**Instructions**

Thank you for taking the time to complete this data analysis exercise. The goal of this is to help both you and Praedicat assess your fit for the Data Analyst role. This will help us by giving us more direct insight into how you work, organise data, and write code. This will help you by presenting you with a (scaled down) real world example of the kinds of problems we regularly face.

The goal of the project is straightforward: create a fused data set using the four input data sets included with this ZIP file. The included data represent information about chemicals from four different sources and we want to unify the data across these sources while keeping track of which source(s) contain each chemical. The data should be output in a single database (or equivalent) where each chemical is represented once along with all its basic properties from all four sources:

* Each chemical should be represented only once in the final data set
* Each chemical should have an associated list of the sources that mentioned it
* All the names/synonyms and CAS numbers for each chemical should be listed
* If you want to extract other properties (physical, regulatory, etc.) you are free to do so, but it is not required

The four data sets are:

1. ChemIDPlus. This is stored in XML format. The Document Type Definition (DTD) is included, and if you need more information on the file structure and elements it contains please email to ask, as there is no easily accessible guide online.
2. MeSH Supplemental Records. This is stored in XML format, with included DTD. For more information on the tags please see <https://www.nlm.nih.gov/mesh/xml_data_elements.html>
3. Pesticide Product Information System (PPIS). This is in text table format across multiple files. Please see ppisdata.pdf for the file formats. Only the chemicals listed with name or CAS number need to be included in the final output, not every product mentioned in the data set.
4. IARC List of Classifications. This is in Excel format. While we are expecting the complete contents of each of the other three data sources to be included in the final output, we are only asking that those chemicals from IARC that intersect with another data set be included.

The output database, whether implemented in flat-file or RDBMS format, should include a documented object model or schema that details each table, its columns and data types, any primary keys, foreign keys, and other constraints you would declare as part of building this database. Feel free to use whatever text or diagrammatic format(s) you feel will clearly represent this information.

We are less interested in the exact content of the results than in your thought processes and your approach to the problem. To that end, please be sure to send your (documented) code along with the result database and any other documentation/explanations you feel are warranted.

If you have any questions, please don’t hesitate to contact Adam.

**Description**

This application combines data from four diverse sources, and when aggregatted by CAS Numbers, generates a single, analytical dataset and loads it to:

1. a table in PostgresSQL database
2. .csv file

to make all the data available to the end user.

The following files should be provided and uploaded to ExperimentalMaterials folder as inputs:

1. chemid.xml
2. MeSHSupplemental.xml
3. List\_of\_Classifications.xls
4. chemname.zip.

**Before running the application:**

Make sure there is a database available in PostgresSQL as the target location for the data. User can choose the table to be created automatically at a runtime (see params.yml)

### User parameters in params.yml:

Under ‘files’ section, enter the path and names of input files to be combined. ‘**postgres’** section lets the user provide PostgresSQL credentials as well as table name. Finally, the name and location of output csv file can be specified in ‘**csv\_directory’ section**. ‘Test’ section applies to run the application in test mode (see below).

### To run the application execute the following command:

* python main.py

### Test mode

A total of two static files have been prepared to run the application in a test mode:

1. List\_of\_Classifications.xls,
2. chemname.zip.

both located in blabla folder.

These files may be replaced by different versions with modified data as long as the overall structure and format remains the same. The other two files can be substituted with any of files available on the following sources:

* MeSH Data from <https://www.nlm.nih.gov/databases/download/mesh.html>
* ChemIDplus Subset Data from <https://www.nlm.nih.gov/databases/download/chemidplus.html>

To run test:

* Set parameters under \*test\* section in params.yml
* execute the following commands:
* \*\*python test.py\*\*