* Developed a Python-based web scraping pipeline utilizing libraries such as requests and BeautifulSoup to extract and preprocess textual content from dynamic websites, enhancing data accessibility for NLP tasks.
* Implemented tokenization strategies using models like Hugging Face's AutoTokenizer for BERT, OpenAI's tiktoken for GPT models, and Microsoft's Phi-3 tokenizer, optimizing input preparation for large language models.
* Engineered a token-based cost estimation module to assess computational expenses of LLM inputs, facilitating resource optimization in AI workflows.
* Integrated multiple AI/NLP frameworks within a unified Python workflow, demonstrating proficiency in leveraging diverse tools for large language model applications.
* Automated end-to-end processes from web scraping to token count analysis, streamlining workflows and enhancing efficiency in data preprocessing.
* Applied transformer-based tokenizers for real-world text preprocessing, contributing to the optimization of LLM inputs.
* Evaluated tokenizer performance and cost implications, providing insights for informed model selection and resource allocation in AI projects.