**Всероссийское СМИ**
**«Академия педагогических идей «НОВАЦИЯ»**
**Свидетельство о регистрации ЭЛ №ФС 77-62011 от 05.06.2015 г.**
**(выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций)**

**Сайт:** akademnova.ru
**e-mail:** akademnova@mail.ru

***Журавлев М.П. Использование языка Python при разработке парсера // Академия педагогических идей «Новация». – 2017. – № 06 (июнь). – АРТ 72-эл. – 0,2 п. л. – URL: http://akademnova.ru/page/875548***

***РУБРИКА: ПРОФЕССИОНАЛЬНОЕ ОБРАЗОВАНИЕ***

**УДК 004**

**Журавлёв Михаил Петрович**

Магистр, 2 курс, факультет дизайна и программной инженерии
ФГБОУ ВО «Казанский Национальный
Исследовательский Технологический Университет»,
г. Казань, Республика Татарстан, Российская Федерация
e-mail: mihuna@yandex.ru

# ИСПОЛЬЗОВАНИЕ ЯЗЫКА PYTHON ДЛЯ ПОИСКА ПРИОРИТЕТНЫХ ЛОТОВ

*Аннотация*: В статье рассматривается применение языка Python для автоматизированного синтаксического анализа больших структурированных объёмов информации.

*Ключевые слова:* python, парсинг, программирование, синтаксический анализ, условные операторы, оптимизация кода, высокопроизводительные вычисления.

**Zhuravlyev Mihail Petrovich**

Master, 2nd year, Design and Software Engineering Faculty
FSEI HE «Kazan National Research Technological University»
Kazan, Republic of Tatarstan, Russian Federation
e-mail: mihuna@yandex.ru

**Всероссийское СМИ**
**«Академия педагогических идей «НОВАЦИЯ»**
**Свидетельство о регистрации ЭЛ №ФС 77-62011 от 05.06.2015 г.**
**(выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций)**
**Сайт:** akademnova.ru
**e-mail:** akademnova@mail.ru

# USE OF THE PYTHON TO SEARCH FOR PRIORITY LOTS

*Abstract:* The article describes the using of Python for automated parsing large amounts of structured data.

*Keywords:* programming, parsing, conditional statements, code optimization, high-performance computing.

Парсинг (по-русски «синтаксический анализ») — задача, результатом решения которой является осмысленный разбор и преобразование в осмысленные единицы блока информации, описанного на некотором фиксированном языке, будь то язык программирования, язык разметки, язык структурированных запросов и т.п. Типичная последовательность этапов решения задачи выглядит примерно так:

1. **Описание языка**. При ясности получаемого результата устанавливаются как правила построения предложений в языке, так и правила определения валидных слов.
2. **Разбиение ввода на токены**. Пишется лексический анализатор, который позволит разбить вводную строку (файл, блок информации и т.п.) на валидные слова.
3. **Построение синтаксического древа**. Этот этап определяет проверку работы, проведенной на предыдущем этапе. Обычно, эта задача решается с помощью метода рекурсивного спуска, с помощью которого строится валидное синтаксическое древо.

Всероссийское СМИ
«Академия педагогических идей «НОВАЦИЯ»
Свидетельство о регистрации ЭЛ №ФС 77-62011 от 05.06.2015 г.
(выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций)
**Сайт:** akademnova.ru
**e-mail:** akademnova@mail.ru

4. **Выполнение операции над информацией.** После построения синтаксического древа информация готова к компиляции, сборке, переводу, отображению и т.п..

В данной работе рассматривается проблемы моделирования системы поиска конкурсов (лотов) на сайте государственных закупок zakupli.gov.ru для создания списка приоритетных лотов.

Создание списка приоритетных лотов при участии в тендере – весьма не редкое явление, тем более, если организация обладает широким спектром услуг, материальной базой и площадью для ведения деятельности, удовлетворяющей требованиям государственных организаций, так как каждый квадратный метр площади предприятия должен нести материальную выгоду и, как минимум, покрывать затраты на её аренду.

Большая площадь предприятия позволяет вести деятельность больших масштабов, одновременно не прекращая обслуживание частных лиц, без ущерба для клиентопотока. Единственная проблема в данном случае – организовать этот поток. Для решения данной проблемы и производится создание данного приложения.

Чтобы решить проблему качественно, необходимо изучить предметную область, так как сфера государственных закупок, а точнее – тендеров на торговых площадках является весьма сложной и запутанной для обывателя.

Весь механизм торговой площадки реализован на идентификации конечного пользователя (предприятия) на основе ключей ЭЦП (Электронная Цифровая Подпись), которая позволяет удаленно идентифицировать покупателя/заказчика или продавца/производителя. В данном случае система торговой площадки идентифицирует пользователя как юридическое лицо и предоставляет ряд функций и уровень доступа для ведения коммерческой

**Всероссийское СМИ**
**«Академия педагогических идей «НОВАЦИЯ»**
**Свидетельство о регистрации ЭЛ №ФС 77-62011 от 05.06.2015 г.**
**(выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций)**
**Сайт:** akademnova.ru
**e-mail:** akademnova@mail.ru

деятельности. Как правило обладателем ЭЦП является генеральный директор организации, который, при необходимости, создает доверенность на ведение тех или иных сделок, предоставляя право подписи от его имени.

После регистрации пользователю предоставляется личный кабинет. Различается два личных кабинета, отличных по функционалу:

1. Поставщика;
2. Заказчика;
3. На следующем этапе вводятся необходимые данные об организации:
4. Полное и краткое наименование организации;
5. ИНН/КПП (Идентификационный номер налогоплательщика)
6. ОГРН
7. Копия ЕГРЮЛ (Выписка из государственного реестра юридических лиц)
8. Копия Устава организации;
9. Копия документа, подтверждающего право подписи (как правило, доверенность или устав);
10. Контактные данные организации;
11. Логин и пароль предприятия;
12. Защитный код.

Далее создается завка/запрос, в зависимости от выбранного режима регистрации. Зарегистрированная заявка попадает в соответствующий параметрам лота раздел, где формируется каталог со всеми файлами, сопроводительной документацию и открытым ключем заявителя, однако названия папок и файлов не говорят о содержании лота, и, зачастую, несут имена из случайных символов.

**Всероссийское СМИ**
**«Академия педагогических идей «НОВАЦИЯ»**
**Свидетельство о регистрации ЭЛ №ФС 77-62011 от 05.06.2015 г.**
**(выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций)**
**Сайт:** akademnova.ru
**e-mail:** akademnova@mail.ru

Данный вариант определяет использование HTTP, как средства получения необходимой информации. Так как торговая площадка представляет собой портал, или, проще говоря, обыкновенный сайт с гипертекстовой разметкой, эта информация хорошо структурирована и, логически, может быть представлена в следующем виде (графе вложенности):

1. Интернет;
2. Ресурс торговой площадки;
3. Страница запроса (перечень лотов);
4. Лот;
5. Сопроводительные документы;
6. Документ;
7. Комментарий к документу.

Однако, при использовании стандартных методов, процесс поиска подходящей заявки может занять от нескольких часов до нескольких суток, к тому же, данные, которые нужно обработать – нужно собирать вручную, сводить в таблицы для наглядности – все это резко снижает производительность труда и КПД сотрудника в целом – и это только на этапе поиска подходящего организации лота, вероятность выигрыша которого, в среднем варьируется около 30%, исходя из данных организации.

Данная проблема и послужила повод к построению разрабатываемой системы, в целях облегчения труда рабочего персонала и минимизации издержек во временных ресурсах.

Так как структура сайта государственных закупок хорошо просматривается, появляется возможность работы с ресурсом с помощью

**Всероссийское СМИ**
**«Академия педагогических идей «НОВАЦИЯ»**
**Свидетельство о регистрации ЭЛ №ФС 77-62011 от 05.06.2015 г.**
**(выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций)**
**Сайт:** akademnova.ru
**e-mail:** akademnova@mail.ru

парсеров и других подобных методов и форм автоматизированного сбора информации.

На рынке, на сегодняшний день, появляются автоматизированные решения данной направленности, однако, ни одно решение не удовлетворяет полностью предъявленным требованиям.

Примером такого решения является система Гарант, предоставляемая предприятием Гарант Софт: система позволяет находить нужные тендеры, но не производит сохранения информации и подписка на такой сервис стоит около 20тыс рублей в месяц для внебюджетных организаций.

По причине глобальной минимизации расходов внутри предприятия, было принято решение разрабатывать приложение своими силами. Код программы и логические связи баз данных из всемирной сети Интернет не получилось, исходя из новизны данного направления.

Далее был проведен опрос работников маркетингового отдела и систематизированы их требования к разрабатываемому приложению:

13. Необходимо автоматизировать поиск лотов на ресурсе https://zakupki.gov.ru;
14. Программа должна производить поиск без вникания в код программы;
15. Должна быть организована автоматизированная передача информации в базу данных или таблицы XLS;
16. Информация должна сохраняться в пригодном для чтения/сравнения виде;
17. Доступ к информации должен быть предоставлен с нескольких рабочих станций одновременно;

Всероссийское СМИ
«Академия педагогических идей «НОВАЦИЯ»
Свидетельство о регистрации ЭЛ №ФС 77-62011 от 05.06.2015 г.
(выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций)
Сайт: akademnova.ru
e-mail: akademnova@mail.ru

18. Программный продукт должен иметь простой интуитивно-понятный интерфейс.

На этапе сбора информации были получены все данные, необходимые для разработки приложения, а так же получены требования к приложению от работающий с порталом закупок менеджеров коммерческого/маркетингового отделов. Данная информация будет использована в дальнейшей разработки, а мнения сотрудников будут учтены при реализации интерфейса разрабатываемого программного продукта.

Для качественного описания работы приложения было создано несколько схем функционирования приложения для полноты описания существующих процессов, возникающих при работе с приложением.

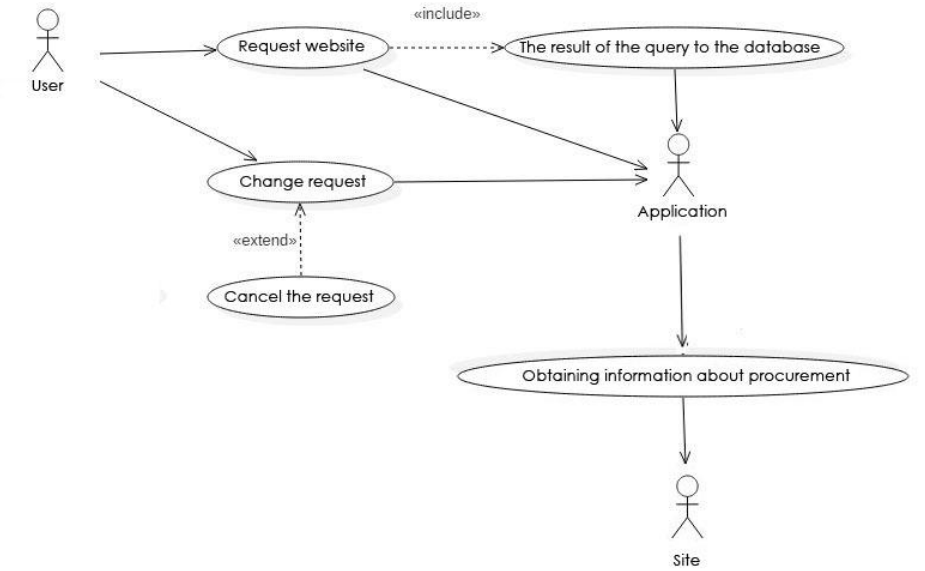


**Схема 1. Диаграмма прецедентов**

Диаграмма прецедентов отображает процесс обмена информацией между действующими объектами. Так как база данных условно входит в приложение и является невидимой для пользователя, данный объект включен в объект «Application».

Следующая схема отображает взаимодействие данных объектов и, условно, показывает последовательность работы, или, проще говоря, ее алгоритм. Таким образом, была составлена диаграмма деятельности:

**Всероссийское СМИ**
**«Академия педагогических идей «НОВАЦИЯ»**
**Свидетельство о регистрации ЭЛ №ФС 77-62011 от 05.06.2015 г.**
**(выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций)**
**Сайт:** akademnova.ru
**e-mail:** akademnova@mail.ru

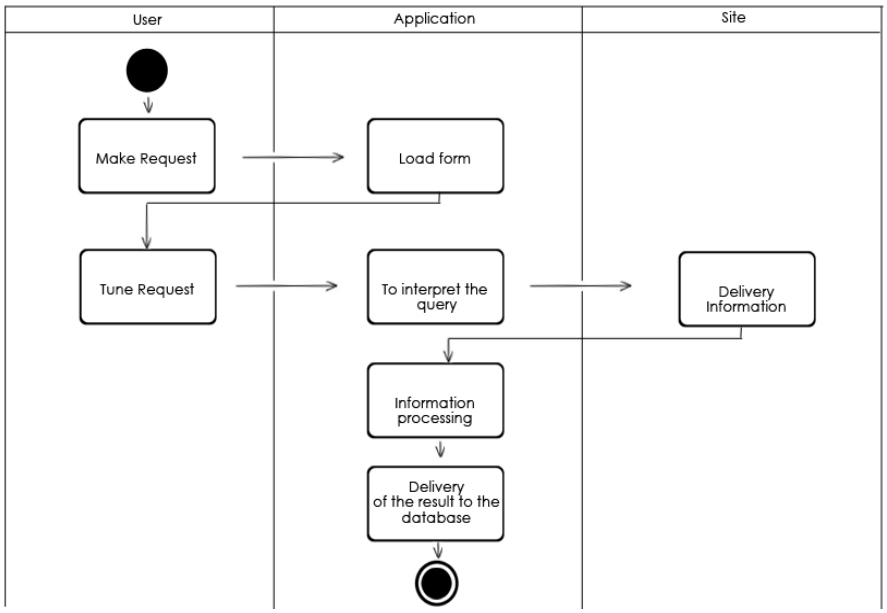


**Схема 2. Диаграмма деятельности**

Функционирование приложения предполагается согласно схеме:


Схема 2.1. Логическая схема функционирования приложения

**Всероссийское СМИ**
**«Академия педагогических идей «НОВАЦИЯ»**
**Свидетельство о регистрации ЭЛ №ФС 77-62011 от 05.06.2015 г.**
**(выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций)**
**Сайт:** akademnova.ru
**e-mail:** akademnova@mail.ru

Таким образом, данный набор схем полностью отображает логическую схему работы приложения, на уровне элементов и их взаимодействия, а так же отображает фукнциональную схему работы на уровне взаимодействия с пользователем.

Стоит отметить, что работа с приложением может быть абсолютно прозрачной для пользователя, и он может быть не посвящен, что приложение является распределенным, так как обработка, хранение результатов запроса и запрашиваемая информация может быть территориально распределена.

В данной статье рассмотрены далеко не все возможные проблемы и пути решения возникающих перед программистом задач при парсинге, однако, представлен весьма элегантный способ решения задач, связанных по разработке алгоритма работы с большими массивами информации на примере портала государственных закупок zakupki.gov.ru, рассматриваемого с точки зрения поставщика медицинских услуг, что позволит рационально использовать время и силы программиста, реализующего подобные задачи.

**Список использованной литературы:**

1. Доусон М. Программируем на Python. – СПб.: Питер, 2014. – 416 с.
2. Лутц М. Изучаем Python, 4-е издание. – Пер. с англ. – СПб.: Символ-Плюс, 2011. – 1280 с.
3. Лутц М. Программирование на Python, том I, 4-е издание. – Пер. с англ. – СПб.: Символ-Плюс, 2011. – 992 с.
4. Лутц М. Программирование на Python, том II, 4-е издание. – Пер. с англ. – СПб.: Символ-Плюс, 2011. – 992 с.
5. Прохоренок Н.А. Python 3 и PyQt. Разработка приложений. – СПб.: БХВ-Петербург, 2012. – 704 с.

**Всероссийское СМИ**
**«Академия педагогических идей «НОВАЦИЯ»**
**Свидетельство о регистрации ЭЛ №ФС 77-62011 от 05.06.2015 г.**
**(выдано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций)**

**Сайт:** akademnova.ru
**e-mail:** akademnova@mail.ru

6. Пилгрим Марк. Погружение в Python 3 (Dive into Python 3 на русском)
7. Прохоренок Н.А. Самое необходимое. — СПб.: БХВ-Петербург, 2011. — 416 с.
8. Хахаев И.А. Практикум по алгоритмизации и программированию на Python. – М.: Альт Линукс, 2010. — 126 с. (Библиотека ALT Linux).
9. Чаплыгин А.Н. Учимся программировать вместе с питоном.
10. Briggs J. R. — Python for Kids — 2012
11. Allen Downey – ThinkPython+Kart[Python_3.2]

***Дата поступления в редакцию: 10.06.2017 г.***
***Опубликовано: 10.06.2017 г.***