Web Mining – Lab 1

Understanding Requests and Re Modules

By Abhijeet Ambadekar (16BCE1156)

The aim of this experiment is to have a general overlook on the python requests module and create a program that find out the urls of all the webpages and images that can be accessed from a given input url.

Program Code:

L1.py

```
import requests
from urllib.parse import urljoin
from pprint import pprint
import re

def pprint_data_to_file(data, fname):
    with open(fname, 'w') as out:
        pprint(data, stream=out)

def get_href_links(url):
    r = requests.get(url)
    pattern = re.compile('(?<=href=").*?(?=")')
    return list(set([urljoin(url, link) for link in pattern.findall(r.text)]))

def get_img_src_links(url):
    r = requests.get(url)
    pattern = re.compile('(?<=src=").*?(?=")')
    return list(set([urljoin(url, link) for link in pattern.findall(r.text)]))


url = "http://www.vit.ac.in"

urls = get_href_links(url)
pprint_data_to_file(urls, "urls.txt")

img_urls = get_img_src_links(url)
pprint_data_to_file(img_urls, "imgurls.txt")
```

Output (urls.txt):


```
['http://www.vit.ac.in/files/admissions/viteee/VITEEE2018_Physics.pdf',
 'http://vtop2.vit.ac.in:8080/VITEEE/',
 'http://www.vit.ac.in/iprcell',
 'http://www.vit.ac.in/',
 'http://www.vit.ac.in/campus/ncc',
 'http://www.vit.ac.in/about/infrastructure',
 'http://www.vit.ac.in/files/VITEEE2017STATUS/Phase2-Day5_19-May-2017.pdf',
 'http://www.vit.ac.in/files/B_Tech_Part-Time-Notification-2017.jpg',
 'http://www.vit.ac.in#counselling1',
 'http://intranet.vit.ac.in/',
 'http://www.vit.ac.in/admissions/research',
 'http://www.vit.ac.in/academics/centers',
 'http://www.vit.ac.in/css/hover-min.css',
 'http://intranet.vit.ac.in',
 'http://www.vit.ac.in/research/sponsoredResearch',
 'http://vtop3.vit.ac.in:8080/VITEEERESULTS/',
 'http://www.vit.ac.in/academics/itp',
 'http://vitbhopal.ac.in/',
 'http://www.vit.ac.in/about/community',
 'http://admission.vit.ac.in/Results/2017/ARC_Results2017/index.asp',
 'http://vtop7.vit.ac.in:8080/paymentfreshers/',
 'https://webmail.vit.ac.in',
 'http://info.vit.ac.in/phddec2017/index.asp',
 'http://www.vit.ac.in/{{ breadcrumb.path }}',
 'http://vit.ac.in/files/viteee2017/Hotels_VLR.pdf',
 'http://www.vit.ac.in/files/admissions/PhysicalFitness_Certificate.pdf',
 'http://www.vitaa.org/',
 'http://www.vit.ac.in/files/admissions/viteee/VITEEE2018_Mathematics.pdf',
 'http://www.vit.ac.in/redressal',
 'http://vtop2.vit.ac.in:8080/UGFOREIGN/',
 'https://vtop9.vit.ac.in/vtop/login/freshers',
```

'http://www.mhrdnats.gov.in/',

'http://www.vit.ac.in/campus/startups',

'http://www.vit.ac.in/campus/hostels',

'http://www.vit.ac.in/placement/consortium',

'http://careers.vit.ac.in/',

'http://www.vit.ac.in/about/leadership',

'http://www.vit.ac.in/icc',

'http://www.vit.ac.in/placement/dreamoffers',

'https://peopleorbit.vit.ac.in/',

'http://www.vit.ac.in/events/eventView/Two Day Workshop on Raspberry Pi '

'Programming for Beginners 2018',

'http://www.vit.ac.in/files/SBST_Freshers_App.rar',

'http://vtop2.vit.ac.in:8080/UGNRI/',

'http://www.vit.ac.in/files/FormatGuidelines.doc',

'http://www.vit.ac.in/campus/fests',

'http://www.vit.ac.in/events',

'http://www.vit.ac.in/files/admissions/Affidavit_Student.pdf',

'http://www.vit.ac.in/research/academic',

'http://www.vit.ac.in/about/mhrd',

'http://www.vit.ac.in/ap/careers',

'http://www.vit.ac.in/campus/sports',

'http://www.vit.ac.in/btechadmissions/viteee2018',

'http://www.vit.ac.in/placement',

'http://www.vit.ac.in/images/favicon.ico',

'http://www.vit.ac.in/files/admissions/viteee/VITEEE2018_Chemistry.pdf',

'http://www.vit.ac.in/about/raac',

'http://www.vit.ac.in/campus/studentchapters/creationLabs',

'http://www.vit.ac.in/files/ug-2018/index.html',

'http://www.vit.ac.in/campus',

'http://www.vit.ac.in/files/hostels/HostelAdmission_InformationSheet.pdf',

'http://academics2.vit.ac.in/onlinewithdraw/',

'http://www.vit.ac.in/academics/internationalRelations',

'https://webmail.vit.ac.in/',

'https://mail.google.com/',

'http://www.vit.ac.in/events/eventView/One Day Workshop on Modelling and '

'simulation of electric motors using ANSYS - MAXWELL for beginners',

'http://www.vit.ac.in/files/wifiservices.pdf',

'http://www.vit.ac.in/academics/ffcs',

'http://www.vit.ac.in/admissions',

'https://plus.google.com/107959047122513483934',

'http://www.vit.ac.in/events/eventView/VIT Summer School ',

'http://www.vit.ac.in/files/admissions/viteee/VITEEE2018_English.pdf',

'http://www.vit.ac.in/css/materialize.min.css',

'http://www.vit.ac.in#counselling',

'http://www.vit.ac.in/files/admissions/Hostel_Affidavit_Ladies_2018.pdf',

'http://www.vit.ac.in/about',

'http://www.vit.ac.in/admissions/pg',

'http://www.vit.ac.in/academics/certificates',

'http://www.vit.ac.in/academics/schools',

'http://www.vit.ac.in/about/administrativeOffices',

'https://vtopbeta.vit.ac.in/studentprofile/',

'http://info.vit.ac.in/guesthouse',

'http://www.vit.ac.in/ap',

'http://www.vit.ac.in/files/admissions/viteee/VITEEE2018_Biology.pdf',

'http://www.vit.ac.in/css/style.css',

'http://www.vit.ac.in/placement/pat',

'http://www.vit.ac.in/admissions/eligibility',

'http://www.vit.ac.in/campus/otheramenities',

'https://fonts.googleapis.com/icon?family=Material+Icons',

'http://www.vit.ac.in',

'http://www.vit.ac.in/admissions/international',

'http://www.vit.ac.in/admissions/testcities',

'http://www.vit.ac.in/campus/studentchapters',

'http://www.vit.ac.in/academics',

'http://www.vit.ac.in/contactus',

'http://www.vit.ac.in/ '

'http:/www.vit.ac.in/files/admissions/Affidavit_Parent.pdf',

'http://www.vit.ac.in/files/admissions/Hostel_Affidavit_Mens_2018.pdf',

'http://chennai.vit.ac.in/',

'http://www.vit.ac.in/files/VITEEE2017STATUS/Phase2-Day1_15-May-2017.pdf',

'http://www.vit.ac.in/campus/studentclubs',

'http://www.vit.ac.in/admissions/postoffices',

'http://www.vit.ac.in/academics/transcripts',

'http://www.vittbi.com/tedp_food.pdf',

'http://vit-otbs-lb.centralindia.cloudapp.azure.com/vitotbs/',

'http://www.vit.ac.in/research',

'http://www.vit.ac.in/placement/advancedTraining',

'https://academics.vit.ac.in/faculty/fac_login.asp',

'http://www.vit.ac.in/css/angular-csp.css',

'http://gmail.vit.ac.in',

'http://www.vit.ac.in/about/news',

'http://www.vit.ac.in/academics/iqac',

'http://www.vit.ac.in/vitunplugged',

'http://www.vit.ac.in/btechadmissions/viteee2017',

'http://www.vit.ac.in/about/sustainability',

'http://www.vit.ac.in/admissions/ug',

'http://careers.vit.ac.in/careers/',

'http://www.vit.ac.in/campus/sae',

'http://www.vit.ac.in/css/home.css',

'http://www.vit.ac.in/research/centers',

'http://www.vit.ac.in/admissions/procedure',

'https://vtopbeta.vit.ac.in/vtop/',

'http://www.vit.ac.in/academics/library',

'http://www.vit.ac.in/admissions/feeStructure',

'http://www.vit.ac.in/academics/coe',

'http://www.vit.ac.in/PGAdmissions2018',

'http://www.vit.ac.in/campus/healthservices',

'http://info.vit.ac.in/map/']

Output (imgurls.txt)

```
['http://www.vit.ac.in/js/angular.js',

 'http://www.vit.ac.in/js/angular-sanitize.js',

 'http://www.vit.ac.in/js/google-map-api.js',

 'http://www.googletagmanager.com/ns.html?id=GTM-WMCMXM',

 'http://www.vit.ac.in/js/jquery-2.1.3.min.js',

 'http://www.vit.ac.in/js/script.js',

 'http://www.vit.ac.in/undefined',

 'http://www.vit.ac.in/js/angular-route.js',

 'http://www.vit.ac.in/js/one-angular.js',

 'http://www.vit.ac.in/images/logo.png',

 'http://www.vit.ac.in/js/ocLazyLoad.min.js',

 'http://www.vit.ac.in/js/ngMap.min.js',

 'http://www.vit.ac.in/js/init.js',

 'http://www.vit.ac.in/images/ripple.svg',

 'http://www.vit.ac.in/js/route-styles.js',

 'http://www.vit.ac.in/images/logo1.png',

 'http://www.vit.ac.in/js/materialize.js',

 'http://code.tidio.co/iuj4oxikuxxrmxjhavcs6yd6us8ijebu.js',

'http://platform-api.sharethis.com/js/
sharethis.js#property=5ae3047f6a9348001198603e&product=sticky-share-buttons']
```