**Программа “Parser”.**

Данная программа представляет из себя утилиту командной строки, принимающую в качестве параметра URL-адрес ресурса.

**>parser https://www.newsru.com/world/31mar2019/salvatormundi.html**

Программа устанавливает соединение, скачивает информацию с указанного ресурса и обрабатывает в соответствии с требованиями, обеспечивающими максимально комфортное восприятие основной текстовой информации, содержащейся на ресурсе.

Алгоритм работы программы:

1. Соединение с сетью интернет.
2. Подключение к указанному ресурсу.
3. Получение HTML кода сайта.
4. Проверка кодировки полученного кода.
5. Если кодировка UTF-8 - смена на ANSI.
6. Обработка полученного кода:

6.1 Сохранение в промежуточный буфер содержимого тегов <title> и <p>.

6.2 Форматирование ссылок, присутствующих в буфере (обрамление в [ ]).

6.3 Удаление из буфера лишних слов, элементов, тегов, мешающих восприятию текста и не несущих никакой полезной информации.

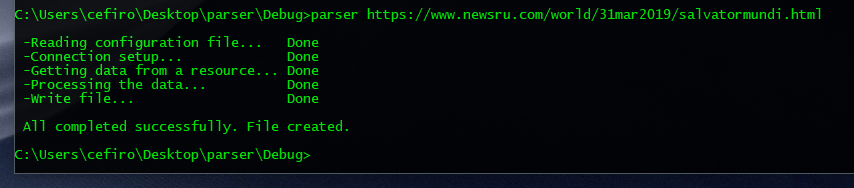
6.4 Выравнивание текста в буфере по ширине с переносом слов.

7. Формирование имени будущего файла в соответствии URL ресурса.

8. Сохранение итогового файла.

Программа имеет файл конфигурации conf.txt, в котором можно изменить максимальную длину строки в итоговом файле, а также добавить новые исключаемые слова, теги, элементы. При отсутствии конфигурационного файла программа использует параметры по умолчанию (80 символов и стандартный список исключений).

В ходе выполнения программы в консоли выводится статусная информация, по которой можно понять, все ли в порядке.



В случае возникновения проблем (нет подключения к интернет или отсутствует конфигурационный файл), программа выдаст сообщение об ошибке.



Список ресурсов, на которых программа работает корректно:

* <https://www.newsru.com/> (некоторые стаьи отображаются некорректно, но читаемы и вся важная информация присутствует)
* <https://tass.ru/> (все статьи на сайте, все корректно)
* <https://www.bbc.com/> (все статьи на сайте, все корректно)
* [https://www.rbc.ru](https://www.rbc.ru/) (некоторые стаьи отображаются некорректно, но читаемы и вся важная информация присутствует)
* <http://www.cnews.ru/> (все статьи на сайте, все корректно)
* <https://lenta.ru/> (все статьи на сайте, все корректно)

Примеры итоговых файлов с этих сайтов в папке проекта.

**Дальнейшее развитие программы**

* ввести в программу возможность устанавливать в конфигурационном файле список нужных новостных ресурсов для обеспечения агрегации свежих новостей по интересующим темам.
* ввести возможность получения не только текстовой, но и другой важной информации (видео, фото) в отдельную папку рядом с текстом, или отдельно.
* ввести машинное обучение, чтобы программа сама распознавала лишние элементы в итоговом тексте и вносила их в список исключения.