

파이썬 크롤링 (API)

최도진

01. 네이버 API를 이용한 크롤링

■ 크롤링이란

■ 크롤링

- 웹에서 데이터를 수집하는 작업
- 크롤러 또는 스파이더라는 프로그램으로 웹 사이트에서 데이터를 추출

• 웹 API

- 웹 API는 일반적으로 HTTP 통신을 사용하는데 사용
- 지도, 검색, 추가, 환율 등 다양한 정보를 가지고 있는 웹 사이트의 기능을 외부에서 쉽게 사용할 수 있도록 사용 절차와 규약을 정의한 것



그림 5-1 웹 API를 이용한 HTTP 요청과 응답

표 5-1 웹 API 제공자

종류	주소
네이버 개발자 센터	https://developers.naver.com
카카오 앱 개발 플랫폼 서비스	https://developers.kakao.com
페이스북 개발자 센터	https://developers.facebook.com
트위터 개발자 센터	https://developer.twitter.com
공공데이터포털	https://www.data.go.kr
세계 날씨	http://openweathermap.org
유료/무료 API 스토어	http://mashup.or.kr http://www.apistore.co.kr/api/apiList.do

01. 네이버 API를 이용한 크롤링

■ 네이버 개발자 가입

1. 네이버 개발자 센터 접속하기

- 네이버 개발자 센터([https:// developers.naver.com](https://developers.naver.com))에 접속

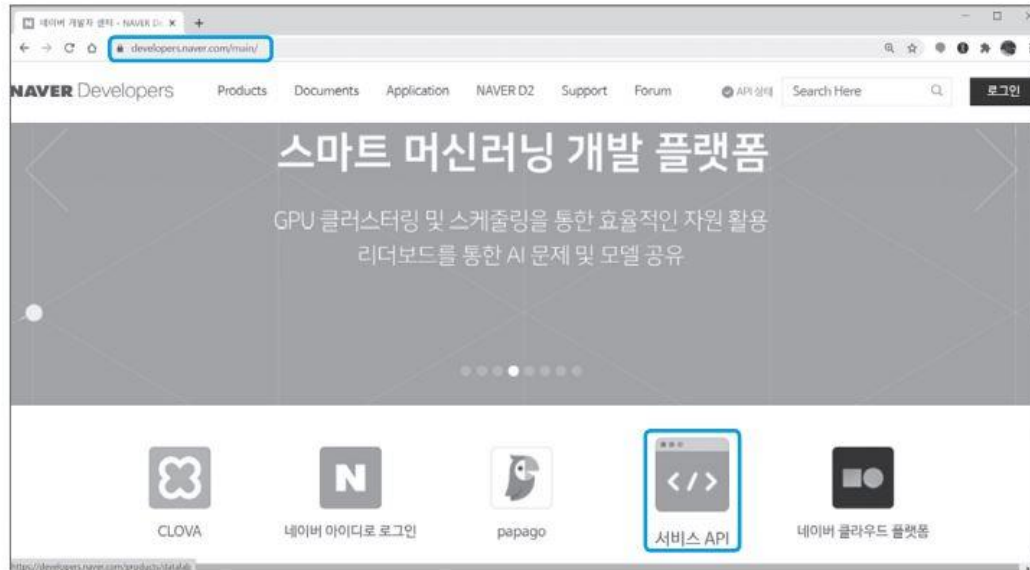


그림 5-2 네이버 개발자 센터

01. 네이버 API를 이용한 크롤링

■ 네이버 개발자 가입

2. 오픈 API 이용 신청하기

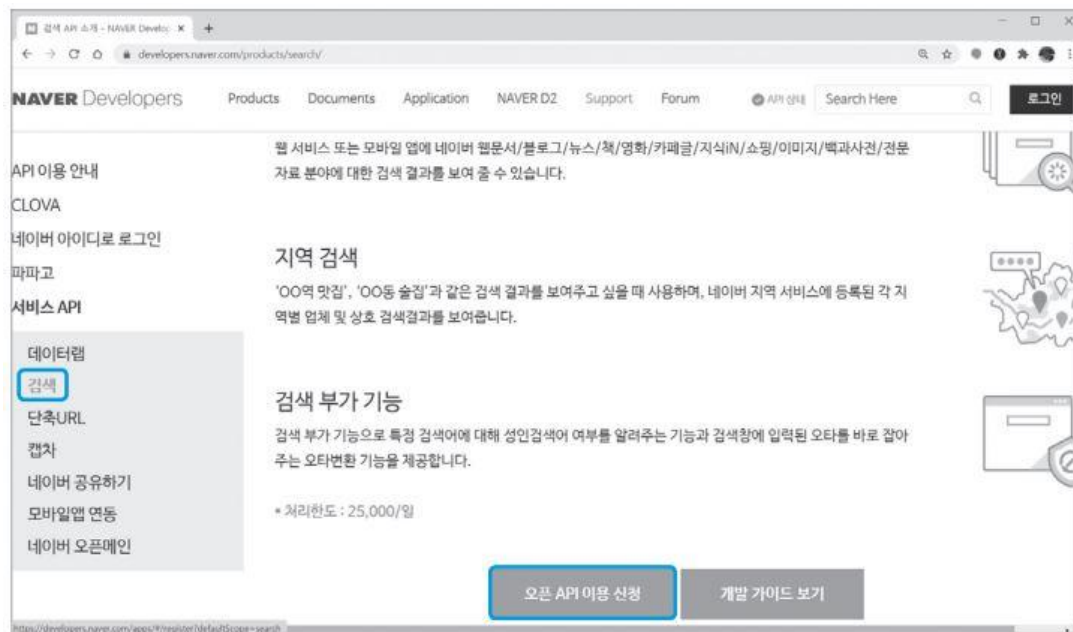


그림 5-3 오픈 API 이용 신청

01. 네이버 API를 이용한 크롤링

■ 네이버 개발자 가입

3. 애플리케이션 등록하기

내 애플리케이션

애플리케이션 등록

CLOVA Platform Console β

API 제휴 신청

계정 설정

애플리케이션 등록 (API 이용신청)

애플리케이션의 기본 정보를 등록하면, 좌측 내 애플리케이션 메뉴의 서브 메뉴에 등록하신 애플리케이션 이름으로 서브 메뉴가 만들어집니다.

애플리케이션 이름	<div>1 nvBj 입력</div> <div>• 네이버 아이디로 로그인할 때 사용자에게 표시되는 이름이므로 가독성 10자 이하의 간결한 이름을 사용해주세요. • 40자 이하의 영문, 한글, 숫자, 공백문자, "-", "_", "."만 입력 가능합니다.</div>
사용 API	<div>선택하세요.</div> <div>검색</div>
비로그인 오픈 API 서비스 환경	<div>2 환경 추가</div> <div>3 http://localhost 입력</div> <div>WEB 설정</div> <div>웹 서비스 URL (최대 10개)</div> <div>• 텍스트 줄 우측 끝의 "+" 버튼을 누르면 행이 추가되며, "-" 버튼을 누르면 행이 삭제됩니다. • http와 https는 구분하지 않습니다. • www는 빼고 입력해 주세요. (예: http://naver.com) • 서브 도메인이 있으면 대표 도메인명만 입력해 주세요. (예: http://naver.com) • 하이퍼링크 주소는 location.href 객체 출력 값을 입력하면 됩니다. (예: files://로컬 URL)</div>

등록하기 취소

그림 5-4 애플리케이션 등록

01. 네이버 API를 이용한 크롤링

■ 네이버 개발자 가입

4. 애플리케이션 정보 확인하기

The screenshot displays the Naver Developer Console (nvBig) interface. On the left, a sidebar lists '내 애플리케이션' (My Applications) with four entries labeled 'nvBig', and a section for '애플리케이션 등록' (Application Registration) including links for 'CLOVA Platform Console β', 'API 제휴 신청', and '계정 설정'. The main area is titled 'nvBig' and features a navigation bar with tabs: '개요' (Overview), 'API 설정' (API Settings), '멤버관리' (Member Management), '로그인 통계' (Login Statistics), 'API 통계' (API Statistics), and 'Playground (Beta)'. The 'API 설정' tab is active, showing '애플리케이션 정보' (Application Information). This section contains a table with two rows: 'Client ID' with the value '0LHQM4VX_MQM6JfkXofa' and 'Client Secret' with a masked value '*****'. A '보기' (View) button is located next to the Client Secret. Below this, there is a section titled 'API 호출 안내' (API Call Guide) with a note: '지도 API 인증실때나 네이버 로그인 이용 제한이 걸렸다면 [API 설정] 탭에서 URL 관련 설정을 수정하시면 정상 이용 가능합니다 !!!'.

애플리케이션 정보	
Client ID	0LHQM4VX_MQM6JfkXofa
Client Secret	***** 보기

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

그림 5-5 애플리케이션 정보 확인

01. 네이버 API를 이용한 크롤링

■ 네이버 개발자 가입

4. 애플리케이션 정보 확인하기



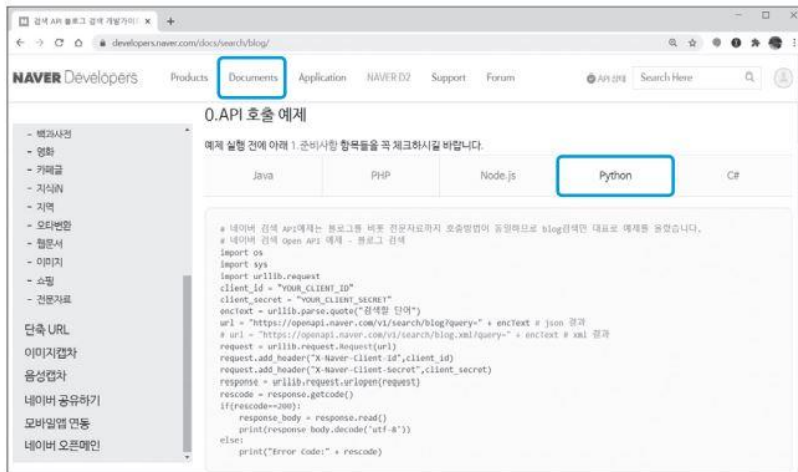
The screenshot displays the Naver Developer Console (nvBig) interface. On the left, a sidebar lists '내 애플리케이션' (My Applications) with four entries labeled 'nvBig', and a section for '애플리케이션 등록' (Application Registration) including links for 'CLOVA Platform Console β', 'API 재휴 신청' (API Re-apply), and '계정 설정' (Account Settings). The main area is titled 'nvBig' and features a navigation bar with tabs: '개요' (Overview), 'API 설정' (API Settings), '멤버관리' (Member Management), '로그인 통계' (Login Statistics), 'API 통계' (API Statistics), and 'Playground (Beta)'. The 'API 설정' tab is active, showing '애플리케이션 정보' (Application Information). This section contains a table with two rows: 'Client ID' with the value '0LHQM4VX_MQM6JfkXofa' and 'Client Secret' with a masked value '*****'. A '보기' (View) button is located next to the Client Secret. Below this, there is a section titled 'API 호출 안내' (API Call Guide) with a note: '지도 API 인증실때나 네이버 로그인 이용 제한이 걸렸다면 [API 설정] 탭에서 URL 관련 설정을 수정하시면 정상 이용 가능합니다 !!!' (If there is a restriction on using the Naver login or the Map API authentication, you can use it normally after modifying the URL-related settings in the [API Settings] tab !!!).

그림 5-5 애플리케이션 정보 확인

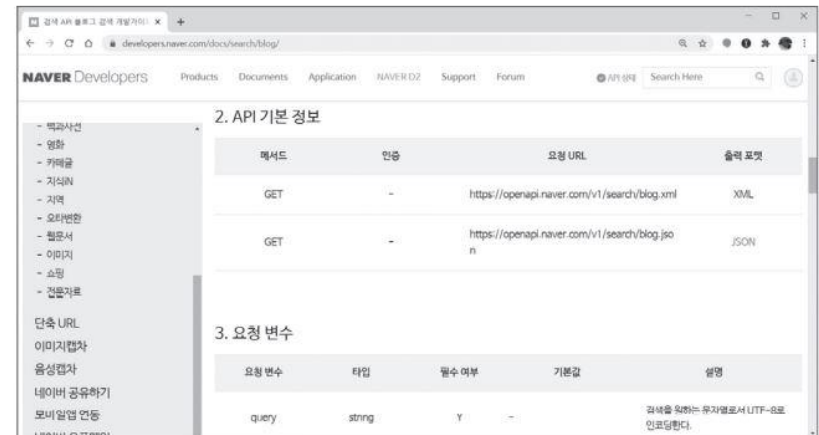
01. 네이버 API를 이용한 크롤링

■ 네이버 개발자 가입

5. 검색 API 이용 안내 페이지 확인하기



(a) 파이썬 코드로 확인



(b) API 기본 정보와 요청 변수 확인

그림 5-6 검색 API 이용 안내 페이지

뉴스 검색 테스트

■ <https://developers.naver.com/docs/serviceapi/search/news/news.md#%EB%89%B4%EC%8A%A4>

Documents > 서비스API > 검색

블로그

뉴스

책

성인 검색어 판별

백과사전

영화

카페글

지식IN

지역


오타변환

웹문서

이미지

쇼핑

전문자료

뉴스 

검색 > 뉴스

네이버 뉴스 검색 결과를 출력해주는 REST API입니다. 비로그인 오픈 API이므로 GET으로 호출할 때 HTTP Header에 애플리케이션 등록 시 발급받은 Client ID와 Client Secret 값을 같이 전송해 주시면 활용 가능합니다.

[오픈 API 이용 신청](#)

1. 준비사항

- 애플리케이션 등록: 네이버 오픈 API로 개발하시려면 먼저 '[Application-애플리케이션 등록](#)' 메뉴에서 애플리케이션을 등록하셔야 합니다.
[\[자세한 방법 보기\]](#)
- 클라이언트 ID와 secret 확인: '[내 애플리케이션](#)'에서 등록된 애플리케이션을 선택하면 Client ID와 Client Secret 값을 확인할 수 있습니다.

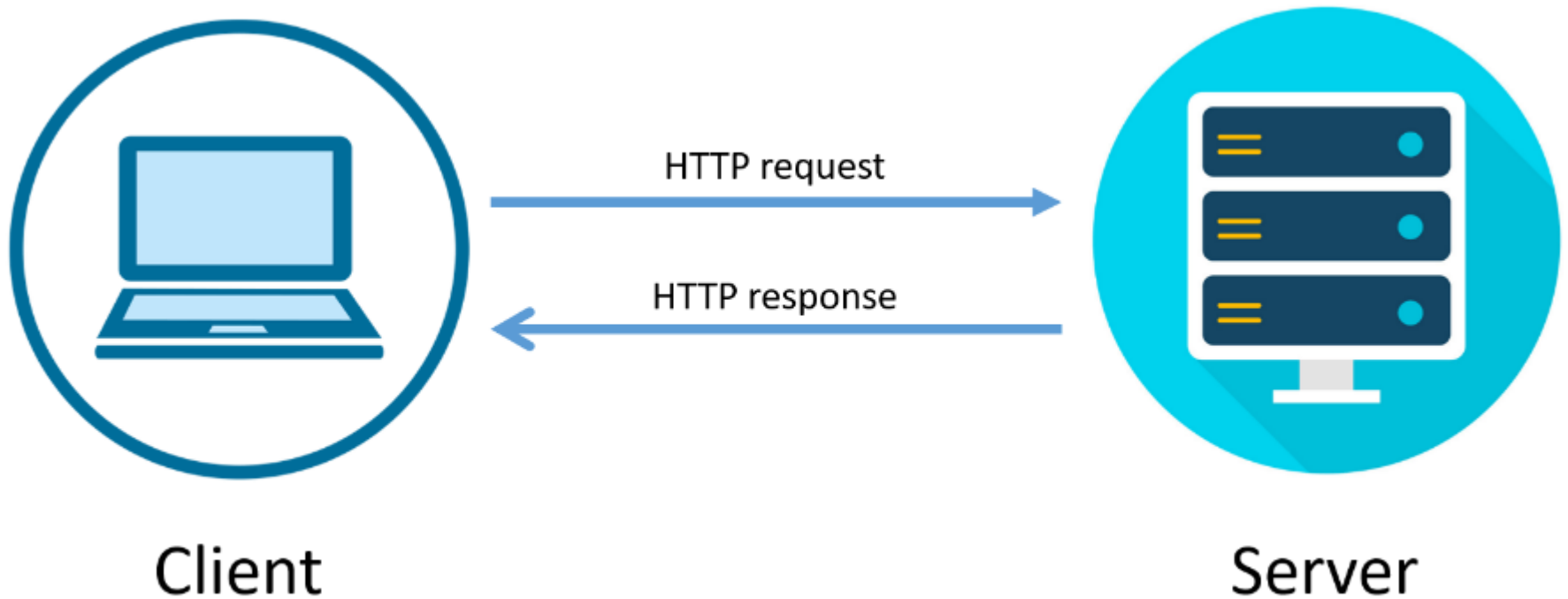
```
import os
import sys
import urllib.request
import urllib
import json
import datetime
```

```
encText = urllib.parse.quote("검색할 단어")
url = "https://openapi.naver.com/v1/search/news?query=" + encText

print(url)
```

```
{
  errorMessage: "Not Exist Client ID : Authentication failed. (인증에 실패했습니다.)",
  errorCode: "024"
}
```

뉴스 검색 테스트

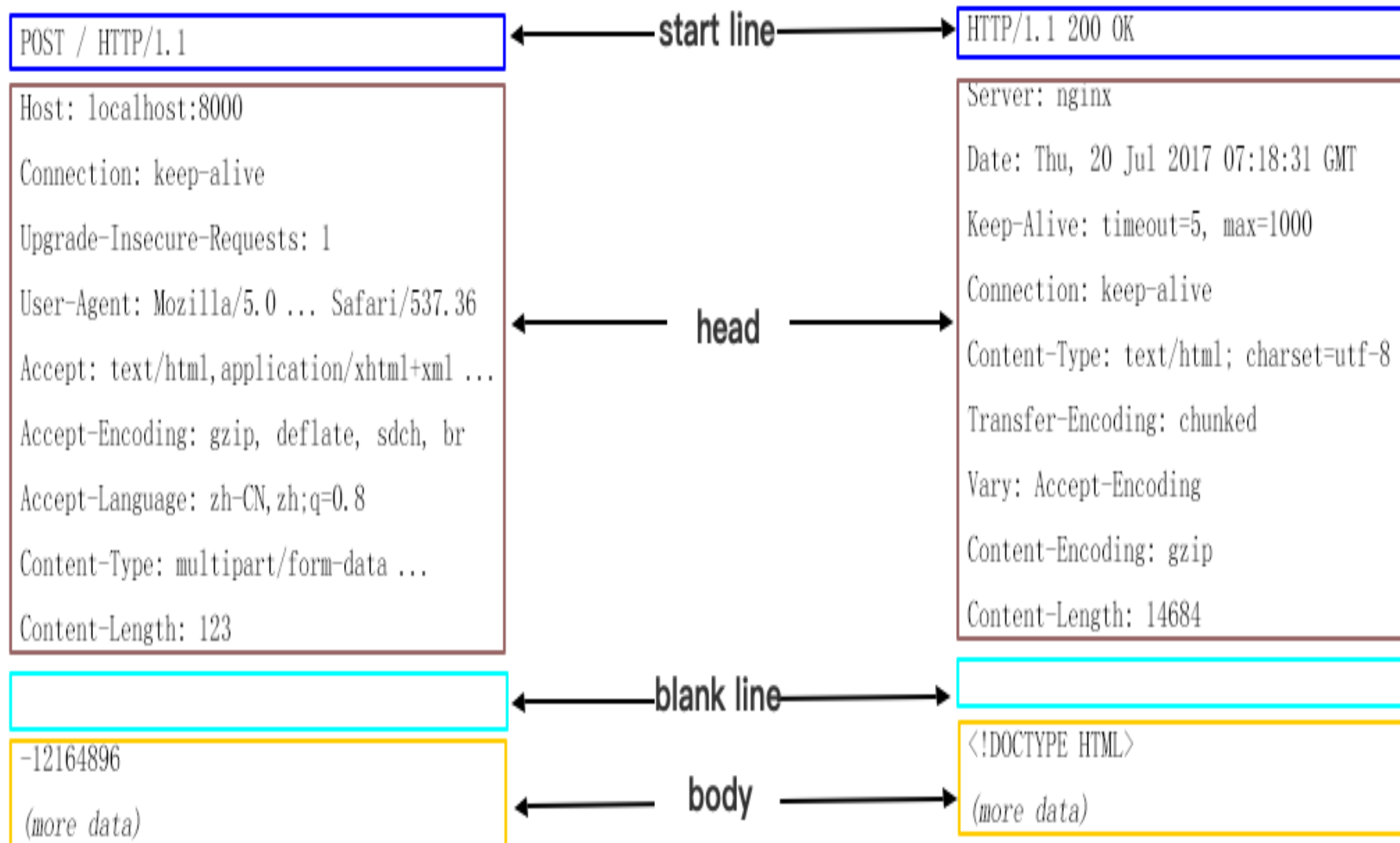


뉴스 검색 테스트

```
request = urllib.request.Request(url)
client_id = "OkToA9BapwFJVzCH2487"
client_secret = "bNNJpjwo51"
request.add_header("X-Naver-Client-Id", client_id)
request.add_header("X-Naver-Client-Secret", client_secret)
```

request

response



뉴스 검색 테스트

```
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)
```

HTTPError: HTTP Error 404: Not Found

HTTP STATUS CODES

2xx Success

200 Success / OK

3xx Redirection

301 Permanent Redirect

302 Temporary Redirect

304 Not Modified

4xx Client Error

401 Unauthorized Error

403 Forbidden

404 Not Found

405 Method Not Allowed

5xx Server Error

501 Not Implemented

502 Bad Gateway

503 Service Unavailable

504 Gateway Timeout

뉴스 검색 테스트

<https://namu.wiki/404.html>

A screenshot of the Chrome DevTools Network tab. The 'Headers' tab is selected. Under the 'General' section, the following information is displayed: Request URL: https://namu.wiki/405.html, Request Method: GET, Status Code: 404 (with a red error icon), Remote Address: 104.16.180.45:443, and Referrer Policy: strict-origin-when-cross-origin.

뉴스 검색 테스트

■ body.decode('utf-8')

character	encoding	bits
A	UTF-8	01000001
A	UTF-16	00000000 01000001
A	UTF-32	00000000 00000000 00000000 01000001
あ	UTF-8	11100011 10000001 10000010
あ	UTF-16	00110000 01000010
あ	UTF-32	00000000 00000000 00110000 01000010

■ notepad test

뉴스 검색 테스트

■ 정상 결과

```
{
  "lastBuildDate": "Tue, 13 Sep 2022 14:18:13 +0900",
  "total": 77345,
  "start": 1,
  "display": 10,
  "items": [
    {
      "title": "유명 영어사전서 'Korea' <b>검색</b>하니...어? 'Choson'으로 나오네",
      "originallink": "http://www.fnnews.com/news/202209131034017219",
      "link": "https://n.news.naver.com/mnews/article/014/0004897569?sid=104",
      "description": "뱅크는 '이번 영어 사전 조사를 통해 상당수의 영어 사전과 백과사전에서 한국사의 왜곡이 심각한 수준이며, 긍정적 기술과 부정적 기술이 더 많은 것을 볼 수 있었다'면서 '각 사전 출판사들이 사전에 <b>단어</b>를 등재할 때...",
      "pubDate": "Tue, 13 Sep 2022 13:31:00 +0900"
    },
    {
      "title": "“인간 중심의 AI 혁신, 알아야 할 ABC는...” 美 특허청 CIO",
      "originallink": "https://www.ciokorea.com/news/254302",
      "link": "https://www.ciokorea.com/news/254302",
      "description": "특허 <b>검색</b> 도구에 AI 구성 요소를 추가하고 있다. 이는 USPTO가 매년 평균적으로 접수 받는 60만 건 이상의 (특허) 신청서에 약 20페이지 분량의 텍스트 및 그림 또는 이를 설명하는 약 1만 <b>단어</b>가 포함돼 있기 때문에...",
      "pubDate": "Tue, 13 Sep 2022 12:38:00 +0900"
    }
  ]
}
```

01. 네이버 API를 이용한 크롤링

■ 네이버 뉴스 크롤링

1. 전체 작업 설계하기

작업 설계	사용할 코드
1. 검색어 지정하기	srcText = '월드컵'
2. 네이버 뉴스 검색하기	getNaverSearch()
2.1 url 구성하기	url = base + node + srcText
2.2 url 접속과 검색 요청하기	urllib.request.urlopen()
2.3 요청 결과를 응답 JSON으로 받기	json.load()
3. 응답 데이터를 정리하여 리스트에 저장하기	getPostData()
4. 리스트를 JSON 파일로 저장하기	json.dumps()

01. 네이버 API를 이용한 크롤링

■ 네이버 뉴스 크롤링

2. 프로그램 구성 설계하기

[CODE 0]

```
def main()
```

1. 검색어 지정

2. 네이버 뉴스 검색

3. 응답 데이터 정리 후
리스트에 저장

4. 리스트를 JSON 파일로 저장

[CODE 2]

getNaverSearch()

json.load(responseDecode)

[CODE 1]

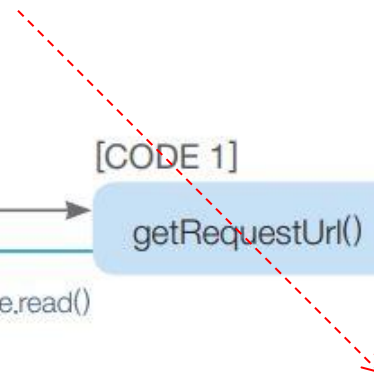
getRequestUrl()

Response.read()

[CODE 3]

getPostData()

jsonResult



뉴스 검색하기

■ 함수 및 파라미터 설정

```
def getNaverSearch(node, srcText, page_start, display) :  
    base = "https://openapi.naver.com/v1/search"  
    node = "/%s.json"% node  
    parameters = "?query=%s&start=%s&display=%s"%(urllib.parse.quote(srcText), page_start, display)  
    url = base + node + parameters
```

3. 요청 변수 (request parameter)

<https://developers.naver.com/docs/service/api/search/news/news.md#%EB%89%B4%EC%8A%A4>

요청 변수	타입	필수 여부	기본값	
query	string	Y	-	검색을 원하는 문자열로서 UTF-8로 인코딩한다.
display	integer	N	10(기본값), 100(최대)	검색 결과 출력 건수 지정
start	integer	N	1(기본값), 1000(최대)	검색 시작 위치로 최대 1000까지 가능
sort	string	N	sim, date(기본값)	정렬 옵션: sim (유사도순), date (날짜순)



뉴스 검색하기

■ json.loads

```
if(rescode==200):  
    response_body = response.read()  
    return json.loads(response_body.decode('utf-8'))  
else:  
    print("Error Code:" + rescode)  
    return None
```

■ main

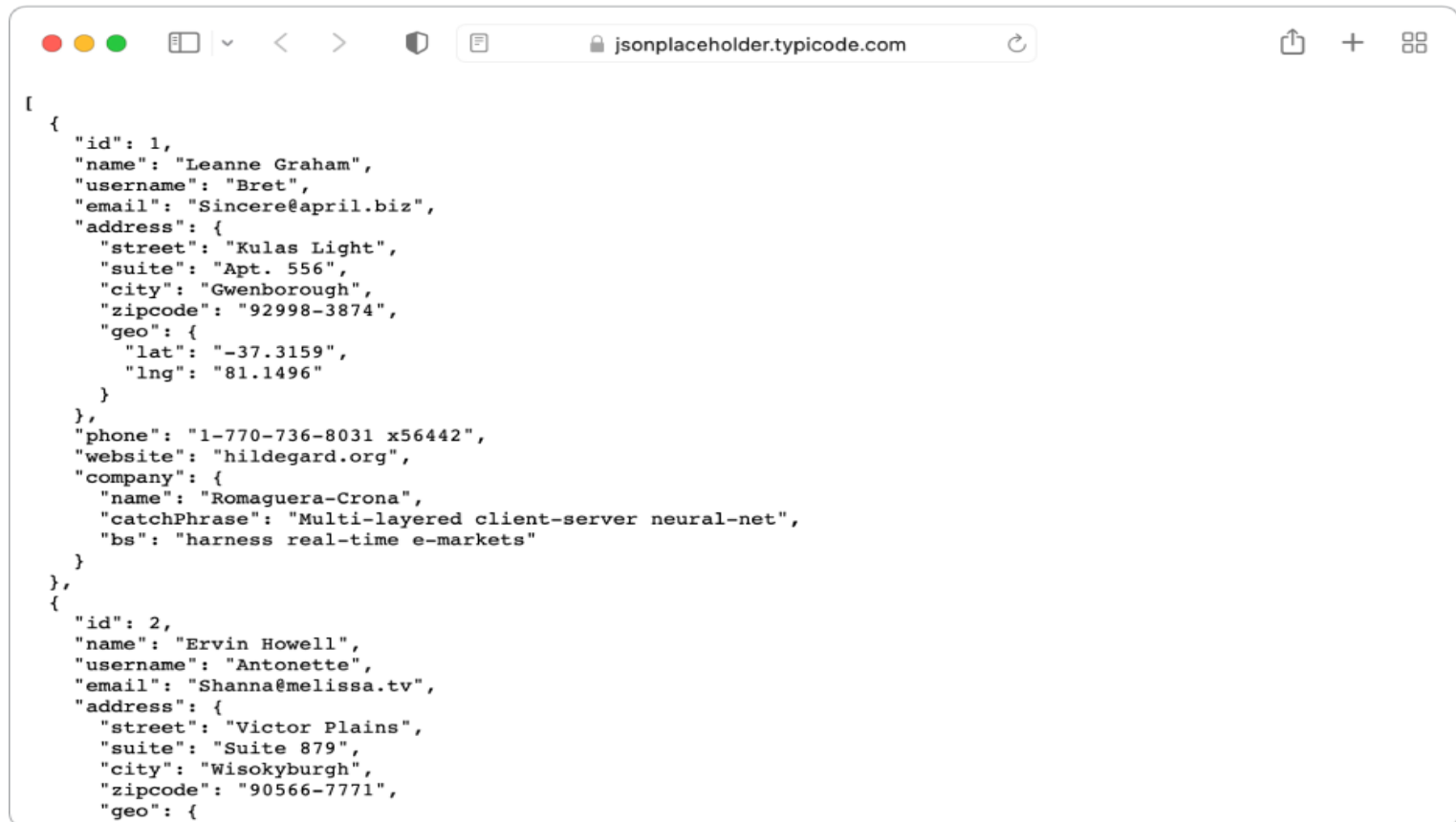
```
node = 'news'  
srcText = "고양이"  
cnt = 0  
jsonResult = []  
jsonResponse = getNaverSearch(node, srcText, 1, 10)  
print(jsonResponse)
```

01. 빅데이터의 이해

■ 빅데이터의 분류

- 반정형 데이터

- 고정된 필드에 저장되어 있지는 않지만 XML, HTML 등의 메타데이터와 스키마를 포함하는 것으로 파일 형태로 저장
- XML, JSON, HTML 등

A screenshot of a web browser window displaying JSON data from the website jsonplaceholder.typicode.com. The browser's address bar shows the URL. The main content area displays a JSON array with two objects. Each object represents a user with various attributes like id, name, username, email, address (including street, suite, city, zipcode, and geo coordinates), phone, website, company, and a business (bs) description.

```
[
  {
    "id": 1,
    "name": "Leanne Graham",
    "username": "Bret",
    "email": "Sincere@april.biz",
    "address": {
      "street": "Kulas Light",
      "suite": "Apt. 556",
      "city": "Gwenborough",
      "zipcode": "92998-3874",
      "geo": {
        "lat": "-37.3159",
        "lng": "81.1496"
      }
    },
    "phone": "1-770-736-8031 x56442",
    "website": "hildegard.org",
    "company": {
      "name": "Romaguera-Crona",
      "catchPhrase": "Multi-layered client-server neural-net",
      "bs": "harness real-time e-markets"
    }
  },
  {
    "id": 2,
    "name": "Ervin Howell",
    "username": "Antonette",
    "email": "Shanna@melissa.tv",
    "address": {
      "street": "Victor Plains",
      "suite": "Suite 879",
      "city": "Wisokyburgh",
      "zipcode": "90566-7771",
      "geo": {
        "lat": "-43.9469",
        "lng": "-120.3349"
      }
    },
    "phone": "010-691-9872",
    "website": "elton-also.com",
    "company": {
      "name": "Corkery-Powell",
      "catchPhrase": "Multi-tiered zero tolerance protocol",
      "bs": "revolutionary practice for distinctive e-markets"
    }
  }
]
```

```
In [20]: import os
import sys
import urllib.request
import urllib
import json
```

```
In [21]: def getNaverSearch(node, srcText, page_start, display) :
    base = "https://openapi.naver.com/v1/search"
    node = "/%s.json"% node
    parameters = "?query=%s&start=%s&display=%s"%(urllib.parse.quote(srcText), page_start, display)
    url = base + node + parameters

    client_id = "OkToA9BapwFJVzCH2487"
    client_secret = "bNNJpjwo51"
    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()
        return json.loads(response_body.decode('utf-8'))
    else:
        print("Error Code:" + rescode)
        return None
```

```
: node = 'news'
srcText = "고양이"
cnt = 0
jsonResult = []
jsonResponse = getNaverSearch(node, srcText, 1, 10)
total = jsonResponse['total']
print(total)
#print(jsonResult)
```


4. 출력 결과

필드	타입	설명
rss	-	디버그를 쉽게 하고 RSS 리더기만으로 이용할 수 있게 하기 위해 만든 RSS 포맷의 컨테이너이며 그 외의 특별한 의미는 없다.
channel	-	검색 결과를 포함하는 컨테이너이다. 이 안에 있는 title, link, description 등의 항목은 참고용으로 무시해도 무방하다.
lastBuildDate	datetime	검색 결과를 생성한 시간이다.
total	integer	검색 결과 문서의 총 개수를 의미한다.
start	integer	검색 결과 문서 중, 문서의 시작점을 의미한다.
display	integer	검색된 검색 결과의 개수이다.
item/items	-	XML 포맷에서는 item 태그로, JSON 포맷에서는 items 속성으로 표현된다. 개별 검색 결과이며 title, originallink, link, description, pubDate를 포함한다.
title	string	개별 검색 결과이며, title, originallink, link, description, pubDate 를 포함한다.
originallink	string	검색 결과 문서의 제공 언론사 하이퍼텍스트 link를 나타낸다.
link	string	검색 결과 문서의 제공 네이버 하이퍼텍스트 link를 나타낸다.
description	string	검색 결과 문서의 내용을 요약한 패시지 정보이다. 문서 전체의 내용은 link를 따라가면 읽을 수 있다. 패시지에서 검색어와 일치하는 부분은 태그로 감싸져 있다.
pubDate	datetime	검색 결과 문서가 네이버에 제공된 시간이다.

뉴스 검색하기

```
import os
import sys
import urllib.request
import urllib
import json
import datetime
import pprint
```

```
node = 'news'
srcText = "고양이"
cnt = 0
jsonResult = []
jsonResponse = getNaverSearch(node, srcText, 1, 10)
# total = jsonResponse['total']
# print(total)
# print(jsonResponse)
pprint.pprint(jsonResponse)
```

```
{'display': 10,
 'items': [{'description': '사진=박수홍 운영 유튜브 채널 &apos;검은<b>고양이</b> Dahong Blackcat '
'Dahong&apos; 캡처 화면 혐의를 받는 방송인 박수홍의 친형 박모씨가 법원 구속영장 '
'심사에 출석했다. 서울서부지검은 13일 오전 116억원에 달하는 박수홍의 출연료 '
'등을... ',
 'link': 'https://www.nbnv.co.kr/news/articleView.html?idxno=992543',
 'originallink': 'https://www.nbnv.co.kr/news/articleView.html?idxno=992543',
 'pubDate': 'Tue, 13 Sep 2022 14:30:00 +0900',
 'title': '박수홍 친형, 법원 출석...구속여부는 언제'},
 {'description': '<b>고양이</b>용품, 패션잡화, 삼베제품, 양평 친환경농산물까지 다양한 품목으로 교육 '
'중 사업자 등록을 내고 사업을 시작한 교육생들은 쇼핑물 상호와 로고, 상품 홍보 동영상까지 '
'직접 제작했다. 또한, 교육 과정 중... ',
 'link': 'https://www.wikitree.co.kr/articles/788708',
 'originallink': 'https://www.wikitree.co.kr/articles/788708',
 'pubDate': 'Tue, 13 Sep 2022 14:30:00 +0900'}
```

01. 네이버 API를 이용한 크롤링

■ 네이버 뉴스 크롤링

1. 전체 작업 설계하기

작업 설계	사용할 코드
1. 검색어 지정하기	srcText = '월드컵'
2. 네이버 뉴스 검색하기	getNaverSearch()
2.1 url 구성하기	url = base + node + srcText
2.2 url 접속과 검색 요청하기	urllib.request.urlopen()
2.3 요청 결과를 응답 JSON으로 받기	json.load()
3. 응답 데이터를 정리하여 리스트에 저장하기	getPostData()
4. 리스트를 JSON 파일로 저장하기	json.dumps()

응답 데이터 리스트 저장

```
def getPostData(post, jsonResult, cnt) :  
    title = post['title']  
    desc = post['description']  
    org_link = post['originallink']  
    link = post['link']
```

p.26 참조

```
{'display': 10,  
'items': [{'description': '사진=박수홍 운영 유튜브 채널 &apos;검은<b>고양이</b> 다홍 Blackcat '  
                'Dahong&apos; 캡처 횡령 혐의를 받는 방송인 박수홍의 친형 박모씨가 법원 구속영장 '  
                '심사에 출석했다. 서울서부지검은 13일 오전 116억원에 달하는 박수홍의 출연료 '  
                '등을... ',  
                'link': 'https://www.nbntv.co.kr/news/articleView.html?idxno=992543',  
                'originallink': 'https://www.nbntv.co.kr/news/articleView.html?idxno=992543',  
                'pubDate': 'Tue, 13 Sep 2022 14:30:00 +0900',  
                'title': '박수홍 친형, 법원 출석...구속여부는 언제'},  
            {'description': '<b>고양이</b>용품, 패션잡화, 잠베제품, 양병 친환경농산물까지 다양한 품목으로 교육 '  
                '중 사업자 등록을 내고 사업을 시작한 교육생들은 쇼핑물 상호와 로고, 상품 홍보 동영상까지 '  
                '직접 제작했다. 또한, 교육 과정 중... ',  
                'link': 'https://www.wikitree.co.kr/articles/788708',  
                'originallink': 'https://www.wikitree.co.kr/articles/788708',  
                'pubDate': 'Tue, 13 Sep 2022 14:30:00 +0900'}
```

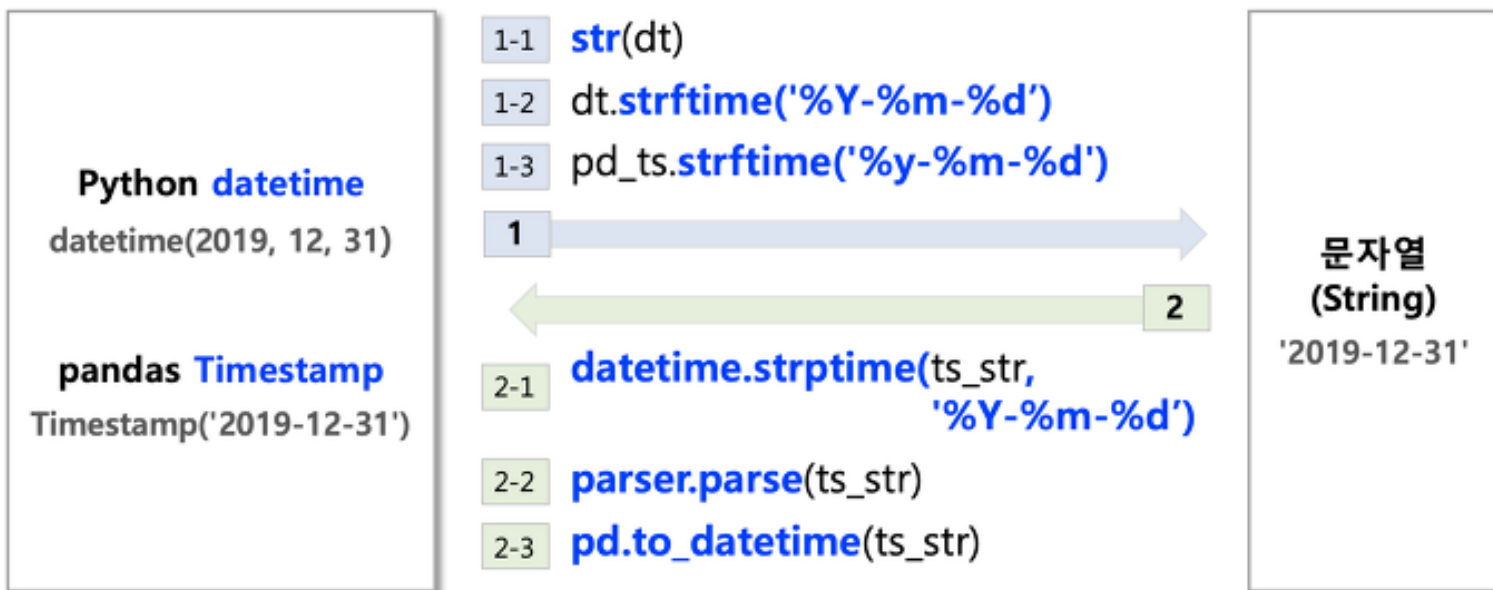
응답 데이터 리스트 저장

```
'pubDate': 'Tue, 13 Sep 2022 14:32:00 +0900',
```

```
pDate = datetime.datetime.strptime(post['pubDate'], '%a, %d %b %Y %H:%M:%S +0900')  
pDate = pDate.strftime('%Y-%m-%d %H:%M:%S')
```

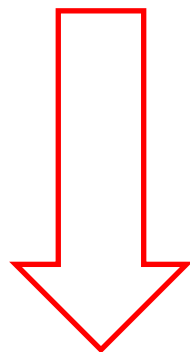


Python datetime, pandas Timestamp과 문자열(string) 간 변환하기
(Converting between Python datetime, pandas Timestamp objects and Strings)



응답 데이터 리스트 저장

```
jsonResult.append({'cnt':cnt, 'title' : title, 'description' : desc, 'org_link' : org_link, 'link':link, 'pDate': pDate})  
return
```



```
: def getPostData(post, jsonResult cnt) :  
    title = post['title']  
    desc = post['description']  
    org_link = post['originallink']  
    link = post['link']  
  
    pDate = datetime.datetime.strptime(post['pubDate'], '%a, %d %b %Y %H:%M:%S +0900')  
    pDate = pDate.strftime('%Y-%m-%d %H:%M:%S')  
  
    jsonResult.append({'cnt':cnt, 'title' : title, 'description' : desc, 'org_link' : org_link, 'link':link, 'pDate': pDate})  
    return
```

응답 데이터 리스트 저장

```
for post in jsonResponse['items'] :  
    cnt += 1  
    getPostData(post, jsonResult, cnt)  
  
pprint.pprint(jsonResult)
```

```
{'cnt': 1,  
 'description': '이어 "그렇지 않고 &apos;누가 <b>고양이</b> 목에 방울을 달지&apos; 정권 눈치만 '  
    '본다면, 돌아선 민심을 회복할 수 없음을 명심해야 한다&quot;고 경고했다. 박 원내대표는 또 '  
    '&quot;국민계선 공정과 도덕성을 상실한 윤석열 정권의 독주에도 불편함이... ',  
 'link': 'http://www.newsway.co.kr/news/view?tp=1&ud=2022091314275611855',  
 'org_link': 'http://www.newsway.co.kr/news/view?tp=1&ud=2022091314275611855',  
 'pDate': '2022-09-13 14:32:00',  
 'title': '민주, &apos;김건희 특검&apos; 강공 &quot;국정 정상화 출발점...반드시 관철시킬 것&quot;'},  
{'cnt': 2,  
 'description': '사실 마스크에서는 마치 지금 개나 <b>고양이</b>를 키우는 문화가 현대에 와서 갑자기 급성장한 문화라고 '  
    '생각을... 그런데 여전히 '남들이 키우니까, 나도 키워볼까?' 아니면 어떤 특정한 강아지나 '  
    '<b>고양이</b>의 외모가 너무... ',  
 'link': 'https://radio.ytn.co.kr/program/?f=2&id=85039&s_mcd=0438&s_hcd=01',  
 'org_link': 'https://radio.ytn.co.kr/program/?f=2&id=85039&s_mcd=0438&s_hcd=01',  
 'pDate': '2022-09-13 14:32:00',  
 'title': '[잠시만요] &quot;개 산책시킬 때 목줄 2m 넘으면 과태료, 짧게 잡아야 안전&quot;'},  
{'cnt': 3,  
 'description': '<b>고양이</b>용품, 패션잡화, 삼베제품, 양평 친환경농산물까지 다양한 품목으로 교육 중 사업자 등록을 '
```

응답 데이터 리스트 저장

```
while ((jsonResponse != None) and jsonResponse['display'] != 0) :  
    for post in jsonResponse['items'] :  
        cnt += 1  
        getPostData(post, jsonResult, cnt)  
  
start = jsonResponse['start'] + jsonResponse['display']  
jsonResponse = getNaverSearch(node,srcText, start, 100)
```

모든 페이지 수집 코드

응답 데이터 리스트 저장

File C:\Program Files\Python310\lib\urllib\request.py:643, in HTTPDefaultErrorHandler.http_error_default(self, req, fp, code, sg, hdrs)

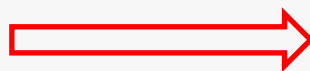
```
642 def http_error_default(self, req, fp, code, msg, hdrs):  
--> 643     raise HTTPError(req.full_url, code, msg, hdrs, fp)
```

HTTPError: HTTP Error 400: Bad Request

```
def getNaverSearch(node, srcText, page_start, display) :  
    base = "https://openapi.naver.com/v1/search"  
    node = "/%s.json"% node  
    parameters = "?query=%s&start=%s&display=%s"%(urllib.parse.quote(srcText), page_start, display)  
    url = base + node + parameters
```

```
    client_id = "OkToA9BapwFJVzCH2487"  
    client_secret = "bNNJpjwo51"  
    request = urllib.request.Request(url)  
    request.add_header("X-Naver-Client-Id", client_id)  
    request.add_header("X-Naver-Client-Secret", client_secret)
```

```
    try :  
        response = urllib.request.urlopen(request)  
        rescode = response.getcode()  
        if(rescode==200):  
            response_body = response.read()  
            return json.loads(response_body.decode('utf-8'))  
    except Exception as e :  
        print(e)  
        return None
```



HTTP Error 400: Bad Request

예외 처리

01. 네이버 API를 이용한 크롤링

■ 네이버 뉴스 크롤링

1. 전체 작업 설계하기

작업 설계	사용할 코드
1. 검색어 지정하기	srcText = '월드컵'
2. 네이버 뉴스 검색하기	getNaverSearch()
2.1 url 구성하기	url = base + node + srcText
2.2 url 접속과 검색 요청하기	urllib.request.urlopen()
2.3 요청 결과를 응답 JSON으로 받기	json.load()
3. 응답 데이터를 정리하여 리스트에 저장하기	getPostData()
4. 리스트를 JSON 파일로 저장하기	json.dumps()

json 파일 저장

다음 아래와 같은 코드를

```
file_data = open('file.txt')
print(file_data.readline(), end='')
file_data.close()
```


아래처럼 바꿀 수 있다.

```
with open('file.txt') as file_data:
    # 기본적으로 사용하는 함수를 with문 안에 사용하면 되며
    # with문을 나을 때 close를 자동으로 불러줍니다.
    print(file_data.readline(), end='')
```

json 파일 저장

```
: with open('%s_naver_%s.json' % (srcText, node), 'w', encoding='utf8') as outfile :  
    outfile.write("test")
```

- `r` : 읽기 모드, 파일을 읽을 때 사용합니다.
- `w` : 쓰기 모드, 파일에 쓸 때 사용하며 파일이 이미 동일한 이름으로 존재한다면 덮어씁니다.
- `a` : 추가 모드, 존재하는 파일에 추가할 때 사용하며 파일이 없다면 생성합니다.
- `r+`, `w+`, `a+` : 읽기모드 + 쓰기모드, `w+` 와 `a+` 의 차이는 위와 같습니다.
- `rb`, `wb`, `ab`, `rb+`, `wb+`, `ab+` : 각각의 모드들은 위와 동일하나 Binary 포맷으로 읽거나 쓰는걸 진행합니다.

 jupyter 고양이_naver_news.json ✓ a few seconds ago

File Edit View Language

1 test

json 파일 저장

```
with open('%s_naver_%s.json' % (srcText, node), 'w', encoding='utf8') as outfile :
    jsonFile = json.dumps(jsonResult)
    outfile.write(jsonFile)
```

{ "cnt": 1, "title": ""#uc9f1#ub3cc#ub85c #ud55c #ubc29#uc5d0"#u2026#ucc9c#uc5f0#uae30#ub150#ubb3c #uc218#ub9ac#ubd80#uc5c9#uc774 #ub0b4#ucad3#uc740 #uc720#ud29c#ubc84 '#ub17c#ub780'", "description": "#uae38#uace0#uc591#uc774#ub97c #ub3cc#ubcf4#ub294 #ucc44#ub110#uc744 #uc6b4#uc601#ud558#ub294 #ud55c #uc720#ud29c#ubc84#uac00 #ucc9c#uc5f0#uae30#ub150#ubb3c#uc778 #uc218#ub9ac#ubd80#uc5c9#uc774#uc5d0#uac8c #ub3cc#uc744 #ub358#uc84c#ub2e4#uace0 #ubc1c#uc5b8#ud574... #uc9c0#ub09c 8#uc77c #uc720#ud29c#ubc84 A#uc528#ub294 #ubc29#uc1a1 #uc911 #uc0c8#ub07c #uace0#uc591#uc774#ub97c #ub178#ub9ac#ub294 #uc218#ub9ac#ubd80#uc5c9#uc774#ub97c #ucad3#uc544#ub0b4#ub294 #uc601#uc0c1#uc744 #uc62c#ub838#ub2e4. #uc601#uc0c1 #uc18d A#uc528#ub294 #ud3c9#uc18c...", "org_link": "https://www.hankyung.com/society/article/2022091363777", "link": "https://n.news.naver.com/mnews/article/015/0004748767?sid=102", "pDate": "2022-09-13 14:57:00"}, { "cnt": 2, "title": ""#ub2f9#uc2e0#uc740 #ub098#uc758 #uc9d1#uc0ac#uac00 #ub420 #uc790#uaca9#uc774 #uc788#uc2b5#ub2c8#uae4c?" #ubc18#ub824#uc778#ub2a5#ub825#uc2dc#ud5d8 '#ub208#uae38'", "description": "#uc62c#ud574 4#ud68c#uc9f8#ub97c #ub9de#uc740 #uc774#ubc88 #uc2dc#ud5d8#uc5d0#uc11c #uc11c#uc6b8#uc2dc#ub294 #uc751#uc2dc#uc778#uc6d0#uc744 6000#uba85(#uac15#uc544#uc9c0 #ubd80#ubb38 3500#uba85, #uace0#uc591#uc774 #ubd80#ubb38 2500#uba85)#uc73c#ub85c... #uc2dc#ud5d8#uc740 #uace0#uc591#uc774 #ubd80#ubb38#uc744 #uc2e0#uc124#ud574 #ube44#ub300#uba74 #ud615#uc2dd#uc73c#ub85c 1004#uba85#uc774 #uc751#uc2dc, #uc81c3#ud68c #uc5ed#uc2dc #ube44#ub300#uba74#uc73c#ub85c 2693#uba85#uc774 #uc9c0#uc5c9#uc774#ub294 #uc6b4#uc601#ud558#ub294 #ud55c #uc720#ud29c#ubc84#uac00 #ucc9c#uc5f0#uae30#ub150#ubb3c#uc778 #uc218#ub9ac#ubd80#uc5c9#uc774#uc5d0#uac8c #ub3cc#uc744 #ub358#uc84c#ub2e4#uace0 #ubc1c#uc5b8#ud574... #uc9c0#ub09c 8#uc77c #uc720#ud29c#ubc84 A#uc528#ub294 #ubc29#uc1a1 #uc911 #uc0c8#ub07c #uace0#uc591#uc774#ub97c #ub178#ub9ac#ub294 #uc218#ub9ac#ubd80#uc5c9#uc774#ub97c #ucad3#uc544#ub0b4#ub294 #uc601#uc0c1#uc744 #uc62c#ub838#ub2e4. #uc601#uc0c1 #uc18d A#uc528#ub294 #ud3c9#uc18c...", "org_link": "https://www.hankyung.com/society/article/view.html?idx=107795", "link": "https://n.news.naver.com/mnews/article/015/0004748767?sid=102", "pDate": "2022-09-13 14:57:00"}]

json 파일 저장

■ ensure_ascii = False

```
1 [{"cnt": 1, "title": "&quot;짱돌로 한 방에&quot;...천연기념물 수리부엉이 내쫓은 유튜버 &apos;논란&apos;", "description": "길<b>고양이</b>를 돌보는 채널을 운영하는 한 유튜버가 천연기념물인 수리부엉이에게 돌을 던졌다고 발언해... 지난 8일 유튜버 A씨는 방송 중 새끼 <b>고양이</b>를 노리는 수리부엉이를 쫓아내는 영상을 올렸다. 영상 속 A씨는 평소... ", "org_link": "https://www.hankyung.com/society/article/2022091363777", "link": "https://n.news.naver.com/mnews/article/015/0004748767?sid=102", "pDate": "2022-09-13 14:57:00"}, {"cnt": 2, "title": "&quot;당신은 나의 집사가 될 자격이 있습니까?&quot; 반려인능력시험 &apos;눈길&apos;", "description": "올해 4회째를 맞은 이번 시험에서 서울시는 응시인원을 6000명(강아지 부문 3500명, <b>고양이</b> 부문 2500명)으로... 시험은 <b>고양이</b> 부문을 신설해 비대면 형식으로 1004명이 응시, 제3회 역시 비대면으로 2693명이 참여하는 등 큰... ", "org_link": "https://www.ibabynews.com/news/articleView.html?idxno=107736", "link": "https://www.ibabynews.com/news/articleView.html?idxno=107736", "pDate": "2022-09-13 14:52:00"}, {"cnt": 3, "title": "&quot;게임이론 활용한 협상가 되기&quot; 안준성의 신간도서 『나는 <b>고양이</b>와도 협상한...&quot;", "description": "『나는 <b>고양이</b>와도 협상한다: 게임이론을 활용한 성공적인 협상가 되기(도서출판 안다, 2022.08.12.)』가... 크게는 국가 간의 자유무역협정 협상으로부터 작게는 집에서 키우는 브리티시 쇼트헤어 <b>고양이</b> &apos;지니&apos;와의 쉼... ", "org_link": "http://www.lecturernews.com/news/articleView.html?idxno=106890", "link": "http://www.lecturernews.com/news/articleView.html?idxno=106890", "pDate": "2022-09-13 14:52:00"}, {"cnt": 4, "title": "광명시 개발지역 동물 돌봄 센터 &apos;길동무&apos; 13일 개소", "description": "길<b>고양이</b> 친구 대표, [사진=광명시] 2022.09.13 1141world@newspim.com 광명시는 구도심 지역의 주거 환경을 개선하기 위한 재건축재개발이 한창이다. 박승원 광명시장은 올해 초 길<b>고양이</b> 보호단체를 방문한 자리에서 도심... ", "org_link": "http://www.newspim.com/news/view/20220913000636", "link": "http://www.newspim.com/news/view/20220913000636", "pDate": "2022-09-13 14:48:00"}, {"cnt": 5, "title": "&quot;길냥이와 공생하는 개발지로&quot; 광명시, 첫 데이프", "description": "박승원 광명시장과 안성환 광명시의회 의장, 유종상, 김용성 도의원, 이흥덕 시의원, 광명길<b>고양이</b>친구 회원... 길<b>고양이</b>와 함께 공생하는 도시로 첫발을 내디뎠다. 광명시는 13일 박승원 광명시장, 안성환 광명시의회 의장... ", "org_link": "http://www.kyeongin.com/main/view.php?key=20220913010001741", "link": "http://www.kyeongin.com/main/view.php?key=20220913010001741", "pDate": "2022-09-13 14:48:00"}, {"cnt": 6, "title": "불안을 담은 밤 풍경...독일 중견 화가 아민 보엠 개인전", "description": "그의 회화에는 전구, 전조등, 네온사인 등 인공적 불빛과 야행성인 <b>고양이</b>, 빛을 가리는 커튼과 같은 밤의 모티브가 등장하며 대중문화와 문학 등도 담아낸다. 그러나 그는 &quot;그림의 주제보다 그림 자체가 더욱 중요하다... ", "org_link": "https://www.yon.co.kr/view/1&KRD202209130936000005?input=1195m", "link": "https://www.yon.co.kr/view/1&KRD202209130936000005?input=1195m", "pDate": "2022-09-13 14:48:00"}]
```

json 파일 저장

■ indent = 4

```
1  [
2    {
3      "cnt": 1,
4      "title": "&quot;짱돌로 한 방에&quot;...천연기념물 수리부엉이 내쫓은 유튜버 &apos;논란&apos;",
5      "description": "길<b>고양이</b>를 돌보는 채널을 운영하는 한 유튜버가 천연기념물인 수리부엉이에게 돌을 던졌다고 발언해... 지난 8일 유튜버 A씨는 방
송 중 새끼 <b>고양이</b>를 노리는 수리부엉이를 쫓아내는 영상을 올렸다. 영상 속 A씨는 평소... ",
6      "org_link": "https://www.hankyung.com/society/article/2022091363777",
7      "link": "https://n.news.naver.com/mnews/article/015/0004748767?sid=102",
8      "pDate": "2022-09-13 14:57:00"
9    },
10   {
11     "cnt": 2,
12     "title": "&quot;당신은 나의 집사가 될 자격이 있습니까?&quot; 반려인능력시험 &apos;눈길&apos;",
13     "description": "올해 4회째를 맞은 이번 시험에서 서울시는 응시인원을 6000명(강아지 부문 3500명, <b>고양이</b> 부문 2500명)으로... 시험은 <b>고양이
</b> 부문을 신설해 비대면 형식으로 1004명이 응시, 제3회 역시 비대면으로 2693명이 참여하는 등 큰... ",
14     "org_link": "https://www.ibabynews.com/news/articleView.html?idxno=107736",
15     "link": "https://www.ibabynews.com/news/articleView.html?idxno=107736",
16     "pDate": "2022-09-13 14:52:00"
17   },
18  ]
```

json 파일 저장

```
with open('%s_naver_%s.json' % (srcText, node), 'w', encoding='utf8') as outfile :  
    jsonFile = json.dumps(jsonResult, ensure_ascii = False, indent=4)  
    outfile.write(jsonFile)  
    print("가져온 데이터 : %d 건" % (cnt))  
    print('%s_naver_%s.json SAVED' % (srcText, node))
```

가져온 데이터 : 1000 건
고양이_naver_news.json SAVED

최종 코드

```
In [18]: import os
import sys
import urllib.request
import urllib
import json
import datetime
import pprint
```

```
In [19]: def getNaverSearch(node, srcText, page_start, display) :
    base = "https://openapi.naver.com/v1/search"
    node = "/%s.json"% node
    parameters = "?query=%s&start=%s&display=%s"%(urllib.parse.quote(srcText), page_start, display)
    url = base + node + parameters

    client_id = "OkToA9BapwFJVzCH2487"
    client_secret = "bNNJpjwo5l"
    request = urllib.request.Request(url)
    request.add_header("X-Naver-Client-Id", client_id)
    request.add_header("X-Naver-Client-Secret", client_secret)

    try :
        response = urllib.request.urlopen(request)
        rescode = response.getcode()
        if(rescode==200):
            response_body = response.read()
            return json.loads(response_body.decode('utf-8'))
    except Exception as e :
        print(e)
        return None
```

```
In [20]: def getPostData(post, jsonResult, cnt) :
    title = post['title']
    desc = post['description']
    org_link = post['originallink']
    link = post['link']

    pDate = datetime.datetime.strptime(post['pubDate'], '%a, %d %b %Y %H:%M:%S +0900')
    pDate = pDate.strftime('%Y-%m-%d %H:%M:%S')

    jsonResult.append({'cnt':cnt, 'title' : title, 'description' : desc, 'org_link' : org_link, 'link':link, 'pDate': pDate})
    return
```

최종 코드

```
In [21]: node = 'news'
srcText = "고양이"
cnt = 0
jsonResult = []
jsonResponse = getNaverSearch(node, srcText, 1, 100)
total = jsonResponse['total']
# print(total)
#print(jsonResponse)
pprint.pprint(jsonResponse)
```

```
In [22]: while ((jsonResponse != None) and jsonResponse['display'] != 0) :
    for post in jsonResponse['items'] :
        cnt += 1
        getPostData(post, jsonResult, cnt)

    start = jsonResponse['start'] + jsonResponse['display']
    jsonResponse = getNaverSearch(node, srcText, start, 100)

# pprint.pprint(jsonResult)
```

최종 코드

```
In [27]: with open('%s_naver_%s.json' % (srcText, node), 'w', encoding='utf8') as outfile :  
        jsonFile = json.dumps(jsonResult, ensure_ascii = False, indent=4)  
        outfile.write(jsonFile)  
        print("가져온 데이터 : %d 건" % (cnt))  
        print('%s_naver_%s.json SAVED' % (srcText, node))
```

가져온 데이터 : 1000 건
고양이_naver_news.json SAVED

