

TagMyBookmark!

북마크 자동 태깅 서비스

Bookmark Auto-tagging Service



NLP-02조 강남특공대

김기범

박희진

이주형

천소영

천재원



강



남



특



공



대

이미지 출처: (주)누리토이즈



INDEX

팀 및 팀원 소개

서비스

서비스 기획

서비스 소개 및 기대효과

서비스 사용 예시

프로젝트

프로젝트 진행현황

프로젝트 아키텍처

프론트엔드 소개

백엔드 소개

데이터 소개

모델 소개



강남특공대: 강남에서 매주 오프라인 모임을 진행하는 엘리트들





팀원 역할

이름	수행 역할
김기범	Django 셋업, 백엔드 기능 구현, REST API 구현, 데이터베이스 구축, 데이터 레이블링
박희진	크롤링, 데이터 수집/분석/정제, 프롬프트 실험
이주형	크롬 익스텐션 북마크 정보 추출, 백엔드 DB 저장 및 조회 기능 개발, REST API 구현 및 개선
천소영	데이터 생성, 프롬프트 실험, 크롬 익스텐션 및 서비스 웹페이지 개발
천재원	AI 태깅 모델 개발, 평가지표 고안, 백엔드 연동, 크롬 익스텐션 및 서비스 웹페이지 개발

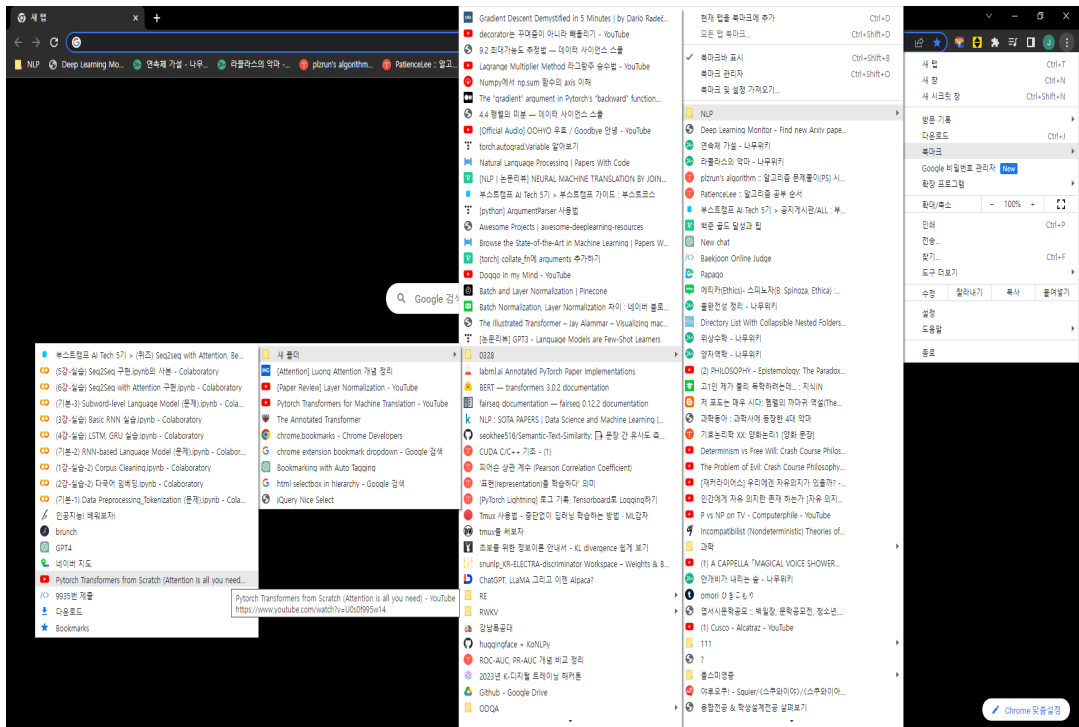


북마크, 어떻게 사용하고 계신가요?





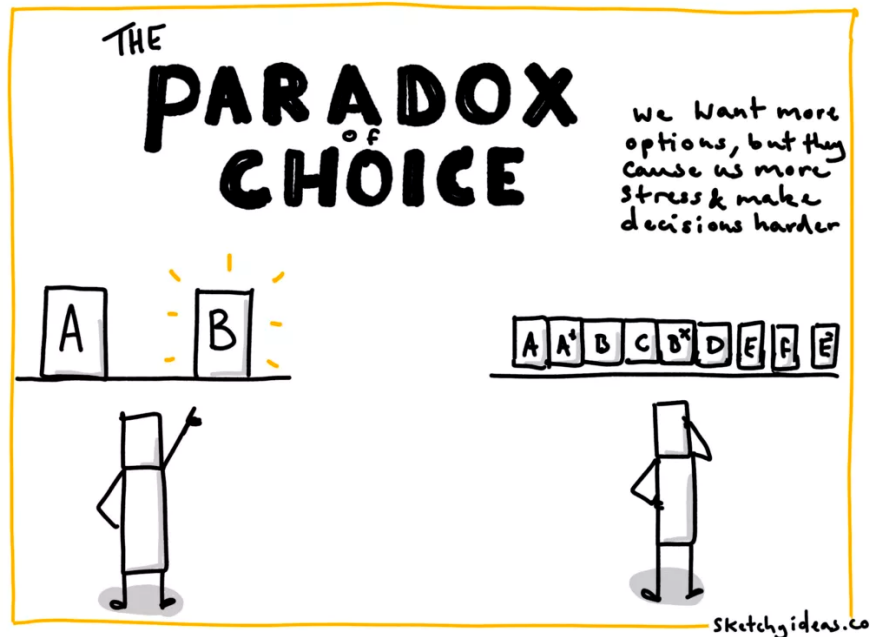
아마도...



눈 깜짝할 새에 쌓여버린 북마크



그때 그 북마크, 어디에 저장했더라?



어디에 정리하죠, 내 북마크?



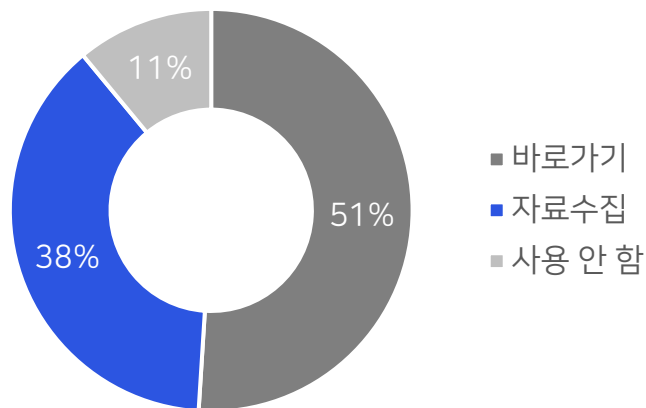
파일 구조로 정리하는 게 최선일까?
자동으로 정리해 주면 좋을 텐데!

북마크 관리, 필요하지 않으세요?

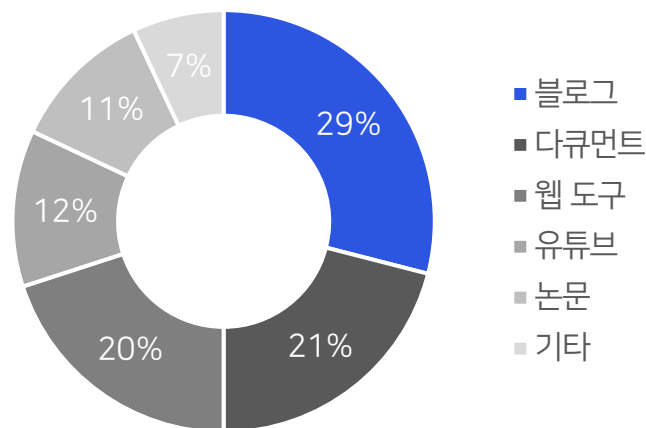


실제 사람들의 목소리

북마크 사용목적

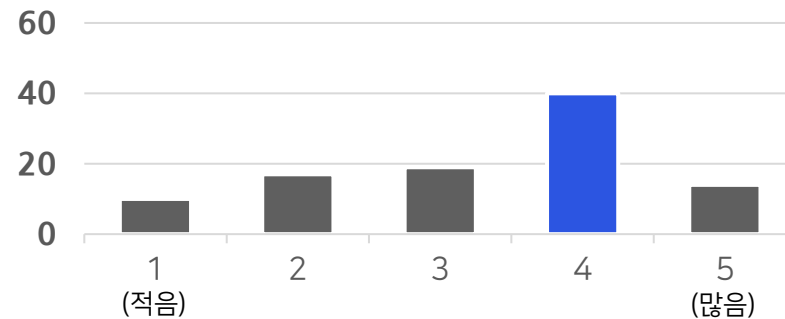


주로 저장하는 출처

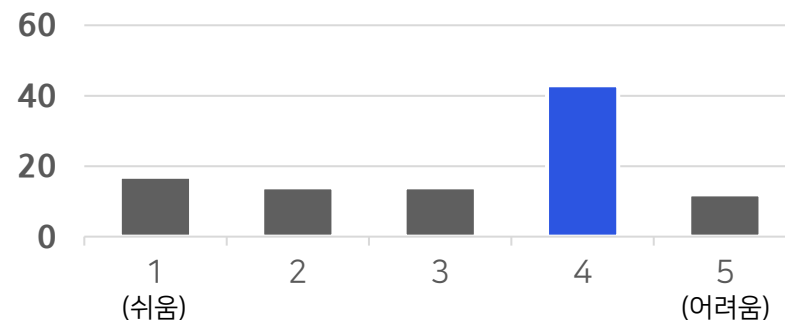


- 블로그
- 다큐먼트
- 웹 도구
- 유튜브
- 논문
- 기타

자료수집 - 북마크 찾기 위해 **해만** 경험



자료수집 - 북마크 저장 **폴더 지정** 어려움



설문 참여자 : 110명



서비스 대상

- ☹️ 북마크를 **자료 수집용**으로 사용하는
- 😞 **블로그 포스트**를 북마킹 하는
- 😞 북마크를 저장할 **폴더 지정이 어려운**
- 😞 북마크 **검색에 어려움**을 겪는





서비스 소개



는 북마크 페이지에 대한 태그를 생성,
효과적인 페이지 관리를 돕는 서비스입니다

유관 문서 묶어보기

As-Is
다른 폴더 속 페이지 묶기 불가능

To-Be
페이지 내용 기반 n개의 태그 생성
다양한 기준으로 묶기 가능

태그기반 검색

As-Is
담아둔 북마크를 하나씩 확인
제목 / 폴더 경로의 세심한 설정 필요

To-Be
자동 생성 태그 기반 검색
쉽고 효과적인 필터링

정리 자동화

As-Is
페이지가 많아질수록 분류의 어려움 ↑
깔끔한 정리를 위한 외부 앱 사용

To-Be
많은 페이지들에 대해서도 자동 분류
크롬 익스텐션 연동 및 웹 기반 서비스



프로토타입 시연 영상



프로젝트 타임라인

	Baseline (23.07.02 - 23.07.07)	Pipeline v1.0 (23.07.08 - 23.07.23)	Pipeline v2.0 (23.07.24 - 23.07.28)
기획	유사 프로그램 서치 북마크 관련 확장 프로그램 분석	파이프라인 점검 및 기능 개선기능 개선 전체적인 서비스 이용 플로우 점검	피드백 기반 개선점 확인 사용자 및 멘토 피드백 의견 수렴
CHORE	깃헙 컨벤션 구축		
프론트엔드	크롬 익스텐션 프로토타입 크롬 북마크 추가 및 페이지 정보 전송	서비스 페이지 프로토타입 메인 웹페이지 구성 및 기능 구현	사용성 개선 UI/UX 개선 및 버그 트러블 슈팅 북마크 삭제, 제목 저장, 실시간 태그 확인 및 사용자 태그 반영
백엔드	Django 프로토타입 DB에 request 추가 및 DL 모델로 페이지 정보 전송	DB 시스템 고도화 유저정보 및 북마크 정보 저장, DL 모델로 정보 전송, 태그 검색	사용성 개선 크롬 익스텐션에서의 전달되는 정보에 맞춰 DB 업데이트
데이터 구축	1차 훈련 데이터 크롤링 블로그 1차 크롤링 Toy data 생성 블로그 텍스트 요약 모델로 요약 후 tag label 생성	Toy data 필터링 훈련에 사용 가능한 샘플만 필터링	태그 품질 지표에 따라 데이터 정제 자체 설정 태그 품질 기준에 맞추어 생성된 태그 직접 정제
메인 모델링		태그 생성 모델 구축 Toy data 이용, base 모델 훈련 및 선정	태그 생성 모델 추론 시간 및 성능 개선 양자화, 코드 리팩토링, 데이터 전처리와 PEFT 최적화 성능 지표 고안 태그 생성 task에 적합한 새로운 성능 지표



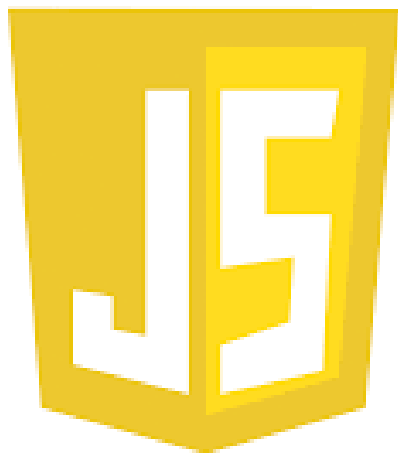
프로젝트 타임라인(To-Do)

	TMB v1.0 (23.07.02 - 23.07.28)	TMB v2.0 (23.07.28 -)
기획	사전 조사, 설문 및 사업 발표 피드백 수렴 시장조사, 북마크 이용자 설문, 발표자료 기반 피드백	실제 서비스 후 유저 피드백 수렴 별점 등 명시적 피드백 및 태그 수정 빈도 등 암묵적 피드백
CHORE	깃헙 컨벤션 구축 팀 내 개발 과정에 맞춘 깃헙 컨벤션	모델 버전 관리 기능 추가 유저의 활동으로 인해 축적되는 데이터로 새로 학습된 모델에 대한 버전관리
프론트엔드	크롬 익스텐션 / 서비스 페이지 필수 기능 구현 프로토타입 시연 영상 및 아래 발표 내용 참고	커스터마이징 / 추천 관련 기능 강화 태그 셋 커스터마이징 및 태그 / 북마크 추천
백엔드	Django 필수 기능 구현 / DB 시스템 고도화 프로토타입 시연 영상 및 아래 발표 내용 참고	데이터 분석 기능 추가 추천 태그 및 북마크 예측 알고리즘 개발
데이터 구축	데이터 20%까지 정제 후 Feeding 완료 평가 기준에 따라 직접 태그 수증	대규모 데이터셋 생성 전체 데이터셋에 대하여 정제 작업 실시
메인 모델링	태그 모델 베이스라인 구축 및 성능지표 고안 추후 추가로 학습되는 모델에 대한 지표	추론 시간 단축 및 모델 성능 지속적 개선 13B -> 7B 모델로 Distillation PPO 방법을 활용한 모델 강화학습, 선형보간으로 Context 길이 연장



기술스택

JavaScript



REST API



Hugging Face

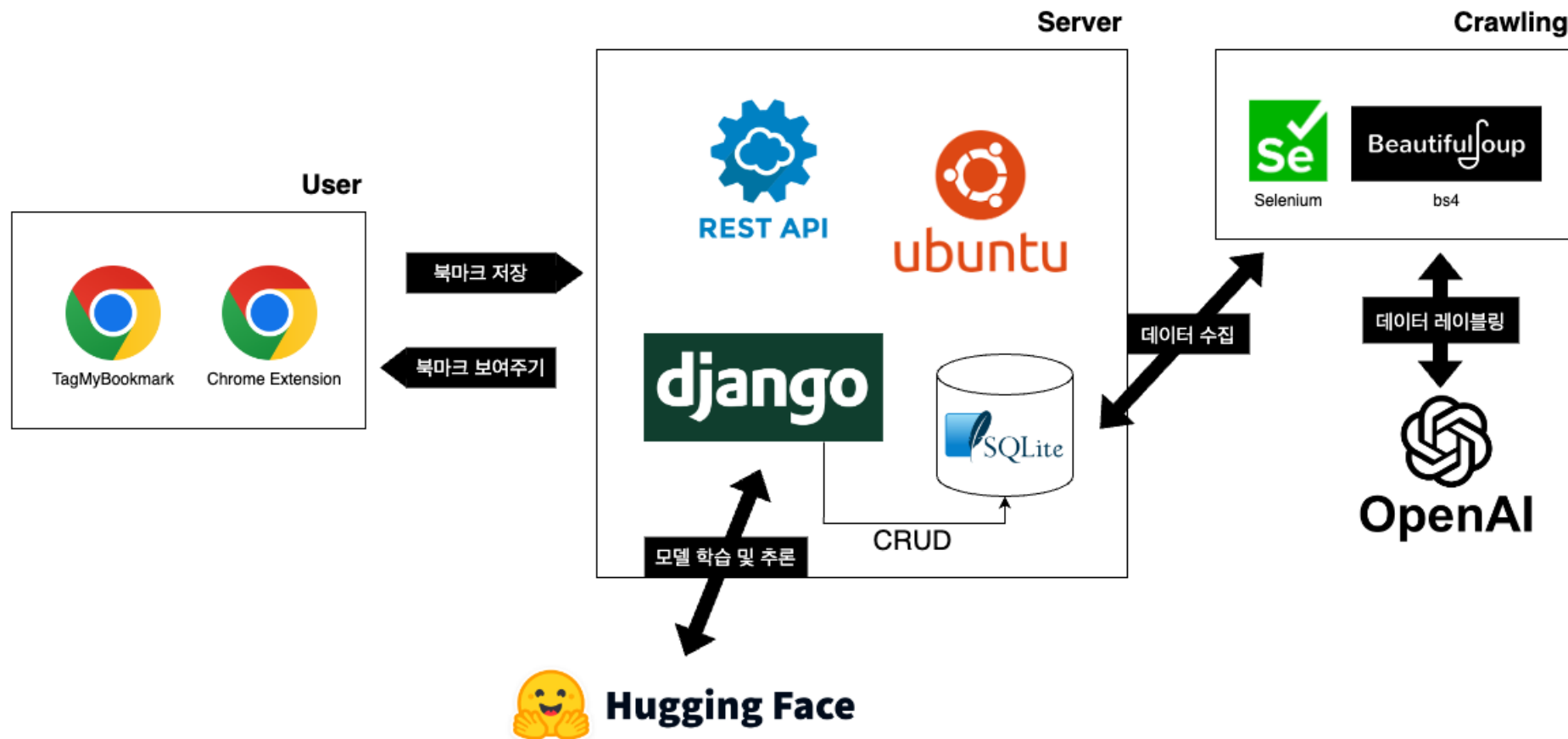
django



SQLite



프로젝트 아키텍처





프로젝트 폴더 구조

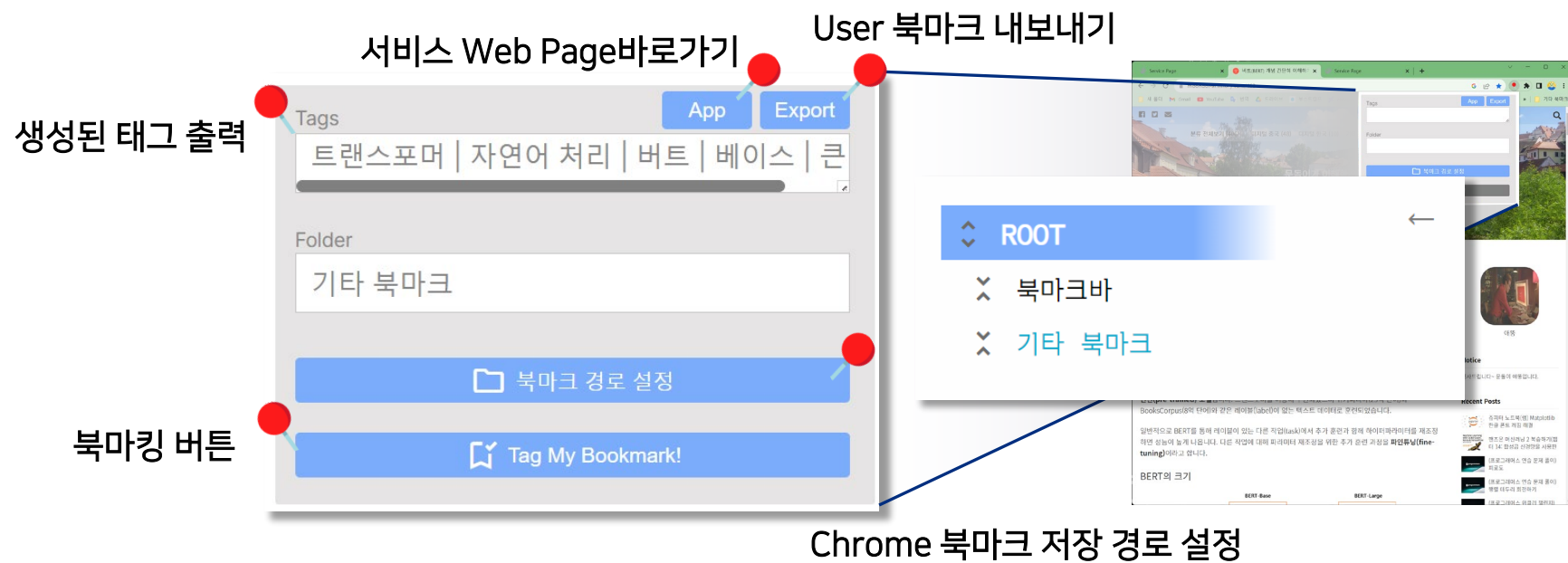
```
level3_nlp_finalproject-nlp-02
├── asset/
├── data/
├── model/
├── web/
│   ├── myapp/
│   │   ├── API/
│   │   ├── SERVICE/
│   │   ├── myapp/
│   │   ├── static/
│   │   ├── README.md
│   │   ├── db.sqlite3
│   │   └── manage.py
│   ├── extension/
│   └── README.md
└── README.md
```

- [Data/](#) : 학습 데이터 생성 관련
- [Model/](#) : 모델 학습 및 추론
- Web/ : 프론트엔드 & 백엔드 구현
 - [Extension/](#) : Chrome Extension 페이지
 - Myapp/ : Django 프로젝트 상위 폴더
 - [API/](#) : REST API 기능 구현
 - [SERVICE/](#) : 유저와 상호작용하는 웹 페이지
 - Myapp/ : Django 프로젝트 설정



크롬 익스텐션

전체 오버뷰



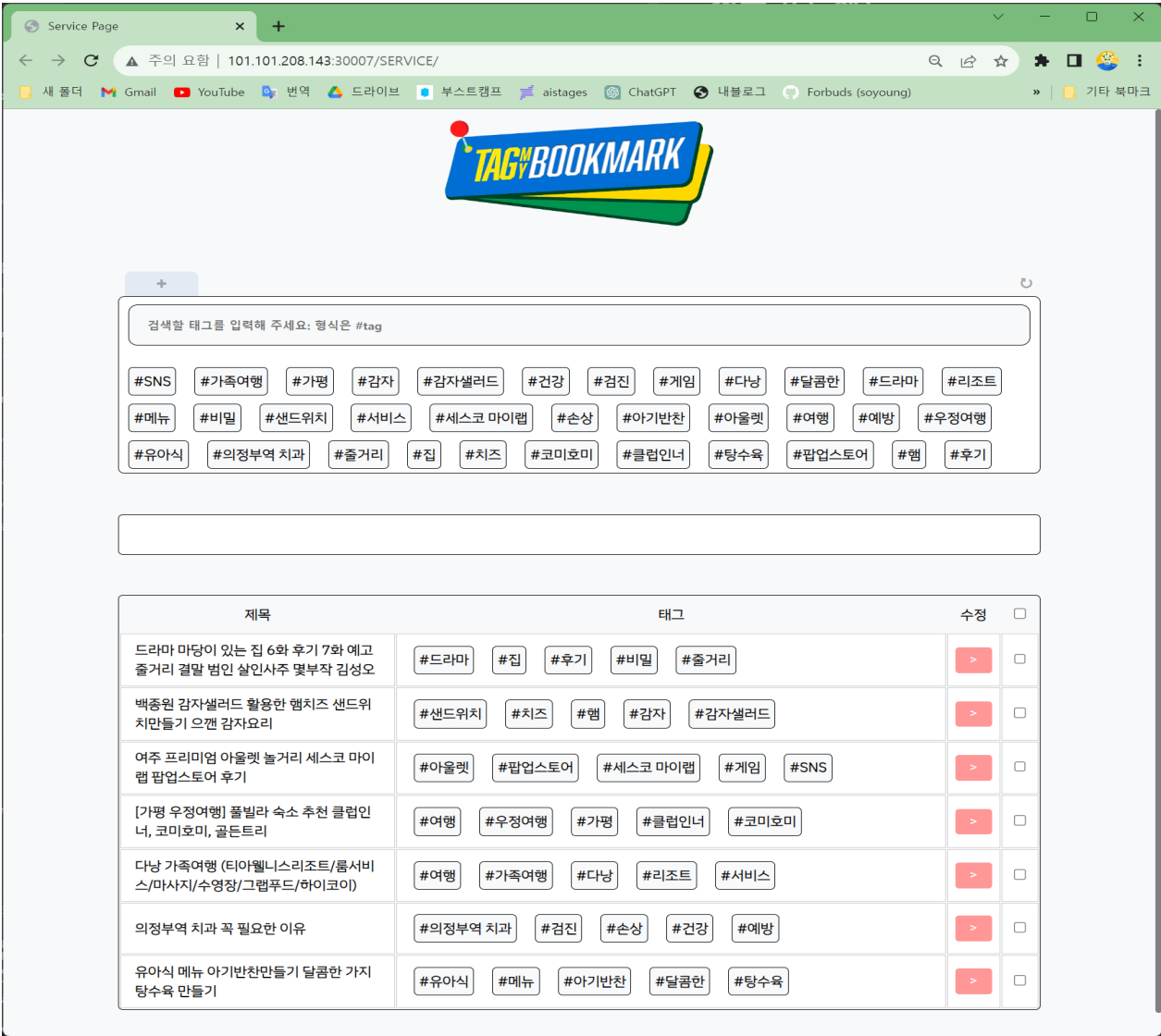


웹 서비스 페이지

전체 오버뷰

태그 선택 섹션


태깅 페이지 섹션





웹 서비스 페이지

태그 선택 섹션



즐거찾기 기능

태그 검색(Feat. 자동 완성)

클릭으로 태그 선택

선택된 태그 확인

#가

#가평

#가족여행

#메뉴

#비밀

#샌드위치

#서비스

#세스코 마이랩

#손상

#아기반찬

#아울렛

#여행

#예방

#우정여행

#유아식

#의정부역 치과

#즐거리

#집

#치즈

#코미호미

#클럽인너

#탕수육

#팝업스토어

#해외여행

#햄

#후기

#메뉴

#즐거리

제목	태그	수정	
드라마 마당이 있는 집 6화 후기 7화 예고 즐거리 결말 범인 살인사주 몇부작 김성오	#드라마 #집 #후기 #비밀 #즐거리	>	
유아식 메뉴 아기반찬만들기 달콤한 가지 탕수육 만들기	#유아식 #메뉴 #아기반찬 #달콤한 #탕수육	>	



웹 서비스 페이지

태깅 페이지 섹션

태깅 페이지 결과



클릭으로
태그 삭제

다낭 가족여행 (티아웰니스리조트/룸
서비스/마사지/수영장/그랩푸드/하이
코이)

#여행 x #가족여행 x #다낭 x #리조트 x #서비스 x

추가할 태그를 입력해 주세요: 형식은 #tag

태그 추가 입력

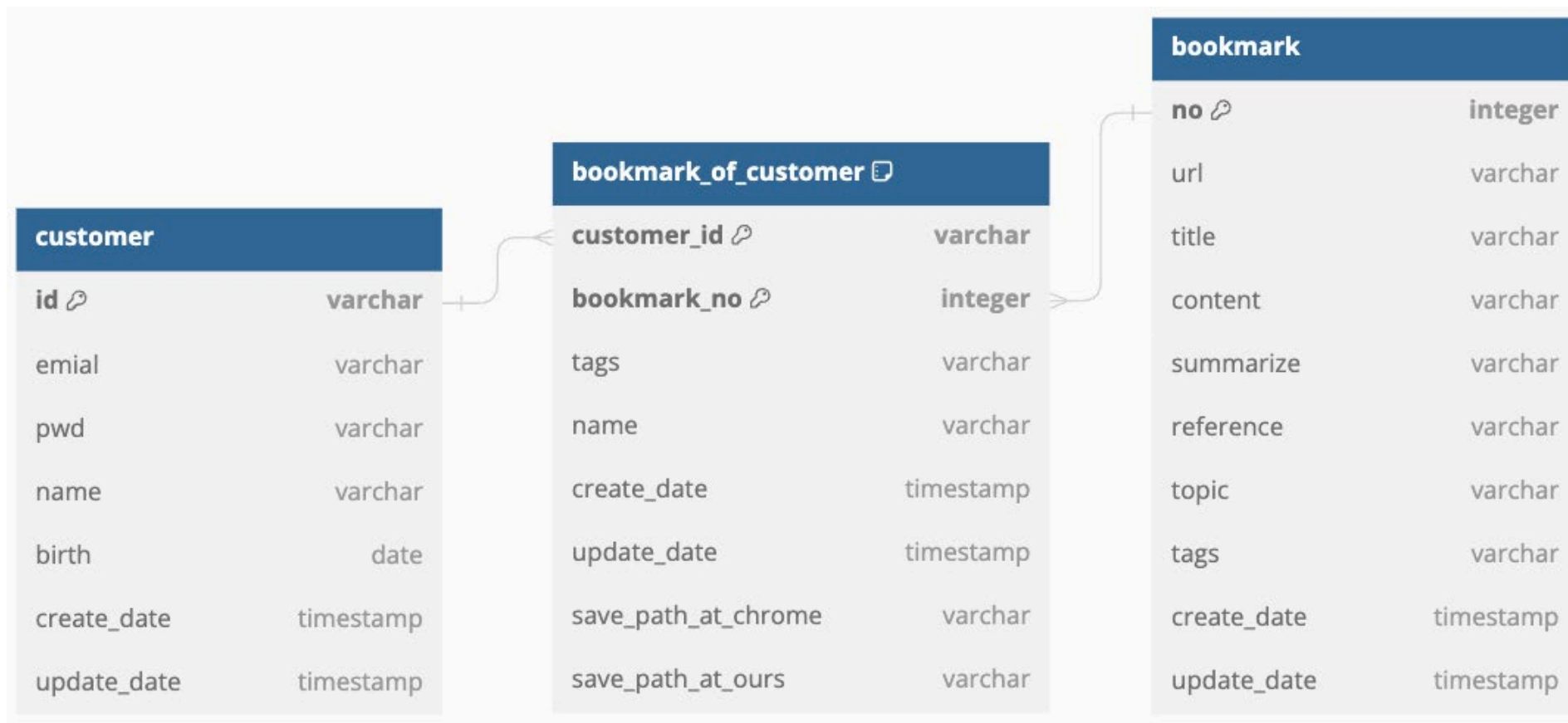
#가족여행 #다낭 #리조트

제목	태그	수정
다낭 가족여행 (티아웰니스리조트/룸서비스/마사지/수영장/그랩푸드/하이코이)	#여행 #가족여행 #다낭 #리조트 #서비스	<div><div></div><div></div><div></div><div></div><div></div></div>

수정 팝업창 열기



ERD





REST API

{URL} # 제목2임 → # 2개

- GET | POST | PATCH
- request

```
{  
  "test": 임의의 값,  
}
```

- response

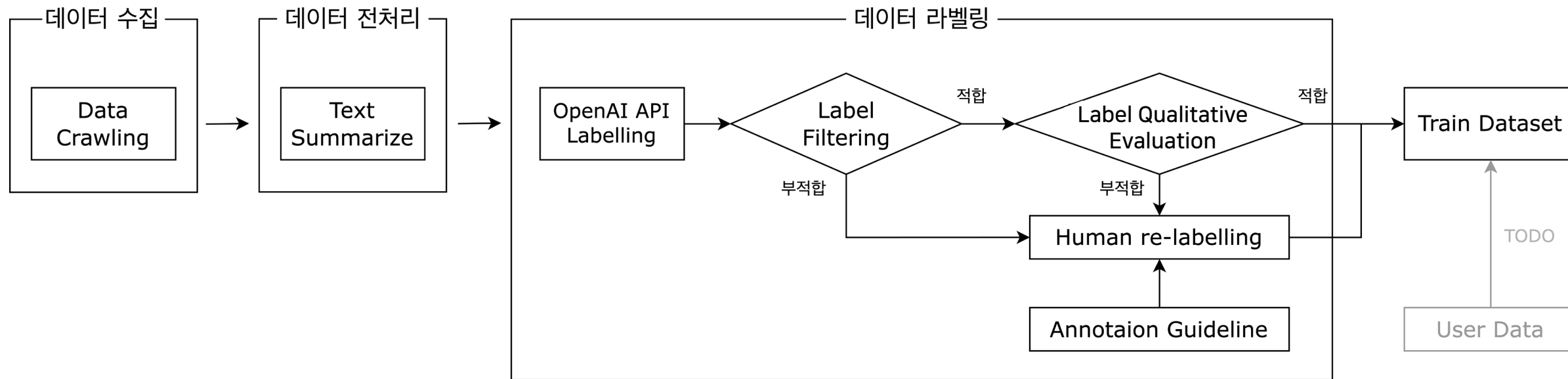
```
{  
  "success": True | False  
}
```

REST API Document 작성

- Request와 Response를 명시해 개발의 편의성을 증진
- 문서화된 호출 방식을 통해 헛갈림 방지



데이터셋 구축 과정





데이터 수집

- 출처: 블로그 포스팅 (Tistory, Naver, Velog)
- 수집 방법: 크롤링 (selenium, beautifulsoup)
- 근거: 설문조사 - 북마킹 출처 중 블로그 비율 ↑

1	url	title	context	big_topic	small_topic	source
2	https://blog.naver.com/s	2023.7.5. 한국 짜장면, 맛있는 역사와 진화 그리고 오늘의 점심	2023.7.5. 한국 짜장면, 맛있는 역사와 진화 그리고 오늘의 점심 외	엔터테인먼트·예술	문학·책	naver_blog
3	https://blog.naver.com/z	괜찮은 서양철학 에세이	평범하게 비범한 철학에세이 저자김필영출판스마트북스발매2023.6	엔터테인먼트·예술	문학·책	naver_blog
4	https://blog.naver.com/t	메타버스 유토피아 마크 반 리메남 (21세기북스) 추천 경제 책	메타버스 유토피아 마크 반 리메남 (21세기북스) 추천 경제 책지	엔터테인먼트·예술	문학·책	naver_blog
5	https://blog.naver.com/jj	역행자 확장판 자정의 22전략 이젠 꼭! 역행을 하면 재미난 세상	역행자 확장판 자정의 22전략 이젠 꼭! 역행을 하면 재미난 세상이	엔터테인먼트·예술	문학·책	naver_blog
6	https://blog.naver.com/t	플러팅 뜻 - 마케팅과 가장 닮아있는 기법	플러팅의 뜻은 이 시대 마케팅과 가장 닮아있는 기법이지 않을까.	엔터테인먼트·예술	문학·책	naver_blog
7	https://blog.naver.com/s	2024년도 한예중 연극원 지정희곡 셰익스피어 <맥베스> 해석과	셰익스피어 맥베스의 키워드 #야망 #욕망 #권력욕셰익스피어 맥베	엔터테인먼트·예술	문학·책	naver_blog
8	https://blog.naver.com/t	[Review] 세계화의 종말과 새로운 시작	세계화의 종말과 새로운 시작 저자마크 레빈슨출판페이지2북스발	엔터테인먼트·예술	문학·책	naver_blog
9	https://blog.naver.com/s	책리뷰 디지털이 할 수 없는 것들 - 인간으로 느끼게 해주는 것	내내내산 <디지털이 할 수 없는 것들> 쪽이책리뷰[디지털이 할 수	엔터테인먼트·예술	문학·책	naver_blog
10	https://blog.naver.com/c	생성 AI 시대 슈퍼 개인의 탄생 : GPT Revolution	AI 시대 슈퍼 개인의 탄생 안녕하세요. 도서 블로거 금소니 입니다	엔터테인먼트·예술	문학·책	naver_blog
11	https://bloq.naver.com/k	상상을 초월하는 인간의 편향적 사고 / 타블로 대학과 타진요 사	타블로와 타진요시작은 열주 2000년대 말이었습니다.인터넷에 흥	엔터테인먼트·예술	문학·책	naver bloq



데이터 라벨링

데이터 전처리



블로그 문서 한 줄 요약

사용 모델: [lcw99/t5-large-korean-text-summary](#)

수행 이유: OpenAI API에 입력 시 발생하는 비용 절감

Input: **블로그 본문**

Output: **요약 text**

제주 여행 코스 짜는 방법

리딩 29 | 지인 4,408

이미지



여행을 할 차례가 올까? 꼭 필요한 과정인 '코스 짜기'. 제주를 관광지이자 맛집이 많아 여행하기에 어려움이 없을 수 있다. 이것 저것 알아보기 번거로운 여행자를 위해 여행 코스 쉽게 짜는 방법을 소개한다. 트립닷컴이 추천하는 대표적인 3박 4일 코스를 참고해보자.

제주 코스 짜는 방법

1단계: 여행 컨셉 정하기

먼저 이번 제주 여행은 어떤 컨셉으로 제주고 싶으시게끔 정하자. 관광, 맛집, 체험 등 컨셉에 따라 코스가 달라진다. 나홀로 여행이 아니라면 친구, 부모님, 연인 등 동행자의 취향도 고려할 것.

2단계: 제주 지역 이해하기

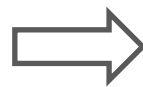
컨셉을 정한 뒤에는 제주를 이해할 차례. 제주를 서울의 3배 크기인 제주 시내와 서귀포, 중문, 성산, 우도 등 6개 지역으로 나눌 수 있다. 각 지역의 특징을 알면 동선 짜기가 수월하다. 여행 중 숙소는 어디에 잡을 것인지 거점을 정한 뒤, 본격적인 일정을 계획해보자.

[제주도를 한눈에, 지역별 소개](#)

요약된 text



제주 여행은 컨셉에 따라 코스가 달라지며 여행 중 숙소를 어디에 잡을 것인지 거점을 정하고 가고 싶은 관광지와 맛집을 일정에 추가하면 나만의 일정이 완성된다.



1차 라벨링



OpenAI API를 이용하여 라벨 생성

사용 모델: [text-davinci-003](#)

프롬프트:

```
Instruction:
  Tell me 5 tags that match the following document.
  Increase the weight of the Title.
  Don't be too descriptive.
  Tags must be noun.
Desired format: #English(Korean), #English(Korean), #English(Korean), #English(Korean), #English(Korean)
Title: {title}
Text: {요약 text}
Tags:
```

Input: **블로그 문서 제목, 요약된 text**

Output: **라벨(해당 페이지를 대표하는 5개의 태그)**

#Jeju(제주)

#Travel(여행)

#Accommodation(숙소)

#Attraction(관광지)



#Food(맛집)



라벨 필터링 후 Human Labeling

Sample_survived: 22268

라벨 필터링 부적합 예시

블로그 제목	부천시청역 현대백화점 맛집 저렴한 스테이크 파스타 시그니처랩				
본문 요약본	부천시청역 현대백화점 내 시그니처랩은 저렴한 금액에 전문 양식 셰프의 요리를 맛볼 수 있는 곳으로 인스타에 태그와 함께 업로드 시 에이드 한잔 무료 서비스가 제공된다.				
생성된 태그	 #부천(부천시청역)	#현대백화점	#맛집	#스테이크	 -
	#영어(한글) 형식에 맞지 않음			5개가 생성되지 않음	



Human Labeling	#Bucheon(부천)	#Hyundai Department Store(현 대백화점)	#Restuarant(맛집)	#Steak(스테이크)	#Pasta(파스타)
----------------	--------------	---	-----------------	--------------	-------------



라벨 정성적 평가 후 Human Labeling

Sample_survived: 371

라벨 정성적 부적합 예시

블로그 제목	Python 프로파일링을 위한 도구들 (Process, Memory, Execution Time)				
본문 요약본	Python에서는 Profiling을 위한 다양한 도구들을 가지고 있어 CPU 사용량을 간접적으로 알 수 있는 memory profiler를 통해 memory 사용량을 측정할 수 있다.				
생성된 태그	#Computer(컴퓨터)	#IndirectCPUCheckUtils (간접적CPU측정도구)	#Process(프로세스)	#Memory(메모리)	#ExecutionTime(실행시간)
	태그가 지나치게 광범위함	태그가 지나치게 특수함			중심적인 내용에서 다소 벗어남

↓

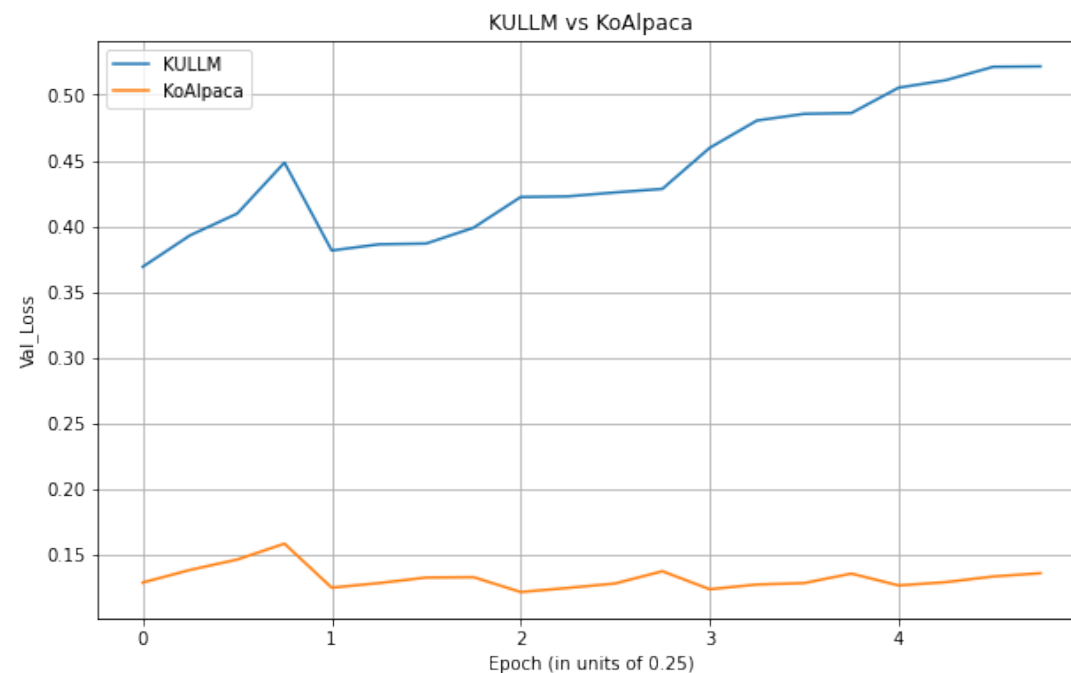
Human Labeling	#Python(파이썬)	#Profiling(프로파일링)	#Usage(사용량)	#CPU(CPU)	#Memory(메모리)
----------------	--------------	-------------------	-------------	-----------	--------------



태그 생성 모델

BackBone 후보 모델

	KULLM	Beomi-KoAlpaca
Backbone	Polyglot-Ko-12.8B	
MaxLength	2048	
PositionalEncoding	RoPE	
Dataset	GPT4ALL, Dolly, Vicuna	KoAlpaca v1.0
Val_Loss	High	Low
Tag_Quality	Low	Adequate
Max_Bsz	1	2
Time_spent	4.26sec/sample	3.77sec/sample





태그 생성 모델

BackBone 모델 선정

Beomi/KoAlpaca-Polyglot-12.8B로 선정

Hyperparameter	Value
$n_{parameters}$	12,898,631,680
n_{layers}	40
d_{model}	5120
d_{ff}	20,480
n_{heads}	40
d_{head}	128
n_{ctx}	2,048
n_{vocab}	30,003 / 30,080
Positional Encoding	Rotary Position Embedding (RoPE)
RoPE Dimensions	64



Hugging Face

Search models, datasets, users...

Hugging Face is way more fun with friends and colleagues! [Join an organization](#)



beomi/ **KoAlpaca-Polyglot-12.8B**

like 34



Text Generation



PyTorch



Safetensors



Transformers



KoAlpaca-v1.1b



License: apache-2.0



Model card



Files and versions



Community



태그 생성 모델

Prompt 소개

Alpaca 데이터셋의 형식처럼 Instruction + Input을 통해 Response를 추론

QLoRA 기법을 활용한 PEFT 방식 사용

Prompt(Input) & Label(Output):

Instruction(명령어):

다음의 블로그 글에 어울리는 태그 5개를 생성하시오. 태그의 형식은 다음과 같음.

[#영어(한글), #영어(한글), #영어(한글), #영어(한글), #영어(한글)]

Input(입력):

주제는 [{self.topics}], 제목은 [{self.title}], 본문은 [{self.content}]이다.

Response(응답):

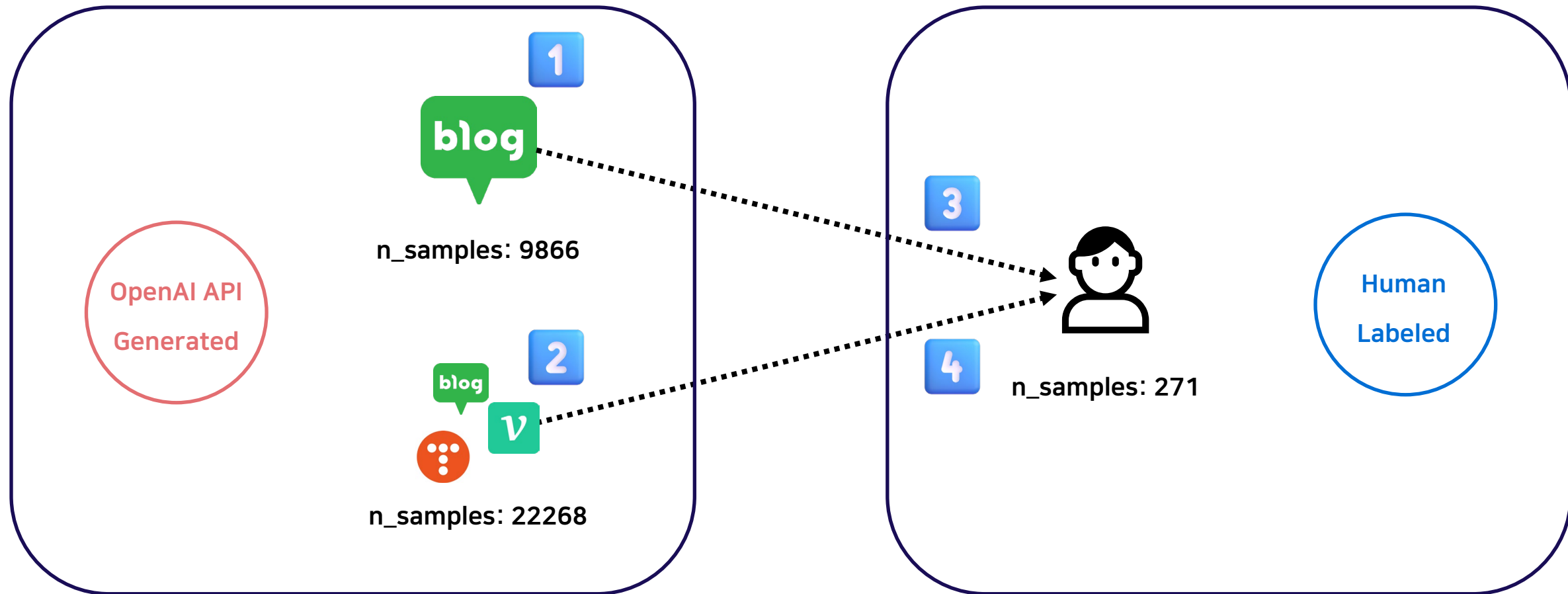
{self.tag1}, {self.tag2}, {self.tag3}, {self.tag4}, {self.tag5}



태그 생성 모델

모델 훈련

Data Feeding 방식에 따라 총 4개의 모델 후보군을 확보





태그 생성 모델

Metric 설명

- Ex)
- **GroundTruth Tag** : #Jeju(제주), #Travel(여행), #Attraction(관광지), #Food(맛집), #Accommodation(숙소)
- **Predicted Tag** : #Tourism(관광), #Jeju(제주도), #Accommodation(숙소), #Trip(여행지), #Food(맛집)

- Let) GroundTruth 에서 Predicted로의 Translation Task라고 생각

- Problem) 단순히 바로 BLEU나 ROUGE를 적용하는 데에 문제가 있음
- ① 한글로 훈련한 모델의 특징 상, 한글 태그는 잘 만들더라도 영어가 다를 경우에 점수가 과도하게 낮게 측정됨
- ② 생성되는 태그의 순서는 상관 없음에도 불구하고, GT와 순서가 다르다는 이유로 점수를 아예 주지 않음
- ③ 거의 유사한 태그를 맞추었으나, 각 태그가 완벽하게 일치하지 않으면 점수를 아예 주지 않음

- Solve)
- ① 성능 평가는 한글 태그에 한해 진행하도록 함
- ② 순서에 무관하게 성능 평가
- ③ a태그와 b태그의 Full-matching에 의한 True/False가 아닌 range(0,1) 사이의 유사도를 구함



태그 생성 모델

Metric 설명

1. 모든 단어 쌍들에 대한 어떠한 방법으로 유사도 계산

Score = []

PRED/GT	제주	여행	관광지	맛집	숙소
관광	0.1	0.8	0.98	0.5	0.15
제주도	0.95	0.05	0.08	0.05	0.02
숙소	0.13	0.3	0.4	0.2	1
여행지	0.5	0.92	0.89	0.6	0.18
맛집	0.06	0.42	0.45	1	0.21



태그 생성 모델

Metric 설명

2. 가장 높은 점수의 칸을 선택 후, Score에 기입

Score = [1]

PRED/GT	제주	여행	관광지	맛집	숙소
관광	0.1	0.8	0.98	0.5	0.15
제주도	0.95	0.05	0.08	0.05	0.02
숙소	0.13	0.3	0.4	0.2	1
여행지	0.5	0.92	0.89	0.6	0.18
맛집	0.06	0.42	0.45	1	0.21



태그 생성 모델

Metric 설명

3. 해당 칸이 속하는 행과 열을 삭제 처리

Score = [1]

PRED/GT	제주	여행	관광지	맛집	숙소
관광	0.1	0.8	0.98	0.5	
제주도	0.95	0.05	0.08	0.05	
숙소					
여행지	0.5	0.92	0.89	0.6	
맛집	0.06	0.42	0.45	1	



태그 생성 모델

Metric 설명

4. 2~3을 반복

Score = [1, 1]

PRED/GT	제주	여행	관광지	맛집	숙소
관광	0.1	0.8	0.98	0.5	
제주도	0.95	0.05	0.08	0.05	
숙소					
여행지	0.5	0.92	0.89	0.6	
맛집	0.06	0.42	0.45	1	

태그 생성 모델

Metric 설명

4. 2~3을 반복

Score = [1, 1, 0.98]

PRED/GT	제주	여행	관광지	맛집	숙소
관광	0.1	0.8	0.98		
제주도	0.95	0.05	0.08		
숙소					
여행지	0.5	0.92	0.89		
맛집					

태그 생성 모델

Metric 설명

4. 2~3을 반복

Score = [1, 1, 0.98, 0.95]

PRED/GT	제주	여행	관광지	맛집	숙소
관광					
제주도	0.95	0.05			
숙소					
여행지	0.5	0.92			
맛집					

태그 생성 모델

Metric 설명

4. 2~3을 반복

Score = [1, 1, 0.98, 0.95, 0.92]

PRED/GT	제주	여행	관광지	맛집	숙소
관광					
제주도					
숙소					
여행지					
맛집					

0.92



태그 생성 모델

Metric 설명

5. Score의 평균이 이 샘플의 점수

Score = [1, 1, 0.98, 0.95, 0.92]

PRED/GT	제주	여행	관광지	맛집	숙소
관광					
제주도					
숙소					
여행지					
맛집					

Score = 0.97



태그 생성 모델

Metric 설명

- 유사도 계산 방법:
- Jaccard 유사도 - 순서에 상관 없이 GT 답의 일부를 가져왔을 때에 대한 보정을 위해 사용
- Rouge.f1 유사도 - Rouge-1으로 글자에 대한 f1, Rouge-L로 순서를 고려한 부분 문자열의 f1을 동시에 측정하기 위해 둘의 평균을 사용
- Levenshtein 유사도 - 구조적으로 최대한 적은 수정을 통해 GT를 복원할 수 있는 Pred는 좋은 태그라고 판단하여 사용
- W2v 유사도 - 앞의 sparse한 방법으로 구하는 것 뿐 아니라, dense한 유사도를 단어 수준에서 간단하게 구해보고자 사용.



태그 생성 모델

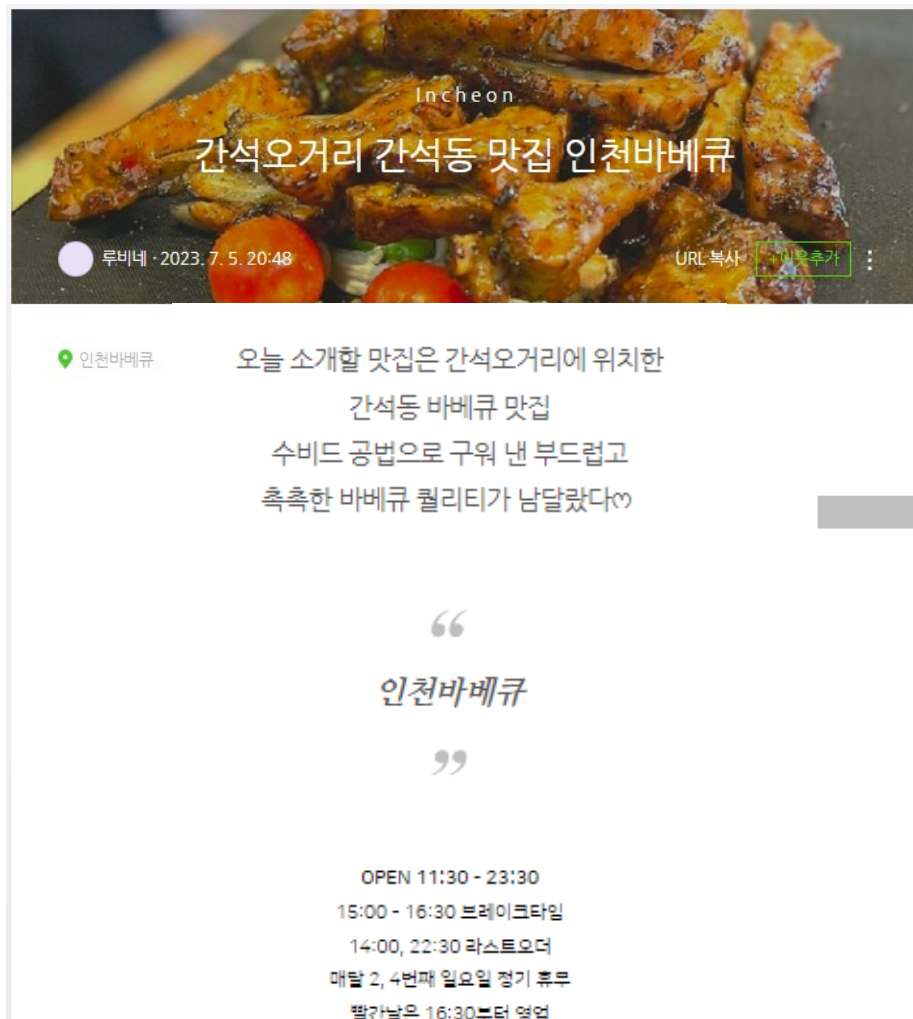
모델 비교

	1	2	3	4
Dataset	Only_Naver	All_Data	Only_Naver + HumanLabel	All_Data + HumanLabel
n_samples	9866	22268	9866 + 271	22268 + 271
Jaccard_score	0.27616	0.27333	0.28994	0.28756
Rouge_score	0.31227	0.30294	0.31879	0.31475
Lev_score	0.30579	0.29940	0.31500	0.31249
W2V_score	0.51933	0.44014	0.43832	0.43624
Sanity	0.83	0.82	0.83	0.83

GT를 직접 수기로 작성해준 것을 고려, **Lexical한 Matching**이 가장 뛰어난 **3번 모델**을 사용하기로 결정



태그 생성 결과 – 예시 1 (Rouge 0.799)



Predict

- #Ganseokdong(간석동)
- #Barbecue(바베큐)
- #Restaurant(맛집)
- #Ganseok(간석)
- #Incheon(인천)

Ground truth

- #Barbecue(바베큐)
- #Restaurant(맛집)
- #Ganseok(간석)
- #Sous Vide(수비드)
- #Incheon(인천)



태그 생성 결과 – 예시 2 (Rouge 0.310)

교육

방통대 유아교육과 가려면?~



HAEUN · 2023. 7. 5. 18:42

URL 복사

+이웃추가



안녕하세요~?

방통대 유아교육과를 알려드리려

찾아온 저인데요~!ㅎㅎ

한국방송통신대는~?

우리나라 최초의 원격 대학으로

사이버대의 시초라고볼수있는데요!

오늘은 그중에서도 유치원교사를

할수있는 유교과에 대해서

집중적으로 입학방식까지 알아보려해요!~ㅎ



Predict

#Teaching(교사)

#University(대학)

#Broadcasting
University
(방통대)

Education(교육)

Education(교육)

Ground truth

#KoreaNationalOpenUniversity
(방송통신대학교)

#ChildhoodEducation
(유아교육)

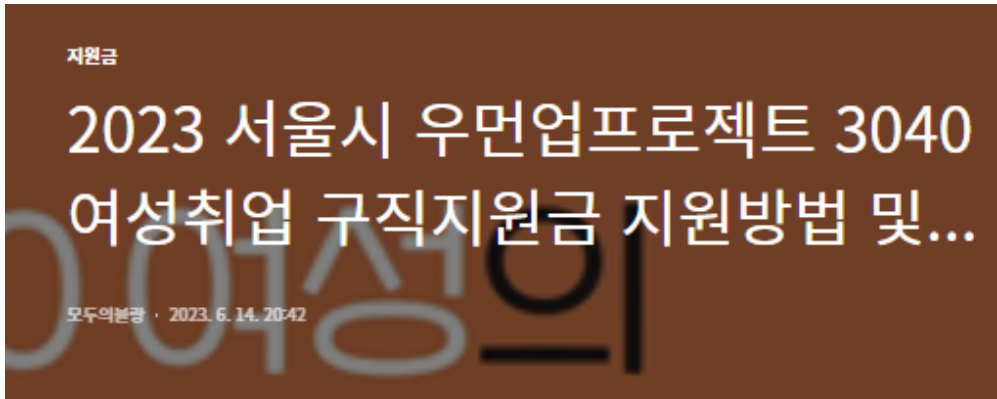
#Transfer(편입)

#AcademicBankSystem
(학점은행제)

#competition
rate(경쟁률)



태그 생성 결과 – 예시 3 (Rouge 0.804)



서울우먼업프로젝트 3040여성취업 주요사업

이번 글에서는 3040 여성 취업의 현실과 어려움에 대해 알아보고, 서울시가 제안하는 '서울우먼업 프로젝트'라는 솔루션에 대해 소개하겠습니다.
서울우먼업프로젝트는 구직지원금과 인턴십, 고용장려금을 통해 여성들이 자신에게 맞는 일자리를 찾고 실질적인 경제활동을 할 수 있도록 지원하는 프로젝트입니다.

목차

- 3040 여성 취업의 현실과 어려움
- 서울우먼업프로젝트란?
- 서울우먼업프로젝트 3가지 주요사업
- 우먼업 구직지원금
- 우먼업 인턴십
- 우먼업 고용장려금



Predict

#WomenUpProject(우먼업프로젝트)

#JobHunting(취업)

#Seoul(서울시)

#Women(여성)

#Support(지원)

Ground truth

#employmen(취업)

#Female(여성)

#Seoul(서울)

#Support(지원)

#Business(사업)



End of Document

Thank You.



강



남



특



호



대

이미지 출처: (주)누리토이즈