BROADWAY INFOSYS NEPAL

TINKUNE, KATHMANDU

A PROJECT ON PYTHON PROGRAMMING

GROUP MEMBERS:

1. Samundra Dahal
2. Kiran Neupane
3. Bidhan Khatri
4. Pranita Karki
5. Sangay Yuten Lama

First of all, we want to thank our sir Anish Chapagain sir for giving us the basic and creative concepts of python programming.On the guidance of our sir we have prepared a project on web scrapping and file handling of the given website on python.

INTRODUCTION:

On our project we have done scrapping on the website http://www.aiovision.org from where we have scrapped the members of that website using scrapping packages. We have scrapped the title of organizations with its address, zip code, telephone, website and gmail of that organization and save to the csv file using python. The tools and packages used on this process are given below:

1. **BEAUTIFUL SOUP:**

There are many scrapping packages in python. Among them we have used Beautiful Soup for scrapping. Beautiful Soup provides a few simple methods and Pythonic idioms for navigating, searching, and modifying a parse tree: a toolkit for dissecting a document and extracting what you need. It doesn't take much code to write an application. Beautiful Soup automatically converts incoming documents to Unicode and outgoing documents to UTF-8. You don't have to think about encodings; unless the document doesn't specify an encoding and Beautiful Soup can't auto detect one. Then you just have to specify the original encoding. Beautiful Soup sits on top of popular Python parsers like lxml and html5lib.So, in this project Beautiful Soup is used as scrapping tools.

2. **REGEX:**

In this project we have use regex to determine the exact details of organizations like telephone, website, zip code and so on. Regular expressions (called REs, or regexes, or regex patterns) are essentially a tiny, highly specialized programming language embedded inside Python and made available through the re module. Using this little language, you specify the rules for the set of possible strings that you want to match; this set might contain English sentences, or e-mail addresses, or TeX commands, or anything you like. Regular expression patterns are compiled into a series of byte codes which are then executed by a matching engine written in C. For advanced use, it may be necessary to pay careful attention to how the engine will execute a given RE, and write the RE in a certain way in order to produce byte code that runs faster. Optimization isn't covered in this document, because it requires that you have a good understanding of the matching engine's internals.

3. FOR LOOP:

For loop is used for iterating over a sequence (that is either a list, a tuple or a string).This is less like the for keyword in other programming language, and works more like an iterator method as found in other object-orientated programming languages.

4. CSV FILE:

Python has a vast library of modules that are included with its distribution. The csv module gives the Python programmer the ability to parse CSV (Comma Separated Values) files. A CSV file is a human readable text file where each line has a number of fields, separated by commas or some other delimiter. You can think of each line as a row and each field as a column. The CSV format has no standard, but they are similar enough that the csv module will be able to read the vast majority of CSV files. You can also write CSV files using the csv module.

SOURCE CODE:

```python
import urllib
import urllib.request as u
import re
from bs4 import BeautifulSoup
import csv
filename = "Scrapping.csv"
with open(filename, 'w+') as f:
    writer = csv.writer(f)
    for i in list(range(1,12)):
        url="http://www.aiovision.org/members-list/page/"+str(i)+"/"
        page= u.urlopen(url)
        soup=BeautifulSoup(page,"html.parser")
        containers=soup.findAll("div",{"class":"one_half"})
        for container in containers:
            gmail=''
            website=''
            telephone=''
            zipcode=''
            address=''
            title=(str(container.div.h2.a["title"]))
            details=list((container.div.div.div.text.split("\n")))
            for detail in details[::-1]:
                if "@" in detail:
                    s=details.index(detail)
                    gmail=(detail)
                    del details[s:-1]
                elif "co.uk" in detail or "www." in detail:
                    website=(detail)
                    s = details.index(detail)
                    del details[s:-1]
                elif "Tel" in detail   or 'TEL' in detail:
                    telephone=detail[4:-1]
                    s = details.index(detail)
                    del details[s:-1]
            for detail in details[::-1]:
                if str(re.findall("[0-9]{5,}",detail)) in detail or "02" in detail or
"01" in detail :
                    telephone=detail
                    s = details.index(telephone)
```

```
                 del details[s:-1]
         try:
             zipcode=details[-2]
             details.__delitem__(-2)
         except:
             print("")
         address=" ".join(details)
         row=[title,address,zipcode,telephone,website,gmail]
         writer.writerow(row)
```
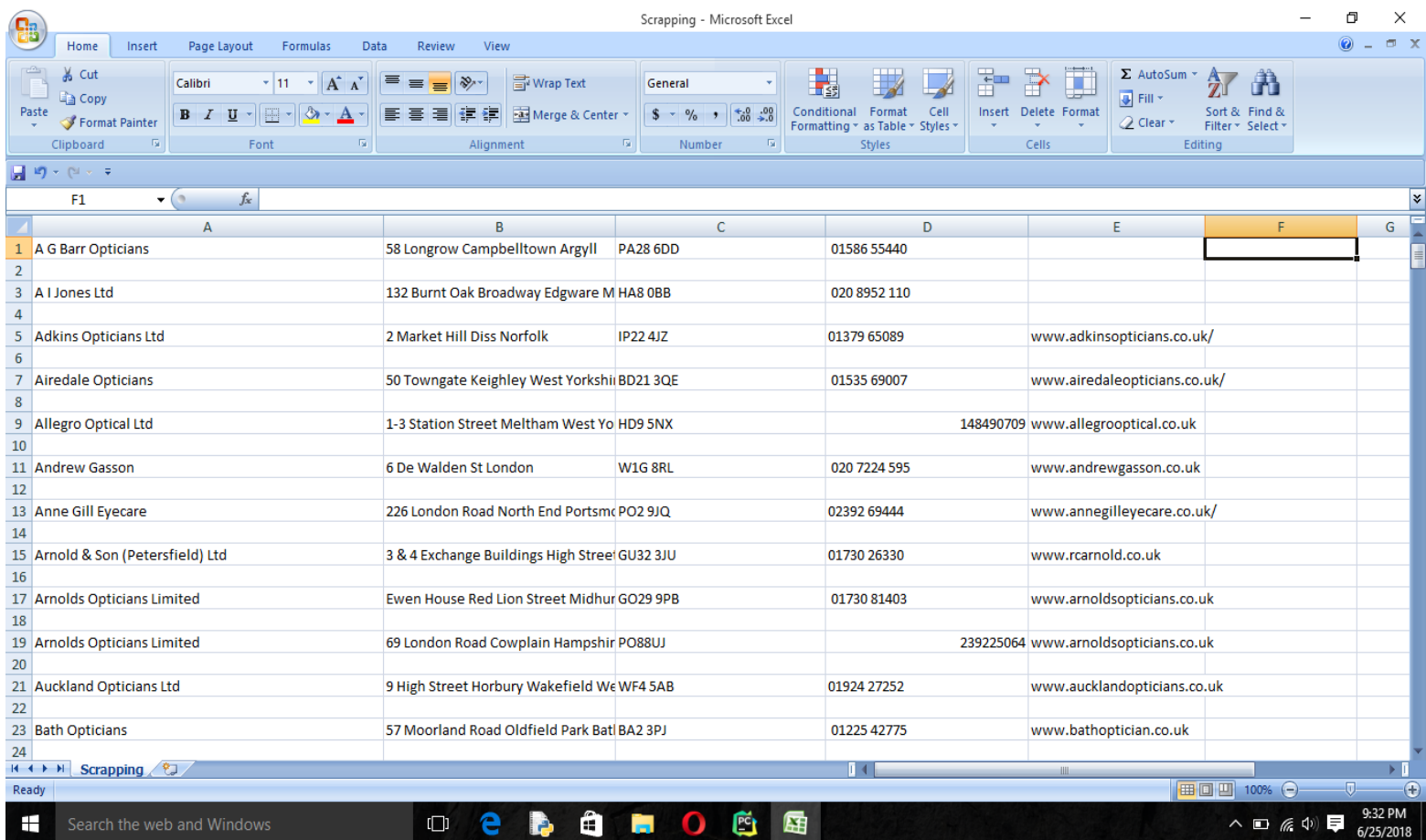
File of source is attached in this documentation.

OUTPUT:

The output in csv file looks like this:

	A	B	C	D	E
1	A G Barr Opticians	58 Longrow Campbelltown Argyll	PA28 6DD	01586 55440	
2					
3	A I Jones Ltd	132 Burnt Oak Broadway Edgware M	HA8 0BB	020 8952 110	
4					
5	Adkins Opticians Ltd	2 Market Hill Diss Norfolk	IP22 4JZ	01379 65089	www.adkinsopticians.co.uk/
6					
7	Airedale Opticians	50 Towngate Keighley West Yorkshi	BD21 3QE	01535 69007	www.airedaleopticians.co.uk/
8					
9	Allegro Optical Ltd	1-3 Station Street Meltham West Yo	HD9 5NX	148490709	www.allegrooptical.co.uk
10					
11	Andrew Gasson	6 De Walden St London	W1G 8RL	020 7224 595	www.andrewgasson.co.uk
12					
13	Anne Gill Eyecare	226 London Road North End Portsmc	PO2 9JQ	02392 69444	www.annegilleyecare.co.uk/
14					
15	Arnold & Son (Petersfield) Ltd	3 & 4 Exchange Buildings High Stree	GU32 3JU	01730 26330	www.rcarnold.co.uk
16					
17	Arnolds Opticians Limited	Ewen House Red Lion Street Midhur	GO29 9PB	01730 81403	www.arnoldsopticians.co.uk
18					
19	Arnolds Opticians Limited	69 London Road Cowplain Hampshir	PO88UJ	239225064	www.arnoldsopticians.co.uk
20					
21	Auckland Opticians Ltd	9 High Street Horbury Wakefield We	WF4 5AB	01924 27252	www.aucklandopticians.co.uk
22					
23	Bath Opticians	57 Moorland Road Oldfield Park Batl	BA2 3PJ	01225 42775	www.bathoptician.co.uk
24					

	A	B	C	D	E	F	G
379	Stephen Wilcox Optometrists	169 Dunbabin Road Liverpool	L16 7QQ	0151 722 118	www.wilcox-optometry.co.uk		
380							
381	Stevenson Jones Opticians	6 Friars Street Newcastle-under-Lyn	ST5 2DZ	1782619175	www.stevensonjonesopticians.co.uk		
382							
383	Stewart Greenberg Opticians	52 High St Llandaff Cardiff	CF5 2DZ	02920 56611	http://www.greenberg.co.uk/		
384							
385	Susan Bowers Optometrists	349 Tile Hill Lane Coventry	CV4 9DU	02476 466 661	www.susanbowers.co.uk		
386							
387	Taylor Biddle Opticians Ltd	39 Bridgnorth Road Compton Wolve	WV6 8AF	01902 89590			
388							
389	The Eye Studio	The Eye Studio 229a Chiswick High R	W4 2DW	020 8742 853	http://www.eyestudio.co.uk		
390							
391	The Eyecare Studio	439 Uxbridge Road Hatch End Pinner	HA5 4JR	020 8421 227			
392							
393	The Optical People- Millington Dawson	6 St. Mary Street Chepstow Sir Fynw	NP16 5EU	01291 62226			
394							
395	The Optical Shop	433 Green Lanes Harringay London	N4 1HA	020 8340 151			
396							
397	The Vision Therapy Clinic	28 The Spain Petersfield Hampshire	GU32 3LA	01730 71017	www.owenleigh.com/		
398							
399	Tombs and Allen Opticians	5 Portland Buildings, Stoke Road, Go	PO12 1JH	02392 58375	www.tombsandallen.com		
400							
401	Vaid Optometrists	132 Station Road Edgware Middlese	HA8 7AA	020 8381 264	www.vaidoptometrists.com		
402							