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

Crawling-team-6

고원진 장지혜

Enjoy your stylish business and campus life with BIZCAM

취미생활 자기계발 트렌드를 반영한 검색/추천 시스템

Enjoy your stylish business and campus life with BIZCAM

예상 시나리오

Enjoy your stylish business and campus life with BIZCAM

필요한 클래스 정보를 추천/비교 해 볼 수 있는 서비스가 있다면 좋겠다!

''간단하게 파이썬을 이용한
자동투자에 대해 알고 싶어! 강의는
어디서 찾아들어야하지?''

CLASS 101

taling

ClassTok

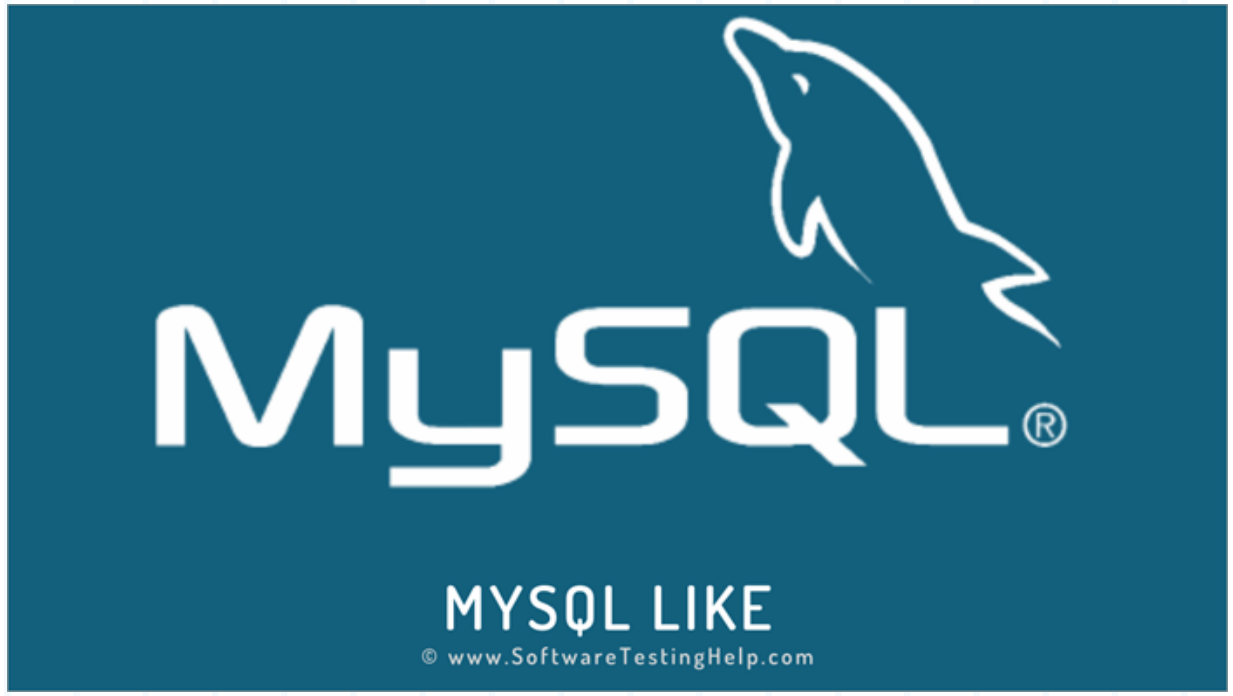
''신사임당 같은 유투버의
노하우를 알고 싶은데
커리큘럼을 확인할 방법은 없을까?''

''홈트로 필라테스
배우고 싶은데 튜터추천을
받고 싶어!''

''글쓰기를 배우고 싶은데
키워드 검색하면 추천해주는
시스템이 있으면 좋겠어.''

Mysql : like -> 검색어/키워드 분류

Enjoy your stylish business and campus life with BIZCAM



태그 -> DB (중복검색 가능성을 염두에 두고 태그 나누기)

클래스 101

Enjoy your stylish business and campus life with BIZCAM

```
cat_ko_ls = ['취미', '수익창출', '직무교육']
cat_eng_ls = ['creative', 'money', 'career']
brands = ["original", "money", "professional"]
categories = list(zip(cat_ko_ls, cat_eng_ls, brands))

req = requests.get('https://class101.net/robots.txt')
prohibit_url = []
for txt in req.text.split('\n'):
    if 'Disallow: ' in txt:
        prohibit_url.append('https://class101.net' + txt.replace('Disallow: ', ''))


class101_df = pd.DataFrame(columns=['site', 'link', 'title', 'teacher', 'category_1', 'category_2', 's_price', 'discount', 'contentment'])

for cat_ko, cat_eng, brand in categories:
    print(f'{cat_ko} / {cat_eng}')
    query = [{"operationName":"InfiniteProductCardsWithLastUpdatedInformation","variables":{"brand":[brand],"offset":offset,"limit":limit},"qu
    req = requests.post(url_graphql, json=query, )
    datas = req.json()


    onetime = []
    for i in range(len(datas[0]['data']['products'])):

        site = '클래스101'
```

클래스 톡

Enjoy your stylish business and campus life with BIZCAM

```
12      datas.append({
13          "title" : element.select_one('h2').text,
14          "category" : element.select_one('span').text.split(' · ')[0],
15          "teacher" : element.select_one('span').text.split(' · ')[1],
16          "price" : element.select_one('.price_info').text.split('\n')[1],
17          "discount" : element.select_one('.price_info').text.split('\n')[2].replace(' 할인',''),
18          "contentment" : likes,
19          "link" : 'https://www.classtok.net' + element.get('href'),
20      })
21
22  classtok_df = pd.DataFrame(datas)
23  classtok_df.tail()
```

	title	category	teacher	price	discount	contentment	link
342	더드림 수화 (베이직싸인)	수화	어구리짱	월30,000원	66%		https://www.classtok.net/class/classDetail/2395
343	English Interview 영어 면접 성공하기	면접	서연 Tiffany	월18,900원	79%		https://www.classtok.net/class/classDetail/3539
344	잘 팔리는 전자책 만들기	출판	박진희	월29,000원	67%		https://www.classtok.net/class/classDetail/3269
345	집에서 배우는 킥복싱 완전 기초 4주 완성 클래스	다이어트	길코치	월29,900원	66.4%	5	https://www.classtok.net/class/classDetail/1216
346	불투명한 수채화, 과슈물감으로 꽃 그리기	그림	J	월10,000원	87%		https://www.classtok.net/class/classDetail/2777

탈잉 : 카테고리 재분류중 -> 클래스 제목에서 검색이 안되는 경우 다수

Enjoy your stylish business and campus life with BIZCAM

```
:

# 탈잉 사이트 크롤링 방식 변경: scrapy -> BeautifulSoup (진행중)

# 키워드 분류/검색을 위한 카테고리 분류: cate1 -> cate2

cat1_beauty_health = list(zip('메이크업', '퍼스털컬러', '패션', '셀프케어', 'PT/GX'))
cat1_activity = list(zip('방송', '댄스', '연기/무용', '스포츠/레저', '이색 액티비티'))
cat1_life = list(zip('인문/교양', '인테리어', '반려동물', '부모/육아', '출판/글쓰기', '사주/타로', '심리상담'))
cat1_hobby = list(zip('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('투잡', '마케팅', '주식투자', '부동산', '금융지식', '창업'))
cat1_career = list(zip('실무역량', '마케팅', '취업/이직/진로', '엑셀', '파워포인트', '스피치', '데이터분석',
                      '웹개발', '앱개발', '컴퓨터공학', '자격증/시험'))
cat1_design = list(zip('건축', '그래픽디자인', 'UX/UI디자인', '제품디자인', '영상편집', '영상제작'))

cat1_language= list(zip('건축', '그래픽디자인', 'UX/UI디자인', '제품디자인', '영상편집', '영상제작'))
```

```
:

categories = [('뷰티/헬스', 3, 29), ('취미/공예', 22, 125)]   # 페이지수 확인하기
start_urls = []
for cat, cat_no, pages in categories:
    for page in range(1, pages+1):
        start_urls.append(f"https://taling.me/Home/Search/?page={page}&cateMain={cat_no}&cateSub=&region=&orderIdx=&quer
```

Issue

Enjoy your stylish business and campus life with BIZCAM

- RDBMS : Mysql -> 컬럼 및 데이터 통일, 수정의 어려움
- encoding = UTF8MB4 적용 예정
- 실시간성 증대를 위한 크롤링 주기 : 4~6시간 간격을 논의했으나
  업데이트 주기를 가늠할 수 없어 1일 기준으로 진행논의
- 검색어 분류를 위한 카테고리 분류 :
  탈잉의 경우 클롤링 방법 변경 후 시도중

취미생활 자기계발 트렌드를 반영한 검색/추천 시스템

Enjoy your stylish business and campus life with BIZCAM

Flask

web development,
one drop at a time

DB 활용 -> 서비스 제공

thank you! ; D

Enjoy your stylish business and campus life with BIZCAM