МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра дискретной математики и информационных технологий

# РЕАЛИЗАЦИЯ КРАУЛИНГА ДЛЯ АНАЛИЗА БИЗНЕС ДАННЫХ

## АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 5 курса 521 группы
направления 09.03.01 «Информатика и вычислительная техника»
факультета компьютерных наук и информационных технологий
Гераськина Владимира Вячеславовича

Научный руководитель
д.э.н., профессор _____ Л.В. Кальянов
подпись, дата

Зав. кафедрой
к. ф.-м.н., доцент _____ Л.Б. Тяпаев
подпись, дата

Саратов 2019

# ВВЕДЕНИЕ

Добывание информации из Всемирной паутины завоевало широкое признание бизнеса и получила большое количество вариантов использования в бизнес-решениях. В настоящее время в Интернете существует более миллиарда страниц информации по всем мыслимым темам. Вопрос в том, как вы можете найти то, что вы хотите? Можно написать компьютерные алгоритмы для поиска в Интернете, но большинство из них не практичны, потому что они должны жертвовать точностью ради покрытия. Тем не менее, несколько движков нашли интересные способы быстрого предоставления высококачественной информации. Ранжирование по значению страницы, поиск по темам и мета-поисковые системы - три из самых популярных, потому что они работают умнее, а не усерднее.

Поисковые системы ежедневно используются миллионами людей по всему миру. Это самый простой и удобный способ поиска контента в Интернете. Они предоставляют бесценную услугу пользователю Интернета, предоставляя огромную базу данных веб-сайтов, которые пользователь может быстро найти. Ключевые слова, которые пользователь вводит в поле поиска, сопоставляются с ресурсами в базе данных ключевых слов поисковой системы. Затем они сортируются по релевантности и представляются пользователю с наиболее релевантными сайтами, появляющимися первыми. Размер различных баз данных поисковых систем может сильно различаться, так как все разные типы систем собирают данные по-разному.

Каждый сайт имеет уникальное строение и индивидуальную верстку. Из-за этого мы не можем создать одну программу, универсальную, подходящую для всех сайтов. Чтобы поисковый робот загружал данные из целевого ресурса, нужно произвести его первичную настройку, которую зачастую выполняют квалифицированные специалисты. Извлечение веб-данных, без сомнения, трудоёмкий процесс. Если целевой сайт является динамическим и на нем используется комбинированная генерация веб-страниц, то сложность разработки поисковых агентов увеличивается в разы.

В данной работе рассматриваются технические нюансы разработки поисковых агентов и методы совершенствования средств, предназначенных для выгрузки информации с веб-сайтов.

Целью выпускной квалификационной работы является реализация различных методов краулинга с сбором информации для последующего анализа. В ходе работы были поставлены следующие задачи:

- изучение состава и принципов работы поисковых систем;
- изучение методов работы web-crawler;
- изучение задач поисковых систем;
- обзор технических нюансов разработки поисковых агентов;
- обзор методов совершенствования средств, предназначенных для выгрузки информации с веб-сайтов.
- Реализация алгоритма краулинга различными методами.

## 1 Поисковые системы

Самые первые поисковые системы появились в сети более пятнадцати лет назад. Тогда они выполняли лишь одну функцию – поиска ссылок к недавно созданным страницам.

Как только начал развиваться интернет информации в нём было мало, как и пользователей сети. Почти все из них были сотрудники либо различных организаций, либо научных университетов. Необходимость поиска информации в сети не был так необходим, как в наше время. В настоящее время поисковые системы трансформировались в многофункциональный инструмент - сервис. Они дают возможность искать и находить в сети Интернет интересующую пользователей информацию, благодаря чему пользуются огромным успехом [4].

Первая попытка создать организационный доступ к информационным ресурсам это сбор тематических каталогов сайтов. Первым, открывшимся в апреле 1994 г, стал Yahoo. Трудно назвать его поисковой системой, в современном понимании, т.к. поиск распространялся только на те сайты, что были зарегистрированными в каталоге Yahoo. Каталоги ссылок до нашего времени использовались довольно обширно, но в наше время практически утратили свою актуальность. Объяснить это просто – сейчас даже каталоги, которые содержат информацию об огромном количестве ресурсов, представляют информацию лишь о довольно малой части сети. Для сравнения - самый полный каталог сети интернет - DMOZ содержит информацию приблизительно о 12.000.000 ресурсов, в то время как база данных самой полной поисковой системы Google состоит более чем из 28.000.000.000 ресурсов.

В 1994 года появился проект под названием WebCrawler. Именно его можно назвать непосредственно первой поисковой. Через год появились поисковые системы AltaVista и Lycos. В 1997 году в Стэнфордском университете, в рамках исследовательского проекта, была создана Google - самая популярная поисковая система на данный момент в мире. Также в этом году появилась поисковая система - Yandex, лидер на рынке русской части

Интернета. На данный момент основными поисковыми системами являются три международных – Google, Yahoo и MSN Search. Остальные используют целиком или частично базы и (или) алгоритмы выше приведенных систем. В Рунете основной поисковой системой является Яндекс, дальше идут Rambler, Google.ru, Mail.ru и Aport.

Поисковая система - это сумма следующих компонентов:

- Web-server (веб-сервер) – является сервером поисковой машины, именно он осуществляет взаимодействие между пользователем и компонентами системы.
- Spider (паук) - программа, написанная по принципу браузера, но работающая напрямую с HTML кодом, предназначенная для парсинга веб-страниц.
- Crawler («путешествующий» паук) – программа, которая в автоматическом режиме ходит по всем внешним ссылкам страницы. Выполняет следующие задачи - поиск неизвестных или измененных документов в сети и расстановка приоритетов, куда дальше должен следовать Паук.
- Indexer (индексатор) – программа анализатор скаченных пауками веб-страниц. Она "разбирает" на части скачанную страницу и анализирует ее элементы, такие как текст, служебные html-теги, заголовки, особенности стилистики и структурные формы.
- Database (база данных) – хранилище для скачанных и уже обработанных страниц - общая база данных поисковой машины.
- Search engine results engine (система выдачи результатов) – извлекает результаты поиска из базы данных поисковой системы. Именно она решает, какие страницы более соответствуют запросу пользователя и отсортировывает их в нужном порядке. Модуль работает согласно заданным поисковой системой алгоритмам ранжирования [2].

# 2 Проблемы и задачи web crawling

Web Scraping – это популярный метод получения контента практически даром. У нас такой метод называется «парсинг контента» или «парсинг сайтов». Метод состоит в том, что специально обученный алгоритм заходит на главную страницу сайта и начинает переходить по всем внутренним ссылкам, тщательно собирая внутренности указанных вами div-ов. В качестве результата работы – готовый CSV файл, в котором вся нужная информация лежит в строгом порядке [1].

Постоянно происходит рост сайтов, которые пользуются средствами, препятствующими web crawling. Зачастую связано это с жесткой конкуренцией на рынке контента. Однако, в Интернете существует огромная база общедоступных и легальных источников информации. Но процесс извлечения данных из таких источников может быть не простым. Сложности могут заключаться во многом. Например, целевой сайт может в любое время изменить верстку веб-страниц. Или ресурс может содержать неисправный JavaScript, который мешает нормальной работе поисковика. Сервер сайта может упасть или может быть недоступным из-за технического обслуживания [11].

Многие потенциальные проблемы могут возникать при длительных сессиях работы поисковых агентов. Поэтому, качественный web-scraper использует множество передовых механизмов обработки ошибок и поддержки стабильности.

Обобщая вышесказанное, можно выделить следующие задачи по извлечению данных из веб-ресурсов, которые являются достаточно трудоемкими для выполнения [10]:

– Извлечение информации из HTML-кода;

– Извлечение информации с динамических веб-страниц;

– Извлечение информации с веб-сайтов, которые используют сдерживающие факторы;

– Извлечение больших объемов информации.

# 3 Реализация процессов web-crawling

В ходе работы были разработаны следующие технологии краулинга:

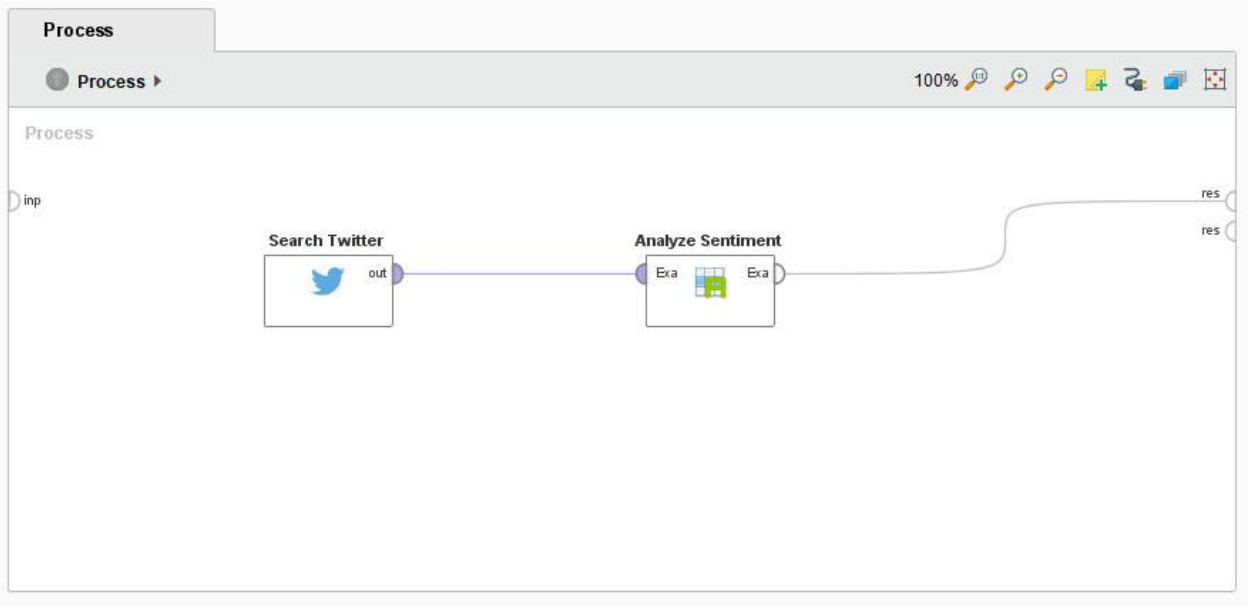


Рисунок 1– Процесс краулинга «твитов» из Twitter

Процесс, изображенный на рисунке 1, позволяет краулить «твиты» из социальной сети Twitter по ключевым словам.

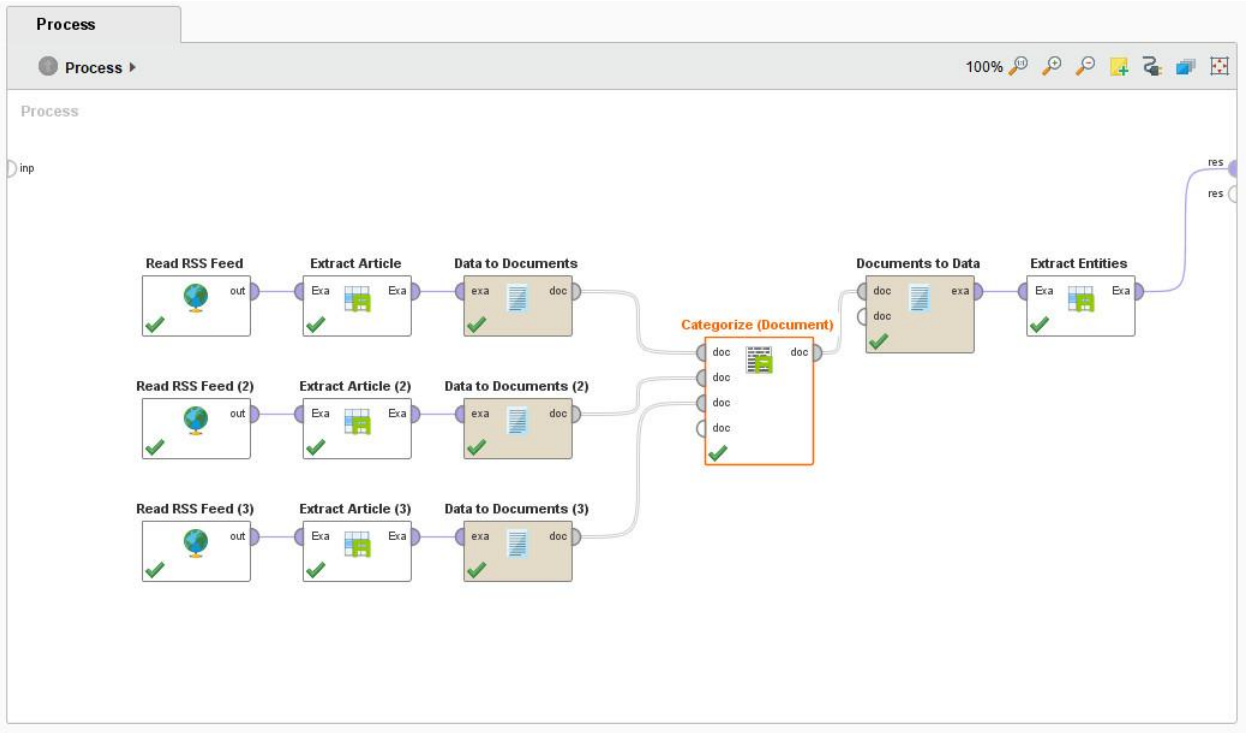


Рисунок 2 – Процесс анализа данных новостных RSS каналов.

Процесс, изображенный на рисунке 2, позволяет краулить новости из различных RSS каналов и после анализировать и структурировать их.

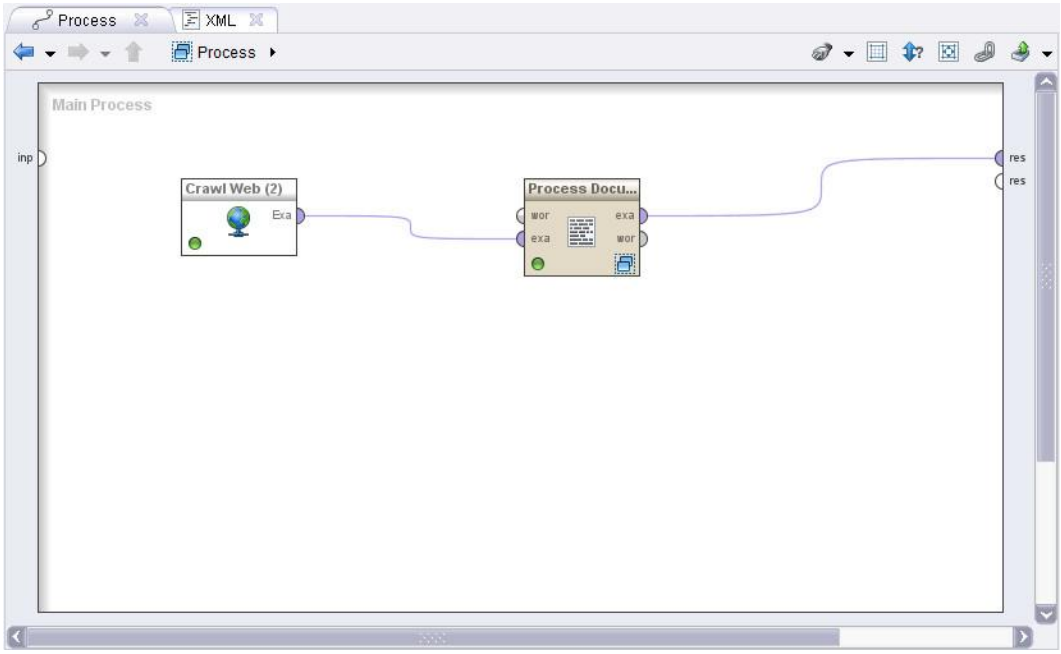


Рисунок 3 – Процесс краулинга

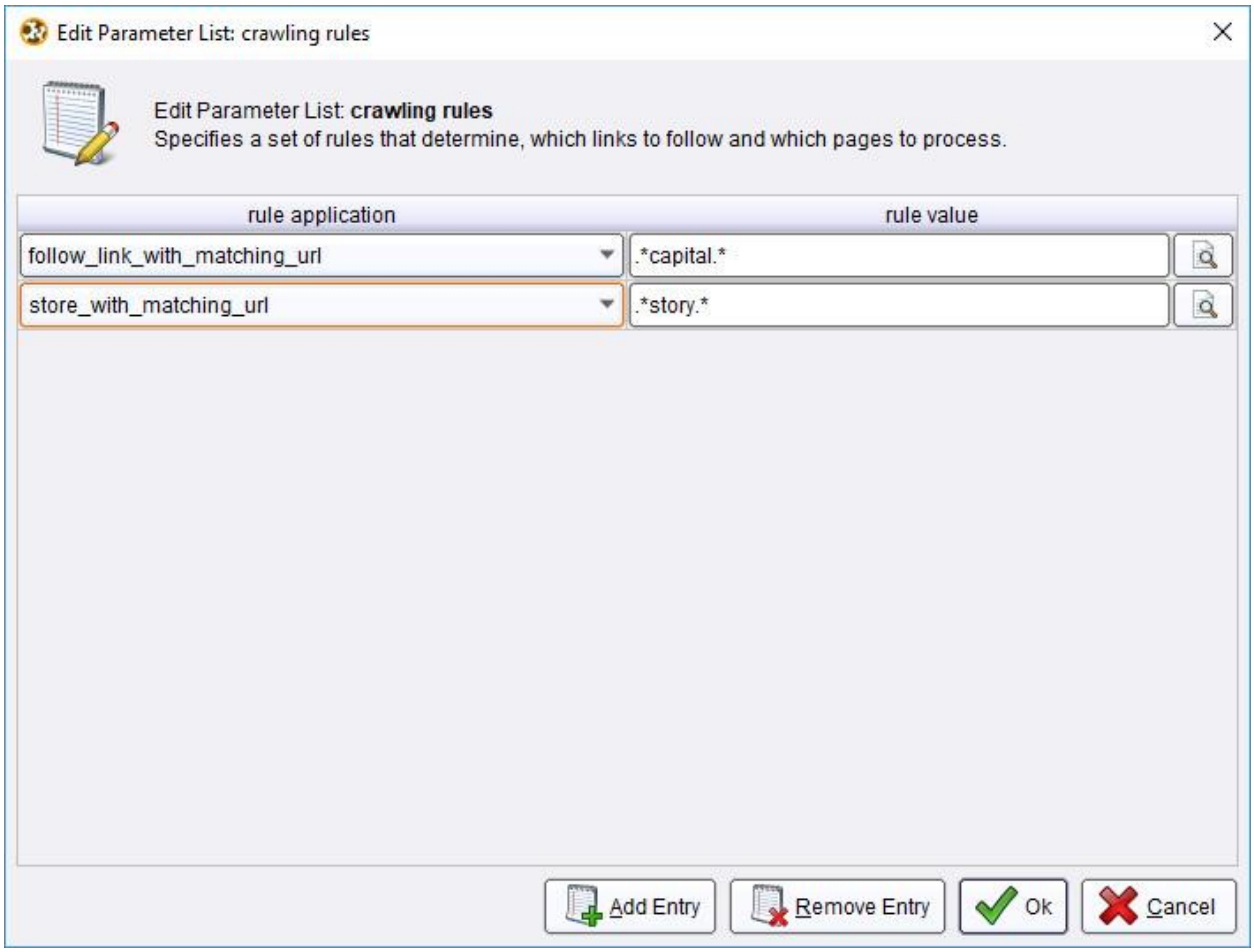


Рисунок 4 – Параметры краулера

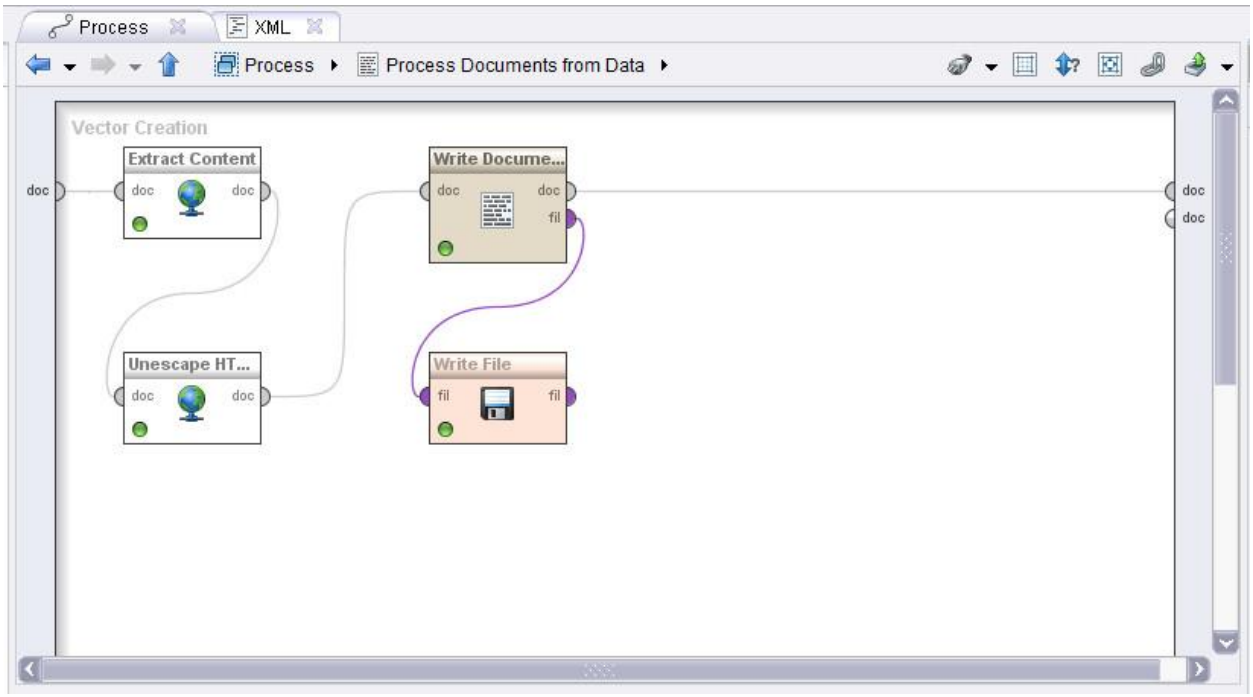


Рисунок 5 – Построение подпроцесса.

Процесс, изображенный на рисунках 3-5, собирает информацию с сайта BBC, раздел Capital и сохраняет её в txt файлы на компьютер.

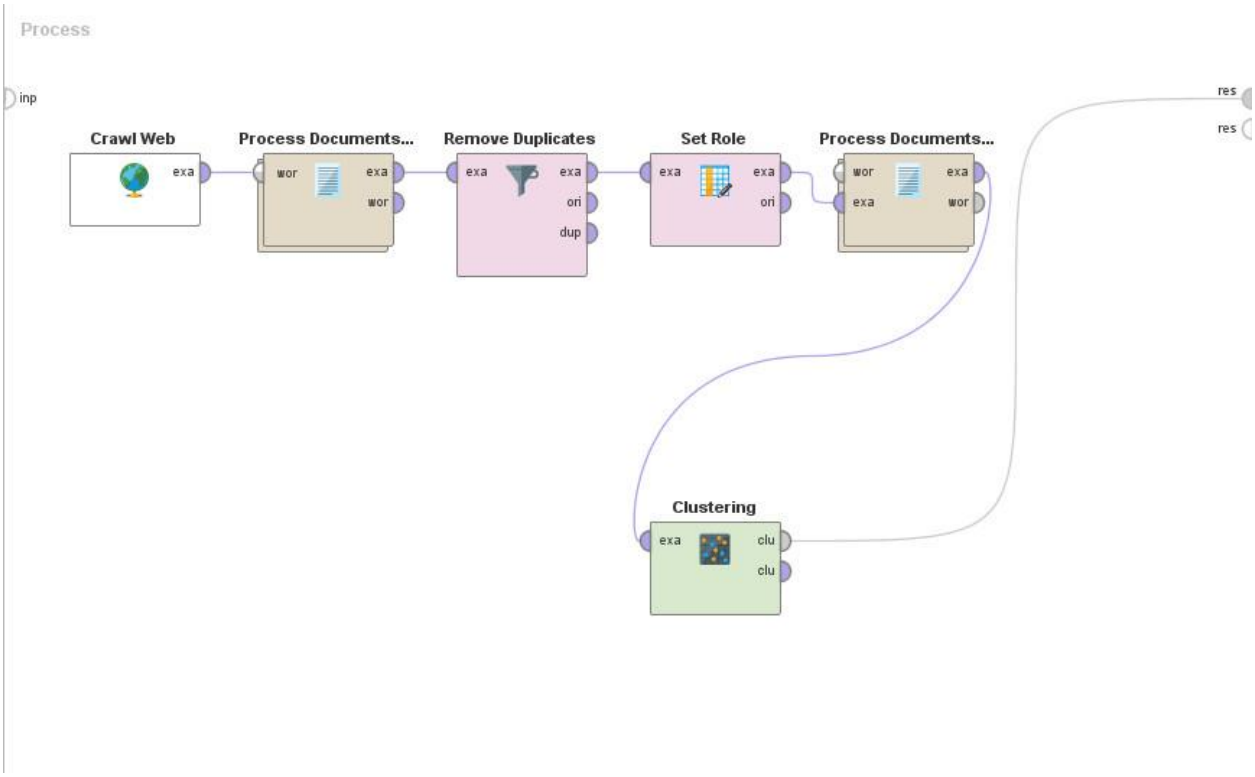


Рисунок 6 – Общий процесс

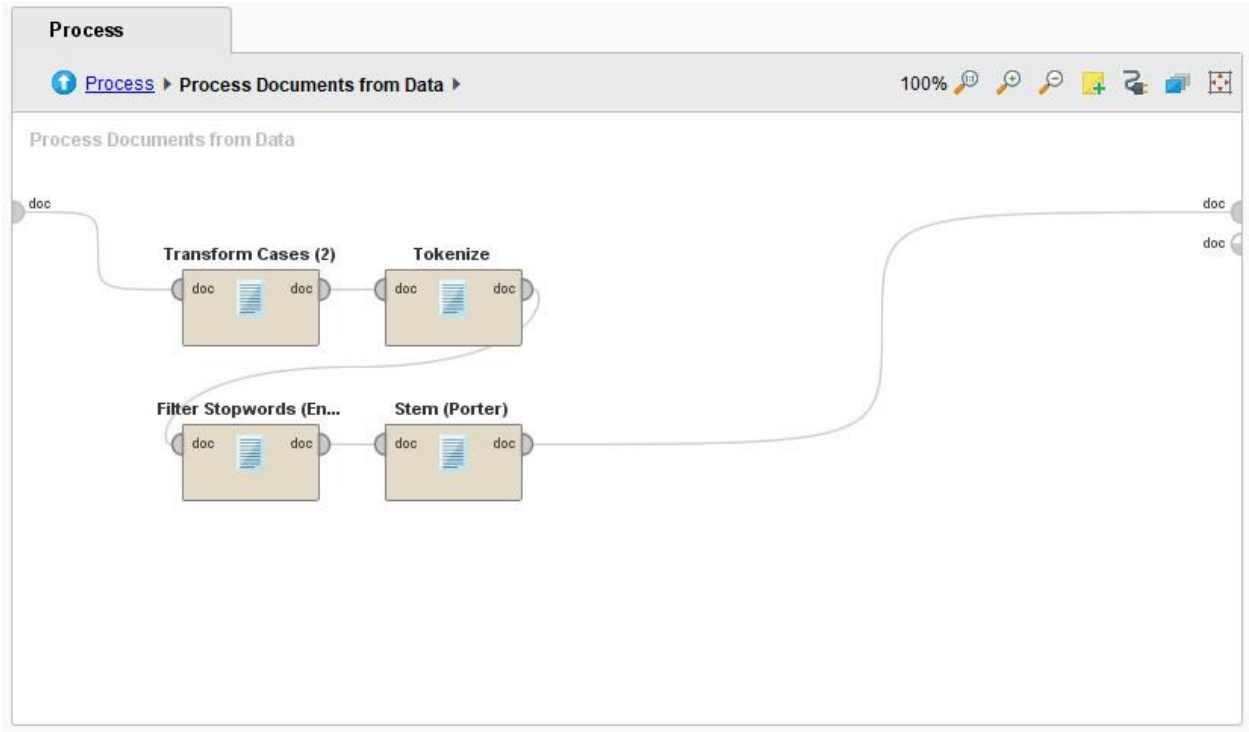


Рисунок 7 – Процесс кластеризации текста.

Процесс, изображенный на рисунках 6-7, кластеризует информацию, содержащуюся на указанном интернет ресурсе. Кластеризация осуществляется на основе ключевых слов с использованием метода k-средних.


Рисунок 8 – процесс работы программы

```
1  Проект,Категории,Цена,Заявки
2  Переделать дополнение от xPay,"Веб-программирование/Скрипты, HTML-верстка",?,3
3  Разработка приложения,ПО для мобильных устройств,?,0
4  Исправить поиск автозапчастей по базе данных Intercars в инт магазине,"Базы данных, Интернет-магазины",?,0
5  Написать курсовую работу,Дипломы/Курсовые/Рефераты,17 USD,0
6  Моделирование круговых волн на воде,Разработка игр,?,0
7  Перевод жилой недвижимости в коммерческую,Копирайтинг/Рерайтинг,?,0
8  Доработки,Веб-программирование/Скрипты,?,1
9  Подключение сервиса toddlahman к сайту wordpress,Веб-программирование/Скрипты,?,0
10 Курсовая работа по экономике.,Дипломы/Курсовые/Рефераты,?,0
11 Дизайн и верстка сайта на DLE,"Дизайн сайтов, HTML-верстка",?,0
12 Портал объявлений Прибалтики,Системы управления (CMS),?,0
13 Лонгрид на wordpress,Веб-программирование/Скрипты,?,0
14 Дизайн листовки-лендинга,Дизайн продукции/упаковки,?,0
15 Оптимизировать РК яндекс,Контекстная реклама,?,0
16 Поднять конверсию лендинга,Контекстная реклама,?,0
17 Перевод с русского на английский описаний достопримечательностей мира,Переводы,?,2
18 Постер для батутного центра,"Верстка/Допечатная подготовка, Дизайн продукции/упаковки",?,2
19 Доредактировать статью!,Редактирование/Корректировка,5 USD,1
20 Визитка + бланк,Фирменный стиль,?,7
21 "Курсовая работа по психологии (укр,). Срочно",Дипломы/Курсовые/Рефераты,15 USD,1
22 Дизайн интернет-магазина детских игрушек,Дизайн сайтов,?,18
23 Настроить гугл аналитик,Веб-программирование/Скрипты,?,0
```

Рисунок 9 – результат работы программы

На рисунке 8 представлен процесс работы программы, написанной на на языке программирования Python, которая собирает информацию о заказах на сайте биржа фрилансеров. На рисунке 9 представлен результат работы программы.

```
7  PROJECT_NAME = 'intel'
8  HOMEPAGE = 'https://www.intel.ru/content/www/ru/ru/homepage.html/'
9  DOMAIN_NAME = get_domain_name(HOMEPAGE)
10 QUEUE_FILE = PROJECT_NAME + '/queue.txt'
11 CRAWLED_FILE = PROJECT_NAME + '/crawled.txt'
12 NUMBER_OF_THREADS = 512
13 queue = Queue()
14 Spider(PROJECT_NAME, HOMEPAGE, DOMAIN_NAME)
```

Рисунок 10 – фрагмент кода программы

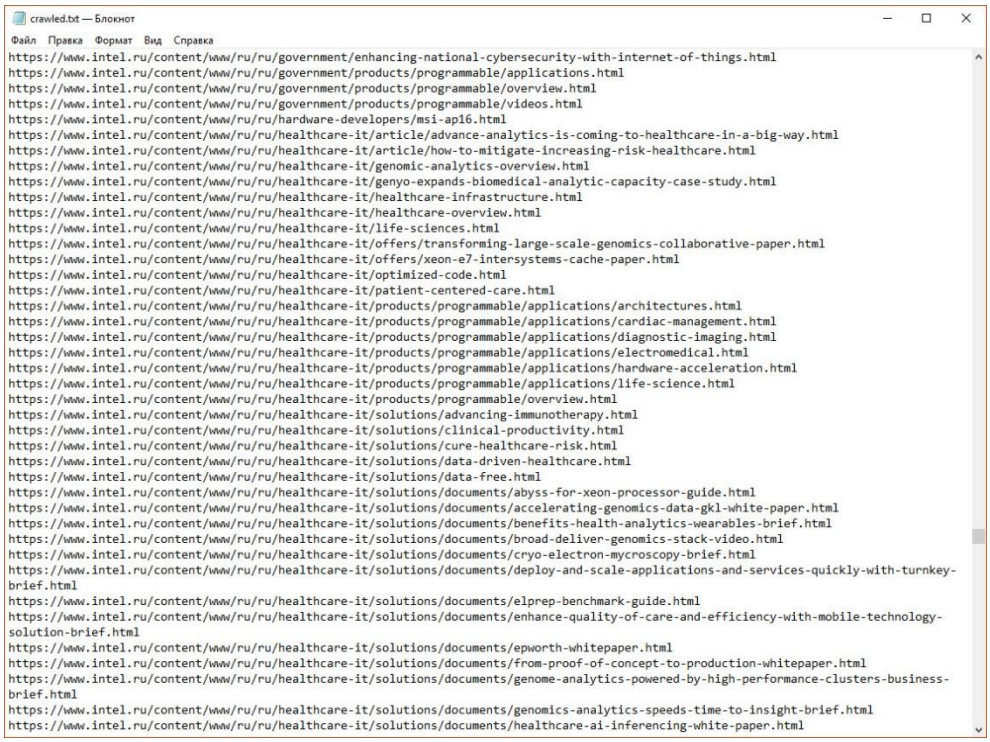


Рисунок 11 – результат программы

Был создан web-crawler по работе аналогичный паукам Google и Яндекс. Программа была написана на языке Python и состоит из 6 файлов: domain.py, general.py, link_finder.py, main.py, spider.py.

Программа на вход получает стартовую страницу, с которой и начинается краулинг. Затем создает 2 файла: crawled и queue. В файле crawled хранятся ссылки страниц уже прошедших сканирование. В файле queue хранится так называемая очередь, адреса на страницы, которые были найдены на уже прошедших сканирование страницах, и которые ждут своей очереди на сканирование.

В программе есть несколько удобных функций, таких как:

- сравнение с доменом стартовой страницы - это нужно для того, чтобы паук не уходил на другие сайты. Можно отключить, если требуется просканировать весь интернет.
- проверка на дубликаты – необходима для того, чтобы паук не тратил время на одни и те же страницы.
- Использование многопоточности – на сканирование одного крупного сайта с количеством страниц более 1000 уйдет много времени. Поэтому внутри программу было реализована многопоточность. Внутри программы сам паук был размножен до 512 (сопоставимо с мощностью

компьютера, на котором был произведен запуск). Но это значение можно менять в зависимости от процессорной мощности компьютера для изменения скорости работы программы.

**Пример работы программы:**

Подаем на вход стартовую страницу-
https://www.intel.ru/content/www/ru/ru/homepage.html/, можно увидеть на рисунке 10.

Так же на рисунке 10 можно увидеть, что значение NUMBER_OF_THREADS равно 512, значит одновременно будут работать 512 пауков.

После запуска, программа проработала около 6 часов и просканировала более 80000 страниц. Все ссылки сохранились в crawled. Как можно увидеть на рисунке 11 просканировались даже все файлы с расширением .pdf .doc .zip .jpg .png. Что даёт огромную базу данных для последующего использования

# ЗАКЛЮЧЕНИЕ

Web-crawling сайтов — это трудоемкая и кропотливая задача, требующая глубоких знаний в области разработки веб-приложений. Также следует отметить то, что наравне с опытом разработчиков web-crawling требует большой стек технологий и инфраструктуры, которые могут решать сложные вопросы, связанные с извлечением веб-данных.

Реализация краулера на языке программирования Python более интересна и позволяет более тонко и полно настраивать краулер конкретно под каждую задачу. А множество библиотек позволяют разнообразить эти задачи.

В ходе работы были реализованы поставленные задачи, а именно: был изучен состав и принципы работы поисковых систем, были изучены методы работы web-crawler, изучены задачи поисковых систем, был произведен обзор технических нюансов разработки поисковых агентов, были рассмотрены методы совершенствования средств, предназначенных для выгрузки информации с веб-сайтов. А также был реализован краулинг различными методами, и получены данные для дальнейшего анализа. Разработанные примеры подходят для работы с различными сайтами и для различных задач.

## СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ


1 [Электронный ресурс]: URL: https://top-technologies.ru/ru/article/view?id=36585 (дата обращения 18.05.2018)

2 RapidMiner Data Mining Use Cases and Business Analytics Applications 2014 Markus Hofmann, Ralf Klinkenberg ISBN 978-1-4822-0550-3.

3 Введение в rapidminer. [Электронный ресурс] сайт. URL: https://habr.com/post/269427/ Дата обращения 16.05.2018.

4 А. А. Барсегян, М. С. Куприянов, И. И. Холод, М. Д. Тесс, С. И. Елизаров. Анализ данных и процессов: учеб. пособие, 3-е изд., перераб., 2009г. ISBN 978-5-9775-0368-6.

5 Официальный сайт RapidMiner. [Электронный ресурс]: URL: https://rapidminer.com (дата обращения 01.03.2018)

6 Принципы решения задач по извлечению данных из веб-ресурсов с помощью web scrapers Аладин Д.В. [Электронный ресурс]: URL: http://www.tpinauka.ru/2017/11/Aladin.pdf (дата обращения 01.03.2018)

7 Michael W. Berry, Survey of Text Mining Clustering, Classification, and Retrieval Scanned by Velocity - University of Tennessee 2004

8 Ronen Feldman, James Sanger, The text mining handbook - Cambridge University 2007

9 Официальный сайт Rosette Text Analytics. [Электронный ресурс]: URL: https://www.rosette.com (дата обращения 10.04.2018)

10 Куприянова, Г.И., "Информационные ресурсы Internet" - М., 2012

11 Метод k-средних. [Электронный ресурс]: URL: https://ru.wikipedia.org/wiki/метод_k-средних (дата обращения 10.04.2018)

12 Microsoft User Group Community. [Электронный ресурс]: URL: http://msugvnua000.web710.discountasp.net/Posts/Details/4213 (дата обращения 02.05.2018)