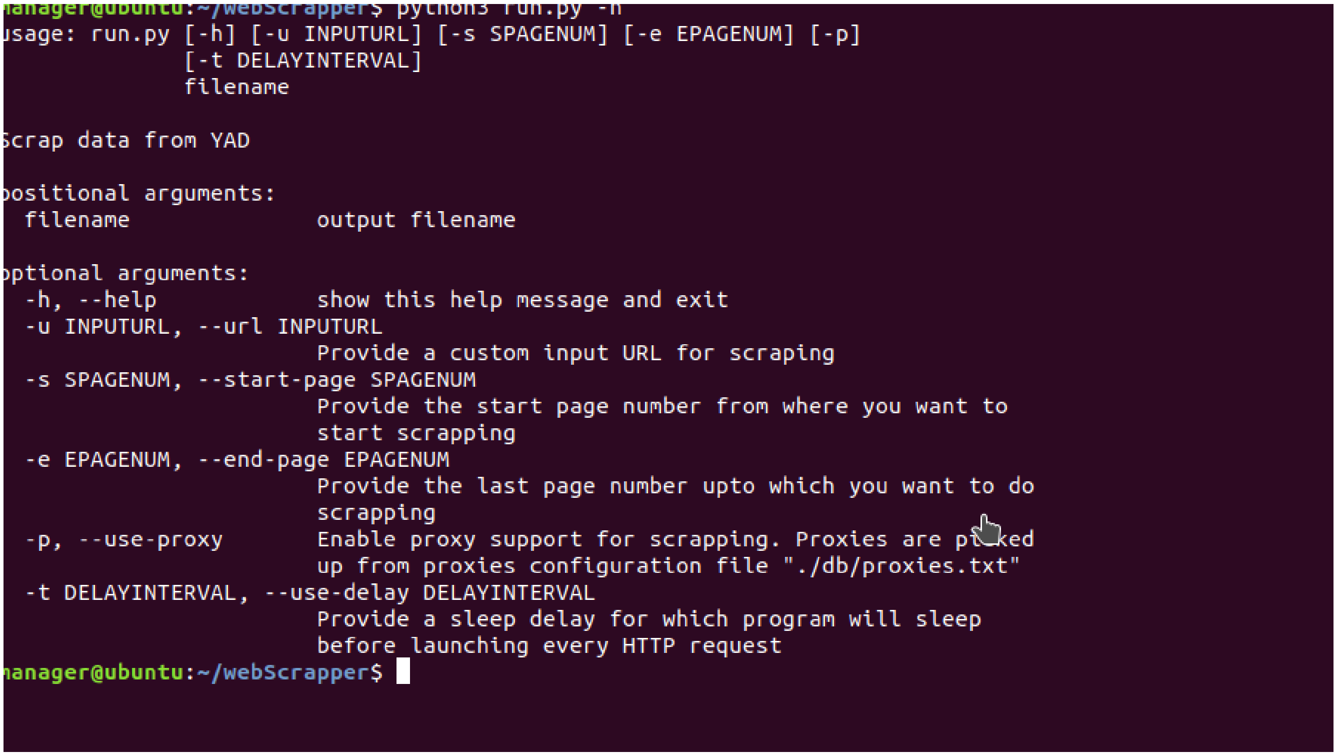
WebScrapper Project

Script provides following few flags which can be used to control script flow based on your need.



```
manager@ubuntu:~/webScrapper$ python3 run.py -h
usage: run.py [-h] [-u INPUTURL] [-s SPAGENUM] [-e EPAGENUM] [-p]
              [-t DELAYINTERVAL]
              filename

Scrap data from YAD

positional arguments:
  filename              output filename

optional arguments:
  -h, --help            show this help message and exit
  -u INPUTURL, --url INPUTURL
                        Provide a custom input URL for scraping
  -s SPAGENUM, --start-page SPAGENUM
                        Provide the start page number from where you want to
                        start scrapping
  -e EPAGENUM, --end-page EPAGENUM
                        Provide the last page number upto which you want to do
                        scrapping
  -p, --use-proxy       Enable proxy support for scrapping. Proxies are picked
                        up from proxies configuration file "./db/proxies.txt"
  -t DELAYINTERVAL, --use-delay DELAYINTERVAL
                        Provide a sleep delay for which program will sleep
                        before launching every HTTP request
manager@ubuntu:~/webScrapper$
```

Output file name a positional arguement which need to be provided for sure but other arguments are optional which can be used based on our need.

-s , --start-page:

With this parameter you can specifiy, starting page number from where scrapping will be started. Default start page is 1. This value is inserted in URL at the very last parameter location.
http://www.yad2.co.il/Nadlan/business.php?AreaID=&City=&Sale=&HomeTypeID=&fromSquareMeter=&untilSquareMeter=&fromPrice=1000&untilPrice=&PriceType=1&fromRooms=&untilRooms=&Info=&PriceOnly=1&Page=1

With -s parameter if you don't specify -e parameter which is end page number, scrapper will scrap only one page.

Example:
python3 run.py -u -s 2 "http://www.yad2.co.il/Nadlan/business.php?AreaID=&City=&Sale=&HomeTypeID=&fromSquareMeter=&untilSquareMeter=&fromPrice=1000&untilPrice=&PriceType=1&fromRooms=&untilRooms=&Info=&PriceOnly=1&Page=1" outputFileName

-e, --end-page:

End page parameter is used to provide end range upto which pages will be scrapped.

If only -e parameter is provided without any -s parameter, script will scrape from page 1 to specified end page number.

Example:

```
python3 run.py -u -e 3 "http://www.yad2.co.il/Nadlan/business.php?
AreaID=&City=&Sale=&HomeTypeID=&fromSquareMeter=&untilSquareMeter=&fromPrice=1000&
untilPrice=&PriceType=1&fromRooms=&untilRooms=&Info=&PriceOnly=1&Page=1"
outputFileName
```

-u or –url flag:
-u is for custom url that you provide for parsing some specific page.

Note: With -u flag, page start and end ranges doesn’t work.

example usage:

```
python3 run.py -u "http://www.yad2.co.il/Nadlan/business.php?
AreaID=&City=&Sale=&HomeTypeID=&fromSquareMeter=&untilSquareMeter=&fromPrice=1000&
untilPrice=&PriceType=1&fromRooms=&untilRooms=&Info=&PriceOnly=1&Page=1"
outputFileName
```

-p or –use-proxy flag:
-p is for enabling proxy support for data scrapping.
Poxies are searched in current project subfolder named ‘db/’ with a name ‘proxies.txt’.
If you want to update proxies, you should update those in the file ‘./db/proxies.txt’

```
python3 run.py -u -p "http://www.yad2.co.il/Nadlan/business.php?
AreaID=&City=&Sale=&HomeTypeID=&fromSquareMeter=&untilSquareMeter=&fromPrice=1000&
untilPrice=&PriceType=1&fromRooms=&untilRooms=&Info=&PriceOnly=1&Page=1"
outputFileName
```

-t, --use-delay flag:
-t is used for delay in between HTTP requests. It might come handy when server is blocking IP based on too many requests from a certain IP.

Example:
Here we are using delay of 3 seconds for requests.

```
python3 run.py -u -t 3 "http://www.yad2.co.il/Nadlan/business.php?
AreaID=&City=&Sale=&HomeTypeID=&fromSquareMeter=&untilSquareMeter=&fromPrice=1000&
untilPrice=&PriceType=1&fromRooms=&untilRooms=&Info=&PriceOnly=1&Page=1"
outputFileName
```

-r, `--residential` flag:
Here we are providing that what kind of data will be scrapped from YAD.
We have hardcoded a URL path in script for this option:

```
http://www.yad2.co.il/Nadlan/business.php?
AreaID=&City=&Sale=&HomeTypeID=&fromSquareMeter=&untilSquareMeter=&fromPrice=10
00&untilPrice=&PriceType=1&fromRooms=&untilRooms=&Info=&PriceOnly=1&Page=1
```

-c, `--commercial` flag:

Here we are providing that what kind of data will be scrapped from YAD.
We have hardcoded a URL path in script for this option:

```
"http://www.yad2.co.il/Nadlan/sales.php?
multiSearch=1&AreaID=&City=&HomeTypeID=&fromRooms=&untilRooms=&fromPrice=50000&
untilPrice=&PriceType=1&fromSquareMeter=&untilSquareMeter=&FromFloor=&ToFloor=&
Info=&PriceOnly=1&Order=price&Page=1"
```

Email Feature:

Email relevent SMTP configuration and receiver email ID can be configured in the file placed at “./db/email.yml”

Email contents are like this for Gmail configuration:

```
---
username: hassanejaz091@gmail.com
password: 888
smtpServer: smtp.gmail.com
smtpPort: 587
receiver: hassanejaz22@gmail.com
subject: Data Scrapping Report
body: Data report has been successfully generated
```