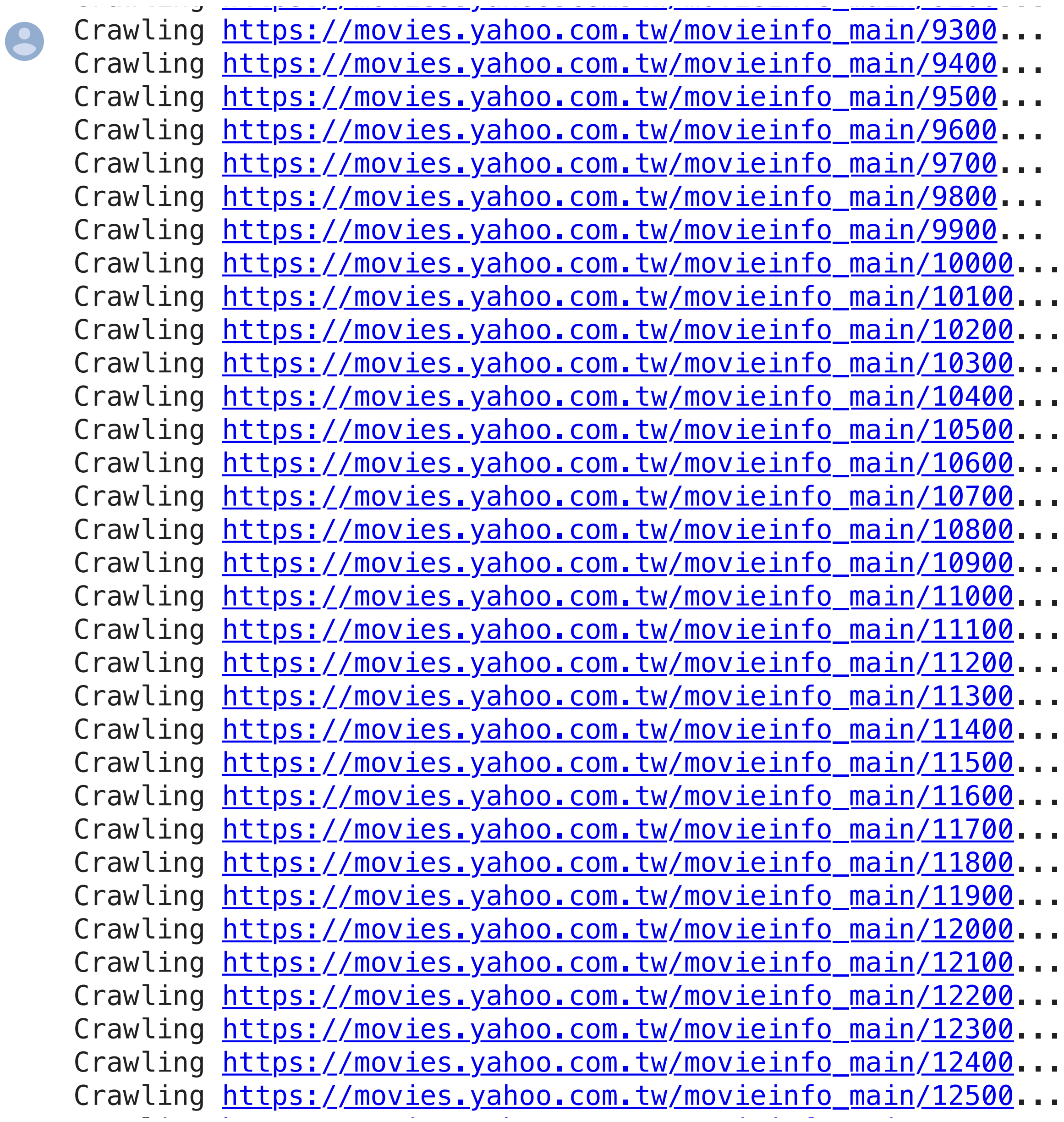
Build search engine

```
import requests
import re
import string
import jieba
import collections
import networkx as nx
from bs4 import BeautifulSoup
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem import SnowballStemmer

class MovieCrawler(object):
    def __init__(self):
        self.movies = []
        self.i = 1
        self.graph = nx.DiGraph()
        self.index = collections.defaultdict(list)
        self.links = collections.defaultdict(list)

    def tokenize_sentence(self, sentence):
        stop_words = stopwords.words("chinese")
        stop_words2 = ("\n", ", ", " ", "\r\n", "，", "。", "…", "★", "、", "《",

        tokens = jieba.cut(sentence, cut_all = False, HMM = True)
        tokens = [i for i in tokens if i not in string.punctuation]
        tokens = [i for i in tokens if i not in stop_words]
        tokens = [i for i in tokens if i not in stop_words2]

        return tokens

    def query(self, target):
        print("您的搜尋結果(Sorting by PageRank Value):")
        print(f"共{len(self.index[target])}筆，符合\"{target}\"---共indexing{len(s
        pagerank = nx.pagerank(self.graph, alpha = 1, tol = 1.0e-3, max_iter = 1
        rank = []
        for i, value in enumerate(self.index[target]):
            rank.append((i, pagerank[value]))
        rank = sorted(rank, key = lambda x: x[1], reverse = True)

        for i, r in rank:
            index = self.index[target][i]
            print(f'{self.movies[index]["doc_id"]}({r}): {self.movies[index]["cn

        n1 = n2 = 0
        for movie in self.movies:
            if (re.search(rf"{target}", movie["cname"] + movie["ename"] + movie[
            if (re.search(rf"{target}", movie["cname"] + movie["ename"] + movie[
                n1 += 1
        for i in self.index[target]:
            if (re.search(rf"{target}", self.movies[i]["cname"] + self.movies[i]
                n2 += 1

        print(f"\nPrecision = {n2 / len(self.index[target]) * 100}%")
        print(f"Recall = {n2 / n1 * 100}%")

    def add_movies(self, page_url):
        resp = requests.get(page_url)
        soup = BeautifulSoup(resp.text, 'html.parser')
        movie = soup.find("div", class_ = "movie_intro_info_r")

        if (movie == None):
            return

        temp = {}

        temp["doc_id"] = self.i
        temp["cname"] = movie.find("h1").text
        temp["ename"] = movie.find("h3").text

        label_info = movie.find("div", class_ = "level_name_box").find_all("div",
        temp["labels"] = []
        for l in label_info:
            label = l.text
            while (label[0] in [" ", "\n"]):
                label = label[1: ]
            while (label[-1] in [" ", "\n"]):
                label = label[: -1]

            temp["labels"].append(label)

        intro = soup.find("span", {"id": "story"}).text
        if (intro != None and intro != "" and len(intro) != 0):
            while (intro and intro[0] in [" ", "\n"]):\
                intro = intro[1: ]
            while (intro and intro[-1] in [" ", "\n"]):
                intro = intro[: -1]
            temp["intro"] = intro
        else:
            temp["intro"] = ""

        temp["release_date"] = movie.find("span").text[5: ]

        tokens = self.tokenize_sentence(temp["cname"]) + self.tokenize_sentence(
        for token in set(tokens):
            self.index[token].append(self.i - 1)

        if (self.i != 1):
            self.links[self.i].append(self.i - 1)
            self.links[self.i - 1].append(self.i)
            self.graph.add_edge(self.i, self.i - 1)
            self.graph.add_edge(self.i - 1, self.i)

        self.i += 1

        self.movies.append(temp)

        return self.movies

movie_url = "https://movies.yahoo.com.tw/movieinfo_main/"

crawler = MovieCrawler()
for i in range(1, 15059):
    if (i % 100 == 0):
        print(f"Crawling {movie_url}{i}...")
    crawler.add_movies(movie_url + str(i))

movies = crawler.movies

print(f"The length of all movies: {len(movies)}")
```



```
Crawling https://movies.yahoo.com.tw/movieinfo_main/9300...
Crawling https://movies.yahoo.com.tw/movieinfo_main/9400...
Crawling https://movies.yahoo.com.tw/movieinfo_main/9500...
Crawling https://movies.yahoo.com.tw/movieinfo_main/9600...
Crawling https://movies.yahoo.com.tw/movieinfo_main/9700...
Crawling https://movies.yahoo.com.tw/movieinfo_main/9800...
Crawling https://movies.yahoo.com.tw/movieinfo_main/9900...
Crawling https://movies.yahoo.com.tw/movieinfo_main/10000...
Crawling https://movies.yahoo.com.tw/movieinfo_main/10100...
Crawling https://movies.yahoo.com.tw/movieinfo_main/10200...
Crawling https://movies.yahoo.com.tw/movieinfo_main/10300...
Crawling https://movies.yahoo.com.tw/movieinfo_main/10400...
Crawling https://movies.yahoo.com.tw/movieinfo_main/10500...
Crawling https://movies.yahoo.com.tw/movieinfo_main/10600...
Crawling https://movies.yahoo.com.tw/movieinfo_main/10700...
Crawling https://movies.yahoo.com.tw/movieinfo_main/10800...
Crawling https://movies.yahoo.com.tw/movieinfo_main/10900...
Crawling https://movies.yahoo.com.tw/movieinfo_main/11000...
Crawling https://movies.yahoo.com.tw/movieinfo_main/11100...
Crawling https://movies.yahoo.com.tw/movieinfo_main/11200...
Crawling https://movies.yahoo.com.tw/movieinfo_main/11300...
Crawling https://movies.yahoo.com.tw/movieinfo_main/11400...
Crawling https://movies.yahoo.com.tw/movieinfo_main/11500...
Crawling https://movies.yahoo.com.tw/movieinfo_main/11600...
Crawling https://movies.yahoo.com.tw/movieinfo_main/11700...
Crawling https://movies.yahoo.com.tw/movieinfo_main/11800...
Crawling https://movies.yahoo.com.tw/movieinfo_main/11900...
Crawling https://movies.yahoo.com.tw/movieinfo_main/12000...
Crawling https://movies.yahoo.com.tw/movieinfo_main/12100...
Crawling https://movies.yahoo.com.tw/movieinfo_main/12200...
Crawling https://movies.yahoo.com.tw/movieinfo_main/12300...
Crawling https://movies.yahoo.com.tw/movieinfo_main/12400...
Crawling https://movies.yahoo.com.tw/movieinfo_main/12500...
Crawling https://movies.yahoo.com.tw/movieinfo_main/12600...
Crawling https://movies.yahoo.com.tw/movieinfo_main/12700...
Crawling https://movies.yahoo.com.tw/movieinfo_main/12800...
Crawling https://movies.yahoo.com.tw/movieinfo_main/12900...
Crawling https://movies.yahoo.com.tw/movieinfo_main/13000...
Crawling https://movies.yahoo.com.tw/movieinfo_main/13100...
Crawling https://movies.yahoo.com.tw/movieinfo_main/13200...
Crawling https://movies.yahoo.com.tw/movieinfo_main/13300...
Crawling https://movies.yahoo.com.tw/movieinfo_main/13400...
Crawling https://movies.yahoo.com.tw/movieinfo_main/13500...
Crawling https://movies.yahoo.com.tw/movieinfo_main/13600...
Crawling https://movies.yahoo.com.tw/movieinfo_main/13700...
Crawling https://movies.yahoo.com.tw/movieinfo_main/13800...
Crawling https://movies.yahoo.com.tw/movieinfo_main/13900...
Crawling https://movies.yahoo.com.tw/movieinfo_main/14000...
Crawling https://movies.yahoo.com.tw/movieinfo_main/14100...
Crawling https://movies.yahoo.com.tw/movieinfo_main/14200...
Crawling https://movies.yahoo.com.tw/movieinfo_main/14300...
Crawling https://movies.yahoo.com.tw/movieinfo_main/14400...
Crawling https://movies.yahoo.com.tw/movieinfo_main/14500...
Crawling https://movies.yahoo.com.tw/movieinfo_main/14600...
Crawling https://movies.yahoo.com.tw/movieinfo_main/14700...
Crawling https://movies.yahoo.com.tw/movieinfo_main/14800...
Crawling https://movies.yahoo.com.tw/movieinfo_main/14900...
Crawling https://movies.yahoo.com.tw/movieinfo_main/15000...
The length of all movies: 12230
```

Display query result

```
crawler.query("音樂")
```

```
經典動畫《大鼻子小英雄》、《大力士阿羅夏》王牌製片 Sergey Selyanov 出手，找來安錫影展入

12118(8.176614881439084e-05): BLACKPINK：The Movie
                                      (2021)
 BLACKPINK：The Movie
瘋迷全球超人氣韓國流行音樂女團「BLACKPINK」，為了慶祝出道五週年，推出《BLACKPINK：The

12122(8.176614881439084e-05): 揮灑心篇章 Maestro(s)
★2022昂古萊姆法語電影節　正式入選
★ 《樂動心旋律》《貝禮一家》製片公司，號召《苦兒流浪記》《獵殺星期一》製片群，再次呈獻撼動
★《找我經紀人》攝影X《緣來想見妳》剪輯X《ANNETTE：星夢戀歌》梳化，合力譜出影史最美音樂饗
★《愛愛後當機》導演夢想溫情心作，改編自榮獲坎城影展最佳編劇、提名奧斯卡最佳外語片提名名作

父與子間充滿矛與盾般的尷尬鴻溝，最後只能靠音符再次拉近彼此心的距離…

杜馬斯家是一對有名的指揮家父子檔，但感情卻並不融洽。爸爸弗朗索瓦正準備結束漫長但精彩的國際

12150(8.176614881439084e-05): 異想羅曼史
                                              (2023)
 Up Here
音樂浪漫喜劇《異想羅曼史》以1999年的紐約市為背景，講述了一對普通情侶不平凡的愛情故事。二人

12154(8.176614881439084e-05): 唐璜羅曼死 Don Juan
```

```
我是情聖唐璜，卻是情場上的輸家...
在舞台劇總飾演「情聖唐璜」的羅宏（塔哈拉辛 Tahar Rahim 飾），即將與心愛的女友茱莉（薇吉

為重拾男性自尊並療癒受傷的心靈，羅宏決定重新投入戀愛，並對他所遇見的每個女子展開追求…但彷

【電影簡介】

凱薩獎影帝后合體《唐璜羅曼死》情聖變情剩顛覆經典
情場左右逢源是男性的專利嗎？由法國名導賽吉波宗（Serge Bozon）改編經典劇作的新片《唐璜羅

近年兩性意識成為話題主流，《唐璜羅曼死》正是導演賽吉波宗展現野心的佳作，他藉由一位形象深植

賽吉波宗的神來一筆，將唐璜從「情聖」變成了「情剩」。把這位家喻戶曉的花花公子，變成了現代的

12184(8.176614881439084e-05): 魔髮精靈：樂團在一起 Trolls Band Together
今年聖誕佳節，夢工廠動畫將推出由安娜坎卓克和賈斯汀提姆布萊克率領全明星配音演員陣容演出，色

在前兩集經過友情的考驗和彼此不斷的打情罵俏之後，波比（安娜坎卓克 配音）和小布（賈斯汀提姆

但是當小布的兄弟佛洛伊因為他的音樂才華被兩個流行音樂的邪惡大壞蛋－凡菲（艾美獎得主及《姐姐

《魔髮精靈：樂團在一起》這部全新音樂喜劇動畫片和前兩集一樣，都有歡樂無窮的全新和經典流行夯

再度回歸為這部全新動畫片獻聲的配音演員則包括葛萊美獎、艾美獎及金球獎入圍者柔伊黛絲香奈，她

導演華特道恩再度回歸執導《魔髮精靈：樂團在一起》，吉娜夏則再度擔任製片，共同導演是提姆海茲

12189(8.176614881439084e-05): SUGA: Road to D-DAY
                                   SUGA: Road to D-DAY
BTS防彈少年團成員中的音樂才子 SUGA在與酷玩樂團Coldplay、韓國「江南大叔」PSY、世界前十大

《SUGA: Road to D-DAY》紀錄了21世紀轟動全球的韓流天團BTS防彈少年團成員SUGA，從首爾、


Precision = 100.0%
Recall = 65.0632911392405%
```

Write json file

```
import json

pagerank = nx.pagerank(crawler.graph, alpha = 1, tol = 1.0e-3, max_iter = 100000

for i in range(len(crawler.movies)):
    crawler.movies[i]["pagerank"] = pagerank[crawler.movies[i]["doc_id"]]
    crawler.movies[i]["links"] = crawler.links[crawler.movies[i]["doc_id"]]

with open("hw2.json", "w", encoding = "utf-8") as f:
    json.dump(crawler.movies, f)
```

Display result by loading json file

```
with open("hw2.json", "r") as f:
    data = json.load(f)

    print(data)
```

```
IOPub data rate exceeded.
The notebook server will temporarily stop sending output
to the client in order to avoid crashing it.
To change this limit, set the config variable
`--NotebookApp.iopub_data_rate_limit`.

Current values:
NotebookApp.iopub_data_rate_limit=1000000.0 (bytes/sec)
NotebookApp.rate_limit_window=3.0 (secs)
```