-Project: Sanitized CSV Creator
-Version: 1.1
-Author: Alexander Kirk

Purpose

To create an easy way for data scientists to share code without exposing sensitive data.

This project will allow the user to create sanitized CSVs automatically, given any CSVs present.

This project does not intend to create CSVs tailored to what a user might need, though an example_csv_creator script is provided that one can model off of, if desired.

If certain columns are auto-sanitized in the result that you'd have liked preserved, no worries! The driver itself has sanitization options you can easily disable/enable.
The program also creates the sanitized CSVs separate from your original CSVs, so you can open both in a sheet editor like Excel and simply copy the original values into the sanitized version, where desired.

Just be sure to hide your original CSVs before sending your project to anyone ;)

Files Included

+sanitized_csv_project.zip
|
|+++README.txt (or README.pdf if you prefer bolding!)
|
|+++License.txt
| So no one gets in trouble for using this software 0:)
|
|+++setup.py
| ONE primary way to install needed modules (discussed in next section)
|
|+++requirements.txt
| ANOTHER good way to install needed modules (discussed in next section)
|
|+++driver.py
| The primary program you'll likely want to run. Creates sanitized CSVs off all existing local
| CSVs.
| You can edit the True/False values directly here to choose what is/isn't sanitized.
|
|+++example_csv_creator.py

|	Creates two example CSVs, if you'd like extra resources seeing how the program works.
|
|+++evaluator.py
|	During testing, I found it useful to be able to see what has been sanitized, without
|	needing to open up both the original and sanitized CSVs. In a terminal environment, this
|	program prints what columns are differentiated, or sanitized. I've left this in here just
|	in case you'd like to use it :)
|
|+++sanitization_pack
	|
	++++++++auto_detector.py
	|	The backbone responsible for automatically grabbing CSV files and applying
	|	sanitization functions -- creating new sanitized CSVs.
	|
	++++++++functions.py
		The actual sanitization functions invoked by our auto_detector and example_csv_creator

---

What Is Needed to Run This Program

A. A CSV file in the same folder.
Note: You will want clear column names for the detector to work properly.

B. Python3 interpreter in a terminal/shell environment

C. INSTALLING MODULES
I have tried to make this as easy as possible!
	1.) IF you have PYTHON 3.6+ and SETUPTOOLS installed:
	python3 setup.py install
	2.) IF you have PYTHON 3.4+ and a pip 9+ that corresponds to python3
	pip install -r requirements.txt

The above contains two easy ways to get everything you need, that should work on most python3 distributions.
We'll now proceed to the operating instructions.

If you are still having trouble, please peer down into the Further Help section to ensure you have a working environment and can run one of the above.

Operating Instructions

WITH ALL MODULES INSTALLED (See Section C. From Above):
python3 driver.py

Yup. That's all there is to it.

As mentioned, you can edit the sanitization options directly in the driver.

If you want some clear example CSVs, feel free to run the example_csv_creator.

FURTHER HELP IF NEEDED

If you are here, neither of the two installation methods worked for you.
You may have been missing setuptools, pip, or had an insufficient python version.
This section will detail one way you can get back on track.

1.) Install Anaconda
https://www.anaconda.com/distribution/#download-section

2.) Once fully installed, you should have a recent python3.7+ distribution, along with pip and setuptools
Which means..
python3 setup.py install

~~~~~~

EVEN FURTHER HELP

If somehow, even after all of the above, you *still* have trouble, then on the command line, run:
conda create -n test_env python=3.6
conda activate test_env
python setup.py install

You should now be able to proceed to the Operating Instructions section.

WHEN YOU ARE DONE RUNNING THE DRIVER PROGRAM, feel free to:
conda deactivate

You can always reactivate the virtual conda environment later via:
conda activate test_env

If you don't plan on reusing this program, you can delete the environment via:
conda env remove -n test_env
~~~~~~

Changes Since Version 1.0:

1.) Added a helper function to ensure* unique IDs are created for a column (even if the odds of making the same one would have been 1/285 decillion)

2.) As Sender/Recpient/Address are ambiguous terms that can refer to Email, Postal, IP, and the like, I've made tweaks to the detector to recognize them as such, and have it look for the first value of the column in question to make that determination.

3.) The auto-detector will no longer try to take in CSV files that begin with a period.

4.) Arguably the most important thing for a public software distribution, a license has been added!

5.) Simple README changes were made, like adding vertical spacing past the "What is Needed to Run" title, informing the user that column names matter for proper detection, adding a changelog of sorts here, etc.

Additional Remarks

Though I've created many tools and projects, this is the first one I'm releasing publicly.
If people find it useful enough, I'll clean up a bit and go the extra distance to make it more easily obtainable, and offer additional support.

And of course, if you find any bugs or issues, dislike or like this design, please let me know!
I can only grow with the feedback provided. Thank you for your time and for using this program.
Hopefully it proves beneficial

END