**Analysis of PPWS Course Data Structure**

The ppws\_courses.txt file includes information about various courses from the Plant Pathology, Physiology, and Weed Science (PPWS) department. The data follows a consistent format for each line, making it a good candidate for transforming into structured formats like CSV or XML. Each record consists of the course code, course title, and credit information. Below is a summary of my observations and analysis based on running a few command-line transformations.

### Observations on Data Structure

1. **Consistent Line Format**:
   * Each course entry follows the pattern: Course Code - Course Title (Credits). This makes it pretty easy to split out different fields.
   * Examples:
     + PPWS 2004 - Mysterious Mushrooms, Malicious Molds (3 credits)
     + PPWS 2964 - Field Study (1-19 credits)
2. **Field Breakdown**:
   * **Course Code**: Starts with PPWS followed by a four-digit number, giving each course a unique identifier.
   * **Course Title**: Comes right after the course code, separated by a dash (-). Titles are consistent in placement but vary in length.
   * **Credits**: Found in parentheses, either as a specific value (3 credits) or a range (1-19 credits).
3. **Command Outputs**:
   * **Uppercase Transformation** (tr command): This standardized the text to uppercase, which helps when eliminating case sensitivity issues.
   * **Removing Leading Whitespace** (sed command): There wasn’t much leading whitespace, but removing it keeps the dataset clean and consistent.
   * **Extracting the Third Word** (awk command): This highlighted the need for more precise separation of fields since the output could vary based on the course title’s length.

### Transformations for Normalization

1. **Converting to CSV**:
   * To get the data into CSV format, we need the following steps:
     + **Field Separation**: Split each line into individual fields. This can be done by using specific delimiters like the dash (-) and parentheses (()).
     + **Remove Extra Text**: Remove the word "credits" from the credit field, leaving only the number. For example, 3 credits should just be 3.
     + **Add Comma Delimiters**: Insert commas between fields to create a CSV structure, e.g.,

PPWS 2004, Mysterious Mushrooms, Malicious Molds, 3

* + **Handling Credit Ranges**: For courses with variable credits (e.g., 1-19 credits), we need a consistent way to represent these. It may make sense to add an extra column for the credit range.

1. **Converting to XML**:
   * To convert the data into XML, each course should be formatted as an XML element with sub-elements for each field:

<course>

<course\_code>PPWS 2004</course\_code>

<course\_title>Mysterious Mushrooms, Malicious Molds</course\_title>

<credits>3</credits>

</course>

* + **Field Tagging**: We need distinct tags for each field (e.g., <course\_code>, <course\_title>, <credits>).
  + **Handling Credit Ranges**: Courses with credit ranges should have a different representation, for example:

<credits min="1" max="19" />

1. **Data Clean-Up Considerations**:
   * **Irregular Titles**: Some course titles include punctuation like commas, which could complicate CSV conversion. These should be wrapped in quotes when creating the CSV.
   * **Optional Fields**: Some courses might have additional information, such as prerequisites or specific contact hours. These should be tagged appropriately if they’re included.

### Summary

The data in ppws\_courses.txt is well-structured for transformation into both CSV and XML formats but does need some modifications. The consistent course codes and credit format make automation feasible. The next step should involve scripting these transformations while making sure special characters, delimiters, and credit ranges are all handled correctly.