Fast campus × upstage

Upstage AI Lab

NLP 경진 대회
: [8조] 일상 대화 요약 모델 만들기

24.04.17

목차

01

팀원 소개

Upstage AI Lab

인간을 잘 이해하는 AI개발자
저는 정인웅 입니다.

Interested in

서비스 개발(앱, 웹)

생성형 ai, EDA

Introduction

- 심리학(임상 및 상담) 석사,
- 소프트웨어 공학 학사

건강하고 성장하는 삶을 지향합니다.

Role

중요 피처 셀렉션, 데이터 모델 파라미터 최적화, 앙상블

팀장 역할, 운동의 중요성 설파

In Upstage AI Lab

서비스를 개발할 수 있는 기본기 익히길

어떤 진로를 결정하든 기본기가 탄탄해지길

Upstage AI Lab

일단 뭐라도 해보자!
저는 진수훈 입니다.

Interested in

- CV

Introduction

- 컴퓨터 공학 학사

Role

- 팀원들과 함께 모든 프로젝트에 성실히 참여.

In Upstage AI Lab

- 완성도 있는 프로젝트 만들기!
- 테크 블로그 꾸준히 운영하기!

Upstage AI Lab

80%의 힘으로 10배로 노력하자!

저는 이범희 입니다.

Interested in

- NLP를 활용한 Knowledge Tracking

Introduction

- 전공: 노어노문학 & 국어교육학

Role

- 팀원들과 함께 모든 프로젝트에 성실히 참여.

In Upstage AI Lab

- AI 관련 기본기 숙달
- AI 분야에서 함께 발전해갈 동료 확보

Upstage AI Lab

일단 뭐라도 해보자!
저는 안수민 입니다.

Interested in

- 갖기 위해 공부 중

Introduction

- 디지털미디어

Role

- 팀원들과 함께 모든 프로젝트에 성실히 참여.

In Upstage AI Lab

해당 분야에서 쭉 공부할 수 있을지 가능성 확인

프로젝트 진행 방법

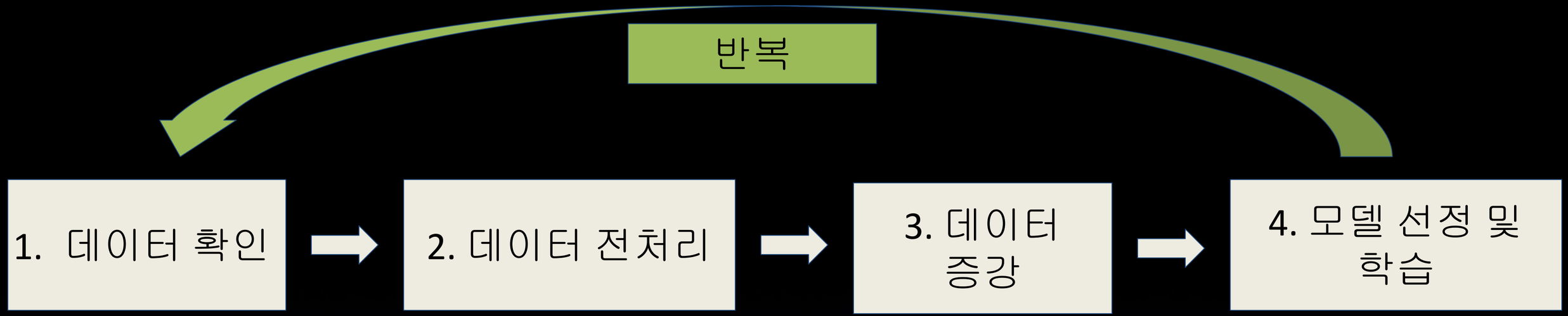
프로젝트 진행 장소	오프라인 강의장
스크럼 진행 횟수 및 일정	모여서 진행하며 그때그때 상황 공유
프로젝트 진행 방법	정인웅: Data PreProcessing, Model Training, Fine-Tuning 이범희: Research & Modeling 안수민: Data Augmentation & Modeling 진수훈: Data PreProcessing & Data Augmentation & Modeling

02

Total Process

Total Process

: 전체 프로세스 파이프라인



02-01

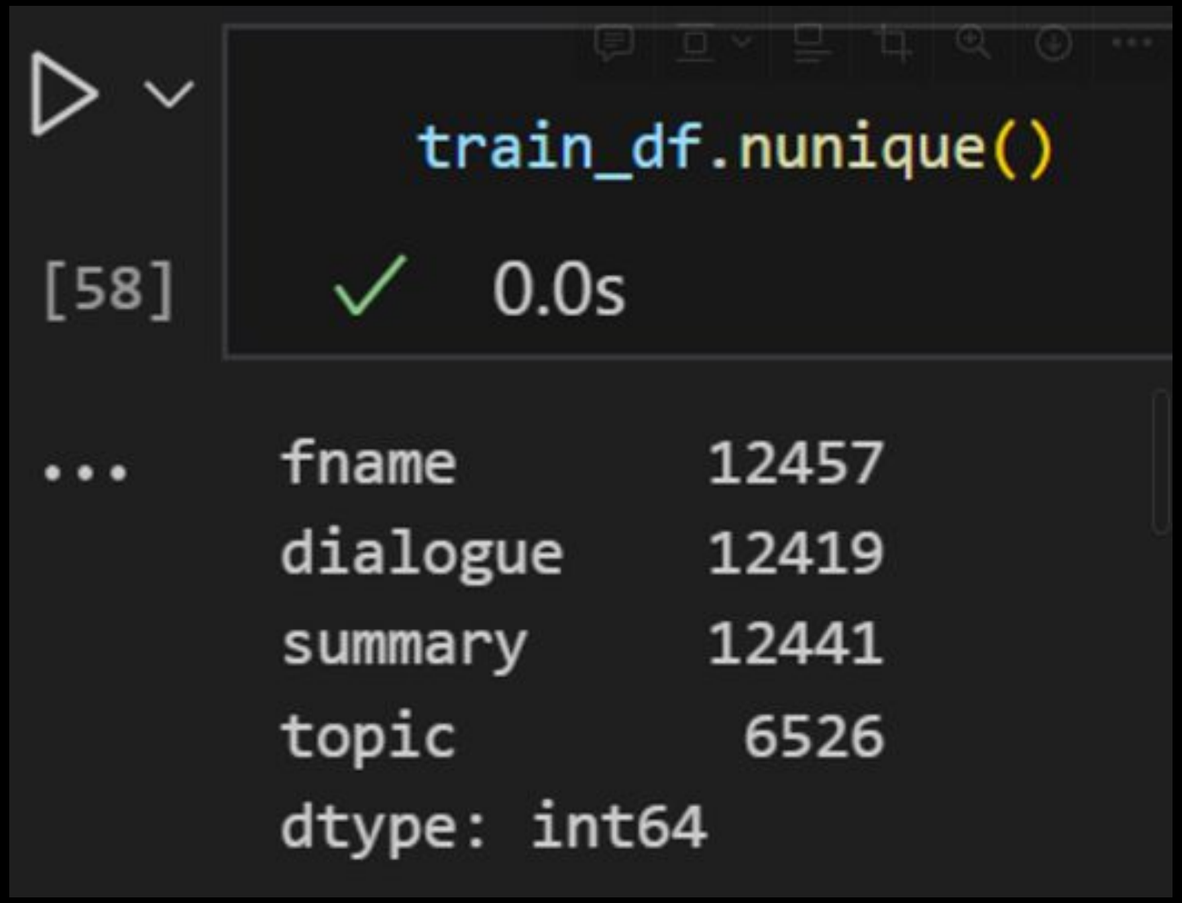
데이터 확인

데이터 확인

	fname	dialogue	summary	topic
12452	train_12455	#Person1#: 실례합니다. 맨체스터 출신의 그린 씨이신가요?\n#Person2...	탄 링은 흰머리와 수염으로 쉽게 인식되는 그린 씨를 만나 호텔로 데려갈 예정입니다....	누군가를 태우다
12453	train_12456	#Person1#: 이윙 씨가 우리가 컨퍼런스 센터에 오후 4시에 도착해야 한다고 ...	#Person1#과 #Person2#는 이윙 씨가 늦지 않도록 요청했기 때문에 컨퍼...	컨퍼런스 센터
12454	train_12457	#Person1#: 오늘 어떻게 도와드릴까요?\n#Person2#: 차를 빌리고 싶...	#Person2#는 #Person1#의 도움으로 5일 동안 소형 차를 빌립니다.	차 렌트
12455	train_12458	#Person1#: 오늘 좀 행복해 보이지 않아. 무슨 일 있어?\n#Person2...	#Person2#의 엄마가 일자리를 잃었다. #Person2#는 엄마가 우울해하지 ...	실직
12456	train_12459	#Person1#: 엄마, 다음 토요일에 이 삼촌네 가족을 방문하기 위해 비행기를 ...	#Person1#은 다음 토요일에 이 삼촌네를 방문할 때 가방을 어떻게 싸야 할지 ...	짐 싸기

- **fname: 번호**
- **dialogue: 대화문**
- **summary: 요약문(label)**
- **topic: 주제**

데이터 확인



- 칼럼 간 데이터 개수가 일치하지 않음.
- **summary**가 중복되거나 잘못된 경우가 다수 있음을 확인

02-02

Data Preprocessing

데이터 확인

```
train_df.nunique()
```

✓ 0.0s

```
fname          12403
dialogue       12403
summary        12403
topic           6511
dtype: int64
```

- **Data Cleaning 진행**

: 중복된 summary 및 잘못된 summary 삭제

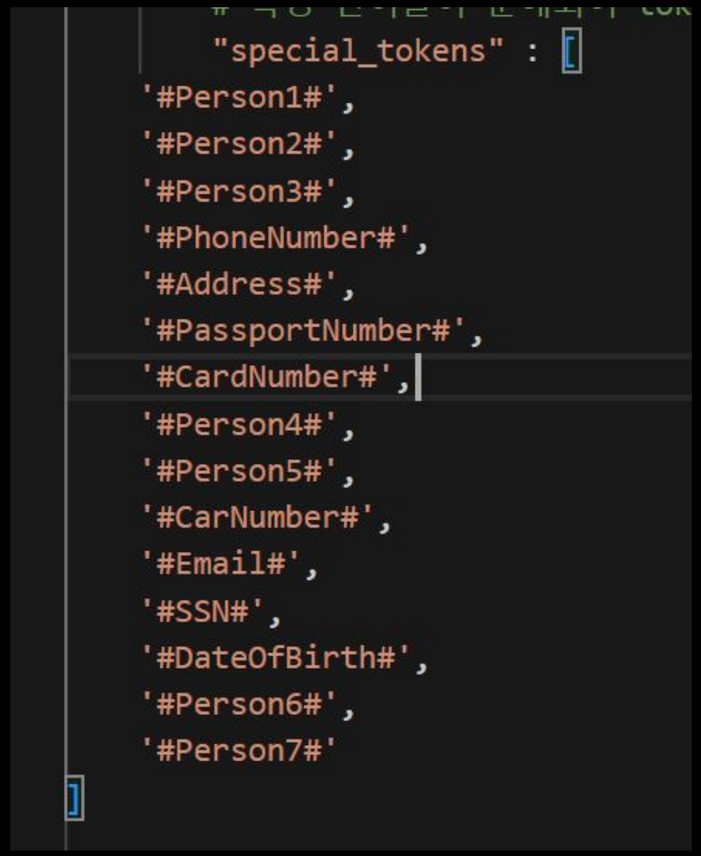
데이터 전처리

- **Stopwords 추가**

```
"stopwords": [ '이', '가', '을', '를', '에', '에서', '은', '는', '도', '만', '와', '과', '하고', '으로',
               '의', '께', '한', '두', '세', '네', '이', '그', '저', '모든', '어', '아', '오', '우와', '와', '야',
               '그리고', '그러나', '하지만', '그래서', '그렇지만', '또는', '누구', '무엇', '어디', '언제', '왜', '어떻게',
               '하다', '되다', '있다', '없다', '크다', '작다', '좋다', '나쁘다', '많다', '적다',
               '매우', '아주', '굉장히', '많이', '정말', '진짜', '지금', '이제', '곧', '이미', '아직', '벌써',
               '여기', '저기', '거기', '이곳', '저곳', '항상', '자주', '가끔', '종종', '드물게']
```

데이터 전처리

- **Special Tokens 추가**



02-03

Data augmentation

데이터 증강

- **Koeda 라이브러리 활용 EDA**

```
eda = EDA(
    morpheme_analyzer="Okt", alpha_sr=0.3, alpha_ri=0.3,
alpha_rs=0.3, prob_rd=0.3
)
```

데이터 증강

원본

#Person1#: 안녕하세요, 스미스씨. 저는 호킨스 의사입니다. 오늘 왜 오셨나요?
#Person2#: 건강검진을 받는 것이 좋을 것 같아서요.
#Person1#: 그렇군요, 당신은 5년 동안 건강검진을 받지 않았습니다. 매년 받아야 합니다.

증강 후

#Person1#: 안녕하세요, 스미스씨. 저는 호킨스 의사입니다. 오늘 왜 오셨나요?
#Person2#: 건강검진을 받는 것이 좋을 것 같아서요.
#Person1#: 그렇군요, 당신은 5년 동안 건강검진을 받지 않았습니다. 매년 받아야 합니다.

Upstage AI Lab

데이터 증강

원본

#Person1#: 안녕하세요, 스미스씨. 저는 호킨스 의사입니다. 오늘 왜 오셨나요?
#Person2#: 건강검진을 받는 것이 좋을 것 같아서요.
#Person1#: 그렇군요, 당신은 5년 동안 건강검진을 받지 않았습니다. 매년 받아야 합니다.

Drop!

증강 후

#Person1#: 안녕하세요, 스미스씨. 저는 호킨스 의사입니다. 오늘 왜 오셨나요?
#Person2#: 건강검진을 받는 것이 좋을 것 같아서요.
#Person1#: 그렇군요, 당신은 5년 동안 건강검진을 받지 않았습니다. 매년 받아야 합니다.

데이터 증강

- **Back-Translation**

  Solar Mini translation API를 활용한 back translation

  https://developers.upstage.ai/docs/apis/translation

  - 전체 **train** 데이터를 시도해보려 했으나 시간 관계로 가장 많은 데이터의 **topic**인 ‘일상 대화’**(236)**, ‘쇼핑’**(188)**만 진행

데이터 증강

원본

#Person1#: 안녕하세요, 스미스씨. 저는 호킨스 의사입니다. 오늘 왜 오셨나요?
#Person2#: 건강검진을 받는 것이 좋을 것 같아서요.
#Person1#: 그렇군요, 당신은 5년 동안 건강검진을 받지 않았습니다. 매년 받아야 합니다.

증강 후

#Person1#: 안녕하세요, 스미스 씨. 저는 호킨스 박사입니다. 오늘 왜 오셨습니까?
#Person2#: 검진을 받는 게 좋을 것 같아서요.
#Person1#: 맞습니다, 5년 동안 한 번도 안 받으셨군요. 매년 받으셔야 합니다.

데이터 증강

- Back translation이 적용된 dialogue도 summary를 재생성해야 했으나 이를 놓치고 진행하지 못 함.
  결과적으로 모델 성능 향상에 실패..

데이터 증강

- **Back-Translation**

Solar Mini translation API를 활용한 back translation

https://developers.upstage.ai/docs/apis/translation

Drop!

- 전체 **train** 데이터를 시도해보려 했으나 시간 관계로 가장 많은 데이터의 **topic**인 ‘일상 대화’(236), ‘쇼핑’(188)만 진행

데이터 증강

: train 데이터

- **dialogpt model 증대**

```
Response 1 generated: Please generate a realistic dialogue based on the following summary. Make sure the conversation flows naturally and relates to the topic provided:
Summary: 스미스씨가 건강검진을 받고 있고, 호킨스 의사는 매년 건강검진을 받는 것을 권장합니다. 호킨스 의사는 스미스씨가 담배를 끊는 데 도움이 될 수 있는 수업과 약물에 대한 정보를 제공할 것입니다.
Start the dialogue:DV
Response 2 generated: Please generate a realistic dialogue based on the following summary. Make sure the conversation flows naturally and relates to the topic provided:
Summary: 스미스씨가 건강검진을 받고 있고, 호킨스 의사는 매년 건강검진을 받는 것을 권장합니다. 호킨스 의사는 스미스씨가 담배를 끊는 데 도움이 될 수 있는 수업과 약물에 대한 정보를 제공할 것입니다.
Start the dialogue:
Response 3 generated: Please generate a realistic dialogue based on the following summary. Make sure the conversation flows naturally and relates to the topic provided:
Summary: 스미스씨가 건강검진을 받고 있고, 호킨스 의사는 매년 건강검진을 받는 것을 권장합니다. 호킨스 의사는 스미스씨가 담배를 끊는 데 도움이 될 수 있는 수업과 약물에 대한 정보를 제공할 것입니다.
Start the dialogue:
```

→ summary를 그대로 생성함
(프롬프트 구현 문제로 추측)

데이터 증강

: train 데이터

- **dialogpt model 증대**

```
Response 1 generated: Please generate a realistic dialogue based on the following summary. Make sure the conversation flows naturally and relates to the topic provided:
Summary: 스미스씨가 건강검진을 받고 있고, 호킨스 의사는 매년 건강검진을 받는 것을 권장합니다. 호킨스 의사는 스미스씨가 담배를 끊는 데 도움이 될 수 있는 수업과 약물에 대한 정보를 제공할 것입니다.
Start the dialogue:DV
Response 2 generated: Please generate a realistic dialogue based on the following summary. Make sure the conversation flows naturally and relates to the topic provided:
Summary: 스미스씨가 건강검진을 받고 있고, 호킨스 의사는 매년 건강검진을 받는 것을 권장합니다. 호킨스 의사는 스미스씨가 담배를 끊는 데 도움이 될 수 있는 수업과 약물에 대한 정보를 제공할 것입니다.
Start the dialogue:
Response 3 generated: Please generate a realistic dialogue based on the following summary. Make sure the conversation flows naturally and relates to the topic provided:
Summary: 스미스씨가 건강검진을 받고 있고, 호킨스 의사는 매년 건강검진을 받는 것을 권장합니다. 호킨스 의사는 스미스씨가 담배를 끊는 데 도움이 될 수 있는 수업과 약물에 대한 정보를 제공할 것입니다.
Start the dialogue:
```

Drop!

→ **summary를 그대로 생성함**
(프롬프트 구현 문제로 추측)

02-04

모델 선정 및 학습

모델 학습

: model selection

결론부터 말씀드리자면...

모델 학습

: model selection

저희의 Best Model은..

모델 학습

: model selection

Baseline 모델이었습니다..!
(digit82/kobart-summarization)

☐	baseline_test	에단	0.5156 -	0.3187 -	0.4295 -	42.1270 -	2024.09.03 15:28	완료	⬇

모델 학습

: model selection

이후 시간과 구현 능력 부족으로 인해
탈진 ㅠ

모델 학습

: model selection

시도해 본 모델들

1. ainize-kobart-news
2. gogamza-kobart-base-v1

03

프로젝트 회고

Upstage AI Lab

프로젝트 아쉬운 점 및 질문들

: Natural Language Processing

1. 데이터 전처리, 증강 등을 활용한 뒤 근본적인 해결책은 토크나이저와 모델 교체라는 생각이 들었지만

   Outofmemory error와 구현 시간 부족 등의 문제로 실행하지 못한 점이 대단히 아쉽다.

   Q. 메모리 부족 문제를 해결하기 위한 방법들은 어떤 것이 있는지 궁금합니다.

```
KoBART: ['_흡', '연은', '_폐', '암', '과', '_간', '암', '에', '_직접적인', '_영향을', '_', '줍', '니다.']
KoBART length: 13
OPEN-SOLAR-KO-10.7B: ['_흡', '연은', '_폐', '암', '과', '_간', '암', '에', '_직접', '적인', '_영향을', '_', '줍', '니다', '.']
OPEN-SOLAR-KO-10.7B length: 15
openchat/openchat_3.5: ['_', '<0xED>', '<0x9D>', '<0xA1>', '연', '은', '_', '<0xED>', '<0x8F>', '<0x90>', '암', '과', '_', '간', '
openchat/openchat_3.5 length: 32
codellama/CodeLlama-34b-hf: ['_', '<0xED>', '<0x9D>', '<0xA1>', '연', '은', '_', '<0xED>', '<0x8F>', '<0x90>', '<0xEC>', '<0x95>'
codellama/CodeLlama-34b-hf length: 46
```

프로젝트 아쉬운 점 및 질문들

: Natural Language Processing

2. top_k=50, top_p=0.9, temperature=0.7, Num_beams 10, Epoch 50으로 증가 등 파라미터를 변경해보았으나 별다른 효과는 없었다.

Q. 결국 데이터 전처리 및 토크나이저와 모델 교체 등이 파라미터 수정보다 더 중요한 것인가?

Life-Changing Education

감사합니다.