SEN163A – Fundamentals of Data Analytics
Formative Assignment - Large-scale Internet Data Analysis

dr. Jacopo De Stefani - Joao Pizani Flor
Based on the material by T.Fiebig

March 21, 2022

Learning Objectives

- Identify fundamental pitfalls inherent to data analysis and identify whether they exist in case-studies.
- Propose and develop suitable data mining pipelines for different case-studies.
- Learn to efficiently load large data sets.
- Combine and execute data queries to find relevant data.
- Analyze a given dataset to answer a given research question.

Disclaimer

All characters and other entities appearing in this work are fictitious. Any resemblance to real persons or other real-life entities is purely coincidental.

Introduction

Tabularazor Inc. is a large national newspaper covering all kind of sectors (from sport to finance). Unfortunately, due to an electrical malfunction, their archive server is not accessible anymore. However, due to their implementation of an effective backup strategy, they have been able to recover the archives of the data of the previous years. You can find the archives of the newspaper at https://jdestefani.github.io/SEN163A-TabularRazorArchives/.

For this assignment, you are required to analyze the available data and metadata, to perform an investigation, to answer to the following questions:

a. Are there couples among the employees. If so, who? Are they still together?

b. Did any of the employees have a child? If so, who?

c. If you would be looking to work for Tabularazor Inc., how many holidays can you expect to get per year?

You need to clearly document your method and code used during the investigation and support your answers for items *a*-*c* with appropriate visualizations.

Deliverables

You will report your investigation through a **Jupyter Notebook**, including both your code and considerations on your findings. The use of figures, formulas, tables and pseudo-code to support your analysis is strongly encouraged. In case you used specific python libraries, you will need to include these libraries in a `requirements.txt` file. A template for the **Jupyter Notebook** and an example of `requirements.txt` file are available on the assignment page.

Evaluation criteria

The final grade for this assignment will be calculated based on the following criteria:

- **Quality of the report** - 35%
  - Reasonable formatting of the document and used citation appropriately
  - Use of proper English (typos, grammar)
  - Code script deliverable
  - Code quality
  - Problem Description
  - Dataset Description
  - Limitations
  - Clear Conclusion/Action recommendations
- **Functional tasks a. to c.** - 65%

Rules for the assignment delivery

To be read carefully !

1. The assignment must be developed in groups of 4 students.
2. The assignment must include the **name** and **student id** of all the students.
3. The assignment must be submitted in **Brightspace** as a **Zip file** containing:
   - A jupyter notebook (`.ipynb`) containing your code and the discussion of the results
   - *(If you used additional libraries)* A file `requirements.txt` containing, one per line, the additional employed libraries to solve the assignment.
4. You have to follow the following constraints:
   - Upload of a file `Group_X.zip` on the Brightspace page of the course, where `X` should be replaced by your actual group number./
   - Date: **Friday 01 April 2022**
   - Time: **Before 18:00**

   After this deadline the assignment will be considered as late and **no feedback will be provided**.
5. Use the feedback you will receive to improve your work for the final version due on **12 April 2022 18:00**.
6. **Knock-off criteria:**
   - Missing names and id on the document/document name.
   - Code not executable: Missing libraries - Errors.