Тема проекта: Анализ киберугроз с использованием методов Natural Language Processing (NLP)

Направление проекта: NLP

Датасет: cyber-threat-intelligence\_all.csv с [Cyber Threat Dataset: Network, Text and Relation](https://www.kaggle.com/datasets/ramoliyafenil/text-based-cyber-threat-detection?resource=download)

Автор: Шукирбай Даурен

# Введение

Целью данного исследования является разработка модели для классификации текстовых данных по различным типам киберугроз. Для достижения этой цели был написан и оптимизирован код, который включает в себя предварительную обработку текста, выбор модели машинного обучения и обучение модели с использованием кросс-валидации.

# Цели и задачи проекта

**Цель:** Создать модель машинного обучения для классификации текстов на различные типы киберугроз.

**Задачи:**

* Собрать и подготовить данные для анализа.
* Провести предварительную обработку текстовых данных.
* Исследовать и выбрать подходящую модель машинного обучения.
* Обучить и настроить модель.
* Оценить точность и производительность модели.

**Бизнес-требования:**

* Высокая точность классификации текстов.
* Быстрое время обработки для реального применения в системах мониторинга киберугроз.
* Гибкость модели для работы с новыми и изменяющимися типами угроз.

**Метрики оценки успеха:**

* **Точность (Accuracy):** Основная метрика для оценки производительности модели.
* **Кросс-валидационная точность (Cross-validation Accuracy):** Для оценки стабильности модели на различных подвыборках данных.
* **Время обработки:** Время, необходимое для предсказания класса текста.

**Необходимые данные и ресурсы:**

* Датасет с текстами, относящимися к различным типам киберугроз.
* Мощности для вычислений: CPU/GPU для обучения модели.

**Exploratory Data Analysis (EDA):** EDA включает в себя следующие этапы:

* **Проверка качества данных:** Удаление пропущенных значений и дубликатов.
* **Обнаружение закономерностей:** Анализ распределения данных, частоты слов и других характеристик текстов.
* **Выявление аномалий:** Обнаружение и анализ выбросов в данных.
* **Проверка гипотез и предположений:** Использование сводной статистики и графических представлений для проверки гипотез.

# Предварительная обработка текстовых данных

Первый этап работы заключался в предварительной обработке текстовых данных. Для этого была написана функция preprocess\_text, которая выполняла токенизацию, удаление стоп-слов и лемматизацию. Затем эта функция была применена к столбцу с текстом данных, и результаты были сохранены в новом столбце clean\_text.

# Выбор модели машинного обучения

Для классификации текстовых данных была выбрана модель логистической регрессии. Эта модель была выбрана из-за своей эффективности в задачах классификации текста и простоты в реализации. Для оптимизации параметров модели использовался метод GridSearchCV, который позволяет исследовать различные комбинации параметров и выбирать наилучшую модель на основе кросс-валидации.

# Разработка и обучение модели

Для разработки и обучения модели был построен Pipeline, который включал в себя TF-IDF векторизатор и модель логистической регрессии. TF-IDF векторизатор был использован для извлечения признаков из текстовых данных, а модель логистической регрессии - для классификации. Затем модель была обучена на предварительно обработанных данных.

# Результаты

После обучения модели с использованием кросс-валидации с помощью GridSearchCV были получены следующие результаты:

Лучшие параметры модели: {'model\_\_C': 100, 'tfidf\_\_max\_features': 3000}

Лучший показатель перекрестной проверки: 0.9414242728184554

# Заключение

В результате данного исследования была разработана и обучена модель для классификации текстовых данных по различным типам киберугроз. Полученная модель достигла высокого показателя точности на перекрестной проверке, что свидетельствует о её эффективности в решении поставленной задачи. Дальнейшие шаги включают в себя тестирование модели на новых данных и её дальнейшее улучшение для повышения качества классификации.