Лабораторная работа №4

Название «веб-кроулер» происходит от английского глагола “to crawl” – ползти (как паук) по веб-ресурсам как по паутине. Центральным элементом таких программ является рекурсия: происходит запрос содержимого по URL, далее идет проверка страницы на наличие другого URL внутри, запрос содержимого по обнаруженного URL и т.д.

Для выполнения данной лабораторной работы рекомендуется использовать язык программирования Python и библиотеку Scrapy. Также допускается использование других вспомогательных библиотек таких как pandas, matplotlib и др.

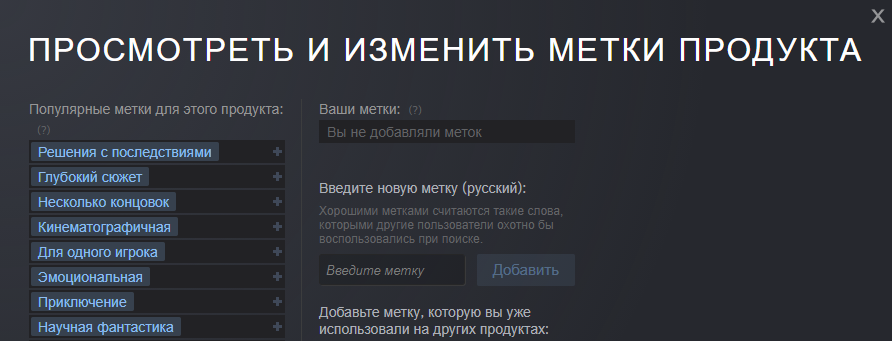
Подробнее о Scrapy можно узнать из источника: <https://doc.scrapy.org/en/latest/topics/spiders.html#crawlspider>

Ход работы:

1. Установить Scrapy
2. Извлечь из веб-ресурса Steam (<https://store.steampowered.com/>). Используя CrawlSpider из библиотеки Scrapy, извлечь информацию: название игры, дата выхода, разработчик, издатель, популярные метки для этого продукта для первых 1000 продуктов из списка «Лидеры продаж». Обратите внимание на продукты, не являющиеся играми, и комплекты (бандлы). Они не должны входить в данный перечень из собранных записей.



Дополнительное задание: получить полный список тегов, доступный по кнопке «+».



Примечание: для извлечения тегов можно использовать CSS и XPath [селекторы](https://docs.scrapy.org/en/latest/topics/selectors.html).

1. Сгруппировать игры по годам. В каждом году определить 3 самых популярных тега и 3 самых непопулярных. Вывети их.
2. Вывести игры, соответствующие самым популярным и непопулярным тегам из п.4 (в соответствии с годом, т.е. если тег принадлежит к одному году, а игра с данным тегом – к другому, то не выводить такую информацию).

Примечание: например, если тег «Решения с последствиями» попало в самые непопулярные теги 2020 года, но не попало в 2019, то из игры с непопулярными тегами такие игры должны быть выведены только для 2020 года.