

**ИНСТИТУТ ИНФОКОММУНИКАЦИОННЫХ**

**СИСТЕМ И ТЕХНОЛОГИЙ**

**Контрольная работа**

**По дисциплине:** «Технологии и системы коллективной разработки программ»

**На тему:** «Сбор и анализ данных новостного источника RIA»

Направление подготовки 01.03.02. Прикладная математика и информатика

**Выполнили:**

студенты группы ПМИ-18

Мамедова Екатерина

Баранников Дмитрий

Лямин Михаил

**Научный руководитель:**

Преподаватель Гусятинер Л.Б.

Королев 2021

**Оглавление**

[**Введение: 3**](#_Toc90937873)

[**1. Парсинг. Основные понятия 4**](#_Toc90937874)

[**1.1. Что такое парcинг? 4**](#_Toc90937875)

[**1.2. Какие задачи решает парсер? 4**](#_Toc90937876)

[**2. Выполнение проекта 5**](#_Toc90937877)

[**2.1. Роли участников 5**](#_Toc90937878)

[**2.2. Временной план реализации проекта 5**](#_Toc90937879)

[**2.3. Техническое задание 6**](#_Toc90937880)

[**2.4. Состав программного приложения 8**](#_Toc90937881)

[**2.5. Результаты 9**](#_Toc90937882)

[**2.6. Вывод 11**](#_Toc90937883)

[**2.7. Пути доработки и возможность использования. 11**](#_Toc90937884)

[**Заключение 11**](#_Toc90937885)

## Введение:

***Цель работы:*** Создание проекта путём коллективной работы, где каждый участник делает свою часть. Наш проект заключается в том, чтобы написать программу на Python, которая выводит новости разных категорий и их количество просмотров, собрать данные в excel и провести анализ полученных данных для улучшение экономических показателей компании заказчика.

***Задачи:***

1. Разработать программное приложение для сбора данных новостного источника RIA
2. С помощью программного приложения записать и сохранить данные в файл Excel
3. На основании полученных результатов, сделать выводы

## Парсинг. Основные понятия

### Что такое парcинг?

Для лучшего понимания темы опишем основные понятия парсинга.

Дословный перевод слова parsing — делать грамматический разбор или структурировать.

**Парсинг** — это **синтаксический анализ информации**.

Под парсингом подразумевается **автоматический сбор информации с последующим сохранением, обработкой и анализом данных**.

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

Источником для парсинга могут быть базы данных, веб-сайты, файлы и другие источники информации.

### Какие задачи решает парсер?

* Парсер может прекрасно справится с задачей сбора информации о товарах и последующем наполнении интернет-магазин.
* С помощью парсера можно проводить аналитику цен, настроив парсер для магазина конкурентов.
* Если работа предусматривает регулярный поиск новых партнёров или клиентов, самостоятельно искать контактные данные в сети, довольно длительный процесс, с автоматизацией сбора контактов тоже поможет парсер, контакты можно собирать с сайтов объявлений, соц. сетей и с любых других сайтов, на которых контактная информация находится в открытом доступе.
* Если стоит задача регулярного мониторинга объявлений, с этим так же может помочь парсер.
* Если нужно создать свой сайт или блог, но не хочется заниматься его наполнение вручную, автоматизировать процесс наполнение поможет парсер.

Парсер поможет с выполнением рутинных задач и значительно повысит эффективность работы.

## Выполнение проекта

### Роли участников

Для того чтобы начать работу в начале необходимо определить роли участников.

Мамедова Екатерина Евгеньевна- заказчик. Задачи в проекте: составление ТЗ, создание и оформление файла News. xlsx для записи и сохранении новостей, даты, категорий и количества просмотров, анализ проделанной работы.

Лямин Михаил Евгеньевич- главный программист. Задачи в проекте: распределение ролей участников проекта, создание временного плана реализации проекта, выполнение ТЗ: разработка парсера новостей.

Баранников Дмитрий Владимирович- программист. Задачи в проекте: выполнение ТЗ: доработка парсера с записью и сохранением данных в файл News. Xlsx

### Временной план реализации проекта

Для реализации проекта был разработан временной план.

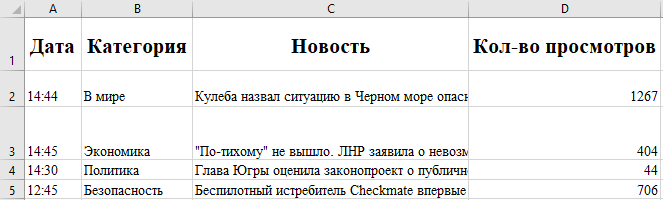
|  |  |  |
| --- | --- | --- |
| Действие | Сроки | Результат |
| Выбор темы проекта | 18 октября | Тема проекта выбрана |
| Распределение обязанностей | 19-20 октября | Распределены обязанности |
| Создание временного плана проекта | 1-7 ноября | Временный план проекта составлен |
| Составление ТЗ и файла News. xlsx | 9-10 ноября | Составлено ТЗ, создан и оформлен файл в excel |
| Начало разработки парсера, ознакомление с элементами сайта | 12 ноября | После ознакомление с элементами сайта, пришли к выводу, что данный сайт не подходит для выполнения нашего проекта |
|  |  | Изменение ТЗ, в связи со сложностью выполнения работы, сайт пришлось изменить (14 ноября) |
| Разработка парсера | 14-21 ноября | Парсер разработан |
| Доработка парсера, запись и сохранение данных в файл News. xlsx | 22-29 ноября | Парсер доработан, теперь он записывает и сохраняет данные в файл News. xlsx |
| Выводы, анализ данных | 30 ноября- 5 декабря | Благодаря данным, которые выводит парсер, сделаны выводы |
| Анализ проделанной работы | 6 декабря-8 декабря | Сделан анализ проделанной работы. |

### Техническое задание

Наша задача будет заключаться в том, чтобы написать парсер, который собирает актуальные новости за день с сайта <https://ria.ru/> по категориям.

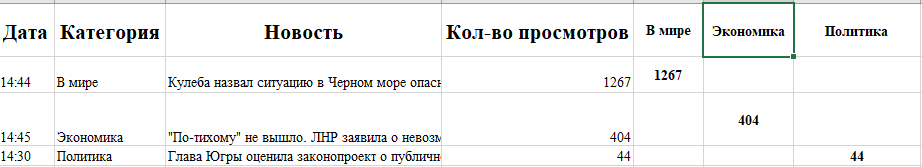
Новости в мире, экономика, политика, безопасность

И сохраняет данные в файл excel (News.xlsx) (Дата, Категория, Новость, Кол-во просмотров). Шрифт Times New Roman, размер 12



1. Пример заполнения новостей, которое должно быть сделано с помощью парсинга

В данном файле автоматически будут заполнятся столбцы категорий В мире, Экономика, Политика, Безопасность. Например, если у новости категория В мире, то в столбец В мире заполняется количество просмотров, если другая, то в этом столбце пусто.



1. Пример автоматического заполнения столбцов по категориям

Также в файле будет автоматически заполняться точечная диаграмма “Кол-во просмотров каждой новости разных категорий.”

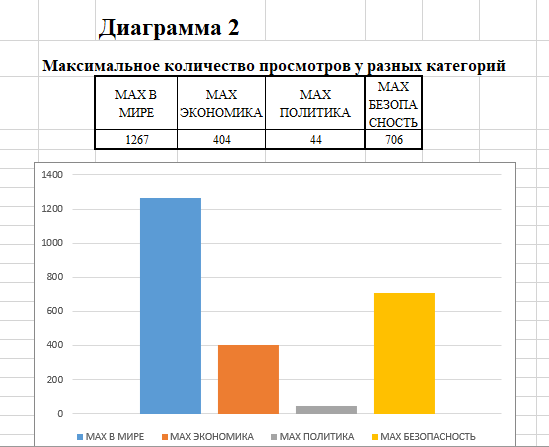
На этой диаграмме каждая точка это новость, разные цвета, это разные категории новостей. Благодаря данной диаграмме мы можем наглядно увидеть каких новостей в категории больше и у каких категорий большее количество просмотров.



1. Пример автоматического заполнения точечной диаграммы.

Также автоматически заполнится максимальное количество просмотров у каждой категории и диаграмма.

Это поможет нам определить в какой категории самая популярная новость.



1. Пример автоматического заполнения диаграммы Максимального количества просмотров у разных категорий.

После того, как данные будут сохранены в файле, мы узнаем, какие новости более популярны, чем интересуются большинство людей.

Разделим новости на более популярные и менее популярные и на основе полученных данных, сделаем выводы.

Благодаря этому мы можем провести анализ, узнать какие категории популярнее и определить о каких темах стоит писать больше новостей, а какие менее интересны и можно ли их сделать популярнее, на основе того, чем интересуются люди.

### Состав программного приложения

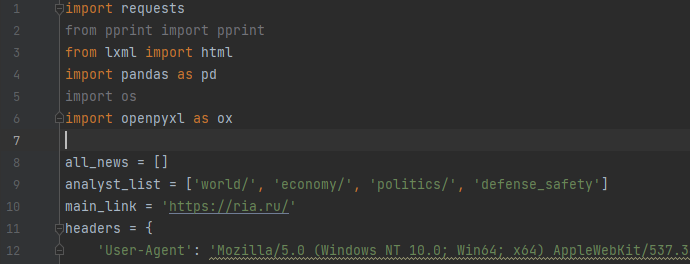
Программное приложение будет написано в интегрированной среде разработки PyCharm на языке программирования Python.

Опишем состав программного приложения.

Для начала импортируем все необходимые библиотеки.

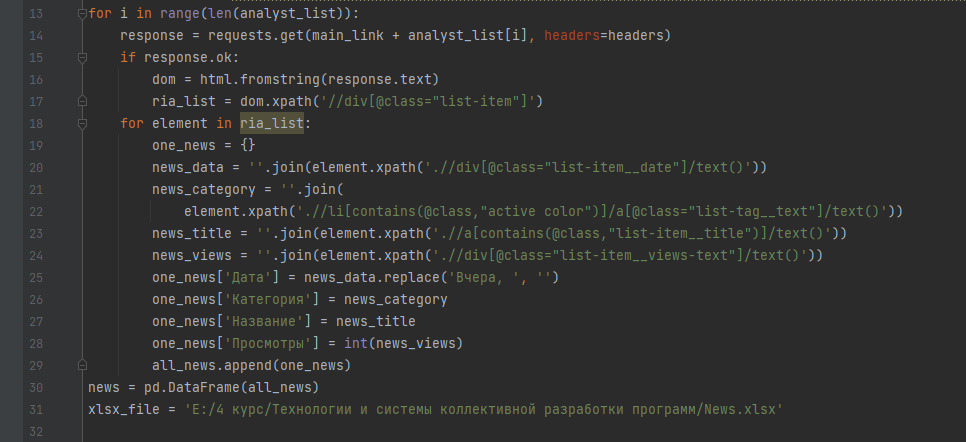
Далее создаём переменную для записи списка новостей, переменную основной ссылки и список категорий.

headers нужна для того, чтобы убрать конфликты некоторых сайтов на анонимные запросы

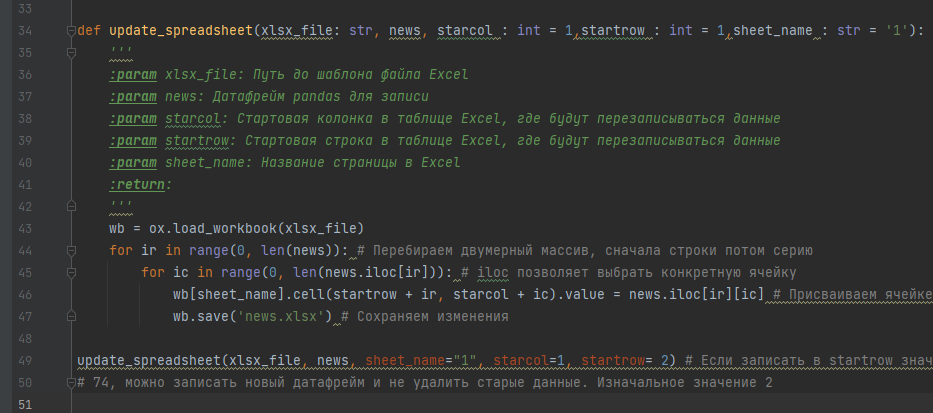


Далее создадим цикл, который проходит по списку категорий. Если сайт работает, то забираем текст html структуры странички и записываем его в html.fromstring, чтобы программа сформировала структуру, по которой сможет найти данные. Далее находим новости, используя путь и для каждой новости парсим словарик “Дата” “Категория” “Название” “Просмотры”

Все данные записываем в словарик, словарик в список и повторяем так по всем категориям. Полученные данные записываем в DataFrame для последующего использования



Далее создаём функцию, в которой используется цикл для перебора двумерного массива. Сначала по индексу, потом по серии. В конце мы сохраняем изменения



### Результаты

Запустим программу и после завершения записи данных, откроем файл News.xlsx и проверим сохранились ли у нас данные.



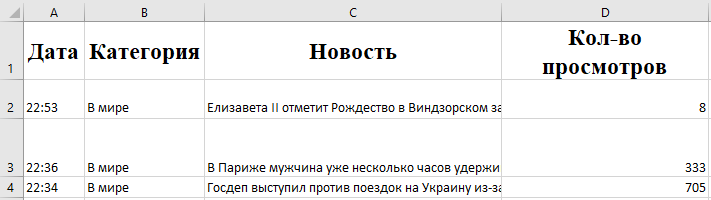
Данные сохранились.

Теперь откроем файл News.xlsx и сайт РИА Новости и сравним правильность и актуальность данных.

Например, первые 3 новости за сегодня категории “В Мире”



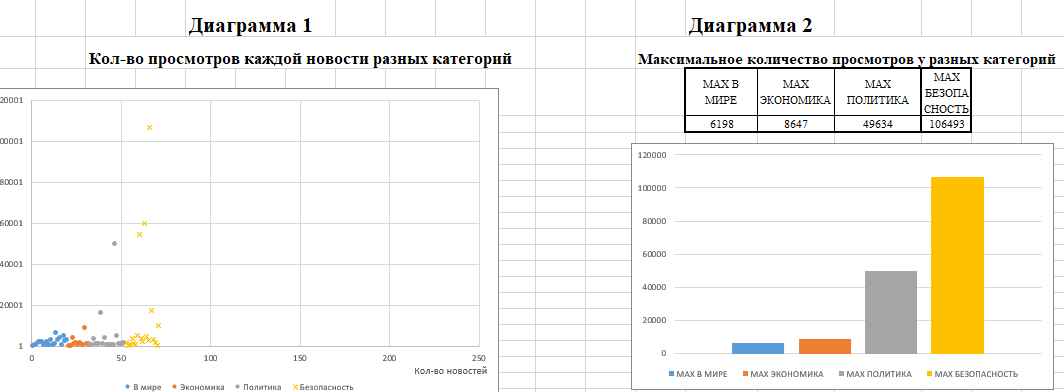
В файле News.xlsx первые 3 новости категории “В Мире”



Следовательно, парсер записал и сохранил новости правильно.

### Вывод

По файлу News.xlsx теперь можно анализировать популярность разных категорий новостей. К примеру, на данный момент, если рассмотреть диаграммы, то можно сделать вывод, что сейчас большинство интересует категория “Безопасности”.



Далее в этом файле мы можем вывести самую популярную новость из каждой категории, вычесть среднее количество просмотров каждой категории, добавить новые диаграммы и графики и многое другое. Благодаря этому мы можем сделать анализ новостного источника, понять какими новостями люди заинтересуются больше.

### Пути доработки и возможность использования.

Пути доработки:

Возможное увеличение кол-ва данных, сбор дополнительных признаков, добавление новых диаграмм и графиков для более конструктивного анализа популярности новостей

Возможности использования:

Система ежедневного анализа трендов

Система долговременного анализа интересов пользователей

## Заключение

Путём коллективной работы, мы создали проект. Программное приложение, позволяющее записывать и сохранять новости в файле excel. Благодаря этому можно анализировать популярность различных категорий новостей.