고원진

장지혜

1

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

Crawling-team-6

Enjoy your stylish business and campus life with BIZCAM

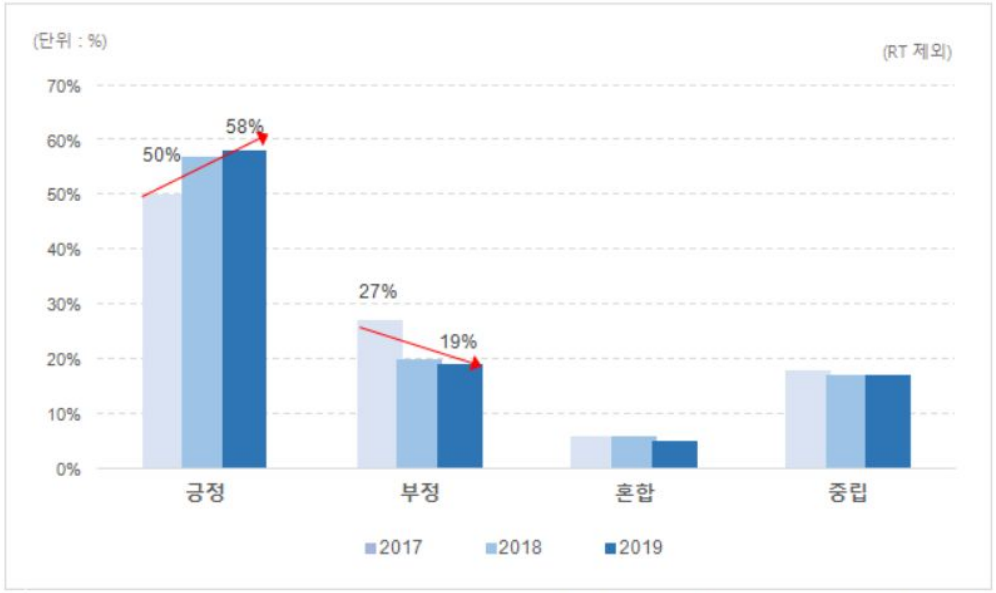
취미생활/자기계발 수요증가

- 주 52시간 근무제 시행
- 코로나로 인한 '집콕족'들의 증가

취미생활/자기계발 플랫폼의 성장세

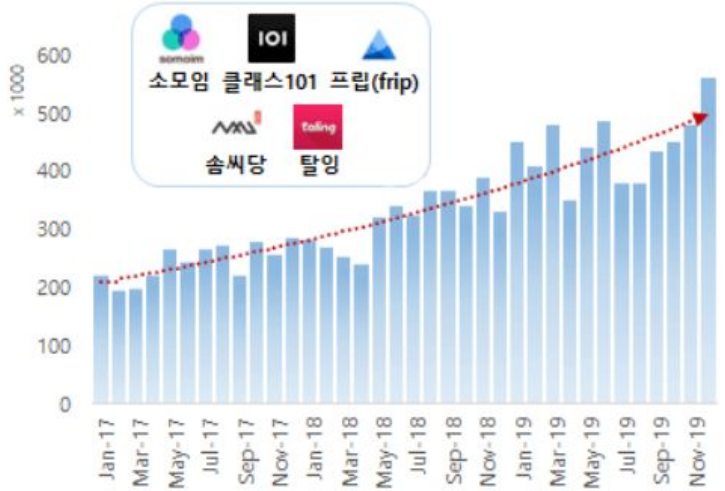
Enjoy your stylish business and campus life with BIZCAM

- '퇴근&취미/여가' 감성분석 -

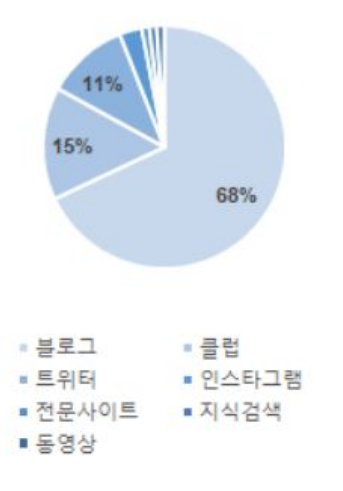


(Data Source : Nielsen Buzzword, 2017.01~2019.12)

'취미'관련 모바일 어플리케이션 이용자 수 추이



퇴근&취미 채널 점유율



(Data Source : ① Nielsen Koreanclick Android Mobile Behavioral Data, 2017.01~2019.12
② Nielsen Buzzword, 2019.06~2019.12)

- 다양한 관심사와 자기계발을 돕는 온/오프라인 플랫폼 수요증가
- N잡러, 프리랜서 마켓으로서의 역할

예상 시나리오 : '소비자가 필요한 클래스 정보를 사이트별로 비교검색 할 수 있는 서비스가 있으면 어떨까?'

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

""홈트로 필라테스
배우고 싶은데 튜터추천을
받고 싶어!""

taling

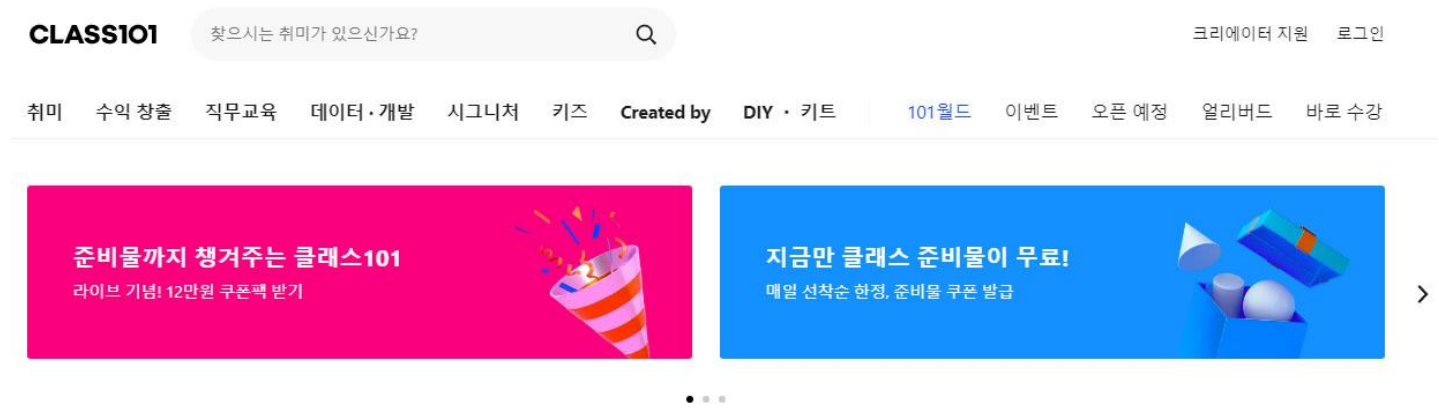
CLASS 101

ClassTok

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

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

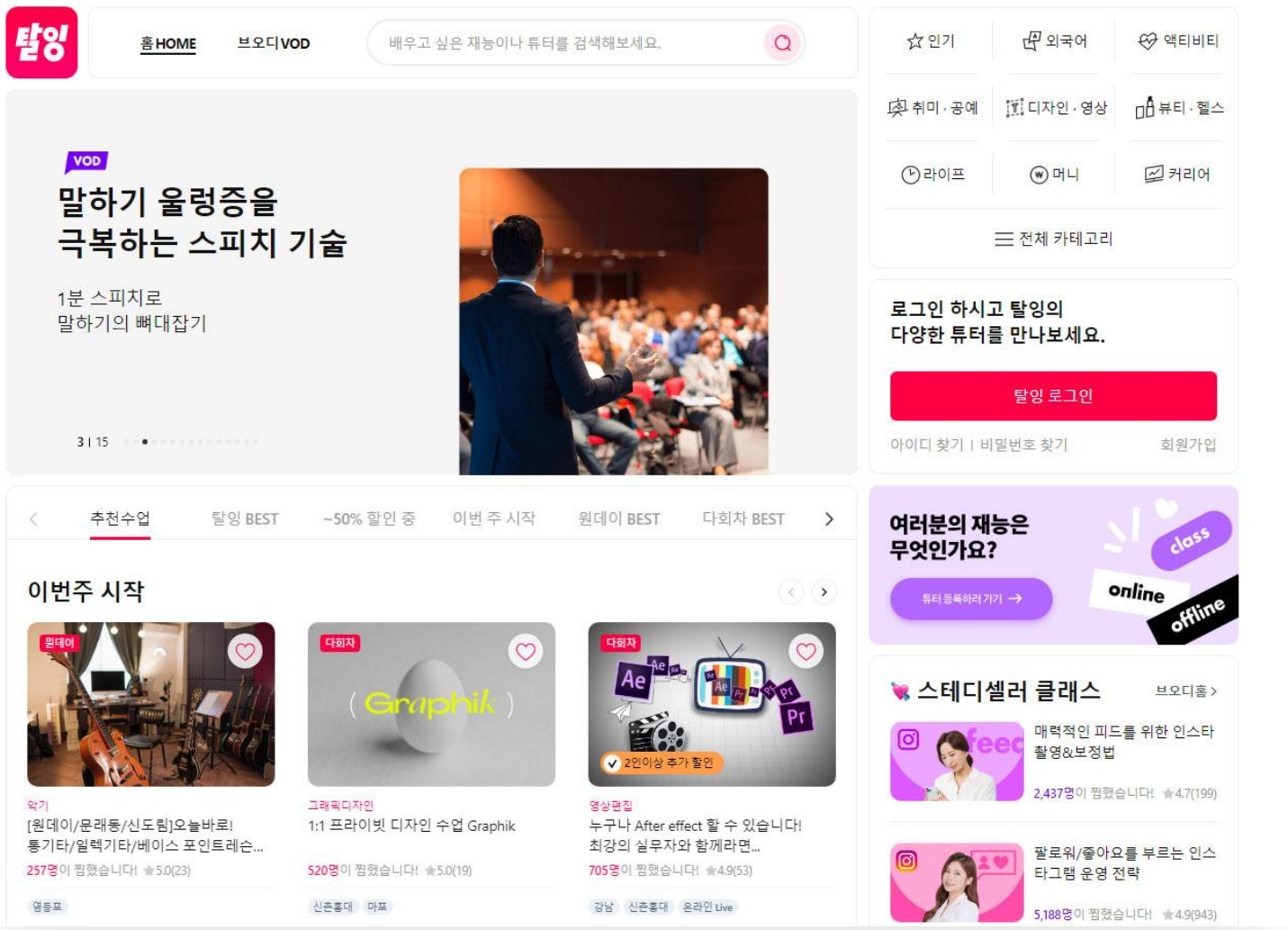
CLASS 101



Crawling Method

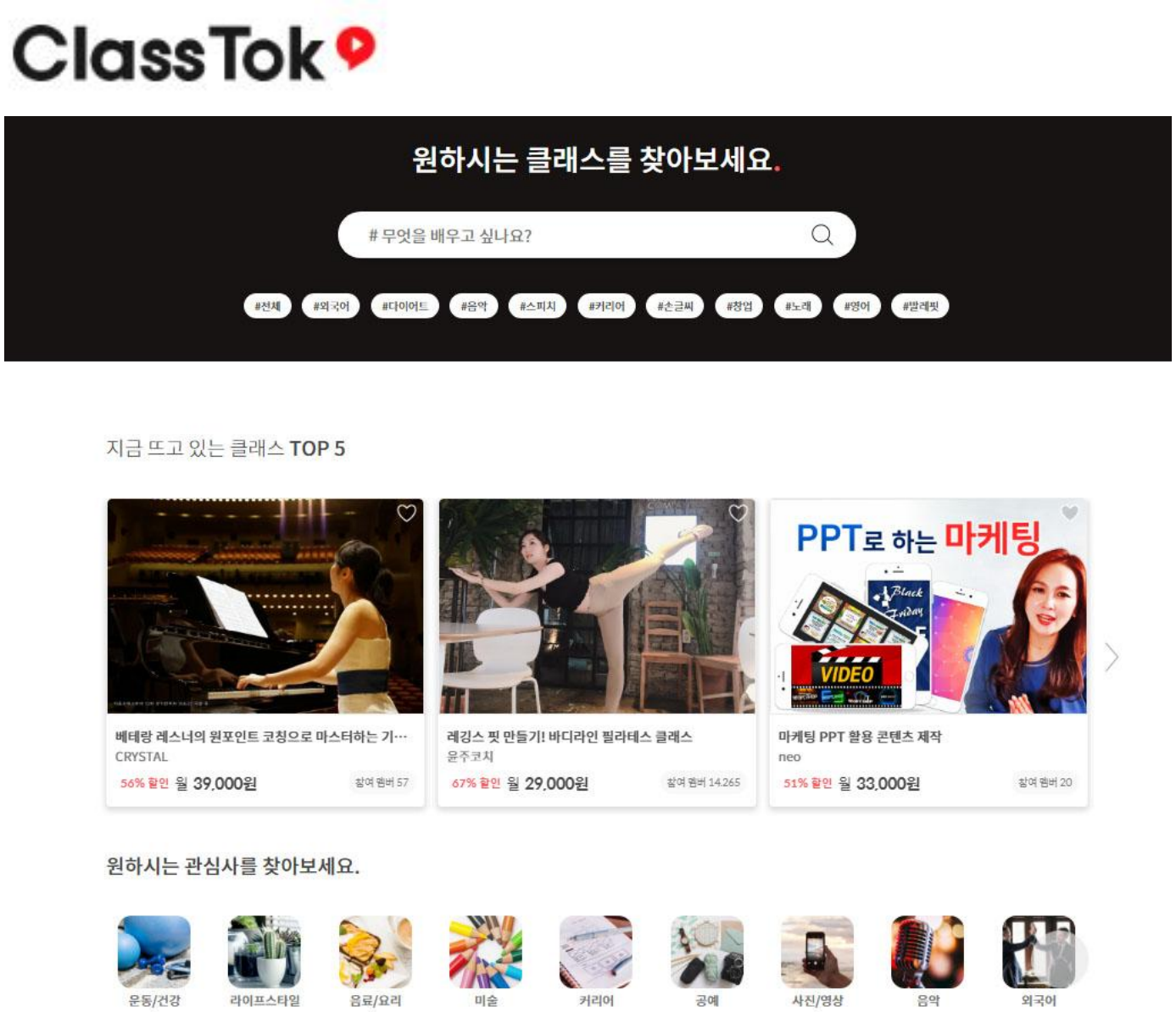
- selenium -> graphql post방식 requests
- 반응형 웹페이지
- 업계 1위 점유율, 가파른 성장세에 주목
- 검색을 위한 카테고리 분류를 위해
  취미, 수익창출, 직무교육 카테고리 크롤링 진행
  (타 사이트와 비교)

taling



Crawling Method

- scrapy -> BeautifulSoup
- 반응형 웹페이지
- 오프라인 수업-> 타이틀에 지역 추가표기
- 'soldout' 수업은 제외
- 세 개의 플랫폼 중 카테고리가 가장 체계적으로 분류되어 있어 해당 사이트를 기준으로 카테고리 분류(검색 시스템)



Crawling Method

- scrapy -> BeautifulSoup
- 반응형 웹페이지
- '#' 해시태그를 이용한 각 상세페이지로의 연결
- 현 상황에서는 구현하기 어려우나 차후 develop 고려

Database

- Mysql (RDBMS): 검색/키워드 추천을 위한 인덱싱의 중요성
- 각 플랫폼 별 컬럼 통일
- Table data 3개로 구분
  (검색용: search/ 저장용: save/ 신규강좌: new_class)

Crawling cycle

- 매일 1시간 간격 or 매일 1회 크롤링을 통한 데이터 수집내용 비교
- Mysql에 기존 데이터는 지우고 최신순으로 저장하는 방식에서 축적으로 변경
- 주기적으로 크롤링 결과 확인하면서 오류 체크

취미생활/자기계발 트렌드를 반영한 검색 시스템 : 서비스 구현

DATA BASE - 웹 페이지 연동

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

classtok: category 추가 - MultinomialNB 모델 사용

```
%%writefile model.py
import pickle
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.pipeline import Pipeline
from sklearn.metrics import classification_report

df = pd.read_csv('./crawling-repo-6/datas/taling_210315194501.csv', index_col=0)

train_x, test_x, train_y, test_y = train_test_split(
    df.title, df.category_1, test_size=0.1, random_state=13)

clf = Pipeline([
    ('vect', TfidfVectorizer()),
    ('clf', MultinomialNB(alpha=0.003))
])

model = clf.fit(train_x, train_y)

pred_y = model.predict(test_x)

with open('./model_cat1_210315194501.pkl', 'wb') as file:
    pickle.dump(model, file)
```

model 성능 평가

```python
# model 성능평가
import pickle
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.pipeline import Pipeline
from sklearn.metrics import classification_report
from sklearn.metrics import accuracy_score

files = ['taling_210318234509.csv', 'taling_210319104511.csv']
test_sizes = [0.05, 0.1, 0.15, 0.2, 0.25, 0.3]
alphas = [0.001, 0.005, 0.01, 0.05, 1, 2, 3]
results = []



# 210315194501 기준 model
with open('/home/ubuntu/notebooks/crawling-repo-6/model_cat1_210315194501.pkl', 'rb') as file:
    load_model = pickle.load(file)

for file in files:
    df = pd.read_csv(f'./crawling-repo-6/datas/{file}', index_col=0)

    for test_size in test_sizes:
        train_x, test_x, train_y, test_y = train_test_split(
        df.title, df.category_1, test_size=test_size, random_state=1)

        # 현재 사용중인 model 평가
        pred_y_train = load_model.predict(train_x)
        pred_y_test = load_model.predict(test_x)
        pred_y_total = load_model.predict(df.title)

        results.append({
            'dataframe': f'model_{file}',
            'Test_Size': test_size,
            'alpha': '없음',
            'Train_Acc': accuracy_score(train_y, pred_y_train),
            'Test_Acc': accuracy_score(test_y, pred_y_test),
            'Total_Acc': accuracy_score(df.category_1, pred_y_total),
        })
```

```python
        # 새로운 model 생성 및 평가

        for alpha in alphas:
            clf = Pipeline([
                ('vect', TfidfVectorizer()),
                ('clf', MultinomialNB(alpha=alpha))
            ])

            model = clf.fit(train_x, train_y)

            pred_y_train = model.predict(train_x)
            pred_y_test = model.predict(test_x)
            pred_y_total = model.predict(df.title)


            results.append({
                'dataframe': file,
                'alpha': alpha,
                'Test_Size': test_size,
                'Train_Acc': accuracy_score(train_y, pred_y_train),
                'Test_Acc': accuracy_score(test_y, pred_y_test),
                'Total_Acc': accuracy_score(df.category_1, pred_y_total),
            })

df_result = pd.DataFrame(results, columns=['dataframe', 'alpha', 'Test_Size',
df_result
```

model 성능 평가

```
1 df_result.sort_values(['Test_Acc'], ascending=False).head(30)
```

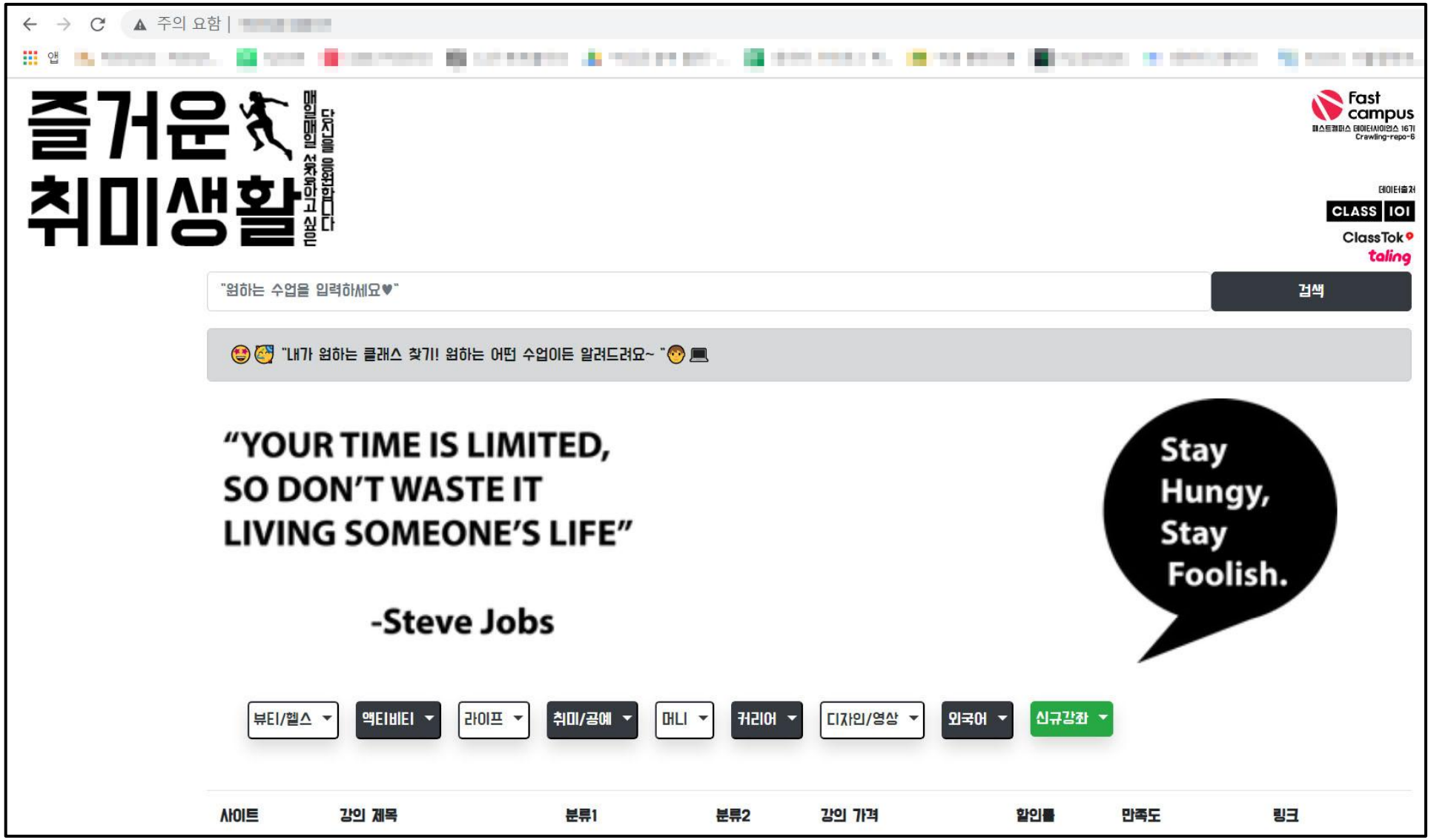
	dataframe	alpha	Test_Size	Train_Acc	Test_Acc	Total_Acc
48	model_taling_210319104511.csv	없음	0.05	0.970184	0.973451	0.970347
80	model_taling_210319104511.csv	없음	0.25	0.970493	0.969912	0.970347
88	model_taling_210319104511.csv	없음	0.30	0.970914	0.969027	0.970347
72	model_taling_210319104511.csv	없음	0.20	0.970678	0.969027	0.970347
56	model_taling_210319104511.csv	없음	0.10	0.970494	0.969027	0.970347
40	model_taling_210318234509.csv	없음	0.30	0.971275	0.967623	0.970179
64	model_taling_210319104511.csv	없음	0.15	0.970841	0.967552	0.970347
24	model_taling_210318234509.csv	없음	0.20	0.971831	0.963576	0.970179
32	model_taling_210318234509.csv	없음	0.25	0.972607	0.962898	0.970179
8	model_taling_210318234509.csv	없음	0.10	0.971036	0.962472	0.970179
16	model_taling_210318234509.csv	없음	0.15	0.971926	0.960294	0.970179
0	model_taling_210318234509.csv	없음	0.05	0.970930	0.955947	0.970179
75	taling_210319104511.csv	0.01	0.20	0.998340	0.745575	0.947776
67	taling_210319104511.csv	0.01	0.15	0.997917	0.744838	0.959947
68	taling_210319104511.csv	0.05	0.15	0.996876	0.743363	0.958840
76	taling_210319104511.csv	0.05	0.20	0.997234	0.740044	0.945784
59	taling_210319104511.csv	0.01	0.10	0.997787	0.736726	0.971675
74	taling_210319104511.csv	0.005	0.20	0.998617	0.735619	0.946006
49	taling_210319104511.csv	0.001	0.05	0.998369	0.734513	0.985174
66	taling_210319104511.csv	0.005	0.15	0.998178	0.734513	0.958619

```
1 df_result[df_result['dataframe']==f'model_{files[0]}']
```

	dataframe	alpha	Test_Size	Train_Acc	Test_Acc	Total_Acc
0	model_taling_210318234509.csv	없음	0.05	0.970930	0.955947	0.970179
8	model_taling_210318234509.csv	없음	0.10	0.971036	0.962472	0.970179
16	model_taling_210318234509.csv	없음	0.15	0.971926	0.960294	0.970179
24	model_taling_210318234509.csv	없음	0.20	0.971831	0.963576	0.970179
32	model_taling_210318234509.csv	없음	0.25	0.972607	0.962898	0.970179
40	model_taling_210318234509.csv	없음	0.30	0.971275	0.967623	0.970179

```
1 df_result[df_result['dataframe']==f'model_{files[1]}']
```

	dataframe	alpha	Test_Size	Train_Acc	Test_Acc	Total_Acc
48	model_taling_210319104511.csv	없음	0.05	0.970184	0.973451	0.970347
56	model_taling_210319104511.csv	없음	0.10	0.970494	0.969027	0.970347
64	model_taling_210319104511.csv	없음	0.15	0.970841	0.967552	0.970347
72	model_taling_210319104511.csv	없음	0.20	0.970678	0.969027	0.970347
80	model_taling_210319104511.csv	없음	0.25	0.970493	0.969912	0.970347
88	model_taling_210319104511.csv	없음	0.30	0.970914	0.969027	0.970347



Flask를 이용한 웹페이지 구현 및 DB 연동

베이킹 ① 검색어 입력

검색

"내가 원하는 클래스 찾기! 원하는 어떤 수업이든 알려드려요~ "

② '검색' 버튼 클릭!

뷰티/헬스 | 액티비티 | 라이프 | 취미/공예 | 머니 | 커리어 | 디자인/영상 | 외국어 | 신규강좌

사이트	강의 제목	분류1	분류2	강의 가격	할인률	만족도	새창보기
클래스101	[sales]도둑이 빵만 먹다가 돌아간 그 곳! 써니브레드의 글루텐프리 베이킹 클래스	취미	요리/음료	총 209,500원	0	98.1%	[sales]도둑이 빵만 먹다가 돌아간 그 곳! 써니브레드의 글루텐프리 베이킹 클래스
클래스101	[sales]유화 물감으로 꾸덕한 크림을 발라요, 그림으로 구워내는 홈베이킹 디저트	취미	미술	총 284,500원	0	100.0%	[sales]유화 물감으로 꾸덕한 크림을 발라요, 그림으로 구워내는 홈베이킹 디저트
클래스101	[sales]비건, 글루텐 프리, 키토제닉 베이킹도 예쁘고 맛있을 권리가 있다!	취미	요리/음료	총 314,500원	0	66.7%	[sales]비건, 글루텐 프리, 키토제닉 베이킹도 예쁘고 맛있을 권리가 있다!
클래스101	[sales]원리부터 이해해요, 감각적인 비건 無밀가루 디저트 베이킹	취미	요리/음료	총 189,500원	0	98.7%	[sales]원리부터 이해해요, 감각적인 비건 無밀가루 디저트 베이킹
클래스101	[funding]르꼬르동블루 출신 쉐프와 함께, 손쉬운 24가지 베이킹 레시피	취미	요리/음료	총 299,500원	0	평가 없음	[funding]르꼬르동블루 출신 쉐프와 함께, 손쉬운 24가지 베이킹 레시피

③ 검색결과 Table data 형태로 출력! (데이터연결)

검색결과: 상세페이지로 연결되는 2가지 방법

뷰티/헬스 | 액티비티 | 라이프 | 취미/공예 | 머니 | 커리어 | 디자인/영상 | 외국어 | 신규강좌

상세페이지로 연결 ②

사이트	강의 제목	분류1	분류2	강의 가격	할인률	만족도	새창보기
클래스101	[sales]도둑이 빵만 먹다가 돌아간 그 곳! 써니브레드의 글루텐프리 베이킹 클래스	취미	요리/음료	총 209,500원	0	98.1%	[sales]도둑이 빵만 먹다가 돌아간 그 곳! 써니브레드의 글루텐프리 베이킹 클래스
클래스101	[sales]유화 물감으로 꾸덕한 크림을 발라요, 그림으로 구워내는 홈베이킹 디저트	취미	미술	총 284,500원	0	100.0%	[sales]유화 물감으로 꾸덕한 크림을 발라요, 그림으로 구워내는 홈베이킹 디저트
클래스101	[sales]비건, 글루텐 프리, 키토제닉 베이킹도 예쁘고 맛있을 권리가 있다!	취미	요리/음료	총 314,500원	0	66.7%	[sales]비건, 글루텐 프리, 키토제닉 베이킹도 예쁘고 맛있을 권리가 있다!
클래스101	[sales]원리부터 이해해요, 감각적인 비건 無밀가루 디저트 베이킹	취미	요리/음료	총 189,500원	0	98.7%	[sales]원리부터 이해해요, 감각적인 비건 無밀가루 디저트 베이킹
클래스101	[funding]르꼬르동블루 출신 쉐프와 함께, 손쉬운 24가지 베이킹 레시피	취미	요리/음료	총 299,500원	0	평가 없음	[funding]르꼬르동블루 출신 쉐프와 함께, 손쉬운 24가지 베이킹 레시피

상세페이지로 연결 ①

페이지내에서 상세페이지로 연결

신규강좌 추가

```
# 3. new_class table : 3일 내 새로운 강좌 table

# 3-1. new_class에서 3일이 지난 데이터 삭제
days_ago = (datetime.datetime.now() - datetime.timedelta(days=3)).strftime("%y%m%d%H%M%S")
QUERY = """
    SELECT *
    FROM crawled.new_class
"""
old_class_df = pd.read_sql(QUERY, engine, index_col=['index'])
old_class_df.reset_index(drop=True)
old_class_df = old_class_df[old_class_df['crawling_time'].astype(str) > days_ago]
old_class_df.reset_index(drop=True)

# 3-2. 지난 데이터에 없는 신규강좌 데이터 확인
new_class_df = pd.DataFrame(columns=['site', 'link', 'title', 'teacher', 'category_1', 'category_2', 's_price', 'discount', 'contentment', 'crawling_time'])
for i in range(len(class101_df)):
    if class101_df['title'].tolist()[i] not in compare_df['title'].tolist():
        new_class_df = new_class_df.append(class101_df.iloc[i])

new_class_df.reset_index(drop=True)

# 3-3. 신규강좌 데이터 저장
new_class_df = new_class_df.append(old_class_df)
new_class_df.reset_index(drop=True)
new_class_df.to_sql(name='new_class', con=engine, if_exists='replace')
```

크롤링 시간을 기준으로 3일 전 데이터 삭제

신규강좌 데이터 확인

```
print('db저장완료')
print('search_df:', len(search_df))
print('class101_df:', len(class101_df))
print('new_class_df:', len(new_class_df))
print('time: ', round((time.time() - start)/60, 1), '분', sep='')
```

신규강좌에 마우스 커서를 올리고

클릭!

새로 추가된 신규강좌 목록이 촤라락

사이트	강의 제목	분류1	분류2	강의 가격	할인률	만족도	새창보기
신규 강좌 **** !!!! "새로 추가된 강좌를 확인하세요!"!!!!							
사이트	강의 제목	분류1	분류2	강의 가격	할인률	만족도	새창보기
신규 강좌 **** !!!! "새로 추가된 강좌를 확인하세요!"!!!!							
사이트	강의 제목	분류1	분류2	강의 가격	할인률	만족도	새창보기
탈잉	[삼성][1:1메이크업레슨] 5주안에 끝내는 나만의 인생메이크업 찾기	뷰티/헬스	메이크업	월 35,000원	0	0	[삼성][1:1메이크업레슨] 5주안에 끝내는 나만의 인생메이크업 찾기
탈잉	[잠실]★2시간 내 논문구조 완벽 이해★ 아무도 알려주지 않는 논문쓰기의 비밀	라이프	출판/글쓰기	월 35,000원	0	5	[잠실]★2시간 내 논문구조 완벽 이해★ 아무도 알려주지 않는 논문쓰기의 비밀
탈잉	[지역 없음]완플리는 내 운, 타고난 내 팔자 궁금해요! 내 명을 알고 준비하자!	라이프	사주/타로	월 30,000원	0	0	[지역 없음]완플리는 내 운, 타고난 내 팔자 궁금해요! 내 명을 알고 준비하자!
탈잉	[은평]타로로 보는 나는? 타로를 통한 내안의 나를 찾아서...	라이프	사주/타로	월 20,000원	0	0	[은평]타로로 보는 나는? 타로를 통한 내안의 나를 찾아서...
탈잉	[온라인 Live,녹화영상][온라인Live] 네이버 카페로 월 300벌기(온라인 건물주 되기)-블로그,인스타그램,주식,전자책 투잡 재테크	머니	마케팅	월 30,000원	0	5	[온라인 Live,녹화영상][온라인Live] 네이버 카페로 월 300벌기(온라인 건물주 되기)-블로그,인스타그램,주식,전자책 투잡 재테크
탈잉	[하남](원데이) 사장님한테 예쁨받는 비법, VBA 쉽고 빠르게 배우기!	커리어	웹개발	월 30,000원	0	5	[하남](원데이) 사장님한테 예쁨받는 비법, VBA 쉽고 빠르게 배우기!
탈잉	[일산,온라인 Live]돈 벌어주는 '포토샵' 노하우 기초 부터 알려드립니다.	디자인/영상	그래픽디자인	월 25,000원	0	0	[일산,온라인 Live]돈 벌어주는 '포토샵' 노하우 기초 부터 알려드립니다.

아쉬운 점

Enjoy your stylish business and campus life with BIZCAM

세 사이트의 카테고리 분류를 나주지 않은 상태여서
버튼 링크를 누르더라도 현상태에서는 연결페이지로의 이동이 불가
차후 Develop 고려

Issue

Fast campus

패스트캠퍼스 데이터사이언스 16기
Crawling-repo-6

데이터출처

CLASS 101

ClassTok

taling

앞으로의 과제

- 정확한 키워드 검색을 위한 자연어 처리 및 형태소 분석
- 카테고리 분류 기준 (세 사이트를 아우를 수 있는 기준)
- DB 관리
- Ini, config 설정법
- 각 카테고리별 연결 페이지 생성
- 웹 페이지 구성요소 고민

고원진

장지혜

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

감사합니다.

Enjoy your stylish business and campus life with BIZCAM

Thank You :D