Algorithm-

START

1. Goto https://www.lib.lsu.edu/
2. Select Databases tab and then ‘w’ from Browse Databases.
3. Click on the ‘Web of Science’ link.
   Note – Instead of above 3 steps, we can directly go to the link - https://sso.paws.lsu.edu/login?service=https%3A%2F%2Fwebauth.shib.lsu.edu%2Fidp%2FAuthn%2FExtCas%3Fconversation%3De1s1&entityId=https%3A%2F%2Flibezp.lib.lsu.edu%2Fshibboleth

4. Login with the username and password.
5. Select the advanced search tab and write the query.
   For example – if we want the institutions that have cited ‘MIT’ in the years from 1998 to 2018 in the field of Astronomy and Astrophysics, then search for ‘MIT’ after selecting ‘OG’ field tag, add the required institution (in our case its MIT) and then click ok. You will be redirected to the query page with the query ‘OG=(Massachusetts Institute of Technology (MIT))’. Now sleect the topic from the WC= Web of Science Category field tag (in our case its Astronomy & Astrophysics) and write a complete query as -
   ‘OG=(Massachusetts Institute of Technology (MIT)) and WC=Astronomy & Astrophysics’
   Click on search.
   Note that time span is selected to be ‘All years (1900 – 2018)’
6. Now click on the results in the search history.
7. After this refine the search by selecting the publication year one by one.
   This can be done by two ways-
   a) Select the publication year from the left menu and click refine
   OR
   b) Goto the link http://apps.webofknowledge.com.libezp.lib.lsu.edu/RAMore.do?product=WOS&search_mode=AdvancedSearch&SID=7BVklzaxLTCSPkDWooE&qid=1&ra_mode=more&ra_name=PublicationYear&colName=WOS&viewType=raMore
   and select years one by one and click refine.
8. Once you get the refined results, click on the ‘Create Citation Report’ option on the right side menu.
9. Select ‘Analyze’ under ‘Citing articles’, then select the ‘Organizations-Enhanced’ option in the left menu. Select the ‘Sort by **Record count**’ as ‘Show **500**’.
10. Copy the url and run the script scraper.py in the terminal with arguments as – url , name of institute you are scraping information for, year for which you are scraping information for.
    For example -

    ```
    python3 scraper.py "http://wcs.webofknowledge.com/RA/analyze.do?product=WOS&SID=7BVklzaxLTCSPkDWooE&field=OG_OrgEnhancedName_OrgEnhancedName_en&yearSort=false" MIT '2018'
    ```

    The above command will make a folder by the name ‘MIT’ in the current directory and make a file 2018.json with the citing institutes list and no of citations to MIT in that particular year.

11. The scraper.py is as follows -

    ```
    import os , sys
    from selenium  import webdriver
    from selenium.webdriver.common.keys import Keys
    import re
    ```

```
from bs4 import BeautifulSoup
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
import json
import errno
browser = webdriver.Firefox()
str1=[]
#yearList=['2018','2017','2016','2015,2014,2013,2012,2011,2010,2009,2008,2007,2006,2005,2004,2003,2002,2001,2000,1999,1998,1997,1996,1995]
def start():
    name=sys.argv[2]
    make_directory(name)
    url=sys.argv[1]
    print(url)
    browser.get(url)
    elem = WebDriverWait(browser,10).until(EC.presence_of_element_located((By.XPATH, '/html/body/div[2]/div[6]/form/div/div[2]/div[2]/div/div[2]/div[4]/table')))
    source = browser.page_source
    scrape(source,name)

def scrape(source,name):
    soup = BeautifulSoup(source,'html.parser')

    file=open('./'+name+'/'+sys.argv[3]+'.json',"w")
    str1=[]
    str1.append("{")
    for tr in soup.findAll("tr",{"class":"RA-NEWRAresultsEvenRow"}):
        tds = tr.find_all('td')
        str1.append("\"")
        str1.append(tds[1].text)
        str1.append("\"")
        str1.append(":")
        str1.append("\"")
        str1.append(tds[2].text.strip())
        str1.append("\"")
        str1.append(",")
    for tr in soup.findAll("tr",{"class":"RA-NEWRAresultsOddRow"}):
        tds = tr.find_all('td')
        str1.append("\"")
        str1.append(tds[1].text)
        str1.append("\"")
        str1.append(":")
        str1.append("\"")
        str1.append(tds[2].text.strip())
        str1.append("\"")
        str1.append(",")
    var = ''.join(str1)
    file.write(var)
    file.close()
    correct_json_data(name)

def correct_json_data(name):
        instituteFileRead=open('./'+name+'/'+sys.argv[3]+'.json',"r")
        finalstr=(instituteFileRead.read())
                finalstr = finalstr[:-1]
        finalstr+="}"
        instituteFileRead.close()
        instituteFileWrite=open('./'+name+'/'+sys.argv[3]+'.json',"w")
        var = ''.join(finalstr)
        instituteFileWrite.write(var)
        instituteFileWrite.close()

def make_directory(name):
    path = "./"+name
    os.makedirs(path, exist_ok=True)
    print ("Path is created")

start()
```

12. Now to get the data for the remaining years follow these steps recusively,

    a)Goto the link http://apps.webofknowledge.com.libezp.lib.lsu.edu/RAMore.do?product=WOS&search_mode=AdvancedSearch&SID=7BVklzaxLTCSPkDWooE&qid=1&ra_mode=more&ra_name=PublicationYear&colName=WOS&viewType=raMore

    b)Select the year and click refine.

    c)Once you get the refined results, click on the ‘Create Citation Report’ option on the right side menu. Select ‘Analyze’ under ‘Citing articles’.

    d)Now execute the same previous command with only difference in the 3rd argument, changing it to the year the results are refined in. Note – the same url can be used once the result is refined manually, under one session.
13. After executing step 12 for all the years we want data for, we will get files 2018.json, 2017.json,2016.json,.... in the MIT directory inside the current directory.
14. Run the script test.py in the same directory

    ```
    python3 test.py
    ```

    The above command will create a directory “CitingInstitutes” inside “MIT” in the current - directory, with InstituteName.json files which have the count of the no. of citations from the year 1998 to 2018 for the InstituteName Institute to MIT.
15. The test.py is as follows

    ```
    import os , sys
    import json
    from pprint import pprint

    def make_directory():
        path = "./MIT/CitingInstitutes"
        os.makedirs(path, exist_ok=True)
        print ("Path is created")

    def makeitwork():
            institutesNameList=['UNIVERSITY OF CALIFORNIA SYSTEM','MASSACHUSETTS INSTITUTE OF TECHNOLOGY MIT','UNITED STATES DEPARTMENT OF ENERGY DOE','ISTITUTO NAZIONALE DI FISICA NUCLEARE','HARVARD
    ```

```
UNIVERSITY','CHINESE ACADEMY OF SCIENCES','CONSEJO
SUPERIOR DE INVESTIGACIONES CIENTIFICAS CSIC','UNIVERSITY
OF CHICAGO','UNIVERSITE PARIS SACLAY','RUSSIAN ACADEMY
OF SCIENCES','CEA','SORBONNE UNIVERSITE','UNIVERSITY OF
CALIFORNIA BERKELEY','CNRS NATIONAL INSTITUTE OF
NUCLEAR PARTICLE PHYSICS IN2P3','GODDARD SPACE FLIGHT
CENTER','SAPIENZA UNIVERSITY ROME','UNIVERSITE SORBONNE
PARIS CITE USPC COMUE','UNIVERSITY OF MICHIGAN','OHIO
STATE UNIVERSITY','JOHNS HOPKINS UNIVERSITY','EUROPEAN
ORGANIZATION FOR NUCLEAR RESEARCH CERN','ISTITUTO
NAZIONALE GEOFISICA E VULCANOLOGIA INGV','UNIVERSITY OF
TORONTO','UNIVERSITY OF OXFORD','DEUTSCHES ELEKTRONEN
SYNCHROTRON DESY','STANFORD UNIVERSITY','UNIVERSITY OF
WASHINGTON','UNIVERSITY OF WISCONSIN
MADISON','NORTHWESTERN UNIVERSITY','INDIAN INSTITUTE OF
TECHNOLOGY SYSTEM IIT SYSTEM','JOINT INSTITUTE FOR
NUCLEAR RESEARCH RUSSIA','UNIVERSITY OF
LONDON','UNIVERSITY OF BIRMINGHAM','CALIFORNIA
INSTITUTE OF TECHNOLOGY','PENNSYLVANIA COMMONWEALTH
SYSTEM OF HIGHER EDUCATION PCSHE']
    stringToWriteInFile=[]
    for i in range (1998,2019):
        make_directory()
        yearFile=open('./MIT/'+str(i)+'.json',"r")
        str1=(yearFile.read())
        with open('./MIT/'+str(i)+'.json') as f:
            data = json.load(f)
        for j in range (35):
            try:
                if (i==1998):
                    stringToWriteInFile.append("{")
                instituteFile=open('./MIT/CitingInstitutes/'+institutesNameList[j]+'.json',"a")
                stringToWriteInFile.append("\""+str(i)+"\""+":"+"\""+data[institutesNameList[j]]+"\",")
                var = ''.join(stringToWriteInFile)
                instituteFile.write(var)
                instituteFile.close()
            except:
                print ("Exception handled")
            stringToWriteInFile=[]

    """	In the code below I have written the code to automatically remove the last comma ',' in the json data and place a '}' in place
    of it using some of the string operations
    """

    for i in range (35):	#35 because I have considered only 35 institutes as of now to get the no of citations from them to MIT and the
                #35 institutes are mentioned in the string institutesNameList
        instituteFileRead=open('./MIT/CitingInstitutes/'+institutesNameList[i]+'.json',"r")
        finalstr=(instituteFileRead.read())
        print (finalstr)
        finalstr = finalstr[:-1]
        finalstr+="}"
        instituteFileRead.close()
        instituteFileWrite=open('./MIT/CitingInstitutes/'+institutesNameList[i]+'.json',"w")
        var = ''.join(finalstr)
        instituteFileWrite.write(var)
        instituteFileWrite.close()
makeitwork()
```

16. Note that only the data of the institutes mentioned in the list by the name ‘institutesNameList’ in the test.py code will be made. Though we have data for 500 institutes per year but here we are only mentioning 35 institute names in the list. Also the variable i of the second loop is hard coded to be 35 i.e. length of string.

17. Once test.py is run, we will get data in json files. Now it can be converted to csv files by the following command -
Example-

```
json2csv -i STANFORD\ UNIVERSITY.json  -o STANFORD\ UNIVERSITY.csv
```

The csv files will be stored in the current directory.

The data is ready to be used for models.

END