Тестовое задание для соискателя на позицию "ML-Инженер"

Производство электронных изделий — сложный процесс, постоянно требующий решения множества интересных технических задач. Наша компания обеспечивает полный цикл производства - от поставки электронных компонентов, печатных плат до сборки конечных изделий для заказчиков. На должности «ML-инженер» необходимо будет разрабатывать различные WEB-сервисы, которые смогут помочь сотрудникам компании в их повседневной рутинной работе.

Одним из основополагающих направлений деятельности нашей компании является поставка электронных компонентов. Это подразумевает их закупку за рубежом, логистику, растаможивание грузов.

Все изделия, пересекающие границу РФ должны иметь ТН ВЭД код, который позволяет сотруднику таможни быстро определить, что за товар перед ним. Также с помощью кода можно быстро рассчитать таможенные платежи (таможенную пошлину, сбор и НДС) и отследить их перечисление.

Определением ТНВЭД кодов занимается отдел логистики нашей компании, это довольно трудозатратный процесс, учитывая количество позиций электронных компонентов, которые мы ввозим. Чтобы частично автоматизировать этот процесс – нами был создан сервис по автоматическому подбору ТНВЭД кодов на основании технической спецификации компонента (datasheet).

В качестве тестового задания – мы бы хотели предложить Вам создать подобную модель на основании наших данных.

Что мы используем

1. Python и библиотеки для анализа данных

- `Pandas` – базовая обработка данных
- `requests`, `BeautifulSoup`, `Selenium` – для поиска и парсинга веб-страниц.
- `pdfplumber`, `PyMuPDF`, `pdfminer` – для извлечения данных из PDF.
- `pytesseract`, `OpenCV` – для обработки изображений в документах.*(bonus)

2. Обработка текста (NLP)

- `spaCy`, `NLTK`, `transformers (Hugging Face)` – для анализа и классификации текстов.
- `TF-IDF`, `word2vec`, `BERT` – для векторизации текстов.

3. Машинное обучение и классификация

- `LLM, scikit-learn, XGBoost, LightGBM` – для построения моделей классификации.
- `PyTorch, TensorFlow` – для работы с нейросетями.

4. Работа с базами данных

- `MSSQL, PostgreSQL` – для хранения данных о компонентах.
- `Pgvector, Qdrant, Milvus` - векторные базы данных, для хранения эмбединнгов.

5. Разработка API и интеграция

- `FastAPI` – для создания API.
- `Docker` - для создания микросервисов
- `Git` - CI/CD и версионность документации

6. Дополнительные компетенции

- `OpenAI API, DeepSeek API` - помощь в разметке данных
- `Google cloud, Yandex cloud и т.п.` - аренда вычислительных мощностей
- `Airflow` - для систематизации сбора данных
- `Power BI или др. BI системы` – для визуализации полученных результатов

Формулировка задания

Мы ожидаем, что соискатель не будет сильно утруждён выполнением задания, потратит на него немного времени (за исключением самого обучения) и получит удовольствие. Мы хотели бы увидеть, какими подходами он пользуется при подготовке данных, анализе, какую библиотеку выберет для создания модели (вы не ограничены указанными выше библиотеками) и как будет оценена её эффективность.

Итак, задание подразумевает создание инструмента автоматического определения **ТНВЭД** кода по следующим входным параметрам: partnumber, manufacturer, текстовое описание компонента.

Датасет представлен по ссылке: https://disk.yandex.ru/d/yfCqYAVZnYgBhw

Описание датасета:

В файле `df_clear_uniq_3.csv`:

- `part_num` - партномер электронного компонента
- `manufacturer` – производитель электронного компонента
- `TNVED` – ТНВЭД код
- `concat_str` – объединённая строка part_num и manufacturer
- `path_ds` – путь ссылки на datasheet
- `id_path` – уникальный guid спецификации компонента

Архив `datasheets_miro_gpt4o_mini_7k.zip` содержит результат суммаризации текстов оригинальных pdf, получен при помощи модели chat-gpt-4o-mini.

Результат

В качестве результата – принимаем Jupiter-notebook-файл с подготовкой, анализом данных, обучением и оценкой результатов обучения.

И саму обученную модель (в архиве.zip)

Задание сформулировано достаточно открыто, соискатель может проинтерпретировать все неуточнённые моменты так как ему будет удобно. Но если захочется получить какое-то уточнение от нас, то мы с радостью ответим. Пишите нам при возникновении каких-либо вопросов! Также приветствуются нестандартные решения подобной задачи. Библиотеку оригинальных даташитов к датасету – можем предоставить.