Trial Task
Due Date: 12th November 2020

Mandatory: Use comment lines, functions and classes where appropriate.
Task 1A: Human Metabolome Database (HMDB) https://hmdb.ca/

1. Download the XML File from the following link (https://hmdb.ca/downloads)
   Filename: "All metabolite" released on 2020-09-09
   Caution: File is too large for the regular browser to process/handle.
   Hint: Split the XML nodes by the tag named <metabolite> into smaller XML files
2. Review the tags and values under normal and abnormal concentrations of tags
3. Calculate the total number of diseases associated with the blood and list all the disease with accession number in CSV file format (important)
   Hint: <biospecimen>blood</biospecimen> and disease will be tagged as <patient_information> under the abnormal concentration tag and <subject_condition> under normal concentrations tag.
4. Filter the XML files with blood as biospecimen on both abnormal and normal concentrations tag
   Caution: Henceforth, we will only work on filtered XML file containing blood as biospecimen on both abnormal and normal concentration tag
5. Convert them into given template A
   Hint: Select normal concentrations, abnormal concentrations and diseases tag from the original file
6. Standardise the age into a numerical format
   Hint: Age format YY, MMDD (Y: Year, M: Months, D: Days)
7. Convert the XML files into given template B
   Hint: The values of the following tags will be the same as given in the template (creation, update, version, completion status and ethnicity tag)

Task 1B:
Note: The following two tasks are independent of the above task. However, they are closely related.

1. List all the different units associated with concentration value and develop a system which can convert the values of other units into the desired value (For example, 6 kilograms can be converted into 6,000 grams and vice-versa)
2. Develop a search system for a synonym to access its root name. For example, if the synonym for the Iron is "Fe" then whenever "Fe" is called it should return Iron.

Task 2: Perform web-scraping on metagene.de (https://www.metagene.de/start.html)

1. Perform web-scraping on the list of all disease in metagene.de and extract all the values associated on the right side of the panel (Example: disease, symptoms etcetera)
2. Convert the extracted values into an XML file.
   Hint: Use template C as a reference
3. Filter diseases with blood or plasma as a specimen (Check step 4 of Task 1A)
4. Propose and standardise the age into the numerical format.
5. Analyse, how many of the diseases and metabolite in the metagene overlaps with the HMDB
6. Standardise the age in numerical format
7. List all the different units associated with concentration value and develop a system which can convert the values of other units into the desired value (For example, 6 kilograms can be converted into 6,000 grams and vice-versa)

Optional tasks:

- Propose a classification technique to classify the diseases using predictor variables such as age, sex, metabolite concentration etcetera from the above data and provide necessary details or justification.
- Develop an auto/runtime application for Task 1A, which performs all the above task