

!pip install selenium

In [10]:

```
%matplotlib inline
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings("ignore")

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

In [11]:

```
executable_path = "chromedriver.exe"
driver = webdriver.Chrome(executable_path = executable_path)

source_url = "https://ko.wikipedia.org/wiki/%ED%8A%B9%EC%88%98:%EC%B5%9C"
driver.get(source_url) #크롬 드라이버를 통해 url의 html 문서 가져옴

req = driver.page_source #전체페이지
print(req)

soup = BeautifulSoup(req, "html.parser") #BeautifulSoup 의 soup로 가공
soup
```

```
<html class="client-js" lang="ko" dir="ltr"><head>
<meta charset="UTF-8">
<title>최근 바뀜 - 위키백과, 우리 모두의 백과사전</title>
<script>document.documentElement.className="client-js";RL
CONF={"wgBreakFrames":true,"wgSeparatorTransformTable":
["",""],"wgDigitTransformTable":["",""],"wgDefaultDateForma
t":"ko","wgMonthNames":["","1월","2월","3월","4월","5월","6
월","7월","8월","9월","10월","11월","12월"],"wgRequestId":"73
9bf4f1-85eb-468a-86a0-437e377030ef","wgCSPNonce":false,"w
gCanonicalNamespace":"Special","wgCanonicalSpecialPageNa
me":"Recentchanges","wgNamespaceNumber":-1,"wgPageNa
me":"특수:최근바뀜","wgTitle":"최근바뀜","wgCurRevisionId":
0,"wgRevisionId":0,"wgArticleId":0,"wgIsArticle":false,"wgIsRe
direct":false,"wgAction":"view","wgUserName":null,"wgUserG
roups":["*"],"wgCategories":[],"wgPageContentLanguage":"k
o","wgPageContentModel":"wikitext","wgRelevantPageNam
e":"특수:최근바뀜","wgRelevantArticleId":0,"wgIsProbablyEdita
ble":false,"wgRelevantPagelsProbablyEditable":false,"wgVisua
lEditor":{"pageLanguageCode":"ko","pageLanguageDir":"lt
r","pageVariantFallbacks":"ko"},
"wgMFDisplayWikibaseDescriptions":{"search":true,"nearby":t
...}
```

In [12]:

```
contents_table=soup.select('.mw-title a')
```

```
contents_table
```

Out[12]:

```
[<a class="mw-changeslist-title" href="/wiki/%EB%9D%BC%E
C%9A%B0%ED%84%B0%EB%B8%8C%EB%A3%A8%EB%84%A
8" title="라우터브루넨">라우터브루넨</a>,
 <a class="mw-changeslist-title" href="/wiki/%EC%82%AC%E
C%9A%A9%EC%9E%90%ED%86%A0%EB%A1%A0:%EC%95%8
8%EB%85%95%ED%95%98%EC%84%B8%EC%9A%94_%E3%
85%8B%E3%85%8B" title="사용자토론:안녕하세요 ㅋㅋ">사용자토
론:안녕하세요 ㅋㅋ</a>,
 <a class="mw-changeslist-title" href="/wiki/%EA%B6%8C%E
A%B7%A0" title="권균">권균</a>,
 <a class="mw-changeslist-title" href="/wiki/%EB%9D%BC%E
C%9D%B4%ED%8A%B8%EC%9B%A8%EC%9D%B4%EB%B8%
8C_3D%EB%A1%9C_%EB%A7%8C%EB%93%A0_%EC%9E%9
1%ED%92%88_%EB%AA%A9%EB%A1%9D" title="라이트웨이브
3D로 만든 작품 목록">라이트웨이브 3D로 만든 작품 목록</a>,
 <a class="mw-changeslist-title" href="/wiki/%EA%B6%8C%E
A%B7%A0" title="권균">권균</a>,
 <a class="mw-changeslist-title" href="/wiki/%EB%9D%BC%E
C%9A%B0%ED%84%B0%EB%B8%8C%EB%A3%A8%EB%84%A
8" title="라우터브루넨">라우터브루넨</a>,
 <a class="mw-changeslist-title" href="/wiki/%EA%B0%95%E
A%B7%80%EC%86%90" title="강귀손">강귀손</a>,
 <a class="mw-changeslist-title" href="/wiki/3ds_%EB%A7%A
5%EC%8A%A4%EB%A1%9C_%EB%A7%8C%EB%93%A0_%E
C%9E%91%ED%92%88_%EB%AA%A9%EB%A1%9D" title="3ds
맥스로 만든 작품 목록">3ds 맥스로 만든 작품 목록</a>,
 <a class="mw-changeslist-title" href="/wiki/%EA%B0%95%E
A%B7%80%EC%86%90" title="강귀손">강귀손</a>,
 <a class="mw-changeslist-title" href="/wiki/%EB%8D%B0%E
B%93%9C_%EC%98%A4%EC%96%B4_%EC%96%BC%EB%9
D%BC%EC%9D%B4%EB%B8%8C_6" title="데드 오어 얼라이브
6">데드 오어 얼라이브 6</a>,
 <a class="mw-changeslist-title" href="/wiki/%EC%A0%95%E
B%AC%B8%ED%98%95" title="정문형">정문형</a>,
 <a class="mw-changeslist-title" href="/wiki/%EC%82%AC%E
C%9A%A9%EC%9E%90:%EC%95%88%EB%85%95%ED%95%9
8%EC%84%B8%EC%9A%94_%E3%85%8B%E3%85%8B" title
="사용자:안녕하세요 ㅋㅋ">사용자:안녕하세요 ㅋㅋ</a>,
 <a class="mw-changeslist-title" href="/wiki/%EB%A7%88%E
C%95%BC%EB%A1%9C_%EB%A7%8C%EB%93%A0_%EC%9
E%91%ED%92%88_%EB%AA%A9%EB%A1%9D" title="마야로
만든 작품 목록">마야로 만든 작품 목록</a>,
 <a class="mw-changeslist-title" href="/wiki/%EB%9D%BC%E
C%9A%B0%ED%84%B0%EB%B8%8C%EB%A3%A8%EB%84%A
8" title="라우터브루넨">라우터브루넨</a>],
```

[<정문형,](/wiki/%EC%A0%95%EB%AC%B8%ED%98%95 "정문형")
[<DC 유니버스 온라인,](/wiki/DC_%EC%9C%A0%EB%8B%88%EB%B2%84%EC%8A%A4_%EC%98%A8%EB%9D%BC%EC%9D%B8 "DC 유니버스 온라인")
[<김치삼,](/wiki/%EA%B9%80%EC%B9%98%EC%82%BC "김치삼")
[<한샤오펑,](/wiki/%ED%95%9C%EC%83%A4%EC%98%A4%ED%8E%91 "한샤오펑")
[>1월 19일,](/wiki/1%EC%9B%94_19%EC%9D%BC "1월 19일")
[>윤호 \(평정공\),](/wiki/%EC%9C%A4%ED%98%B8_(%ED%8F%89%EC%A0%95%EA%B3%B5) "윤호 (평정공)")
[>다윈 프로젝트,](/wiki/%EB%8B%A4%EC%9C%88_%ED%94%84%EB%A1%9C%EC%A0%9D%ED%8A%B8 "다윈 프로젝트")
[>위키백과:사랑방/2022년 제3주,](/wiki/%EC%9C%84%ED%82%A4%EB%B0%B1%EA%B3%BC:%EC%82%AC%EB%9E%91%EB%B0%A9/2022%EB%85%84_%EC%A0%9C3%EC%A3%BC "위키백과:사랑방/2022년 제3주")
[>위키백과:사용자 관리 요청/2022년 제3주,](/wiki/%EC%9C%84%ED%82%A4%EB%B0%B1%EA%B3%BC:%EC%82%AC%EC%9A%A9%EC%9E%90_%EA%B4%80%EB%A6%AC_%EC%9A%94%EC%B2%AD/2022%EB%85%84_%EC%A0%9C3%EC%A3%BC "위키백과:사용자 관리 요청/2022년 제3주")
[>허종 \(1434년\),](/wiki/%ED%97%88%EC%A2%85_(1434%EB%85%84) "허종 (1434년)")
[>김치삼,](/wiki/%EA%B9%80%EC%B9%98%EC%82%BC "김치삼")
[>D4DJ,](/wiki/D4DJ "D4DJ")
[>사용자:안녕하세요 ㅋㅋ,](/wiki/%EC%82%AC%EC%9A%A9%EC%9E%90:%EC%95%88%EB%85%95%ED%95%98%EC%84%B8%EC%9A%94_%E3%85%8B%E3%85%8B "사용자:안녕하세요 ㅋㅋ")
[>윤사흔,](/wiki/%EC%9C%A4%EC%82%AC%ED%9D%94 "윤사흔")
[>성봉조,](/wiki/%EC%84%B1%EB%B4%89%EC%A1%B0 "성봉조")
[>11월 25일,](/wiki/11%EC%9B%94_25%EC%9D%BC "11월 25일")
[>사용자:안녕하세요 ㅋㅋ,](/wiki/%EC%82%AC%EC%9A%A9%EC%9E%90:%EC%95%88%EB%85%95%ED%95%98%EC%84%B8%EC%9A%94_%E3%85%8B%E3%85%8B "사용자:안녕하세요 ㅋㅋ")
[>](/wiki/%EC%B9%B4%EC%9A%B4%ED%84%B0-%EC%8A%A4%ED%8A%B8%EB%9D%BC%EC%9D%B4%ED%81%AC:%_EA%B8%80%EB%A1%9C%E)

B%B2%8C_%EC%98%A4%ED%8E%9C%EC%8B%9C%EB%B8%8C" title="카운터-스트라이크: 글로벌 오펜시브">카운터-스트라이크: 글로벌 오펜시브,

한백륜,

아침마당,

Next Level (aespa의 노래),

코즈믹브레이크,

사용자토론:Ha98574,

윤사분,

안나 라자레바,

강순,

커맨드 앤 컨커 (취소된 비디오 게임),

강순,

아침마당의 에피소드 목록 (2022년),

이인손,

컴뱃암즈,

라우터브루넨,

이인손,

Savage (aespa의 E P),

정분,

<a class="mw-changeslist-title" href="/wiki/%EC%BD%9C_%E C%98%A4%EB%B8%8C_%EB%93%80%ED%8B%B0:_%EC%9

B%8C%EC%A1%B4" title="콜 오브 듀티: 워존">콜 오브 듀티: 워존]

In [13]:

```
urlList=[] #url 을 담을 빈 리스트
titleList=[] #title 담을 빈 리스트
for i in contents_table:
    page_url_base = "https://ko.wikipedia.org" #상대 주소에 덧붙이기 위한 기본 주소
    titles=i.text #제목
    titleList.append(titles)
    print(i.text)
    url =page_url_base + i["href"] # 리스트[i]의 속성값
    urlList.append(url)
    print(url)
```

라우터브루넨

<https://ko.wikipedia.org/wiki/%EB%9D%BC%EC%9A%B0%ED%84%B0%EB%B8%8C%EB%A3%A8%EB%84%A8> (<https://ko.wikipedia.org/wiki/%EB%9D%BC%EC%9A%B0%ED%84%B0%EB%B8%8C%EB%A3%A8%EB%84%A8>)

사용자토론:안녕하세요 ㅋㅋ

https://ko.wikipedia.org/wiki/%EC%82%AC%EC%9A%A9%EC%9E%90%ED%86%A0%EB%A1%A0:%EC%95%88%EB%85%95%ED%95%98%EC%84%B8%EC%9A%94_%E3%85%8B%E3%85%8B (https://ko.wikipedia.org/wiki/%EC%82%AC%EC%9A%A9%EC%9E%90%ED%86%A0%EB%A1%A0:%EC%95%88%EB%85%95%ED%95%98%EC%84%B8%EC%9A%94_%E3%85%8B%E3%85%8B)

권균

<https://ko.wikipedia.org/wiki/%EA%B6%8C%EA%B7%A0> ([http s://ko.wikipedia.org/wiki/%EA%B6%8C%EA%B7%A0](https://ko.wikipedia.org/wiki/%EA%B6%8C%EA%B7%A0))

라이트웨이브 3D로 만든 작품 목록

https://ko.wikipedia.org/wiki/%EB%9D%BC%EC%9D%B4%ED%8A%B8%EC%9B%A8%EC%9D%B4%EB%B8%8C_3D%EB%A1%9C_%EB%A7%8C%EB%93%A0_%EC%9E%91%ED%92%88_%EB%AA%A9%EB%A1%9D ([https://ko.wikipedia.org/wiki/%E](https://ko.wikipedia.org/wiki/%EB%9D%BC%EC%9D%B4%ED%8A%B8%EC%9B%A8%EC%9D%B4%EB%B8%8C_3D%EB%A1%9C_%EB%A7%8C%EB%93%A0_%EC%9E%91%ED%92%88_%EB%AA%A9%EB%A1%9D)

In [14]:

```
#print(urlList)
print(titleList)
```

```
['라우터브루넨', '사용자토론:안녕하세요 ㅋㅋ', '권균', '라이트웨이브 3D로
만든 작품 목록', '권균', '라우터브루넨', '강귀손', '3ds 맥스로 만든 작품 목
록', '강귀손', '데드 오어 얼라이브 6', '정문형', '사용자:안녕하세요 ㅋㅋ', '마
야로 만든 작품 목록', '라우터브루넨', '정문형', 'DC 유니버스 온라인', '김치
삼', '한샤오펑', '1월 19일', '윤호 (평정공)', '다윈 프로젝트', '위키백과:사랑
방/2022년 제3주', '위키백과:사용자 관리 요청/2022년 제3주', '허종 (1434
년)', '김치삼', 'D4DJ', '사용자:안녕하세요 ㅋㅋ', '윤사훈', '성봉조', '11월
25일', '사용자:안녕하세요 ㅋㅋ', '카운터-스트라이크: 글로벌 오픈시브', '한
백륜', '아침마당', 'Next Level (aespa의 노래)', '코즈믹브레이크', '사용자
토론:Ha98574', '윤사분', '안나 라자레바', '강순', '커맨드 앤 컨커 (최소된
비디오 게임)', '강순', '아침마당의 에피소드 목록 (2022년)', '이인손', '컴뱃
암즈', '라우터브루넨', '이인손', 'Savage (aespa의 EP)', '정분', '콜 오브 듀
티: 워존']
```

In [15]:

```
urlList
```

Out[15]:

```
['https://ko.wikipedia.org/wiki/%EB%9D%BC%EC%9A%B0%E
D%84%B0%EB%B8%8C%EB%A3%A8%EB%84%A8',
'https://ko.wikipedia.org/wiki/%EC%82%AC%EC%9A%A9%E
C%9E%90%ED%86%A0%EB%A1%A0:%EC%95%88%EB%85%9
5%ED%95%98%EC%84%B8%EC%9A%94_%E3%85%8B%E3%
85%8B',
'https://ko.wikipedia.org/wiki/%EA%B6%8C%EA%B7%A0',
'https://ko.wikipedia.org/wiki/%EB%9D%BC%EC%9D%B4%E
D%8A%B8%EC%9B%A8%EC%9D%B4%EB%B8%8C_3D%EB%A
1%9C_%EB%A7%8C%EB%93%A0_%EC%9E%91%ED%92%88
_%EB%AA%A9%EB%A1%9D',
'https://ko.wikipedia.org/wiki/%EA%B6%8C%EA%B7%A0',
'https://ko.wikipedia.org/wiki/%EB%9D%BC%EC%9A%B0%E
D%84%B0%EB%B8%8C%EB%A3%A8%EB%84%A8',
'https://ko.wikipedia.org/wiki/%EA%B0%95%EA%B7%80%E
C%86%90',
'https://ko.wikipedia.org/wiki/3ds_%EB%A7%A5%EC%8A%A
4%EB%A1%9C_%EB%A7%8C%EB%93%A0_%EC%9E%91%E
D%92%88_%EB%AA%A9%EB%A1%9D',
'htnns://ko wikinedia org/wiki/%FA%R0%95%FA%R7%80%F
```

데이트 프레임으로 만들기

In [19]:

```
columns = ["title", "category", "content_text"]
df = pd.DataFrame(columns=columns)
```

In [20]:

```
for i in range(10):
    excutable_path = "chromedriver.exe"
    driver = webdriver.Chrome(executable_path=excutable_path)
    driver.get(urlList[i])
    req = driver.page_source
    soup = BeautifulSoup(req, 'html.parser')

    contents_table = soup.find(name="main")

    ### 타이틀 추출
    title = contents_table.find_all('h1')[0]
    if title is not None: #타이틀에 아무내용도 없으면
        row_title = title.text.replace("\n", " ") #빈 칸으로 교체
    else:
        row_title = ""

    ### 카테고리 추출

    if len(contents_table.select("#mw-normal-catlinks")) > 0: # article ul 로 검색한 결과
        category = contents_table.select("#mw-normal-catlinks")[0] # 제일 첫번째 article
    else:
        category = None

    if category is not None: #카테고리가 비어 있으면
        row_category = category.text.replace("\n", " ") #빈 칸으로 교체
    else:
        row_category = ""

    ### 내용 추출
    content_paragraphs = contents_table.select("#mw-content-text > div.mw-parser-o
    content_corpus_list = []
    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("")

    row = [row_title, row_category, "".join(content_corpus_list)]
    series = pd.Series(row, index=df.columns)
    df = df.append(series, ignore_index=True)
    driver.close()
```


In [21]:

```
df
```

Out[21]:

	title	category	content_text
0	라우터브루넨	분류: 스위스의 도시베른주베르너 오버란트	라우터브루넨(독일어: Lauterbrunnen)는 스위스 베른주에 위치한 도시로, ...
1	사용자토론:안녕하세요 ㅋㅋ		
2	권균	분류: 1464년 출생1526년 사망안동 권씨조선의 문신우의정15세기 한국 사람16...	
3	라이트웨이브 3D로 만든 작품 목록	분류: 3차원 그래픽 소프트웨어매킨토시용 소프트웨어컴퓨터 지원 설계 소프트웨어3차원...	
4	권균	분류: 1464년 출생1526년 사망안동 권씨조선의 문신우의정15세기 한국 사람16...	
5	라우터브루넨	분류: 스위스의 도시베른주베르너 오버란트	라우터브루넨(독일어: Lauterbrunnen)는 스위스 베른주에 위치한 도시로, ...

In []:

In []: