**Описание программы «parser\_site\_truspilot(by company)»**

* Файл «chromedriver» используется программой для открытия и получения данных с помошью урезанной версии CHROME. Данная сборка работает по ОС Windows ( для других ОС скачать можно тут (<https://sites.google.com/a/chromium.org/chromedriver/home>)
* Файл «proxies» содержит в себе список ПРОКСИ серверов в формаье «IP:PORT» для произведения get запросов на сайте trustpilot. Сервера должны поддерживать https протокол и Html5, физических ограничений на их местоположение нет. Данный файл может обновляться и дополняться для работы (сервера переодически отключаются)
* Файл «user\_agent» содержит в себе список физических описании машин, для эмуляции пользовательских машин осуществляющих запрос на сайте.
* Файл «parser\_site\_trustpilot(by company)» содержит в себе весь исполняющий код.

**Описание процесса запуска парсера**

1. В интерпритаторе требуется указать название нового файла, который будет содержать все конечные данные (расширение файла не указывается, по умолчанию .csv)
2. Требуется указать ссылку на все категории сайта trustpilot (Пример: <https://au.trustpilot.com/categories>). При вводе допускаются пробелы до и после ссылки.
3. Ожидайте завершения работы программы.

**Описание процесса работы**

После указания ссылки (пункт 2-Описания процесса запуска парсера), со страницы копуруются в память все ссылки на которые ссылаются категории. Далее к каждой из полученной ссылке поочереди прибавляется значение из списка(открывается\закрывается окно браузера Сhrome):

['?status=claimed', '?status=unclaimed', '?status=collecting']

при этом пролистывая до конца все страницы открытых категорий со статусами из списка выше. Таким образом формируется файл «Url\_Company.csv», содержащий в себе строки :

наименование категории; оценка trustpilot(по 10-ти бальной шкале); статус; ссылку на компанию на сайте trustpilot.

После того как предыдущий шаг закончен, в фоновом режиме программа запрашивает данные у сайта trustpilot , используя прокси сервера(proxies.txt) и пользовательские данные(user\_agent.txt) для обхода блокировки со стороны сайта. При этом данные, добавленные в конечный файл будут уникальные(не повторяющиеся) все повторы при попадании удаляются, оставляя последнюю актуальную информацию.

**Конфигурирование**

* Можно изменять список статусов для выборки конкретного статуса(строка 155):

list\_status = ['?status=claimed', '?status=unclaimed', '?status=collecting']

* Для актуализации рабочих прокси серверов периодически их необходимо удалять\добавлять(файл proxies)\ проверять на сайте «https://hidemyname.org/ru»