In [ ]:
```
Crawler(HyperLink 따라다니면서 수집) + DB
0. URL pool <= seed(구글 검색결과)
===> DFS(Stack, LIFO), BFS(Queue, FIFO)
===> Focused(전략:depth,domain제한,id/class제한)
1. robots.txt 파싱
   user-agent\nallow:?\ndisallow:?
   ---------- tuple => DB(범용관계형X)
2. status_code, content-type = text/html or xml
===> Scraping(별도)
3. DOM <= bs4, find*, select(css selector)
4. HTML Tags[attribute ㄷ href, src, (action-Form, method)]
5. Absolute Path(URL/I)
6. #, about:, mailto:, data:, javascript, ... => 제거
7. 갔다왔는지, 갈예정인지, 둘다 없으면 URL pool 추가
8. 0부터시작(URL pool에 더이상 없을때까지)
```

In [6]:
```
from requests import request
from requests.compat import *
from bs4 import BeautifulSoup

headers = {
    'user-agent':'Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7) AppleWebKit/537.36 (KH
}
```

In [2]:
```
import re
from requests.exceptions import HTTPError

ROBOT = dict()

def canFetch(url, agent='*', path='/'):
    k = urlparse(url).netloc

    if k not in ROBOT:
        robot = urljoin(url, '/robots.txt')
        resp = request('GET', robot)

        if resp.status_code == 200:
            ua = '*'
            ROBOT[k] = list()

            for l in resp.text.lower().splitlines():
                a = re.search(r'user-agent:(.+)', l, re.IGNORECASE)

                if a:
                    ua = a.group(1).strip()
                else:
                    p = re.search(
                        r'((?:dis)?allow):(.+)', l, re.IGNORECASE)

                    if p:
                        allow = False if p.group(1) == 'disallow' else True
                        ROBOT[k].append((ua, p.group(2).strip(), allow))

    if k in ROBOT:
        candidates = list(filter(lambda l:l[0] == agent and \
                     re.match(l[1].replace('*', '[^/]+?'), path),
                                 ROBOT[k]))
```

```
        candidates = sorted(candidates, key=lambda c:len(c[1]),
                            reverse=True)

        if len(candidates) > 0:
            return candidates[0][-1]
        return True

    return True
```

In [29]:

```
urls = [{'url':'https://search.naver.com/search.naver?where=nexearch&query=%EB%89%B4%EC%
         'depth':0}]
visited = []
focused1 = 3
focused2 = ['blog.naver.com']

while urls:
    url = urls.pop(0)

    if url['depth'] > focused1:
        visited.append(url)
        continue

    try:
        print(urlparse(url['url']).path, \
              canFetch(url['url'], path=urlparse(url['url']).path))
    except:
        print(url['url'])

    resp = request('GET', url['url'], headers=headers)

    if resp.status_code == 200:
        visited.append(url)

        if re.search('xml|html', resp.headers['content-type']):
            dom = BeautifulSoup(resp.text, 'lxml')

            a = list()
            for node in dom.select('[href], [src]'):
                temp = node.attrs['href'
                                  if node.has_attr('href')
                                  else 'src']

                if not re.match(r'#|javascript|data:|about:', temp)\
                   and urlparse(temp).netloc in focused2:
                    newurl = urljoin(url['url'], temp)
                    a.append({'url':newurl,'depth':url['depth']+1})

            urls += list(filter(
                lambda link:link['url'] not in [v['url'] for v in visited] \
                and link['url'] not in [u['url'] for u in urls], a))
    else:
        if resp.status_code >= 500:
            urls.append(url)
        else:
            visited.append(url)
```

```
/search.naver False
/lo_meo True
/lo_meo True
/lo_meo/223146020328 True
```

/lo_meo/223146020328 True
/lo_meo/223146020328 True
/lo_meo/223146020328 True
/lo_meo/223146020328 True
/lo_meo/223146020328 True
/lo_meo/223146020328 True
/lo_meo/223146020328 True
/lo_meo/223146020328 True
/lo_meo/223146020328 True
/lo_meo/223146020328 True
/lo_meo/223146020328 True
/bbq0928 True
/bbq0928 True
/bbq0928/223144819220 True
/bbq0928/223144819220 True
/bbq0928/223144819220 True
/hello_jooya00 True
/hello_jooya00 True
/hello_jooya00/223006254339 True
/hello_jooya00/223006254339 True
/hello_jooya00/223006254339 True
/hello_jooya00/223006254339 True
/hello_jooya00/223006254339 True
/hello_jooya00/223006254339 True
/hello_jooya00/223006254339 True
/hello_jooya00/223006254339 True
/hello_jooya00/223006254339 True
/hello_jooya00/223006254339 True
/hello_jooya00/223006254339 True
/hello_jooya00/223006254339 True
/secretinto True
/secretinto True
/secretinto/222867286470 True
/secretinto/222867286470 True
/secretinto/222867286470 True
/thrue111 True
/thrue111 True
/thrue111/222933256319 True
/thrue111/222933256319 True
/thrue111/222933256319 True
/hello_jooya00 True
/hello_jooya00 True
/hello_jooya00/223135372609 True
/hello_jooya00/223135372609 True
/hello_jooya00/223135372609 True
/hello_jooya00/223135372609 True
/hello_jooya00/223135372609 True
/hello_jooya00/223135372609 True
/hello_jooya00/223135372609 True
/hello_jooya00/223135372609 True
/hello_jooya00/223135372609 True
/hello_jooya00/223135372609 True
/hello_jooya00/223135372609 True
/hello_jooya00/223135372609 True
/icecreamyam222 True
/icecreamyam222 True
/icecreamyam222/223101858534 True
/icecreamyam222/223101858534 True
/icecreamyam222/223101858534 True

```
/yelim1014 True
/yelim1014 True
/yelim1014/223018294878 True
/yelim1014/223018294878 True
/yelim1014/223018294878 True
/naver_search/221086300708 True
/NBlogWlwLayout.naver True
/NBlogWlwLayout.naver True
/opt/homebrew/anaconda3/lib/python3.9/site-packages/bs4/builder/__init__.py:545: XMLPars
edAsHTMLWarning: It looks like you're parsing an XML document using an HTML parser. If t
his really is an HTML document (maybe it's XHTML?), you can ignore or filter this warnin
g. If it's XML, you should know that using an XML parser will be more reliable. To parse
this document as XML, make sure you have the lxml package installed, and pass the keywor
d argument `features="xml"` into the BeautifulSoup constructor.
  warnings.warn(
/NBlogWlwLayout.naver True
/prologue/PrologueList.naver True
/NBlogWlwLayout.naver True
/NBlogWlwLayout.naver True
/NBlogWlwLayout.naver True
/NBlogWlwLayout.naver True
/prologue/PrologueList.naver True
/NBlogWlwLayout.naver True
/MyBlog.naver True
/prologue/PrologueList.naver True
/prologue/PrologueList.naver True
/PostList.naver True
/PostList.naver True
/PostList.naver True
/PostList.naver True
/PostList.naver True
/mapview/MapView.naver True
/guestbook/GuestBookList.naver False
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
```

```
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/prologue/PrologueList.naver True
/PostList.naver True
/PostList.naver True
/PostList.naver True
/PostList.naver True
/PostList.naver True
/guestbook/GuestBookList.naver False
/prologue/PrologueList.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
```

```
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
/PostView.naver True
```

In [39]:
```
list(map(lambda url:(urlparse(url['url']).path, url['depth']),
         visited))
```

Out[39]:
```
[('/search.naver', 0),
```

('/lo_meo', 1),
('/lo_meo', 1),
('/lo_meo/223146020328', 1),
('/lo_meo/223146020328', 1),
('/lo_meo/223146020328', 1),
('/lo_meo/223146020328', 1),
('/lo_meo/223146020328', 1),
('/lo_meo/223146020328', 1),
('/lo_meo/223146020328', 1),
('/lo_meo/223146020328', 1),
('/lo_meo/223146020328', 1),
('/lo_meo/223146020328', 1),
('/lo_meo/223146020328', 1),
('/lo_meo/223146020328', 1),
('/bbq0928', 1),
('/bbq0928', 1),
('/bbq0928/223144819220', 1),
('/bbq0928/223144819220', 1),
('/bbq0928/223144819220', 1),
('/hello_jooya00', 1),
('/hello_jooya00', 1),
('/hello_jooya00/223006254339', 1),
('/hello_jooya00/223006254339', 1),
('/hello_jooya00/223006254339', 1),
('/hello_jooya00/223006254339', 1),
('/hello_jooya00/223006254339', 1),
('/hello_jooya00/223006254339', 1),
('/hello_jooya00/223006254339', 1),
('/hello_jooya00/223006254339', 1),
('/hello_jooya00/223006254339', 1),
('/hello_jooya00/223006254339', 1),
('/hello_jooya00/223006254339', 1),
('/secretinto', 1),
('/secretinto', 1),
('/secretinto/222867286470', 1),
('/secretinto/222867286470', 1),
('/secretinto/222867286470', 1),
('/thrue111', 1),
('/thrue111', 1),
('/thrue111/222933256319', 1),
('/thrue111/222933256319', 1),
('/thrue111/222933256319', 1),
('/hello_jooya00', 1),
('/hello_jooya00', 1),
('/hello_jooya00/223135372609', 1),
('/hello_jooya00/223135372609', 1),
('/hello_jooya00/223135372609', 1),
('/hello_jooya00/223135372609', 1),
('/hello_jooya00/223135372609', 1),
('/hello_jooya00/223135372609', 1),
('/hello_jooya00/223135372609', 1),
('/hello_jooya00/223135372609', 1),
('/hello_jooya00/223135372609', 1),
('/hello_jooya00/223135372609', 1),
('/hello_jooya00/223135372609', 1),
('/hello_jooya00/223135372609', 1),
('/icecreamyam222', 1),
('/icecreamyam222', 1),

```
('/icecreamyam222/223101858534', 1),
('/icecreamyam222/223101858534', 1),
('/icecreamyam222/223101858534', 1),
('/yelim1014', 1),
('/yelim1014', 1),
('/yelim1014/223018294878', 1),
('/yelim1014/223018294878', 1),
('/yelim1014/223018294878', 1),
('/naver_search/221086300708', 1),
('/NBlogWlwLayout.naver', 2),
('/NBlogWlwLayout.naver', 2),
('/NBlogWlwLayout.naver', 2),
('/prologue/PrologueList.naver', 2),
('/NBlogWlwLayout.naver', 2),
('/NBlogWlwLayout.naver', 2),
('/NBlogWlwLayout.naver', 2),
('/NBlogWlwLayout.naver', 2),
('/prologue/PrologueList.naver', 2),
('/NBlogWlwLayout.naver', 2),
('/MyBlog.naver', 3),
('/prologue/PrologueList.naver', 3),
('/prologue/PrologueList.naver', 3),
('/PostList.naver', 3),
('/PostList.naver', 3),
('/PostList.naver', 3),
('/PostList.naver', 3),
('/PostList.naver', 3),
('/mapview/MapView.naver', 3),
('/guestbook/GuestBookList.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
```

```
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/prologue/PrologueList.naver', 3),
('/PostList.naver', 3),
('/PostList.naver', 3),
('/PostList.naver', 3),
('/PostList.naver', 3),
('/PostList.naver', 3),
('/guestbook/GuestBookList.naver', 3),
('/prologue/PrologueList.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
```

```
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('/PostView.naver', 3),
('', 4),
('/tae9853dnjs', 4),
('/cookiebear1', 4),
('/dufla92', 4),
('/sarangmom1004', 4),
('/cbj_moim', 4),
('/house_28', 4),
('/oooooni0', 4),
```

```
 ('/mr_vip', 4),
 ('/mr_vip', 4),
 ('/superman_cnm', 4),
 ('/profile/intro.naver', 4),
 ('/profile/history.naver', 4),
 ('/profile/comment.naver', 4),
 ('/prologue/PrologueList.naver', 4),
 ('/profile/intro.naver', 4),
 ('/profile/history.naver', 4),
 ('/profile/comment.naver', 4),
 ('/yelim1014/222811824806', 4),
 ('/yelim1014/222811824806', 4),
 ('/yelim1014/222802830696', 4),
 ('/yelim1014/222802830696', 4)]
```

In [40]:
```python
import sqlite3

con = sqlite3.connect('blog.db')
cur = con.cursor()
```

In [90]:
```python
#             A => TableA
#     1   2    3   4   5 ... => TableB
# TableA
# PK netloc
# 1  blog.naver.com

# TableB
# path, qs, FK(referer)
# A         1
# A         1
# B         1

# 1 => A (2) => 2/N
# 1 => B (1) => 1/N
# group by path, count() => path별로 들어온 링크의 갯수(inboud)
# sum(FK) => 1번 사이트에서 나온 모든 갯수(Outboud)
# 1 => sum => 전체 링크갯수/N

cur.executescript('''
    DROP TABLE IF EXISTS T_A;
    CREATE TABLE T_A(
        pk INTEGER PRIMARY KEY,
        netloc TEXT NOT NULL UNIQUE
    );

    DROP TABLE IF EXISTS T_B;
    CREATE TABLE T_B(
        pk INTEGER PRIMARY KEY,
        path TEXT,
        qs TEXT,
        visited TEXT DEFAULT 'N',
        regdate DATETIME,
        fk INTEGER NOT NULL
    );
''')
```

Out[90]:
```
<sqlite3.Cursor at 0x117909810>
```

In [43]:
```python
# 기존 테이블 컬럼 추가
```

```
cur.execute('''
    ALTER TABLE T_B
    ADD COLUMN visited TEXT DEFAULT 'N'
''')
con.commit()
```

In [91]:
```
url = 'https://search.naver.com/search.naver?where=nexearch&query=%EB%89%B4%EC%A7%84%EC%
cur.execute('''
    INSERT INTO T_A(netloc) VALUES(?)
''', [urlparse(url).netloc])
con.commit()
```

In [92]:
```
cur.execute('''
    INSERT INTO T_B(path, qs, regdate, fk)
    VALUES(?, ?, CURRENT_TIMESTAMP, (
        SELECT PK FROM T_A WHERE netloc=?
    ))
''', [urlparse(url).path, urlparse(url).query, urlparse(url).netloc])
con.commit()
```

In [71]:
```
cur.execute('''
    SELECT T_B.pk, T_A.netloc, T_B.path, T_B.qs FROM T_A
    INNER JOIN T_B
    ON T_B.fk = T_A.pk
    WHERE T_B.visited = 'N'
    ORDER BY T_B.regdate ASC
''')
comp = cur.fetchall()
urlunparse(('https', comp[0][1], comp[0][2], None,
            comp[0][3], None))
```

Out[71]:
```
'https://search.naver.com/search.naver?where=nexearch&query=%EB%89%B4%EC%A7%84%EC%8A%A4&
oquery=%EB%89%B4%EC%A7%84%EC%8A%A4'
```

In [93]:
```
focused = ['blog.naver.com']
limit = 20
while True:
    # 20개 넘어가면 종료
    cur.execute('SELECT COUNT(pk) FROM T_A')
    if cur.fetchone()[0] > 20:
        break

    # 더이상 N이 없으면 종료
    cur.execute('''
        SELECT T_B.pk, T_A.netloc, T_B.path, T_B.qs FROM T_A
        INNER JOIN T_B
        ON T_B.fk = T_A.pk
        WHERE T_B.visited = 'N'
        ORDER BY T_B.regdate ASC
        LIMIT 0,1
    ''')
    comp = cur.fetchone()

    if not comp:
        break

    pk = comp[0]

    url = urlunparse(('https', comp[1], comp[2], None, comp[3], None))
```

```
try:
    print(urlparse(url).path, \
          canFetch(url, path=urlparse(url).path))
except:
    print(url)

resp = request('GET', url, headers=headers)

if resp.status_code == 200:
    if re.search('xml|html', resp.headers['content-type']):
        dom = BeautifulSoup(resp.text, 'lxml')

        for node in dom.select('[href], [src]'):
            temp = node.attrs['href'
                              if node.has_attr('href')
                              else 'src']

            if not re.match(r'#|javascript|data:|about:', temp)\
               and urlparse(temp).netloc in focused2:
                nurl = urljoin(url, temp)
                comp = urlparse(nurl)
                s = comp.path.split('/')

                if len(s) > 1: #(netloc/id)
                    cur.execute('''
                        SELECT pk FROM T_A
                        WHERE netloc=?
                    ''', [comp.netloc+'/'+s[1]])

                    if not cur.fetchone():
                        cur.execute('''
                            INSERT INTO T_A(netloc) VALUES(?)
                        ''', [comp.netloc+'/'+s[1]])
                        con.commit()

                    cur.execute('''
                        INSERT INTO T_B(path, qs, regdate, fk)
                        VALUES(?, ?, CURRENT_TIMESTAMP, (
                            SELECT PK FROM T_A WHERE netloc=?
                        ))
                    ''', [comp.path, comp.query, comp.netloc+'/'+s[1]])
                    con.commit()

                    cur.execute('''
                        UPDATE T_B
                        SET visited='Y'
                        WHERE pk=?
                    ''', [pk])
                    con.commit()
                else:
                    cur.execute('''
                        SELECT pk FROM T_A
                        WHERE netloc=?
                    ''', [comp.netloc])

                    if not cur.fetchone():
                        cur.execute('''
                            INSERT INTO T_A(netloc) VALUES(?)
```

```
                        ''', [comp.netloc])
                        con.commit()

                    cur.execute('''
                        INSERT INTO T_B(path, qs, regdate, fk)
                        VALUES(?, ?, CURRENT_TIMESTAMP, (
                            SELECT PK FROM T_A WHERE netloc=?
                        ))
                    ''', [comp.path, comp.query, comp.netloc])
                    con.commit()

                    cur.execute('''
                        UPDATE T_B
                        SET visited='Y'
                        WHERE pk=?
                    ''', [pk])
                    con.commit()
    else:
        if resp.status_code >= 500:
            cur.execute('''
                UPDATE T_B
                SET regdate=CURRENT_TIMESTAMP
                WHERE pk=?
            ''', [pk])
            con.commit()
        else:
            cur.execute('''
                UPDATE T_B
                SET visited='Y'
                WHERE pk=?
            ''', [pk])
            con.commit()
```

```
/search.naver False
/hello_jooya00/hello_jooya00 True
/hello_jooya00/hello_jooya00 True
/hello_jooya00/hello_jooya00/223006254339 True
/hello_jooya00/hello_jooya00/223006254339 True
/hello_jooya00/hello_jooya00/223006254339 True
/hello_jooya00/hello_jooya00/223006254339 True
/hello_jooya00/hello_jooya00/223006254339 True
/hello_jooya00/hello_jooya00/223006254339 True
/hello_jooya00/hello_jooya00/223006254339 True
/hello_jooya00/hello_jooya00/223006254339 True
/hello_jooya00/hello_jooya00/223006254339 True
/hello_jooya00/hello_jooya00/223006254339 True
/hello_jooya00/hello_jooya00/223006254339 True
/hello_jooya00/hello_jooya00/223006254339 True
/secretinto/secretinto True
/secretinto/secretinto True
/secretinto/secretinto/222867286470 True
/secretinto/secretinto/222867286470 True
/secretinto/secretinto/222867286470 True
/thrue111/thrue111 True
/thrue111/thrue111 True
/thrue111/thrue111/222933256319 True
/thrue111/thrue111/222933256319 True
/thrue111/thrue111/222933256319 True
/lo_meo/lo_meo True
/lo_meo/lo_meo True
```

```
/lo_meo/lo_meo/223146020328 True
/lo_meo/lo_meo/223146020328 True
/lo_meo/lo_meo/223146020328 True
/lo_meo/lo_meo/223146020328 True
/lo_meo/lo_meo/223146020328 True
/lo_meo/lo_meo/223146020328 True
/lo_meo/lo_meo/223146020328 True
/lo_meo/lo_meo/223146020328 True
/lo_meo/lo_meo/223146020328 True
/lo_meo/lo_meo/223146020328 True
/lo_meo/lo_meo/223146020328 True
/lo_meo/lo_meo/223146020328 True
/bbq0928/bbq0928 True
/bbq0928/bbq0928 True
/bbq0928/bbq0928/223144819220 True
/bbq0928/bbq0928/223144819220 True
/bbq0928/bbq0928/223144819220 True
/hello_jooya00/hello_jooya00 True
/hello_jooya00/hello_jooya00 True
/hello_jooya00/hello_jooya00/223135372609 True
/hello_jooya00/hello_jooya00/223135372609 True
/hello_jooya00/hello_jooya00/223135372609 True
/hello_jooya00/hello_jooya00/223135372609 True
/hello_jooya00/hello_jooya00/223135372609 True
/hello_jooya00/hello_jooya00/223135372609 True
/hello_jooya00/hello_jooya00/223135372609 True
/hello_jooya00/hello_jooya00/223135372609 True
/hello_jooya00/hello_jooya00/223135372609 True
/hello_jooya00/hello_jooya00/223135372609 True
/hello_jooya00/hello_jooya00/223135372609 True
/icecreamyam222/icecreamyam222 True
/icecreamyam222/icecreamyam222 True
/icecreamyam222/icecreamyam222/223101858534 True
/icecreamyam222/icecreamyam222/223101858534 True
/icecreamyam222/icecreamyam222/223101858534 True
/yelim1014/yelim1014 True
/yelim1014/yelim1014 True
/yelim1014/yelim1014/223018294878 True
/yelim1014/yelim1014/223018294878 True
/yelim1014/yelim1014/223018294878 True
/naver_search/naver_search/221086300708 True
---------------------------------------------------------------------------
TypeError                                 Traceback (most recent call last)
Cell In [93], line 19
     10 cur.execute('''
     11     SELECT T_B.pk, T_A.netloc, T_B.path, T_B.qs FROM T_A
     12     INNER JOIN T_B
   (...)
     16     LIMIT 0,1
     17 ''')
     18 comp = cur.fetchone()
---> 19 pk = comp[0]
     21 if not comp:
     22     break

TypeError: 'NoneType' object is not subscriptable
```

In [94]:
```
cur.execute('''
```

```
        SELECT * FROM T_A
    ''')
cur.fetchall()
```

Out[94]:
```
[(1, 'search.naver.com'),
 (2, 'blog.naver.com/hello_jooya00'),
 (3, 'blog.naver.com/secretinto'),
 (4, 'blog.naver.com/thrue111'),
 (5, 'blog.naver.com/lo_meo'),
 (6, 'blog.naver.com/bbq0928'),
 (7, 'blog.naver.com/icecreamyam222'),
 (8, 'blog.naver.com/yelim1014'),
 (9, 'blog.naver.com/naver_search')]
```

In [95]:
```
cur.execute('''
        SELECT * FROM T_B
    ''')
cur.fetchall()
```

Out[95]:
```
[(1,
  '/search.naver',
  'where=nexearch&query=%EB%89%B4%EC%A7%84%EC%8A%A4&oquery=%EB%89%B4%EC%A7%84%EC%8A%A4',
  'Y',
  '2023-07-11 03:00:03',
  1),
 (2, '/hello_jooya00', '', 'Y', '2023-07-11 03:00:07', 2),
 (3, '/hello_jooya00', '', 'Y', '2023-07-11 03:00:07', 2),
 (4, '/hello_jooya00/223006254339', '', 'Y', '2023-07-11 03:00:07', 2),
 (5, '/hello_jooya00/223006254339', '', 'Y', '2023-07-11 03:00:07', 2),
 (6, '/hello_jooya00/223006254339', '', 'Y', '2023-07-11 03:00:07', 2),
 (7, '/hello_jooya00/223006254339', '', 'Y', '2023-07-11 03:00:07', 2),
 (8, '/hello_jooya00/223006254339', '', 'Y', '2023-07-11 03:00:07', 2),
 (9, '/hello_jooya00/223006254339', '', 'Y', '2023-07-11 03:00:07', 2),
 (10, '/hello_jooya00/223006254339', '', 'Y', '2023-07-11 03:00:07', 2),
 (11, '/hello_jooya00/223006254339', '', 'Y', '2023-07-11 03:00:07', 2),
 (12, '/hello_jooya00/223006254339', '', 'Y', '2023-07-11 03:00:07', 2),
 (13, '/hello_jooya00/223006254339', '', 'Y', '2023-07-11 03:00:07', 2),
 (14, '/hello_jooya00/223006254339', '', 'Y', '2023-07-11 03:00:07', 2),
 (15, '/hello_jooya00/223006254339', '', 'Y', '2023-07-11 03:00:07', 2),
 (16, '/secretinto', '', 'Y', '2023-07-11 03:00:07', 3),
 (17, '/secretinto', '', 'Y', '2023-07-11 03:00:07', 3),
 (18, '/secretinto/222867286470', '', 'Y', '2023-07-11 03:00:07', 3),
 (19, '/secretinto/222867286470', '', 'Y', '2023-07-11 03:00:07', 3),
 (20, '/secretinto/222867286470', '', 'Y', '2023-07-11 03:00:07', 3),
 (21, '/thrue111', '', 'Y', '2023-07-11 03:00:07', 4),
 (22, '/thrue111', '', 'Y', '2023-07-11 03:00:07', 4),
 (23, '/thrue111/222933256319', '', 'Y', '2023-07-11 03:00:07', 4),
 (24, '/thrue111/222933256319', '', 'Y', '2023-07-11 03:00:07', 4),
 (25, '/thrue111/222933256319', '', 'Y', '2023-07-11 03:00:07', 4),
 (26, '/lo_meo', '', 'Y', '2023-07-11 03:00:07', 5),
 (27, '/lo_meo', '', 'Y', '2023-07-11 03:00:07', 5),
 (28, '/lo_meo/223146020328', '', 'Y', '2023-07-11 03:00:07', 5),
 (29, '/lo_meo/223146020328', '', 'Y', '2023-07-11 03:00:07', 5),
 (30, '/lo_meo/223146020328', '', 'Y', '2023-07-11 03:00:07', 5),
 (31, '/lo_meo/223146020328', '', 'Y', '2023-07-11 03:00:07', 5),
 (32, '/lo_meo/223146020328', '', 'Y', '2023-07-11 03:00:07', 5),
 (33, '/lo_meo/223146020328', '', 'Y', '2023-07-11 03:00:07', 5),
 (34, '/lo_meo/223146020328', '', 'Y', '2023-07-11 03:00:07', 5),
 (35, '/lo_meo/223146020328', '', 'Y', '2023-07-11 03:00:07', 5),
 (36, '/lo_meo/223146020328', '', 'Y', '2023-07-11 03:00:07', 5),
```

```
(37, '/lo_meo/223146020328', '', 'Y', '2023-07-11 03:00:07', 5),
(38, '/lo_meo/223146020328', '', 'Y', '2023-07-11 03:00:07', 5),
(39, '/lo_meo/223146020328', '', 'Y', '2023-07-11 03:00:07', 5),
(40, '/bbq0928', '', 'Y', '2023-07-11 03:00:07', 6),
(41, '/bbq0928', '', 'Y', '2023-07-11 03:00:07', 6),
(42, '/bbq0928/223144819220', '', 'Y', '2023-07-11 03:00:07', 6),
(43, '/bbq0928/223144819220', '', 'Y', '2023-07-11 03:00:07', 6),
(44, '/bbq0928/223144819220', '', 'Y', '2023-07-11 03:00:07', 6),
(45, '/hello_jooya00', '', 'Y', '2023-07-11 03:00:07', 2),
(46, '/hello_jooya00', '', 'Y', '2023-07-11 03:00:07', 2),
(47, '/hello_jooya00/223135372609', '', 'Y', '2023-07-11 03:00:07', 2),
(48, '/hello_jooya00/223135372609', '', 'Y', '2023-07-11 03:00:07', 2),
(49, '/hello_jooya00/223135372609', '', 'Y', '2023-07-11 03:00:07', 2),
(50, '/hello_jooya00/223135372609', '', 'Y', '2023-07-11 03:00:07', 2),
(51, '/hello_jooya00/223135372609', '', 'Y', '2023-07-11 03:00:07', 2),
(52, '/hello_jooya00/223135372609', '', 'Y', '2023-07-11 03:00:07', 2),
(53, '/hello_jooya00/223135372609', '', 'Y', '2023-07-11 03:00:07', 2),
(54, '/hello_jooya00/223135372609', '', 'Y', '2023-07-11 03:00:07', 2),
(55, '/hello_jooya00/223135372609', '', 'Y', '2023-07-11 03:00:07', 2),
(56, '/hello_jooya00/223135372609', '', 'Y', '2023-07-11 03:00:07', 2),
(57, '/hello_jooya00/223135372609', '', 'Y', '2023-07-11 03:00:07', 2),
(58, '/hello_jooya00/223135372609', '', 'Y', '2023-07-11 03:00:07', 2),
(59, '/icecreamyam222', '', 'Y', '2023-07-11 03:00:07', 7),
(60, '/icecreamyam222', '', 'Y', '2023-07-11 03:00:07', 7),
(61, '/icecreamyam222/223101858534', '', 'Y', '2023-07-11 03:00:07', 7),
(62, '/icecreamyam222/223101858534', '', 'Y', '2023-07-11 03:00:07', 7),
(63, '/icecreamyam222/223101858534', '', 'Y', '2023-07-11 03:00:07', 7),
(64, '/yelim1014', '', 'Y', '2023-07-11 03:00:07', 8),
(65, '/yelim1014', '', 'Y', '2023-07-11 03:00:07', 8),
(66, '/yelim1014/223018294878', '', 'Y', '2023-07-11 03:00:07', 8),
(67, '/yelim1014/223018294878', '', 'Y', '2023-07-11 03:00:07', 8),
(68, '/yelim1014/223018294878', '', 'Y', '2023-07-11 03:00:07', 8),
(69, '/naver_search/221086300708', '', 'Y', '2023-07-11 03:00:07', 9)]
```

In [84]:
```
cur.execute('''
        SELECT T_B.pk, T_A.netloc, T_B.path, T_B.qs FROM T_A
        INNER JOIN T_B
        ON T_B.fk = T_A.pk
        ORDER BY T_B.regdate ASC
        LIMIT 0,1
    ''')
cur.fetchall()
```

Out[84]:
```
[(1,
  'search.naver.com',
  '/search.naver',
  'where=nexearch&query=%EB%89%B4%EC%A7%84%EC%8A%A4&oquery=%EB%89%B4%EC%A7%84%EC%8A%A
4')]
```

In [96]:
```
cur.execute('''
    SELECT fk, COUNT(fk) FROM T_B
    GROUP BY fk
''')
cur.fetchall()
```

Out[96]:
```
[(1, 1), (2, 28), (3, 5), (4, 5), (5, 14), (6, 5), (7, 5), (8, 5), (9, 1)]
```

In [110]…
```
import numpy as np
```

In [126]…
```
cur.execute('''
    SELECT COUNT(pk) FROM T_A
''')
dim = cur.fetchone()[0]
PR = np.zeros((dim, dim))
```

In [127]…
```
cur.execute('''
    SELECT T_A.netloc, T_A.pk, COUNT(T_B.pk) FROM T_A
    LEFT JOIN T_B
    ON T_B.fk = T_A.pk
    GROUP BY T_B.fk
    ORDER BY T_A.pk
''')
for r in cur.fetchall():
    PR[r[1]-1,:] = r[2]
```

In [128]…
```
SITE = np.zeros((dim,1))
SITE[:,:] = 1/dim
```

In [129]…
```
for _ in range(10):
    SITE = .15 * .85*PR.dot(SITE)
```

In [130]…
```
SITE
```

Out[130]:
```
array([[4.02479407e+07],
       [1.12694234e+09],
       [2.01239703e+08],
       [2.01239703e+08],
       [5.63471170e+08],
       [2.01239703e+08],
       [2.01239703e+08],
       [2.01239703e+08],
       [4.02479407e+07]])
```

In [ ]:
원래는 위에처럼(PageRank), 이제는 현실적으로 Crawling+Scraping

In [156]…
```
urls = ['https://news.daum.net/'] # 앞으로 방문할곳
visited = [] # 이미 방문한곳

while urls: # 더이상 방문할 곳이 없을때까지
    url = urls.pop(0) # 앞에서부터 하나씩 꺼내고(FIFO)

    # robots.txt 검사
    try:
        print(urlparse(url).path, \
              canFetch(url, path=urlparse(url).path))
    except:
        print(url)

    # HTTP 받아오고
    resp = request('GET', url, headers=headers)

    # 잘 받았으면
    if resp.status_code == 200:
        # 방문한곳에 추가해주고
        visited.append(url)

        # Scarping(대상: text/html)
```

```
        if re.search('html', resp.headers['content-type']):
            # DOM 만들어주고
            dom = BeautifulSoup(resp.text, 'lxml')

            # 전략3-1. 특정 영역 - 뉴스 목록
            news = dom.select('''#mainContent + .main-content [href],
                                 #mainContent + .main-content [src]''')

            # 링크 찾기
            if len(news) > 0:
                a = []
                for temp in news:
                    temp = temp.attrs['href' if temp.has_attr('href') else 'src']
                    if not re.match(r'#|javascript|data:|about:', temp):
                        newurl = urljoin(url, temp)
                        a.append(newurl)

                    urls += list(filter(
                    lambda link:link not in visited and \
                                link not in urls, a))

            # 전략3-2. 뉴스 본문만
            news = dom.select_one('#mArticle')

            if news:
                # https://v.daum.net/v/20230711122803039
                with open(url.split('/')[-1].replace('?', '')+'.txt',
                          'w', encoding='utf8') as fp:
                    fp.write(news.text)
        # 전략 3-3. 뉴스 삽화(이미지) 저장
        elif re.search('image', resp.headers['content-type']):
            ext = re.search('(jpeg|png|bmp|webp|gif)',
                            resp.headers['content-type']).group(1)
            with open(url.split('/')[-1].replace('?', '')+'.'+ext, 'wb') as fp:
                    fp.write(resp.content)
    # 못받았으면
    else:
        # FIFO, 앞으로 방문할 곳 제일 뒤에 붙여주고
        if resp.status_code >= 500:
            urls.append(url)
        # 방문했다고 치고, 방문안하게
        else:
            visited.append(url)
```

```
/ True
/v/20230711123907181 True
/thumb/S96x60ht.u/ True
/media/news/news2016/cp/cp_kukminilbo.gif True
/v/20230711124500280 True
/thumb/S96x60ht.u/ True
/media/news/news2016/cp/cp_imbc.gif True
/v/20230711105114094 True
/thumb/S96x60ht.u/ True
/media/news/news2016/cp/cp_akn.gif True
/v/20230711125905454 True
/thumb/S96x60ht.u/ True
/media/news/news2016/cp/cp_yonhap.gif True
/v/20230711112811950 True
/thumb/S96x60ht.u/ True
/media/news/news2016/cp/cp_chosun.gif True
```

/v/20230711123707151 True
/thumb/S96x60ht.u/ True
/media/news/2020/cp/cp_ytn.gif True
/v/20230711113603332 True
/thumb/S96x60ht.u/ True
/media/news/news2016/cp/cp_ked.gif True
/v/20230711124818319 True
/thumb/S96x60ht.u/ True
/media/news/news2016/cp/cp_news1.gif True
/v/20230711123101079 True
/thumb/S96x60ht.u/ True
/v/20230711124839322 True
/thumb/S96x60ht.u/ True
/v/20230711120029334 True
/thumb/S96x60ht.u/ True
/media/news/news2016/cp/cp_munhwa.gif True
/v/20230711120021313 True
/thumb/S96x60ht.u/ True
/v/20230711120230478 True
/thumb/S96x60ht.u/ True
/media/news/news2016/cp/cp_moneytoday.gif True
/v/20230711123702148 True
/thumb/S96x60ht.u/ True
/v/20230711115110948 True
/thumb/S96x60ht.u/ True
/media/news/news2016/cp/cp_khan.gif True
/v/20230711094600042 True
/thumb/S96x60ht.u/ True
/v/20230711124016205 True
/thumb/S96x60ht.u/ True
/media/news/news2016/cp/cp_newsis.gif True
/v/20230711123758162 True
/thumb/S96x60ht.u/ True
/media/news/news2016/cp/cp_dailian.gif True
/v/20230711111017086 True
/thumb/S96x60ht.u/ True
/v/20230711124029214 True
/thumb/S96x60ht.u/ True
/series/5204526 True
/v/20230711125807444 True
/thumb/S112x70ht.u/ True
/series/5509370 True
/v/20230711125528400 True
/thumb/S112x70ht.u/ True
/series/5016367 True
/v/20230711115627144 True
/thumb/S112x70ht.u/ True
/series/1283511 True
/v/20230711113547315 True
/thumb/S112x70ht.u/ True
/series/272122 True
/v/20230711120319530 True
/thumb/S112x70ht.u/ True
/series/5811806 True
/v/20230711100221880 True
/thumb/S112x70ht.u/ True
/series/5575783 True
/v/20230711100139843 True
/thumb/S112x70ht.u/ True

```
/series/5714559 True
/v/20230711120009278 True
/thumb/S112x70ht.u/ True
/series/ True
/domestic/kospi True
/domestic/kosdaq True
/p/viewer/379/uODvWCZd8q True
/thumb/S302x170ht.u/ True
/p/viewer/379/ True
/p/viewer/5615618/FOsRKU4IwN True
/thumb/S302x170ht.u/ True
/p/viewer/5615618/ True
/p/viewer/674200/uuHl6KRyhi True
/thumb/S302x170ht.u/ True
/p/viewer/674200/ True
/p/viewer/1471/K5uSMfiHyd True
/thumb/S302x170ht.u/ True
/p/viewer/1471/ True
/info/correct True
/info/revision True
/info/unfair True
```

In [157]…
```
from os import listdir
len(list(filter(lambda f:re.search('[.]txt$', f), listdir('.'))))
```

Out[157]: 40

In [158]…
```
from os import listdir
len(list(filter(lambda f:re.search('[.](!:png|jpeg|bmp|gif|wepb)$', f), listdir('.'))))
```

Out[158]: 46