Milestone 7 due Friday, 11/16.

Part 1:

Think of 4 interesting queries that span your *dataset1* and *dataset2*. These queries should use a join or a set operation to combine the data from the two datasets and the source data for these queries should require some prior transformation.

For each query:

- Briefly describe what results the query will produce and what SQL operations it will use to produce those results (1-2 sentences).
- Briefly describe what type(s) of transforms to the data are required to successfully implement the query (1-2 sentences).

Create a file `CROSS-DATASETS.txt` and add your descriptions and explanations to this file.

Part 2:

Keeping your cross-dataset queries in mind from Part 1, choose 2 transforms from the list of Core Beam transforms below. Try to choose the transforms which are most relevant to your cross-dataset queries.

1. `ParDo`
2. `GroupByKey`
3. `CoGroupByKey`
4. `Combine`
5. `Flatten`

For each transform, write a short beam program that applies your chosen transform. The program should create a Beam pipeline that contains the following logic:

- reads the contents of a local text file, `input.txt`, that has a few lines of sample data or runs a BigQuery query on one of your datasets that returns a few rows from a table
- makes a `PCollection` from the file contents or BigQuery results
- applies your chosen transform on the `PCollection` and outputs a new `PCollection`
- writes the output `PCollection` to a local text file, `output.txt`
- writes the output `PCollection` to a BigQuery table in your project

Naming Conventions:

- Both Beam programs should be self-contained in their own files. The files should be named `test_<transform>.py`. For example, the program that applies `ParDo` should be named `test_ParDo.py`. Be sure to debug and test your programs before pushing your `.py` files to your GitHub repo.

- Both BigQuery output tables should be created in a new `beam` dataset. The tables should be named after the transform that generated the output data. For example, the table that is produced by a `ParDo` transform should be named `beam.ParDo`.

CS 327E Milestone 7 Rubric
Due Date: 11/16/18

Part 1 - Create a file `./CROSS-DATASETS.txt` containing query and transformation information for 4 queries, as described in the outline. **-40** `./CROSS-DATASETS.txt` does not exist **-10** for each missing pair of query description and required transformation(s) description, up to **-40**	40
Part 2 - Create two files, `./test_<transform>.py` that each demonstrate the proper application of a transform to a PCollection using some sample data from a text file. The transformed collections should be saved as a new BigQuery table. **-30** each missing `./test_<transform>.py` file, up to **-60** **-30** each script that does not apply proper transform in name **-20** each script that fails to run from an error **-10** each new table missing from BigQuery	60

`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**