MCDA 5560, Summer 2019

# Assignment 1

(Issue: May 27, Due: June 15, 11:59PM)

- Instructor: Trishla Shah (trishla.shah@smu.ca)

---

## Question-1: Data Preparation using OpenRefine

1. **Objectives:**
   1) Learn how to prepare data for any business
2. **Software Requirements:**
   1) For this task you will require OpenRefine tool
   2) Signup for free OpenCage Reverse Geocoding API (https://opencagedata.com/api)
3. **Get Help - Resources (Created by: Trishla Shah)**
   1) https://www.youtube.com/watch?v=hx4kXxAIWLw – This video will show you Installation of OpenRefine, how to create a new project and upload dataset
   2) https://www.youtube.com/watch?v=u7Emm5zvOdg – This Video will show you how basic operations in OpenRefine such as how to convert text to number and vice versa. Also, it will show how to group records with the same name.
   3) https://www.youtube.com/watch?v=80_iYWnVJG4 – This video will show you sorting, custom text transformation by writing expression, how to split/merge two columns.
   4) https://www.youtube.com/watch?v=bf1faDPOVsk – This video will show you how to formulate the URLs and fetch JSON Data. It will also show you how to create column from JSON data.
4. **Task: Data Preparation:**
   1) Install **OpenRefine** as shown in the tutorial
   2) Create Project **question1** in OpenRefine
   3) Load **question1.csv**
   4) Remove all rows in which **id** is empty
   5) Remove all rows in which **email** is empty
   6) Split Full Name in to First name and Last Name. Store first name into **f_name** and last name into **l_name** (you will create these two columns).
   7) Code Gender (**sex** column) into **M** and **F** instead of Male and Female (Do not create a new column).
   8) Calculate age of the person on 24th May 2019 from the **birth_date** column. Store calculated age in **age** column (You will create this column).
   9) Only keep last 4 digits of credit card in the **credit_card** column (Do not create a new column).
   10) Delete **full_name** column.
   11) Delete **birth_date** column.

12) Signup for OpenCage Reverse Geocoding API (https://opencagedata.com/api) and get API Key.
13) URL format to retrieve location (JSON Data): https://api.opencagedata.com/geocode/v1/json?q=LAT+LNG&key=YOUR-API-KEY
14) You can replace the text highlighted in red using expression and retrieve location detail (JSON data). You can store this JSON data into **raw_data** column (You will create this column).
15) Filter continent and country from the JSON data and store into **continent and country** columns (You will create these two columns).
16) Delete **raw_data** column.
17) Export project as well as CSV file. Rename project file as **solution1.tar.gz** and rename CSV file as **solution1.csv**
18) Include screenshot of steps shown in the OpenRefine in the report. Include count of rows (total number of rows. For example – 980 rows) in the report. Also, you need to include description of all important steps.

## Question-2: Data Preparation using Google Cloud Dataprep

1. **Objectives:**
   a. Learn how to prepare data on cloud platform for any business
2. **Software Requirements:**
   a. Signup for Google Cloud Dataprep (https://clouddataprep.com). Note: You will require credit/debit card to register for a trial. You will get $300 credit so you will not pay anything to do this task.
3. **Get Help - Resources (Created by: Trishla Shah)**
   a. https://www.youtube.com/watch?v=w4NGD2GryJk – This video will show you Signup process and all operations including joins on Google Cloud Dataprep.
4. **Task: Data Preparation:**
   a. Signup for Google Cloud Dataprep
   b. Load **question2_1.csv**, **question2_2.csv**, **question2_3.csv**
   c. In this question, we want to generate 3 output datasets which are as follows:
      i. Output Dataset 1 (**solution2_1.csv):** This dataset should contain all stocks related to **finance sector**. Also, the stock market capital of those stocks must be equal or more than $30M. This dataset should contain the following columns: stock_market, stock_symbol and stock_market_cap.

ii. Output Dataset 2 (**solution2_2.csv):** This dataset should contain all stocks related to **technology sector**. Also, the stock market capital of those stocks must be equal or more than $30M. This dataset should contain the following columns: stock_market, stock_symbol and stock_market_cap.
iii. Output Dataset 3 (**solution2_3.csv):** This dataset should contain all stocks related to **health care sector**. Also, the stock market capital of those stocks must be equal or more than $30M. This dataset should contain the following columns: stock_market, stock_symbol and stock_market_cap.

d. Please do all necessary steps to generate the above-mentioned datasets such as merge/split operation, conversion of billion to million, remove stocks with blank record or stocks for which stock market capital is not applicable (n/a) and other operations (if applicable). Remove all the duplicate stocks of that sector.
e. Include screenshot of steps and the flow the report. Include count of rows for each output dataset (total number of rows. For example – 980 rows) in the report. Also, you need to include description of all important steps.

## Question-3: Data Preparation (ETL Pipeline) using Microsoft SSIS

1. **Objectives:**
   1) Learn how to build ETL data preparation pipeline from data sources.
2. **Software Requirements:**
   1) For thiss you will require Microsoft sql server and management studio, SQL Server Integration Services and Visual Studio
3. **Get Help - Resources (Created by: Trishla Shah)**
   1) https://www.youtube.com/watch?v=orsLUinyyjo - This video will show you the complete installation of all the required software of this assignment.
   2) https://www.youtube.com/watch?v=uO5iiaMeNoE – Example (part-1)
   3) https://www.youtube.com/watch?v=1BlIK34_120 – Example (part-2)
   4) https://www.youtube.com/watch?v=kY1pzTbSs1U – Example (part-3)
   5) https://www.youtube.com/watch?v=EJ4GB1OUQbY – Example (part-4)
4. **Task: Data Preparation:**
   1) Choose appropriate application data sources.
   2) Identify data preprocessing tasks and design an ETL pipeline.
   3) Use an ETL tool to implement the pipeline for generating target dataset(s). Your ETL implementation should perform the following tasks:

   -All SQL tables created from the source datasets should be in either 3rd or

BC normal form, with identifiable relationship with other tables.

- A workflow should also allow you to move data from the SQL tables to target flat file(s) based on partitioning predicates (merge, conditional splits, derived columns and aggregations) for the analytical task.

4) A brief description on the steps of using the tool to generate your result.

## Report and Submission Guidelines

- Write description of each question along with screenshots in the report. Also, give a comparison of all tools in the report (DO NOT GOOGLE IT. WRITE COMPARISION FROM YOUR OWN EXPERIENCE).
- Submit a zip file **assignment1.zip** which should contain assignment1.pdf, all output csv files and any other files/script (if any). Do not include word file.
- Also, mention your team members' name in the report as well as in the comment/note (you will see one description box in the submission folder.)