[웹크롤링 _ 위키피디아 사이트 분석 및 시각화]

In [109…

```
# -*- coding: utf-8 -*-

%matplotlib inline

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

import warnings
warnings.filterwarnings("ignore")
```

In [110…

```
from selenium import webdriver
from bs4 import BeautifulSoup
import re # 정규식 표현을 위한 모듈
```

In [114…

```
# 크롤링한 데이터를 데이터 프레임으로 만들기 위해 준비
excutable_path = "chromedriver.exe"
base_url = 'https://ko.wikipedia.org/'
driver = webdriver.Chrome(executable_path=excutable_path)
columns = ["title", "category", "content_text"]
df = pd.DataFrame(columns=columns)
title_list = urlmake(base_link).select('.mw-title')
searchnum = input("검색 수 : ")

num = 0
for i in range(int(searchnum)):
    title = title_list[i].select_one('.mw-changeslist-title').text
    link = title_list[i].select_one('a').attrs['href']
    contents = urlmake(link).select_one('.mw-parser-output>p')
    category = urlmake(link).select_one('.mw-normal-catlinks > ul
    num += 1

    if category == None:
        category = ''
    else:
        category = category.text

    if contents == None:
        contents = ''
    else:
        contents = contents.text
```

```
    row = [title,category,re.sub(r'[₩n]',' , ',contents)]
    series = pd.Series(row, index=df.columns)
    df = df.append(series, ignore_index=True)
df
```

검색 수 : 10

Out[114...

	title	category	content_text
0	영인운수	서울특별시의 시내버스 기업	영인운수(永仁運輸)는 서울특별시의 시내버스 업체다. ,
1	이대연	1966년 출생	이대연(李大淵[1], 1966년 11월 13일 ~)은 대한민국의 배우이다. 198...
2	영인운수	서울특별시의 시내버스 기업	영인운수(永仁運輸)는 서울특별시의 시내버스 업체다. ,
3	두샨 블라호비치	2000년 출생	두샨 블라호비치(세르비아어: Душан Влаховић, 영어: Dušan Vlah...
4	월드 오브 워쉽	2015년 비디오 게임	《월드 오브 워쉽》(영어: World of Warships)은 20세기의 역사적인 ...
5	이청청	대한민국의 패션 디자이너	이청청(李청청)은 대한민국의 패션 디자이너 겸 기업가이다. ,
6	사용자:최고의 편집자/참가한 에디터돈		최고의 편집자가 참가할 또는 참가한 에디터돈 목록입니다. 인증을 위한 목적이며 저의...
7	월드 오브 워플레인	2013년 비디오 게임	《월드 오브 워플레인》 (영어 : World of Warplanes)는 Persha...
8	사용자:최고의 편집자/참가한 에디터돈		최고의 편집자가 참가할 또는 참가한 에디터돈 목록입니다. 인증을 위한 목적이며 저의...
9	위키백과:사용자 관리 요청/2022년 제3주		좇념파닭 (토론 · 기여[전체 위키 기여 · 삭제된 기여] · 기록[차단 기록 · ...

In [112...

```
category = urlmake(link).select('#mw-normal-catlinks > ul > li')
category
```

Out[112...

```
[<li><a href="/wiki/%EB%B6%84%EB%A5%98:%EC%9D%8C%EB%A3%8C" title="분류:음료">음료</a></li>,
 <li><a href="/wiki/%EB%B6%84%EB%A5%98:%EC%9D%8C%EB%A3%8C%EC%88%98" title="분류:음료수">음료수</a></li>,
 <li><a href="/wiki/%EB%B6%84%EB%A5%98:%EC%BD%94%EC%B9%B4%EC%BD%9C%EB%9D%BC%EC%9D%98_%EC%A0%9C%ED%92%88" title="분류:코카콜라의 제품">코카콜라의 제품</a></li>]
```

In [ ]: