ITEC649 2018 Python Assignment

This assignment involves writing Python code to extract information about jobs, people and companies from data files and load them into a consistent SQL database. It is an example of an Extract-Transform-Load (ETL) task.

You have been given the task of generating some normalised data on job postings given some data files in different formats. You are given:

- An HTML file downloaded from the Jobs! website that lists 50 jobs
- A CSV spreadsheet containing details of companies
- A CSV spreadsheet containing details of people

The HTML job listing mentions the title of each job and the company it is with. The CSV companies list includes the company name and some contact details including the name of a contact person. The CSV people list includes more complete details of those contacts plus some other people. Your task is to read the data from all of these files and add it into an SQL database.

The schema for the SQL database is provided for you in the file database.py. You can run this file to create the database. Your code will then add data to it. Note that:

- The companies and people tables are related through the contact field. In the companies table the value of contact should be the id of the corresponding person.
- The positions and companies tables are related through the company field. In the positions table the value of company should be the id of the corresponding company.
- In the companies CSV file, contact names are given in full but in the people CSV file they are split into first, last and middle names. You need to match up these records.

Useful Python Modules

Python has many useful modules for this task. You will want to look at: * the **csv** module for reading and writing CSV files * the **bs4** module (BeautifulSoup) for reading HTML files

and of course you will use the **sqlite3** module for handling the database.

Required Output

To show that you have completed the task successfully, you will generate a single CSV file report that contains the following fields: * company name * position title * company location * contact first name * contact last name * contact email

There should be one row in your output CSV file for every job in the HTML file.

You will also submit the code you have written to solve this problem. Your code **must** use functions and every function **must** include a suitable docstring that describes what it does. Each function should implement a logical part of the overall ETL process.

Additional Report

You should also submit a brief (1 page) report on the following topic:

This is just a trial data set for this task. The real data is much bigger, with around 10,000 people, 5,000 companies and 50,000 job listings. Thinking about how you have implemented the ETL process, describe any problems that might arise with such a large data set and how you might have to modify your implementation to address these problems.

Submission Summary

You should submit: * your code to read and process the data and generate the output CSV file (a zip file) * the final CSV file report * your report on the larger scale problem (PDF, 1 page)

Grading

The assignment is worth 15% in total:

- Correctly generated CSV output file (5 marks)
    - CSV will be re-generated from your code
- Code quality and presentation (5 marks)
    - Clear and readable
    - Appropriate use of functions to break down the problem
    - Documentation for all functions
- Written report (5 marks)
    - Clear writing
    - Good analysis of the problem
    - Good solutions proposed

Acknowledgements

The people CSV file is derived from one distributed via github: https://github.com/lawlesst/vivo-sample-data/blob/master/data/csv/people.csv