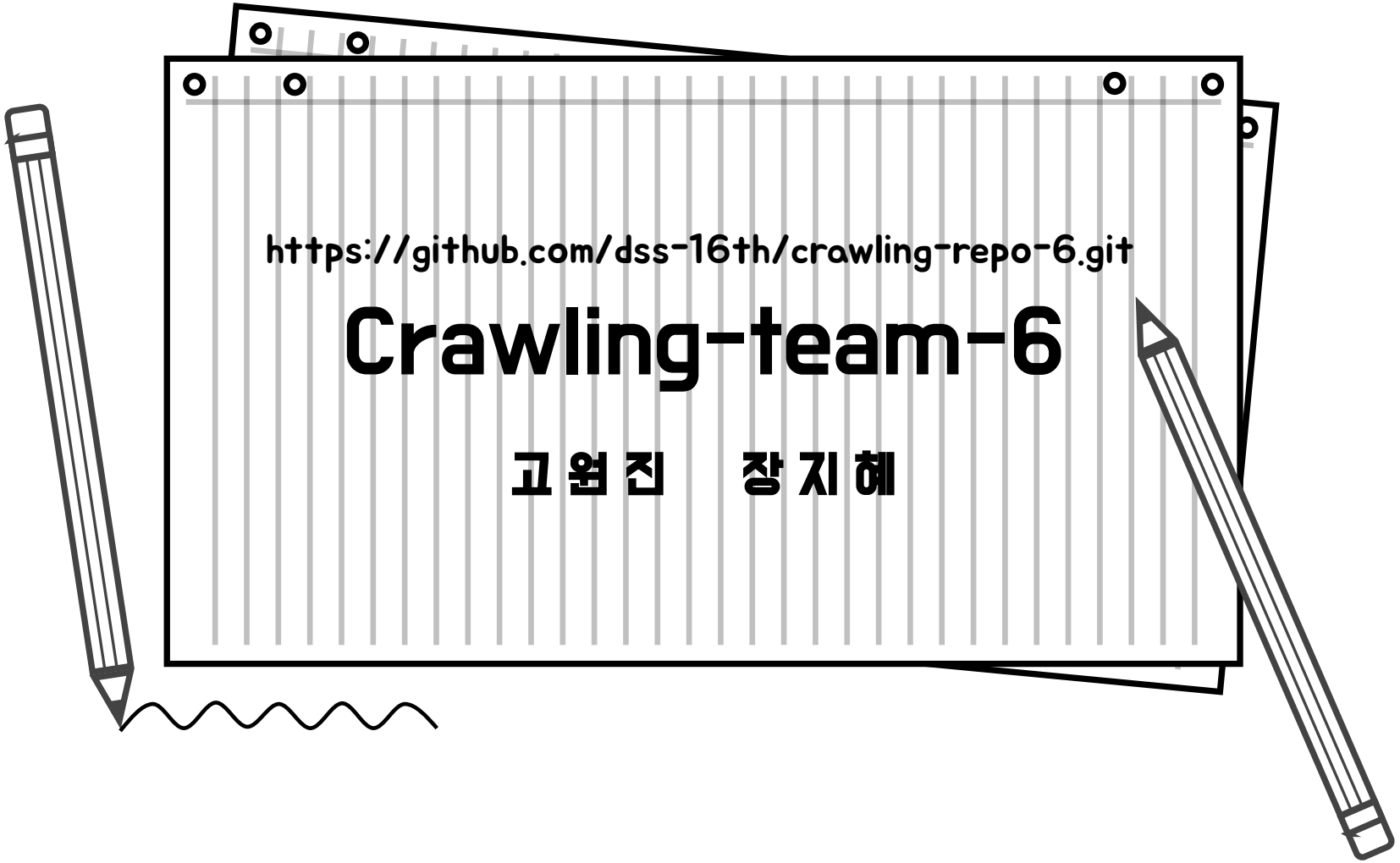


https://github.com/dss-16th/crawling-repo-6.git

Crawling-team-6

고원진 장지혜

취미생활 자기계발 트렌드를 반영한 검색/추천 시스템

Goal : ① 각 플랫폼 데이터를 Mysql DB 축적 및 키워드 분류

② 검색엔진 구현 및 웹 서비스 제공

클래스톡/탈잉 크롤링 방식 변경 : '검색'을 위한 카테고리 분류

taling ClassTok

```
cat1_beauty_health = list(zip([28, 32, 31, 33, 27], ['메이크업', '퍼스털컬러', '패션', '셀프케어', 'PT/GX']))
cat1_activity = list(zip([78, 235, 123, 217, 240], ['방송', '댄스', '연기/무용', '스포츠/레저', '이색 액티비티']))
cat1_life = list(zip([233, 246, 88, 248, 80, 127, 103], ['인문/교양', '인테리어', '반려동물', '부모/육아', '출판/글쓰기', '사주/타로', '심리상담']))
cat1_hobby = list(zip([81, 79, 222, 232, 84, 83, 60, 59, 61, 76, 125, 249, 126], ['이색취미/공예', '사진', '취미미술', '디지털드로잉', '요리/베이킹', '커피/차/술', '보컬',
                    '악기', '작곡/디제잉','캘리그래피', '플라워', '조향/캔들/비누', '가죽/목공/도예']))
cat1_money = list(zip([214, 188, 116, 244, 213, 15], ['투잡', '마케팅', '주식투자', '부동산', '금융지식', '창업']))
cat1_career = list(zip([239, 250, 17, 13, 12, 14, 11, 34, 35, 54, 182], ['실무역량', '마케팅', '취업/이직/진로', '엑셀', '파워포인트', '스피치', '데이터분석',
                    '웹개발', '앱개발', '컴퓨터공학', '자격증/시험']))
cat1_design = list(zip([3, 201, 206, 209, 193, 199], ['건축', '그래픽디자인', 'UX/UI디자인', '제품디자인', '영상편집', '영상제작']))

cat1_language = list(zip([41, 42, 43, 44, 51], ['영어회화', '중국어회화', '일본어회화', '어학자격증', '기타 외국어']))
cat1s = [cat1_beauty_health, cat1_activity, cat1_life, cat1_hobby, cat1_money, cat1_career, cat1_design, cat1_language]
cat1s_ko = ['뷰티/헬스', '액티비티', '라이프', '취미/공예', '머니', '커리어', '디자인/영상', '외국어']
taling_df = pd.DataFrame(columns=['site', 'link', 'title', 'teacher', 'category_1', 'category_2', 's_price', 'discount', 'contentment'])

soldout = 0
```

CLASS 101

	Unnamed: 0	site	link	title	teacher	category_1	category_2	s_price	discount	contentment
1440	229	클래스 101	https://class101.net/products/5f7193a0133d4a00...	[sales]유럽 건축사: 로마 vs 비로마	클래스101[클래스101]	직무교육	글쓰기/콘텐츠	총 9,900원	없음	91.3%
1441	230	클래스 101	https://class101.net/products/5f8f3701b0a15c00...	[sales]기획자들: 종합광고 대행사 <스튜디오좋>	클래스101[클래스101]	직무교육	글쓰기/콘텐츠	총 9,900원	없음	100.0%
1442	231	클래스 101	https://class101.net/products/5f6abac6a34a5800...	[sales]SNS 디지털 퍼포먼스 마케팅 기초부터 심화까지!	노아노마드 에듀[최윤영]	직무교육	비즈니스/생산성	총 309,000원	없음	81.2%
1443	232	클래스 101	https://class101.net/products/5e8e7fb05a025f41...	[sales][101원] 사무실 없는 회사, 로켓펀치가 알려주는 원격 근무 잘 하는 법	로켓펀치[로켓펀치]	직무교육	비즈니스/생산성	총 101원	없음	96.8%
1444	233	클래스 101	https://class101.net/products/5f4f1d374dc39e00...	[sales]원격 근무, 이론부터 실무까지! 원격 근무 솔루션의 모든 것	리모트워크[리모트워크]	직무교육	비즈니스/생산성	총 69,000원	없음	0

taling

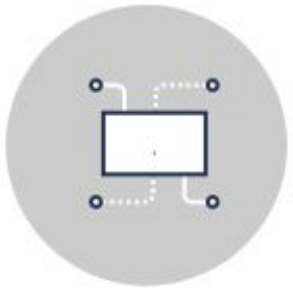
	Unnamed: 0	site	link	title	teacher	category_1	category_2	s_price	discount	contentment
4478	5	탈잉	https://taling.me/Talent/Detail/14568	[종로]여행 중국어 회화 ! 배우고 가자 중국 여행 !	Xiaotuyaa[이혜린]	외국어	기타 외국어	월 10,000원	0	0
4479	6	탈잉	https://taling.me/Talent/Detail/21789	[강남][B급 생존스페인어] 까미나 꼼미고! 스페인어로 함께 걸어가요 :D !!	지쥴[지주련]	외국어	기타 외국어	월 20,000원	0	0
4480	7	탈잉	https://taling.me/Talent/Detail/1781	[전남대]독일을 꿈꾸는 모든 이를 위한 수업.	차니[임경찬]	외국어	기타 외국어	월 25,000원	0	0
4481	8	탈잉	https://taling.me/Talent/Detail/19778	[잠실][여행작가/ 스페인어 전공생] TOMAS 쌤의 세상에서 가장 쉬운 스페인...	Tomas 쌤[이승민]	외국어	기타 외국어	월 12,000원	0	0
4482	9	탈잉	https://taling.me/Talent/Detail/28503	[시흥][한 방에 배우는 베트남어] #1:1 맞춤형수업 #빠른피드백 #체계적수업 ...	THIÊN ÂN[지은(Thiên ân)]	외국어	기타 외국어	월 30,000원	0	0

ClassTok

-_- 이런

	Unnamed: 0	site	link	title	teacher	category_1	category_2	s_price	discount	contentment
344	344	클래스톡	https://www.classtok.net/class/classDetail/1647	세상 단 하나뿐인, 나만의 유니크한 캔버스 조명 만들기	소니아	NaN	DIY	월 39,800 원	55%	4.7
345	345	클래스톡	https://www.classtok.net/class/classDetail/2730	박스필라테스로 만드는 아름다운 라인	Judy	NaN	운동	월 20,000 원	75%	NaN
346	346	클래스톡	https://www.classtok.net/class/classDetail/3295	씨리얼 토익 800점을 위한 LC 특강!	토익강사Erin	NaN	영어	월 20,000 원	60%	NaN
347	347	클래스톡	https://www.classtok.net/class/classDetail/431	하루20분!!!! 다이어트+체력 향상 이 영상 하나면 끝!!!!!	버닝쭌	NaN	다이어트	월 29,000 원	57.9%	4.9
348	348	클래스톡	https://www.classtok.net/class/classDetail/3205	발성전문가의 뛰어나되 유연한 진짜 스피치 강의	다이아제이	NaN	스피치	월 49,000 원	62%	NaN

물리적 시나리오 (예 상)

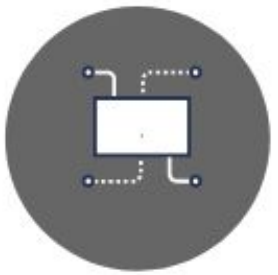


HTTP

Control Server

Crawling Server

Data Base



Backup DB

물리적 시나리오 (실 제)

taling

CLASS 101

ClassTok

Control Server

Crawling Server

Backup DB

Data Base

I. Crawling Method

- Class101(클래스101) : selenium -> graphql post방식 requests
- Taling(탈잉) : Scrapy -> BeautifulSoup 으로 크롤링 방식 변경
  -(Category_1, Category_2 분류 -> 키워드검색)
- ClassTok(클래스톡) : Scrapy -> BeautifulSoup

II. DataBase

- Mysql (RDBMS): 검색/키워드 추천을 위한 인덱싱의 중요성
- Flask를 통한 서비스 구현을 위한 DB 연동 (업데이트)
- Backup DB에 관한 논의 필

III. Crawling Cycle

- 실시간성을 높이기 위해 6시간 간격 (하루 3번 정도 업데이트 : 클래스톡, 탈잉의 경우)
  - 매일 1시간 간격 or 매일 1회 크롤링을 통한 데이터 수집내용 비교
  - Mysql에는 기존 데이터 지우고 업데이하는 형식으로 크롤링 진행중
  - csv 포맷으로도 축적중

- 서버를 늘려서 실시간성 증대
  - 각 서비스별 정보만은 text로 제공시 서버를 분산하지 않아도 된다는 판단

취미생활 자기계발 트렌드를 반영한 검색/추천 시스템 : 검색엔진 웹 서비스 제공

Mysql query 문법 -> LIKE

SQLAlchemy: ORM(Object-Relational Mapping, 객체 관계 매핑)

DB 연동방법 고민중 : 1) 직접 연동 2) Flask확장 모듈 사용

ISSUE

① Flask로 검색 사이트 만들기

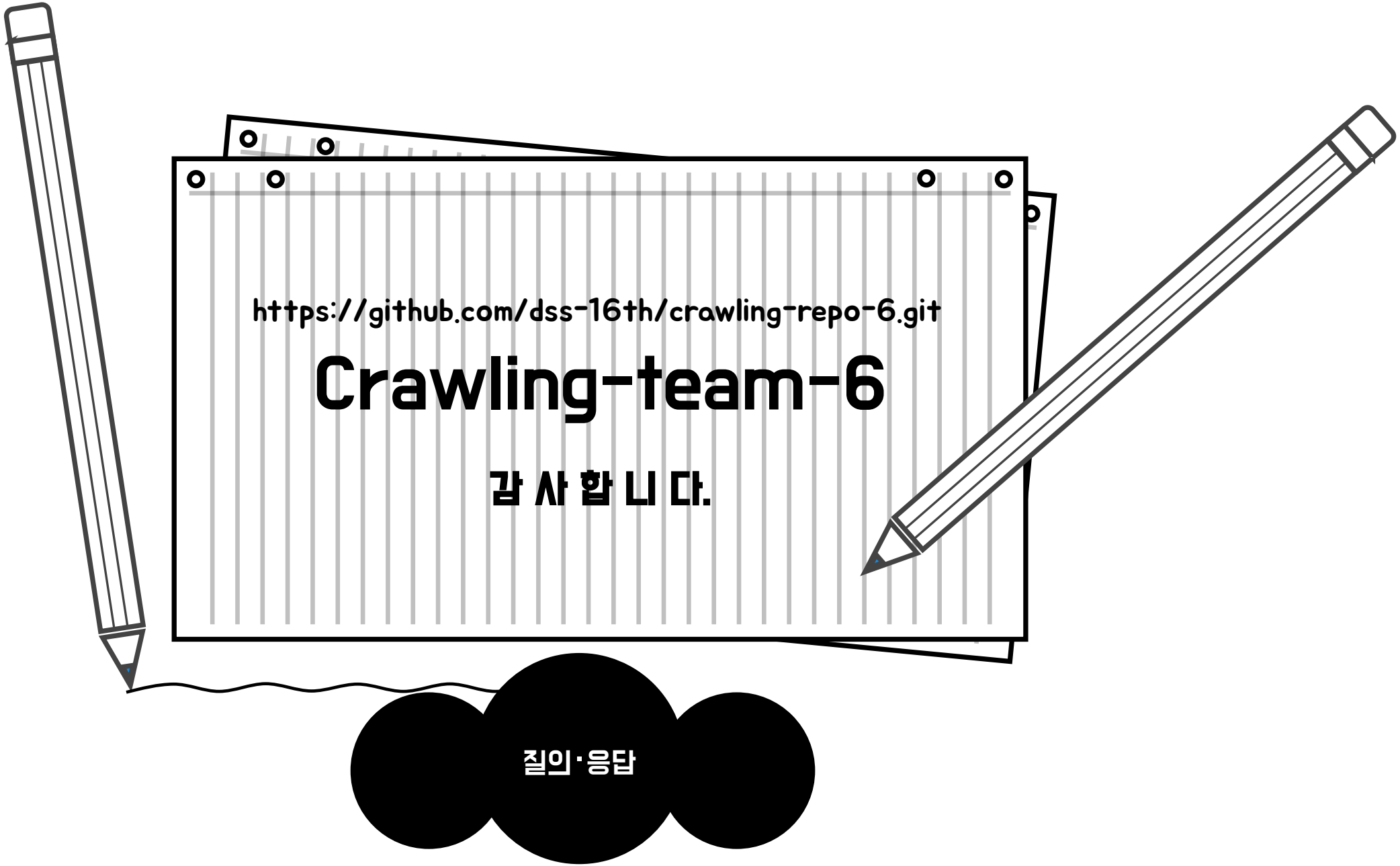
② Flask-mysql 연동
(추가적인 서버의 필요성)

③ 추가 기능: 웹 페이지상에 신규강좌 노출시키기
(코딩으로 DataFrame. csv 파일 만들기까지 진행)

④ DB 백업의 필요성 (현재까지는 그 필요성을 못느끼고 있으나 Flask 구현시 문제발생 염려)

⑤ .py 모듈화의 필요성 (Pipeline)

⑥ 정확한 키워드 검색을 위한 자연어 처리 및 문자열 형태소분류
(차후 학습이 더 진행된 상태에서 Develop 논의 : 현재는 카테고리로 키워드 분류)


https://github.com/dss-16th/crawling-repo-6.git
Crawling-team-6
감 사 합 니 다.
질의·응답