Спайдер сбора ссылок комментариев для создания своей тематической базы дропов под названием Black Widow Spider. Сбор своей базы дропов по своей тематике. Существует обмен комментариями сайтов по своей тематике, например, на сайте с рецептами комментируют люди, имеющие близкую тематику и оставляют ссылку, ведущую на сайт, это практически сообщество, довольно узкий круг, без залетных.

Для чего это нужно?

Многие сайты уходят в небытие - люди забрасывают сайты, переезжают, меняют работы - и это, как правило, очень хорошие сайты с авторскими текстами - я работаю именно так, только отбираю вручную, отдаю программе и работаю с этими забытыми доменами, руки не доходят автоматизировать, позже приведу пример, если не ясно. А просто скаченный список доменов по ключевым словам - это 80% мусора, который программа перелопачивает, прежде чем наткнутся на что-то стоящее.

Что он делает?

Паук ходит по тематическим сайтам (например, строительной, женской или сайтам другой тематики) и собирает ссылки, находящиеся в никах пользователей, ведущей на их сайт.

Как работает?

1. БОТ берет ссылки с любых сайтов на любой платформе и на любых языках

2. БОТ самообучается - во время поиска ссылок пишет данные всех сайтов, на которые заходил. Если он зашел на Фейсбук или Твиттер и ему там не понравилось, при следующем случайном заходе он сравнит список и на них не пойдет.

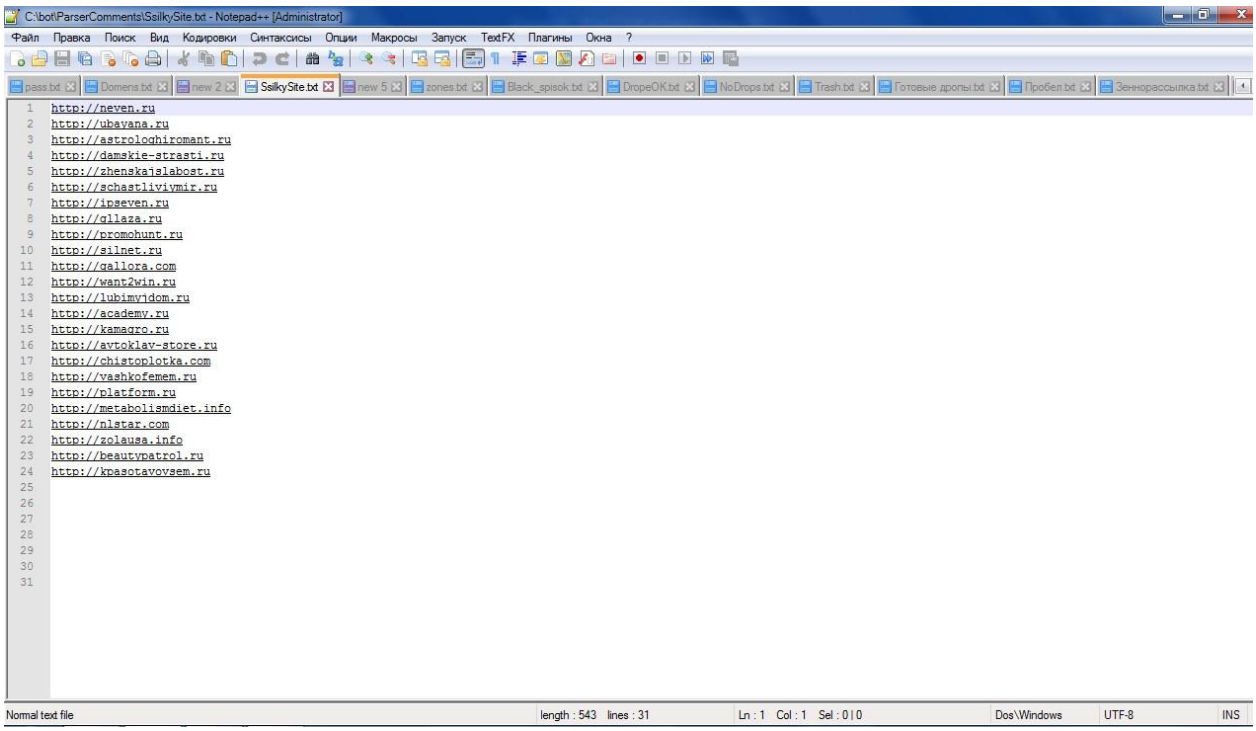
3. Примерный принцип работы БОТА (алгоритм очень сложный) - Бот берет данные из файла (начальные URL, которые вы ему скормите, смотрите видео), создает карту сайта из данного урл и начинает ходить по страницам в поисках ссылок, берет ссылки из комментариев и начинает их проверять на ответ сервера, если ответа нет, бот идет в Вебархив и проверяет, есть ли там сайт. Если ответ положительный, смотрит количество ссылок, и если считает, что сайт перспективный, заносит его в список дропов.

Эта программа идет в связке с WebArchiveMasters - спайдер собирает тематические дропы, а парсер берет с них текста. Теперь не нужны списки доменов, программа найдет всё сама.

Что нужно сделать для работы паука.

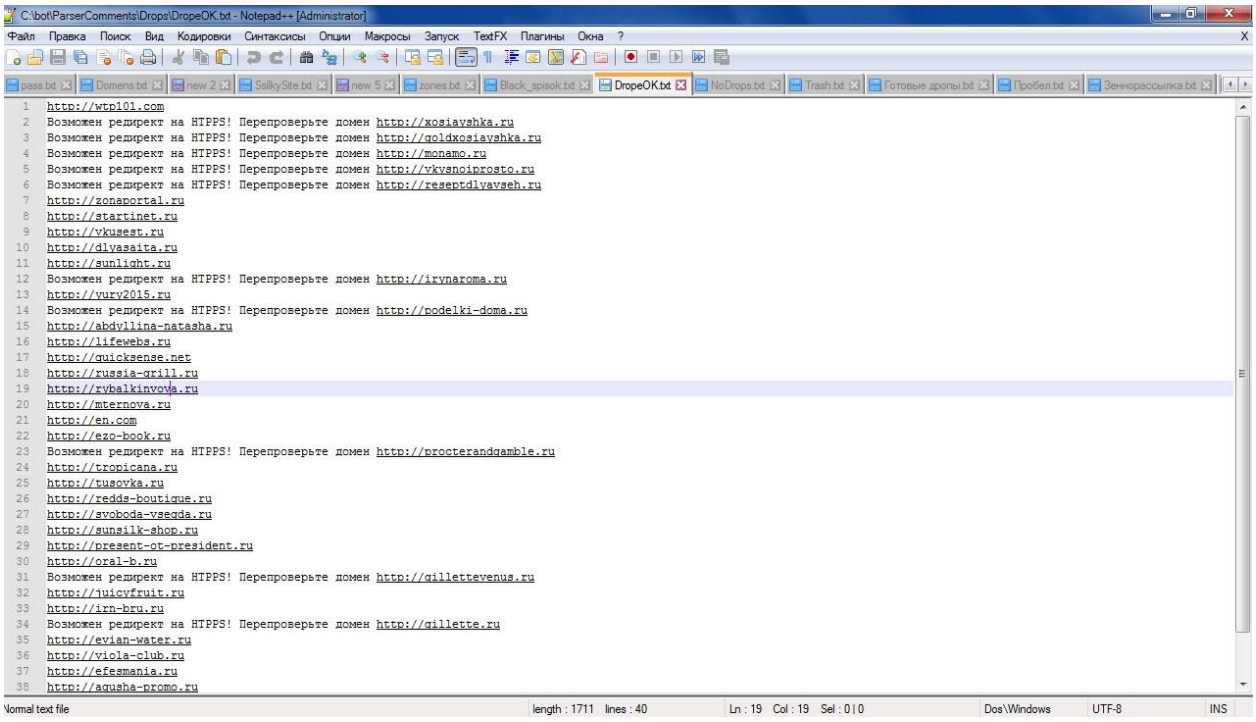
Ввести в поисковый ключевой запрос по вашей тематике - так-как я работаю по женской тематике, я ввожу запрос: "женский сайт свежие комментарии". После этого пройтись по выдаче и найти 5-6 сайтов с комментариями.

После этого скопировать УРЛы сайтов и записать их в файл **SsilkySite**.txt



1. В папке **Drops** находятся файлы **DropeOK.txt**, куда записываются найденные .

Пример файла DropeOK.txt



За час работы найдено около 40 тематических дропа.

3. в ФАЙЛЕ **NoDrops.txt** находятся существующие сайты, в подавляющем большинстве вашей тематики, этот список используется для фильтрации, как и файл **Trash.txt**, но вы можете использовать их также для себя. Как пример, вы можете скопировать их в главный файл **SsilkySite.txt**, чтобы не искать вручную, и программа пройдется по ним.

В ***Black_spisok.txt*** заносятся данные доменов, по которым бот уже проходил, чтобы не дублироваться. Бота нужно погонять не менее 2-3 часов, чтобы он набрал необходимые данные.