GeekBrains

Дипломный проект на тему:

«Применение наиболее значимых компонентов освоенного в процессе обучения технологического стека для разработки поисково-аналитического комплекса»

Команда проекта

Пиев Андрей

Аналитик больших данных
Группа 5321
Зона ответственности: идея и управление проектом, первичный парсер на языке Python с загрузкой данных в SQL формат

Фирюза Лаптева

Аналитик больших данных
Группа 5321
Зона ответственности: Реализация на Python многопоточной обработки данных для парсинга сайтов с последующим глубоким анализом в Jupyter Notebook

Анна Афанасьева

BI аналитик
Группа 5295
Зона ответственности: Анализ полученных данных средствами SQL, визуализация в Power BI

Юлия Евстифеева

BI аналитик
Группа 5295
Зона ответственности: Визуализация в Power BI, оценка и представление коммерческой ценности проекта

Максим Майбродский

Аналитик больших данных
Группа 5321
Зона ответственности: Разработка аналитического парсер-бота с помощью Python

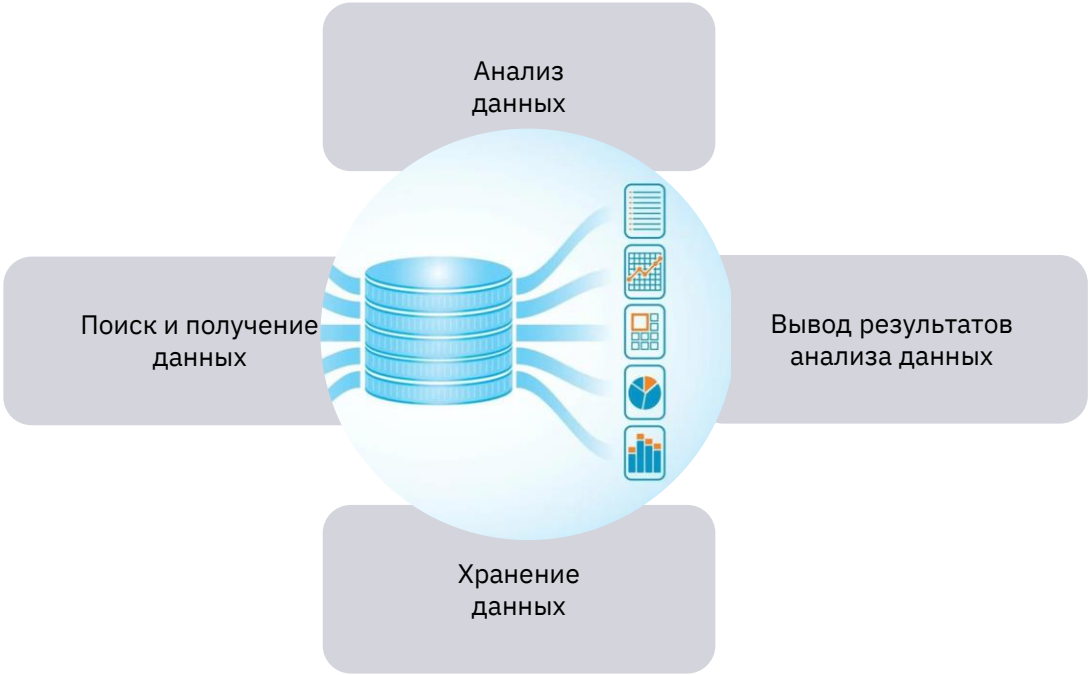
Владимир Богинский

Аналитик больших данных
Группа 5321
Зона ответственности: Создание сайта проекта на HTML/CSS, обработка данных

Поставленная задача

Команда проекта поставила перед собой задачу практической обкатки освоенного в процессе обучения технологического стека для создании механизмов поиска и анализа полученной информации.

Кратко проект можно описать как получение программой из первичного источника данных об адресе сайта в сети Интернет, скачивании html содержимого сайта в файл, дальнейшая оптимизация и анализ полученных данных.



Основные задействованные технологии

- Python (requests, sqlite3, pandas, numpy, pymorphy2, gensim, pyldavis, matplotlib, re, nltk, пр.)
- SQL
- Power BI
- HTML / CSS
- Docker
- PySpark
- CSV, JSON
- Git

Достижения проекта

- Краулером пройдено более 100 000 сайтов из первичной базы в 5 млн. адресов
- Скачано и проанализированно более 11 000 страниц
- Опробовано совместное использование нескольких технологий и форматов файлов
- Получен практический опыт многопоточной обработки данных
- Осознано коммерческое применение продукта
- Опробованы технологии ML
- Создан аналитический чат-бот
- Создан сайт проекта
- Создан дашборд

Практическое применение результатов проекта

Отдельные компоненты дипломной работы могут, при дальнейшей доработке, иметь как самостоятельное применение, так и стать частью исследовательских продуктов в следующих сферах:

- Поисковые системы
- Социально-политические исследования
- Мониторинг СМИ
- Маркетинговые исследования
- Мониторинг конкурентов
- Исследование социальных сетей

Технологические ограничения проекта

Отсутствие закупленных коммерческих серверных мощностей априроири ограничило возможность проекта по следующим направлениям:

- Возможности соединия компонентной базы в единый поисково-аналитический комплекс
- Реализации проекта на полном массиве базы адрессов
- Более глубокого прохождения (кроулинга) «внутрь» структуры сайтов
- Полноценной работы с по-настоящему большими данными (изображения, аудио, пр.) и реализации более совершенного ML

Как мы будем преодолевать барьеры и трудности

- В рамках новых курсов мы получим тестовый доступ к достаточно мощным ресурсам по хранению и обработке данных, и далее закупим необходимые мощности, когда точно поймем, что наш продукт актуален для рынка
- Более глубоко изучим методы ML и библиотеки для краулинга, обкатаем все это на практике
- Усовершенствуем бота и сделаем интерактивный сайт

Что нас ждет в будущем

В перспективе проект будет реализован командой уже не на локальных машинах. Планируется доработка отдельных компонент и разворачивание поисково-аналитического механизма на серверах Yandex Cloud, что придаст проекту совершенно иные возможности.

Также в перспективе будут задействованы механизмы сбора, хранения и анализа больших данных.

Также прорабатывается идея получить на выходе полностью автоматизированную систему, где будет возможно на сайте или в телеграм-канале проекта задать рамки исследования и получить конкретные, коммерчески ценные результаты для своей предметной области.

GeekBrains

БЛАГОДАРИМ ЗА ВНИМАНИЕ!