Amazon-ASIN

Description

Appends data from amazon ASIN information.

Reads data from a dropbox folder and append the information into 5 GB csv files.
The columns that the file contains are: 'asin', 'manufacturer','invalid'.
These files are saven in a dropbox directory.

The python script reads a dropbox folder and list all the files that are inside this folder,
as well as the folders inside. Download each file and append the information with the rest.
As the output file is necessary to be less than 5 GB, once reached this limit the result file
is upload to the result's dropbox folder and a new file is started to be generated with the
information of the following input files.

As the input files are in different formats, a regular expression is used to validate the
expected name and extension of the file. In case that the file does not matches with this
regular expression the file will not be processed.

Configuration

The configuration is placed in the file `config.properties`.

- log_level: Level of log (DEBUG,WARNING,INFO)
- access_token: Token to access dropbox folders
- dropbox_folder_download: Path in dropbox for the folder to download the input data
- dropbox_folder_upload: Path in dropbox to upload the results
- dropbox_timeout: Timeout for HTTP connection with dropbox
- dropbox_chunck: Chunck size to upload data to dropbox
- file_regex: Regular expression that determines a valid name for the input file
- encoding_input: Encoding for input files
- data_folder: Local folder to save temporary the input files
- output_size_mb: Limit in MG for the output files
- result_prefix: Prefix used for the output files
- result_extension: Extension used for the output files
- result_folder: Path for the saving temporary the output files

Installation

```
python -r requirements.txt
```

Execution

```
python amazon-asin.py -c "config.properties" &> out.txt
```

Error Handling

If there is any error processing the input file, it is saved in the local directory data_folder.
If there is any error uploading the output file, it is saved in the result_folder.
The program will continue executing until processing the last input file.
In the log, the error messages could be seen.

Observation: As there are different type of formats and encodings, there might be some warnings because the script tries with different encodings to read correctly the file. If you see a warning related to this issue, it is not an error; it only means that the script tried with the incorrect encoding, so then try again with a different one.