Алгоритм работы

1. В качестве первого аргумента в командной строке принимается URL
2. Читается конфигурационный файл config.json. Далее идет проверка имеются ли индивидуальные шаблоны для запрашиваемого сайта. Если таких нет, то к сайту применяются стандартные шаблоны “common”. Если есть, то они добавляются к стандартным.
3. Происходит чистка документа от элементов из черного списка загруженных в шаблоне.
4. Извлекаются и форматируются ссылки согласно заданию.
5. Из документа извлекается текст и добавляется отступ между параграфами.
6. Полученный текст записывается в файл, который сохраняется в сформированной по URL директории.

Список URL на которых проверялась работа.

<https://72.ru/text/health/69388360/> - Основной текст сохранен. Потеряно оглавление. Немного мусора в конце.

<https://lenta.ru/news/2020/07/28/kinzhal/> - Основной текст сохранен. Потеряно оглавление.

<https://ura.news/news/1052442572> - Основной текст сохранен. Потеряно оглавление. Немного мусора.

<https://www.gazeta.ru/business/2020/07/28/13168093.shtml> - Основной текст сохранен.

<https://www.foxnews.com/world/israel-hezbollah-trade-fire-near-lebanese-border> - Основной текст сохранен. Немного мусора.

<https://www.bbc.com/news/entertainment-arts-53557876> - Основной текст сохранен. Мусор в конце.

Направление для дальнейшего развития/улучшения.

1. Перестать терять оглавление.
2. Внедрить больше методов для чистки мусора.