Milestone 5 due Friday, 03/01.

This is the first of two milestones that involves cleansing your main dataset (aka `dataset1`) using Apache Beam.

1. Make a list of all the data formatting issues which could not be resolved through standard SQL as part of Milestone 4. For example, duplicate records, non-conforming dates, etc. Add the issues to the `TRANSFORMS.txt` file.

2. Choose one of the tables you identified in `TRANSFORMS.txt`. Write a short beam pipeline that cleanses the data in this table using a `ParDo`. The pipeline should contain the following logic:

- runs a BigQuery query on your main dataset
- makes an input `PCollection` from the BigQuery results
- writes the input `PCollection` to a local file named `input.txt`
- applies your custom `DoFn` through a ParDo
- writes the output `PCollection` to a local file named `output.txt`
- writes the output `PCollection` to a new BigQuery table in your main dataset

Coding Conventions:

- The Beam pipeline should be in a file named `pardo_<table>.py where <table>` is the name of the table that is being transformed.
- The BigQuery input and output tables should reside in your main dataset.
- The `ParDo` code should be commented sufficiently to understand the main logic of the transform.

CS 327E Milestone 5 Rubric
Due Date: 03/01/19

Create a file `pardo_<table>.py` that takes in data from your dataset1, performs a ParDo transform on the data, and writes it back out into another table. Sufficiently comment the code to show understanding of the Apache Beam pipeline. In addition, a `TRANSFORMS.txt` file should now be present for all groups. If a transformation could not be found, please refer to the TAs for assistance. **-100** missing `pardo.py` from repository **-50** code does not implement the ParDo transform **-50** code does not pull from or write back to your dataset **-40** code does not write to two output files `input.txt` and `output.txt` (these text files need not be pushed to your repo) **-30** code missing comments **-40** missing `TRANSFORMS.txt`	100

`submission.json` submitted into Canvas. Your project **will not** be graded without this submission. The file should have the following schema:
 	
`{`
`"commit-id": "your most recent commit ID from Github",`
`"project-id": "your project ID from GCP"`
`}`
 	
Example:
 	
`{`
`"commit-id": "dab96492ac7d906368ac9c7a17cb0dbd670923d9",`
`"project-id": "some-project-id"`
`}`
	Required

Total Credit:	**100**