Data Engineering Project

OliveAI - Provider - Surgical Supply Variation

Purpose of This Project

Rather than test algorithms or arcane coding skills, we will assess **how you structure a project**. Ensure you spend enough time thinking about how to organize the code so it is extensible and scalable, rather than diving right in with the content of the functions themselves.

To that end, this should constitute a rough proof-of-concept; as much as can be reasonably accomplished with under 4 hours of work.

Have enough code stubbed out so we can anticipate where you will go next. Empty or partially written code with populated docstrings are acceptable.

For your own sanity, please don't spend more than 4 hours on this. Feel free to use all of the resources at your disposal, as you would normally do in the course of a project for work.

Deliverables

You can either post your results to Github and share them with me (bdoremus), or zip/tar/whatever the project and email it to me at ben.doremus@oliveai.com.

Please use Python 3.x.

Include instructions on how to run the code and a brief explanation of what to expect. We will be executing the script on our end to test the functionality, so a few sentences should suffice.

Scope

Assume that the problem as defined here is only the starting point for the project.

Given the business needs, you know that there will be:

- additional functions that need to manipulate this data both before and after the functions you write, though you do not yet know what those functions will be.
- multiple clients who may need slightly different implementations of this code.

Problem Definition

Input

./input/*.csv:

- Zero or more csv files that contain some IDs, a description of that ID, and the most recent modification date for that id's description.
- Has three columns:

  id
  description
  timestamp

Output

./output/id_descriptions.csv

- A flat file that combines together all of the files in /input/*.csv. **For each id, there should only be one description**. If there is more than one description per id, take the description from the entry with the most recent timestamp.
- Has these four columns:

  id
  description
  timestamp
  source_filename