Current Selection

Current PROBLEMS

Current Solutions

Seek Scrape Development

# Concept

* Scrape a batch of 1000+ SEEK job posts to extract the description for software engineer skills via web scraper.
* Use spacy machine learning NLP model to extract the keyword for job skills.
* Collect the occurrence stats to graph the pattern for market demanding skills in software engineer.

## Current Development Status:

### Scope:

* A project to showcase the ability to scrape job market desired skills + learning process.
* Scope limited to:
  + Most occurrence of words for each skill ( including being a part of other skills counts).
  + Not considering smart recognition of a skill being a child of other skills without featuring the other skill’s name in there.

### Achievement:

* Able to scrape the correct SEEK web api for job posts (both entire list of jobs and each job specification)
* Decent trained spaCy NER model for recognizing software engineer terms
* Ranking most skills occurence in descending order on python GUI tkinter via matplotlib.

### Challenge:

* Spacy models haven’t yielded efficiently correct outcome (<90%) as of now.
* Training Dataset is not good or adequate in term of batch size and selection:
  + Might require data to be pre-cleaned.
  + Fix training dataset with better annotation
* 2 approaches for choosing the right model:
  + spaCy custom en\_core\_web\_md model for light computation combined with effective word vector:
  + Current issues:
    - [] empty job tuple in training data set causing training failure.
    - ?? general area of coding isn’t picked up by DeepSeek ( worth to include or no ?)
      * Classifying into group would need us to include.
      * Generate prompt:
        + DeepSeek.
        + Chatgpt.
        + Mistral?
        + CodeLlama?
  + spaCy lg model with transformer:
    - better accuracy
    - quicker to achieve ( less learning + more off the shell model)
    - BUT slower computation.
* Using DeepSeek AI model to extract software engineer skills and terms
  + - Better at classifying which skills fall under which umbrella of bigger skills