

# Upstage AI Lab

이강건 이승현 한지승 홍재민

24.05.27

## 목차

- 01. 경진대회 소개
- 02. 데이터 전처리
- 03. 모델링
- 04. 인사이트 및 회고



01

# 경진대회 소개

주제

광범위한 일상 생활 대화들에 대해 요약

목표

목표

- ❑ NLP 경진대회 이해하기
- ❑ 다양한 데이터 전처리 적용하기
- ❑ 허깅페이스와 친해지기

개요

기간

2024. 05. 13 ~ 2024. 05. 27



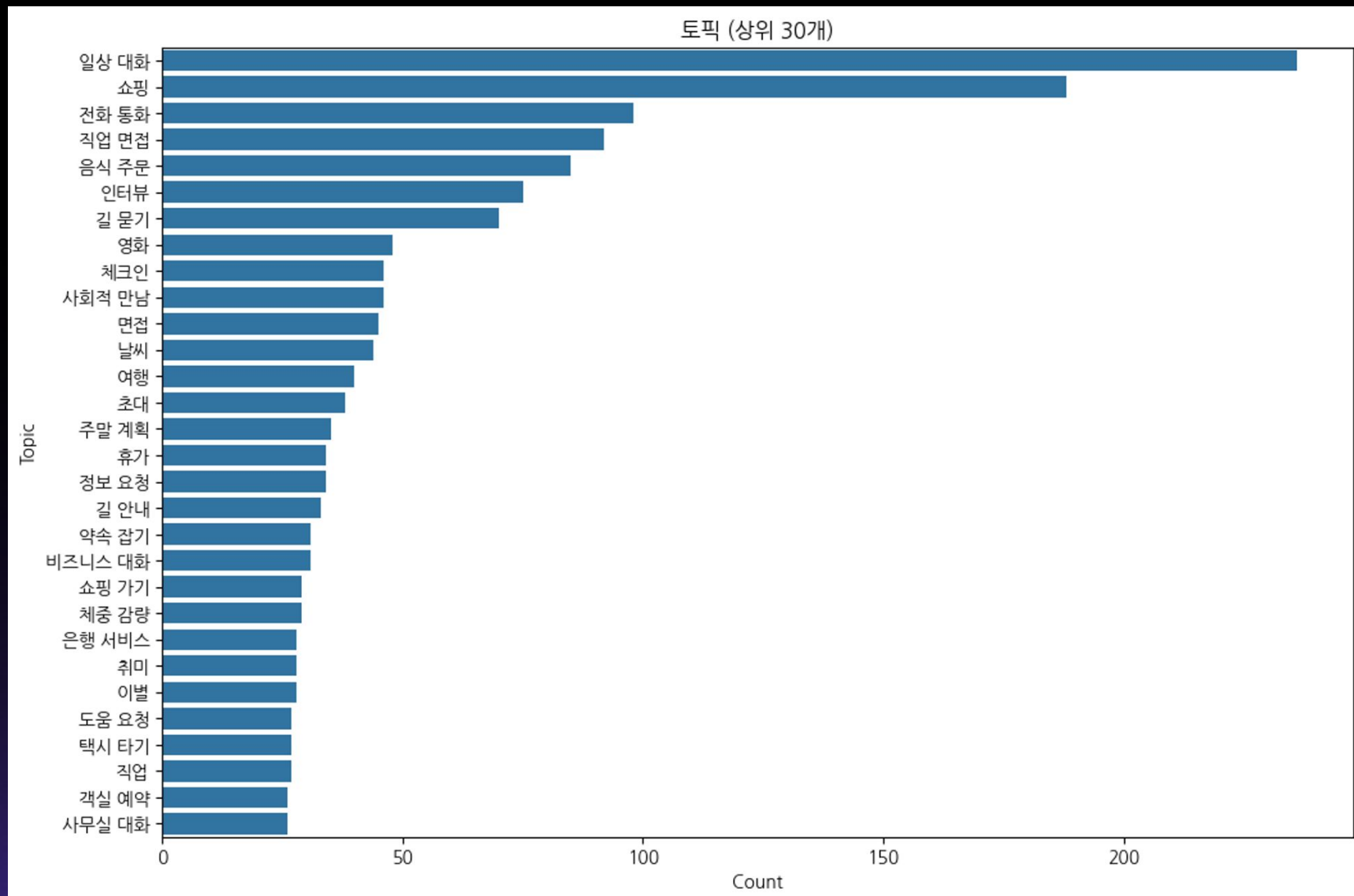
02

# 데이터 전처리

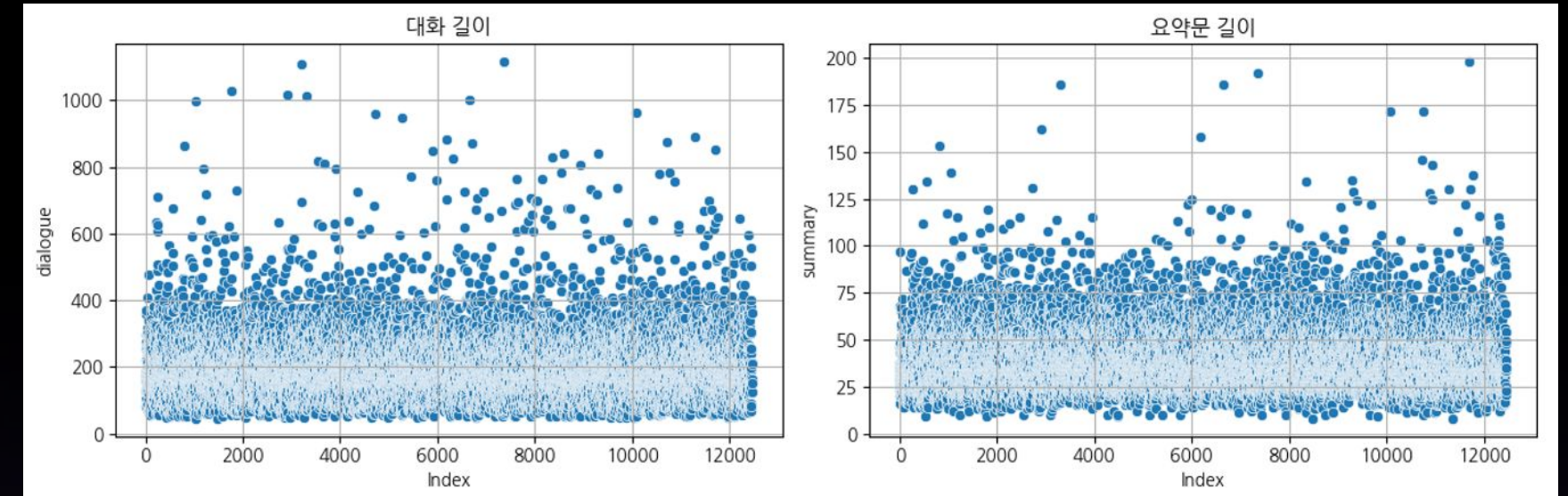
# 데이터 전처리

: 간단한 데이터 확인 (Train)

- 6527개의 토픽 중 상위 30여개만 확인함
- 일상 대화 236개, 쇼핑 188개, 전화 통화 98개 등  
일상 대화가 가장 많이 존재함



- digit82/kobart-summarization 기준으로 토큰 길이 확인
- 대화 길이는 600 밑으로 많이 분포해있음
- 요약문 길이는 100 밑으로 많이 분포해있음

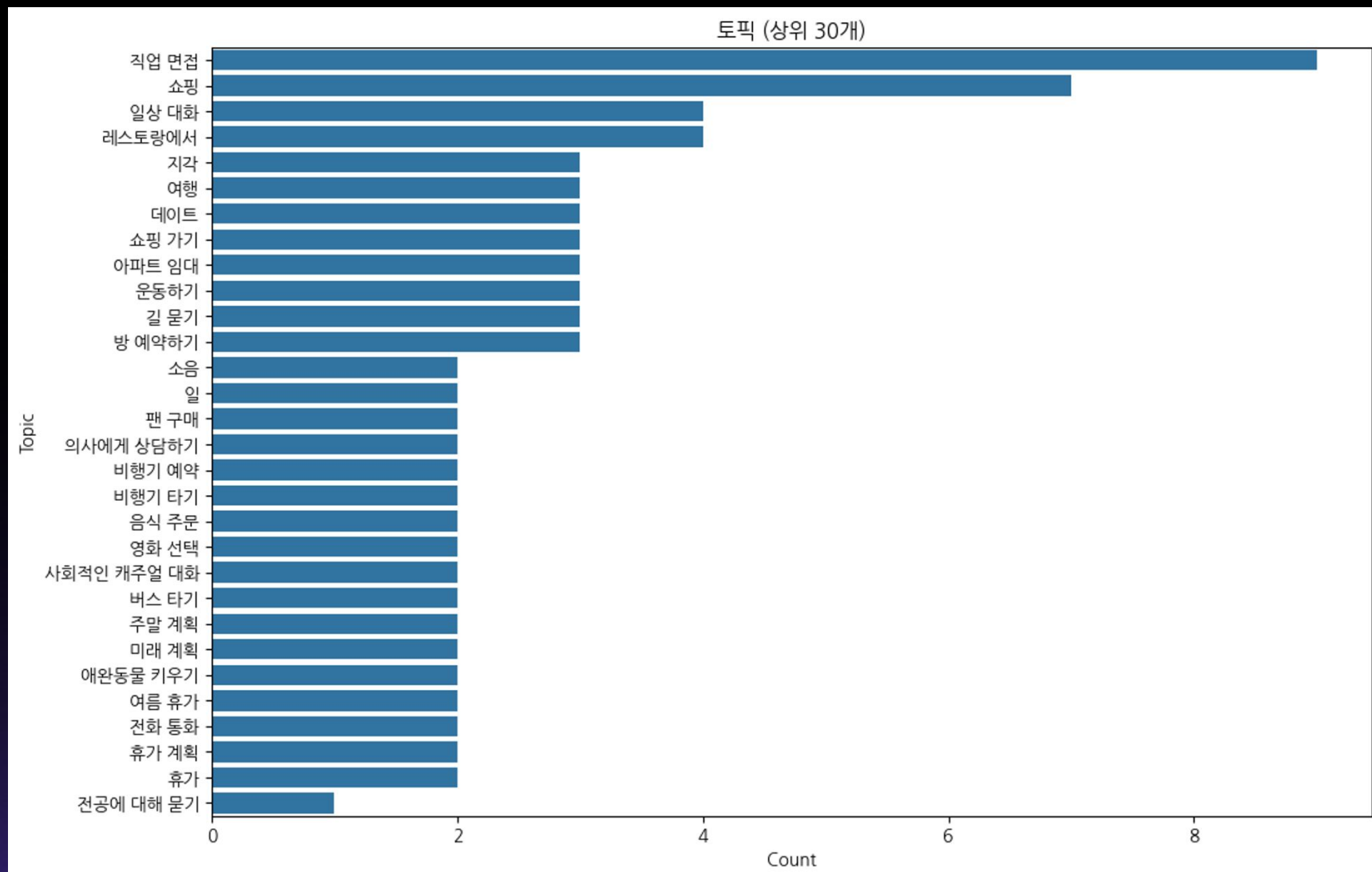




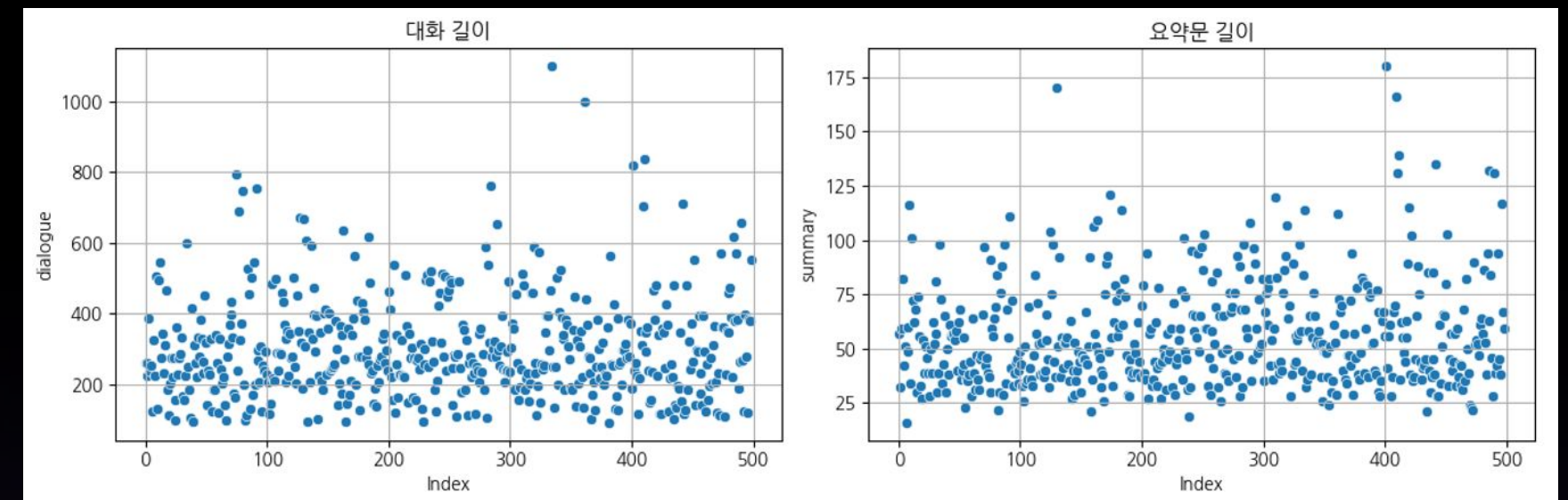
# 데이터 전처리

: 간단한 데이터 확인 (Valid)

- 446개의 토픽 중 상위 30여개만 확인함
- 직업 면접 9개, 쇼핑 7개, 일상 대화 4개 등 존재함



- digit82/kobart-summarization 기준으로 토큰 길이 확인
- 대화 길이는 600 밑으로 많이 분포해있음
- 요약문 길이는 100 밑으로 많이 분포해있음



Upstage AI Lab

# 데이터 전처리

: 좌우 공백 제거

좌우 공백 제거 전

	index	dialogue	summary	topic
0	count	12457.000000	12457.000000	12457.000000
1	mean	206.195874	40.671430	2.434455
2	std	98.807088	16.983884	1.517949
3	min	46.000000	8.000000	1.000000
4	25%	144.000000	29.000000	2.000000
5	50%	188.000000	37.000000	2.000000
6	75%	252.000000	49.000000	3.000000
7	max	1115.000000	198.000000	94.000000

좌우 공백 제거 후

	index	dialogue	summary	topic
0	count	12457.000000	12457.000000	12457.000000
1	mean	206.161516	40.663242	2.422092
2	std	98.795178	16.985430	1.273105
3	min	46.000000	8.000000	1.000000
4	25%	144.000000	29.000000	2.000000
5	50%	188.000000	37.000000	2.000000
6	75%	252.000000	49.000000	3.000000
7	max	1115.000000	198.000000	53.000000



# 데이터 전처리

: 자음, 모음 대체

❑ 대화에서 모음만으로 구성된 경우

❑ 편집장이 제1 다른 잡지에서 편집자로 일했던 경험이 있다는 걸 듣고,

❑ 대화에서 자음만으로 구성된 경우

❑ 나는 CD 가게에 들어가서 CD를 보는 척했ㄴ거든.

❑ 이것은 19세기 초 배경으로 설정된 로맨스 소설이에요.

❑ 속았어! ㅋㅋ..

❑ 너는 아직표알맞는 사람을 만나지 못했을 뿐이고,



- 오타 수정

- ㅋㅋ 는 웃기다로 대체

❑ 요약문에서 모음만으로 구성된 경우

❑ 머라이어 H리를 들어본 적이 있는지

# 데이터 전처리

: 대화 중 괄호 안에 있는 내용 제거

#Person1#: 실례합니다 이 버스는 센트럴 파크로 가나요?  
#Person2#: 네 이 버스가 맞아요.  
#Person1#: 센트럴 파크에 도착하면 알려주실 수 있나요?  
#Person2#: 걱정하지 마세요. 정거장을 알려드릴게요.  
#Person1#: (몇 분 후.) 다음 정거장에서 내려야 하나요?  
#Person2#: 아니요 걱정하지 마세요. 도착하면 알려드릴게요.  
#Person1#: 오래 걸리나요?  
#Person2#: 아니요 그렇게 오래 걸리지 않아요. 두 정거장 더 가면 내리시면 돼요 선생님.  
#Person1#: 알겠습니다. 감사합니다.  
#Person2#: 천만에요.

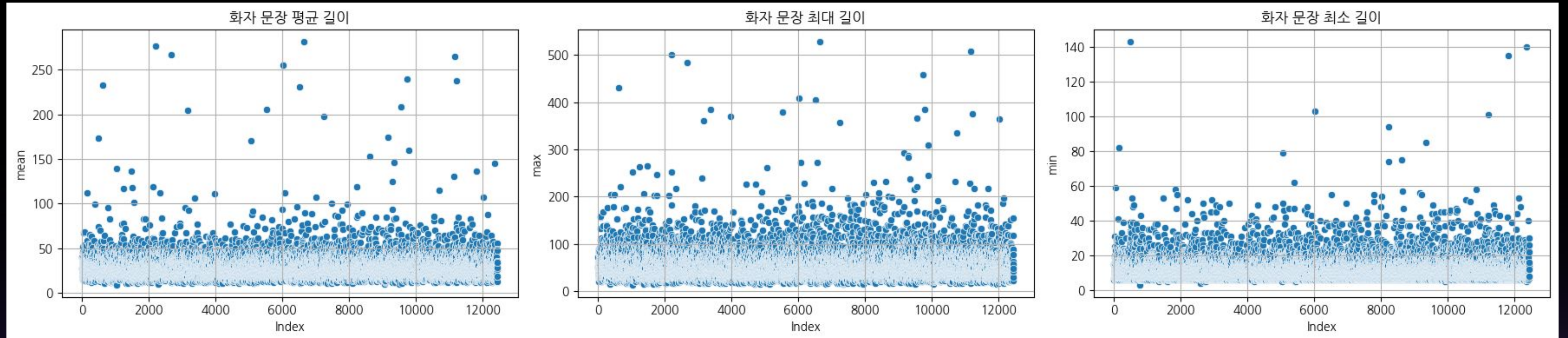
-----  
#Person1#: 실례합니다 이 버스는 센트럴 파크로 가나요?  
#Person2#: 네 이 버스가 맞아요.  
#Person1#: 센트럴 파크에 도착하면 알려주실 수 있나요?  
#Person2#: 걱정하지 마세요. 정거장을 알려드릴게요.  
#Person1#: 다음 정거장에서 내려야 하나요?  
#Person2#: 아니요 걱정하지 마세요. 도착하면 알려드릴게요.  
#Person1#: 오래 걸리나요?  
#Person2#: 아니요 그렇게 오래 걸리지 않아요. 두 정거장 더 가면 내리시면 돼요 선생님.  
#Person1#: 알겠습니다. 감사합니다.  
#Person2#: 천만에요.



# 데이터 전처리

: 화자 문장 긴 것 삭제

- ❑ 데이터마다 화자별 문장 길이를 mean, max, min으로 확인함
- ❑ 화자 문장을 최대 길이로 확인해보니, 300 이하에 많이 분포하는 것을 알게 됨
- ❑ 화자 문장 최대 길이가 300 을 넘어가는 데이터는 삭제



# 데이터 전처리

: 동일한 화자가 연속으로 말하는 경우

```
1 num = 756
2 print(train.iloc[num, 1])
3 print('-'*30)
4 print(train.iloc[num, 2])
```

✓ 0.0s

```
#Person1#: 안녕, 메이슨!
#Person1#: 오, 안녕, 피비!
#Person2#: 오늘 수업에서 너에게 대가족이 있다고 들었어.
#Person1#: 맞아. 나는 다섯 형제와 여섯 자매가 있어.
#Person2#: 와! 엄청난 대가족이네! 너는 맏이야, 아니면 막내야?
#Person1#: 둘 다 아니야. 나는 세 번째로 많아.
```

-----  
메이슨이 피비에게 그가 대가족이 있다고 말한다.

아래 두 가지 조건을 고려하여 수정

1. Person1, Person2, ...가 번갈아가며, 대화하는 것이 아니라 연속으로 말함
2. 대화문을 읽어봤을 때, 수정하지 않으면, 부자연스러울 것 같음



# 데이터 전처리

: 잘못된 화자 이름 수정

- ❑ Person이 아니라 사람인 경우

#사람1# 반기 시 계정 갱신 서비스는 만  
#Person2#: 그게 저한테 딱 맞아요.

- ❑ Person에 숫자가 없는 경우

#Person1#: 지미의 성적표가 오늘 왔어.  
#Person#: 한번 봐볼까. 이게 뭐야? 성적

- ❑ #이 없는 경우

Person1#: 이번 여름에 당신의  
#Person2#: 감사합니다. 저는

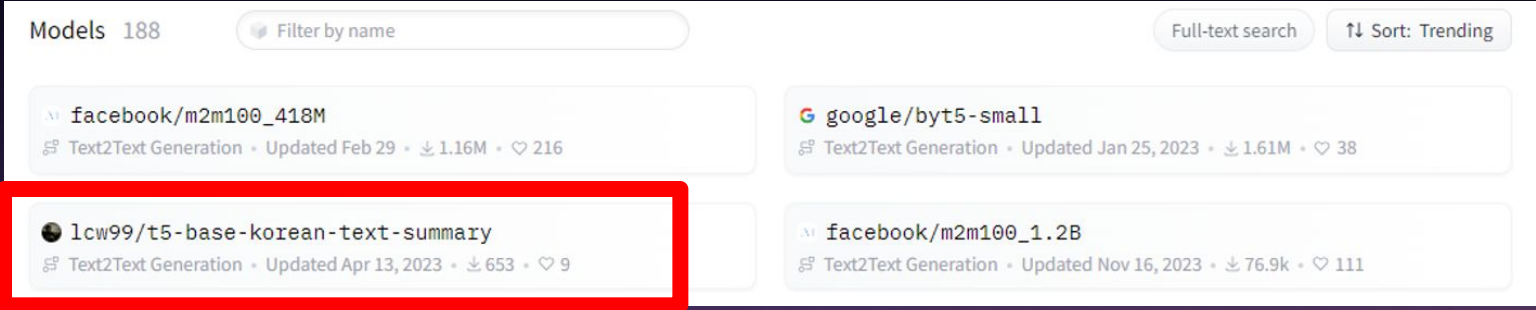
- ❑ Person이 표기되지 않은 경우

#Person1#: 네, 코닥 컬러 필름 두 롤 주세요.  
#여기 있습니다.

03

# 모델링



모델	선정 방식
Baseline digit82/kobart-summarization	다양한 전처리와 Config 수정을 통해 실험
paust/pko-t5-large	<div>1. kobart 모델과 비교해보기 위해서 선정함</div> <div>2. t5 small, base 이후에 더 큰 모델로 실험해보려고 선정함</div>
lcw99/t5-base-korean-text-summary	<div>1. Tasks: Text2Text Generation</div> <div>2. Libraries: PyTorch, Transformers</div> <div>3. Languages: Korean</div> <div>위 3가지 기준으로 찾았을 때 나온 모델로 선정함</div> <div></div>

# 모델링

: 실험 정리 (5개로 추림)

모델	전처리	Rouge-1, 2, L (Local)	Rouge-1, 2, L (LB)	Final Result
Baseline digit82/kobart-summarization	無	0.3733 0.1385 0.3593	0.5121 0.3159 0.4245	41.7506
Baseline digit82/kobart-summarization	train dialogue special token 오타 수정	0.3780 0.1381 0.3614	0.5157 0.3208 0.4262	42.0899
paust/pko-t5-large	<ul style="list-style-type: none"><li>- 좌우 공백 제거</li><li>- 자음, 모음 대체</li><li>- 동일한 화자 연속 대화 수정</li><li>- 잘못된 화자 이름 수정 (예: #사람1)</li></ul>	0.3239 0.0982 0.3150	0.5288 0.3343 0.4401	43.4391
lcw99/t5-base-korean-text-summary	<ul style="list-style-type: none"><li>- 좌우 공백 제거</li><li>- 자음, 모음 대체</li></ul>	0.3333 0.1004 0.3223	0.5243 0.3220 0.4277	42.4679
lcw99/t5-base-korean-text-summary	<ul style="list-style-type: none"><li>- 좌우 공백 제거</li><li>- 자음, 모음 대체</li><li>- 동일한 화자 연속 대화 수정</li><li>- 잘못된 화자 이름 수정</li><li>- 화자 문장이 특정값 이상으로 긴 데이터 삭제</li><li>- 괄호 안에 있는 내용 제거</li></ul>	0.3188 0.0921 0.3090	0.5232 0.3193 0.4268	42.3087



04

# 인사이트 및 회고

# 인사이드 및 회고

## Point 1

### 성능 향상에 도움이 된 전처리

- ❑ 좌우 공백 제거
- ❑ 자음, 모음 대체
- ❑ 동일한 화자 연속 대화 수정
- ❑ 잘못된 화자 이름 수정

## Point 2

### 큰 모델보다 '전처리'에 집중

- ❑ 모델이 커질수록 학습 시간이 너무 길어져서, 모델을 가볍게 쓰되 데이터의 품질 개선에 신경을 많이 씀.

## Point 3

### '서버 용량 초과 사용' 자주 발생

- ❑ 서버가 터져서 코드가 날아간 것이 큰 충격이었음. 이 이후론 로컬 및 깃허브에 코드를 습관적으로 백업함.



# 인사이드 및 회고

: 소감

---

이강권    Hugging Face 사용법을 익히게 되어 유익한 시간이었다.

---

이승현    쉽지 않다.

---

한지승    cuda out of memory 너무 싫다

---

홍재민    NLP 경진대회에 어떻게 접근해야 하는지 알게 되었다.

---

Life-Changing Education

감사합니다.

---