

Задача C++ | Python – Mini readability

*Внимательно прочитайте формулировку и требования к выполнению задачи!
Оценивается не только собственно программа, но и умение понять задачу и
требования заказчика. Все фразы ниже значимы!*

Формулировка задачи.....	1
Дополнения к задаче.....	1
Требования к выполнению задачи	1

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.

Дополнения к задаче

Выполнение дополнений – сугубо на ваше усмотрение.

Дополнение 1: Реализуйте решение для удобного хранения файлов, полученных в результате обработки веб-страниц. Обоснуйте, почему ваше решение будет удобно.

Дополнение 2: Выделять значимые картинки – те, которые относятся к тексту статьи. Выкладывать эти картинки рядом с файлом-результатом, а в файле оставлять ссылки на картинки в квадратных скобках.

Требования к выполнению задачи

1. Задача выполняется на C++ или Python с использованием классов. Использование сторонних библиотек не запрещается, кроме библиотек, напрямую 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.