Grepsr: Assessment

Assessment Overview

The objective of this assessment is to evaluate the candidate's ability to apply their machine learning skills, particularly in natural language processing, to solve a real-world problem. The candidate is required to develop an API that extracts meaningful attributes from an HTML block of an e-commerce website using an open-source large language model (LLM) and returns the relevant and meaningful information from the page in JSON format. The use of the ChatGPT API is prohibited for this task; candidates must utilize other available open-source models. We are looking for creative ways to solve this problem.

Task Description

1. **Development Environment Setup:**
    - Select and configure an open-source LLM such as LLaMA 3, BERT, Vicuna or any other suitable alternative.
    - Set up a development environment for API implementation using a framework like Flask or FastAPI.
2. **API Implementation:**
    - Create an API endpoint that accepts an HTML block as input.
    - Implement functionality within the API to utilize the chosen LLM for processing the HTML content.
3. **HTML Processing:**
    - Using ML, parse the HTML to identify and extract meaningful attributes relevant to e-commerce contexts (e.g., product names, prices, descriptions, images).
    - Identify the CSS selectors or Xpaths for each extracted attribute to pinpoint their location within the HTML structure.
4. **Data Formatting and Output:**
    - Format the extracted attributes and their respective selectors into a JSON structure.
    - Ensure the API returns this JSON formatted data as a response to requests.

Deliverables

- Source code for the API, including all dependencies and necessary configuration files.
- A brief report documenting:
    - The choice of LLM and rationale.
    - Details of the API's design and implementation.

grepsr

- Instructions for setting up and testing the API.
- Examples of input HTML blocks and the corresponding JSON outputs.

Evaluation Criteria

- **Creativity**: Candidates are evaluated on their innovative use of open-source LLMs to address the task of parsing and extracting structured data from HTML blocks of e-commerce sites. This includes creative solutions for handling different document structures, and extracting meaningful attributes effectively.
- **Decision Making and Problem Solving Skills**: This criterion assesses the candidate's ability to make sound decisions throughout the development process, including choosing the appropriate LLM, designing the API, and implementing solutions to potential challenges such as varying HTML structures and ambiguous data points.
- **Understanding of LLMs**: Candidates must demonstrate a deep understanding of large language models. This includes knowledge of how LLMs can be applied to natural language understanding tasks, their limitations, and the ability to integrate them into a functional system that operates on HTML content.

This assessment tests the candidate's technical skills and their ability to apply these skills to develop practical solutions. There is no single correct approach to this assessment. Candidates are encouraged to showcase their creativity and innovation in solving the task. They may use any tools, libraries, and services they consider appropriate to aid in the development and enhancement of the assessment, as long as they adhere to the constraint of not using the ChatGPT API. This flexibility is intended to allow candidates to demonstrate their unique problem-solving skills and technical prowess.