# Web Mining – Lab 2

## Generating a Mini-Crawler

By Abhijeet Ambadekar (16BCE1156)

The aim of this experiment is to use the python requests and BeautifulSoup modules to create a mini crawler that can crawl over a given seed url to generate a page table containg the list of all the urls that can be accessed from the seed url and recursively to those urls and similarly a reverse page table to backtrack this data.

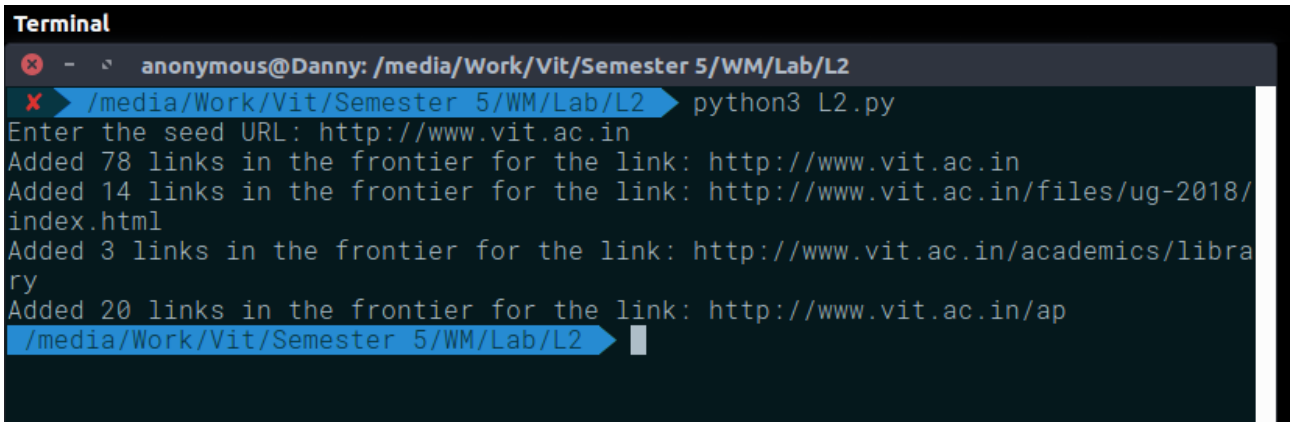
**Program Code:**

```
import requests
from bs4 import BeautifulSoup
from pprint import pprint

try:
	# Python 3
	from urllib.parse import urlparse, urljoin
except ImportError:
	from urlparse import urljoin, urlparse



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


class Crawler(object):
	def __init__(self):
		self.pagetable = {}

	def get_seed(self):
		self.seed_url = input("Enter the seed URL: ").strip()
		# self.seed_url = "http://www.vit.ac.in"
		self.hostname = urlparse(self.seed_url).hostname
		self.frontier = [self.seed_url]
		self.rev_pagetable = {}

	def seed_test(self, url):
		return True if self.hostname in url else False

	def get_all_urls(self, url):
		try:
			page = requests.get(url)
		except Exception as e:
			try:
				print("Failed to reached {}".format(url))
```

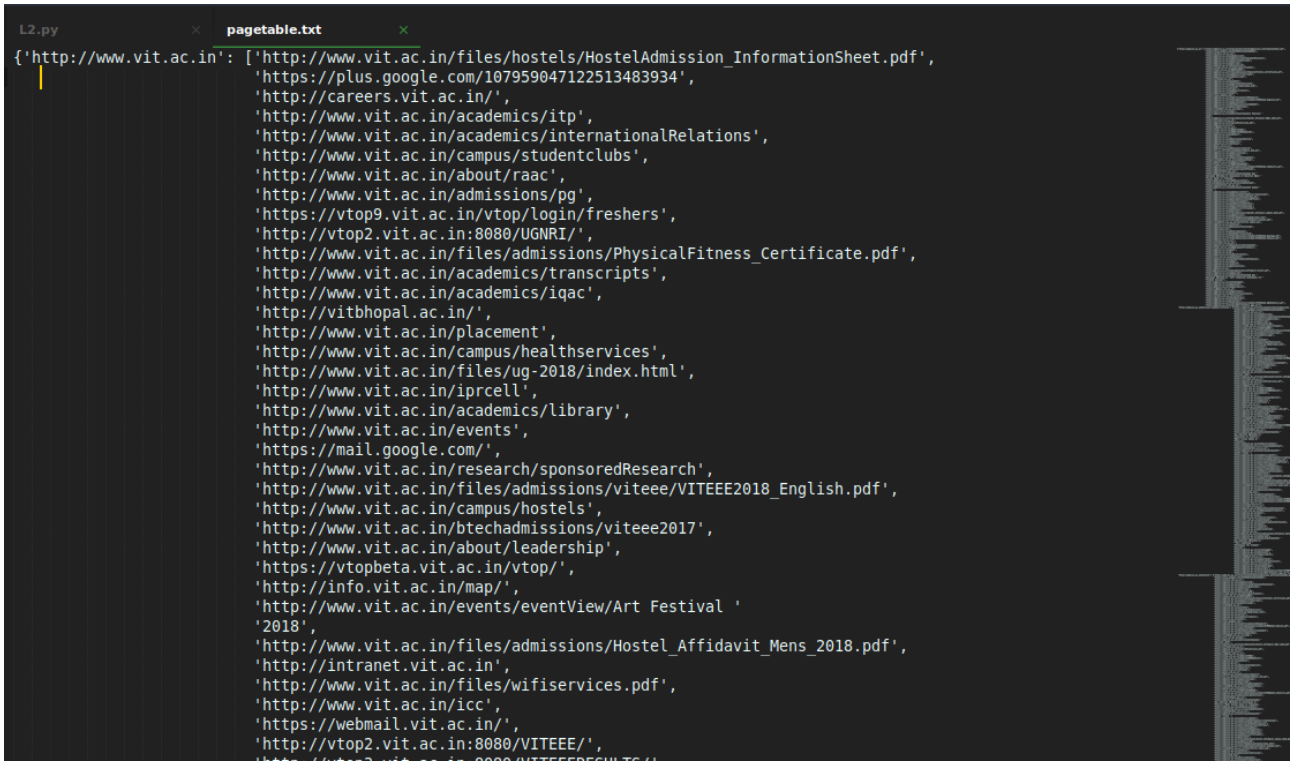
```
                except UnicodeEncodeError:
                        print("Failed to reached and cant show the URL")
                return None

        soup = BeautifulSoup(page.text, 'html.parser')

        urls = []
        for link in soup.find_all('a', href=True):
                if link.get('href') in [None, "#", ""]:
                        continue
                urls.append(urljoin(self.seed_url, link.get('href')))

        return list(set(urls))


    def crawl(self):
        for curr_url in self.frontier:

                urls = self.get_all_urls(curr_url)
                if not urls:
                        continue

                oldFronLen = len(self.frontier)

                for url in urls:
                        if curr_url in self.pagetable:
                                self.pagetable[curr_url].append(url)
                        else:
                                self.pagetable[curr_url] = [url]

                        if url not in self.frontier and self.seed_test(url):
                                self.frontier.append(url)

                if len(self.frontier) > oldFronLen:
                        try:
                                print("Added {} links in the frontier for the
link: {}".format(len(self.frontier)-oldFronLen, curr_url))
                        except UnicodeEncodeError:
                                print("Added {} links in the frontier for the
link: NOT ABLE TO PRINT".format(len(self.frontier)-oldFronLen))


    def rev_table_gen(self):
        for url in self.pagetable:
                for target in self.pagetable[url]:
                        if target in self.rev_pagetable:
                                if url in self.rev_pagetable[target]:
                                        continue
                                else:
                                        self.rev_pagetable[target].append(url)
                        else:
```

```
                                self.rev_pagetable[target] = [url]


vitc = Crawler()
vitc.get_seed()
vitc.crawl()

pprint_data_to_file(vitc.pagetable, "pagetable.txt")
pprint_data_to_file(vitc.rev_pagetable, "rev_pagetable.txt")
pprint_data_to_file(vitc.frontier, "frontier.txt")
```

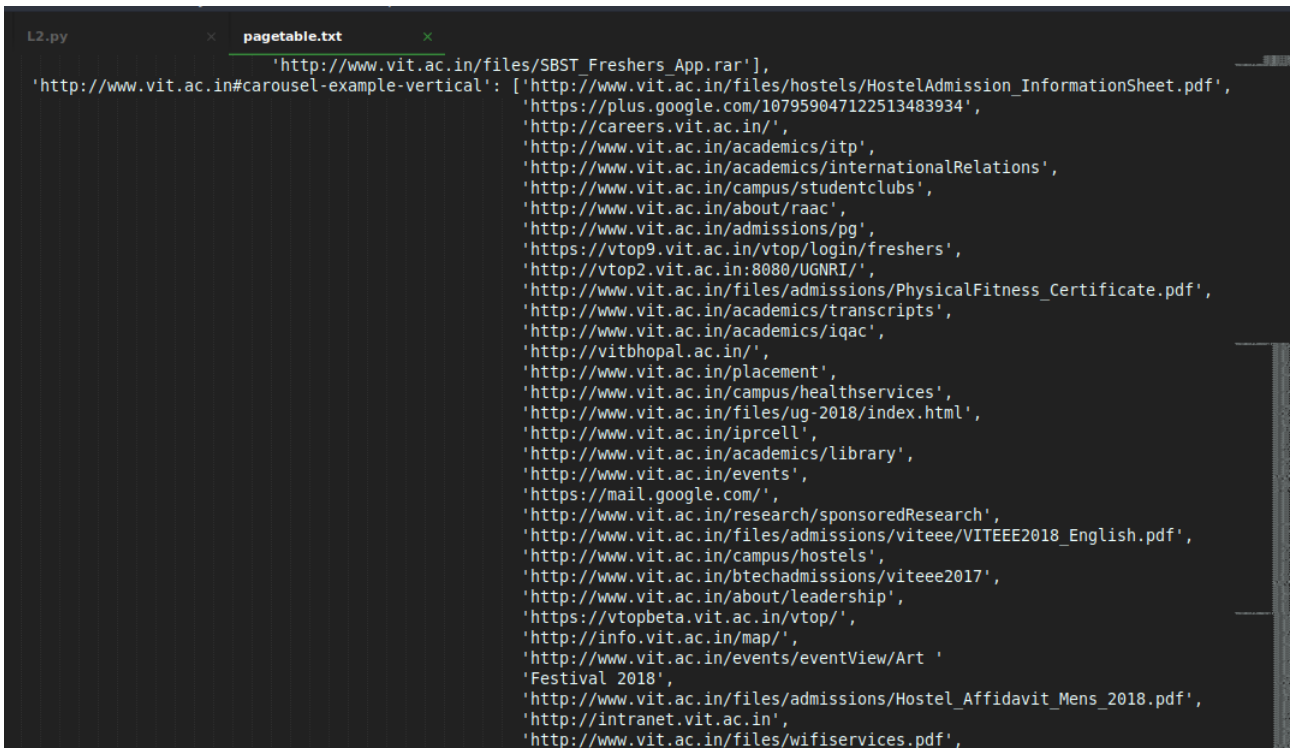
**Output:**



**Output (pagetable.txt)** (A file of 10000 lines, hence discrete output)**:**



```
L2.py    pagetable.txt
{'http://www.vit.ac.in': ['http://www.vit.ac.in/files/hostels/HostelAdmission_InformationSheet.pdf',
                          'https://plus.google.com/107959047122513483934',
                          'http://careers.vit.ac.in/',
                          'http://www.vit.ac.in/academics/itp',
                          'http://www.vit.ac.in/academics/internationalRelations',
                          'http://www.vit.ac.in/campus/studentclubs',
                          'http://www.vit.ac.in/about/raac',
                          'http://www.vit.ac.in/admissions/pg',
                          'https://vtop9.vit.ac.in/vtop/login/freshers',
                          'http://vtop2.vit.ac.in:8080/UGNRI/',
                          'http://www.vit.ac.in/files/admissions/PhysicalFitness_Certificate.pdf',
                          'http://www.vit.ac.in/academics/transcripts',
                          'http://www.vit.ac.in/academics/iqac',
                          'http://vitbhopal.ac.in/',
                          'http://www.vit.ac.in/placement',
                          'http://www.vit.ac.in/campus/healthservices',
                          'http://www.vit.ac.in/files/ug-2018/index.html',
                          'http://www.vit.ac.in/iprcell',
                          'http://www.vit.ac.in/academics/library',
                          'http://www.vit.ac.in/events',
                          'https://mail.google.com/',
                          'http://www.vit.ac.in/research/sponsoredResearch',
                          'http://www.vit.ac.in/files/admissions/viteee/VITEEE2018_English.pdf',
                          'http://www.vit.ac.in/campus/hostels',
                          'http://www.vit.ac.in/btechadmissions/viteee2017',
                          'http://www.vit.ac.in/about/leadership',
                          'https://vtopbeta.vit.ac.in/vtop/',
                          'http://info.vit.ac.in/map/',
                          'http://www.vit.ac.in/events/eventView/Art Festival '
                          '2018',
                          'http://www.vit.ac.in/files/admissions/Hostel_Affidavit_Mens_2018.pdf',
                          'http://intranet.vit.ac.in',
                          'http://www.vit.ac.in/files/wifiservices.pdf',
                          'http://www.vit.ac.in/icc',
                          'https://webmail.vit.ac.in/',
                          'http://vtop2.vit.ac.in:8080/VITEEE/',
```



```
L2.py    pagetable.txt
                          'http://www.vit.ac.in/files/SBST_Freshers_App.rar'],
 'http://www.vit.ac.in#carousel-example-vertical': ['http://www.vit.ac.in/files/hostels/HostelAdmission_InformationSheet.pdf',
                                                    'https://plus.google.com/107959047122513483934',
                                                    'http://careers.vit.ac.in/',
                                                    'http://www.vit.ac.in/academics/itp',
                                                    'http://www.vit.ac.in/academics/internationalRelations',
                                                    'http://www.vit.ac.in/campus/studentclubs',
                                                    'http://www.vit.ac.in/about/raac',
                                                    'http://www.vit.ac.in/admissions/pg',
                                                    'https://vtop9.vit.ac.in/vtop/login/freshers',
                                                    'http://vtop2.vit.ac.in:8080/UGNRI/',
                                                    'http://www.vit.ac.in/files/admissions/PhysicalFitness_Certificate.pdf',
                                                    'http://www.vit.ac.in/academics/transcripts',
                                                    'http://www.vit.ac.in/academics/iqac',
                                                    'http://vitbhopal.ac.in/',
                                                    'http://www.vit.ac.in/placement',
                                                    'http://www.vit.ac.in/campus/healthservices',
                                                    'http://www.vit.ac.in/files/ug-2018/index.html',
                                                    'http://www.vit.ac.in/iprcell',
                                                    'http://www.vit.ac.in/academics/library',
                                                    'http://www.vit.ac.in/events',
                                                    'https://mail.google.com/',
                                                    'http://www.vit.ac.in/research/sponsoredResearch',
                                                    'http://www.vit.ac.in/files/admissions/viteee/VITEEE2018_English.pdf',
                                                    'http://www.vit.ac.in/campus/hostels',
                                                    'http://www.vit.ac.in/btechadmissions/viteee2017',
                                                    'http://www.vit.ac.in/about/leadership',
                                                    'https://vtopbeta.vit.ac.in/vtop/',
                                                    'http://info.vit.ac.in/map/',
                                                    'http://www.vit.ac.in/events/eventView/Art '
                                                    'Festival 2018',
                                                    'http://www.vit.ac.in/files/admissions/Hostel_Affidavit_Mens_2018.pdf',
                                                    'http://intranet.vit.ac.in',
                                                    'http://www.vit.ac.in/files/wifiservices.pdf',
```

## Output (rev_pagetable.txt):

```
                                        'http://www.vitaa.org/',
                                        'http://www.vit.ac.in/academics/ffcs'],
'http://www.vit.ac.in/campus/studentchapters/creationLabs': ['http://www.vit.ac.in/files/admissions/viteee/
VITEEE2018_Mathematics.pdf',
                                                             'http://www.vit.ac.in/academics/schools',
                                                             'http://www.vit.ac.in/admissions/international',
                                                             'http://www.vit.ac.in/files/FormatGuidelines.doc',
                                                             'https://webmail.vit.ac.in/',
                                                             'http://careers.vit.ac.in/careers/',
                                                             'http://www.vit.ac.in/files/admissions/viteee/
                                                             VITEEE2018_Chemistry.pdf',
                                                             'http://www.vit.ac.in/academics/internationalRelations',
                                                             'http://www.vit.ac.in/events/eventView/Two '
                                                             'Day Workshop on '
                                                             'Raspberry Pi '
                                                             'Programming for '
                                                             'Beginners 2018',
                                                             'http://www.vit.ac.in/research/sponsoredResearch',
                                                             'http://www.vit.ac.in/admissions/postoffices',
                                                             'http://careers.vit.ac.in/',
                                                             'http://www.vit.ac.in/research/academic',
                                                             'http://www.vit.ac.in/campus/studentchapters',
                                                             'http://www.vit.ac.in/campus/sae',
                                                             'http://www.vit.ac.in/about/raac',
                                                             'http://www.vit.ac.in/campus/ncc',
                                                             'http://www.vit.ac.in/research',
                                                             'http://www.vit.ac.in/about/news',
                                                             'http://www.vit.ac.in/placement/pat',
                                                             'https://peopleorbit.vit.ac.in/',
                                                             'http://www.vit.ac.in/iprcell',
                                                             'http://www.vit.ac.in/placement/advancedTraining',
                                                             'http://www.vit.ac.in/files/wifiservices.pdf',
                                                             'http://www.vit.ac.in/admissions/pg',
                                                             'http://www.vit.ac.in/placement/consortium',
                                                             'http://www.vit.ac.in/research/centers',
                                                             'http://www.vit.ac.in/academics/coe',
                                                             'http://www.vit.ac.in/btechadmissions/viteee2017'.
```

```
                                                            'http://www.vit.ac.in/academics/ffcs'],
'http://www.vit.ac.in/campus/studentclubs': ['http://www.vit.ac.in/files/admissions/viteee/VITEEE2018_Mathematics.pdf',
                                             'http://www.vit.ac.in/academics/schools',
                                             'http://www.vit.ac.in/admissions/international',
                                             'http://www.vit.ac.in/files/FormatGuidelines.doc',
                                             'https://webmail.vit.ac.in/',
                                             'http://careers.vit.ac.in/careers/',
                                             'http://www.vit.ac.in/files/admissions/viteee/VITEEE2018_Chemistry.pdf',
                                             'http://www.vit.ac.in/academics/internationalRelations',
                                             'http://www.vit.ac.in/events/eventView/Two '
                                             'Day Workshop on Raspberry Pi '
                                             'Programming for Beginners 2018',
                                             'http://www.vit.ac.in/research/sponsoredResearch',
                                             'http://www.vit.ac.in/admissions/postoffices',
                                             'http://careers.vit.ac.in/',
                                             'http://www.vit.ac.in/research/academic',
                                             'http://www.vit.ac.in/campus/studentchapters',
                                             'http://www.vit.ac.in/campus/sae',
                                             'http://www.vit.ac.in/about/raac',
                                             'http://www.vit.ac.in/campus/ncc',
                                             'http://www.vit.ac.in/research',
                                             'http://www.vit.ac.in/about/news',
                                             'http://www.vit.ac.in/placement/pat',
                                             'https://peopleorbit.vit.ac.in/',
                                             'http://www.vit.ac.in/iprcell',
                                             'http://www.vit.ac.in/placement/advancedTraining',
                                             'http://www.vit.ac.in/files/wifiservices.pdf',
                                             'http://www.vit.ac.in/admissions/pg',
                                             'http://www.vit.ac.in/placement/consortium',
                                             'http://www.vit.ac.in/research/centers',
                                             'http://www.vit.ac.in/academics/coe',
                                             'http://www.vit.ac.in/btechadmissions/viteee2017',
                                             'http://www.vit.ac.in/campus',
                                             'http://www.vit.ac.in/placement/dreamoffers',
                                             'https://vtop9.vit.ac.in/vtop/login/freshers',
                                                                                    Spaces: 2
```

**Output (frontier.txt):**

```
['http://www.vit.ac.in',
 'http://www.vit.ac.in/files/admissions/viteee/VITEEE2018_Mathematics.pdf',
 'http://www.vit.ac.in/academics/schools',
 'http://www.vit.ac.in/admissions/international',
 'http://www.vit.ac.in/files/FormatGuidelines.doc',
 'http://www.vit.ac.in/files/admissions/viteee/VITEEE2018_Chemistry.pdf',
 'http://www.vit.ac.in/academics/internationalRelations',
 'http://www.vit.ac.in/events/eventView/Two Day Workshop on Raspberry Pi '
 'Programming for Beginners 2018',
 'http://www.vit.ac.in/research/sponsoredResearch',
 'http://www.vit.ac.in/admissions/postoffices',
 'http://www.vit.ac.in/research/academic',
 'http://www.vit.ac.in/campus/studentchapters',
 'http://www.vit.ac.in/campus/sae',
 'http://www.vit.ac.in/about/raac',
 'http://www.vit.ac.in/campus/ncc',
 'http://www.vit.ac.in/research',
 'http://www.vit.ac.in/about/news',
 'http://www.vit.ac.in/placement/pat',
 'http://www.vit.ac.in/iprcell',
 'http://www.vit.ac.in/placement/advancedTraining',
 'http://www.vit.ac.in/files/wifiservices.pdf',
 'http://www.vit.ac.in/admissions/pg',
 'http://www.vit.ac.in/placement/consortium',
 'http://www.vit.ac.in/research/centers',
 'http://www.vit.ac.in/academics/coe',
 'http://www.vit.ac.in/btechadmissions/viteee2017',
 'http://www.vit.ac.in/campus',
 'http://www.vit.ac.in/placement/dreamoffers',
 'http://www.vit.ac.in/academics/centers',
```

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

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

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

'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/academics',

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

'http://www.vit.ac.in/files/PGAdmissions2018/index.html',

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

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

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

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

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

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

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

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

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

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

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

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

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

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

'http://www.vit.ac.in/files/admissions/viteee/VITEEE2018_Physics.pdf',

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

'http://www.vit.ac.in/eventslist?status=publish',

'http://www.vit.ac.in/programmes-offered',

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

'http://www.vit.ac.in/visual-media/',

'http://www.vit.ac.in/campus-life/clubs-chapters/',

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

'http://www.vit.ac.in#carousel-example-vertical',

'http://www.vit.ac.in/vit-ap-advantage/#vitutp',

'http://www.vit.ac.in/vit-ap-advantage/#vitco-op',

'http://www.vit.ac.in/vit-ap-advantage/#vitsp',

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

'http://www.vit.ac.in/vit-ap-advantage/#vitqf',

'http://www.vit.ac.in/international-collaborations/',

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

'http://www.vit.ac.in/vit-ap-advantage/#vitms',

'http://www.vit.ac.in/academics/engineering-clinics/',

'http://www.vit.ac.in/eventslist?status=future',

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

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

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

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

'http://www.vit.ac.in#portfolio-modal-5',

'http://www.vit.ac.in#portfolio-modal-2',

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

'http://www.vit.ac.in#portfolio-modal-4',

'http://www.vit.ac.in/files/vsparc/B.Arch_Information_Brochure-2018.pdf',

'http://www.vit.ac.in/files/llb/Law_Info_2018.pdf',

'http://www.vit.ac.in/files/VFIT_Information_Brochure_2018.pdf',

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

'http://www.vit.ac.in#portfolio-modal-6',

'http://www.vit.ac.in#portfolio-modal-3',

'http://www.vit.ac.in#page-top',

'http://www.vit.ac.in#portfolio-modal-1',

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

'http://www.vit.ac.in/files/ug-2018/UG-2018_brochure.pdf']