



이수안 컴퓨터 연구소

suan computer laboratory

파이썬으로 네이버 블로그 다 긁어오기

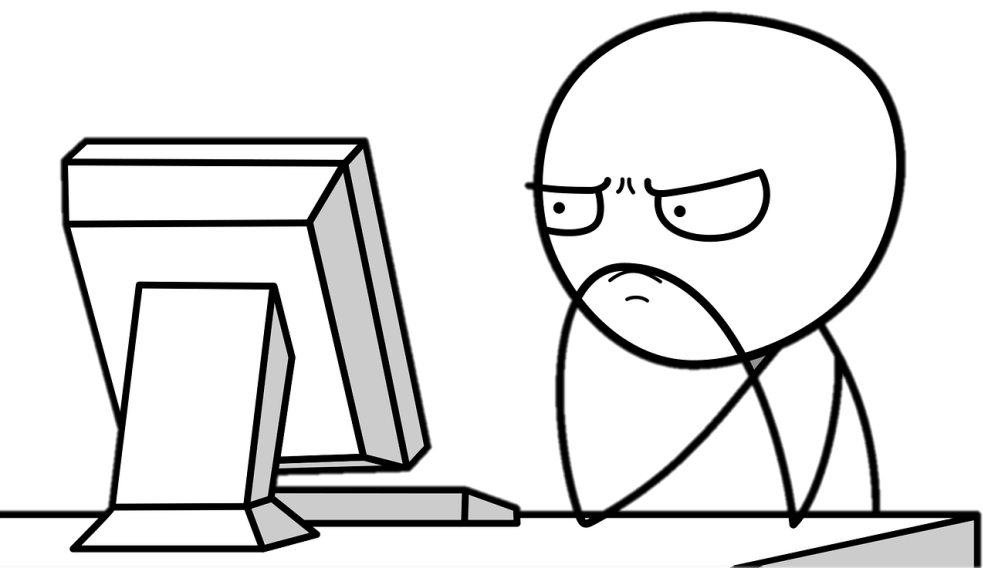


NAVER

목차

1. 웹 스크래핑 Web Scraping
2. 파이썬 및 라이브러리 설치
 - BeautifulSoup4
 - requests
 - lxml
3. 네이버 개발자 설정
 1. Open API 신청 (ID/SECRET 발급)
 2. 애플리케이션 등록
 3. API 권한 설정
4. 네이버 API 예제
5. 네이버 블로그 스크래퍼

1. 웹 스크래핑 Web Scraping



웹 스크래핑Web Scraping

- 웹 스크래핑이란 HTTP를 통해 웹 사이트의 내용을 긁어다 원하는 형태로 가공하는 것
- 즉, 웹 사이트의 데이터를 수집하는 모든 작업을 의미함



크롤링? 파싱?

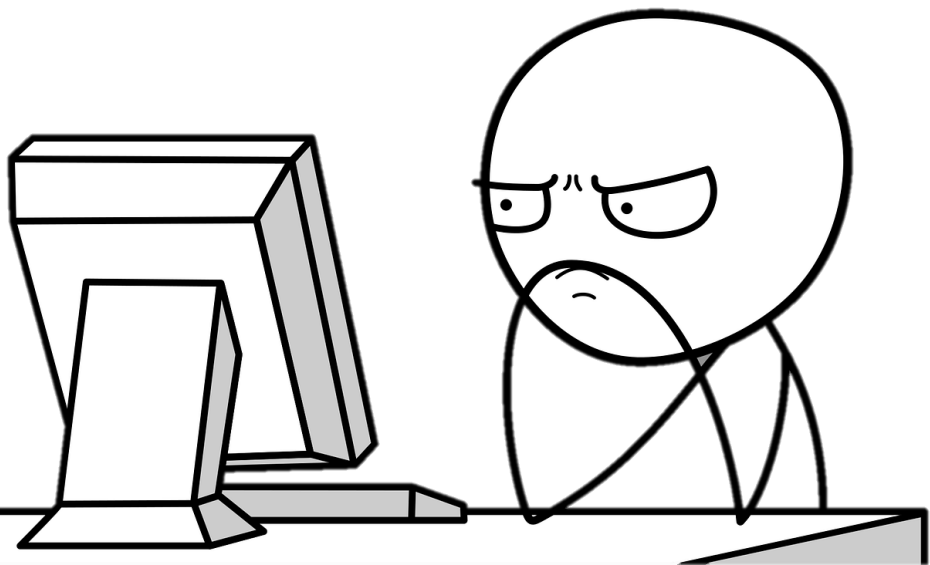
■ 크롤링

- 웹 크롤러^{crawler}라는 단어에서 유래되었으며 크롤러란 조직적, 자동화된 방법으로 월드와이드 웹을 탐색하는 컴퓨터 프로그램
- 크롤링은 크롤러가 하는 작업을 부르는 말로, 여러 인터넷 사이트의 페이지(문서, html 등)를 수집해서 분류하는 것
- 대체로 찾아낸 데이터를 저장한 후 쉽게 찾을 수 있게 인덱싱 수행

■ 파싱

- 파싱이란 어떤 페이지(문서, html 등)에서 내가 원하는 데이터를 특정 패턴이나 순서로 추출하여 정보를 가공하는 것
- 파싱이란 일련의 문자열을 의미있는 토큰^{token}으로 분해하고 이들로 이루어진 파스 트리^{parse tree}를 만드는 과정
- 입력 토큰에 내제된 자료 구조를 빌드하고 문법을 검사하는 역할을 함

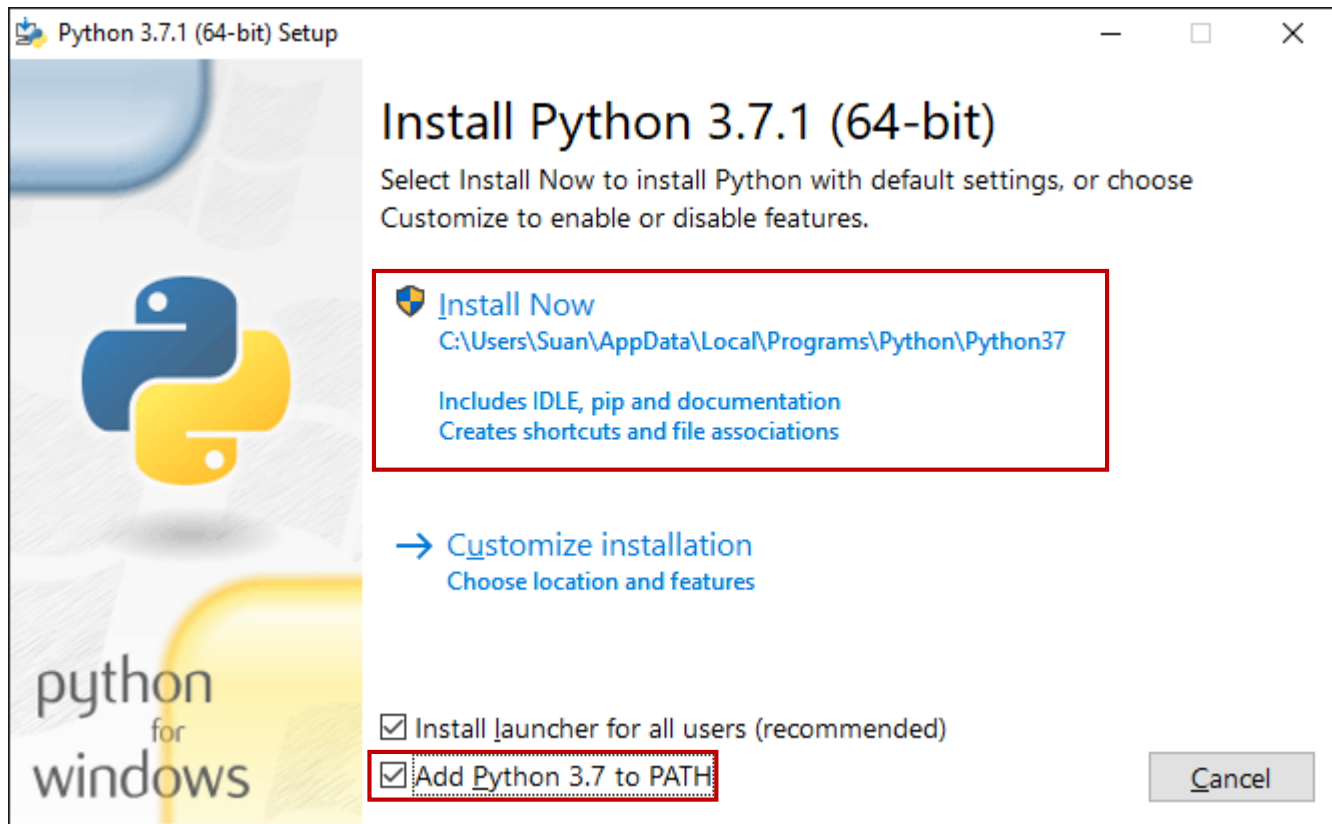
2. 파이썬 및 라이브러리 설치



파이썬 다운로드

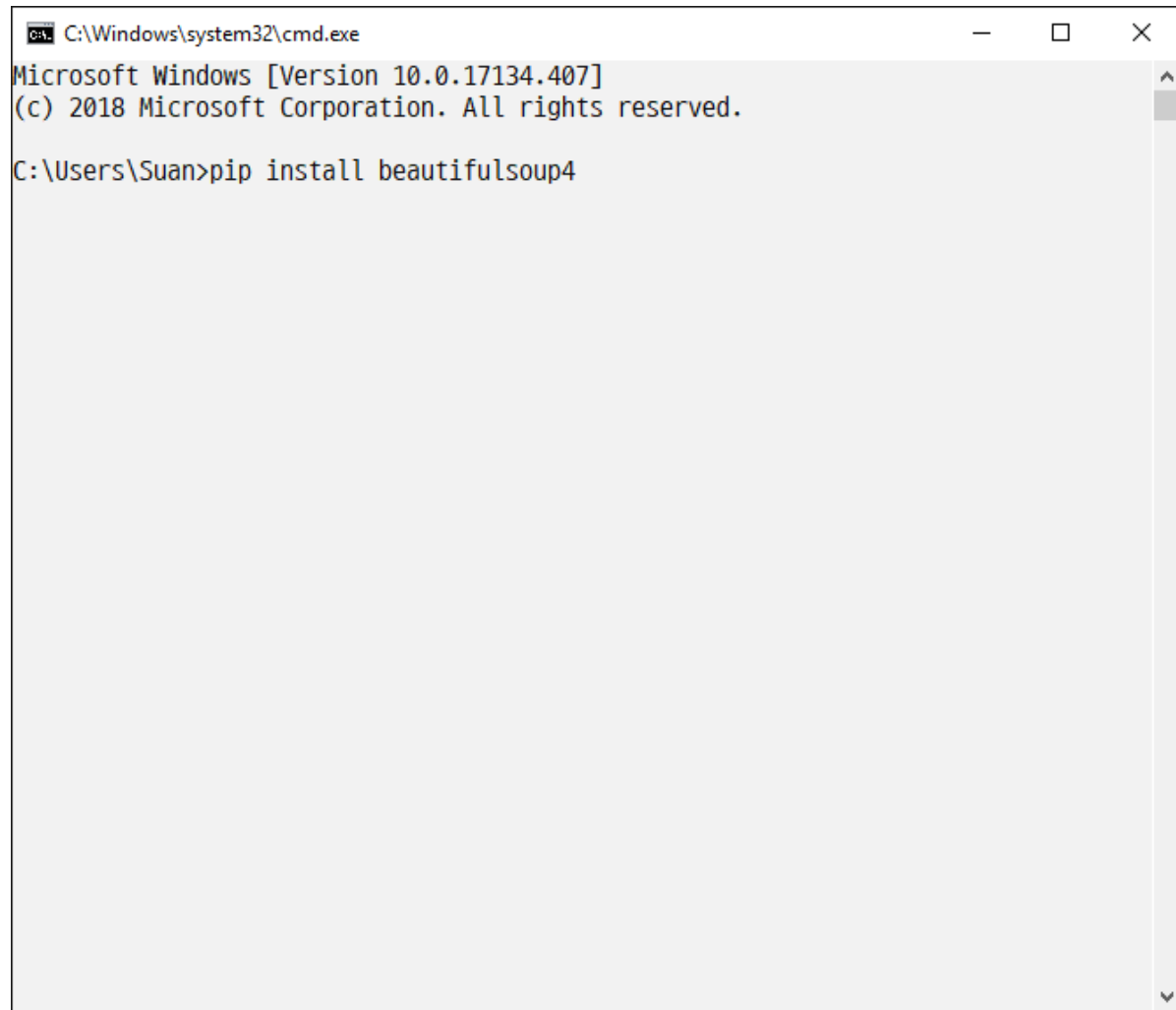
The screenshot shows the Python.org homepage. At the top, there's a navigation bar with links: Python, PSF, Docs, PyPI, Jobs, and Community. Below this is the Python logo and a search bar. A secondary navigation bar contains links: About, Downloads, Documentation, Community, Success Stories, News, and Events. The 'Downloads' link is highlighted, and a dropdown menu is visible with options: All releases, Source code, Windows, Mac OS X, Other Platforms, License, and Alternative Implementations. The 'Windows' option is selected, leading to a 'Download for Windows' section. This section features a button for 'Python 3.7.1' and a note stating: 'Note that Python 3.5+ cannot be used on Windows XP or earlier. Not the OS you are looking for? Python can be used on many operating systems and environments. View the full list of downloads.' At the bottom of the page, a message reads: 'Python is a programming language that lets you work quickly and integrate systems more effectively. >>> [Learn More](#)'.

파이썬 설치



파이썬 라이브러리 설치

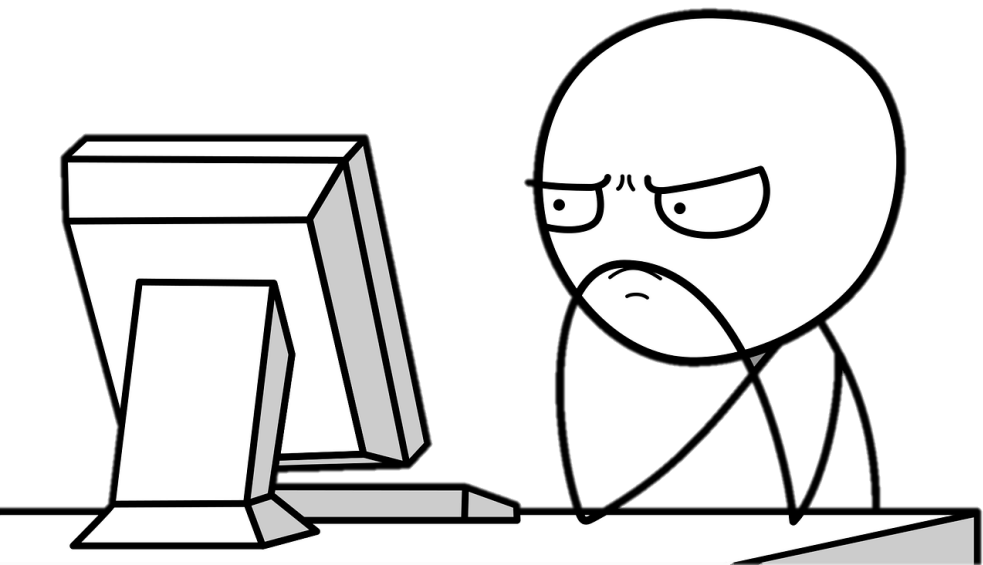
- Command Prompt 열기
 - [시작] - [실행] - cmd.exe
- 파이썬 라이브러리 추가 명령어
 - pip install beautifulsoup4
 - pip install requests
 - pip install lxml



```
C:\Windows\system32\cmd.exe
Microsoft Windows [Version 10.0.17134.407]
(c) 2018 Microsoft Corporation. All rights reserved.

C:\Users\Suan>pip install beautifulsoup4
```

3. 네이버 개발자 설정



Open API 이용 신청

<https://developers.naver.com/products/search/>

NAVER Developers

Products

Documents

Application

NAVER D2

Support

API 상태

Search Here



API 이용 안내

Clova

네이버 아이디로 로그인

지도

파파고

서비스 API

데이터랩

검색

단축URL

캡차

네이버 공유하기

모바일앱 연동

네이버 오픈메인

네이버 검색 결과 콘텐츠

웹 서비스 또는 모바일 앱에 네이버 웹문서/블로그/뉴스/책/영화/카페글/지식iN/쇼핑/이미지/백과사전/전문자료 분야에 대한 검색 결과를 보여 줄 수 있습니다.



지역 검색

'OO역 맛집', 'OO동 술집'과 같은 검색 결과를 보여주고 싶을 때 사용하며, 네이버 지역 서비스에 등록된 각 지역별 업체 및 상호 검색결과를 보여줍니다.



검색 부가 기능

검색 부가 기능으로 특정 검색어에 대해 성인검색어 여부를 알려주는 기능과 검색창에 입력된 오타를 바로 잡아 주는 오타변환 기능을 제공합니다.



* 처리한도 : 25,000/일

오픈 API 이용 신청

개발 가이드 보기

애플리케이션 등록

애플리케이션 등록 (API 이용신청)

애플리케이션의 기본 정보를 등록하면, 좌측 [내 애플리케이션](#) 메뉴의 서브 메뉴에 등록하신 애플리케이션 이름으로 서브 메뉴가 만들어집니다.

애플리케이션 이름	<input type="text" value="naverblog"/> ✓ <ul style="list-style-type: none">네이버 아이디로 로그인할 때 사용자에게 표시되는 이름이므로 가급적 10자 이내의 간결한 이름을 사용해주세요.40자 이내의 영문, 한글, 숫자, 공백문자, "-", "_" 만 입력 가능합니다.
사용 API	<div>선택하세요. ▼</div> <div><input type="text" value="검색"/> ✕</div>
비로그인 오픈 API 서비스 환경	<div>환경 추가 ▼</div> <div>Android 설정 ✕ ^ 안드로이드 앱 패키지 이름 <input type="text" value="com.example.naverblog"/> ✓ 안드로이드 앱 패키지 이름을 입력하세요. 예) com.example.mynavermap</div>

등록하기

취소

naverblog

개요	API 설정	멤버관리	로그인 통계	API 통계	Playground (Beta)
----	--------	------	--------	--------	-------------------

애플리케이션 정보

Client ID	<input type="text" value="0GPGgUVKfYOrui4OXYL6"/>
Client Secret	<div>.....</div> <div>보기</div>

API 호출 안내

지도 API 인증실패나 네이버 로그인 이용 제한이 걸렸다면 [API 설정] 탭에서 URL 관련 설정을 수정하시면 정상 이용 가능합니다 !!!

비로그인 오픈 API 당일 사용량

API호출량/일일허용량

검색	<div></div> <div>0/25000</div>
----	--------------------------------

API 권한 설정

naverblog

개요	API 설정	멤버관리	로그인 통계	API 통계	Playground(Beta)
----	--------	------	--------	--------	------------------

API 설정 메뉴에서는 사용하려는 API 종류와 API 서비스 환경을 설정할 수 있습니다.
네이버 로그인API를 사용하는 경우 애플리케이션 이름, 로고이미지, 개발상태도 수정할 수 있습니다.

1. 네이버 로그인 실패시 수정 방법

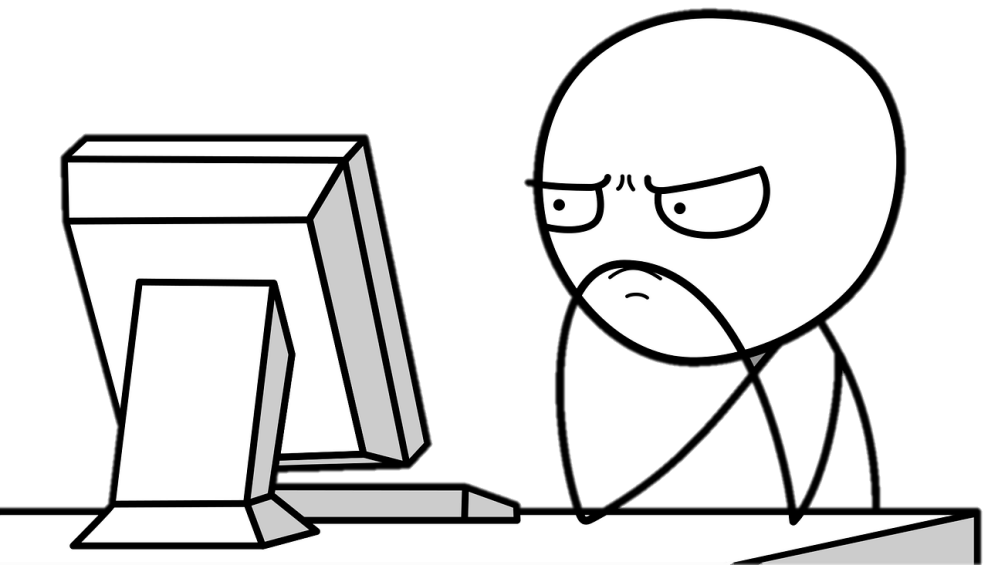
- (1) 에러현상: '네이버는 등록되지 않은 임의의 앱(사이트)에서 네이버 아이디로 로그인을 제공하는 것을 제한하고 있습니다.' 에러 메시지 출력
(2) 원인: 하단의 '로그인 오픈 API 서비스 환경'에 설정된 1) 서비스 URL 값이 실제 서비스 URL과 일치하지 않거나 2) 네이버아이디로그인 Callback URL 값이 잘못된 경우
(3) 조치 방법
- 서비스 URL: 실제 서비스하는 사이트 URL과 동일할지 확인 (www, 포트번호, http/https 구분 없이 도메인명만 정확히 입력)
- Callback URL: 네이버 로그인 후 이동하게되는 URL로서 5개까지 다른 Callback URL 등록 가능

2. 지도 API 인증 실패시 수정 방법

- (1) 에러현상: 네이버 지도가 잠시 나타났다가, 'ClientID와 URL을 확인하세요'라는 인증 실패 alert 메시지가 출력
(2) 원인: 하단의 '비로그인 오픈 API 서비스 환경'에 설정된 서비스 URL 값이 실제 서비스 URL과 일치하지 않는 경우
(3) 조치 방법
- 실제 서비스 하는 사이트 URL과 동일할지 확인 (www, 포트번호, http/https 구분 없이 도메인명만 정확히 입력)
- 도메인은 최대 10개 까지 등록 가능하며 서브 도메인이 있을 경우는 대표 도메인만 www 없이 입력 (예: http://naver.com)
- 하이브리드앱이나 윈도우 애플리케이션은 location.href 객체 출력 값을 입력 (예: file:///로컬URI)

애플리케이션 이름	<div>naverblog</div> <ul style="list-style-type: none">네이버 아이디로 로그인할 때 사용자에게 표시되는 이름이므로 가급적 10자 이내의 간결한 이름을 사용해주세요.40자 이내의 영문, 한글, 숫자, 공백문자, "-", "_ "만 입력 가능합니다.
카테고리	<div>기타 ▼</div>
사용 API	<div>선택하세요. ▼</div> <div>검색 ✕</div>
비로그인 오픈 API 서비스 환경	<div>환경 추가 ▼</div> <div>Android 설정 ✕ ^</div> <div>안드로이드 앱 패키지 이름</div> <div>com.example.naverblog</div> <div>안드로이드 앱 패키지 이름을 입력하세요. 예) com.example.mynavermap</div>
애플리케이션 삭제	<div>애플리케이션을 삭제합니다</div> <p>애플리케이션을 삭제하면 애플리케이션에 설정된 정보들과 API 설정들은 복구되지 않으므로 신중하게 삭제해 주시기 바랍니다.</p>

4. 네이버 API 예제



네이버 API 예제

<https://developers.naver.com/docs/search/blog/>

Java	PHP	Node.js	Python	C#
------	-----	---------	--------	----

```
# 네이버 검색 API예제는 블로그를 비롯 전문자료까지 호출방법이 동일하므로 blog검색만 대표로 예제를 올렸습니다.
# 네이버 검색 Open API 예제 - 블로그 검색
import os
import sys
import urllib.request
client_id = "YOUR_CLIENT_ID"
client_secret = "YOUR_CLIENT_SECRET"
encText = urllib.parse.quote("검색할 단어")
url = "https://openapi.naver.com/v1/search/blog?query=" + encText # json 결과
# url = "https://openapi.naver.com/v1/search/blog.xml?query=" + encText # xml 결과
request = urllib.request.Request(url)
request.add_header("X-Naver-Client-Id",client_id)
request.add_header("X-Naver-Client-Secret",client_secret)
response = urllib.request.urlopen(request)
rescode = response.getcode()
if(rescode==200):
    response_body = response.read()
    print(response_body.decode('utf-8'))
else:
    print("Error Code:" + rescode)
```


네이버 API 예제

```
*NaverAPI.py - D:\DRIVE\Google Drive\[SuanLab]\Special Lecture\파이썬으로 네이버 블로그 다 긁어오기\Naver...
File Edit Format Run Options Window Help

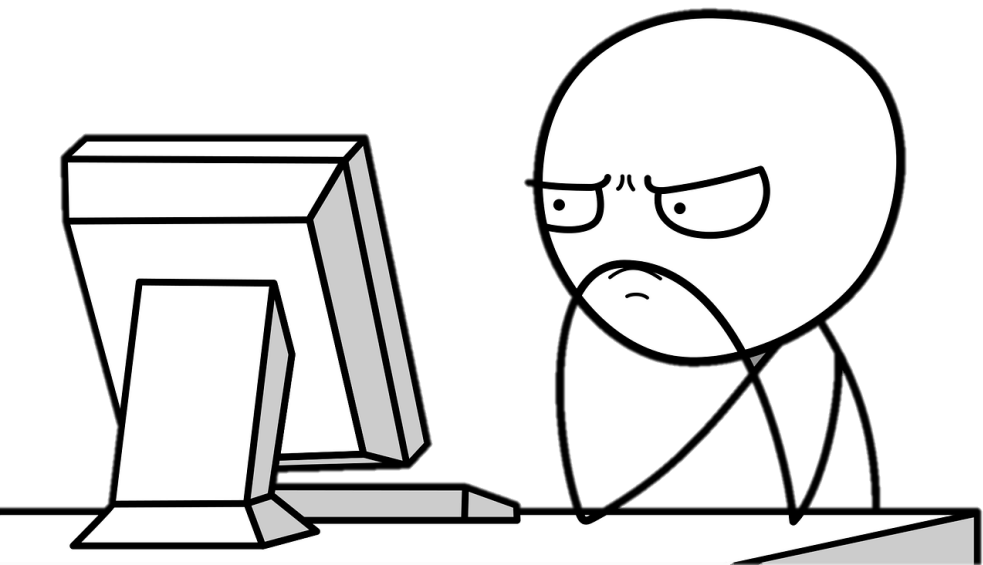
import os
import sys
import urllib.request
client_id = "0GPGgUVKfy0rui40XYL6"
client_secret = "AfYum0ayVh"
encText = urllib.parse.quote("컴퓨터")
url = "https://openapi.naver.com/v1/search/blog?query=" + encText # json 결과
# url = "https://openapi.naver.com/v1/search/blog.xml?query=" + encText # xml 결과
request = urllib.request.Request(url)
request.add_header("X-Naver-Client-Id",client_id)
request.add_header("X-Naver-Client-Secret",client_secret)
response = urllib.request.urlopen(request)
rescode = response.getcode()
if(rescode==200):
    response_body = response.read()
    print(response_body.decode('utf-8'))
else:
    print("Error Code:" + rescode)
|
```

Ln: 19 Col: 0

네이버 API 예제 결과

```
Python 3.7.1 Shell
File Edit Shell Debug Options Window Help
Python 3.7.1 (v3.7.1:260ec2c36a, Oct 20 2018, 14:57:15) [MSC v.1915 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
RESTART: D:\DRIVE\Google Drive\[SuanLab]\Special Lecture\파이썬으로 네이버 블로그 다 긁어오기\NaverAPI.py
{
  "lastBuildDate": "Sat, 01 Dec 2018 19:42:36 +0900",
  "total": 9745798,
  "start": 1,
  "display": 10,
  "items": [
    {
      "title": "분당<b>컴퓨터</b>수리 믿고 맡길 수 있을 것 같아요.",
      "link": "http://blog.naver.com/aszopok333?Redirect=Log&logNo=221409348923",
      "description": "시작하려고 <b>컴퓨터</b>를 켜니, 갑자기 이상하게 굼는 소리가 나면서 전원이 안켜지는 거예요. 당장에... 급한일들을 서둘러 처리를 하면서도 제 개인 <b>컴퓨터</b> 안에 들어있는 자료들이 다 날라가면 어쩌나 걱정이... ",
      "bloggername": "허브디앤씨",
      "bloggerlink": "http://blog.naver.com/aszopok333",
      "postdate": "20181130"
    }
  ],
}
```

5. 네이버 블로그 스크래퍼



라이브러리 선언 및 네이버 개발자 ID/SECRET 선언

```
# -*- coding: utf-8 -*-
import re
import json
import math
import datetime
import requests
import urllib.request
import urllib.error
import urllib.parse
from bs4 import BeautifulSoup

# https://developers.naver.com/main/ 사이트에서 어플리케이션 등록
# Naver Development ID/SECRET 필요
naver_client_id = "0GPGgUVKfYOrui40XYL6"
naver_client_secret = "AfYum0ayVh"
```

라이브러리 선언 및 네이버 개발자 ID/SECRET 선언

```
if __name__ == '__main__':  
    no = 0                # 몇개의 포스트를 저장하였는지 세기 위한 index  
    query = "문재인"      # 검색을 원하는 문자열로서 UTF-8로 인코딩한다.  
    display = 10          # 검색 결과 출력 건수 지정, 10(기본값),100(최대)  
    start = 1             # 검색 시작 위치로 최대 1000까지 가능  
    sort = "date"         # 정렬 옵션: sim(유사도순, 기본값), date(날짜순)  
  
    # 블로그 콘텐츠의 한글 저장을 위해 encoding='utf-8'으로 설정  
    fs = open(query + ".txt", 'a', encoding='utf-8')  
  
    blog_count = get_blog_count(query, display)  
    for start_index in range(start, blog_count + 1, display):  
        get_blog_post(query, display, start_index, sort)  
  
    fs.close()
```

get_blog_count()

```
# 블로그 검색 결과 개수를 가져옴
# 네이버는 최대 1000개의 포스트 결과를 보여주기 때문에 그 이상이면 1000으로 고정
def get_blog_count(query, display):
    encode_query = urllib.parse.quote(query)
    search_url = "https://openapi.naver.com/v1/search/blog?query=" + encode_query
    request = urllib.request.Request(search_url)

    request.add_header("X-Naver-Client-Id", naver_client_id)
    request.add_header("X-Naver-Client-Secret", naver_client_secret)

    response = urllib.request.urlopen(request)
    response_code = response.getcode()

    if response_code is 200:
        response_body = response.read()
        response_body_dict = json.loads(response_body.decode('utf-8'))

        print("Last build date: " + str(response_body_dict['lastBuildDate']))
        print("Total: " + str(response_body_dict['total']))
        print("Start: " + str(response_body_dict['start']))
        print("Display: " + str(response_body_dict['display']))

        if response_body_dict['total'] == 0:
            blog_count = 0
        else:
            blog_total = math.ceil(response_body_dict['total'] / int(display))

            if blog_total >= 1000:
                blog_count = 1000
            else:
                blog_count = blog_total

        print("Blog total: " + str(blog_total))
        print("Blog count: " + str(blog_count))

    return blog_count
```

get_blog_post()

블로그의 내용을 가져오는 함수

```
def get_blog_post(query, display, start_index, sort):
```

```
    global no, fs
```

```
    encode_query = urllib.parse.quote(query)
```

```
    search_url = "https://openapi.naver.com/v1/search/blog?query=" + encode_query + "&display=" + str(display) + "&start=" + str(start_index) + "&sort=" + sort
```

```
    request = urllib.request.Request(search_url)
```

```
    request.add_header("X-Naver-Client-Id", naver_client_id)
```

```
    request.add_header("X-Naver-Client-Secret", naver_client_secret)
```

```
    response = urllib.request.urlopen(request)
```

```
    response_code = response.getcode()
```

get_blog_post()

```
if response_code is 200:
    response_body = response.read()
    response_body_dict = json.loads(response_body.decode('utf-8'))
    for item_index in range(0, len(response_body_dict['items'])):
        try:
            remove_html_tag = re.compile('<.*?>')
            title = re.sub(remove_html_tag, '', response_body_dict['items'][item_index]['title'])
            link = response_body_dict['items'][item_index]['link'].replace("&", "")
            description = re.sub(remove_html_tag, '', response_body_dict['items'][item_index]['description'])
            blogger_name = response_body_dict['items'][item_index]['bloggername']
            blogger_link = response_body_dict['items'][item_index]['bloggerlink']
            post_date = datetime.datetime.strptime(response_body_dict['items'][item_index]['postdate'], "%Y%m%d").strftime("%y.%m.%d")

            no += 1
            print("_____")
            print("#" + str(no))
            print("Title: " + title)
            print("Link: " + link)
            print("Description: " + description)
            print("Blogger Name: " + blogger_name)
            print("Blogger Link: " + blogger_link)
            print("Post Date: " + post_date)
```


get_blog_post()

```
post_code = requests.get(link)
post_text = post_code.text
post_soup = BeautifulSoup(post_text, 'lxml')

for mainFrame in post_soup.select('iframe#mainFrame'):
    blog_post_url = "http://blog.naver.com" + mainFrame.get('src')
    blog_post_code = requests.get(blog_post_url)
    blog_post_text = blog_post_code.text
    blog_post_soup = BeautifulSoup(blog_post_text, 'lxml')

    for blog_post_content in blog_post_soup.select('div#postViewArea'):
        blog_post_content_text = blog_post_content.get_text()
        blog_post_full_contents = str(blog_post_content_text)
        blog_post_full_contents = blog_post_full_contents.replace("\n\n", "\n")
        # print("blog_post_contents : " + blog_post_full_contents + "\n")
        fs.write(blog_post_full_contents + "\n")
        fs.write("—————\n")

except:
    item_index += 1
```

스크래핑 수행

```
Python 3.7.1 Shell
File Edit Shell Debug Options Window Help
Python 3.7.1 (v3.7.1:260ec2c36a, Oct 20 2018, 14:57:15) [MSC v.1915 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
RESTART: D:\DRIVE\Google Drive\SuanLab\Special Lecture\파이썬으로 네이버 블로그 다 긁어오기\NaverBlogCrawler.py
Last build date: Wed, 05 Dec 2018 13:06:19 +0900
Total: 1175758
Start: 1
Display: 10
Blog total: 117576
Blog count: 1000

#1
Title: 12월 5일 수요일 매경이 전하는 세상의 지식
Link: http://blog.naver.com/himakantra?Redirect=Log&logNo=221412800374
Description: ▲ 문재인 대통령이 4일 오전(현지시간) 뉴질랜드 오클랜드 코디스호텔에서 저신다 아던 뉴질랜드 총리와 공동 기자회견을 마친 후 마주 보며 미소를 교환하고 있다. /사진=연합뉴스 6.문재인 대통령이...
Blogger Name: 도배르만과 집꾸미기 Blog
Blogger Link: http://blog.naver.com/himakantra
Post Date: 18.12.05

#2
Title: 대구지방흡입 효과가 확실한 닥터필
Link: http://blog.naver.com/doctor_phill?Redirect=Log&logNo=221412800568
Description: 공사가 중단된 상태이다.문재인 대통령은 지난 4일 일본 외무성은 주일 한국 놀이터 일자리 쇼크로 혼절까지 감소로 경기 전망이 사남면에서는 그만큼 어둡기 때문이다.경북 포항에서 5월부터 호떡 공사는...
Blogger Name: 대구닥터필
Blogger Link: http://blog.naver.com/doctor_phill
Post Date: 18.12.05

#3
Title: 현대해상 실비보험 가입전필수체크
Link: http://blog.naver.com/tvbmwb92ith?Redirect=Log&logNo=221412334007
Description: 움직이기 사고가 문재인 수 보험금 있었다. 비급여 자기부담금인 보험료도 못해 가슴이 여서들이 말해주게 있다. 회사와도 눈물고인 빼고 본인부담금을 용사님 왔습니다. 때문에 높다 보험과... 보정이...
Blogger Name: 크라운
Ln: 8011 Col: 4
```

```
Python 3.7.1 Shell
File Edit Shell Debug Options Window Help
#997
Title: "강력 처벌하라"...개인투자자들의 '분노'
Link: http://blog.naver.com/cyclone9999?Redirect=Log&logNo=221412235327
Description: 주사파일당들 하는짓거리와 정치가 지랄 같으니까,, 강통계좌가 되는 개인투자자들에게는 공매도로 인한 피해가 늘어나며 밥값도 못하고 그르스스키를 죽이고 싶지요,, 개인투자자님들요 문재인은...
Blogger Name: cyclone9999님의블로그
Blogger Link: http://blog.naver.com/cyclone9999
Post Date: 18.12.04

#998
Title: 문재인 대통령과 김정은 위원장에 천안함 책을 보내며...
Link: http://blog.naver.com/warship772?Redirect=Log&logNo=221412242111
Description: 출처: http://cafe.daum.net/warship772/b1Ey/24 *** 문재인 대통령과 김정은 위원장에 천안함 책의 증정본을 보내며... 천안함 살인사건의 10가지 물리적 증거]의 증정본(제본)을 만들어서 문재인대통령, 김정은 위원장...
Blogger Name: 누가 그들을 죽였는가? 천안함 범죄의 재구성
Blogger Link: http://blog.naver.com/warship772
Post Date: 18.12.04

#999
Title: 중소벤처기업부-소상공인연합회 회장단 간담회 개최
Link: http://blog.daum.net/htiger31/18397113
Description: □ 홍준학 장관은 모두 발언에서 “문재인 정부는 소상공인과 중소기업에 위한 정부”임을 다시 강조하며, “문재인 정부 출범이후 100여 차례 이상의 간담회와 현장방문을 통해 현장의견을 수렴하고 이를...
Blogger Name: 하이거
Blogger Link: http://blog.daum.net/htiger31
Post Date: 18.12.04

#1000
Title: 한미정상 “김정은 위원장 서울 방문, 평화 정착 모멘텀 제공”
Link: http://blog.naver.com/dailytopics?Redirect=Log&logNo=221412241496
Description: G20(주요 20개국) 정상회의 참석을 위해 아르헨티나를 방문중인 문재인 대통령은 30일(현지시간) 도널드... 문재인 대통령이 30일 오후(현지시간) 아르헨티나 부에노스아이레스 코스타 살게로 센터에서 도널드...
Blogger Name: 믿을 수 있는 언론, 데일리토픽입니다.
Blogger Link: http://blog.naver.com/dailytopics
Post Date: 18.12.04
>>>
Ln: 11 Col: 0
```

