Assignment 4

Web crawlers are called such because they crawl across the Web. At their core is an element of recursion. They must retrieve page contents for a URL, examine that page for another URL, and retrieve that page, ad infinitum.

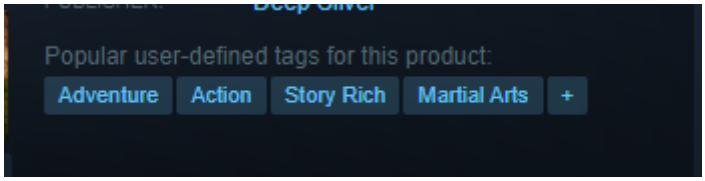
The recommended Python library for this assignment is Scrapy. Using other libraries like pandas and matplotlib is allowed.

More about Scapy crawlers: https://doc.scrapy.org/en/latest/topics/spiders.html#crawlspider

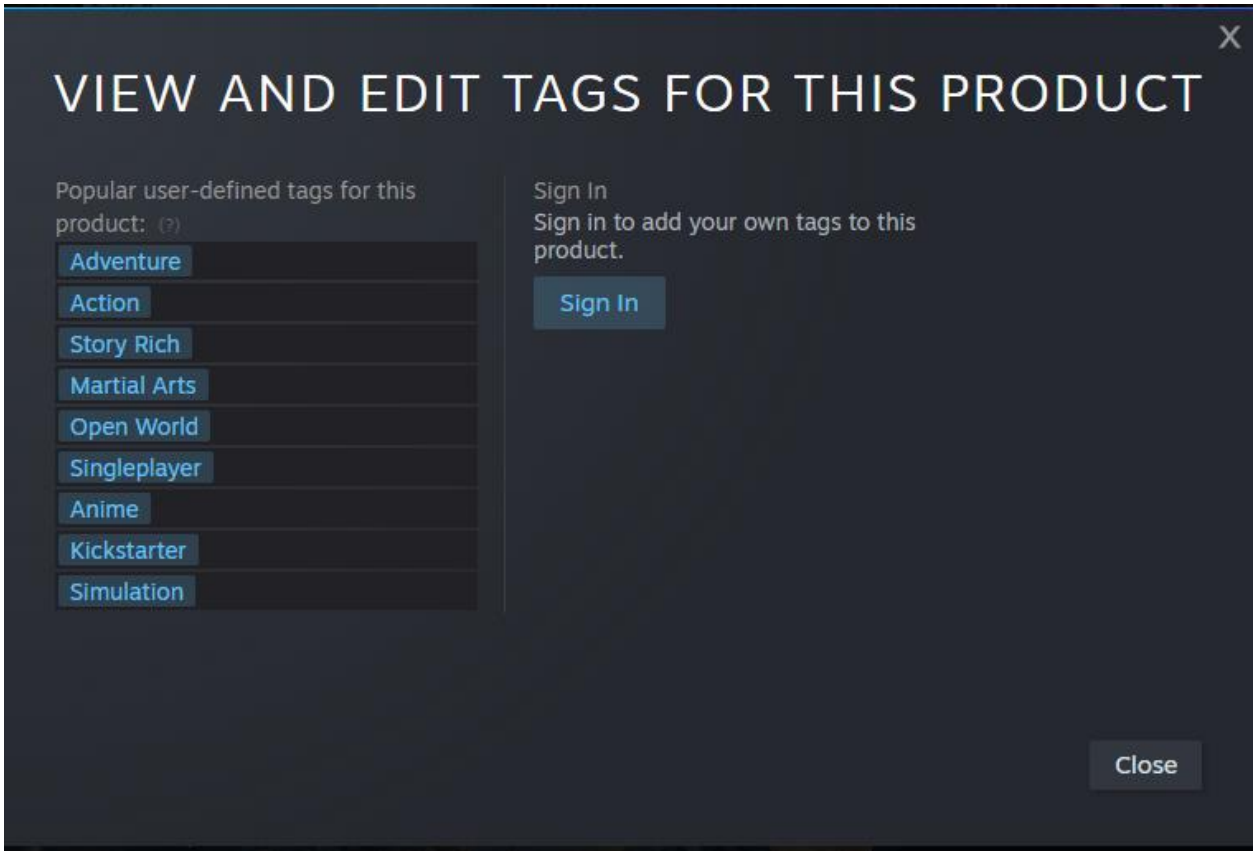
Working process:

1. Install Scrapy.

2. Extract data from Steam (https://store.steampowered.com/). Using CrawlSpider from Scrapy get the following info: Game title, release date, developer, publisher, popular user-defined tags for a product from first 1000 products of “Top Sellers”.



Advanced task: to get all the whole tags list available by “+” button.



Note: For extracting tags, you can use CSS and XPath selectors.

4. Group your games dataset by year. Name 3 the most popular game tags in each year. Name the least popular game tags in each year.

5. For these tags, print game names that have such tags (considering the proper year).

Note: e.g., if tag “Open World” was popular in 2020, but was not in a top in 2019, only game titles with tag “Open World” from 2020 should be printed.