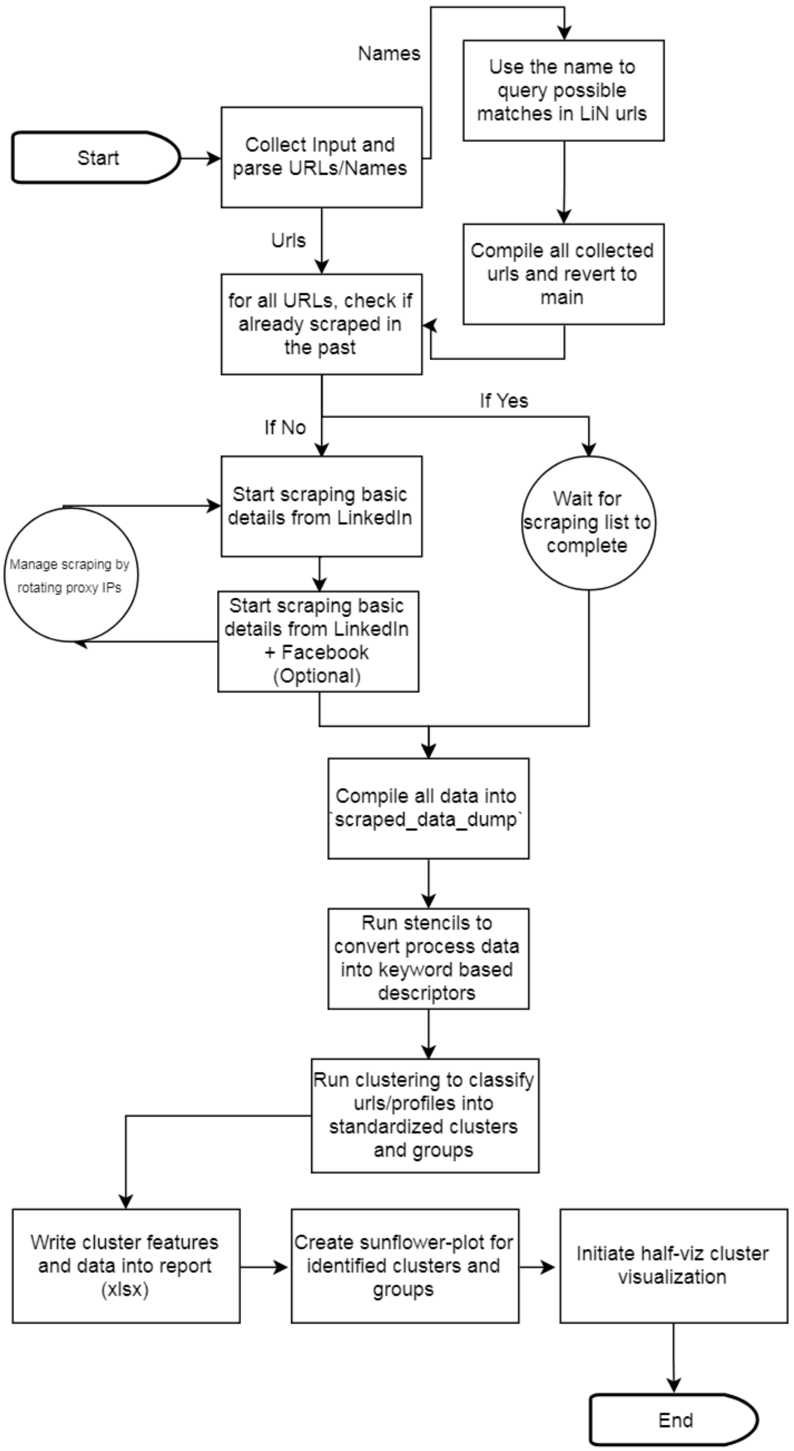
protoSCAP – An integrated scraping and data-visualization tool.

Index

Process Flow


Start
Collect Input and parse URLs/Names
Names
Use the name to query possible matches in LiN urls
Compile all collected urls and revert to main
Urls
for all URLs, check if already scraped in the past
If No
If Yes
Start scraping basic details from LinkedIn
Manage scraping by rotating proxy IPs
Start scraping basic details from LinkedIn + Facebook (Optional)
Wait for scraping list to complete
Compile all data into `scraped_data_dump`
Run stencils to convert process data into keyword based descriptors
Run clustering to classify urls/profiles into standardized clusters and groups
Write cluster features and data into report (xlsx)
Create sunflower-plot for identified clusters and groups
Initiate half-viz cluster visualization
End

Key Component Programs & Controllers

controller.py

This is the main process that controls and orchestrates the other component functions. The main program uses the controller to check existing data and proceed/manage scraping process accordingly.

geckodriver

geckodriver in firefox's native browser available as a binary. The selenium library communicates with geckoriver to load urls and its respective doms to capture data from the HTTP/HTTPS

stencils.pyc

stencils is a (very) rudimentary natural language processing tool written specifically for protoSCAP. This program captures sentences and strategically converts the data into descriptors.

cluster.py

The clustering algorithm uses descriptors identified for each url and cluster the profiles accordingly. The program uses basic details and scraped/identified interests to form multi-dimensional clusters.

plotter.py

The plotter draws sunflower plots of identified clusters and writes in into the final report. The program also invokes halfviz visualization program for cluster/attribute visualization.

Software requirements

- Python2.7.x

Python modules

- Selenium (for python)
- Parsel
- Bs4 (BeautifulSoup)
- Requests
- Xlsxwriter

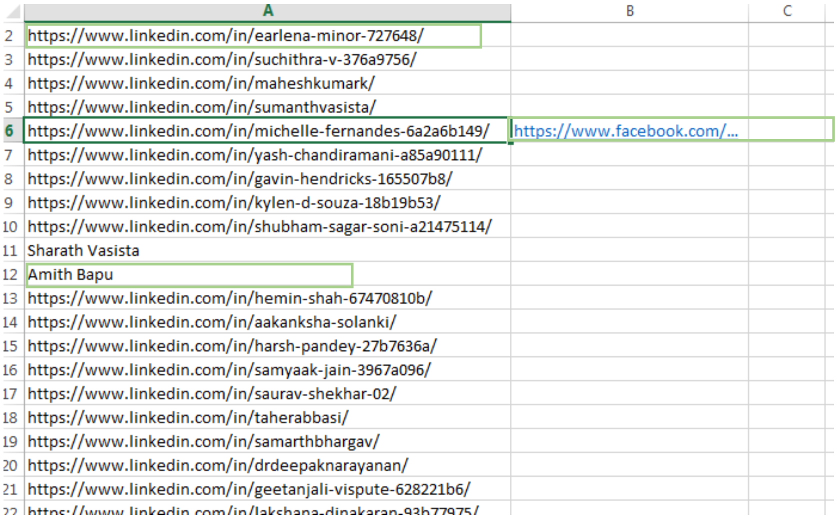
Browser modules

- Firefox 66.+
- Geckodriver v0.24

Installation and execution

Download the latest source files from the location – https://github.com/SMVasista/protoSCAP. Ensure that all the above software requirements are satisfied. To execute a scraping process follow the steps as below:

1) Extract protoSCAP from its zipped folder (gunzip <zip folder> in linux or right-click -> Extract Here in windows)
2) Extract the geckodriver binary within the protoSCAP folder using the same process. Ensure the main.py or protoSCAP binary AND the geckodriver binary are present in the same folder/location
3) Create an input file to be passed into the program. The input file can be a simple list of urls/names – if both facebook and linkedIn URL is to be provided, use comma-separated entries.



	A	B	C
2	https://www.linkedin.com/in/earlena-minor-727648/		
3	https://www.linkedin.com/in/suchithra-v-376a9756/		
4	https://www.linkedin.com/in/maheshkumark/		
5	https://www.linkedin.com/in/sumanthvasista/		
6	https://www.linkedin.com/in/michelle-fernandes-6a2a6b149/	https://www.facebook.com/...	
7	https://www.linkedin.com/in/yash-chandiramani-a85a90111/		
8	https://www.linkedin.com/in/gavin-hendricks-165507b8/		
9	https://www.linkedin.com/in/kylen-d-souza-18b19b53/		
10	https://www.linkedin.com/in/shubham-sagar-soni-a21475114/		
11	Sharath Vasista		
12	Amith Bapu		
13	https://www.linkedin.com/in/hemin-shah-67470810b/		
14	https://www.linkedin.com/in/aakanksha-solanki/		
15	https://www.linkedin.com/in/harsh-pandey-27b7636a/		
16	https://www.linkedin.com/in/samyaak-jain-3967a096/		
17	https://www.linkedin.com/in/saurav-shekhar-02/		
18	https://www.linkedin.com/in/taherabbasi/		
19	https://www.linkedin.com/in/samarthbhargav/		
20	https://www.linkedin.com/in/drdeepaknarayanan/		
21	https://www.linkedin.com/in/geetanjali-vispute-628221b6/		
22	https://www.linkedin.com/in/lakshana-dinakaran-93b77975/		

4) To invoke the program open a terminal (windows/linux) in the current working directory/folder and type the following command.

   To run using the source files:

```
python/python.exe main.py <input-file>
```

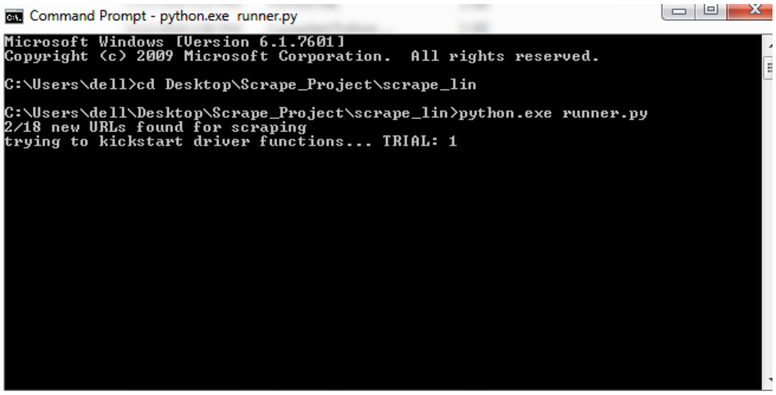
To run using the binary (.exe):

```
protoscap.exe <input-file>
```

To run using the linux binary:

```
protoscap <input-file>
```

Execution outputs

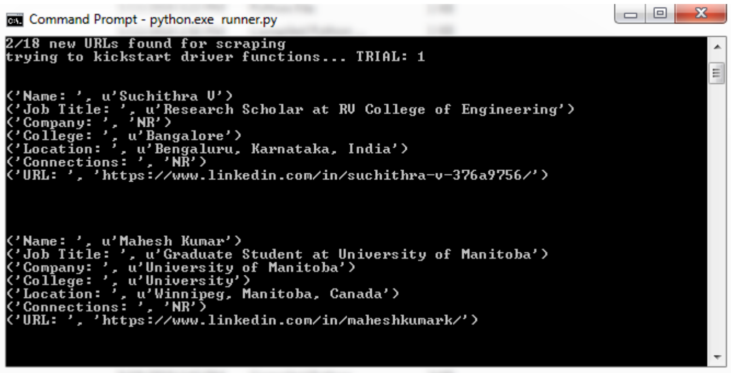


```
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation.  All rights reserved.

C:\Users\dell>cd Desktop\Scrape_Project\scrape_lin

C:\Users\dell\Desktop\Scrape_Project\scrape_lin>python.exe runner.py
2/18 new URLs found for scraping
trying to kickstart driver functions... TRIAL: 1
```

Execution begins by running unit tests on each component functions to ensure things are working properly. Once the signal is cleared, the process begins. During listing, the software classified the input list into already scraped and to be scraped. Only the new links are passed for scraping.



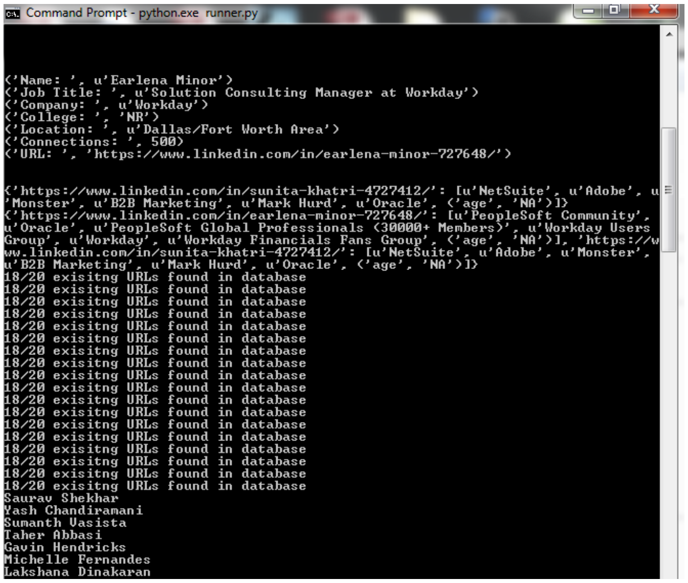
```
2/18 new URLs found for scraping
trying to kickstart driver functions... TRIAL: 1


('Name: ', u'Suchithra V')
('Job Title: ', u'Research Scholar at RV College of Engineering')
('Company: ', 'NR')
('College: ', u'Bangalore')
('Location: ', u'Bengaluru, Karnataka, India')
('Connections: ', 'NR')
('URL: ', 'https://www.linkedin.com/in/suchithra-v-376a9756/')



('Name: ', u'Mahesh Kumar')
('Job Title: ', u'Graduate Student at University of Manitoba')
('Company: ', u'University of Manitoba')
('College: ', u'University')
('Location: ', u'Winnipeg, Manitoba, Canada')
('Connections: ', 'NR')
('URL: ', 'https://www.linkedin.com/in/maheshkumark/')
```

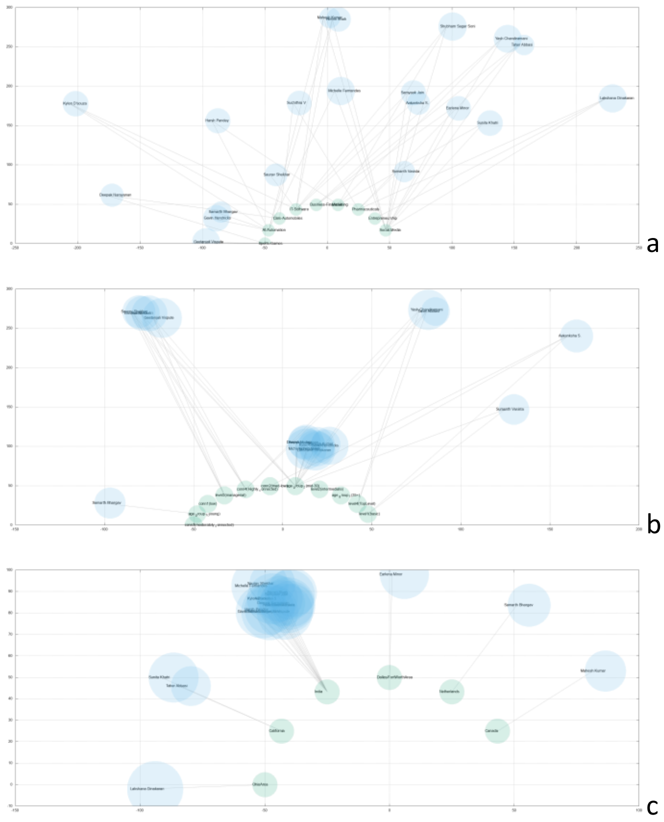
First round of scraping, the basic details (including geo-location, title/organization, connections, age etc are scraped for each profile. Interests are scraped in the second round of scraping.



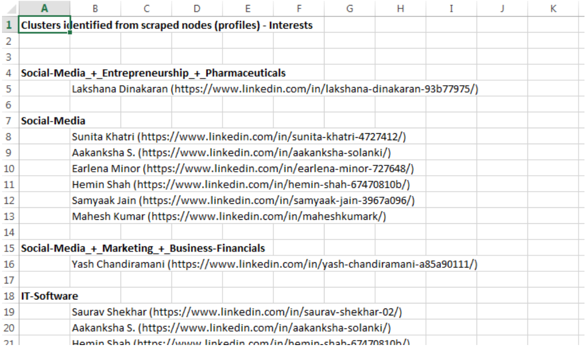
```
('Name: ', u'Earlena Minor')
('Job Title: ', u'Solution Consulting Manager at Workday')
('Company: ', u'Workday')
('College: ', 'NR')
('Location: ', u'Dallas/Fort Worth Area')
('Connections: ', 500)
('URL: ', 'https://www.linkedin.com/in/earlena-minor-727648/')


('https://www.linkedin.com/in/sunita-khatri-4727412/': [u'NetSuite', u'Adobe', u
'Monster', u'B2B Marketing', u'Mark Hurd', u'Oracle', ('age', 'NA')])
('https://www.linkedin.com/in/earlena-minor-727648/': [u'PeopleSoft Community',
u'Oracle', u'PeopleSoft Global Professionals (30000+ Members)', u'Workday Users
Group', u'Workday', u'Workday Financials Fans Group', ('age', 'NA')], 'https://w
ww.linkedin.com/in/sunita-khatri-4727412/': [u'NetSuite', u'Adobe', u'Monster',
u'B2B Marketing', u'Mark Hurd', u'Oracle', ('age', 'NA')])
18/20 exisitng URLs found in database
18/20 exisitng URLs found in database
18/20 exisitng URLs found in database
18/20 exisitng URLs found in database
18/20 exisitng URLs found in database
18/20 exisitng URLs found in database
18/20 exisitng URLs found in database
18/20 exisitng URLs found in database
18/20 exisitng URLs found in database
18/20 exisitng URLs found in database
18/20 exisitng URLs found in database
18/20 exisitng URLs found in database
18/20 exisitng URLs found in database
18/20 exisitng URLs found in database
18/20 exisitng URLs found in database
18/20 exisitng URLs found in database
18/20 exisitng URLs found in database
18/20 exisitng URLs found in database
Saurav Shekhar
Yash Chandiramani
Sumanth Vasista
Taher Abbasi
Gavin Hendricks
Michelle Fernandes
Lakshana Dinakaran
```

Once the scraping process is completed, the existing data is reloaded from dump and compiled together to be passed into the stencils program. This is then clustered based on descriptors and reports are generate.

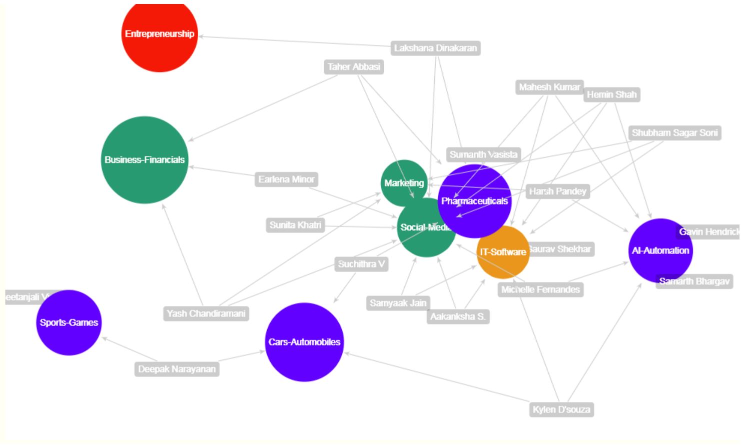


Sunflower-plots of clusters identified based on interests(a) and groups(b, c) (connectedness, geographical location, age-group, designation etc) are generated in the report.



	A	B
1	**Clusters identified from scraped nodes (profiles) - Interests**	
2		
3		
4	**Social-Media_+_Entrepreneurship_+_Pharmaceuticals**	
5		Lakshana Dinakaran (https://www.linkedin.com/in/lakshana-dinakaran-93b77975/)
6		
7	**Social-Media**	
8		Sunita Khatri (https://www.linkedin.com/in/sunita-khatri-4727412/)
9		Aakanksha S. (https://www.linkedin.com/in/aakanksha-solanki/)
10		Earlena Minor (https://www.linkedin.com/in/earlena-minor-727648/)
11		Hemin Shah (https://www.linkedin.com/in/hemin-shah-67470810b/)
12		Samyaak Jain (https://www.linkedin.com/in/samyaak-jain-3967a096/)
13		Mahesh Kumar (https://www.linkedin.com/in/maheshkumark/)
14		
15	**Social-Media_+_Marketing_+_Business-Financials**	
16		Yash Chandiramani (https://www.linkedin.com/in/yash-chandiramani-a85a90111/)
17		
18	**IT-Software**	
19		Saurav Shekhar (https://www.linkedin.com/in/saurav-shekhar-02/)
20		Aakanksha S. (https://www.linkedin.com/in/aakanksha-solanki/)
21		Hemin Shah (https://www.linkedin.com/in/hemin-shah-67470810b/)

The report (xlsx) also contains detailed clusters and lists of profiles identified in textual format indexed with the input url links.



Visualization of data through halfviz cluster-visualization libraries.