AI Engineer - Technical Test

Curate a large raw dataset into several subsets suitable for validating a language model. The large source dataset contains about 200,000 Jeopardy Questions from which you will form subsets. The data are currently linked from this Reddit post (navigated from a larger set of interesting text data on GitHub.)

Requirements

- Start with the file JEOPARDY_QUESTIONS1.json, downloaded from the link above
- We need strata suitable for comparing performance of a named entity recognition (NER) algorithm over three cases
  - phrases containing numbers
  - phrases containing non-English words
  - phrases containing unusual proper nouns
- Consider any or all of the features of the Jeopardy questions (e.g. categories, difficulty, question text, answer text)
- Create datasets of 1000 examples each
- Estimate how many examples of each type there are in total among the 200K possible examples

Specifications

- Present the curated datasets as json or jsonl files: there is no need to store the data in a database
- You do not have to label the data for the downstream NER task
- Write a short description about how your curation process works overall
- Document any helper functions you write
- Use any third party libraries you wish, but write a few comments about what the library is doing: assume that we are not familiar with the third party tools you choose
- The use of AI coding assistants (e.g. Copilot, Cursor, Windsurf) is encouraged
- Tie your work together with Python, but feel free to use command-line tools (e.g. jq, head) if they help with your analysis

Presentation

- Publish any scripts, notebooks and final datasets on Github
- Please make small incremental commits to show your process
- Submit by 30/09/2025T00:00:00Z