Sprint 1 - Html Parser – Research Documentation

What is the best way to parser the information?

1. BeautifulSoup: a library that can be installed and is specific to parsing HTML in Python
    a. The fact that it has to be installed is not ideal
    b. Includes various features that make parsing efficient and easy
2. HTML parser module: built in to Python, and provides a way to parse HTML in Python
3. RE: Regular Expression library, can used to find patterns of text, such as tags in HTML
    a. Video: https://www.youtube.com/watch?v=GEshegZzt3M discusses this

** Since the HTML parser module did not have to be installed and the amount of HTML that had to be parsed was minimal, it is the most ideal option

<u>HTML Parser in Python:</u>

1. About:
    a. html.parser.HTMLParser - provides a very simple and efficient way for coders to read through HTML code. This library is pre-installed in the stdlib.

    b. It provides a way to understand HTML code in a “nested fashion”.

    c. The module has methods that are automatically called when specific HTML elements are encountered. It also simplifies HTML tags and data identification.

    d. it is fed HTML code, and reads it tag by tag, from the beginning to end of the tag.

2. How to use:
    a. Note: HTMLparser doesn’t output any data, it just interprets the data in the tags fed to it

    b. Functionality must be added to methods in order to output the information that was found
        i. Subclassing/Inheritance

    c. By updating the default version of methods, through subclassing, these methods become more useful

    d. Create a parser class and defining the methods that are needed to parse but include what the output should be. This means the output will be whatever information we need in our processed file.

e. You can also feed the parser the URL of the website you would like to parse (not a feature necessary for this)

3. Important Methods:

→ HTMLParser.**handle_starttag(tag, attrs)** – Called when start tags are found *(example <html>, <head>, <body>)*

→ HTMLParser.**handle_endtag(tag)** – Called when end tags are found *(example <html/>, <head/>, <body/>)*

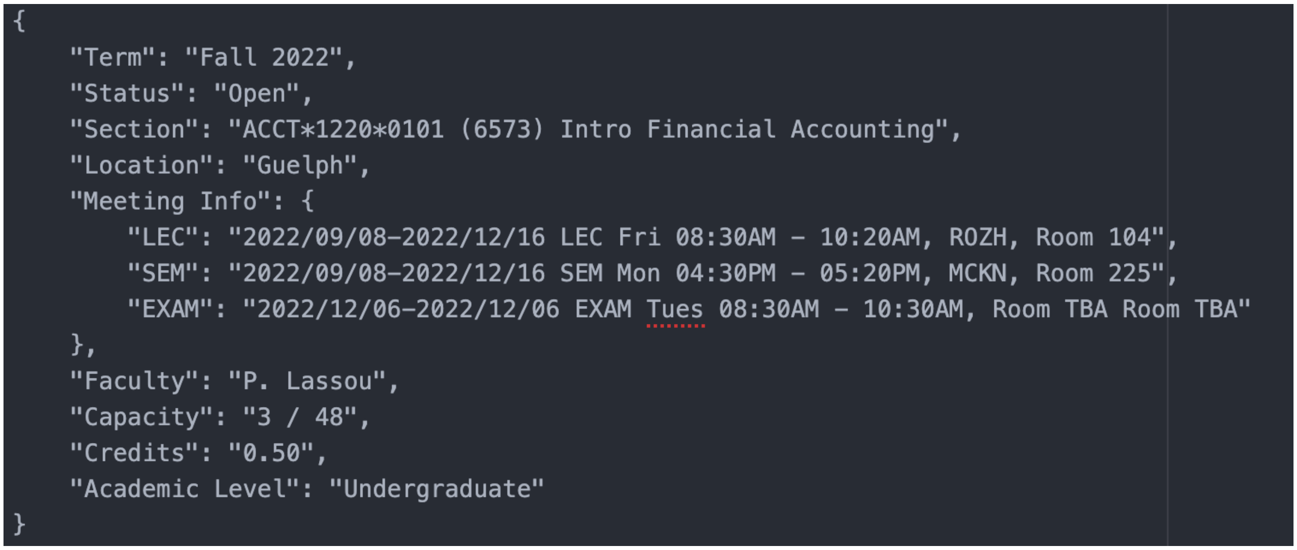
→ HTMLParser.**handle_data(data)** – Called when data is found *(example <a href =#> **data** </a>)*

→ HTMLParser.**handle_comment(data)** – Called when comments are found *(example <!–This is a comment–>)*

→ HTMLParser.**handle_decl(decl)** – Called when declarations are found *(example <!DOCTYPE html>*)

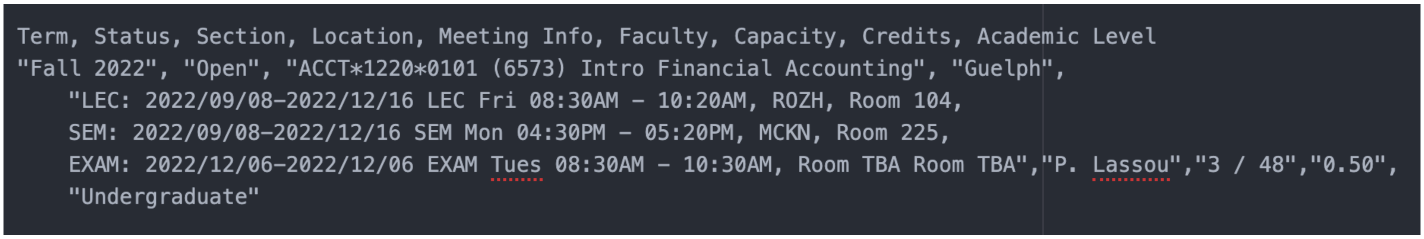
Which file type will the processed file be?

1. **JSON**
   a. Json is a module in Python that allows you to read/write/create JSON objects out of strings
   b. https://www.programiz.com/python-programming/json
   c. It would be possible to create a .json file containing all courses as JSON objects



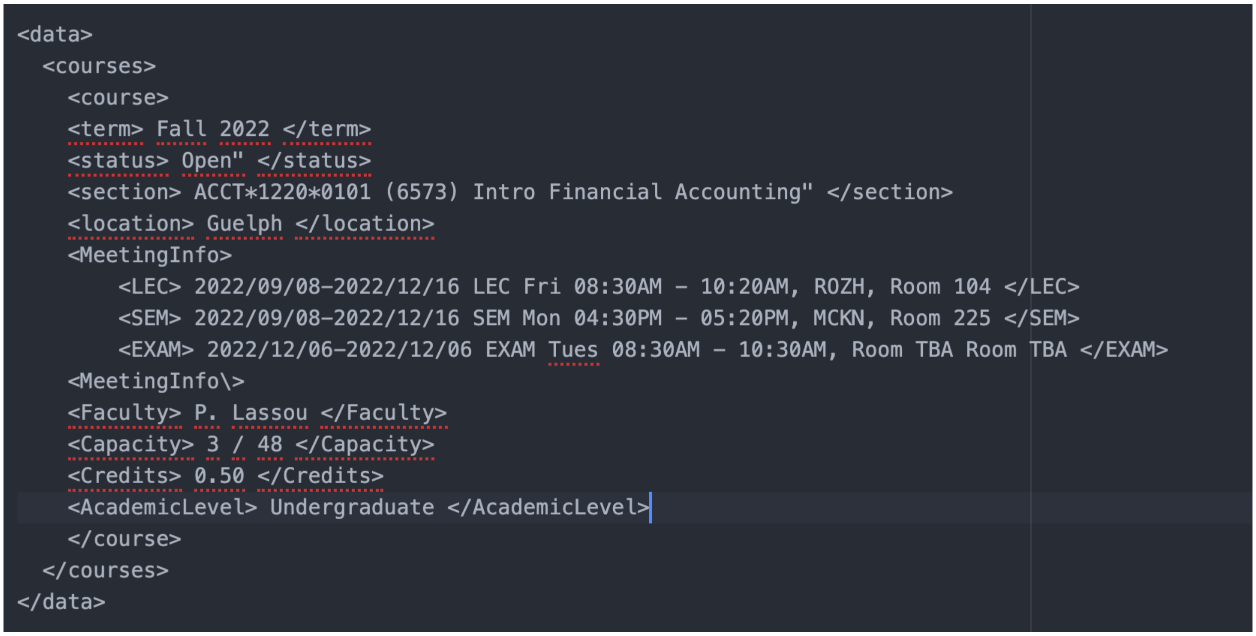
```
{
    "Term": "Fall 2022",
    "Status": "Open",
    "Section": "ACCT*1220*0101 (6573) Intro Financial Accounting",
    "Location": "Guelph",
    "Meeting Info": {
        "LEC": "2022/09/08-2022/12/16 LEC Fri 08:30AM - 10:20AM, ROZH, Room 104",
        "SEM": "2022/09/08-2022/12/16 SEM Mon 04:30PM - 05:20PM, MCKN, Room 225",
        "EXAM": "2022/12/06-2022/12/06 EXAM Tues 08:30AM - 10:30AM, Room TBA Room TBA"
    },
    "Faculty": "P. Lassou",
    "Capacity": "3 / 48",
    "Credits": "0.50",
    "Academic Level": "Undergraduate"
}
```

2. **csv**
    a. csv is another module in Python that allows you to read and write from JSON objects
    b. https://www.pythontutorial.net/python-basics/python-write-csv-file/
    c. It is possible to pull the information from the html file and write it to a csv file with headers to distinguish the data. The data would have to be separated by commas. It could it complicated when dealing with meeting info.



```
Term, Status, Section, Location, Meeting Info, Faculty, Capacity, Credits, Academic Level
"Fall 2022", "Open", "ACCT*1220*0101 (6573) Intro Financial Accounting", "Guelph",
    "LEC: 2022/09/08-2022/12/16 LEC Fri 08:30AM - 10:20AM, ROZH, Room 104,
    SEM: 2022/09/08-2022/12/16 SEM Mon 04:30PM - 05:20PM, MCKN, Room 225,
    EXAM: 2022/12/06-2022/12/06 EXAM Tues 08:30AM - 10:30AM, Room TBA Room TBA","P. Lassou","3 / 48","0.50",
    "Undergraduate"
```

3. **XML**
    a. ElementTree is a module in Python that allows you to parse and create XML files.
    b. https://docs.python.org/3/library/xml.etree.elementtree.html
    c. https://stackabuse.com/reading-and-writing-xml-files-in-python/
    d. After while parsing the tags, the data could be written to an XML file.



```
<data>
  <courses>
    <course>
    <term> Fall 2022 </term>
    <status> Open" </status>
    <section> ACCT*1220*0101 (6573) Intro Financial Accounting" </section>
    <location> Guelph </location>
    <MeetingInfo>
        <LEC> 2022/09/08-2022/12/16 LEC Fri 08:30AM - 10:20AM, ROZH, Room 104 </LEC>
        <SEM> 2022/09/08-2022/12/16 SEM Mon 04:30PM - 05:20PM, MCKN, Room 225 </SEM>
        <EXAM> 2022/12/06-2022/12/06 EXAM Tues 08:30AM - 10:30AM, Room TBA Room TBA </EXAM>
    <MeetingInfo\>
    <Faculty> P. Lassou </Faculty>
    <Capacity> 3 / 48 </Capacity>
    <Credits> 0.50 </Credits>
    <AcademicLevel> Undergraduate </AcademicLevel>
    </course>
  </courses>
</data>
```

Sources

https://www.askpython.com/python-modules/htmlparser-in-python
https://linuxhint.com/parsing-html-python/