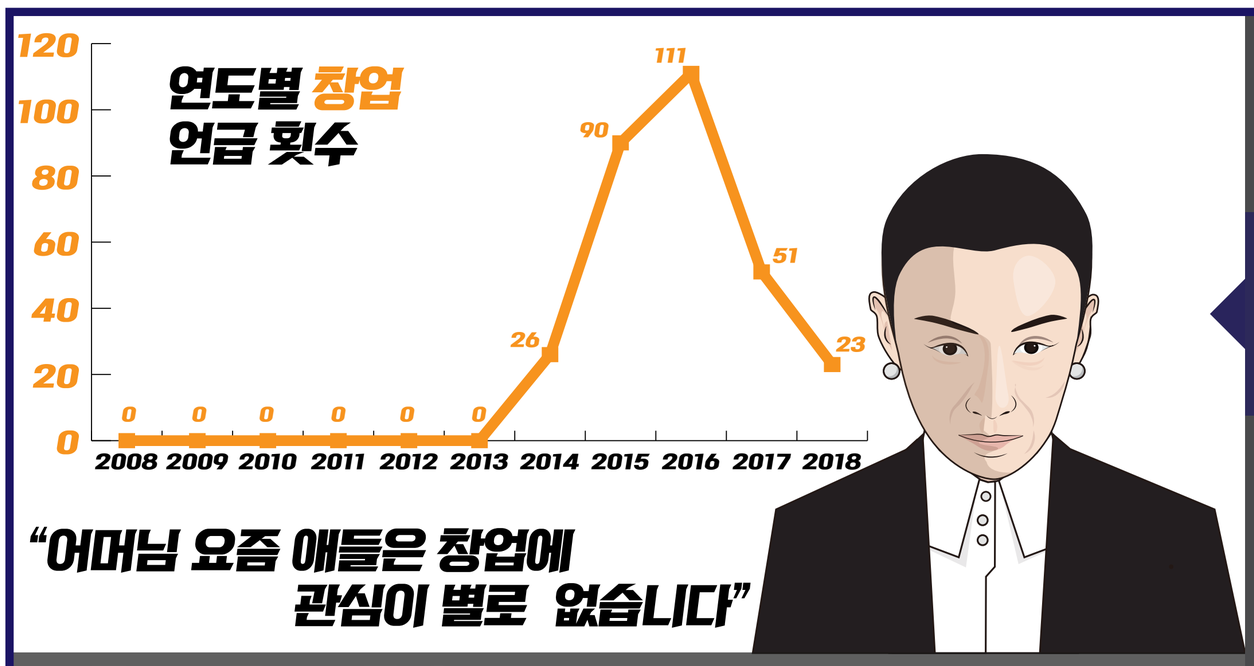
LangCon 2019 키워드로 본 대학신문

파이썬으로 한림학보 분석하기

한림 데이터랩 / 박형민, 이상규

대학보도 | 심층보도 | 기 획 | 취 업 | 학술교양 | 생활문화 | 오피니언 | 한림학보 검색



대학생 청년 창업 2018년 현재 광풍은 지속되는가?

2013년 시작된 대학가 창업 열풍은 2016년 절정을 찍은 뒤 점차 감소하고 있는 것으로 보인다.

다사다난한 2018년 올해의 단어 Top 8

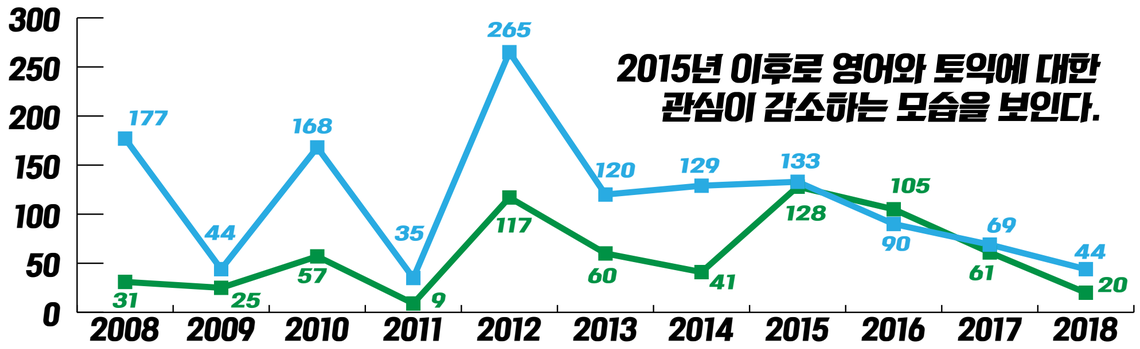
대학생 청년 창업 지금도 광풍은 지속되는가

음주와 학업 대학생의 필수 딜레마

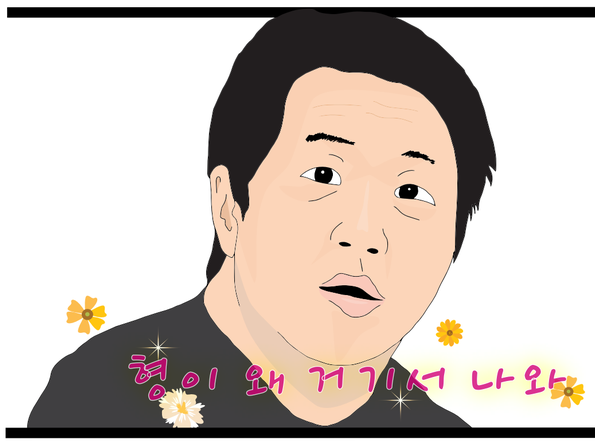
대학생활 라이프 자취인가 기숙사인가

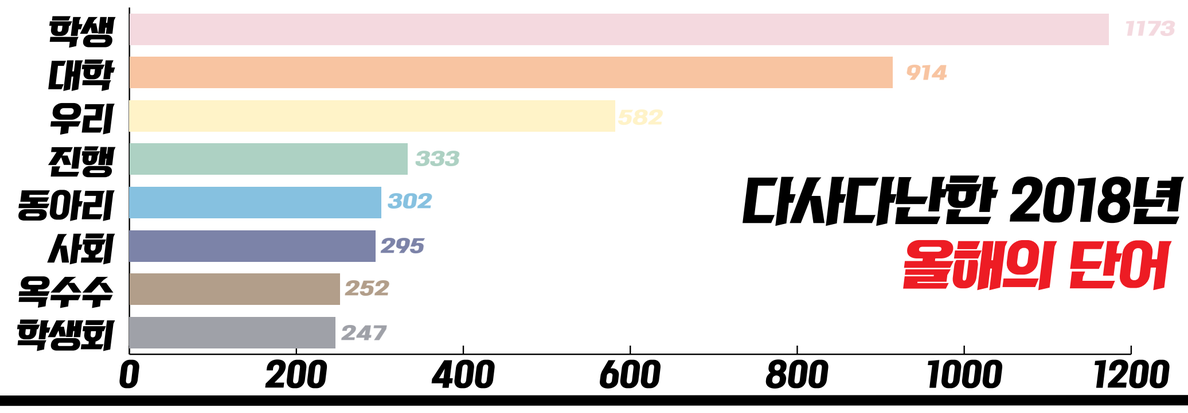
전공이냐 교양이냐 그것이 문제로다

영어와 토익, 학생들의 관심은?



2015년 이후로 영어와 토익에 대한 관심이 감소하는 모습을 보인다.

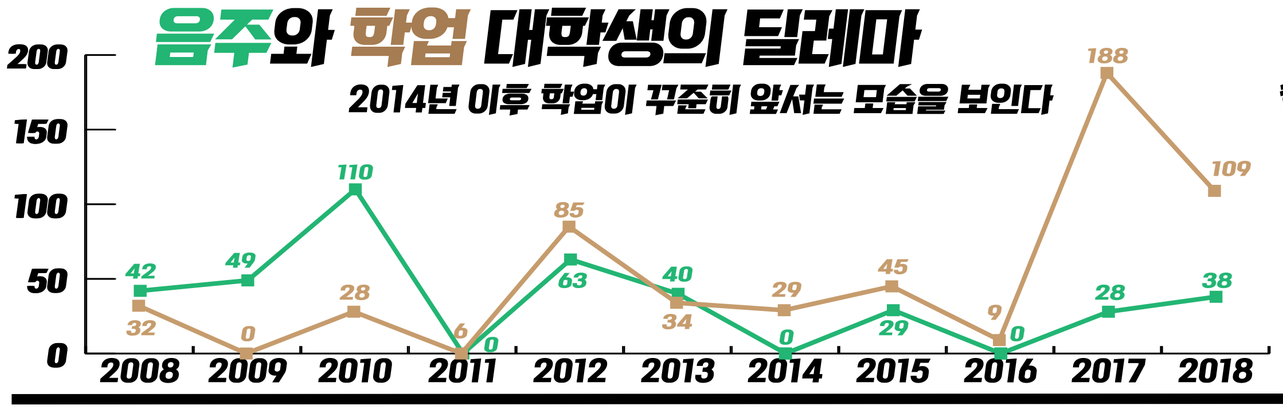




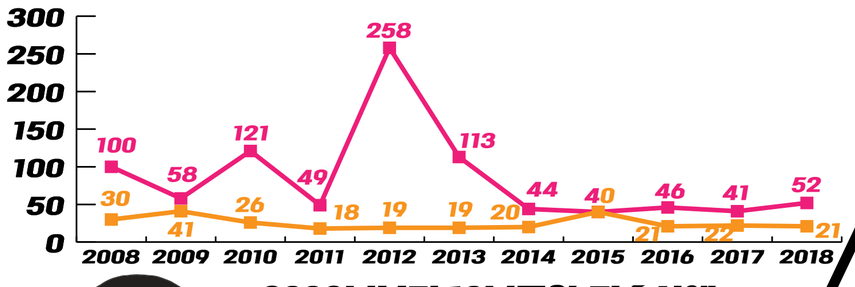
다사다난한 2018년 올해의 단어

음주와 학업 대학생의 딜레마

2014년 이후 학업이 꾸준히 앞서는 모습을 보인다

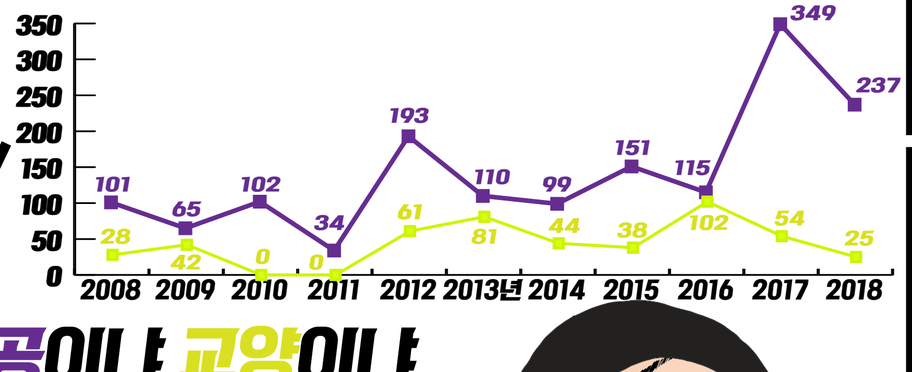


대학생활 자취인가 기숙사인가



2008년부터 10년동안 기숙사에 대한 관심이 더 높았다.

혹시 너 자취하니?



전공이냐 교양이냐 그것이 문제로다

10년 내내 전공이 교양을 앞서는 흐름이 계속 되어왔다.

어서와 대학강의는 처음이지?

음주와 학업 대학생의 딜레마

파이썬 / 사용 코드

판다스를 활용한 웹크롤링

```
In [1]: import pandas as pd
        from bs4 import BeautifulSoup
        import requests
        import time
```

```
In [2]: PREFIX = 'http://news.hallym.ac.kr/news/'

        category_dict = {
            '대학보도':'sc_sub_section_code=S2N1',
            '심층보도':'sc_sub_section_code=S2N20',
            '기획':'sc_sub_section_code=S2N2',
            '학술교양':'sc_section_code=combine',
            '생활문화':'sc_sub_section_code=MM',
            '오피니언':'sc_sub_section_code=S2N3'
        }
```

```
In [4]: def get_doc(doc_url):
            '''
            입력받은 url의 기사내용 및 메타정보 크롤링하여 반환
            '''
            category_rule = 'td.View_Local'
            date_rule='div.View_Time'
            title_rule='div.View_Title'
            article_rule='td#articleBody'

            r = requests.get(doc_url, timeout=30)
            soup = BeautifulSoup(r.content, 'lxml')
            category = soup.select_one(category_rule).text.split('>')[-1].strip()
            date = soup.select_one(date_rule).text.split()[1].strip()
            title = soup.select_one(title_rule).text
            article = soup.select_one(article_rule).text

            doc = {}
            doc['category'] = category.strip()
            doc['date'] = date.strip()
            doc['title'] = title.strip()
            doc['article'] = article.strip()
            doc['url'] = doc_url
            return doc
```

```
In [5]: def get_doc_urls(list_url):
            '''
            기사 리스트 화면에서 기사 url 리스트 반환
            '''
            r = requests.get(list_url, timeout=30)
            soup = BeautifulSoup(r.content, 'lxml')
            list_rule = 'td.ArtList_Title a'
            items = soup.select(list_rule)
            doc_urls = [PREFIX + item.attrs['href'].strip() for item in items]
            return doc_urls
```

Konlpy - 형태소 분석기 활용

```
]: tokenized = []
   for element in docs.values:
       tokenized.append([element[0], okt.nouns(element[3]), okt.nouns(element[4])])
```

```
]: pd.DataFrame(docs).to_excel("hallym_news.xlsx")
```

```
]: pd.DataFrame(tokenized).to_excel("hallym_news.tokenized.xlsx")
```

```
]: docs = pd.read_excel("hallym_news.xlsx")
```

```
]: docs
```

]:

	0	1	2	
0	http://news.hallym.ac.kr/news/articleView.html...	보도	2018.12.08	[한림의 천사를 찾 지 방법
1	http://news.hallym.ac.kr/news/articleView.html...	보도	2018.12.08	1인미디어실 생 트홀 · 산학협력
2	http://news.hallym.ac.kr/news/articleView.html...	보도	2018.12.08	소프트웨어 주간 품 논란

파이썬 함수를 활용한 데이터 분석

전체

```
In [4]: al = pd.read_excel("news_token.xlsx")

        s = []
        for row in al.values:
            s.append(eval(row[1])+eval(row[2]))

        ss = ",".join(map(str,s))

        getTokens(ss)
        getTypeFreq(ss)

        sss = getTypeFreq(ss)

        sss.items()
        sorted(sss.items(), key=lambda t : t[1], reverse=True)
```

```
Out[4]: [("'", 4402050),
         (',', 2201024),
         ('것', 39428),
         ('학생', 28897),
         ('수', 24012)]
```

LangCon 2019

X

한림대학교 HALLYM UNIVERSITY