{Tender Hack}

Создание механизма контроля недопустимого контента

Команда [LIFE]

Глеб Илья Дарья

Новосибирск, 2021

Общая информация

Задача: создать систему комплексной проверки заявки на заведение новой СТЕ на портале поставщиков и оповещении поставщика и модератора о возможных ошибках.

Портал: https://edu.pp24.dev

Метрика: уменьшение времени между подачей заявки и ее опубликованием на портале.

Данные

Номер заявки	0
Форма заявки	Полная
Наименование	Изолента
Единица измерения	Штука
Изображение	158586639.0
Классификация ГОСТ/ТУ	NaN
Описание	NaN
Статус	Отклонена
Причина отказа	наименование заявки должно однозначно описыват...
Комментарий	Истек регламентный срок ответа
Ид оферты	68977414
Наименование оферты	Изолента
Артикул оферты	21146628-19
Регион поставки	Москва
Срок поставки в днях От	3
Срок поставки в днях До	7
Доступное количество От	102.0
Доступное количество До	102.0
Исходные характеристики	[{"name":"Производитель","value":"ИЭК"},{"name...
Категория оферты	Товары-Товары для ремонта или строительства-Ма...
Категория справочника	Электроизоляционные материалы
Вид продукции	Товары
Количество эталонных утвержденных характеристик в категории	12
Количество использованных поставщиком эталонных утвержденных характеристик	0

Задачи команды

- Анализ данных: табличные данные и изображения, форма заведения заявки;
- Выделение существующих проблем и постановка задач;
- Распределение задач между участниками команды;
- Построение единого пайплайна
- Презентация проекта
- Победа

Решение задач заполнения формы

1) Опечатки в тексте
2) Наличие обсценной лексики
3) Наличие наименования запрещенных товаров
4) Некорректное заполнение полей (символов табуляции)
5) Количество заполненных характеристик товара не менее 4-х
6) Наименование во множественном числе
7) Наименование полностью на английском языке
8) Наличие единиц измерения в наименовании характеристик, в значении характеристик, наименовании товара
9) Наличие артиклей в начале наименования товара
10) Название страны в графе производитель
11) Дублирование характеристик

Решение задач с использованием изображения

1) Проблема определения на изображении:
    - водяных знаков,
    - логотипов,
    - оттисков печатей,
    - ссылок на веб-ресурсы,
    - QR кодов.

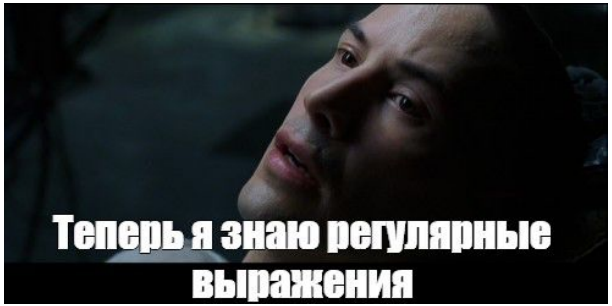
2) Проблема соответствия наименования СТЕ и изображения (изображение соответствует поставляемому товару)

Методы решения

- Обработка текста:
    - токенизация: NLTK
    - лемматизация: Mystem
    - транслитерация: Translit
    - приведение к нижнему регистру
    - удаление знаков пунктуации: regex
- Опечатки в тексте: YandexSpeller
- Наличие обсценной лексики: поиск по словарю
- Наличие наименования запрещенных товаров: поиск по словарю
- Некорректное заполнение полей (символов табуляции)
- Количество заполненных характеристик товара не менее 4-х
- Наименование во множественном числе: SpaCy
- Наименование полностью на английском языке

Методы решения

- Наличие единиц измерения в наименовании и значении характеристик и в наименовании товара: регулярные выражения/поиск по словарю
- Наличие артиклей в начале наименования товара: регулярные выражения/поиск по словарю
- Название страны в графе производитель: регулярные выражения/поиск по словарю
- Дублирование характеристик: регулярные выражения/поиск по словарю



Методы решения

1) Проблема определения на изображении:

   водяных знаков,

   логотипов,

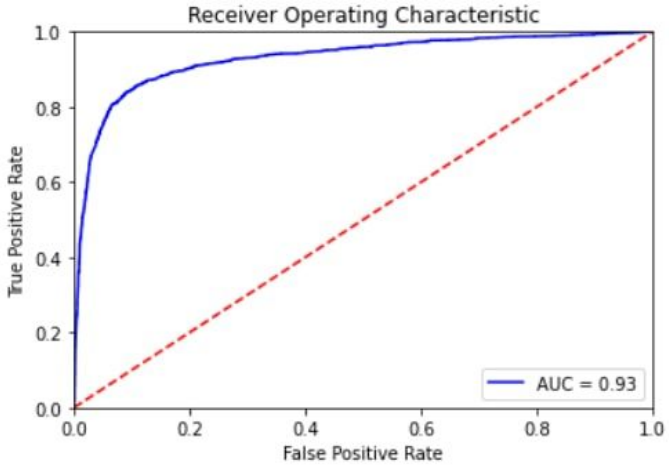
   оттисков печатей,

   ссылок на веб-ресурсы,

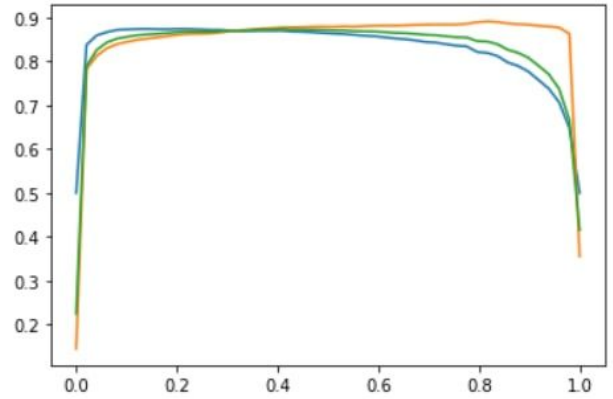
   QR кодов. — модель: NFnet_l0

2) Проблема соответствия наименования СТЕ и изображения (изображение соответствует поставляемому товару): — модель для текста: rubert-tiny
   модель для изображения: NFnet_l0

Результаты на тестовом наборе для определения водяных знаков




```
optimal threshold:  0.4081632653061224
Max F1_score:  0.8734226629459836
```



threshold: 0.4081632653061224

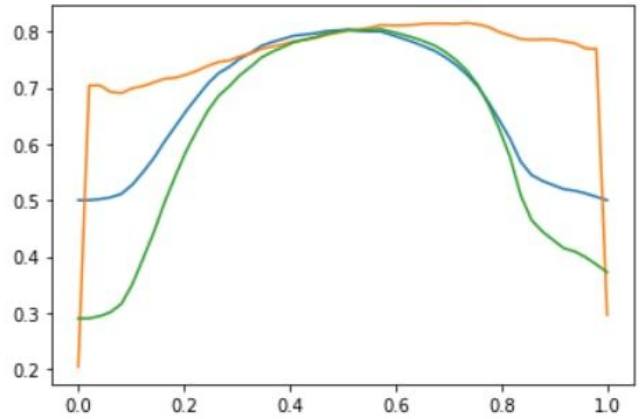
	precision	recall	f1-score	support
0	0.92	0.93	0.93	4984
1	0.83	0.81	0.82	2033
accuracy			0.90	7017
macro avg	0.88	0.87	0.87	7017
weighted avg	0.90	0.90	0.90	7017

Модуль для матчинга картинок и текстового описания

roc_auc_score: 0.8777400847562373

	precision	recall	f1-score	support
0	0.85	0.83	0.84	3737
1	0.76	0.78	0.77	2572
accuracy			0.81	6309
macro avg	0.80	0.80	0.80	6309
weighted avg	0.81	0.81	0.81	6309

Optimal threshold: 0.5714285714285714
Max F1_score: 0.8044525945876185



Результат

Номер заявки	197762
Форма заявки	Полная
Наименование	Настольная игра "Играем в математику"
Единица измерения	Штука
Изображение	2100917597.0
Классификация ГОСТ/ТУ	NaN
Описание	NaN
Статус	Принята
Причина отказа	предложенных характеристик недостаточно для со...
Комментарий	Добавьте характеристику: Состав набора/или Ком...
Ид оферты	840545233
Наименование оферты	Настольная игра "Играем в математику"
Артикул оферты	127
Регион поставки	Ямало-Ненецкий
Срок поставки в днях От	30
Срок поставки в днях До	45
Доступное количество От	1.0
Доступное количество До	5.0
Исходные характеристики	[{"name":"производитель","unitCode":"","newUni...
Категория оферты	Товары-Водогрейные котлы-Товары для детей-Игру...
Категория справочника	Наборы игрушек и настольных игр
Вид продукции	Товары
Количество эталонных утвержденных характеристик в категории	1
Количество использованных поставщиком эталонных утвержденных характеристик	1

```
results = form_checking_tests(d1, vocab_obscene, vocab_forbidden)
text_results(results, provider_comments, moderator_comments)
```

```
"Результат проверки заявки для поставщика: ['Необходимо заполнить минимум 4 характеристики.', 'Удовлетворение всем требованиям и нормам при заполнении заявки позволит ускорить процесс ее утверждения. В случае наличия ошибок заявка может быть отклонена.']. Результат проверки заявки для модератора: ['Заполнено менее 4 характеристик.']"
```

Спасибо за внимание!