Jonathan Smoley
Professor Bowers
CPSC 324
5 April 2024

Project Proposal

Abstract

In the interest of pushing ahead with an ongoing project external to the Big Data Analytics course at Gonzaga University, namely a Discord-native chat bot that uses the Ollama infrastructure, this project poses a unique opportunity to train a Large Language Model (LLM) to the needs of the project. As such, I propose to utilize Google Cloud Platform (GCP) tools to develop a pipeline for ingesting and preparing data for use in training an LLM on the Gonzaga University public website.

Description

For the final project of the Big Data Analytics course at Gonzaga University I will demonstrate how GCP can be used for web scraping data from online resources and preparing said data for use in training an LLM. This is an end-to-end product that will require a resulting visualization in the form of a context string that will be provided as the system message template on top of an existing Ollama LLM. In other words, only the top layer of the LLM will be modified with the new context provided by this pipeline. Although the use of GCP tools for further LLM training is enticing, this is beyond the current scope of the project and will not be pursued.

Details

After an initial exploration of the tools necessary to build such a pipeline there are a number of services that could be useful. With Cloud Functions I can write a web scraper in Python (using Beautiful Soup) and deploy it. This will provide an ability to trigger the script along a specified interval or at a custom time. The trigger will likely be controlled by Cloud Pub/Sub so the script can run when another part of the pipeline requests new data. But, this trigger still will need to be scheduled, so something like Cloud Scheduler to set a time for this trigger could be useful. For storing the results of this operation, Cloud Storage will be the likely solution as many different types of objects can be sent to it. As for the initial dataset, the cloud function will be collecting HTML documents from the web which can be later formulated into whatever form is desired.

Services: Cloud Functions, Cloud Pub/Sub, Cloud Schedule, and Cloud Storage
Dataset: HTML, text documents from www.gonzaga.edu