CS 327E Milestone 3 due Thursday, 12/03.

This is the first of two milestones that makes use of Apache Beam. The goal of this milestone is to code the Beam pipelines which will be needed to implement your cross-dataset queries.

1. Create a new Jupyter notebook named `milestone3.ipynb`.

2. Identify all the tables from your refined datasets which contain data that needs to be cleansed. The tables you select should contain either some duplicate records or some fields which have non-conforming data or both. For example, non standard dates, addresses, person names, employer names, locations, genres, etc.

3. For each source table, write a Beam pipeline that normalizes the data from the table and creates a new table with the cleansed data. The pipeline should satisfy the following requirements:
    - It should be executed with the Direct Runner
    - It should run a BigQuery query with a `LIMIT` clause over the source table in your `refined` dataset such that the number of records being processed is < 500
    - It should make a `PCollection` from the BigQuery query results
    - It should implement at least one `DoFn`
    - It should apply at least one pardo to the `PCollection`
    - It should write the output to a local file by the name of `output.txt`
    - It should also write the output to a new BigQuery table in your `<source>_refined` dataset
    - It should be called from your `milestone3.ipynb` notebook.

4. Verify that the BigQuery output tables from the previous step contain a valid primary key. If the output tables are child tables, they should also have a foreign key. Run the SQL statements to verify these constraints from your `milestone3.ipynb` notebook.

Coding Conventions:

- Each Beam pipeline should be in a file named `<table>_beam.py` where `<table>` is the name of the table being transformed.
- The BigQuery output tables should be named `<table>_Beam` and reside in your `refined` dataset.
- The DoFn code should be commented sufficiently to show understanding of the transform(s).

CS 327E Milestone 3 Rubric

Due Date: 12/03/20

Implement one or more Beam pipelines that transform the data from the refined dataset. Sufficiently comment the code to show understanding of the Apache Beam pipeline. **-100** missing `<table>_beam.py` from repository **-50** code does not implement a DoFn transform **-50** code does not pull from or write back to your refined dataset **-20** missing `LIMIT` clause from input query **-40** missing or incorrect `pardo` call **-40** code does not write to output file `output.txt` **-50** code does not write to output table `<table>_Beam` **-30** code missing comments **-40** missing pipeline run call from `milestone3.ipynb` **-20** missing or incorrect primary key verification on output table `<table>_Beam` from `milestone3.ipynb` **-20** missing or incorrect foreign key verification on child output table `<table>_Beam` from `milestone3.ipynb` (assuming output table is a child table)	100
`milestone3.ipynb` and `<table>_beam.py` for each beam pipeline pushed to your group's private repo on GitHub. Your project **will not** be graded without this submission.	**Required**

`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**