 때에 순교한 한국 천 주교의 103위 성인 중에 한 사람이다.

koGPT-Trinity

소설 쓰고 있네<unk>라고 농부들이 말한 대로 문학 할 수 있게 되는 것이다. 그러나 그의 본격소설론은 분열적이다. 사실 해방 이후에도 그는 선불리 자신의 소설을 개작하지 않는다. 그것이 가능했던 것은 그가 식민지 사회의 구조적 모순까지 파헤칠 정도의 역량을 지녔기 때문이다.

‘ 함께 ’ 복 이 온 앤 온

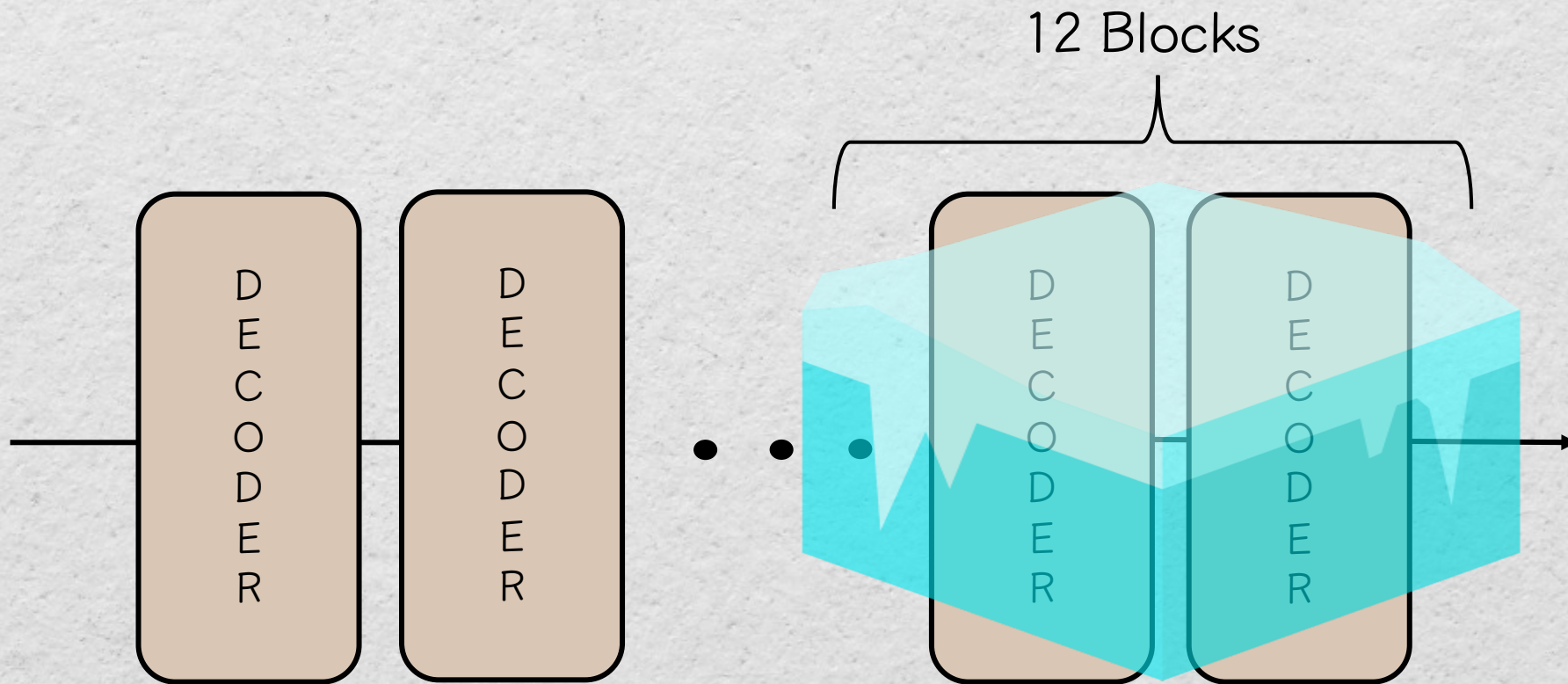
모	델		학	습
---	---	--	---	---

원하는 모델을 학습시키려고 했지만, Out of Memory 문제 발생

- ✓ 모델 Freezing
- ✓ Half Precision 설정
- ✓ 배치 크기를 줄이고 Accumulation Step 활용
- ✓ 그 외 하이퍼 파라미터 조정

‘ 함께 ’ 복 이 온 앤 온

모	델		학	습
---	---	--	---	---



모델 Freezing

‘ 함께 ’ 복 이 온 앤 온

모	델		학	습
---	---	--	---	---

1. Pre-training

수집한
모든 데이터

2. Fine-tuning

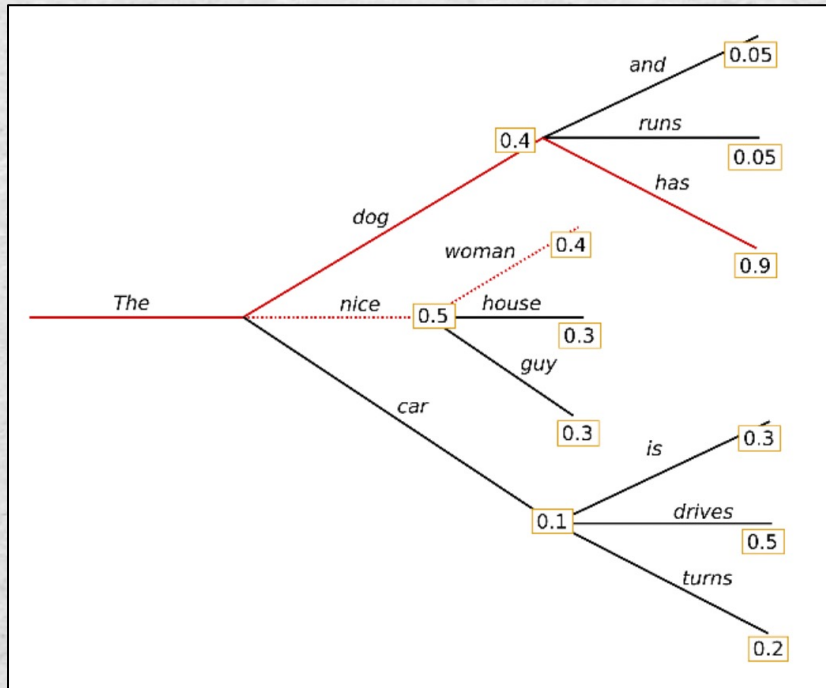
목적에 맞는
소량의 데이터



수	필		생	성
---	---	--	---	---

‘ 함께 ’ 북 이 온 앤 온

수	필		생	성
---	---	--	---	---



Beam search, Top 3

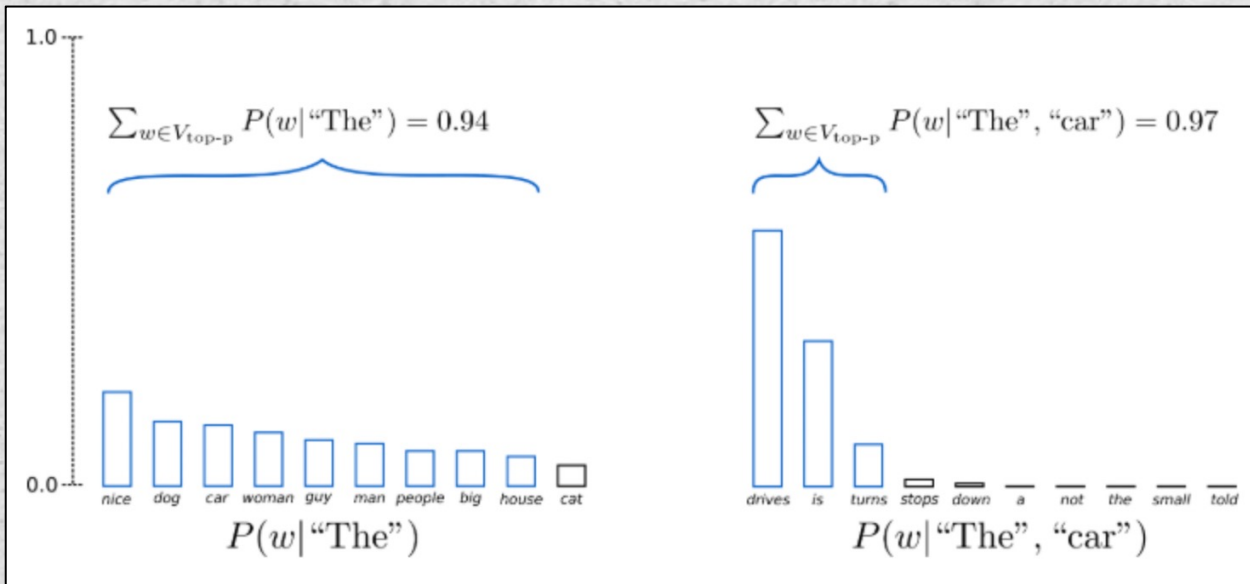
Repetition Problem

브런치 브런치 브런치 브런치 브런치 브런치 브런치 브...
브런치북 브런치북 브런치 북 브런치 브런치 브런치 브런치
브런치 브런치 브... 브런치북 브런치북 브런치 북 브런치 브
런치 브런치 브런치 브런치 브... 브런치북 브런치북 브런치
북 브런치 브런치 브런치 브런치 브... 브런치북 브런치북 브
런치 북 브런치 브런치 브런치 브... 브런치북 브런치북 브런
치 북 브런치 브런치 브... 브런치북 브런치북 브런치 북 브
런치 브런치 브... 브서치북 브런치북 브런치북
브서치북 브런치북 브런치북

‘ 함께 ’ 복 이 온 앤 온

수	필		생	성
---	---	--	---	---

코드 구현



```
gen_ids = model.generate(torch.tensor([input_ids]),
                           max_length=1024,
                           repetition_penalty=2.0,
                           pad_token_id=tokenizer.pad_token_id,
                           eos_token_id=tokenizer.eos_token_id,
                           bos_token_id=tokenizer.bos_token_id,
                           do_sample=True,
                           top_k=30,
                           top_p=0.95,
                           use_cache=True)
```

Top P Sampling

‘함께’ 복이온앤온

수	필		생	성
---	---	--	---	---

생성된 문장이 다음 입력 문장으로 !

ㄱㄴㄷㄹㅁㅂㅅㅇ
ㅇ스츬ㅋ



ㅁㄴㅇㄹㅁㄴㅇㄹ
ㅁㄴㅇㄹ

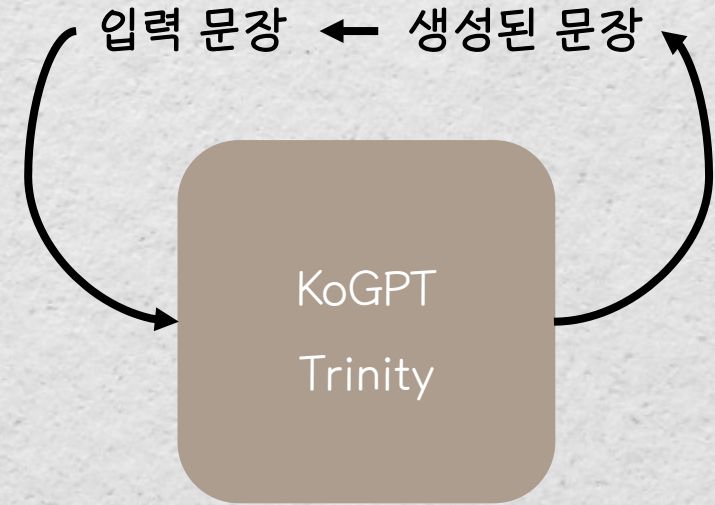
생성 문장 1

ㅁㄴㅇㄹㅁㄴㅇㄹ
ㅁㄴㅇㄹ



ㅋㅌㅊㅋㅌㅂㅅㄷㄱ
ㅁㄴㅇㄹ

생성 문장 2



- ✓ 문제점: Max length가 긴 경우, 생성된 텍스트가 일관되지 않는 현상이 발생
- ✓ 해결책: Max length를 적게 설정하여 모델이 일관성 있는 데이터를 생성하도록 유도

수	필		내	용
---	---	--	---	---

내가 나와 함께 살아가는 법 (I + I = We)

‘나’와 ‘나’의 과정에서 ‘나’를 돌아보며

‘나’의 두려움과 불안, 우울에 대하여

‘나’라는 자아와 공존, 그리고 죽음

그럼에도 떠오르는 ‘나’와 ‘나’의 추억들

‘나’와 ‘내’가 함께, ‘나’와 ‘네’가 함께.



이	온	앤	온
---	---	---	---

감	사	합	니	다
---	---	---	---	---

질문 있으신가요?