This documentation is regarding the web scraper automation of the amazonbusiness.in website. Our script is named “**web\_scraper.py**” and it uses python.

Before the script is run on a system, we need :

1. The latest version of Google Chrome installed on the system

2. The latest selenium web-driver available for Google Chrome. It is available here : <https://sites.google.com/a/chromium.org/chromedriver/downloads>

3. Python 2.x

The script makes use of the selenium web-scraping tool. If not installed, it can be installed using the command : “**pip install selenium**” on Linux.

**pip install pyvirtualdisplay**

**sudo apt-get install xvfb**

On the server which runs the script, the following additional commands need to be added:

**sudo apt-get install libxss1 libappindicator1 libindicator7**

**wget https://dl.google.com/linux/direct/google-chrome-stable\_current\_amd64.deb**

**sudo apt-get install fonts-liberation**

**sudo apt-get install -f**

**sudo dpkg -i google-chrome\*.deb**

**~./boto to be added onto the server with the following content:**

[Credentials]

aws\_access\_key\_id = <aws access key>

aws\_secret\_access\_key = <secret access key>

We use the packages “xlrd” and “xlwt” to read and write excel files.

We can use the following commands on the Linux terminal to install these:

“**pip install xlrd**” and “**pip install xlwt**”

The format to run the script is as follows: python path/to/the/python/file “path/to/the/excel/file”

The script imports the product names from the specified excel file, and produces an output excel file with the name in the format “**Scraped\_on\_ddmmyyyy.xls**”

The output excel file conatins the product details that was available on the amazon website. If a product was not found, or if the product was found, but some detail was not present, it will be indicated by a “N/A' in the created excel file. Also, if the “quantity” column is empty, that means the amazon didn't show any upper or lower limit on the quantity of the item. However, the quantity option on the website takes a maximum value of 999.

The resultant excel is mailed to a google group with the mail id- [**amazon.webscraping.data@10i.co.in**](mailto:amazon.webscraping.data@10i.co.in)

If in case of any error, the error mail is sent to the google group with the mail id-

[**webscraping.errorreport@10i.co.in**](mailto:webscraping.errorreport@10i.co.in)

The mail comes from the mail id – **amazonwebscraping@10i.co.in**