Методические материалы к хакатону

**Zakupki Hack 2.0**

**Цель и задачи**

**Создать лучший интеллектуальный кластер товаров, работ и услуг (ТРУ) с учетом всех потребностей заказчиков и участников электронных торгов**

Кейс будет опубликован за сутки до официального начала хакатона. Data-сет будет представлять собой 4 выгрузки в формате csv.

**В рамках хакатона участникам предстоит реализовать задачи:**

Проанализировать бизнес-процессы, осуществляемые заказчиками и участниками закупок при подготовке к осуществлению закупочного процесса.

Собрать типовые аналитические потребности заказчиков и участников закупок при мониторинге рынка.

Проанализировать представленные data-сеты.

Разобрать такие сущности как ТРУ, товары-аналоги, сопутствующие товары. Обработать характеристики ТРУ, их компоновку, страну происхождения, поставщиков, ценовые предложения, конкуренцию и т. д.

Создать работающий прототип, с помощью которого можно найти необходимые ТРУ, их свойства, производителей, уровни цен и т.д.

Продемонстрировать работоспособность проекта.

Определить возможности по масштабированию решения и следующим его доработкам.

Презентовать проект.

**Описание полей data-сетов:**

product\_name - Наименование ТРУ

price - Цена за единицу товара с учетом НДС в рублях

product\_vat\_rate - НДС

product\_msr - единица измерения ТРУ

product\_characteristics - в этой строке указаны характеристики товара со следующей структурой: *Наименование: характеристика || Наименование: характеристика*

*Пример: Количество контактов: 102 || Номинальное напряжение: 100*

okpd2\_code - Код ОКПД2

okpd2\_name - Наименование ОКПД2

inn - хешированный инн поставщика

country\_code - код страны происхождения (если их несколько, то запись идет через ‘’ | ‘’; если кода страны нет, то значит, что она неизвестна)

*Все колонки находятся в текстовом формате (varchar)*

*Формат данных - csv, разделитель - ‘’ ; ‘’, кодировка UTF-8*

**Этапы защиты**

**Первый этап**

Защита проектов (презентация до 5 минут). Параллельно второй

участник команды демонстрирует работоспособность системы.

Максимальное количество баллов по критериям — 100.

**Второй этап**

Отправка ссылок на репозитории с кодом, визуалом и презентациями

(или гугл диск). Защита топ-5 решений (презентация — до 10 минут, затем 10

минут на вопросы жюри).

Максимальное количество баллов по критериям — 100. Баллы

суммируются с баллами первого этапа.

**Критерии оценки**

1. Работоспособность прототипа - 50

2. Оригинальность идеи - 10

3. Масштабируемость - 10

4. Уровень понимания участников сферы закупок - 10

5. Применение передовых Computer Science подходов - 20

**Библиотеки Python**

Numpy – Быстрая обработки массивов

Pandas – Работы с табличными данными

Natasha – Обработка текста на русском языке

Gensim – Обработка текста

scikit-learn – Реализованные алгоритмы машинного обучения

Keras – Фраемворк для проектирования нейронных сетей

PyTorch – Фраемворк для проектирования нейронных сетей

**Полезные материалы**

<https://habr.com/ru/post/244561/> - Разработка системы поиска

https://colab.research.google.com – Среда разработки от Google;

https://scikit-learn.org/ – Python-библиотека для машинного обучения;

https://radimrehurek.com/gensim\_3.8.3/index.html – Python-библиотека для

работы с текстом и создания текстовых моделей;

https://fasttext.cc/ – Python-библиотека для создания текстовых моделей;

https://docs.python.org/3.8/ – Документация Python;

https://www.anaconda.com/ – Дистрибутив языков Python и R,

включающий библиотеки для машинного обучения;

<https://habr.com/ru/company/ruvds/blog/494720/> - Шпаргалка по Pandas

<https://habr.com/ru/post/501214/> - Агрегация в Pandas

<https://fasttext.cc/docs/en/python-module.html> - fasttext

<https://pythonru.com/biblioteki/gensim> - Gensim

<http://nlpx.net/archives/57> - TF-IDF

<https://melaniewalsh.github.io/Intro-Cultural-Analytics/05-Text-Analysis/03-TF-IDF-Scikit-Learn.html> - TF-IDF Sklearn

<https://habr.com/ru/post/516098/> - Natasha