Objective:

Traverse through the parent directory and all child subdirectories parsing all HTML files found and append them to an output file (preferably an XLSX file format, but a CSV file format would work too). The output file should have a header to identify the data in the associated column.

Java program Input parameters:

1. Supplier (static string passed to java program)
2. Input (Source) Directory. Parent folder, which may have subfolders
3. Output File (XLSX or CSV, with header)

Output file structure (column names):

1. SKU
2. UPC
3. vendor
4. name
5. short description
6. description
7. supplier
8. file date

Details:

How the file should be parsed:

The first <h2> tag is the contents for the “name” column. This is the product name used for WooCommerce.

The first <h3> tag is the contents for the “SKU” column. Within this tag the literal “SKU:” or “ID:” will always found and be associated to the “SKU” column.

The second <h3> tag is the contents for the "UPC" column. The literal “UPC:” will always be present.

The third <h3> tag is the contents for the "vendor" column. The literal “vendor:” will always be present

The next part of the parsing of the file might be little complicated so I will do my best to explain.

We should be including the html tags as part of the contents of the “description” and “short description” columns.

The first <p> tag can be assigned to one of two columns. It can be either the contents for "short description" or "description column”.

Rule:

If only ONE <p> tag found in the entire file, then this, including the tag and any remaining text are contents of the "description" column. In other words, we might not always have a “short description” for a given product.

If more the ONE <p> tag is found the FIRST <p> tag is the contents of the "short description" column and the remaining <p> tags and any remaining text are the contents of the "description" column.

Please note: Some of the HTML files do not have ending paragraph </p> tags while others do.

Current State (Existing java code):

The existing JAR file (classes) are working fine until multiple <p> tags are found. This is the part of the code which needs to be fixed. In addition, the current java program would need the header changed to reflect the column names mentioned above.

The current source code used “apache poi excel report” api and “html parser” api

Current way to invoke:

1. CD to the directory for the jar file location. Example: C:\....\\Inv\_ExcelReport
2. C:\....\\Inv\_ExcelReport>java -classpath .;excelReporter.jar;commons-logging-1.1.1.jar;jericho-html-3.2.jar;poi-3.9.jar com.excel.report.ExcelSheetWriter Royal c:\aquarium C:\outPut.xls