Datacleansing ReadMe

Introduction: This part converts the original JSON format files to CSV format.

But after conversion, in business dataset, every business's 'full address' field contains a Mac newline character '\r' which results in wrong line break. So I used Linux tr command to replace all '\r' with white space.

In review data set, Newline causes wrong line break in the CSV file. And since newline character in review text are newline '\n', and newline at the end of a record is '\r\n'. So I use tr command to first replace '\n' with whitespace and change remaining '\r' to '\n' to ensure that the resulting file works in linux.

Except for newline, there're commas in business dataset's 'neighborhood', and 'category' column because they are arrays. Also, there're commas in 'review text' column of review dataset. So I wrote changeDelimiter.py to change delimiter of the datasets from comma to tab.

For four commands in LinuxCommand_datacleansing file, the following are their inputs and outputs file names (all are in this folder):

#	Input	Output
(1)	yelp_academic_dataset_business.json	yelp_academic_dataset_business.csv
(2)	yelp_academic_dataset_review.json	yelp_academic_dataset_review.csv
(3)	yelp_academic_dataset_business.csv	yelp_business_clean.csv
(4)	yelp_academic_dataset_review.csv	yelp_review_clean.csv

For changeDelimiter.py:
Input: yelp_review_clean.csv, yelp_business_clean.csv

Output: yelp_review_clean_tab.csv, yelp_business_clean_tab.csv