**About:**

This project is a custom web crawler written in [Python 3](https://www.python.org/downloads/release/python-364/). It was meant for browsing content on the IIITD website.

Main external libraries:

* [BeautifulSoup 4](https://www.crummy.com/software/BeautifulSoup/bs4/doc/) for HTML parsing.
* [NLTK Porter Stemmer](http://www.nltk.org/api/nltk.stem.html) for stemming.
* [SciKit-Learn](http://scikit-learn.org/stable/) for Euclidean distance calculation.
* [Pickle](https://wiki.python.org/moin/UsingPickle) for importing and exporting the index.

**How to Run**

Step 1

Download the Python 3.6. This is required to run the crawler

Step 2

Install Python setup tools. On Linux, run the following command

sudo apt-get install python3-setuptools

Step 3

Navigate to the project folder and run the following script.

cd Web-Crawler/

sudo python3 setup.py install

This will install the necessary dependencies to your local version of python3. If you still have problems after running the setup file, you can install the dependencies using pip.

Step 4

Now run the project:

**References Used:**

1. https://stackoverflow.com/a/43086135/8853372
2. https://stackoverflow.com/a/7160778/8853372
3. <https://www.youtube.com/watch?v=2RRSw7Ycv0c>
4. <https://www.youtube.com/watch?v=m_3gjHGxIJc>
5. <https://www.blog.datahut.co/post/how-to-build-a-web-crawler-from-scratch>
6. <https://www.topcoder.com/thrive/articles/web-crawler-in-python>