ASSIGNMENT 2

İSTANBUL TECHNICAL UNIVERSITY
Department of Computer Engineering
BIL103E – Introduction to Information Systems – Fall 2017
Assignment 2

High-level Instructions

✻ See page 2 for step-by-step instructions. ✻

- STEP 1 (Week 10: 13 November approximately):
  - Use the *below link* to make a GitHub repository to hold your assignment files and for submitting them.
  - Download a dataset related to a topic in which you are interested.
  - Load the dataset into the spreadsheet program LibreOffice Calc and do some basic data processing.
  - Ensure you have a version of the dataset in CSV format.
- STEP 2 (Week 11: 20 November approximately):
  - Take the provided template Python program that can load a CSV file and run on your spreadsheet.
  - Change the program to do some simple calculations on the data, and print out some values.
  - Push your code to GitHub.
- STEP 3 (Week 12: 27 November approximately):
  - Change your Python program so that it can load the CSV file and output an HTML file that contains some of the data and the results of some calculations.
  - Push the Python program and the output HTML to GitHub.
- STEP 4 (Week 13: **4 December 9:30am**):
  - Change the Python program so that that can load the CSV and generate some statistics on the file.
  - Push the spreadsheet, CSV, Python program, and HTML output all to GitHub **by the due date and time** so that it can be collected for marking.
- Ensure the files are in the correct place so that your submission can be marked.

Submission Notes

- All steps of this assignment will be evaluated only after the final submission time (4 December 9:30am) has passed. However, it is recommended that you get the earlier steps done by their respective due dates.
- **By the final due date and time, you need to push your files to the repository that you create using the link below so that they can be downloaded by our markers.** The files required in your repository (not in a subdirectory) are:
  - `a2_input.csv`
  - `a2_output.html`
  - `a2_process_data.py`
  - `a2_processing.ods`

  To check which files are ready for submission, use the following link:

  http://web.itu.edu.tr/djduff/BIL103E2017/assignment2_student_links.html
- ***Keep your files small in size, otherwise they may not be collected.***
- Check the separate evaluation form to see on what basis your markers will be grading you.
  - Use all of the data processing techniques shown in the evaluation form.
  - Use all of the Python techniques shown in the evaluation form.
- Have fun.

✻ Keep your eye on the separate evaluation form, for the marks. ✻

Step 1

Finish by approximately 13 November.

Make your repository

Click on this link:

https://classroom.github.com/a/ZTCraGll

Follow the instructions for making your repository. Use the same GitHub username that you have previously registered with us via the Ninova upload of GitHub usernames (assignment 1 step 0). You will use your Git knowledge to clone this repository to your computer, commit your work to the repository as you go, and push it regularly to GitHub. The GitHub repository is the place from which your files will be collected for marking.

Download a dataset

Search the internet for a spreadsheet dataset that contains data about a topic in which you are interested (preferably, but not necessarily related to your first fan-page website). For example, if you are interested in the subject of climate change you might check the data at http://datahub.io/organization/climatedata, or if you are interested in socio-economic gender equality, take a look at http://datacatalog.worldbank.org/?Topics=Gender. The Turkish government maintains some statistics online at http://www.turkstat.gov.tr. And so on. Most data in this kind of repository can be exported in at least one of CSV, ODS, XLS or XLSX file formats.

You will need to convert the data to the formats you need for this assignment (you may use LibreOffice Calc for this task). In particular you will need the data as:

- The CSV file for processing with Python (see steps 2-4)
- The ODS file for processing with LibreOffice Calc.

When it comes time to submit your assignment, the resulting CSV file should be named `a2_input.csv` and the ODS file (which should also include the data processing activities from this assignment) should be named `a2_processing.ods`.

✯

When you click the above link to create your repository, CSV and ODS files with the desired names will be created (you will need to replace these with the files that you download).

Calculate summary statistics

Paying attention to the evaluation criteria in the separate evaluation criteria sheet if you are interested in the marks, load your ODS file into LibreOffice Calc and do some exploratory data analysis.

For example, try calculating some summary statistics using functions such as COUNT, SUM, AVERAGE, or similar (the easy set), as well as COUNTIF, SUMIF, SUMPRODUCT, or similar (the difficult set). Also, try sorting, filtering, making charts and formatting the spreadsheet for readability. Save the results in the file `a2_processing.ods`.

✯

You can create multiple "worksheets" in the spreadsheet if you wish to do different operations on different worksheets.

Think about what these statistics and visualisations might tell us about the topic of the data.

Step 2

Finish by approximately 20 November.

Use Python to run the program that loads the file

In your GitHub repository, there should be a Python program file called `a2_process_data.py`. Ensure that that file, together with the CSV data file from the previous step (`a2_input.csv`) are in the current directory.

Now if you run the Python program file, you should obtain the following output:

```
This assignment (assignment 2) hasn't been finished.
All it can do is print out the contents of a couple of cells of the file a2_input.csv:
Cell at index 0,0:
Ada
Cell at index 0,1:
Lovelace
Cell at index 1,0:
Grace
```

ℹ️**Note**: To run the Python program from the command line you would use the `cd` command to reach the correct directory and then run the Python command – for example, if your files are in the directory */home/damien/a2-damien*, you would write:

```
cd /home/damien/a2-damien
python3 a2_process_data.py
```

Note that we use the *python3* command as we are using Python version 3.

✋

FOR THIS ASSIGNMENT WE WILL BE MARKING USING PYTHON VERSION 3.5.2.
Do not develop your code for any version less than 3.3 because it probably won't work for the markers.

Make some changes to the Python program

Now you can edit the Python program file to change its output. In the program, the code to read the CSV file is already written and you do not need to change it. Thereafter, the variable `contents` will contain the content of the CSV file. `contents` will be a "list of lists".

✻ Make sure that you read the comments in the Python file so that you know which parts you should change. ✻

Try making the following changes:

- Print the variable `contents`.
- Print the variable `contents[0]`.
- Print the variable `contents[0][0]`.
- Print the type of the variable `contents` using the `type()` function.
- Print the type of the variable `contents[0]`.
- Print the type of the variable `contents[0][0]`.
- Print some other members of the `contents` list of lists, if they exist. E.g. `contents[5][6]`.
- Print the type of some other members of the `contents` list of lists, if they exist. E.g. `contents[5][6]`.
- Multiply, subtract, exponentiate, and add some of the numerical values; save the result in variables with descriptive names (like `agetotal`, for example).

✻ Saving a value to a variable saves it in memory; it is not the same as saving to a file. ✻

- Try to calculate a mean average of some variables and save the result in a variable.
- Print the variables that you have saved.
- See what happens if you try to print variables that don't exist, such as `contents["chickenchicken"]`. Keep a note of the error message.
- Concatenate some of the strings and print the result.
- See what happens when you try to concatenate a number and a string. Keep a note of the error message.
- See what happens when you try to multiply a number with a string.
- Print the types of the results of these calculations.
- Run the `help()` function on some of the variables you have created.

Note: Some interactive Python programs make life easier for you, such as:

▹ jupyter-console or jupyter-notebook
▹ spyder

If you do this, the easiest way of making sure the program can access your data files and Python programs is to run the interactive program from the same directory as your data files and Python programs.

For example, if your files are in the directory */home/damien/a2-damien*, and you are using jupyter-console, you would write:

```
cd /home/damien/a2-damien
jupyter-console
```

Then in the case of jupyter-console you would run the following command to run your Python program from within jupyter-console:

```
%run a2_process_data.py
```

Thereafter, the `contents` variable will be available to you to manipulate as you please. For example, try typing:

```
contents
contents[0]
contents[0][0]
```

Step 3

Finish by approximately 27 November.

Output HTML from your Python program

Edit your program `a2_process_data.py` so that it produces valid HTML 5 as output instead of plain text. If the output of the program is saved to file (called `a2_output.html`), that file can then be loaded into a browser and displayed as a web-page or uploaded to an HTML 5 validator like http://validator.w3c.org/.

Using the command-line the necessary command would be:

```
python a2_process_data.py > a2_output.html
```

✻ The ">" character is called a "redirection" and redirects program output to the given file-name. ✻

As an example, if your program previously outputted the text on the left, it should now output the text on the right:

Non-HTML	HTML

Mean temperature in India for 2012 was 24.23.
Mean temperature in India for 2013 was 24.69.
	<!DOCTYPE html>
	<html lang="en">
	  <head>
	    <meta charset="utf-8">
	    <title>India Temperature Analysis</title>
	  </head>
	  <body>
	    <p>Mean temperature in India for 2012 was 24.23.</p>
	    <p>Mean temperature in India for 2013 was 24.69.</p>
	  </body>
	</html>

Push everything to GitHub

Ensure you commit your files and push them to GitHub. You learnt how do do this for assignment 1.

To check which files are ready for submission, use the following link and search for your student ID:

http://web.itu.edu.tr/djduff/BIL103E2017/assignment2_student_links.html

Step 4

Finish by 4 December 9:30am. This is when your assignments will be collected.

Output the whole input CSV as HTML

In the previous steps you changed the provided Python program that loaded the input CSV so that it did some calculations and outputted some HTML. Now you will add to that program another feature.

Now change the the program so that it loads the input CSV and prints the *whole file* as appropriately formatted HTML. Since the CSV represents a grid of data it would be advisable to print HTML that presents the data in a grid.

To do this properly you will need to use flow-control. In particular, you will need to write either `while` loops or `for` loops.

Calculate and output summary statistics

In previous steps you have not been asked to calculate summary statistics on the data because doing so properly requires you to use flow control – in particular, `while` loops or `for` loops.

In the first step you will calculate a summary statistic (such as a spreadsheet may calculate in a `SUM` or COUNT or `AVERAGE` function). For example, you might average a column of numbers.

Next, you will calculate a more sophisticated summary statistic (such as a spreadsheet may calculate in one of the more advanced functions like `SUMIF`, `COUNTIF` or `SUMPRODUCT`). For example, you might average numbers in a column matching a condition..

The output of this will be formatted as HTML as you have been doing since step 3.

Make the output into a readable document

You now have a program that outputs HTML containing the whole dataset presented nicely, as well as some statistics calculated on that dataset. Now change the program so that the output HTML is in the form of a report containing the data as a table, the summary statistics and a brief (1-2 paragraphs) report about this investigation. In particular, write about what questions about the data itself your brief investigation may have answered. For example, if your dataset was about temperatures in India for the last 100 years, has India warmed a lot since the 1960s? Do summer and winter temperature trends follow the same pattern? Etc.

Push to GitHub

Ensure you commit your files and push them to GitHub. You learnt how do do this for assignment 1.

To check which files are ready for submission, use the following link and search for your student ID:

http://web.itu.edu.tr/djduff/BIL103E2017/assignment2_student_links.html

✯

If you get the code and output pushed to GitHub in the correct location by the due date and time, it will be marked.