FIT5196 Task 1 - Assessment 1 (Development History)

Student Name: Alexandra Goh & Sothearith Tith

Student ID: 29796431 & 27208001

TASK 1

Date: 19/03/2024

Today, we met up to review the assignment specifications and discussed our approach to analyzing the first task's data. We also recognised that we needed to remove XML special characters from the text, such as converting "&amp" to "&", as well as specific patterns within our input data.

Contribution:

Alexandra:

- Explored the data using Notepad++ to understand its structure.
- Identified data patterns:
  - Observed that `<last-update-date>` records span from approximately the early 1990s to 2023.
  - Noticed that `<date-recorded>` records mostly range from the early 1990s to 2023, with a few exceptions dating even earlier (e.g., spotted records from 1955, 1963, 1977, 1980 etc.).
  - Recognised that "&amp" in XML is used to represent the character "&".

  - Identified special characters (accented letters), such as "SOCIÉTÉ," present within the `<person-or-organization-name>` field.
  - Some organization names end with S.A. and S.A.S., which represent types of corporations (e.g. simplified stock corporation) and are commonly used in many countries in Latin America.
  - Some organization names (especially German ones) have ING. contained within their names.
  - All dates are in the format "YYYYMMDD", with no spaces in between.

Sothearith:

- Proposed using Notepad++ for a closer examination of the data's structure
- Explored the data using Notepad++ to understand its structure.
- Identified data patterns:
  - Instances where USA states were labeled under the `<nationality>` attribute, while for other records, other countries like Canada, China, Germany, and Israel were labeled as "nationality."
  - Almost all of the `<state>` attributes correspond to USA states.

Date: 23/03/2024

We initiated the process of parsing through the data and extracting key attribute fields through Python. Using regular expressions to identify relevant data entries and process the content of the input file, we also ensured proper handling of special characters (e.g. `&amp`) and specific formats (e.g. `YYYY-MM-DD` for dates).

Contribution:

Alexandra:

- Facilitated data extraction and structuring processes by reading in the input data.
- Ensured careful handling of XML special characters, missing attribute elements and relevant data formats in parsing and extraction efforts.
- Prepared comprehensive documentation for both Tasks 1 and Tasks 2.

Sothearith:

- Contributed to data reading and structuring operations.
- Identified the most efficient method for data reading and loading operations (i.e. format and store extracted data in a dictionary).

Date: 01/04/2024

We focused on refining the extraction of the 'country' attribute from the data and ensured that each party's country was accurately determined, considering factors like nationality and state information. We also worked on incorporating the extraction of property counts by using regular expressions to find matches for the content within tags and storing them in a list of strings, followed by calculating the number of elements within the list itself.

Contribution:

Alexandra:

- Improved the method for inferring the country attribute, considering both nationality and state details to ensure accuracy in output.
- Maintained consistency and precision in determining country values, including handling cases where the 'nationality' attribute was labeled as "NOT PROVIDED" by appropriately inferring the country based on available state information.
- Conducted thorough testing to validate the updated extraction method and addressed any issues, such as printing unique values of the extracted 'country' attribute for both assignors and assignees to verify accuracy.
- Calculated property counts and double-checked accuracy with input file.

Date: 02/04/2024

Upon reviewing the output data, it came to our attention that the states and nationalities also included Canadian provinces. After discussing this matter in the forum and receiving confirmation from tutors, we realized the need to account for Canadian states in our data extraction and parsing methods as well. Subsequently, we made the necessary amendments to ensure that the extraction process accurately reflects the presence of Canadian states in the data.

Contribution:

Alexandra:

- Identified the existence of Canadian states within the dataset.
- Updated the extraction process to ensure that the attribute is appropriately stored as "Canada" when any Canadian province is encountered in assignment records.
- Sought confirmation regarding the treatment of British overseas territories and Crown dependencies.

Date: 06/04/2024

Today's focus was on extracting and organizing the trademark assignment entry data into a structured dictionary format. We tackled the crucial task of processing assignor and assignee information, as well as extracting unique identifiers and other main attributes from each assignment entry. This involved utilizing regular expressions to parse through the data effectively, ensuring accurate extraction of relevant information.

Contribution:

Alexandra:

- Implemented the functions for extracting assignor and assignee information from each assignment entry.
- Developed the logic for identifying unique identifiers and extracting other main attributes such as reel number, frame number, and last update date.
- Refined regular expressions and fine-tuned extraction methodologies to enhance accuracy and efficiency in data processing.

Date: 09/04/2024

After extracting and organizing all the trademark assignment entry data into a structured dictionary format, today marked the final step, which was saving the data into a JSON file.

Contribution:

Alexandra:

- Created and implemented the `save_to_json` function to facilitate the storage of structured data into a JSON file.
- Ensured the proper handling of encoding and formatting parameters, including UTF-8 encoding and indentation settings, to enhance the readability and compatibility of the JSON file.
- Tested and verified the functionality of the `save_to_json` function to ensure seamless execution.
- Completed documentation for Task 1 as well as updated references.

Google Colab Workbook Link

Task 1: https://colab.research.google.com/drive/15HZJdgqO7EETbS7c69833Cm4xRBJOP4T?usp=sharing