X5 Tech

X5 Tech AI Hack

СОЗДАВАЙ РИТЕЙЛ БУДУЩЕГО

Команда «Дайте два»

Описание решения

Наше решение для маскирования включает следующие этапы:
1. Предобработка данных - очистка и нормализация текста (удаление лишних пробелов, знаков пунктуации).
2. Использование предобученной модели («xml-roberta-large-ner-russian») для поиска «чувствительных данных».
3. Использование регулярных выражений для «чувствительных данных», которые не были предсказаны предобученной моделью.
4. Оценка результатов работы модели - сравнение предсказанных «чувствительных данных» с истинными данными.
Наше решение гибкое и масштабируемое, его легко адаптировать к различным типам текста.

Проверяемые гипотезы

Нами были проверены следующие гипотезы:
1. Предобученные трансформеры могут эффективно использоваться для определения «чувствительных данных».
2. Использование регулярных выражений, в сочетании с предобученными моделями, улучшает точность распознавания «чувствительных данных».
3. Предобработка данных, перед использованием предобученных моделей, повышает качество распознавания «чувствительных данных».

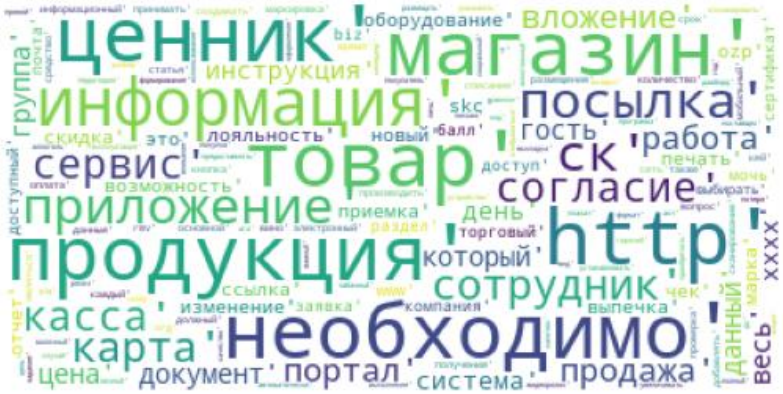
Используемые технологии

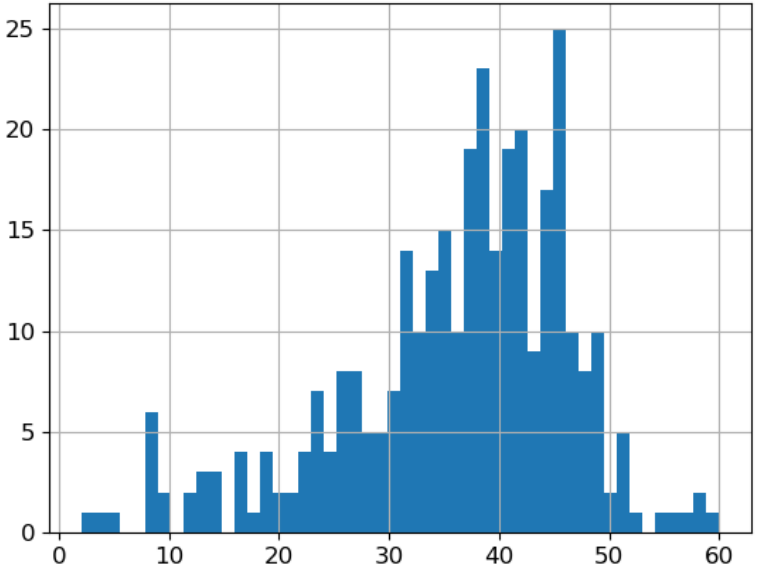
Язык программирования - Python.
Фреймворк машинного обучения для языка Python с открытым исходным кодом - PyTorch.
Программная библиотека на языке Python для обработки и анализа данных - Pandas.
Библиотека для работы с массивами и набором математических функций - Numpy.
Библиотека для машинного обучения - Scikit-learn.

Предобработка данных

1. Предобработка данных: очистка и нормализация теста (удаление лишних пробелов, знаков пунктуации).
2. Постобработка данных: удаление лишних символов начала и окончания найденных «чувствительных данных» (захваченные моделью знаки пунктуации, латинские буквы для русскоязычного текста).

Аналитика данных





Анализ получаемых результатов

В экспериментах были опробованы модели «из коробки»:

- yqelz/xml-roberta-large-ner-russian
- dbmdz/bert-large-cased-finetuned-conll03-english
- Jean-Baptiste/roberta-large-ner-english
- DeepPavlov/rubert-base-cased-conversational

Наилучший результат выдала «xml-roberta-large-ner-russian».

Были опробованы все стратегии объединения выдаваемых моделью токенов, наилучший результат достигнут со стратегией «none» с самодельным объединением токенов.

Модель уверенно распознавала персоны и организации, все остальные сущности были получены с помощью регулярных выражений.

Анализ получаемых результатов

Анализ получаемых результатов состоял из таких шагов:
Оценка точности - сравниваем предсказанные «чувствительные данные» с истинными значениями «чувствительных данных».
Точность обычно рассчитывается как отношение количества правильно определенных к общему количеству в наборе данных.
Анализ ошибок - проанализировать ошибки, чтобы понять, какие типы данных решение определяет неправильно. Это помогло определить области для улучшения и внести соответствующие корректировки в модель. xml-roberta-large-ner-russian была обучена на предоставленных данных + коллекция Named_Entities_5. На проверку результатов не хватило времени…

Q&A-сессия

Команда «Дайте два»