Тестовое задание: Задача Python - Mini readability

# Алгоритм решения задачи

1. Получение html-страницы с помощью requests и BeautyfulSoup.
2. Получение списка тэгов из внешних файлов(header\_tags, content\_tags).
3. Выборка информации из html-страницы по указанным тэгам.
4. Формирование пути выходного файла посредством создания необходимых каталогов.
5. Запись результата в файл.

# Исходный код программы

#!/usr/bin/python3

import requests

import sys

from bs4 import BeautifulSoup

import os

# длина строки

LINE\_LENGTH = 80

# имя файла результата

FILE\_NAME = 'output.txt'

# класс для простого текста

class Text:

def \_\_init\_\_(self, html):

self.html = html

self.text = html.text

def add\_line\_breaks(self): # функция разбивания текста на строки

result = ''

index = LINE\_LENGTH

while (len(self.text) / index > 0):

while (index < len(self.text) and self.text[index] != ' ' and index > 0):

index -= 1

result += self.text[0 : index]

result += '\n'

self.text = self.text[index + 1 : len(self.text)]

index = LINE\_LENGTH

result += self.text

self.text = result

# класс для контента со ссылками

class Content(Text):

def add\_links(self): # функция добавлени адресов ссылок в текст

for a in self.html.findAll('a'):

if a.string == None:

a.string = ''

a.string = a.string + "[{}]".format(a['href'])

self.text = self.html.text

def get\_tags\_from\_file(file\_name, default\_values): # функция получения тэгов из файла

tags = []

try:

file = open(file\_name, "r")

while file and True:

line = file.readline()

if not line:

break

tags.append(line.strip())

except FileNotFoundError:

tags = default\_values

return tags

def create\_path(url):

url = url[url.find('://') + 3 : len(url)]

dirs = url.split('/')

path = ''

for dir in dirs:

path = dir + '/'

if not(os.path.exists(path)):

os.mkdir(dir)

os.chdir(os.getcwd() + '/' + dir)

# извлечение html-страницы по url

url = sys.argv[1]

response = requests.get(url)

soup = BeautifulSoup(response.text, 'html.parser')

# получение тэгов для поиска информации

header\_tags = get\_tags\_from\_file('header\_tags.txt', ['h1', 'h2'])

content\_tags = get\_tags\_from\_file('content\_tags.txt', ['p'])

# получение заголовков и контента

if (soup.article):

headers = soup.article.find\_all(header\_tags)

contents = soup.article.find\_all(content\_tags)

else:

headers = soup.find\_all(header\_tags)

contents = soup.find\_all(content\_tags)

# обработка заголовков и контента, формирование результата

result = ''

for title in headers:

header = Text(title)

header.add\_line\_breaks()

result += f'{header.text}\n'

for p in contents:

content = Content(p)

content.add\_links()

content.add\_line\_breaks()

result += f'{content.text}\n'

# формирование пути для выходного файла по url

create\_path(url)

print(f'Результат: {os.getcwd()}/{FILE\_NAME}')

# запись результата в файл

file = open(FILE\_NAME, "w")

file.write(result)

file.close()

# Результаты тестирования

* <https://lenta.ru/news/2023/07/01/orlov>

Обладатель Кубка Стэнли Орлов пожаловался на неопределенность будущего в НХЛ

Фото: Григорий Сысоев / РИА Новости

Российский защитник «Бостон Брюинс» Дмитрий Орлов[/tags/persons/orlov-dmitriy/]

пожаловался на неопределенность своего будущего в Национальной хоккейной лиге

(НХЛ). Его слова приводит

YouTube[https://www.youtube.com/watch?v=4Os7CaU-a9k&ab\_channel=NHL.online]-канал

NHL.online.

Контракт хоккеиста с бостоном закончился 1 июля, на данный момент он не нашел

новую команду. «На самом деле самое тяжелое — это неопределенность. Не знаешь,

куда ты можешь поехать, так как нет разговоров. Ждешь 1 июля — и тогда уже

начнется какое-то движение, звонки агенту», — рассказал Орлов.

Хоккеист также поделился, какие факторы являются для него определяющими при

выборе команды. «Чтобы лучшие условия были для меня, чтобы были хорошая команда,

где бы я мог играть в плей-офф и бороться за кубок. Ну и также чтобы получать

удовольствие от хоккея», — пояснил он.

24 февраля Орлова обменяли в «Бостон» вместе с американским нападающим Гарнетом

Хэтэуэем из «Вашингтон Кэпиталс». В составе нового клуба россиянин провел 23

матча в регулярном чемпионате, в которых забросил четыре шайбы и отдал 13

результативных передач.

Защитник выступал за «Вашингтон» с 2011 года. Вместе с командой он стал

обладателем Кубка Стэнли в 2018-м. В составе сборной

России[/tags/organizations/sbornaya-rossii/] хоккеист выиграл чемпионат мира.

* <https://sport.rambler.ru/autosport/51029151-chempion-formuly-1-nazval-edinstvennogo-pilota-kotoryy-mozhet-sravnitsya-s-ferstappenom>

Чемпион Формулы-1 назвал единственного пилота, который может сравниться

с Ферстаппеном

Чемпион Формулы-1 Жак Вильнёв[/person/vilnev-zhak/] комплиментарно высказался о

голландском гонщике «Ред Булл» Максе Ферстаппене, предположив, что лишь испанец

Фернандо Алонсо[https://www.championat.com/tags/449-fernando-alonso/] («Астон

Мартин») может сравниться с ним.

«То, что проделывает Ферстаппен, просто невероятно. Быть может, только Алонсо

может сравниться с ним. Если посмотрить на большинство других пилотов, то у всех

них бывают взлёты и падения. Но только не у Макса. Я действительно не знаю,

откуда он берёт всю эту энергию», — приводит слова Вильнёва PlanetF1 со ссылкой

на La Gazzetta dello Sport.

В настоящий момент Ферстаппен уверенно лидирует в общем зачёте пилотов, имея в

активе 229 очков. На втором месте товарищ Ферстаппена по команде Серхио

Перес[/person/peres-serhio/], на счету которого 148 баллов. Алонсо замыкает

топ-3 (131 очко).

* <https://www.gazeta.ru/sport/news/2023/07/03/20798738.shtml>

Рублев стартовал с победы на Уимблдоне

Россиянин Рублев победил австралийца Перселла на старте Уимблдона

Российский теннисист Андрей Рублев[/tags/person/andrei\_rublev.shtml] вышел во

второй круг Уимблдона-2023, который пройдет на травяных кортах Всеанглийского

клуба лаун-тенниса и крокета в Лондоне[/tags/geo/london.shtml].

В матче первого круга 25-летний россиянин победил Макса Перселла

из Австралии[/tags/geo/avstraliya.shtml]. Встреча продолжалась 1 час 33 минуты и

завершилась со счетом 6:3, 7:5, 6:4.

Во втором круге третьего в сезоне турнира Большого шлема Рублев встретится

с победителем матча между французом Люкой Ван Ашем и россиянином Асланом

Карацевым.

Уимблдон-2023 стартовал 3 июля и продлится до 16 июля. Россияне и белорусы

допущены на турнир в нейтральном статусе. В 2022 году российским и белорусским

теннисистам отказали в праве принять участие в Уимблдоне из-за проведения СВО

на территории Украины[/tags/geo/ukraina.shtml].

Сайт «Би-би-си» сообщал, что спортсменам из этих стран также пришлось

предоставить доказательства отказа от спонсорской поддержки компаний, попавших

под британские санкции.

Ранее двукратный чемпион Уимблдона

допустил[http://www.gazeta.ru/sport/news/2023/07/03/20796764.shtml] срыв матчей

турнира.

Также тестирование проводилось для следующих url:

* <https://cheremuha.com/2023/06/29/mne-protivno-uchastvovat-v-etom-deputat-vladimir-paxarev-o-procedure-vybora-kandidatov-dlya-doski-pochyota.html>
* <https://kub.media/news/6621-est-tolko-dva-gendera-rossii/>
* <https://sportrbc.ru/news/64a2b1369a79478f47539099>