**Generic Web Scrapping Framework**

Problem Statement: -

Design a python web scraping framework which can be used to extract documents from any website.

Assumptions: -

1. User has basic understanding of html/css required for providing inputs to the framework specific to the seed website.

Features of the framework: -

1. This framework is designed to scrape below specified 3 websites –

https://www.sec.gov/edgar/searchedgar/companysearch.html

https://www.iana.org/domains/reserved

https://www.aer.gov.au/publications

1. Framework is scalable and can be extended to scrape any other websites up to depth of 3 pages.
2. Framework has configurable options to specify the html/css attributes which allow to quickly reconfiguration for any website.
3. Framework can be extended to accommodate more complex websites and additional depth of pages.

Steps to configure website for scraping: -

Generic crawler can be configured to scrape documents from any website which has below like searching mechanism –

1. Search Box
2. Links from one page to other
3. Direct document links

In case of differently structured website, scrapper can be enhanced to handle other structures too.

Configuring the websites consists of below steps –

1. Identify how many page hits(steps) required to reach the page with downloadable documents.
2. Categorize the steps in above specified search options.
3. Update the config.yaml file with html/css selectors as specified in below snapshot.
4. Crawler doesn’t have dependency on type of html/css selectors. We can use id, class, name or any other attribute to select the element from page.

Source Code: -



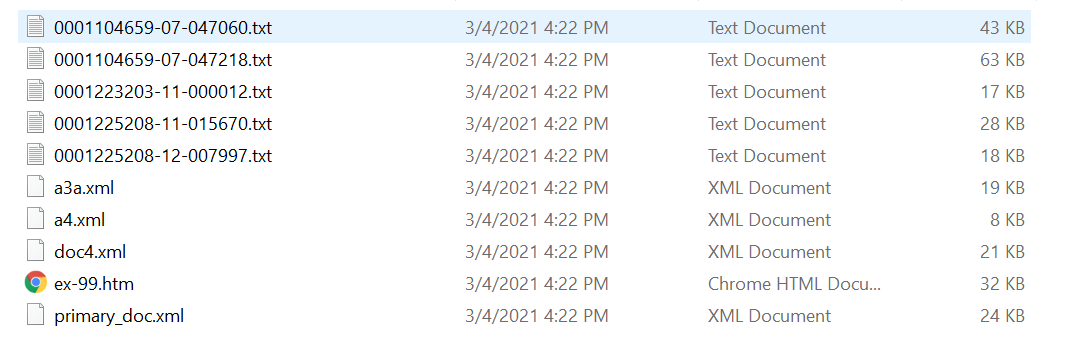


Website 1: https://www.sec.gov/edgar/searchedgar/companysearch.html

Graphical user interface, text, application

Description automatically generated

Downloaded Files:

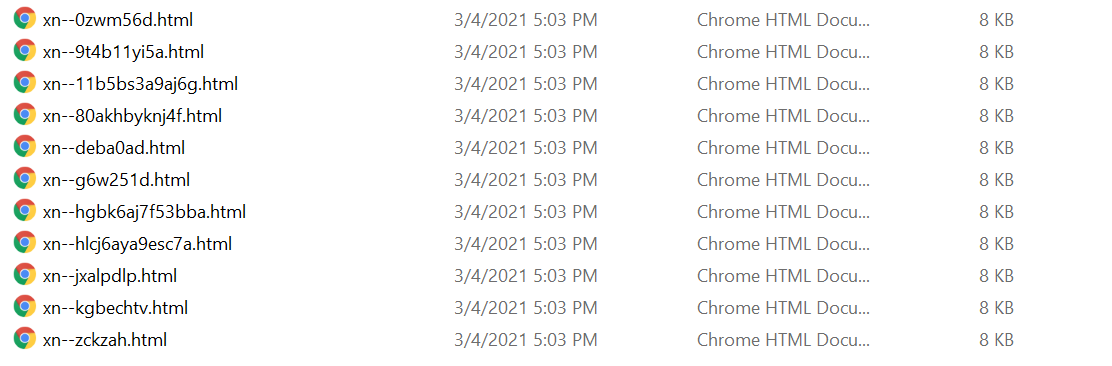


Website 2: https://www.iana.org/domains/reserved

Graphical user interface, text

Description automatically generated

Downloaded Files:



Website 3: https://www.iana.org/domains/reserved

Graphical user interface, text, application

Description automatically generated

Downloaded Files:

