МОСКОВСКАЯ ПРЕДПРОФЕССИОНАЛЬНАЯ ОЛИМПИАДА ШКОЛЬНИКОВ

Профиль «Информационные технологии»

Командный кейс №1 «Классификатор по тематикам интернет ресурсов»

1. Условия

В наше время количество информации, получаемой человеком, растёт невероятно быстрыми темпами. Чтобы оптимизировать наше время и снизить нагрузку на мозг, людям проще классифицировать определенные данные, чтобы была возможность, например, не вчитываясь в новость, понять, с чем она связана.

Большинство объектов в информационном поле сейчас классифицировано, есть специальные датасеты в свободном доступе, начиная от квартир до категоризации эмоций людей на фотографиях. Такие данные при использовании машинного обучения дают огромный пул инструментов, позволяющих создавать невероятные приложения для оптимизации некоторых процессов жизни человека.

На данный момент классификаторы по тематикам ресурсов имеются, но у них очень много недостатков, например, низкая точность или же скудная функциональность.

Участникам Олимпиады предлагается разработать программу для классификации интернет-ресурсов по тематикам (например: образовательный, развлекательный, финансы и т. д.). Участникам нужно внести не менее 6 классов. Для обучения и тестирования программы можно использовать датасеты с открытым доступом. В программе необходимо реализовать классификацию ресурсов с применением машинного обучения и хранением обработанных и классифицированных данных в БД.

2. Техническое задание

Требуется разработать программное приложение, которое должно запускаться как минимум на одной из популярных операционных систем: Windows 10, дистрибутив Linux, MacOS.

Функциональность программы:

1. Внесение ссылки на сайт с целью его классификации;
2. Реализация базы данных для хранения классифицированных и обработанных данных интернет-ресурсов;
3. Возможность выгрузки данных сайта в формате CSV;
4. Возможность подгрузки файла с ссылками и данными для классификации и выгрузка файла с уже классифицированными данными офлайн.

5. Возможность увидеть метрики при прогоне тестовой выборки, например: accuracy, precision, recall.
6. Программа может иметь дополнительную функциональность (например, реализован выбор алгоритма, по которому происходит классификация: линейная регрессия, логистическая регрессия и т.д.) по желанию участника.
7. Интерфейс программы должен быть простым и интуитивным, не требующим дополнительного обучения.

3. Требования к документации

- Титульный лист (с указанием названия кейса и перечислением членов команды);
- Анализ технических требований;
- Структурная и функциональная схемы программного продукта;
- Блок-схема работы основного алгоритма;
- Описание проведённых испытаний в соответствии с регламентом кейса (снимки экрана и/или запись экрана с работой);
- Программный код (ссылка на репозиторий).

4. Регламент испытаний

1. Установка из файла;
2. Выбор и ввод ссылки для дальнейшей классификации;
3. Классификация, просмотр результатов;
4. Выгрузка результатов в таблицу;
5. Проверка работоспособности функций программы, описанного в документации.
6. Импорт ссылок из CSV для дальнейшей классификации.

5. Примерный перечень средств и инструментов для выполнения задания

_https://www.python.org
_https://scikit-learn.org/stable/
_https://www.tensorflow.org
_https://aws.amazon.com/ru/machine-learning/