**Global Transfer System(GTS)**

**Arizona State University(ASU) Website Scraper Documentation:**

* **Project Goal:**

The goal of this project is to scrape and collect all the data related to university courses such as Degree,Program,location etc. This data can be viewed on the Global transfer System(GTS) UI which helps the client to choose and compare the programs with respect to the Institution/University.

* **Generic** – Generic means a single script having multiple functionalities and different level architecture in a single(.py file).
* We have developed a generic script to scrape the **Arizona State University** Website by collecting different program\_levels such as “Undergraduate, Graduate & Undergraduate Minors and Certificates” .
* **Sequence -** The sequence generator is a service that generates a sequence of integer numbers.
* **DataBase:** In this project we are using a relational **database management system** i.e **MariaDb,** the main reason behind using this as a database over Mongodb, Mariadb supports “sequence” which is an alternative to **AUTO\_INCREMENT.**

The main advantage of using sequence, it shows up in POST/SHOW tables and one can also create a sequence with CREATE table and also SELECT from it as from any other table.

**The technical requirements for this project are :-**

* + Python 3.7.0
  + Pycharm-IDLE – Python Software for learning and development environment.
  + MariaDb 10.6.5

**How to Install Python and Run the Project**

Most factory versions of Ubuntu 18.04 or Ubuntu 20.04 come with Python pre-installed. [Check your version of Python](https://phoenixnap.com/kb/check-python-version) by entering the following:

python --version or python -V

### **Update and Refresh Repository Lists**

**sudo apt update**

1. **Install Supporting Software**

The software-properties-common package gives you better control over your package manager by letting you add PPA (Personal Package Archive) repositories. Install the supporting software with the command:

**sudo apt install software-properties-common**

### **Add Deadsnakes PPA**

Deadsnakes is a PPA with newer releases than the default Ubuntu repositories. Add the PPA by entering the following:

**sudo add-apt-repository ppa:deadsnakes/ppa**

The system will prompt you to press enter to continue. Do so, and allow it to finish. Refresh the package lists again:

**sudo apt update**

### **Install Python 3**

**sudo apt install python3.7**

Allow the process to complete and verify the Python version was installed sucessfully::

**Python -V**

**Install virtual-environment:**

Update your system:

**sudo apt update**

1. Install the virtualenv tool using your package manager:  
   **sudo apt install virtualenv**
2. Create a python-environments directory in your user’s home directory and navigate to it:
3. Create a Python virtual environment.Since Ubuntu 18.04 does not have Python 2 installed, you should use the --python option to tell virtualenv to use your system’s Python 3.6 interpreter. Replace env with the name you would like to assign to your virtual environment.

**sudo virtualenv --python=python3.7 env**

1. Validate that your environment is installed with the version of Python that you expect:

**ls env/lib**

1. Activate the newly created virtual environment:

**source env/bin/activate**

1. Check python version after activating the virtual environment

**\*\*\*** To deactivate an active virtual environment, issue the following command: **\*\*\***

**Deactivate**

After activating the virtual-environment , Go to the project directory, find and install the requirement file in your environment.

**python -m pip install -r requirements.txt**

\*\*\* Try to install mariadb manually in order to avoid missing packages \*\*\*

After running the requirements.txt , install these requirements manually , required for Mariadb database.

**If you face any of these errors such as “ModuleNotFoundError: No module named 'mariadb.\_mariadb' or “ModuleNotFoundError: No module named 'mariadb’ ”**

Try to reinstall mariab module:

**$ pip3 uninstall mariadb &**

**$ pip3 install --no-cache mariadb**

**Install Mariadb server 10.6.5**(if needed)

How to install MariaDb in Ubuntu 18.04 or 20.04 & linux:

**Ref:** <https://computingforgeeks.com/how-to-install-mariadb-on-ubuntu-lts-linux/>

# **Install Mozilla Firefox in Ubuntu**

Enter the following command to update your system’s repository index with that of the internet repositories:

**$ sudo apt-get update**

Then enter the following command in order to install Mozilla Firefox from the official Ubuntu repository:

**$ sudo apt install firefox**

* **Install Gecodriver for Ubuntu , if not installed**

1. Download Gecko-Driver in your system

**wget** [**https://github.com/mozilla/geckodriver/releases/tag/v0.31.0/geckodriver-v0.31.0-linux64.tar.gz**](https://github.com/mozilla/geckodriver/releases/tag/v0.31.0/geckodriver-v0.31.0-linux64.tar.gz) **Or download manually**

2. Unzip tar file

**sudo tar -xvf geckodriver-v0.26.0-linux64.tar.gz**

3. Move Gecko Driver to Binary Location

**sudo mv geckodriver to/your/specified/path**

4. Make Executable Permission to 'geckodriver'

**sudo chmod 775 geckodriver**

Reference:<https://github.com/timgrossmann/InstaPy/issues/5282#issuecomment-666283451>

\*\*\* Before running the python file, please make sure that “**driver.headless = True**” \*\*\*

>>Run your python\_file

Activate your environment , before running your python-file

(env)filepath/> **python python\_file.py**

* **This website provides three different program\_levels:**
  + **Undergraduate :** This contains all the available programs under Undergraduate program\_level.
  + **Graduate :** This contains all the available programs under Graduate program\_level.
  + **Undergraduate Minors and Certificate :**This contains all the available programs under Undergraduate Minors and Certificate program\_level.
* **Data generates after scraping:**

By scraping the different program such as Undergraduate , Graduate and Undergraduate Minors and Certificate, we can generate data with respective to the different tables such as:

* Institution: University/Institution name, parent\_id
* Area: program\_name
* Degree: program\_degree
* Location: program\_location
* Program: progam\_name,description,math\_intensity, language-requirement , program\_fee
* Term: term\_name,credit\_type,total\_hours
* Requirement:
* requirement\_text,expression,credits,subject\_hours,critical course ,necessary course
* Course: title,description,general\_studies,college\_name
* **Python Libraries used for developing this script are:**
* **beautifulsoup4==4.10.0** : HTML Parser which is used to extract data from HTML Pages.
* **lxml==4.7.1** : Powerful and Pythonic XML processing library combining libxml2/libxslt with the ElementTree API.
* **PyYAML==6.0** : PyYAML is a YAML parser and emitter for Python.
* **requests==2.27.1** : Requests library is one of the integral part of Python for making HTTP requests to a specified URL.
* **selector lib==0.16.0** : A library to read a YML file with Xpath or CSS Selectors and extract data from HTML pages using it.
* **selenium==4.1.0** : Python language bindings for Selenium WebDriver.
* **SQLAlchemy==1.4.31** : SQLAlchemy is a library that facilitates the communication between Python programs and databases.
* **urllib3==1.26.8** :urllib3 is a powerful,sanity-friendlyHTTP client for Python. Much of the Python ecosystem already uses urllib3 and you should too. urllib3 brings many critical features that are missing from the Python standard libraries.
* **Configurations:**

The configurations file is **selectors.yaml** file which has all the configurations of the websites to be passed for scraping.

* Configurations in the sense

**WEBSITE\_NAME:**

**CREATED\_BY:** Developer's name

**TYPE:** Type of website

**DOMAIN:**

**PROJECT:**

**URL:** "website url"

**Pagination\_urls:** Fetching pagination urls from a website.

**Urls:**

css: ”selector name”

xpath:

multiple: “single of multiple urls”

type: “type of data to be extracted e.g: text,link.

attribute: “css attribute

**Post\_course:**

**Title:** "extracts title of all posts"

css: ”selector name”

type: “type of data to be extracted e.g: text,link.

**Degree:** "extracts body content of all posts in the website"

css: ”selector name”

multiple:“If multiple fields to be extracted the give

True”

type: “type of data to be extracted e.g: text,link.

**Post\_summary:**

**Title:** "extracts title of all posts"

css: ”selector name”

type: “type of data to be extracted e.g: text,link.

**Degree:** "extracts body content of all posts in the website"

css: ”selector name”

multiple:“If multiple fields to be extracted the give

True”

type: “type of data to be extracted e.g: text,link.

**Post\_term\_heading:**

**Title:** "extracts title of all posts"

css: ”selector name”

type: “type of data to be extracted e.g: text,link.

**Degree:** "extracts body content of all posts in the website"

css: ”selector name”

multiple:“If multiple fields to be extracted the give

True”

type: “type of data to be extracted e.g: text,link.

**Post\_curriculum\_requirement\_hours:**

**Title:** "extracts title of all posts"

css: ”selector name”

type: “type of data to be extracted e.g: text,link.

**Degree:** "extracts body content of all posts in the website"

css: ”selector name”

multiple:“If multiple fields to be extracted the give

True”

type: “type of data to be extracted e.g: text,link.

* **DataBase Configurations :**
  + The DB Configurations file is **instance\_settings.yaml** which have all the configurations related to db e.g- Hostname,db\_name, username, password:

engine=create\_engine(‘mysql+mysqldb://username:password@localhost:port\_number/database\_name)

**List of websites configured till now are listed below:**

1. "https://webapp4.asu.edu/programs/t5/programs/Adv/-UndergradAdd:accelerated/undergrad/false",
2. "https://webapp4.asu.edu/programs/t5/programs/Adv/-UndergradAdd:accelerated/graduate/false",
3. "https://webapp4.asu.edu/programs/t5/programs/Keyword/minor/undergrad/true",
4. "https://webapp4.asu.edu/programs/t5/programs/Keyword/%20Certificate%20/undergrad/true"