Research Impact Modeling

Self-Contained Setup Guide

- **General Information:**

  The project code environment setup is logically divided into three structures:

  - Environment setup for the Web Scraping
  - Environment setup for the Project Development
  - Environment setup for the Dashboard

- **Environment setup for the Web Scraping:**

  Our web scraper is fully developed in python language, so to run the scraper user must have python 3.x installed on the computer. The two libraries used to scrape/crawl the data are as follows:

  1) **Scrapy:**
     The main job of scrapy in our crawler is to fetch the data from the given page using specified xpath. To install the scrapy, user should execute the following command in the PowerShell or on conda command prompt if using the Anaconda:

     Command: `pip install scrapy`

     If you are using Anaconda, then

     ```
     conda install -c conda-forge scrapy
     ```

     (Note: Visual C++ build tools is prerequisites to install scrapy otherwise scrapy installation might fail at installation of “twisted” module)

  2) **Selenium:**
     The main job of selenium in our crawler is to provide the automated chrome driver to open the all pages specified. It also helps to execute the java script events at runtime to scrape the data dynamically. To install selenium, user have to enter the following command:

     Command:
     ```
     pip install selenium
     pip install webdriver-manager
     ```

     If you are using Anaconda, then

     ```
     conda install -c conda-forge selenium
     ```

     (Note: selenium for conda comes with web driver manager)

Next, user also have to use specific webdriver for browser which they want use for selenium, for example for chrome user can download it from here : http://chromedriver.chromium.org/

3) Execution of Scraper:

After the installation, user must place the "RIM" folder to the scrapy installed folder and then double click python file to start the scraping for URLs

RIM → RIM → spiders → **`author_profile_urls.py`**

To run it from command shell type:

```
scrapy crawl author_profile_urls -o xyz_unversity_name.csv
```

Once the URLs are scraped, it will store that URLs to csv file, which is given as input to next scraper,

RIM → RIM → spiders → **`generic_author_profile_scraping.py`**

To run it from command shell, type:

```
scrapy crawl generic_author_profile_scraping
```

- **Environment setup for the Project Development:**

For the development of the project we are using the google colab as our environment as it provides most of all the libraries in-built and notebook is always available online for collaborative working. It also provides very easy management with GitHub for version control. Google Colab also provides GPU and TPUs for heavy processing and training of the model.

To run/open project notebook, user should click 'open with colab' at following page or can download the notebook/ can also copy paste the code cell by cell to colab notebook.
Notebook URL:
https://github.com/HarshPatel-HP/Research-Impact-Modelling-RIM/blob/master/Research_Impact_Modeling.ipynb

Before running the notebook, the dataset file should be uploaded to drive. Our dataset is also available at:
Dataset URL:
https://github.com/HarshPatel-HP/Research-Impact-Modelling-RIM/blob/master/Author_Profile_Master.csv

- **Environment setup for the Dashboard:**
  For visualizing our predictions, we have used the Tableau software. User must install the Tableau Desktop version for accessing our dashboard. Open the Tableau workbook on clicking the dashboard URL.
  Files for Dashboard:
  1) https://github.com/HarshPatel-HP/Research-Impact-Modelling-RIM/blob/master/Author_Profile_Master_Processed.csv
  2) https://github.com/HarshPatel-HP/Research-Impact-Modelling-RIM/blob/master/Cad%20Uni%20GEo%20info.xlsx

  Dashboard URL:
  https://github.com/HarshPatel-HP/Research-Impact-Modelling-RIM/blob/master/ideas-rima.twb

Once the workbook is opened in the Tableau, user must link the two source files using Left Outer Join under Data Source tab. Once the setup is completed, dashboard can be visualized on the Dashboard tab.