[웹크롤링_나무위키 사이트 분석 및 시각화]

<step1.크롤링>:크롤링으로 웹데이터 가져오기

[웹크롤링 라이브러리 사용하기]

- 파이썬에서는 BeautifulSoup과 requests 라이브러리로 웹 크롤러를 만들 수 있음
- requests는 특정 URL로부터 HTML 문서를 가져오는 작업을 수행
- 나무위키와 같은 페이지는 HTML 문서가 Javascript로 동적 로딩되는 경우가 있음
- requests대신 selenium 라이브러리를 이용해 크롬 브라우저로 동적 웹 크롤링 수행
- selenium은 웹 브라우저를 자동으로 구성해주는 라이브러리
- selenium을 사용하기 위해 크롬 드라이버를 이용해 크롬 브라우저 자동으로 구동 => 크롬 드라이버 필요

[BeautifulSoup과 selenium을 이용한 웹 크롤링]

- (env_name) `pip install selenium`

[크롬 브라우저 업데이트 및 크롬 드라이버 설치]

- 크롬 브라우저 설정에서 최신 버전으로 업데이트
- 크롬 드라이버 사이트에서 브라우저 버전에 맞는 드라이버 다운로드
  - https://chromedriver.chromium.org/downloads (https://chromedriver.chromium.org/downloads)
- chromedriver.exe 파일을 노트북 파일 경로에 저장

In [1]:

```
pip install selenium
```

```
Requirement already satisfied: trio-websocket~=0.9 in c:₩users₩yj₩anaconda3₩lib₩si
te-packages (from selenium) (0.9.2)
Requirement already satisfied: urllib3[secure]~=1.26 in c:₩users₩yj₩anaconda3₩lib
₩site-packages (from selenium) (1.26.7)
Requirement already satisfied: cffi>=1.14 in c:₩users₩yj₩anaconda3₩lib₩site-packag
es (from trio~=0.17->selenium) (1.14.6)
Requirement already satisfied: idna in c:₩users₩yj₩anaconda3₩lib₩site-packages (fr
om trio~=0.17->selenium) (3.2)
Requirement already satisfied: sniffio in c:₩users₩yj₩anaconda3₩lib₩site-packages
(from trio~=0.17->selenium) (1.2.0)
Requirement already satisfied: async-generator>=1.9 in c:₩users₩yj₩anaconda3₩lib₩s
ite-packages (from trio~=0.17->selenium) (1.10)
Requirement already satisfied: outcome in c:₩users₩yj₩anaconda3₩lib₩site-packages
(from trio~=0.17->selenium) (1.1.0)
Requirement already satisfied: sortedcontainers in c:₩users₩yj₩anaconda3₩lib₩site-
packages (from trio~=0.17->selenium) (2.4.0)
Requirement already satisfied: attrs>=19.2.0 in c:₩users₩yj₩anaconda3₩lib₩site-pac
kages (from trio~=0.17->selenium) (21.2.0)
Requirement already satisfied: pycparser in c:₩users₩yj₩anaconda3₩lib₩site-package
```

In [40]:

```
from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from bs4 import BeautifulSoup as bts
import re #정규식 표현을 위한 모듈
import warnings
import pandas as pd
warnings.filterwarnings('ignore')
```

In [24]:

```
#윈도우용 크롬 웹 드라이버 실행 경로(window)지정
executable_path="chromedriver.exe"
driver=webdriver.Chrome(executable_path=executable_path)

#사이트의 html구조에 기반하여 크롤링을 수행
source_url="https://namu.wiki/RecentChanges"
driver.get(source_url)

element=WebDriverWait(driver,10).until(EC.presence_of_element_located((By.CLASS_NAME,"app")))
req=driver.page_source
req
```

Out[24]:


```
'<html><head><link href="/skins/senkawa/6.0ec579cd0a387a25b691.css" rel="styleshee
t"><link href="/skins/senkawa/3.c2f4326b616fb16062e8.css" rel="stylesheet"><script
async="" src="/cdn-cgi/bm/cv/669835187/api.js"></script><style type="text/css">.re
size-observer[data-v-b329ee4c]{position:absolute;top:0;left:0;z-index:-1;width:10
0%;height:100%;border:none;background-color:transparent;pointer-events:none;displa
y:block;overflow:hidden;opacity:0}.resize-observer[data-v-b329ee4c] object{displa
y:block;position:absolute;top:0;left:0;height:100%;width:100%;overflow:hidden;poin
ter-events:none;z-index:-1}</style><link rel="stylesheet" type="text/css" href="/s
kins/senkawa/10.4d02833f9fb9e7f9340e.css"><script charset="utf-8" src="/skins/senk
awa/10.4d02833f9fb9e7f9340e.js"></script><title>최근 변경내역 - 나무위키</title><l
ink data-n-head="1" rel="canonical" href="https://namu.wiki/RecentChanges"><link d
ata-n-head="1" rel="search" type="application/opensearchdescription+xml" title="나
무위키" href="/opensearch.xml"><link data-n-head="1" rel="icon" href="/favicon.sv
g" sizes="any" type="image/svg+xml"><link data-n-head="1" rel="icon" href="/favico
n-32.png" sizes="32x32" type="image/png"><link data-n-head="1" rel="icon" href="/f
avicon-192.png" sizes="192x192" type="image/png"><link data-n-head="1" rel="apple-
touch-icon" href="/img/apple_icon.png"><meta data-n-head="1" charset="utf-8"><meta
data-n-head="1" name="viewport" content="user-scalable=no, initial-scale=1.0, maxi
```

In [25]:

```
soup=bts(req,'html.parser')
soup
```

Out[25]:

```
<html><head><link href="/skins/senkawa/6.0ec579cd0a387a25b691.css" rel="styleshee
t"/><link href="/skins/senkawa/3.c2f4326b616fb16062e8.css" rel="stylesheet"/><scri
pt async="" src="/cdn-cgi/bm/cv/669835187/api.js"></script><style type="text/cs
s">.resize-observer[data-v-b329ee4c]{position:absolute;top:0;left:0;z-index:-1;wid
th:100%;height:100%;border:none;background-color:transparent;pointer-events:none;d
isplay:block;overflow:hidden;opacity:0}.resize-observer[data-v-b329ee4c] object{di
splay:block;position:absolute;top:0;left:0;height:100%;width:100%;overflow:hidden;
pointer-events:none;z-index:-1}</style><link href="/skins/senkawa/10.4d02833f9fb9e
7f9340e.css" rel="stylesheet" type="text/css"/><script charset="utf-8" src="/skin
s/senkawa/10.4d02833f9fb9e7f9340e.js"></script><title>최근 변경내역 - 나무위키</ti
tle><link data-n-head="1" href="https://namu.wiki/RecentChanges" rel="canonical"/>
<link data-n-head="1" href="/opensearch.xml" rel="search" title="나무위키" type="a
pplication/opensearchdescription+xml"/><link data-n-head="1" href="/favicon.svg" r
el="icon" sizes="any" type="image/svg+xml"/><link data-n-head="1" href="/favicon-3
2.png" rel="icon" sizes="32x32" type="image/png"/><link data-n-head="1" href="/fav
icon-192.png" rel="icon" sizes="192x192" type="image/png"/><link data-n-head="1" h
ref="/img/apple_icon.png" rel="apple-touch-icon"/><meta charset="utf-8" data-n-hea
d="1"/><meta content="user-scalable=no, initial-scale=1.0, maximum-scale=5.0, mini
```

In [26]:

```
contents_table = soup.find(name="table") #find 함수를 이용해 태그명이 table인 것을 찾기
table_body = contents_table.find(name= "tbody") #table 안 tbody 태그인 것 찾기
table_rows = table_body.find_all(name="tr") #table tbody 안 tr태그인 것 찾기 => [ ]의 요소로 담김
#table_body
table_rows[0]
```

Out[26]:

```
<tr class="" data-v-349171da=""><td data-v-349171da=""><a data-v-349171da="" href="/
w/%EB%B0%B1%ED%98%84%EC%A4%91%ED%95%99%EA%B5%90">백현중학교</a> <a data-v-349171da
="" href="/history/%EB%B0%B1%ED%98%84%EC%A4%91%ED%95%99%EA%B5%90">[역사]</a> <a data
-v-349171da="" href="/diff/%EB%B0%B1%ED%98%84%EC%A4%91%ED%95%99%EA%B5%90?rev=315&am
p;oldrev=314">[비교]</a> <a data-v-349171da="" href="/discuss/%EB%B0%B1%ED%98%84%EC%
A4%91%ED%95%99%EA%B5%90">[토론]</a> <span data-v-349171da="">(<span class="" data-v-
349171da="" data-v-6cbb5b59="">0</span>)</span></td> <td data-v-349171da=""><div cla
ss="v-popover" data-v-349171da="" data-v-9a113440=""><div aria-describedby="popover_
qoeocvfqwj" class="trigger" style="display: inline-block;"><a class="IM9WYpTN" data-
v-9a113440="">kimapple</a> </div> </div> <!-- --></td> <td data-v-349171da=""><time
data-v-349171da="" datetime="2022-01-18T05:16:46.000Z">2022-01-18 14:16:46</time></t
d></tr>
```

attrs와 get의 차이

- attrs['']: 원하는 정보를 딕셔너리 형태로 가져와서 key와 value로 값을 가져옴.
- get(''): 원하는 정보의 속성값을 가져옴

In [36]:

```
#특성 속성 값을 추출
page_url_base = "https://namu.wiki" #베이스 url 정의
page_urls = [] # href 속성값을 담기 위한 빈 리스트

for i in range(0, len(table_rows)): #table_rows의 길이만큼 반복
    first_td = table_rows[i].find_all('td')[0] #td가 3개 있는데 0번째에 원하는 href 가 있음
    td_url = first_td.find_all('a')
    if len(td_url) > 0:
        page_url = page_url_base + td_url[0].get('href') #나무위키주소+get() 태그가 가지고 있는 속성
        #print(page_url)
        if "png" not in page_url:
            page_urls.append(page_url)



for page in page_urls:
    print(page)
#     page_urls = list(set(page_urls)) #중복 url 제거

# print(len(page_urls))
# for page in page_urls:
#     print(page)
```

```
https://namu.wiki/w/%EB%B0%B1%ED%98%84%EC%A4%91%ED%95%99%EA%B5%90 (https://namu.wi
ki/w/%EB%B0%B1%ED%98%84%EC%A4%91%ED%95%99%EA%B5%90)
https://namu.wiki/w/%ED%99%94%EC%84%B1%20%EB%B2%84%EC%8A%A4%2073 (https://namu.wik
i/w/%ED%99%94%EC%84%B1%20%EB%B2%84%EC%8A%A4%2073)
https://namu.wiki/w/%EB%82%98%20%ED%98%BC%EC%9E%90%EB%A7%8C%20%EB%A0%88%EB%B2%A8%E
C%97%85 (https://namu.wiki/w/%EB%82%98%20%ED%98%BC%EC%9E%90%EB%A7%8C%20%EB%A0%88%E
B%B2%A8%EC%97%85)
https://namu.wiki/w/%EC%A7%84(%EB%B0%A9%ED%83%84%EC%86%8C%EB%85%84%EB%8B%A8) (http
s://namu.wiki/w/%EC%A7%84(%EB%B0%A9%ED%83%84%EC%86%8C%EB%85%84%EB%8B%A8))
https://namu.wiki/w/%ED%8C%8C%EC%9D%BC:COMPLETE%20WITH%20YOU%20KIM%20DONG%20HYUN%2
02.jpg (https://namu.wiki/w/%ED%8C%8C%EC%9D%BC:COMPLETE%20WITH%20YOU%20KIM%20DONG%
20HYUN%202.jpg)
https://namu.wiki/w/%EC%9D%B4%EC%88%98%EC%A0%95(%EA%B5%90%EC%88%98) (https://namu.
wiki/w/%EC%9D%B4%EC%88%98%EC%A0%95(%EA%B5%90%EC%88%98))
https://namu.wiki/w/%EC%B5%9C%EA%B0%95%EC%B0%BD%EB%AF%BC (https://namu.wiki/w/%EC%
B5%9C%EA%B0%95%EC%B0%BD%EB%AF%BC)
https://namu.wiki/w/%ED%96%89%EB%B3%B5%20(Happiness) (https://namu.wiki/w/%ED%96%8
9%EB%B3%B5%20(Happiness))
https://namu.wiki/w/WWE%20%ED%8F%90%EA%B8%B0%20%EA%B0%81%EB%B3%B8 (https://namu.wi
```

드라이버를 for문을 돌려 필요한 자료 추출

In [42]:

```
# 크롤링한 데이터를 데이터 프레임으로 만들기 위해 준비
columns = ["title", "category", "content_text"]
df = pd.DataFrame(columns=columns)

#for page_url in page_urls:
for i in range(10):
    # 윈도우용 크롬 웹드라이버 실행 경로 (Windows) 지정
    excutable_path = "chromedriver.exe"
    driver = webdriver.Chrome(executable_path=excutable_path)
    # 크롬 드라이버를 통해 page_urls[0]번째 사이트의 HTML 문서 가져옴
    #driver.get(page_url)  # page_urls[i],  page_url의 정보를 가져옴
    driver.get(page_urls[i])  # page_urls[i],  page_url의 정보를 가져옴
    req = driver.page_source # 페이지 소스를 req에 저장
    soup = bts(req, 'html.parser') # html.parser로 파싱
    contents_table = soup.find(name="article") #  불러온 소스에서 태그명이 article인 요소 하나만 추출

    ### 타이틀 추출
    title = contents_table.find_all('h1')[0] # 태그명이 h1인 모든 태그 추출, article h1
    if title is not None:
        row_title = title.text.replace("₩n", " ")
    else:
        row_title = ""

    ### 카테고리 추출
    # 카테고리 정보가 없는 경우를 확인합니다.
    if len(contents_table.find_all("ul")) > 0: # article ul 로 검색한 결과 여러 ul 결과가 나올 경우
        category = contents_table.find_all("ul")[0] # 제일 첫번째 article ul 을 category로 설정
    else:
        category = None

    if category is not None:
        row_category = category.text.replace("₩n", " ")
    else:
        row_category = ""

    ### 내용 추출
    #contents_table.find_all(name="div", attrs={"class":"wiki-paragraph"})
    #div  태그 중  class 속성값이  wiki-paragraph인 요소를 추출
    content_paragraphs = contents_table.select("div.wiki-paragraph")
    #  내용으로 추출한 리스트를 하나의 문자열로 전처리
    content_corpus_list = [] # 내용 중 텍스트만 담을 빈 리스트 생성

    # content_paragraphs 리스트의 값을 순서대로 paragraphs에 대입
    if content_paragraphs is not None:
        for paragraphs in content_paragraphs:
            if paragraphs is not None:
                content_corpus_list.append(paragraphs.text.replace("₩n", " "))
            else:
                content_corpus_list.append("")
    else:
        content_corpus_list.append("")

    # 모든 정보를 하나의 데이터 프레임에 저장하기 위해서 시리즈 생성
    # 각 페이지의 정보를 추출하여 제목, 카테고리, 내용 순으로 행을 생성
    row = [row_title, row_category, "".join(content_corpus_list)]
    # 시리즈로 만듦
    series = pd.Series(row, index=df.columns)
    # 데이터 프레임에 시리즈를 추가, 한 페이지 당 하나의 행 추가
    df = df.append(series, ignore_index=True)
```

```
# 크롤링에 사용한 브라우저를 종료합니다.
driver.close()
```

In [43]:

```
df
```

Out[43]:

	title	category	content_text
0	백현중학교	성남시의 중학교1995년 개교나무위키 교육기관 프로젝트	은(는) 여기로 연결됩니다. 용인시 소재인 동명의 중학교에 대한 내용은 용인백...
1	화성 버스 73	화성시의 시내버스2008년 개업한 버스 노선2015년 개업한 버스 노선2022년 폐...	화성시 시내버스+ [펄치기 · 접기]광역급행M4130M4137M4434M444...
2	나 혼자만 레벨업	나 혼자만 레벨업웹소설/목록	나 혼자만 레벨업Only I level up장르현대 판타지, 헌터, 어반 판타지작가...
3	진(방탄소년단)	방탄소년단한국 남가수1992년 출생송파구 출신 인물안양시 출신 인물2013년 데뷔광...	은(는) 여기로 연결됩니다. 동음이의어에 대한 내용은 진 문서를의 번 문단을 의...
4	파일:COMPLETE WITH YOU KIM DONG HYUN 2.jpg	파일/AB6IX	이 파일은 나무위키에서 제한된 한도 안에서 쓰입니다.본 이미지는 퍼블릭 도메인 혹은...
5	이수정(교수)	1964년 출생서초구 출신 인물연세대학교 출신리라초등학교 출신대한민국의 사회과학 교...	로그인 후 편집 가능한 문서입니다.이 문서는이 문단은 토론을 통해 소속 정당이...
6	최강창민	최강창민	유노윤호최강창민 [전 멤버]영웅재중믹키유천시아준수 [한국 음반]앨범TRI-A...
7	행복 (Happiness)	Red Velvet/음반2014년 싱글	은(는) 여기로 연결됩니다. H.O.T.의 동명의 곡에 대한 내용은 행복(H....
8	WWE 폐기 각본	WWE	1. 개요2. 목록다음은 WWE에서 계획했으나 시행되지 못한채 폐기된 각본들이다.1...
9	파일:COMPLETE WITH YOU KIM DONG HYUN 1.jpg	파일/AB6IX	이 파일은 나무위키에서 제한된 한도 안에서 쓰입니다.본 이미지는 퍼블릭 도메인 혹은...

In [44]:

```
pip install konlpy
```

```
Collecting konlpy
  Downloading konlpy-0.6.0-py2.py3-none-any.whl (19.4 MB)
Collecting JPype1>=0.7.0
  Downloading JPype1-1.3.0-cp39-cp39-win_amd64.whl (362 kB)
Requirement already satisfied: numpy>=1.6 in c:₩users₩yj₩anaconda3₩lib₩site-packages
(from konlpy) (1.20.3)
Requirement already satisfied: lxml>=4.1.0 in c:₩users₩yj₩anaconda3₩lib₩site-package
s (from konlpy) (4.6.3)
Installing collected packages: JPype1, konlpy
Successfully installed JPype1-1.3.0 konlpy-0.6.0
Note: you may need to restart the kernel to use updated packages.
```

In [45]:

```
import konlpy
```

In [46]:

```
!pip install pytagcloud pygame simplejson
```

```
Collecting pytagcloud
  Downloading pytagcloud-0.3.5.tar.gz (754 kB)
Collecting pygame
  Downloading pygame-2.1.2-cp39-cp39-win_amd64.whl (8.4 MB)
Collecting simplejson
  Downloading simplejson-3.17.6-cp39-cp39-win_amd64.whl (75 kB)
Building wheels for collected packages: pytagcloud
  Building wheel for pytagcloud (setup.py): started
  Building wheel for pytagcloud (setup.py): finished with status 'done'
  Created wheel for pytagcloud: filename=pytagcloud-0.3.5-py3-none-any.whl size=7598
70 sha256=624f7645531bd658dedc095f1b10f0c43b3dda3ade04fd478199434e50928dce
  Stored in directory: c:₩users₩yj₩appdata₩local₩pip₩cache₩wheels₩74₩9f₩93₩6322d7ac8
b7c348b7d625f95919691d20cd46d2989dc61b165
Successfully built pytagcloud
Installing collected packages: simplejson, pytagcloud, pygame
Successfully installed pygame-2.1.2 pytagcloud-0.3.5 simplejson-3.17.6
```

In [47]:

```
import pytagcloud
```

```
pygame 2.1.2 (SDL 2.0.18, Python 3.9.7)
Hello from the pygame community. https://www.pygame.org/contribute.html (https://ww
w.pygame.org/contribute.html)
```

[각 링크 페이지내 텍스트 구조 확인 => 제목, 카테고리, 내용 출력]

In [28]:

```
driver=webdriver.Chrome(executable_path=executable_path)
#크롬 드라이버를 통해 page_urls[0]번째 사이트의 html 문서 가져옴
driver.get(page_urls[0])#page_urls[0]의 정보를 가져옴
req=driver.page_source#페이지 소스를 req에 저장
req##소스가 문제없이 불러지는지 확인.
```


```
="#008275"><meta data-n-head="ssr" name="googlebot" content="noarchive"><link data
-n-head="ssr" rel="canonical" href="https://namu.wiki/w/%EB%B0%B1%ED%98%84%EC%A4%9
1%ED%95%99%EA%B5%90"><link data-n-head="ssr" rel="search" type="application/opense
archdescription+xml" title="나무위키" href="/opensearch.xml"><link data-n-head="ss
r" rel="copyright" href="//creativecommons.org/licenses/by-nc-sa/2.0/kr/"><link da
ta-n-head="ssr" rel="icon" href="/favicon.svg" sizes="any" type="image/svg+xml"><l
ink data-n-head="ssr" rel="icon" href="/favicon-32.png" sizes="32x32" type="image/
png"><link data-n-head="ssr" rel="icon" href="/favicon-192.png" sizes="192x192" ty
pe="image/png"><link data-n-head="ssr" rel="apple-touch-icon" href="/img/apple_ico
n.png"><script data-n-head="ssr" async="" src="https://securepubads.g.doubleclick.
net/tag/js/gpt.js"></script><link rel="preload" href="/skins/senkawa/manifest.0c9a
a297b5b49b628f61.js" as="script"><link rel="preload" href="/skins/senkawa/6.0ec579
cd0a387a25b691.css" as="style"><link rel="preload" href="/skins/senkawa/vendor.0ec
579cd0a387a25b691.js" as="script"><link rel="preload" href="/skins/senkawa/3.c2f43
26b616fb16062e8.css" as="style"><link rel="preload" href="/skins/senkawa/main.c2f4
326b616fb16062e8.js" as="script"><link rel="stylesheet" href="/skins/senkawa/6.0ec
579cd0a387a25b691.css"><link rel="stylesheet" href="/skins/senkawa/3.c2f4326b616fb
16062e8.css"><script async="" src="/cdn-cgi/bm/cv/669835187/api.js"></script><styl
e type="text/css">.resize-observer[data-v-b329ee4c]{position:absolute;top:0;left:
```


In [29]:

```
soup=bts(req,'html.parser')
soup
```

Out[29]:


```
<html><head><title>백현중학교 - 나무위키</title><meta charset="utf-8" data-n-head
="ssr"/><meta content="user-scalable=no, initial-scale=1.0, maximum-scale=5.0, min
imum-scale=1.0, width=device-width" data-n-head="ssr" name="viewport"/><meta conte
nt="ie=edge" data-n-head="ssr" http-equv="x-ua-compatible"/><meta content="the see
d" data-n-head="ssr" name="generator"/><meta content="yes" data-n-head="ssr" name
="mobile-web-app-capable"/><meta content="나무위키" data-n-head="ssr" name="applic
ation-name"/><meta content="나무위키" data-n-head="ssr" name="msapplication-toolti
p"/><meta content="/w/%EB%82%98%EB%AC%B4%EC%9C%84%ED%82%A4:%EB%8C%80%EB%AC%B8" dat
a-n-head="ssr" name="msapplication-starturl"/><meta content="max-image-preview:lar
ge" data-n-head="ssr" name="robots"/><meta content="#008275" data-n-head="ssr" nam
e="theme-color"/><meta content="noarchive" data-n-head="ssr" name="googlebot"/><li
nk data-n-head="ssr" href="https://namu.wiki/w/%EB%B0%B1%ED%98%84%EC%A4%91%ED%95%9
9%EA%B5%90" rel="canonical"/><link data-n-head="ssr" href="/opensearch.xml" rel="s
earch" title="나무위키" type="application/opensearchdescription+xml"/><link data-n
-head="ssr" href="//creativecommons.org/licenses/by-nc-sa/2.0/kr/" rel="copyrigh
t"/><link data-n-head="ssr" href="/favicon.svg" rel="icon" sizes="any" type="imag
e/svg+xml"/><link data-n-head="ssr" href="/favicon-32.png" rel="icon" sizes="32x3
```

In [30]:

```
contents_article=soup.find(name='article')
contents_article
```

Out[30]:

<article data-v-47329d14="" data-v-e9de27d8=""><!-- --> <!-- --> <div class="8z71a
wy+" data-v-dc897e4a="" data-v-e9de27d8=""><div class="eFDRAnEj" data-v-dc897e4a
=""><a class="has-tooltip" data-original-title="null" data-v-dc897e4a="" href="/me
mber/star/%EB%B0%B1%ED%98%84%EC%A4%91%ED%95%99%EA%B5%90"><span class="ion-ios-star
-outline 7/cj/n5l" data-v-dc897e4a=""> 4 <!-- --> <a data-v-dc897e4a=""
href="/backlink/%EB%B0%B1%ED%98%84%EC%A4%91%ED%95%99%EA%B5%90">역링크 <a class
="" data-v-dc897e4a="" href="/discuss/%EB%B0%B1%ED%98%84%EC%A4%91%ED%95%99%EA%B5%9
0">토론 <a data-v-dc897e4a="" href="/edit/%EB%B0%B1%ED%98%84%EC%A4%91%ED%95%9
9%EA%B5%90" rel="nofollow">편집 <a data-v-dc897e4a="" href="/history/%EB%B0%B
1%ED%98%84%EC%A4%91%ED%95%99%EA%B5%90">역사 <!-- --> <a data-v-dc897e4a="" hre
f="/acl/%EB%B0%B1%ED%98%84%EC%A4%91%ED%95%99%EA%B5%90" rel="nofollow">ACL </di
v> <h1 data-v-e9de27d8=""><a data-v-e9de27d8="" href="/w/%EB%B0%B1%ED%98%84%
EC%A4%91%ED%95%99%EA%B5%90"><!-- -->백현중학교 <!-- --></h1> <p data-v-e9de27d
8="">최근 수정 시각: <time data-v-e9de27d8="" datetime="2022-01-18T05:16:46.000Z">
2022-01-18 14:16:46</time></p> <!-- --> <!-- --> <div class="9hHDDOo/" data-v-e9de
27d8="" id="fScKR8hUD"><div data-google-query-id="CM7w8IPGuvUCFRdmYAodUqQM1Q" data
-v-e9de27d8="" id="0vnKQn4YN"><div id="google_ads_iframe_/21952009853/namuwiki/top
_0__container__" style="border: 0pt none;"><iframe allow="attribution-reporting" a

In [31]:

```
##타이틀 추출
title=contents_article.find_all('h1')[0]#태그명이 h1인 모든 태그 추출, => select('article h1')
title
```

Out[31]:

<h1 data-v-e9de27d8=""><a data-v-e9de27d8="" href="/w/%EB%B0%B1%ED%98%84%EC%A4%91%E
D%95%99%EA%B5%90"><!-- -->백현중학교 <!-- --></h1>

In [32]:

```
category=contents_article.find_all('ul')[0]
category
```

Out[32]:

<ul data-v-9c687c72=""><li data-v-9c687c72=""><a class="" data-v-9c687c72="" href="/
w/%EB%B6%84%EB%A5%98:%EC%84%B1%EB%82%A8%EC%8B%9C%EC%9D%98%20%EC%A4%91%ED%95%99%EA%B
5%90">성남시의 중학교
<li data-v-9c687c72=""><a class="" data-v-9c687c72=""
href="/w/%EB%B6%84%EB%A5%98:1995%EB%85%84%20%EA%B0%9C%EA%B5%90">1995년 개교
<li data-v-9c687c72=""><a class="" data-v-9c687c72="" href="/w/%EB%B6%84%EB%A5%98:%E
B%82%98%EB%AC%B4%EC%9C%84%ED%82%A4%20%EA%B5%90%EC%9C%A1%EA%B8%B0%EA%B4%80%20%ED%94%8
4%EB%A1%9C%EC%A0%9D%ED%8A%B8">나무위키 교육기관 프로젝트

In [33]:

```
#내용추출
paragraphs=contents_article.select('div.wiki-paragraph')#class가 wiki-paragraph인 div를 추출
#내용으로 추출한 리스트를 하나의 문자열로 전처리
contents_corpus_list=[]

for paragraph in paragraphs:
    contents_corpus_list.append(paragraph.text)

contents_corpus=" ".join(contents_corpus_list) #"텍스트".join(리스트명)=>리스트의 요소를 '텍스트'로
contents_corpus
```

```
대한 내용은  문서를의 번 문단을의 번 문단을의  부분을의  부분을, 에 대한 내용은
문서를의 번 문단을의 번 문단을의  부분을의  부분을, 에 대한 내용은  문서를의 번 문
단을의 번 문단을의  부분을의  부분을, 에 대한 내용은  문서를의 번 문단을의 번 문단
을의  부분을의  부분을, 에 대한 내용은  문서를의 번 문단을의 번 문단을의  부분을의
부분을, 에 대한 내용은  문서를의 번 문단을의 번 문단을의  부분을의  부분을 참고하
십시오.  ₩xa0 가입 후 15일이 지나야 편집 가능한 문서입니다. 이 문서는 학교 관련 문
서이며, 로그인 후 수정 가능합니다. 다음과 같은 서술은 작성 시 제재될 수 있습니다.
[ 펼치기 · 접기 ]현황과 역사를 불문하고 학교 간 우열이나 서열을 확정·조장하는 서
술과 학교 수준을 비하하는 서술학교 내 특정 임의 단체(급식, 매점 등)와 교내 학생 단
체(동아리, 학생회 등)에 관한 서술언론에 보도되지 않은 학교 관련 사건 사고에 관한
서술학교 밖에서 알려지지 않은 교사나 학생에 관한 서술템플릿:학교에 존재하지 않는
문단에 관한 서술편집지침에 어긋나는 서술 현황과 역사를 불문하고 학교 간 우열이나
서열을 확정·조장하는 서술과 학교 수준을 비하하는 서술 학교 내 특정 임의 단체(급
식, 매점 등)와 교내 학생 단체(동아리, 학생회 등)에 관한 서술 언론에 보도되지 않은
학교 관련 사건 사고에 관한 서술 학교 밖에서 알려지지 않은 교사나 학생에 관한 서술
템플릿:학교에 존재하지 않는 문단에 관한 서술 편집지침에 어긋나는 서술  경기도 성남
시 분당구의 중학교  [ 펼치기 · 접기 ]구미중학교낙원중학교내정중학교늘푸른중학교매
송중학교백현중학교보평중학교분당중학교불곡중학교삼평중학교샛별중학교서현중학교송림
중학교수내중학교신백현중학교야탑중학교양영중학교운중중학교이매중학교이우중학교장안
중학교정자중학교청솔중학교판교대장중학교판교중학교하탑중학교흐르샘 바탕은 남자중학
```

In [34]:

```
print(title.text.strip())
print('₩n')
print(category.text)
print('₩n')
print(contents_corpus)
#크롤링에 사용한 브라우저를 종료합니다.
driver.close()
```

```
매점 등)와 교내 학생 단체(동아리, 학생회 등)에 관한 서술 언론에 보도되지 않은 학교
관련 사건 사고에 관한 서술 학교 밖에서 알려지지 않은 교사나 학생에 관한 서술 템플
릿:학교에 존재하지 않는 문단에 관한 서술 편집지침에 어긋나는 서술  경기도 성남시
분당구의 중학교  [ 펼치기 · 접기 ]구미중학교낙원중학교내정중학교늘푸른중학교매송
중학교백현중학교보평중학교분당중학교불곡중학교삼평중학교샛별중학교서현중학교송림중
학교수내중학교신백현중학교야탑중학교양영중학교운중중학교이매중학교이우중학교장안중
학교정자중학교청솔중학교판교대장중학교판교중학교하탑중학교푸른색 바탕은 남자중학
교, 붉은색 바탕은 여자중학교. 흰색 바탕은 남녀공학. 구미중학교 낙원중학교 내정중학
교 늘푸른중학교 매송중학교 백현중학교 보평중학교 분당중학교 불곡중학교 삼평중학교
샛별중학교 서현중학교 송림중학교 수내중학교 신백현중학교 야탑중학교 양영중학교 운
중중학교 이매중학교 이우중학교 장안중학교 정자중학교 청솔중학교 판교대장중학교 판
교중학교 하탑중학교 푸른색 바탕은 남자중학교, 붉은색 바탕은 여자중학교. 흰색 바탕
은 남녀공학. 초등학교 · 중학교 · 고등학교 · 각종학교 학교 틀 둘러보기  백현중학
교  Bakhyun Middle School  栢峴伯賢中學校  학교로고  학교전경  학교위치  개교  1
995년 3월 2일 유형  일반계 중학교 성별  남녀공학 운영형태  공립 교장  제 12대 "홍
향표" 선생님 교감  제 11대 "이옥경" 선생님 교무부장  염유선 선생님 교훈  밝고 맑고
바르게 상징  교목: 잣나무, 교화: 철쭉, 교조: 비둘기 학생 수  810명2021.12.09 교직
원 수  58명2021.12.09 관할교육청  경기도성남교육지원청 소재지  경기도 성남시 분당
구 황새울로 124 (정자동) 홈페이지  1. 개요2. 역사2.1. 학교 연혁3. 교훈 및 상징3.
1. 교육목표 및 백현교육의 약속3.2. 학교 상징3.2.1. 교표3.2.2. 교복3.2.3. 교가3.2.
```

네이버 크롤링

In [14]:

```
#윈도우용 크롬 웹 드라이버 실행 경로(window)지정
executable_path="chromedriver.exe"
driver=webdriver.Chrome(executable_path=executable_path)

#사이트의 html구조에 기반하여 크롤링을 수행
source_url="https://kin.naver.com/qna/kinupList.naver"
driver.get(source_url)

#element=WebDriverWait(driver,5).until(EC.presence_of_element_located((By.CLASS_NAME,"app")))
req=driver.page_source
req
```

```
x;margin:0;padding:0;font-size:0;line-height:0}Wn#gnb .gnb_my_membership{padding:
0;display: block;width: 64px;height: 16px;background-position: -296px -359px;margi
n: 0;}Wn#gnb .gnb_my_interface{padding:5px;position:absolute;top:12px;right:8px;di
splay:block;width:17px;height:16px;background-position:-90px 5px}Wn#gnb .gnb_my_in
terface:hover{background-position:-90px -20px}Wn#gnb .gnb_my_interface:focus{backg
round-position:-90px -20px}Wn#gnb .gnb_pad_lyr{position:absolute}Wn#gnb .gnb_ico_n
um{display:block;position:absolute;top:1px;width:40px;text-align:center}Wn#gnb .gn
b_ico_num .gnb_ico_new{height:15px;display:inline-block;background-position:-331px
0;zoom:1}Wn#gnb .gnb_ico_num .gnb_ico_new .gnb_count{position:relative;top:0;righ
t:-5px;height:15px;margin:0;padding:0 4px 0 1px;display:inline-block;*display:inli
ne;vertical-align:top;background-position:100% 0;text-indent:-2px;font-family:taho
ma !important;font-weight:bold;color:#fff;zoom:1}Wn#gnb .gnb_ico_num .gnb_ico_new
.plus{margin:1px -1px 0 2px;font-size:8px;display:inline-block;color:#fff;vertical
-align:top}Wn:root #gnb .gnb_pad_lyr{opacity:1 !important;/* background:#fff */}
Wn.gnb_lst{margin:0;padding:0;zoom:1}Wn.gnb_lst:after{display:block;clear:both;con
tent:'W'}Wn.gnb_lst ul{margin:0;padding:0}Wn.gnb_lst .ico_arrow{display:none;posi
tion:absolute;left:50%;top:27px;width:10px;height:8px;margin-left:-5px;background-
position:-175px -10px}Wn.gnb_lyr_opened .gnb_my_lyr, .gnb_lyr_opened .gnb_service_
lyr, .gnb_lyr_opened .gnb_notice_lyr, .gnb_lyr_opened .ico_arrow{display:block !im
```

In [15]:

```
soup=bts(req,'html.parser')
soup
```

```
        blsActionNoticeDisplay : ("false" == "" || "false" == "false") ? false : t
rue

};




var standardReportPopupUrl = "https://srp2.naver.com/report";
</script>
<style id="gnb_style" type="text/css">@charset "UTF-8";
/* NTS UIT Development Office YJH 140717 */

a.gnb_my, .gnb_icon, #gnb .gnb_my_interface, .gnb_my_li .gnb_my_content .gnb_membe
rship, #gnb .gnb_my_membership, #gnb .gnb_ico_num .gnb_ico_new, #gnb .gnb_ico_num
.gnb_ico_new .gnb_count, .gnb_lst .ico_arrow, a.gnb_my .filter_mask, .gnb_my_lyr,
.gnb_my_li .gnb_my_content .gnb_mask, .gnb_my_li .gnb_my_content .gnb_change, .gnb
_my_li .gnb_my_content .gnb_edit_lst li, .gnb_my_li .gnb_my_content .gnb_pay_check
em, #gnb .gnb_my_li .gnb_my_community a.gnb_pay span, .gnb_notice_li .gnb_notice_l
```

In [16]:

```
contents_board=soup.select('td.title a')
contents_board
```

Out[16]:

```
[<a class="_nclicks:kls_bst.list,r:1,i:100412_409937830" href="/qna/detail.naver?d
1id=10&amp;dirId=100412&amp;docId=409937830" rel="KIN">유치원 신원 진술서 판매처,
작성 방법</a>,
 <a class="_nclicks:kls_bst.list,r:2,i:40102_409937747" href="/qna/detail.naver?d1
id=4&amp;dirId=40102&amp;docId=409937747" rel="KIN">확정기여형DC 퇴직연금 주식 되
나요?</a>,
 <a class="_nclicks:kls_bst.list,r:3,i:12091501_409937692" href="/qna/detail.nave
r?d1id=12&amp;dirId=12091501&amp;docId=409937692" rel="KIN">허리디스크 한의원 치료
는 어떨까요? 분당 근처로 알려주...</a>,
 <a class="_nclicks:kls_bst.list,r:4,i:7011401_409937656" href="/qna/detail.naver?
d1id=7&amp;dirId=7011401&amp;docId=409937656" rel="KIN">임신 여부</a>,
 <a class="_nclicks:kls_bst.list,r:5,i:20132_409937653" href="/qna/detail.naver?d1
id=2&amp;dirId=20132&amp;docId=409937653" rel="KIN">롤 미드 라인</a>,
 <a class="_nclicks:kls_bst.list,r:6,i:5010601_409937600" href="/qna/detail.naver?
d1id=5&amp;dirId=5010601&amp;docId=409937600" rel="KIN">저 진짜 급한데 도와주세요
아까 갑자</a>,
 <a class="_nclicks:kls_bst.list,r:7,i:70701_409937590" href="/qna/detail.naver?d1
id=7&amp;dirId=70701&amp;docId=409937590" rel="KIN">제주도한달살기비용 숙소에 어느
```

In [35]:

```
page_url_base = "https://kin.naver.com" #베이스 url 정의
page_urls = [] # href 속성값을 담기 위한 빈 리스트

for content in contents_board:
    if "png" not in page_url:
            page_urls.append(page_url)
    page_urls.append(page_url_base+content['href'])

for page in page_urls:
    print(page)
```

```
https://namu.wiki/w/%EB%AC%B8%20%EB%82%98%EC%9D%B4%ED%8A%B8(%EB%93%9C%EB%9
A7%88) (https://namu.wiki/w/%EB%AC%B8%20%EB%82%98%EC%9D%B4%ED%8A%B8(%EB%93%9C%EB%9

D%BC%EB%A7%88))
https://kin.naver.com/qna/detail.naver?d1id=7&dirId=7011401&docId=409937656 (http
s://kin.naver.com/qna/detail.naver?d1id=7&dirId=7011401&docId=409937656)
https://namu.wiki/w/%EB%AC%B8%20%EB%82%98%EC%9D%B4%ED%8A%B8(%EB%93%9C%EB%9D%BC%EB%
A7%88) (https://namu.wiki/w/%EB%AC%B8%20%EB%82%98%EC%9D%B4%ED%8A%B8(%EB%93%9C%EB%9
D%BC%EB%A7%88))
https://kin.naver.com/qna/detail.naver?d1id=2&dirId=20132&docId=409937653 (http
s://kin.naver.com/qna/detail.naver?d1id=2&dirId=20132&docId=409937653)
https://namu.wiki/w/%EB%AC%B8%20%EB%82%98%EC%9D%B4%ED%8A%B8(%EB%93%9C%EB%9D%BC%EB%
A7%88) (https://namu.wiki/w/%EB%AC%B8%20%EB%82%98%EC%9D%B4%ED%8A%B8(%EB%93%9C%EB%9
D%BC%EB%A7%88))
https://kin.naver.com/qna/detail.naver?d1id=5&dirId=5010601&docId=409937600 (http
s://kin.naver.com/qna/detail.naver?d1id=5&dirId=5010601&docId=409937600)
https://namu.wiki/w/%EB%AC%B8%20%EB%82%98%EC%9D%B4%ED%8A%B8(%EB%93%9C%EB%9D%BC%EB%
A7%88) (https://namu.wiki/w/%EB%AC%B8%20%EB%82%98%EC%9D%B4%ED%8A%B8(%EB%93%9C%EB%9
D%BC%EB%A7%88))
https://kin.naver.com/qna/detail.naver?d1id=7&dirId=70701&docId=409937590 (http
s://kin.naver.com/qna/detail.naver?d1id=7&dirId=70701&docId=409937590)
```

In [18]:

```
driver=webdriver.Chrome(executable_path=executable_path)
#크롬 드라이버를 통해 page_urls[0]번째 사이트의 html 문서 가져옴
driver.get(page_urls[0])#page_urls[0]의 정보를 가져옴
req=driver.page_source#페이지 소스를 req에 저장
req##소스가 문제없이 불러지는지 확인.
```

```
Lwgj9ePt66XMEvNkVW0E0WPd7TP9sBQ25X0Q15Lr1Nn4oGFQkAAACceyJvcmlnaW4iOiJodHRwczovL2dv
b2dsZXN5bmRpY2F0aW9uLmNvbTo0NDMiLCJmZWF0dXJlIjoiQ29udmVyc2lvbk1lYXN1cmVtZW50Iiwi
hwaXJ5IjoxNjQzMTU1MTk5LCJpc1N1YmRvbWFpbiI6dHJ1ZSwiaXNUaGlyZFBhcnR5Ijp0cnVlLCJ1c2Fn
ZSI6InN1YnNldCJ9"><meta http-equiv="origin-trial" content="A4/Htern2udN9w3yJK9QgWQ
xQFrux0XsXL7cW60DyCI0EZFGCSme/J33Q/WzF7bBkVvhEWDIcBiUyZaim5CpFQwAAACceyJvcmlnaW4i0
iJodHRwczovL2dvb2dsZXRhZ3NlcnZpY2VzLmNvbTo0NDMiLCJmZWF0dXJlIjoiQ29udmVyc2lvbk1lYXN
1cmVtZW50Iiwi ZXhwaXJ5IjoxNjQzMTU1MTk5LCJpc1N1YmRvbWFpbiI6dHJ1ZSwiaXNUaGlyZFBhcnR5I
jp0cnVlLCJ1c2FnZSI6InN1YnNldCJ9"><script src="https://securepubads.g.doubleclick.n
et/gpt/pubads_impl_2022011002.js" async=""></script><meta http-equiv="origin-tria
l" content="A0Bg2nddUj4Nw6FzsXudBXHZs1aAzlg0+UGzfJGkC1f4J56ghvJ6TCirjdt8BUwsK14sBB
jWGm0Y+QCTr2HrBQoAAACBeyJvcmlnaW4iOiJodHRwczovL3NlY3VyZXB1YmFkcy5nLmRvdWJsZWNsaWNr
Lm5ldDo0NDMiLCJmZWF0dXJlIjoiU3VicmVzb3VyY2VXZWJCdW5kbGVzIiwiZXhwaXJ5IjoxNjUy0DMx0T
k5LCJpc1RoaXJkUGFydHki0nRydWV9"><link rel="preload" href="https://adservice.googl
e.co.kr/adsid/integrator.js?domain=namu.wiki" as="script"><script type="text/javas
cript" src="https://adservice.google.co.kr/adsid/integrator.js?domain=namu.wiki">
</script><link rel="preload" href="https://adservice.google.com/adsid/integrator.j
s?domain=namu.wiki" as="script"><script type="text/javascript" src="https://adserv
ice.google.com/adsid/integrator.js?domain=namu.wiki"></script><script src="urn:uui
d:204c26d8-0ef1-48f5-8763-9e83ba6b9a89"></script></head><body><div id="app"><!----
><div id="BiHgQzR8G" class="app senkawa-fixed-size senkawa-fixed-1300" data-v-4732
```

In [19]:

```
soup=bts(req,'html.parser')
soup
```

Out[19]:

```
<html><head><title>마법사의 밤 - 나무위키</title><meta charset="utf-8" data-n-head
="ssr"/><meta content="user-scalable=no, initial-scale=1.0, maximum-scale=5.0, min
imum-scale=1.0, width=device-width" data-n-head="ssr" name="viewport"/><meta conte
nt="ie=edge" data-n-head="ssr" http-equv="x-ua-compatible"/><meta content="the see
d" data-n-head="ssr" name="generator"/><meta content="yes" data-n-head="ssr" name
="mobile-web-app-capable"/><meta content="나무위키" data-n-head="ssr" name="applic
ation-name"/><meta content="나무위키" data-n-head="ssr" name="msapplication-toolti
p"/><meta content="/w/%EB%82%98%EB%AC%B4%EC%9C%84%ED%82%A4:%EB%8C%80%EB%AC%B8" dat
a-n-head="ssr" name="msapplication-starturl"/><meta content="max-image-preview:lar
ge" data-n-head="ssr" name="robots"/><meta content="#008275" data-n-head="ssr" nam
e="theme-color"/><meta content="noarchive" data-n-head="ssr" name="googlebot"/><li
nk data-n-head="ssr" href="https://namu.wiki/w/%EB%A7%88%EB%B2%95%EC%82%AC%EC%9D%9
8%20%EB%B0%A4" rel="canonical"/><link data-n-head="ssr" href="/opensearch.xml" rel
="search" title="나무위키" type="application/opensearchdescription+xml"/><link dat
a-n-head="ssr" href="//creativecommons.org/licenses/by-nc-sa/2.0/kr/" rel="copyrig
ht"/><link data-n-head="ssr" href="/favicon.svg" rel="icon" sizes="any" type="imag
e/svg+xml"/><link data-n-head="ssr" href="/favicon-32.png" rel="icon" sizes="32x3
2" type="image/png"/><link data-n-head="ssr" href="/favicon-192.png" rel="icon" si
```

In [20]:

```
title=soup.find(name='title')
title.text.strip()
```

Out[20]:

```
'마법사의 밤 - 나무위키'
```

In [21]:

```
for i in range(0, len(table_rows)): #table_rows의 길이만큼 반복
    first_td = table_rows[i].find_all('td')[0] #td가 3개 있는데 0번째에 원하는 href 가 있음
    #print(first_td)
    td_url = first_td.find_all('a')
    #print(td_url)
    if len(td_url) > 0:
        page_url = page_url_base + td_url[0].get('href') #나무위키주소+get() 태그가 가지고 있는 속성
        #print(page_url)
        if "png" not in page_url:
            page_urls.append(page_url)
```

네이버크롤링

In [22]:

```
#윈도우용 크롬 웹 드라이버 실행 경로(window)지정
executable_path="chromedriver.exe"
driver=webdriver.Chrome(executable_path=executable_path)

#사이트의 html구조에 기반하여 크롤링을 수행
source_url="https://kin.naver.com/qna/kinupList.naver"
driver.get(source_url)

#element=WebDriverWait(driver,5).until(EC.presence_of_element_located((By.CLASS_NAME,"app")))
req=driver.page_source

soup=bts(req,'html.parser')

contents_board=soup.select('td.title a')

page_url_base = "https://kin.naver.com" #베이스 url 정의
page_urls = [] # href 속성값을 담기 위한 빈 리스트

for content in contents_board:
    page_url=page_url_base+content['href']
    if "png" not in page_url:
            page_urls.append(page_url)

for page in page_urls:
    print(page)

driver=webdriver.Chrome(executable_path=executable_path)
#크롬 드라이버를 통해 page_urls[0]번째 사이트의 html 문서 가져옴
for page_url in page_urls:
    driver.get(page_url)
    req=driver.page_source
    soup=bts(req,'html.parser')
    qtime=soup.select('.c-userinfo__info')#질문 작성일
    print(f'{qtime[0].text} : {qtime[1].text}')
    print('₩n')
    #print(qtime[2])
    title=soup.find(name='title')#질문 제목
    print('제목 : '+title.text.strip())
    print('₩n')
    question=soup.select('div.c-heading__content')#질문내용
    print('질문 : '+question[0].text.strip())#이미지가 있거나 하면 오류 발생
    print('₩n')
    answer=soup.find('div',{'class':'_endContents c-heading-answer__content'})#답변 div
    an=[]#답변 div의 텍스트만 저장할 빈 리스트
    for ans in answer:
        an.append(ans.text.strip())#답변div에서 text만 추출하고 공백제거
    atime=soup.select('.c-heading-answer__content-date')#답변 작성일
    print(f'작성일 : {atime[0].text}')
    print('₩n')
    print('답변 : '+an[1].replace('    ',''))#an리스트의1번째에 답변 내용이 저장됨 그후 공백을 replac
    print('₩n')
    print('======================')
```

```
https://kin.naver.com/qna/detail.naver?d1id=10&dirId=100412&docId=409937830 (http
s://kin.naver.com/qna/detail.naver?d1id=10&dirId=100412&docId=409937830)
https://kin.naver.com/qna/detail.naver?d1id=8&dirId=8040106&docId=409937753 (http
s://kin.naver.com/qna/detail.naver?d1id=8&dirId=8040106&docId=409937753)
https://kin.naver.com/qna/detail.naver?d1id=4&dirId=40102&docId=409937747 (http
s://kin.naver.com/qna/detail.naver?d1id=4&dirId=40102&docId=409937747)
```

```
https://kin.naver.com/qna/detail.naver?d1id=12&dirId=12091501&docId=409937692 (htt
ps://kin.naver.com/qna/detail.naver?d1id=12&dirId=12091501&docId=409937692)
https://kin.naver.com/qna/detail.naver?d1id=7&dirId=7011401&docId=409937656 (http
s://kin.naver.com/qna/detail.naver?d1id=7&dirId=7011401&docId=409937656)
https://kin.naver.com/qna/detail.naver?d1id=2&dirId=20132&docId=409937653 (http
s://kin.naver.com/qna/detail.naver?d1id=2&dirId=20132&docId=409937653)
https://kin.naver.com/qna/detail.naver?d1id=5&dirId=5010601&docId=409937600 (http
s://kin.naver.com/qna/detail.naver?d1id=5&dirId=5010601&docId=409937600)
https://kin.naver.com/qna/detail.naver?d1id=7&dirId=70701&docId=409937590 (http
s://kin.naver.com/qna/detail.naver?d1id=7&dirId=70701&docId=409937590)
https://kin.naver.com/qna/detail.naver?d1id=8&dirId=81302&docId=409937587 (http
s://kin.naver.com/qna/detail.naver?d1id=8&dirId=81302&docId=409937587)
https://kin.naver.com/qna/detail.naver?d1id=6&dirId=61303&docId=409937565 (http
```

위키백과 크롤링

In [23]:

```
#윈도우용 크롬 웹 드라이버 실행 경로(window)지정
executable_path="chromedriver.exe"
driver=webdriver.Chrome(executable_path=executable_path)

#사이트의 html구조에 기반하여 크롤링을 수행
source_url="https://ko.wikipedia.org/wiki/%ED%8A%B9%EC%88%98:%EC%B5%9C%EA%B7%BC%EB%B0%94%EB%80%9C?hi
driver.get(source_url)

#element=WebDriverWait(driver,5).until(EC.presence_of_element_located((By.CLASS_NAME,"app")))
req=driver.page_source

soup=bts(req,'html.parser')

atags=soup.select('.mw-title a')#제목과 url주소가 모두 들어있는 a태그 추출

base_url='https://ko.wikipedia.org'#앞에 기본으로 붙는 wikipedia 주소

page_urls=[]
for a in atags:
    print(a.text)
    page_urls.append(base_url+a['href'])
    print(base_url+a['href'])
    print('============================')
```

```
동우에이앤이
https://ko.wikipedia.org/wiki/%EB%8F%99%EC%9A%B0%EC%97%90%EC%9D%B4%EC%95%A4%EC%9D%
B4 (https://ko.wikipedia.org/wiki/%EB%8F%99%EC%9A%B0%EC%97%90%EC%9D%B4%EC%95%A4%E
C%9D%B4)
============================
권경훈
https://ko.wikipedia.org/wiki/%EA%B6%8C%EA%B2%BD%ED%9B%88 (https://ko.wikipedia.or
g/wiki/%EA%B6%8C%EA%B2%BD%ED%9B%88)
============================
김채하
https://ko.wikipedia.org/wiki/%EA%B9%80%EC%B1%84%ED%95%98 (https://ko.wikipedia.or
g/wiki/%EA%B9%80%EC%B1%84%ED%95%98)
============================
사용자:베리슈아
https://ko.wikipedia.org/wiki/%EC%82%AC%EC%9A%A9%EC%9E%90:%EB%B2%A0%EB%A6%AC%EC%8
A%88%EC%95%84 (https://ko.wikipedia.org/wiki/%EC%82%AC%EC%9A%A9%EC%9E%90:%EB%B2%A
0%EB%A6%AC%EC%8A%88%EC%95%84)
============================
전해리
```

In [ ]: