Instructions for ZIP-File named: Lab 3 - Group 11

**--- STEP 0: Importing packages ---------------------------------------------------------------------------------------------------------**

The following packages and modules are used:

-requests

- pandas

- time

- numpy

- bs4 and BeautifulSoup

- concurrent.futures

- seaboard

- matplotlib.pyplot

**--- STEP 1: Opening the zip ---------------------------------------------------------------------------------------------------------------**

When opening Lab 3 - Group 11.zip the following files be visible:

- README.txt

- scraping.py

- resultScraping.py

- analysis.py

- SEN163\_Lab3\_Group1.pdf

- Structure site.pptx

- Graphs

- ResultsCSV

**--- STEP 2: Handling the folder ----------------------------------------------------------------------------------------------------------**

The textual assignment is can be accessed by opening: SEN163\_Lab3\_Group11.pdf

The scraping code can be accessed and run by opening: scraping.py

The analysis code can be accessed and run by opening: analysis.py

**--- STEP 3: Running scraping.py --------------------------------------------------------------------------------------------------------**

**Prerequisites:**

**STEP 0: Importing packages**

Use the scraping.py file for the web-scraping procedure on the https://news.tabularazor.org/ website.

This allows for retrieving all data used within the analysis, which should take a little over 1 hour.

**Returns:**

resultScraping.csv

**--- STEP 4: Running analysis.py ---------------------------------------------------------------------------------------------------------**

**Prerequisites:**

**STEP 4: Running scraping.py**

The analysis.py file is used for performing the anlaysis of the data retrieved from webscraping. When runn-

ing this file without having performed the webscraping first, it is important to have the resultScraping.csv

in the ResultCSV folder that is within the zip file as the the analysis.py calls upon this folder to retrieve

the csv script from. The Graphs resulting from the analysis are stored within the Graph folder that is also

already present within the Zip file.

**Returns:**

- Graphs

AuthorPublicationPlot.png

Boxplot.png

Heatmap.png

MaternityCheck.png

- ResultsCSV

Couples over years.csv

PublishingScheduleDays.csv

Down below a table is made which refers to the given feedback, whether it has been improved, and refers to the corresponding parts in the report:

|  |  |  |
| --- | --- | --- |
| Feedback for possible improvement | Improved? | Where |
| \* Very good report structure, only suggestion is to rename the section "Performing Web Scraping" into "Scraping and Dataset description" | Yes | In the title of Section 2.2 |
| \* There are some inconsistencies in writing style and grammar, some excessive passive voice. Figure 1 is mentioned but can't be found, you probably mean Table 1. | Yes | General grammar, spelling, and figure consistencies have been improved in the entire report. Also, extra attention was given to the excessive passive voice. |
| \* Explains well the site strucutre and the process of scraping, but could have given a bit more statistics about the gathered dataset. E.g. | Yes | Section 2.2 is now touching upon general statistics about Tabularazor and their employees. |
| \* Identifies relevant limitations of the analysis coming from the given data, but doesn't try and suggest some possible workarounds. | Yes | Section 5 now introduces suggestions and possible workarounds for the limitations of the analysis. |
| \* The correct couples are found but you don't talk explicitly about a couple's separation period while working at Tabularazor. | Yes | Section 3.2 now explicitly talks about a particular couple separation that has been observed. |
| \* Your expected number of holiday is in a reasonable range but a bit too low, probably because you discarded too many employees from the analysis. Only considering those who worked the full 7 years at the company is too restrictive... | No |  |