We have a dataset at Cogo Labs that is used for job market analysis. It is used to compare and contrast different types of jobs and to identify the types of jobs that someone may be in market for. We've made this data available at the following API:

* Host: EC2Co-EcsEl-MT18UEPNPS93-1308701016.us-east-1.elb.amazonaws.com
* Port: 8000
* Path: /records
* Pagination is available by including a page query parameter. Each page will return 10000 records.
* The returned results are in JSON format
* No authentication is required (¯\\_(ツ)\_/¯)

We have an exciting new opportunity with a partner named Liveworks. They have additional job market data they've made available to us and we'd like to do an analysis on how their data compares to our data. Their data is stored remotely in a MySQL database:

* Host: data-engineer-rds.czmkgxqloose.us-east-1.rds.amazonaws.com
* Port: 3306
* DB: liveworks
* User: cogo\_read\_only
* Pass: N&f#vSq9

All of the data they're sharing is located in the cogo\_list\_v1 table. Its schema is:

+-----------+--------------+------+-----+---------+----------------+

| Field | Type | Null | Key | Default | Extra |

+-----------+--------------+------+-----+---------+----------------+

| id | int(11) | NO | PRI | NULL | auto\_increment |

| name | varchar(64) | NO | | NULL | |

| address | varchar(128) | NO | | NULL | |

| birthdate | varchar(10) | NO | | NULL | |

| sex | varchar(1) | NO | | NULL | |

| job | varchar(64) | NO | | NULL | |

| company | varchar(64) | NO | | NULL | |

| emd5 | varchar(32) | NO | MUL | NULL | |

+-----------+--------------+------+-----+---------+----------------+

The column names in this table should be 1-to-1 with the keys in the JSON from the API.

**Your task:**

Please write a Python script that interacts with both datasets. It should answer the following questions:

1. How many users in the Liveworks dataset are also in Cogo's dataset? How many are just in Cogo's? How many are just in Liveworks'?
2. For the users within that intersection, what percent have different job titles between the two sets?

In addition to answering those questions, your script should output data for the users within the intersection. Please output the following three column CSV:

emd5, JSON list of "job" and "company" key/value pairs for Cogo, JSON list of "job" and "company" key/value pairs for Liveworks

Finally, please output a CREATE TABLE statement for MySQL (5.7) that could be created and populated with the CSV data generated by this script.

**Additional notes and requirements:**

* The emd5 value can be considered a unique identifier for each user.
  + That is, two records with different names but the same emd5 value should be considered one person.
  + Or, two records with the same name but different emd5 values should be considered two people.
* **Your script cannot read either source fully into memory at any point**. Data must be read in chunks. How this is implemented is up to you.
* Please properly test your code.
* Please write your solution in Python.

**Output**

Your job should output the following:

* **Line 1**: Total runtime of your script up to the time of output
* **Line 2**: The number of users within both datasets
* **Line 3**: The number of users unique to Cogo's dataset
* **Line 4**: The number of users unique to Livework's dataset
* **Line 5**: For users within both datasets, the % of users with different job titles between them
* **Lines 6 - 15**: A 10 line sample of your CSV output
* **Remaining lines**: The CREATE TABLE statement for the table that could house the CSV output.

**Submission**

* Please redirect the output of your script into results.txt
* Please add any required libraries into requirements.txt
* Please create a document named notes.txt with the following:
  + Instructions on how to run your script.
  + Instructions on how to run your tests.
  + Any additional info or comments you'd like to share on your submission.

When your script is complete and the text files are ready for submission please create a gzipped file named <firstname>\_<lastname>\_<year>\_<month>\_<day>\_data\_engineer.gz containing your source code and text files and email this to the recruiter you're working with.

Again, thank you very much for your interest in the Data Engineer position and for working on this assignment.