Александр Алексеев.

Задача C++|Python – Mini readability.

Формулировка задачи

Большинство веб-страниц сейчас перегружено всевозможной рекламой… Наша задача «вытащить» из веб-страницы только полезную информацию, отбросив весь «мусор» (навигацию, рекламу и тд).

Полученный текст нужно отформатировать для максимально комфортного чтения в любом текстовом редакторе. Правила форматирования: ширина строки не больше 80 символов (если больше, переносим по словам), абзацы и заголовки отбиваются пустой строкой. Если в тексте встречаются ссылки, то URL вставить в текст в квадратных скобках. Остальные правила на ваше усмотрение.

Программа оформляется в виде утилиты командной строки, которой в качестве параметра указывается произвольный URL. Она извлекает по этому URL страницу, обрабатывает ее и формирует текстовый файл с текстом статьи, представленной на данной странице.

В качестве примера можно взять любую статью на lenta.ru, gazeta.ru и тд

Алгоритм должен быть максимально универсальным, то есть работать на большинстве сайтов.

Усложнение задачи 1: Имя выходного файла должно формироваться автоматически по URL. Примерно так: http://lenta.ru/news/2013/03/dtp/index.html=>[CUR\_DIR]/lenta.ru/news/2013/03/dtp/index.txt

Усложнение задачи 2: Программа должна поддаваться настройке – в отдельном файле/файлах задаются шаблоны обработки страниц.

Требования к выполнению задачи

1. Задача выполняется на С++|Python с использованием классов. Не должно использоваться сторонних библиотек, впрямую решающих задачу.

2. Предпочтительная среда выполнения – MS Windows.

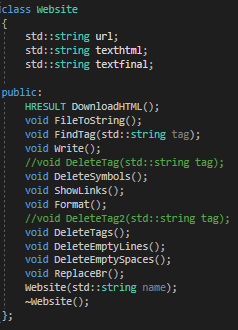
3. Решение должно состоять из документа, описывающего алгоритм, исходных кодов программы, исполняемого модуля.

4. Приложите список URL, на которых вы проверяли свое решение. И результаты проверки.

5. Желательно указать направление дальнейшего улучшения/развития программы.

1. Класс

Имеется класс Website, в котором будет хранится url сайта, который передается, а также текст со скачанной страницы texthtml и отформатированный итоговый текст textfinal



Website.h

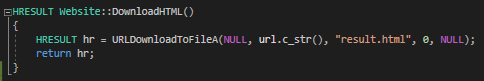
1. Алгоритм

Так выглядит основной алгоритм:



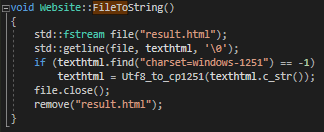
Source.cpp

Сначала скачивается HTML страница

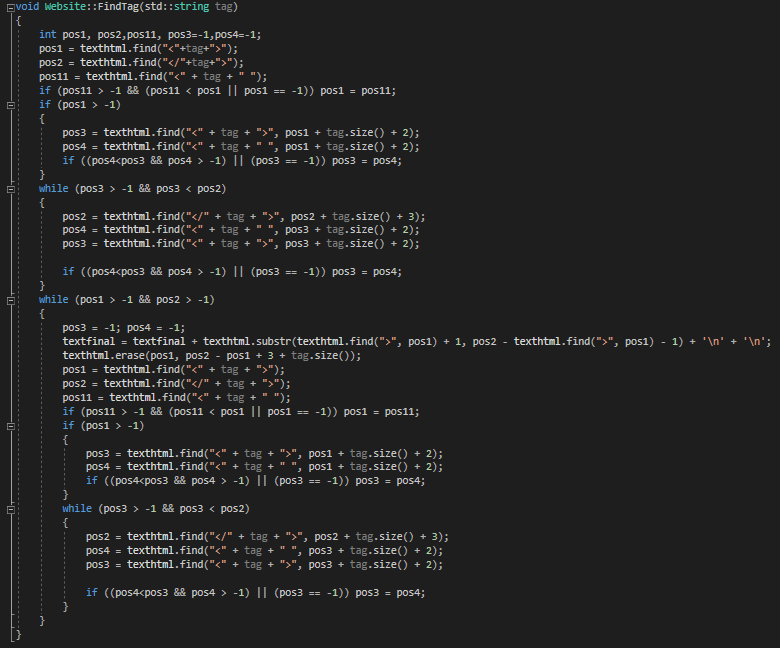


Если успешно скачалась, то идем дальше, иначе выводим ошибку.

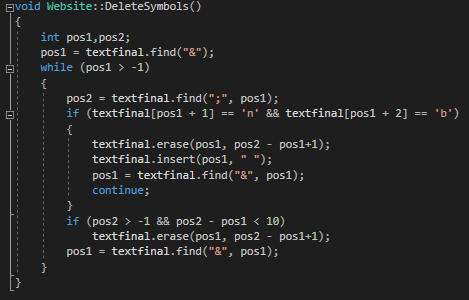
Затем заносим из файла текст и удаляем файл



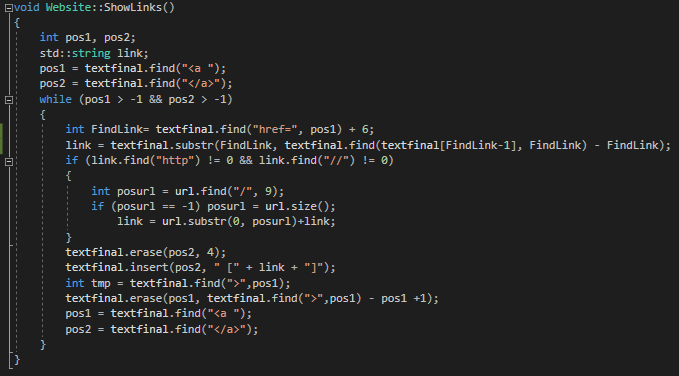
Из texthtml вырываем текст под тегами <title> и <p> и вставляем в textfinal



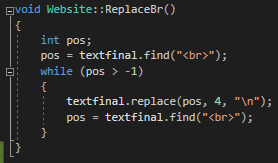
Удаляем спецсимволы (&nbsp; заменяем на пробел)



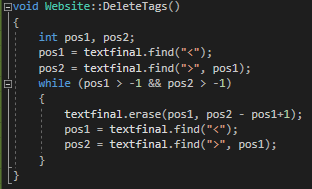
Если в тексте встречаются ссылки, то URL вставить в текст в квадратных скобках.



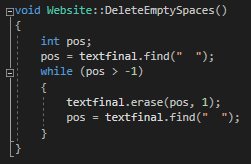
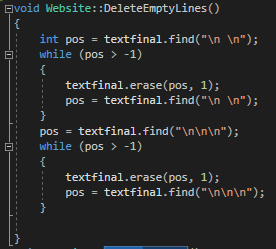
Заменяем <br> переносом на новую строку



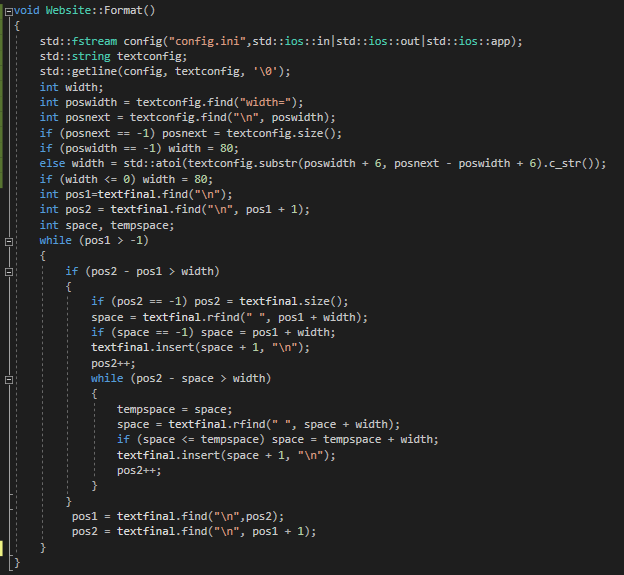
Удаляем все оставшиеся теги



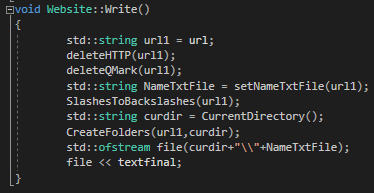
Убираем ненужные пробелы

Форматируем строки (ширину строк берем из отдельного файла)

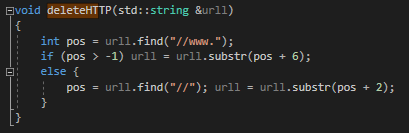


Запись в файл

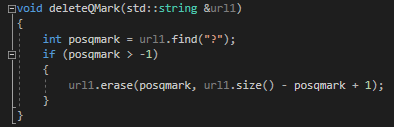


Рассмотрим алгоритм записи в файл:

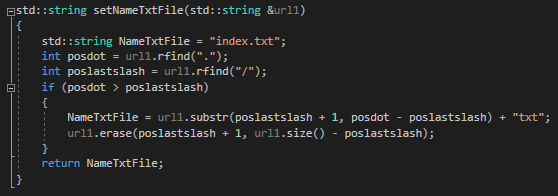
Из url сначала удаляем http://



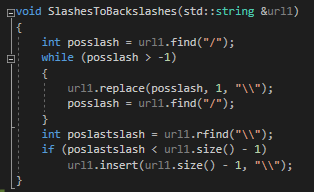
Затем, если встречается GET запрос стираем ненужное



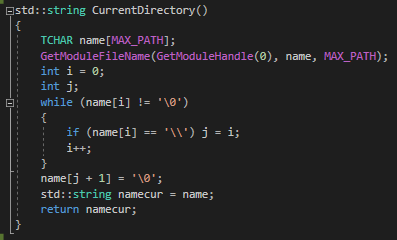
Задаем имя txt файла который мы впоследствии создадим



Заменяем слеши обратными слешами

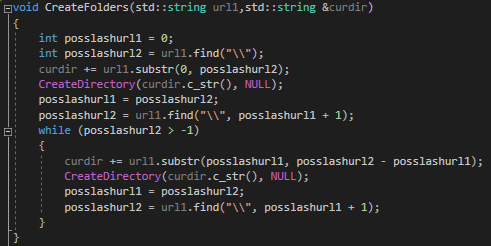


Узнаем текущую директорию



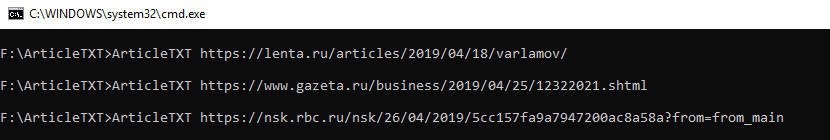
Имя выходного файла должно формироваться автоматически по URL.

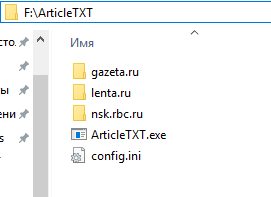
Создаем папки

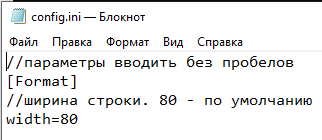


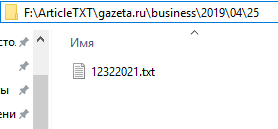
И в итоге в заданной директории сохраняем файл с названием NameTxtFile и записываем туда textfinal

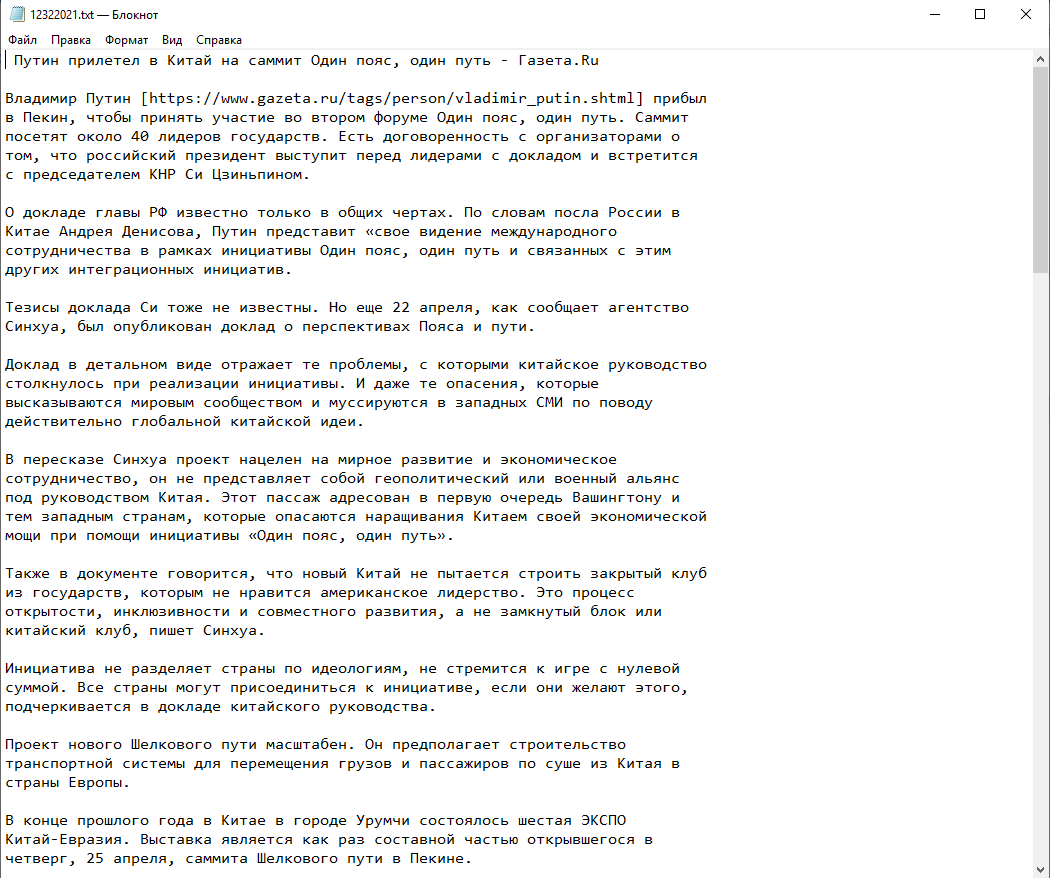
1. Результаты проверки

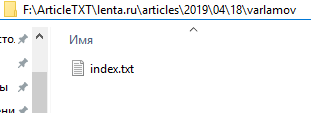


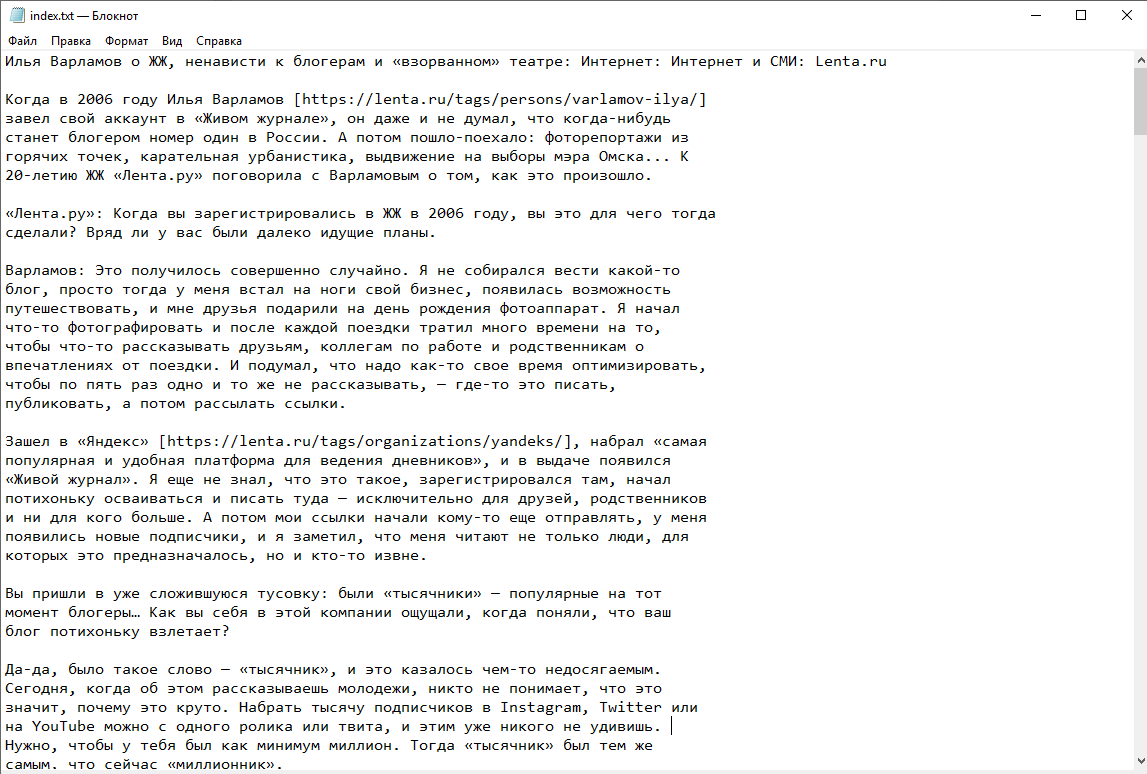


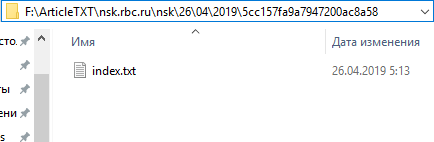


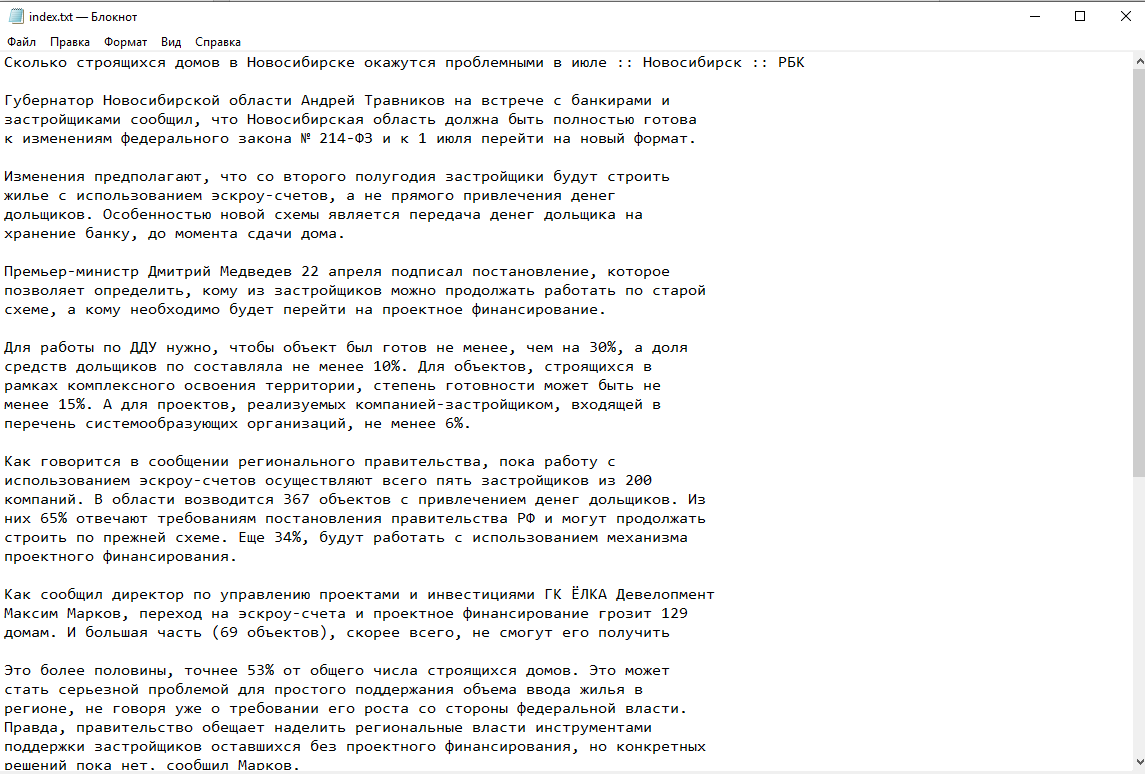












Работает на большинстве сайтов.

Проверялось на lenta.ru, gazeta.ru, rg.ru, kinopoisk.ru, kanobu.ru, the-flow.ru, psychologies.ru, Pravda.ru, rbc.ru, riafan.ru

1. Желательно указать направление дальнейшего улучшения/развития программы.

В дальнейшем можно заменять спецсимволы, в моем алгоритме я удаляю спецсимволы и только &nbsp; заменяю на пробел.

В дальнейшем рассматривается более гибкая настройка программы с отдельного файла: например брать оттуда список тегов, текст внутри которых мы хотим оставить. Допустим в файле мы задаем findtags=title,h1,h2,h3,p и в программе работает функция FindTag(tag) для каждого тега.