Author: Ellis
Task: Skill Extraction
Scripts:

- `PhaseTwo/Utility/llm_remote_skill_extractor.ipynb`

Data:

- Overwrote “skills” column in: `PhaseTwo/Data/new_england_indeed_jobs.csv`
- `PhaseTwo/Data/skills.csv`

Methodology:

- Researched sentence transformers and keyword extraction models (like YAKE), but these wouldn’t take out just job skills.
- Then I looked at named-entity-recognition (NER).
- Was able to find various models on HuggingFace (HF), like a BERT based model, however this would be a victim of the same issue and recognize non-job skills.
- I stumbled upon SkillNER which is a job specific skill NER tool. However, this hasn’t been updated much in the past 4 years, and is bloated with dependencies. Most likely was “out of commission” with LLMs.
- Decided to research more for using a LLM. I was hesitant at first because I wanted to find a more specialized model that was light weight.
- Started to use the HF inference API with a free account using the `gpt-oss-20b` model with the `PhaseTwo/Utility/llm_remote_skill_extractor.ipynb`
- This required frequent timeouts and writes to the CSV to keep progress. Essentially a very manual process, but I wanted a way to avoid needing the actual tensors on my local machine…This was a waste of time and energy but I am glad I still attempted it as I learned a lot in the process and refreshed my memory on a lot of topics.
- If I wanted to actually take advantage of the benefits that `gpt-oss-20b` provides, I should have spent implementing the harmony renderer and having it locally installed.
- LLaMa CPP https://blog.steelph0enix.dev/posts/llama-cpp-guide/#prerequisites

Notes:

- Notably does not have University of Southern Maine