Web Mining Lab – 3

Indexing data from files and web pages

By Abhijeet Ambadekar (16BCE1156)

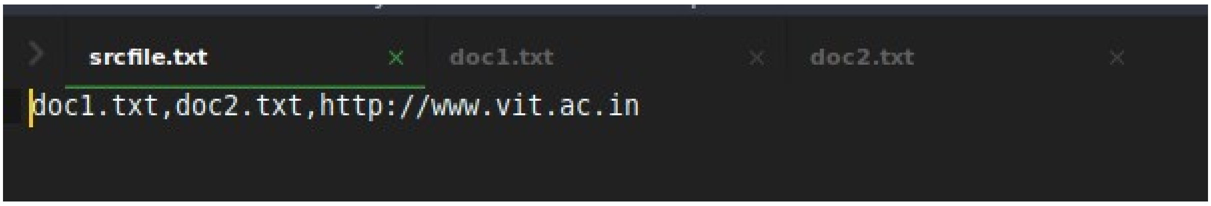
The aim of this experiment is to create a program that indexes all the words from multiple sources (can be a local file or a wepage url) into an index table that maintains the frequency of the words along with th offset at which the words are present in each source.

Program Code:

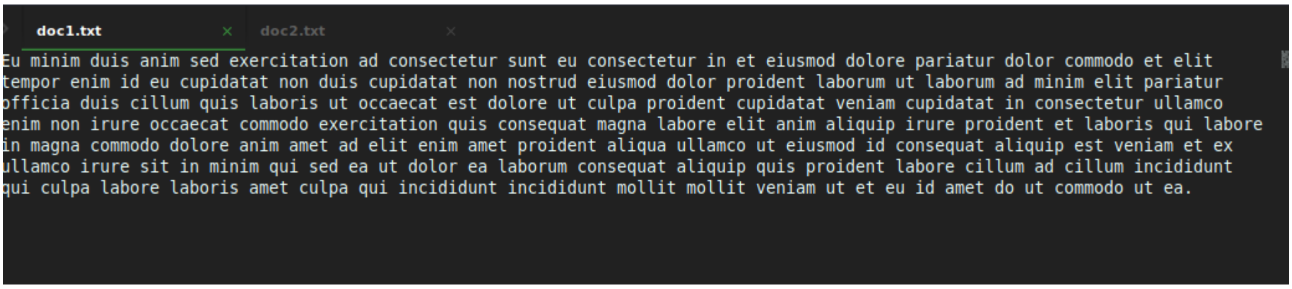
```
import requests
from bs4 import BeautifulSoup

import string
import re
from pprint import pprint

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


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


class Indexer():
      def __init__(self):
            self.index_list = {}
            self.data_dict = {}

      def get_words(self, text):
            text = re.sub(r'[^\w\s]','',text)
            return re.split(r'\s*', text)

      def get_data_from_file(self, fname):
            try:
                  with open(fname) as file:
                        self.data_dict[fname] = 
self.get_words(file.read().lower())
            except IOError:
                  print("File does not exist!")
```

```
    def get_data_from_url(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')
        self.data_dict[url] = self.get_words(soup.text.lower())

    def add_word_to_index(self, off, fname, word):
        if word not in self.index_list:
            self.index_list[word] = [1, [(fname, off)]]
        else:
            self.index_list[word][0] += 1
            self.index_list[word][1].append((fname, off))

    def index_data(self):
        for fname in self.data_dict.keys():
            print("Indexing {} words from the source: 
{}".format(len(self.data_dict[fname]), fname))
            for i in range(len(self.data_dict[fname])):
                self.add_word_to_index(i, fname, self.data_dict[fname]
[i])


    def get_data(self, srcfname):
        try:
            with open(srcfname, 'r') as infile:
                srcs = re.split(r',', infile.read())
        except IOError:
            print("Source file not found.")

        for src in srcs:
            if urlparse(src).scheme == 'http':
                self.get_data_from_url(src)
            else:
                self.get_data_from_file(src)


viti = Indexer()
viti.get_data("srcfile.txt")
viti.index_data()

pprint_data_to_file(viti.index_list, 'index.idx')
```

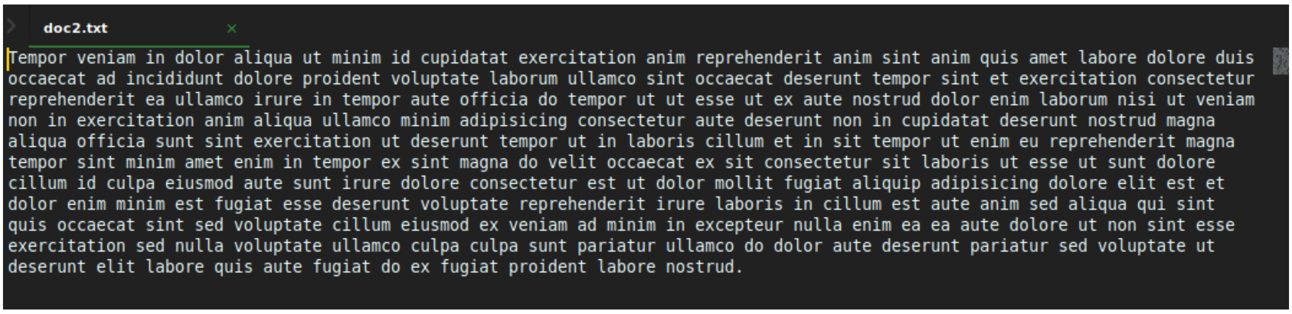
Sourcefiles used:



```
doc1.txt,doc2.txt,http://www.vit.ac.in
```

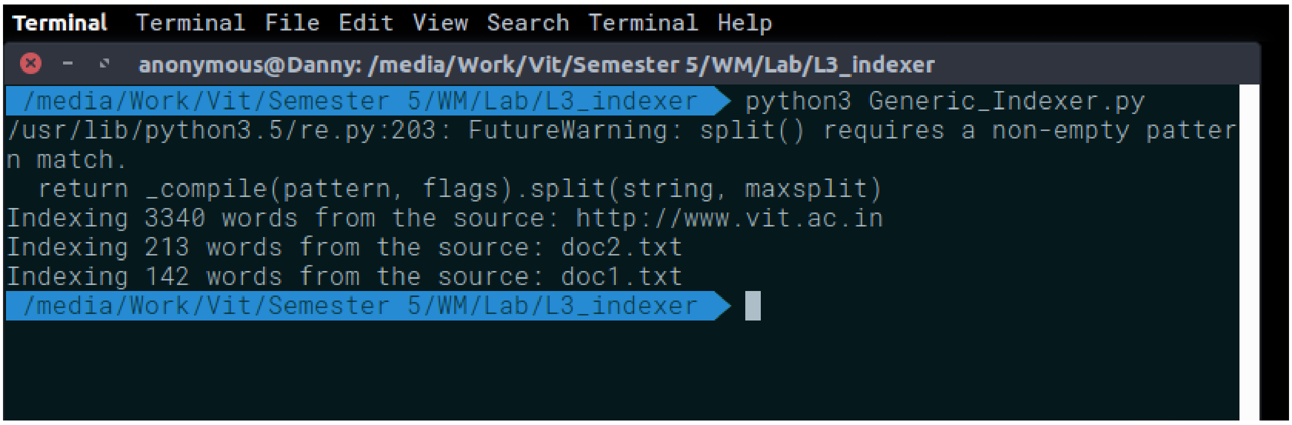


```
Eu minim duis anim sed exercitation ad consectetur sunt eu consectetur in et eiusmod dolore pariatur dolor commodo et elit
tempor enim id eu cupidatat non duis cupidatat non nostrud eiusmod dolor proident laborum ut laborum ad minim elit pariatur
officia duis cillum quis laboris ut occaecat est dolore ut culpa proident cupidatat veniam cupidatat in consectetur ullamco
enim non irure occaecat commodo exercitation quis consequat magna labore elit anim aliquip irure proident et laboris qui labore
in magna commodo dolore anim amet ad elit enim amet proident aliqua ullamco ut eiusmod id consequat aliquip est veniam et ex
ullamco irure sit in minim qui sed ea ut dolor ea laborum consequat aliquip quis proident labore cillum ad cillum incididunt
qui culpa labore laboris amet culpa qui incididunt incididunt mollit mollit veniam ut et eu id amet do ut commodo ut ea.
```



```
Tempor veniam in dolor aliqua ut minim id cupidatat exercitation anim reprehenderit anim sint anim quis amet labore dolore duis
occaecat ad incididunt dolore proident voluptate laborum ullamco sint occaecat deserunt tempor sint et exercitation consectetur
reprehenderit ea ullamco irure in tempor aute officia do tempor ut ut esse ut ex aute nostrud dolor enim laborum nisi ut veniam
non in exercitation anim aliqua ullamco minim adipisicing consectetur aute deserunt non in cupidatat deserunt nostrud magna
aliqua officia sunt sint exercitation ut deserunt tempor ut in laboris cillum et in sit tempor ut enim eu reprehenderit magna
tempor sint minim amet enim in tempor ex sint magna do velit occaecat ex sit consectetur sit laboris ut esse ut sunt dolore
cillum id culpa eiusmod aute sunt irure dolore consectetur est ut dolor mollit fugiat aliquip adipisicing dolore elit est et
dolor enim minim est fugiat esse deserunt voluptate reprehenderit irure laboris in cillum est aute anim sed aliqua qui sint
quis occaecat sint sed voluptate cillum eiusmod ex veniam ad minim in excepteur nulla enim ea ea aute dolore ut non sint esse
exercitation sed nulla voluptate ullamco culpa culpa sunt pariatur ullamco do dolor aute deserunt pariatur sed voluptate ut
deserunt elit labore quis aute fugiat do ex fugiat proident labore nostrud.
```

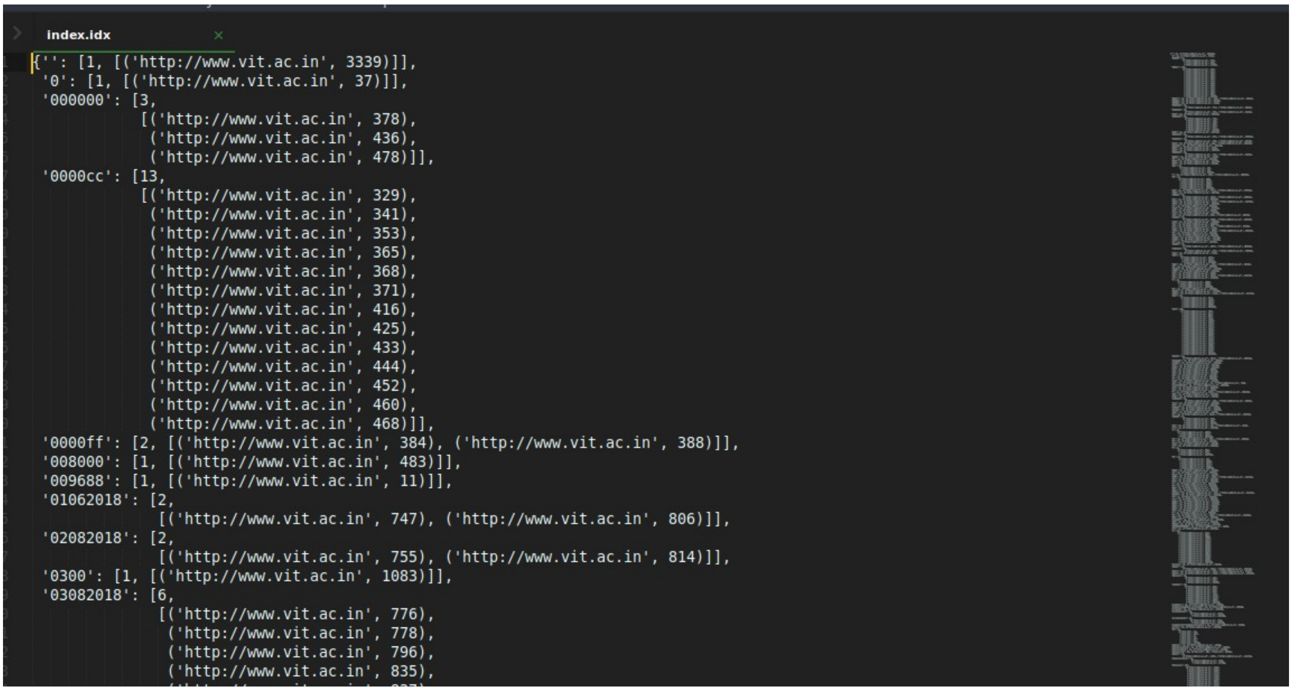
Output:


Terminal
Terminal File Edit View Search Terminal Help
anonymous@Danny: /media/Work/Vit/Semester 5/WM/Lab/L3_indexer

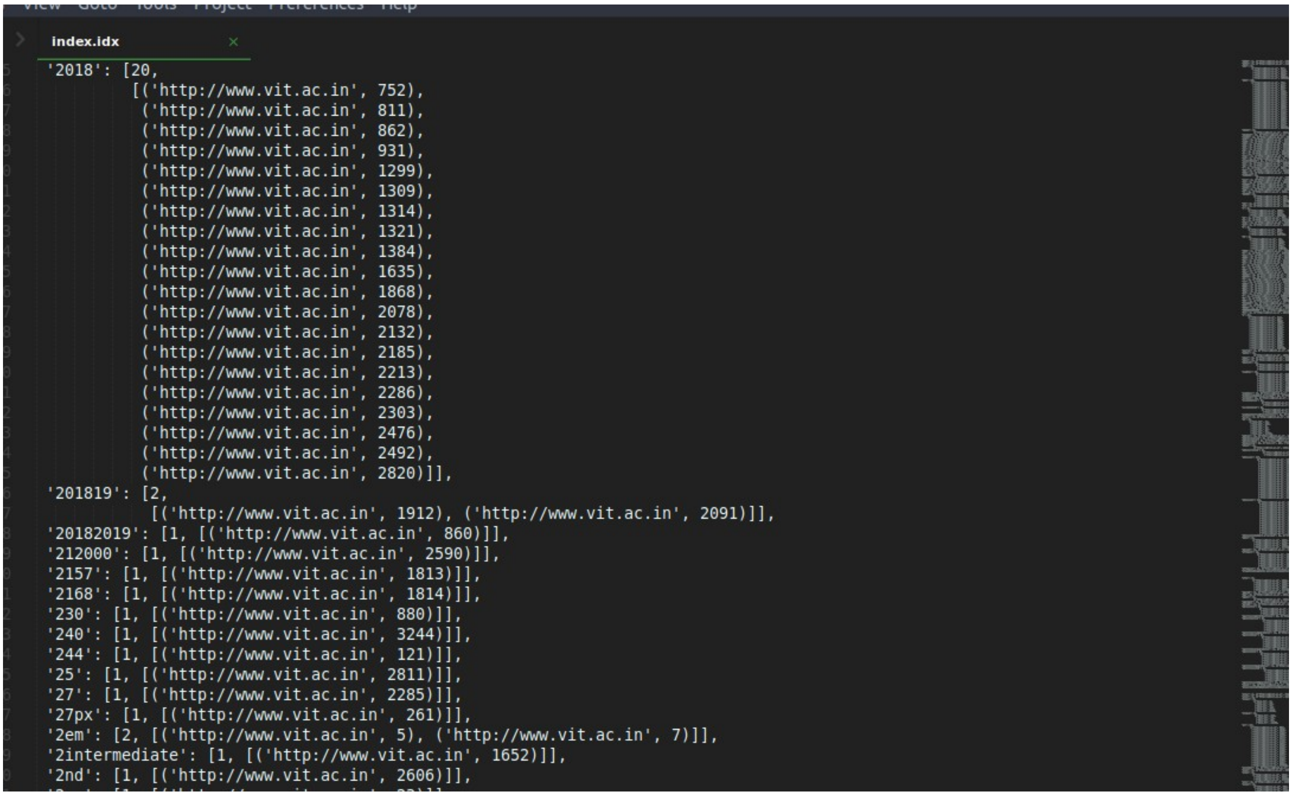
```
 /media/Work/Vit/Semester 5/WM/Lab/L3_indexer  python3 Generic_Indexer.py
/usr/lib/python3.5/re.py:203: FutureWarning: split() requires a non-empty patter
n match.
  return _compile(pattern, flags).split(string, maxsplit)
Indexing 3340 words from the source: http://www.vit.ac.in
Indexing 213 words from the source: doc2.txt
Indexing 142 words from the source: doc1.txt
 /media/Work/Vit/Semester 5/WM/Lab/L3_indexer
```

Output (index.idx):

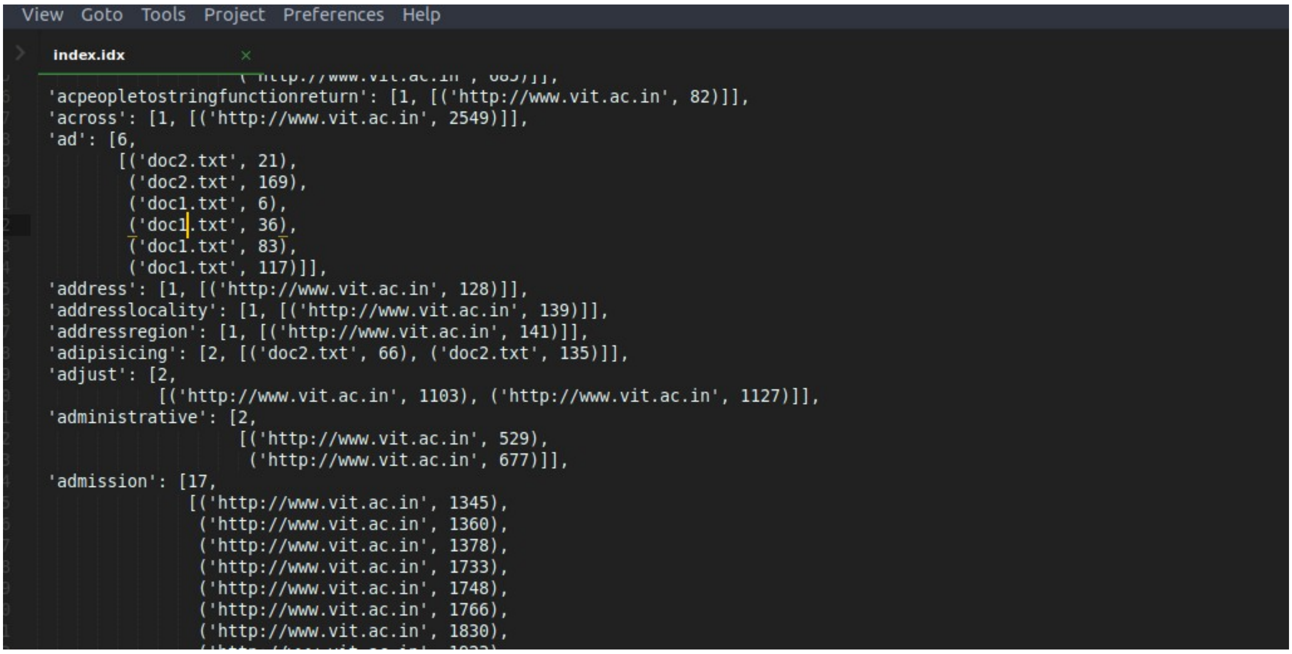
//Since the file has an output of 4000 lines, noly screenshots of some of the output is shown.



```
{'': [1, [('http://www.vit.ac.in', 3339)]],
 '0': [1, [('http://www.vit.ac.in', 37)]],
 '000000': [3,
            [('http://www.vit.ac.in', 378),
             ('http://www.vit.ac.in', 436),
             ('http://www.vit.ac.in', 478)]],
 '0000cc': [13,
            [('http://www.vit.ac.in', 329),
             ('http://www.vit.ac.in', 341),
             ('http://www.vit.ac.in', 353),
             ('http://www.vit.ac.in', 365),
             ('http://www.vit.ac.in', 368),
             ('http://www.vit.ac.in', 371),
             ('http://www.vit.ac.in', 416),
             ('http://www.vit.ac.in', 425),
             ('http://www.vit.ac.in', 433),
             ('http://www.vit.ac.in', 444),
             ('http://www.vit.ac.in', 452),
             ('http://www.vit.ac.in', 460),
             ('http://www.vit.ac.in', 468)]],
 '0000ff': [2, [('http://www.vit.ac.in', 384), ('http://www.vit.ac.in', 388)]],
 '008000': [1, [('http://www.vit.ac.in', 483)]],
 '009688': [1, [('http://www.vit.ac.in', 11)]],
 '01062018': [2,
              [('http://www.vit.ac.in', 747), ('http://www.vit.ac.in', 806)]],
 '02082018': [2,
              [('http://www.vit.ac.in', 755), ('http://www.vit.ac.in', 814)]],
 '0300': [1, [('http://www.vit.ac.in', 1083)]],
 '03082018': [6,
              [('http://www.vit.ac.in', 776),
               ('http://www.vit.ac.in', 778),
               ('http://www.vit.ac.in', 796),
               ('http://www.vit.ac.in', 835),
```



```
 '2018': [20,
          [('http://www.vit.ac.in', 752),
           ('http://www.vit.ac.in', 811),
           ('http://www.vit.ac.in', 862),
           ('http://www.vit.ac.in', 931),
           ('http://www.vit.ac.in', 1299),
           ('http://www.vit.ac.in', 1309),
           ('http://www.vit.ac.in', 1314),
           ('http://www.vit.ac.in', 1321),
           ('http://www.vit.ac.in', 1384),
           ('http://www.vit.ac.in', 1635),
           ('http://www.vit.ac.in', 1868),
           ('http://www.vit.ac.in', 2078),
           ('http://www.vit.ac.in', 2132),
           ('http://www.vit.ac.in', 2185),
           ('http://www.vit.ac.in', 2213),
           ('http://www.vit.ac.in', 2286),
           ('http://www.vit.ac.in', 2303),
           ('http://www.vit.ac.in', 2476),
           ('http://www.vit.ac.in', 2492),
           ('http://www.vit.ac.in', 2820)]],
 '201819': [2,
            [('http://www.vit.ac.in', 1912), ('http://www.vit.ac.in', 2091)]],
 '20182019': [1, [('http://www.vit.ac.in', 860)]],
 '212000': [1, [('http://www.vit.ac.in', 2590)]],
 '2157': [1, [('http://www.vit.ac.in', 1813)]],
 '2168': [1, [('http://www.vit.ac.in', 1814)]],
 '230': [1, [('http://www.vit.ac.in', 880)]],
 '240': [1, [('http://www.vit.ac.in', 3244)]],
 '244': [1, [('http://www.vit.ac.in', 121)]],
 '25': [1, [('http://www.vit.ac.in', 2811)]],
 '27': [1, [('http://www.vit.ac.in', 2285)]],
 '27px': [1, [('http://www.vit.ac.in', 261)]],
 '2em': [2, [('http://www.vit.ac.in', 5), ('http://www.vit.ac.in', 7)]],
 '2intermediate': [1, [('http://www.vit.ac.in', 1652)]],
 '2nd': [1, [('http://www.vit.ac.in', 2606)]],
```



```
                    ('http://www.vit.ac.in', 685)]],
'acpeopletostringfunctionreturn': [1, [('http://www.vit.ac.in', 82)]],
'across': [1, [('http://www.vit.ac.in', 2549)]],
'ad': [6,
       [('doc2.txt', 21),
        ('doc2.txt', 169),
        ('doc1.txt', 6),
        ('doc1.txt', 36),
        ('doc1.txt', 83),
        ('doc1.txt', 117)]],
'address': [1, [('http://www.vit.ac.in', 128)]],
'addresslocality': [1, [('http://www.vit.ac.in', 139)]],
'addressregion': [1, [('http://www.vit.ac.in', 141)]],
'adipisicing': [2, [('doc2.txt', 66), ('doc2.txt', 135)]],
'adjust': [2,
           [('http://www.vit.ac.in', 1103), ('http://www.vit.ac.in', 1127)]],
'administrative': [2,
                   [('http://www.vit.ac.in', 529),
                    ('http://www.vit.ac.in', 677)]],
'admission': [17,
              [('http://www.vit.ac.in', 1345),
               ('http://www.vit.ac.in', 1360),
               ('http://www.vit.ac.in', 1378),
               ('http://www.vit.ac.in', 1733),
               ('http://www.vit.ac.in', 1748),
               ('http://www.vit.ac.in', 1766),
               ('http://www.vit.ac.in', 1830),
               ('http://www.vit.ac.in', 1833),
```