Web-crawler’s implementation

Существует много видов краулеров, которые используются для разных целей, например, поиск страниц с 404, просмотром индексации и с поиском доступных страниц.

В моей программе реализован сканер страниц, которые возвращают ответ с ошибкой, например, 404.

Для решения данной проблемы используем библиотеку “net/http”, в данном пакете есть функция http.Get, которая вернет тело нашей страницы с заголовками.

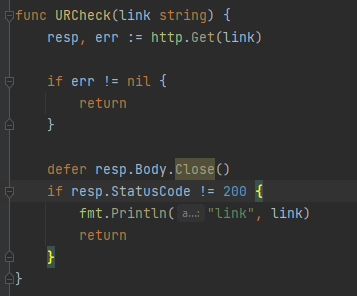
Чтобы вставить ссылку в html страницу, используется атрибут href. Его мы будем искать и заносить в базу данных (в данной реализации в массив) наши ссылки, чтобы в дальнейшем не переходить на уже проверенные ссылки.

Алгоритм:

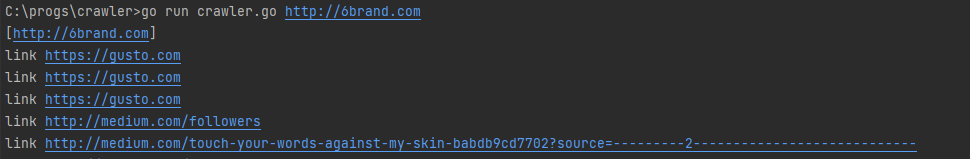
1. Построение бд с URL с добавлением их в очередь
2. Проходимся по каждой ссылке
   1. Загружаем для URL HTML страницу
   2. Если статус страницы: 404, 500 и т.д., но не 200, то идем на шаг с и выводим URL
   3. Если в очереди есть ссылки, то идем на шаг 2.

Данный краулер основан на <https://jdanger.com/build-a-web-crawler-in-go.html>

Написана функция URCheck, которая проверяет статус страницы и печатает, если он отличен от 200.



Вывод:



При переходе на <http://medium.com/followers> видим, что страница не найдена.

