

Колледж космического машиностроения и технологий

КУРСОВОЙ ПРОЕКТ

По МДК.01.02 «Прикладное программирование»

Тема: «Разработка парсера HTML на Python»

Выполнил студент

Пищулин И.А.

Группа П1-17

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_(Подпись)

\_\_\_\_\_\_\_\_\_\_\_ (Дата сдачи работы)

Принял преподаватель

Гусятинер Л.Б.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_(Оценка)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (Подпись)

Королёв 2020 г.

Оглавление

[Введение 3](#_Toc43504838)

[1. Теоретическая часть 4](#_Toc43504839)

[1.1. Описание предметной области 4](#_Toc43504840)

[1.2. Описание существующих разработок 8](#_Toc43504841)

[1.2.1. ParserOK. 8](#_Toc43504842)

[1.2.2. Datacol 9](#_Toc43504843)

[2. Проектная часть 11](#_Toc43504844)

[2.1. Диаграмма прецедентов 11](#_Toc43504845)

[2.2. Выбор инструментов 11](#_Toc43504846)

[2.3. Проектирование сценария 12](#_Toc43504847)

[2.4. Диаграмма классов 12](#_Toc43504848)

[2.5. Описание главного модуля 13](#_Toc43504849)

[2.6. Описание спецификаций к модулям 16](#_Toc43504850)

[2.8. Описание тестовых наборов модулей 20](#_Toc43504851)

[2.9. Описание применения средств отладки 21](#_Toc43504852)

[2.10. Анализ оптимальности использования памяти и быстродействия 21](#_Toc43504853)

[3. Эксплуатационная часть 22](#_Toc43504854)

[3.1. Руководство операторa 22](#_Toc43504855)

[3.2. To-Do лист 22](#_Toc43504856)

[Заключение 23](#_Toc43504857)

[Список литературы и интернет-источников 24](#_Toc43504858)

[Приложения 25](#_Toc43504859)

# **Введение**

Целью данной курсового проекта является разработка парсера HTML. Эта тема является актуальной на данный момент, потому что в наше время web разработки выходят на один из первых планов в программировании. Данная курсовая работа позволит получить из HTML кода абстрактное дерево классов, которое в дальнейшем можно использовать в создании браузеров.

В первой части будет рассмотрена предметная область данной темы, а также несколько продуктов по данной теме.

Во второй части будут рассмотрены классы и методы, которые были разработаны, структура программной части и листинги ключевых частей программных модулей.

В третьей части будет рассмотрено руководство для пользователей, а также подробная инструкция по использованию программы.

В заключительной части будет приведен To-do лист с планами по доработке программы, а также сделаны общие выводы о получившемся проекте.

# **Теоретическая часть**

## **Описание предметной области**

Парсер — это программа, сервис или скрипт, который собирает данные с указанных веб-ресурсов, анализирует их и выдает в нужном формате.

С помощью парсеров можно делать много полезных задач:

* **Цены**. Актуальная задача для интернет-магазинов. Например, с помощью парсинга вы можете регулярно отслеживать цены конкурентов по тем товарам, которые продаются у вас. Или актуализировать цены на своем сайте в соответствии с ценами поставщика (если у него есть свой сайт).
* **Товарные позиции**: названия, артикулы, описания, характеристики и фото. Например, если у вашего поставщика есть сайт с каталогом, но нет выгрузки для вашего магазина, вы можете спарсить все нужные позиции, а не добавлять их вручную. Это экономит время.
* **Метаданные**: SEO-специалисты могут парсить содержимое тегов title, description и другие метаданные.
* **Анализ сайта**. Так можно быстро находить страницы с ошибкой 404, редиректы, неработающие ссылки и т. д.

**Для справки**. Есть еще серый парсинг. Сюда относится скачивание контента конкурентов или сайтов целиком. Или сбор контактных данных с агрегаторов и сервисов по типу Яндекс.Карт или 2Гис (для спам-рассылок и звонков). Но мы будем говорить только о белом парсинге, из-за которого у вас не будет проблем.

**HTML** (от [англ.](https://ru.wikipedia.org/wiki/%D0%90%D0%BD%D0%B3%D0%BB%D0%B8%D0%B9%D1%81%D0%BA%D0%B8%D0%B9_%D1%8F%D0%B7%D1%8B%D0%BA) *HyperText Markup Language* — «язык [гипертекстовой](https://ru.wikipedia.org/wiki/%D0%93%D0%B8%D0%BF%D0%B5%D1%80%D1%82%D0%B5%D0%BA%D1%81%D1%82) разметки») — стандартизированный язык разметки документов во [Всемирной паутине](https://ru.wikipedia.org/wiki/%D0%92%D1%81%D0%B5%D0%BC%D0%B8%D1%80%D0%BD%D0%B0%D1%8F_%D0%BF%D0%B0%D1%83%D1%82%D0%B8%D0%BD%D0%B0). Большинство [веб-страниц](https://ru.wikipedia.org/wiki/%D0%92%D0%B5%D0%B1-%D1%81%D1%82%D1%80%D0%B0%D0%BD%D0%B8%D1%86%D0%B0) содержат описание разметки на языке HTML (или [XHTML](https://ru.wikipedia.org/wiki/XHTML)). Язык HTML интерпретируется [браузерами](https://ru.wikipedia.org/wiki/%D0%91%D1%80%D0%B0%D1%83%D0%B7%D0%B5%D1%80); полученный в результате интерпретации форматированный текст отображается на экране монитора компьютера или мобильного устройства.

Язык HTML до 5-й версии определялся как приложение [SGML](https://ru.wikipedia.org/wiki/SGML) (стандартного обобщённого языка разметки по стандарту [ISO](https://ru.wikipedia.org/wiki/ISO) 8879). Спецификации HTML5 формулируются в терминах [DOM](https://ru.wikipedia.org/wiki/Document_Object_Model) (объектной модели документа).

Язык [XHTML](https://ru.wikipedia.org/wiki/XHTML) является более строгим вариантом HTML, он следует синтаксису [XML](https://ru.wikipedia.org/wiki/XML) и является приложением языка XML в области разметки гипертекста.

Во всемирной паутине HTML-страницы, как правило, передаются браузерам от сервера по протоколам [HTTP](https://ru.wikipedia.org/wiki/HTTP) или [HTTPS](https://ru.wikipedia.org/wiki/HTTPS), в виде простого текста или с использованием [шифрования](https://ru.wikipedia.org/wiki/%D0%A8%D0%B8%D1%84%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5).

Язык [гипертекстовой](https://ru.wikipedia.org/wiki/%D0%93%D0%B8%D0%BF%D0%B5%D1%80%D1%82%D0%B5%D0%BA%D1%81%D1%82) разметки HTML был разработан [британским](https://ru.wikipedia.org/wiki/%D0%92%D0%B5%D0%BB%D0%B8%D0%BA%D0%BE%D0%B1%D1%80%D0%B8%D1%82%D0%B0%D0%BD%D0%B8%D1%8F) учёным [Тимом Бернерсом-Ли](https://ru.wikipedia.org/wiki/%D0%91%D0%B5%D1%80%D0%BD%D0%B5%D1%80%D1%81-%D0%9B%D0%B8,_%D0%A2%D0%B8%D0%BC) приблизительно в [1986](https://ru.wikipedia.org/wiki/1986_%D0%B3%D0%BE%D0%B4)—[1991 годах](https://ru.wikipedia.org/wiki/1991_%D0%B3%D0%BE%D0%B4) в стенах [ЦЕРНа](https://ru.wikipedia.org/wiki/%D0%A6%D0%95%D0%A0%D0%9D) в [Женеве](https://ru.wikipedia.org/wiki/%D0%96%D0%B5%D0%BD%D0%B5%D0%B2%D0%B0) в [Швейцарии](https://ru.wikipedia.org/wiki/%D0%A8%D0%B2%D0%B5%D0%B9%D1%86%D0%B0%D1%80%D0%B8%D1%8F). HTML создавался как язык для обмена научной и технической документацией, пригодный для использования людьми, не являющимися специалистами в области [вёрстки](https://ru.wikipedia.org/wiki/%D0%92%D1%91%D1%80%D1%81%D1%82%D0%BA%D0%B0). HTML успешно справлялся с проблемой сложности SGML путём определения небольшого набора структурных и [семантических](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BC%D0%B0%D0%BD%D1%82%D0%B8%D0%BA%D0%B0) элементов — дескрипторов. Дескрипторы также часто называют «[тегами](https://ru.wikipedia.org/wiki/%D0%A2%D0%B5%D0%B3_(%D1%8F%D0%B7%D1%8B%D0%BA%D0%B8_%D1%80%D0%B0%D0%B7%D0%BC%D0%B5%D1%82%D0%BA%D0%B8))». С помощью HTML можно легко создать относительно простой, но красиво оформленный документ. Помимо упрощения структуры документа, в HTML внесена поддержка [гипертекста](https://ru.wikipedia.org/wiki/%D0%93%D0%B8%D0%BF%D0%B5%D1%80%D1%82%D0%B5%D0%BA%D1%81%D1%82). [Мультимедийные](https://ru.wikipedia.org/wiki/%D0%9C%D1%83%D0%BB%D1%8C%D1%82%D0%B8%D0%BC%D0%B5%D0%B4%D0%B8%D0%B0) возможности были добавлены позже.

Первым общедоступным описанием HTML был документ «Теги HTML», впервые упомянутый в Интернете Тимом Бернерсом-Ли в конце 1991 года,. В нём описываются 18 элементов, составляющих первоначальный, относительно простой дизайн HTML. За исключением тега гиперссылки, на них сильно повлиял SGMLguid, внутренний формат документации, основанный на стандартном обобщенном языке разметки (SGML), в CERN. Одиннадцать из этих элементов всё ещё существуют в HTML 4.

Изначально язык HTML был задуман и создан как средство структурирования и форматирования документов без их привязки к средствам воспроизведения (отображения). В идеале, текст с разметкой HTML должен был без стилистических и структурных искажений воспроизводиться на оборудовании с различной технической оснащённостью (цветной экран современного компьютера, монохромный экран органайзера, ограниченный по размерам экран мобильного телефона или устройства и программы голосового воспроизведения текстов). Однако современное применение HTML очень далеко от его изначальной задачи. Например, тег <table> предназначен для создания в документах таблиц, но иногда используется и для оформления размещения элементов на странице. С течением времени основная идея платформонезависимости языка HTML была принесена в жертву современным потребностям в мультимедийном и графическом оформлении.

**Абстрактное синтаксическое дерево** (АСД) — в [информатике](https://ru.wikipedia.org/wiki/%D0%98%D0%BD%D1%84%D0%BE%D1%80%D0%BC%D0%B0%D1%82%D0%B8%D0%BA%D0%B0) [конечное](https://ru.wikipedia.org/wiki/%D0%9A%D0%BE%D0%BD%D0%B5%D1%87%D0%BD%D0%BE%D0%B5_%D0%BC%D0%BD%D0%BE%D0%B6%D0%B5%D1%81%D1%82%D0%B2%D0%BE) помеченное ориентированное [дерево](https://ru.wikipedia.org/wiki/%D0%94%D0%B5%D1%80%D0%B5%D0%B2%D0%BE_(%D1%82%D0%B5%D0%BE%D1%80%D0%B8%D1%8F_%D0%B3%D1%80%D0%B0%D1%84%D0%BE%D0%B2)), в котором внутренние вершины сопоставлены (помечены) с [операторами](https://ru.wikipedia.org/wiki/%D0%9E%D0%BF%D0%B5%D1%80%D0%B0%D1%82%D0%BE%D1%80_(%D0%BF%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D0%BC%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5)) языка программирования, а листья — с соответствующими [операндами](https://ru.wikipedia.org/wiki/%D0%9E%D0%BF%D0%B5%D1%80%D0%B0%D0%BD%D0%B4). Таким образом, листья являются пустыми операторами и представляют только переменные и константы.

Синтаксические деревья используются в [парсерах](https://ru.wikipedia.org/wiki/%D0%A1%D0%B8%D0%BD%D1%82%D0%B0%D0%BA%D1%81%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B8%D0%B9_%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7) для [промежуточного представления](https://ru.wikipedia.org/w/index.php?title=%D0%9F%D1%80%D0%BE%D0%BC%D0%B5%D0%B6%D1%83%D1%82%D0%BE%D1%87%D0%BD%D0%BE%D0%B5_%D0%BF%D1%80%D0%B5%D0%B4%D1%81%D1%82%D0%B0%D0%B2%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5&action=edit&redlink=1) программы между [деревом разбора](https://ru.wikipedia.org/w/index.php?title=%D0%94%D0%B5%D1%80%D0%B5%D0%B2%D0%BE_%D1%80%D0%B0%D0%B7%D0%B1%D0%BE%D1%80%D0%B0&action=edit&redlink=1) (конкретным синтаксическим деревом) и [структурой данных](https://ru.wikipedia.org/wiki/%D0%A1%D1%82%D1%80%D1%83%D0%BA%D1%82%D1%83%D1%80%D0%B0_%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D1%85), которая затем используется в качестве внутреннего представления в [компиляторе](https://ru.wikipedia.org/wiki/%D0%9A%D0%BE%D0%BC%D0%BF%D0%B8%D0%BB%D1%8F%D1%82%D0%BE%D1%80) или [интерпретаторе](https://ru.wikipedia.org/wiki/%D0%98%D0%BD%D1%82%D0%B5%D1%80%D0%BF%D1%80%D0%B5%D1%82%D0%B0%D1%82%D0%BE%D1%80) [компьютерной программы](https://ru.wikipedia.org/wiki/%D0%9A%D0%BE%D0%BC%D0%BF%D1%8C%D1%8E%D1%82%D0%B5%D1%80%D0%BD%D0%B0%D1%8F_%D0%BF%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D0%BC%D0%B0) для оптимизации и генерации кода. Возможные варианты подобных структур описываются абстрактным синтаксисом.

Абстрактное синтаксическое дерево отличается от дерева разбора тем, что в нём отсутствуют узлы и рёбра для тех синтаксических правил, которые не влияют на семантику программы. Классическим примером такого отсутствия являются группирующие скобки, так как в АСД группировка операндов явно задаётся структурой дерева.

Для языка, который описывается [контекстно-свободной грамматикой](https://ru.wikipedia.org/wiki/%D0%9A%D0%BE%D0%BD%D1%82%D0%B5%D0%BA%D1%81%D1%82%D0%BD%D0%BE-%D1%81%D0%B2%D0%BE%D0%B1%D0%BE%D0%B4%D0%BD%D0%B0%D1%8F_%D0%B3%D1%80%D0%B0%D0%BC%D0%BC%D0%B0%D1%82%D0%B8%D0%BA%D0%B0), какими являются почти все языки программирования, создание абстрактного дерева в [синтаксическом анализаторе](https://ru.wikipedia.org/wiki/%D0%A1%D0%B8%D0%BD%D1%82%D0%B0%D0%BA%D1%81%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B8%D0%B9_%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7%D0%B0%D1%82%D0%BE%D1%80) является тривиальной задачей. Большинство правил в грамматике создают новую вершину, а символы в правиле становятся рёбрами. Правила, которые ничего не привносят в АСД (например, группирующие правила), просто заменяются в вершине одним из своих символов. Кроме того, анализатор может создать полное дерево разбора и затем пройти по нему, удаляя узлы и рёбра, которые не используются в абстрактном синтаксисе, для получения АСД.



Рисунок 1. AST дерево.

## **Описание существующих разработок**

Есть несколько вариантов:

1. Оптимальный — если в штате есть программист (а еще лучше — несколько программистов). Поставьте задачу, опишите требования и получите готовый инструмент, заточенный конкретно под ваши задачи. Инструмент можно будет донастраивать и улучшать при необходимости.
2. Воспользоваться готовыми облачными парсерами (есть как бесплатные, так и платные сервисы).
3. Десктопные парсеры — как правило, программы с мощным функционалом и возможностью гибкой настройки. Но почти все — платные.
4. Заказать разработку парсера «под себя» у компаний, специализирующихся на разработке (этот вариант явно не для желающих сэкономить).

Первый вариант подойдет далеко не всем, а последний вариант может оказаться слишком дорогим.

Что касается готовых решений, их достаточно много, и если вы раньше не сталкивались с парсингом, может быть сложно выбрать.

### **ParserOK.**

**Программа «Парсер сайтов»** разработана для сбора, анализа, выборки, группировки, структуризации, трансформации данных с последующим выводом данных в таблицу **Excel** в форматах **xls\*** и **csv**.

Парсер создан на **VBA** (Visual Basic for Applications) и представлен в виде **надстройки** для **MS Excel**, по сути это набор макросов, каждый набор отвечает за выполнение определенных функций при обработке.

Таким образом, для работы программы необходимы: файл надстройки **Parser.xla** и файл управления надстройкой **Name.xlp** (Name — имя файла).

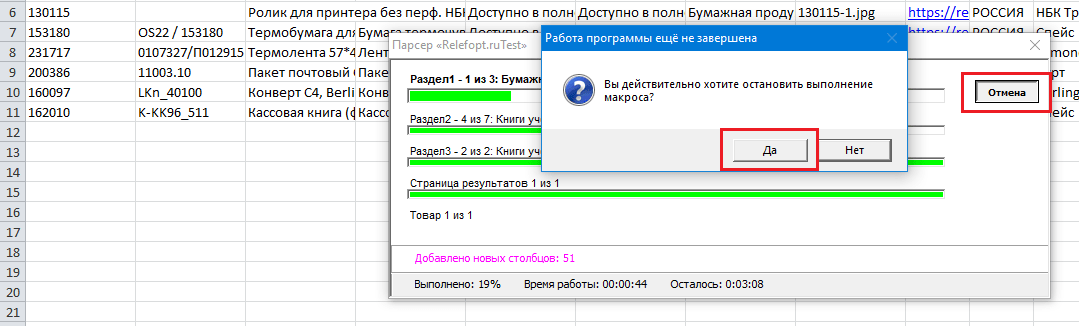


Рисунок 2. Работа с ParserOK

### **Datacol**

Это универсальный парсер, который умеет собирать данные с любого сайта или файла, дополнительно их обработать и сохранять результат работы в файл, базу данных или сразу грузить на ваш сайт.

Перед запуском кампании задайте входные данные - название товаров, прайс лист или ссылки на категории сайта, из которых нужно спарсить данные.

После запуска кампании за ее работой можно следить в блоке “Статистика работы”. Если все настроено верно, первые результаты вскоре появятся внизу программы.

Собранные данные можно [дополнительно обрабатывать](http://web-data-extractor.net/obrabotka-dannyh/): [переводить](http://web-data-extractor.net/obrabotka-dannyh/avtomaticheskij-perevodchik-eaws/), [добавлять наценку](http://web-data-extractor.net/plugins/d5plugin_pricechanger/) к цене поставщика, [уникализировать](http://web-data-extractor.net/d5plugin_synonymaru/) текст, [переводить](http://web-data-extractor.net/dynamic-plugins/preobrazovaniya-chislovogo-znacheniya/) в другую валюту и многое другое. Это возможно благодаря плагинам, которые дополняют и расширяют функционал Datacol. Доступные плагины можно найти [здесь](http://web-data-extractor.net/plugins/) и [здесь](http://web-data-extractor.net/dynamic-plugins/). Если нужного вам плагина нет, его разработку [можно заказать](http://web-data-extractor.net/services/) у наших специалистов.

По умолчанию, результаты парсинга в базовых кампаниях автоматически сохраняются в EXCEL файл с таким же названием, что и кампания парсинга. Файл можно найти в папке “Мои документы”. При необходимости, в настройках кампании можно изменить как параметры сохранения, так их формат данных. Можно отправить данные на FTP, в базу данных MySQL, экспортировать в потоковом режиме или сохранить все в CSV файл нужной структуры.

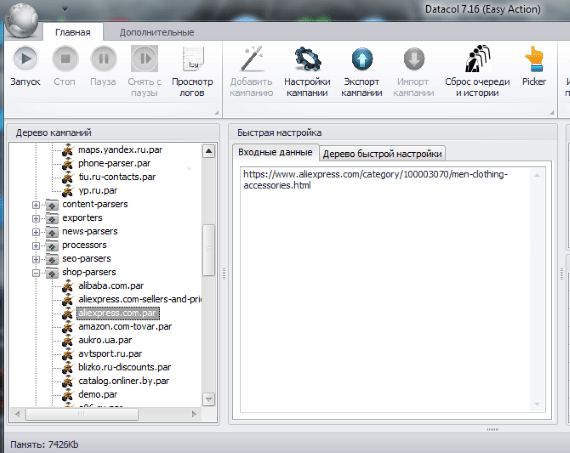


Рисунок 3. Работа в программе Datacol

# **Проектная часть**

## **Диаграмма прецедентов**

В этом разделе представлена диаграмма прецедентов. На диаграмме (Рисунок 3) показаны все возможные функциональные отношения.

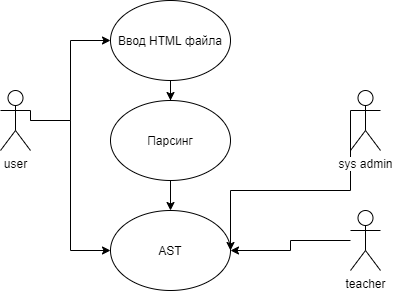
**

Рисунок 4. Диаграмма прецедентов работы парсера

## **Выбор инструментов**

При выборе инструментов я исходил из некоторых критериев.

Важность критерия я выбирал из: низкая, ниже средней, средняя, ниже высокой, высокая.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Критерий | Участие в корпоративном проекте | Простота сопровождения | Наличие библиотек | Наличие документации на русском языке | Скорость разработки |
| Важность критерия | Ниже средней | Средняя | Высокая | Ниже высокой | Ниже высокой |

Исходя из этих критериев, я сравнил 3 языка программирования от 0 до 10 баллов за критерий.

|  |  |  |  |
| --- | --- | --- | --- |
|  | C++ | Python | Object Pascal |
| Участие в корпоративном проекте | 5 | 10 | 1 |
| Простота сопровождения | 2 | 10 | 2 |
| Наличие библиотек | 5 | 10 | 3 |
| Наличие документации на русском языке | 7 | 10 | 7 |
| Скорость разработки | 1 | 10 | 3 |
| Итого баллов | 20 | 50 | 16 |

По результатам сравнения был выбран язык программирования Python.

## **Проектирование сценария**

В данном разделе приведен сценарий использования программы пользователем (Рисунок 4).

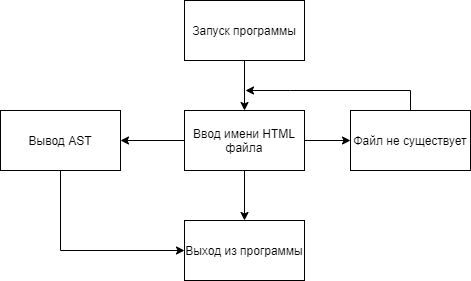


Рисунок 5. Сценарий использования

Пользователь после запуска программы вводит название HTML файла или может выйти из программы.

При выборе выхода программа заканчивает свою работу, при вводе вводе имени начинается проверка ну существование файла, если файл не существует, то предлагается ввести имя еще раз, если существует, то выводится сгенерированное AST дерево HTML файла и выходит из программы.

## **Диаграмма классов**

В данном разделе представлены все классы, использующиеся в проекте, (Рисунок 6).

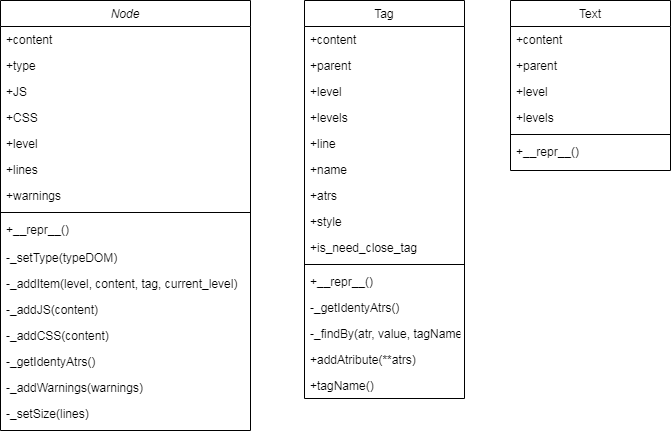


Рисунок 6. Диаграмма классов для проекта

## **Описание главного модуля**

В главный модуль подключается файл parserHTML, в котором прописаны все связи остальных файлов и функции. Далее в главном модуле находится функция получения AST дерева из имени файла, получение имени файла, проверка ошибок и вывод его.

**Листинг 1. Функция получения AST дерева.**

def getDocument():

filename = input("\nName of HTML file ('q' - exit): ")

if filename == 'q':

sys.exit()

while not os.path.isfile(filename) or not 'html' in filename.split('.') or len(filename.split('.')) < 2:

if filename == 'q':

sys.exit()

if not 'html' in filename.split('.') or len(filename.split('.')) < 2:

filename = input("\n\033[41m{}\033[40m\n".format("File '" + filename +"' is not HTML file.") + "\nPlease enter correct name of HTML file ('q' - exit): ")

else:

filename = input("\n\033[41m{}\033[40m\n".format("File '" + filename +"' is not exist.") + "\nPlease enter correct name of HTML file ('q' - exit): ")

document = parserHTML(filename) #парсируем файл

if document:

(document, (len\_err, err)) = document

print("\n\033[42m{}\033[40m\n".format("Parsed completed!"))

if len\_err:

print("\033[30m\033[43m{}\033[37m\033[40m\n".format(

f"{len\_err} warning in {','.join(' < ' + e + ' > on line ' + str(l) for (e, l) in err)}")

)

return document

else:

return None

Данный метод получает имя файла, открывает его, читает из него текст, передает в лексер и получает из него токены, их же в свою очередь передает в парсер, который возвращает AST дерево.

**Листинг 2. Главная функция.**

if \_\_name\_\_ == "\_\_main\_\_":

print("HTML Parser v.2.3.1.610 (released 13.05.2020). Created by OVOSKOP.")

print('Type "help" for more information.')

document = getDocument()

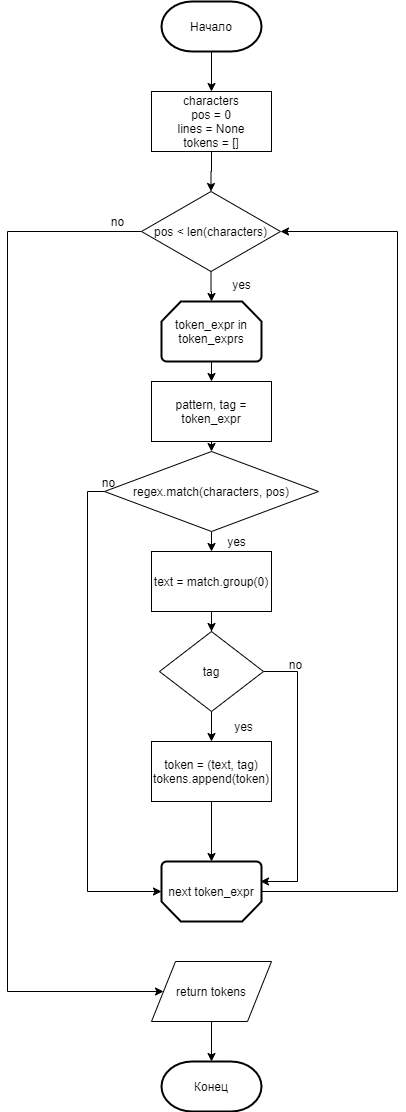
while not document:

document = getDocument()

print('\n')

print(document)

**Блок-схема лексера**

****

## **Описание спецификаций к модулям**

Разработаны 5 модулей, не включая главный:

* Классы и методы самого парсера
* Функция парсера
* Конфигурации лексера
* Функция лексера
* Подключение лексера и парсера в одну функцию

Структура подключения модулей изображена на Рисунке 7.

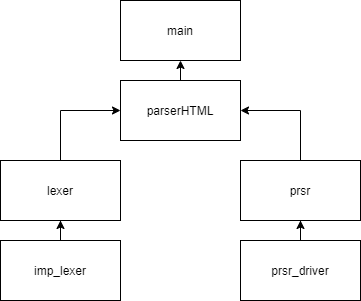


Рисунок 7 .Структура подключения модулей

* 1. **Описание спецификаций к модулям**

**Модуль 1. Классы и методы самого парсера**

Прописываются классы Node, Tag, Text. Реализуются поля и методы классов.

Листинг 1. Создание класса Text

class Text:

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

self.content = text

self.parent = None

self.level = 0

self.levels = []

def \_\_repr\_\_(self):

return self.content

**Модуль 2. Функция парсера**

Прописывается основная функция парсера.

Листинг 2. Функция парсера

def parser(tokens):

buff = []

level = 0

err = []

lines = 1

doc = Node()

for token in tokens:

# print(token)

if token[1] == 'OPEN\_TAG':

tag = Tag(token[0], lines)

doc.\_addItem(level, tag)

buff.append(tag)

level += 1

elif token[1] == 'CLOSE\_TAG':

if buff[-1].tagName() == token[0][0][0]:

buff.pop()

level -= 1

else:

# print( buff[-1])

while buff[-1].tagName() != token[0][0][0]:

err.append((buff[-1].name + buff[-1].\_getIdentyAtrs(), buff[-1].line))

buff.pop()

level -= 1

buff.pop()

level -= 1

# sys.stderr.write('Illegal token: %s\n' % str(token))

# sys.exit(1)

else:

if token[1] == 'TAG':

doc.\_addItem(level, Tag(token[0], lines, is\_need\_close\_tag=False))

elif token[1] == 'SCRIPT':

doc.\_addJS(token[0])

doc.\_addItem(level, Tag([['script']], lines))

elif token[1] == 'STYLE':

doc.\_addCSS(token[0])

doc.\_addItem(level, Tag([['style']], lines))

elif token[1] == 'TYPE':

doc.\_setType(token[0])

elif token[1] == 'LINE':

lines += 1

else:

doc.\_addItem(level, Text(token[0]))

# print(buff)

# print(level, err)

# print(lines)

doc.\_setSize(lines)

for item in buff:

err.append((item.name + item.\_getIdentyAtrs(), item.line))

doc.\_addWarnings(err)

return (doc, (len(err), err))

**Модуль 3. Конфигурации лексера**

В данном модуле прописываются все паттерны и конфиги для лексера.

Листинг 3. Список токенов для лексера

token\_exprs = [

(r'[ \t]+', None),

(r'\n', LINE),

(r'<!--', COMMENT),

(r'<script( |\n)\*([-\w;: ]+(=(")?[^>]\*(")?)?)\*(( )?/)?>', SCRIPT),

(r'<style( |\n)\*([-\w;: ]+(=(")?[^>]\*(")?)?)\*(( )?/)?>', STYLE),

(r'<\/[\w-]\*>', CLOSE\_TAG),

(r'<!DOCTYPE [ \w.:\/\-\"]+>', TYPE),

(r'(([^<>\s])|( )|( )|( )|( ))+', CONTENT),

(r'<((area)|(base)|(br)|(col(?!g))|(command)|(embed)|(hr)|(img)|(input)|(keygen)|(link)|(meta)|' +

r'(param)|(source)|(track)|(wbr))( |\n)\*([-\w;: ]+(=(")?[^>]\*(")?)?)\*(( )?/)?>', TAG),

(r'<([\w-]+)( |\n)\*([-\w;: ]+(=(")?[^>]\*(")?)?)\*(( )?/)?>', OPEN\_TAG),

]

**Модуль 4. Функция лексера**

В данном модуле реализована функция лексера, которая создает из текста HTML многомерный массив, который в последствии превращается в AST

Листинг 4. Функция лексирования тегов

def lexTag(characters, token\_exprs):

pos = 0

lines = 0

tokens = []

while pos < len(characters):

match = None

for token\_expr in token\_exprs:

pattern, tag = token\_expr

regex = re.compile(pattern)

match = regex.match(characters, pos)

if match:

text = match.group(0)

if tag:

if tag == 'LINE':

lines += 1

else:

token = (text, tag)

tokens.append(token)

break

if not match:

print("\n\033[41m{}\033[40m".format('Invalid tag: "%s" ' % (characters[pos:pos+10:]) + 'in pos: %s' % str(pos)))

return None

else:

pos = match.end(0)

return (tokens, lines)

**Модуль 5. Подключение лексера и парсера в одну функцию**

В данном модуле все функции соединяются в одну, которая получает текст HTML, а возвращает AST

Листинг 5. Функция полного парсинга HTML и отслеживание ошибок

def parserHTML(filename):

if filename:

for enc in encoding:

try:

file = open(filename, encoding=enc)

characters = file.read()

file.close()

except (UnicodeDecodeError, LookupError):

pass

else:

break

document = ""

if characters:

tokens = imp\_lex(characters) #лексируем файл

if tokens:

document = parser(tokens) #парсируем файл

# for token in tokens:

# print (token)

if document != "":

return document

print("\n\033[41m{}\033[40m\n".format("Parsed not completed! Invalid File!"))

return None

## **Описание тестовых наборов модулей**

В каждом модуле требуется вводить определенные значения. Там, где нужно ввести имя файла, идет проверка на существование такого файла, после идет проверка на правильность кодировки, далее на валидность HTML файла лексером, а после парсером.

**Тест 1. Ввод неправильного имени**

Если файл не существует, то вызовется ошибка (рис. 8), и потребуется заново писать имя файла.

****

Рисунок 8. Ошибка при вводе неверного пути к файлу

**Тест 2. Ввод файла в неправильной кодировке**

Eсли файл в неправильной кодировке, то вызовется ошибка, что файл не корректный

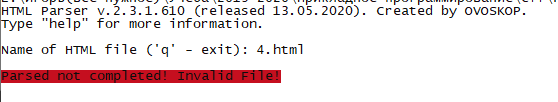


Рисунок 9. Ошибка при вводе файла в неправильной кодировке

**Тест 3. Ввод файла с ошибкой**

При вводе файла с ошибкой, парсер выдаст предупреждение об ошибке, в каком теге и на какой строке

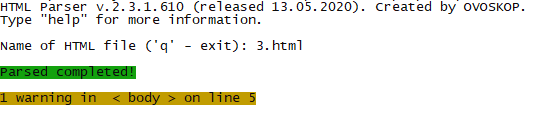


Рисунок 10. Ввод файла с ошибкой

## **Описание применения средств отладки**

В ходе написания курсового проекта при попытке запустить программу были получены ошибки:

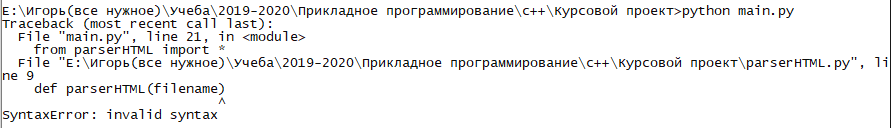


Рисунок 11. Ошибки

При проверке кода были исправлены найденные ошибки, в результате при запуске программы ошибок не было:

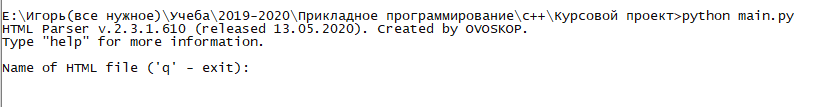


Рисунок 12. Сообщение об ошибках

## **Анализ оптимальности использования памяти и быстродействия**

Сложность нерекурсивного алгоритма равна O(n), соответственно, сложность алгоритма добавления тега в AST равна O(n), что при очень больших кодах сайта заметно тормозит работу программы.

# **Эксплуатационная часть**

## **Руководство операторa**

Программа main.py выдаст сообщение об ошибке, показанное на рис. 13

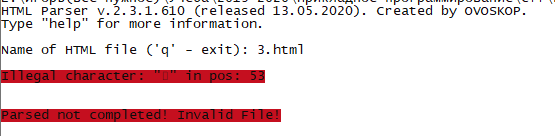


Рисунок 13. Ошибка

ПРИЧИНА. В коде HTML присутствует символ, не прописанный в темплейтах лексера.

ДЕЙСТВИЯ ПРОГРАММЫ. Программа не парсит данный сайт.

ДЕЙСТВИЯ ОПЕРАТОРА. Добавить соответствующий символ в темплейты лексера.

## **To-Do лист**

* Создать интерактивную консольную программу для работы с HTML и модифицированием его

# **Заключение**

В результате выполнения курсовой работы был разработан парсер HTML. Данная программа имеет потенциал для будущих разработок и ее можно будет использовать для более крупных проектов. Разработанная программа выполнена в соответствии с требованиями технического задания.

# **Список литературы и интернет-источников**

1. Что такое Datacol? [1.2.2]

<http://web-data-extractor.net/datacol-s-chego-nachat/>

1. Парсер сайтов [1.2.1]

<https://parserok.ru/>

1. 30+ парсеров для сбора данных с любого cайта [1]

<https://habr.com/ru/company/click/blog/494020/#:~:text=%D0%9F%D0%B0%D1%80%D1%81%D0%B5%D1%80%20%E2%80%94%20%D1%8D%D1%82%D0%BE%20%D0%BF%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D0%BC%D0%B0%2C%20%D1%81%D0%B5%D1%80%D0%B2%D0%B8%D1%81%20%D0%B8%D0%BB%D0%B8,%D0%A6%D0%B5%D0%BD%D1%8B.>

1. Абстрактное синтаксическое дерево [1]

<https://ru.wikipedia.org/wiki/%D0%90%D0%B1%D1%81%D1%82%D1%80%D0%B0%D0%BA%D1%82%D0%BD%D0%BE%D0%B5_%D1%81%D0%B8%D0%BD%D1%82%D0%B0%D0%BA%D1%81%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%BE%D0%B5_%D0%B4%D0%B5%D1%80%D0%B5%D0%B2%D0%BE>

1. HTML [1]

<https://ru.wikipedia.org/wiki/HTML>

# **Приложения**

**Приложение 1. Код главного модуля.**

## HTML PARSER V.2.3.1.610 (Count Lines, Fixed bugs, Correct output of warnings)

##

## DEVELOPER: OVOSKOP

##

## COPYRIGHT. ALL RIGHT RESERVED.

##

## CLASSES:

##

## TAG - Object of Tag

## Node - Document Object Model or Node of Tags

##

##

import sys

import os.path

import re

import inspect

import ctypes

from parserHTML import \*

kernel32 = ctypes.windll.kernel32

kernel32.SetConsoleMode(kernel32.GetStdHandle(-11), 7)

def getDocument():

filename = input("\nName of HTML file ('q' - exit): ")

if filename == 'q':

sys.exit()

while not os.path.isfile(filename) or not 'html' in filename.split('.') or len(filename.split('.')) < 2:

if filename == 'q':

sys.exit()

if not 'html' in filename.split('.') or len(filename.split('.')) < 2:

filename = input("\n\033[41m{}\033[40m\n".format("File '" + filename +"' is not HTML file.") + "\nPlease enter correct name of HTML file ('q' - exit): ")

else:

filename = input("\n\033[41m{}\033[40m\n".format("File '" + filename +"' is not exist.") + "\nPlease enter correct name of HTML file ('q' - exit): ")

document = parserHTML(filename) #парсируем файл

if document:

(document, (len\_err, err)) = document

print("\n\033[42m{}\033[40m\n".format("Parsed completed!"))

if len\_err:

print("\033[30m\033[43m{}\033[37m\033[40m\n".format(

f"{len\_err} warning in {','.join(' < ' + e + ' > on line ' + str(l) for (e, l) in err)}")

)

return document

else:

return None

if \_\_name\_\_ == "\_\_main\_\_":

print("HTML Parser v.2.3.1.610 (released 13.05.2020). Created by OVOSKOP.")

print('Type "help" for more information.')

document = getDocument()

while not document:

document = getDocument()

print('\n')

print(document)

**Приложение 2. Код модуля parserHTML.**

from imp\_lexer import \*

from prsr import \*

encoding = [

'utf-8',

'windows-1251',

]

def parserHTML(filename):

if filename:

for enc in encoding:

try:

file = open(filename, encoding=enc)

characters = file.read()

file.close()

except (UnicodeDecodeError, LookupError):

pass

else:

break

document = ""

if characters:

tokens = imp\_lex(characters) #лексируем файл

if tokens:

document = parser(tokens) #парсируем файл

# for token in tokens:

# print (token)

if document != "":

return document

print("\n\033[41m{}\033[40m\n".format("Parsed not completed! Invalid File!"))

return None

**Приложение 3. Код модуля imp\_lexer.**

import lexer

TAG = 'TAG'

OPEN\_TAG = 'OPEN\_TAG'

CLOSE\_TAG = 'CLOSE\_TAG'

TYPE = 'TYPE'

CONTENT = 'CONTENT'

SCRIPT = 'SCRIPT'

STYLE = 'STYLE'

COMMENT = 'COMMENT'

LINE = 'LINE'

TAG\_NAME = 'TAG\_NAME'

ATRIBUTE = 'ATRIBUTE'

ATRIBUTE\_NAME = 'ATRIBUTE\_NAME'

VALUE = 'VALUE'

token\_exprs = [

(r'[ \t]+', None),

(r'\n', LINE),

(r'<!--', COMMENT),

(r'<script( |\n)\*([-\w;: ]+(=(")?[^>]\*(")?)?)\*(( )?/)?>', SCRIPT),

(r'<style( |\n)\*([-\w;: ]+(=(")?[^>]\*(")?)?)\*(( )?/)?>', STYLE),

(r'<\/[\w-]\*>', CLOSE\_TAG),

(r'<!DOCTYPE [ \w.:\/\-\"]+>', TYPE),

(r'(([^<>\s])|( )|( )|( )|( ))+', CONTENT),

(r'<((area)|(base)|(br)|(col(?!g))|(command)|(embed)|(hr)|(img)|(input)|(keygen)|(link)|(meta)|' +

r'(param)|(source)|(track)|(wbr))( |\n)\*([-\w;: ]+(=(")?[^>]\*(")?)?)\*(( )?/)?>', TAG),

(r'<([\w-]+)( |\n)\*([-\w;: ]+(=(")?[^>]\*(")?)?)\*(( )?/)?>', OPEN\_TAG),

]

token\_tag = [

[

(r'[ \t<>/]+', None),

(r'\n', LINE),

(r'(?<=<)[\w-]+', TAG\_NAME),

(r'(?<=</)[\w-]+', TAG\_NAME),

(r'["\w;:-]+(=("|\')?[^\'"]\*("|\')?)?', ATRIBUTE),

],

]

def imp\_lex(characters):

return lexer.lex(characters, token\_exprs, token\_tag)

**Приложение 4. Код модуля prsr.**

def parser(tokens):

buff = []

level = 0

err = []

lines = 1

doc = Node()

for token in tokens:

# print(token)

if token[1] == 'OPEN\_TAG':

tag = Tag(token[0], lines)

doc.\_addItem(level, tag)

buff.append(tag)

level += 1

elif token[1] == 'CLOSE\_TAG':

if buff[-1].tagName() == token[0][0][0]:

buff.pop()

level -= 1

else:

# print( buff[-1])

while buff[-1].tagName() != token[0][0][0]:

err.append((buff[-1].name + buff[-1].\_getIdentyAtrs(), buff[-1].line))

buff.pop()

level -= 1

buff.pop()

level -= 1

# sys.stderr.write('Illegal token: %s\n' % str(token))

# sys.exit(1)

else:

if token[1] == 'TAG':

doc.\_addItem(level, Tag(token[0], lines, is\_need\_close\_tag=False))

elif token[1] == 'SCRIPT':

doc.\_addJS(token[0])

doc.\_addItem(level, Tag([['script']], lines))

elif token[1] == 'STYLE':

doc.\_addCSS(token[0])

doc.\_addItem(level, Tag([['style']], lines))

elif token[1] == 'TYPE':

doc.\_setType(token[0])

elif token[1] == 'LINE':

lines += 1

else:

doc.\_addItem(level, Text(token[0]))

# print(buff)

# print(level, err)

# print(lines)

doc.\_setSize(lines)

for item in buff:

err.append((item.name + item.\_getIdentyAtrs(), item.line))

doc.\_addWarnings(err)

return (doc, (len(err), err))

**Приложение 5. Код модуля prsr\_driver.**

# \*\*\*\*\* CLASSES \*\*\*\*\*

class Text:

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

self.content = text

self.parent = None

self.level = 0

self.levels = []

def \_\_repr\_\_(self):

return self.content

# Tag

class Tag:

def \_\_init\_\_(self, args, line, is\_need\_close\_tag = True):

# global styles

self.content = []

self.parent = None

self.level = 0

self.levels = []

self.line = line

self.name = args[0][0]

self.atrs = {}

self.style = {}

# if '\*' in styles:

# self.style.update(styles['\*'])

# if self.name in styles:

# self.style.update(styles[self.name])

self.is\_need\_close\_tag = is\_need\_close\_tag

for atr in args[1:]:

if atr[1] == "ATRIBUTE":

atr\_value = None

if '=' in atr[0]:

atr\_name = atr[0].split('=')[0].replace(' ', '')

atr\_value = atr[0].split('=')[1]

else:

atr\_name = atr[0].replace(' ', '')

if atr\_value:

if atr\_value[0] == '"' and atr\_value[-1] == '"' or atr\_value[0] == "'" and atr\_value[-1] == "'":

atr\_value = atr\_value[1:-1]

self.addAtribute(\*\*{atr\_name: (atr\_value if atr\_value else "")})

# \*\*\*\* STYLES \*\*\*\*

# if self.name == "link" and "rel" in self.atrs:

# if self.atrs["rel"] == "stylesheet":

# styles = updateDict(styles, getStyle(self.atrs["href"]))

def \_\_repr\_\_(self):

# print(self.name, self.level, self.levels)

tabs = ""

if self.level >= 1:

tabs = ''.join(['| ' if self.levels[i] else ' ' for i in range(self.level)])

line = f"{self.name + self.\_getIdentyAtrs()}"

# atributes

# for atr in self.atrs:

# line += f"\n{tabs} {atr + ' : ' + self.atrs[atr]}"

for item in self.content:

typeElem = str(type(item)).split("'")[1].split(".")

typeElem = typeElem[1] if len(typeElem) > 1 else typeElem[0]

item.level = self.level + 1

if self.content.index(item) < len(self.content) - 1:

item.levels = [\*self.levels, 1]

else:

item.levels = [\*self.levels, 0]

line += f"\n{tabs}|\_\_\_\_" + '%r' % item

self.level = 0

self.levels = []

return line

def \_getIdentyAtrs(self):

identy = ""

if 'id' in self.atrs:

for ids in self.atrs['id']:

if len(ids) > 0:

identy += "#" + ids

if 'class' in self.atrs:

for classes in self.atrs['class']:

if len(classes) > 0:

identy += "." + classes

return identy

def \_findBy(self, atr = None, value = None, tagName = None):

elems = []

for elem in self.content:

typeElem = str(type(elem)).split("'")[1].split(".")

typeElem = typeElem[1] if len(typeElem) > 1 else typeElem[0]

# print(typeElem, value)

if typeElem == "Tag":

if not tagName:

if atr in elem.atrs:

if value:

if elem.atrs[atr] == value or \

value in elem.atrs[atr]:

elems.append(elem)

else:

elems.append(elem)

elems.extend(elem.\_findBy(atr, value))

else:

if elem.name == tagName:

elems.append(elem)

elems.extend(elem.\_findBy(tagName=tagName))

return elems

def addAtribute(self, \*\*atrs):

for atr, value in atrs.items():

if atr == 'style':

pass

elif atr == "class" or atr == "id":

self.atrs[atr].extend(value.split(" ")) if (atr in self.atrs) else self.atrs.update({atr: value.split(" ")})

else:

self.atrs[atr] = value

return True

def tagName(self):

return self.name

class Node:

def \_\_init\_\_(self, content = None):

self.content = []

self.type = None

self.JS = []

self.CSS = []

self.level = -1

self.lines = 0

self.warnings = None

if content:

self.content.append(content)

def \_\_repr\_\_(self):

line = ""

for item in self.content:

line += f"{str(item)}\n"

# line += str(styles)

return line

def \_setType(self, typeDOM):

self.type = typeDOM.split(" ")[1].split(">")[0]

def \_addItem(self, level, content, tag = None, current\_level = 0):

if not tag:

tag = self

if level == current\_level:

typeElem = str(type(content)).split("'")[1].split(".")

typeElem = typeElem[1] if len(typeElem) > 1 else typeElem[0]

if typeElem == "Tag":

content.parent = tag

tag.content.append(content)

return tag

tag.content[len(tag.content) - 1] = self.\_addItem(level, content, tag.content[len(tag.content) - 1], current\_level + 1)

return tag

def \_addJS(self, content):

self.JS.append(content)

def \_addCSS(self, content):

self.CSS.append(content)

def \_getIdentyAtrs(self):

return ""

def \_addWarnings(self, warnings):

self.warnings = warnings

def \_setSize(self, lines):

self.lines = lines

**Приложение 6. Код модуля lexer.**

import re

def lex(characters, token\_exprs, sub\_token\_exprs=None):

pos = 0

lines = None

tokens = []

while pos < len(characters):

l = 0

match = None

for token\_expr in token\_exprs:

# print(token\_exprs)

#print()

pattern, tag = token\_expr

# print('pattern' + pattern)

# print('pos = ' + str(pos), 'char ' + characters[pos])

regex = re.compile(pattern)

# print(regex)

match = regex.match(characters, pos)

# print(match)

if match:

# print('match' + match)

text = match.group(0)

if tag:

# print(text, tag)

if (tag == "TAG" or tag == "CLOSE\_TAG" or tag == "OPEN\_TAG") and sub\_token\_exprs:

(text, l) = lexTag(text, sub\_token\_exprs[0])

# print(l)

if tag == "SCRIPT" or tag == "STYLE":

start = match.end(0)

regex = re.compile(r'<\/(script|style)>')

match = regex.search(characters, match.end(0))

if match:

end = match.start(0)

text = characters[start:end:]

l = text.count('\n')

if tag == 'COMMENT':

regex = re.compile(r'-->')

start = match.end(0)

match = regex.search(characters, start)

if match:

end = match.start(0)

text = characters[start:end:]

l = text.count('\n')

for i in range(l):

tokens.append(('\n', 'LINE'))

break

token = (text, tag)

# print("\n\n")

# print(token)

tokens.append(token)

for i in range(l):

tokens.append(('\n', 'LINE'))

# print(('\n', 'LINE'))

break

if not match:

print("\n\033[41m{}\033[40m".format('Illegal character: "%s" ' % (characters[pos:pos+10:]) + 'in pos: %s' % str(pos)))

return None

else:

pos = match.end(0)

return tokens

def lexTag(characters, token\_exprs):

pos = 0

lines = 0

tokens = []

while pos < len(characters):

match = None

for token\_expr in token\_exprs:

pattern, tag = token\_expr

regex = re.compile(pattern)

match = regex.match(characters, pos)

if match:

text = match.group(0)

if tag:

if tag == 'LINE':

lines += 1

else:

token = (text, tag)

tokens.append(token)

break

if not match:

print("\n\033[41m{}\033[40m".format('Invalid tag: "%s" ' % (characters[pos:pos+10:]) + 'in pos: %s' % str(pos)))

return None

else:

pos = match.end(0)

return (tokens, lines)