**CS 5800**

**Experiential Project**

**Cordiance**

**CS 5800 Project**

**Project Description**

In this project, you will help a company “cordiance” map categories between 2 files.

The first file, Avalara\_goods\_and\_services.xlsx, is a well-structured file of goods and services with somewhat uniform fields; whereas the second file, UNSPSC\_English.csv is hierarchical, but the fields are messy and non-uniform.

Your task is to map categories **from the reference file: Avalara\_goods\_and\_services.xlsx** to the file UNSPSC\_English.csv.

Look for out-right spelling errors and look for data truncation. Check to see if the spelling error is fixable by appending the category name with the non-matching portion of another category. Hint: enough must match to make it a good candidate.

Report the number of correct categories and which categories have spelling/data truncation issues and candidate fixes for each one.

Each of these files is structured as a forest or set of trees with each node representing a category and has several subcategories as children.

The parent node in the Avalara file is written in blue. You don’t need to map the nodes written in blue.

The trees in the UNSPSC\_English.csv are composed of 4 levels:

Segment (code) Segment Title Segment Definition: This is the parent node

Family (code) Family Title Family Definition

Class (code) Class Title Class Definition

Commodity (code) Commodity Title Commodity definition and Synonym

Sample of UNSPSC file

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Key | Segment | Segment Title | Segment Definition | Family | Family Title | Family Def | Class | Class Title | Class Definition | Commodity | Commodity Title | Definition |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 114670 | 50000000 | Food Beverage and Tobacco Products | This segment includes human food and beverages as well as condiments, colorings, flavorings and preservatives used in the preparation of food or beverages. This segment also includes plant extracts, mineral supplements and vitamins for human consumption. | 50180000 | Bread and bakery products |  | 50181900 | | Bread and biscuits and cookies |  | | 50181901 | | Fresh bread |  |  |
| 106263 | 50000000 | Food Beverage and Tobacco Products | This segment includes human food and beverages as well as condiments, colorings, flavorings and preservatives used in the preparation of food or beverages. This segment also includes plant extracts, mineral supplements and vitamins for human consumption. | 50180000 | Bread and bakery products |  | 50181900 | Bread and biscuits and cookies | | |  | | 50181905 | Sweet biscuits or cookies |  |

For example, after reading the Avalara….csv file, you come across Food & food ingredients -biscuits/cookies other than those covered or partly covered in chocolate. Now, you would walk the UNSPSC tree to the “Bread and Bakery products”  branch, then the “Bread and biscuits and cookies” branch, then the Sweet biscuits or cookies, which has a UNSPSC code of 50181905. So the **Alavara** branch of **biscuits/cookies** would get the **UNSPSC code of 50181905**.

Sample of Avalara file

|  |  |
| --- | --- |
| PF050067 | Food & food ingredients - ice cream - containers less than 200 ml (approximately 0.42 pint) |
| PF050068 | Food & food ingredients -ice cream - containers 200 ml (approximately 0.42 pint) or more |
| PF050069 | Food & food ingredients -potato snacks, unshelled roasted and/or salted nuts, and similar products |
| PF050070 | Food & food ingredients -fruit snacks, non-roasted nuts and similar products. |
| PF050071 | Food & food ingredients -biscuits/cookies other than those covered or partly covered in chocolate |

**Methodology**

You should meet with your groupmates as often as you can to decide on the plan or methodology that you will follow in designing your project. Divide your work equally among the members of your group.

Manually scan the data to decide on the approach that you should use to scan the two files.

For each non-blue node in the Avalara file, search the UNSPSC data (filename: UNSPSC\_English.csv) one tree at a time and try to match the Avalara node with a Commodity node in the UNSPSC.

You need to design the performance measures and the tests that should yield the best decisions to choose the best mapping.

After you complete your preliminary implementation, you can test it on the given files and submit your results in the form of a progress report. Most probably, additional updates and iterations will be needed to improve those results.

Note:

Not all UNSPSC categories will need to be mapped to the target tree categories.

You can write your project in the language that you want.

Your code should be as efficient as possible.

Your code should maximize the number of matching categories between the two files.

No interactions with other groups are allowed.

All group members should effectively contribute to the development of the project. A student’s grade depends on their work and contribution.

**Deliverables**

Weekly progress reports should be submitted describing your progress, and the contribution of each student.

An excel or csv file mapping categories from Avalara file (and corresponding code numbers) to UNSPSC codes (and corresponding Commodity title).

You should submit a final report describing your design methodology, the algorithm used, its complexity, and statistics on the output of your program.

More details on the final report and the final presentation can be provided later.

**Progress Report 1**

In this progress report, explain the methodology that you will follow in your project design.

Explain your plan for the next 4 weeks.

Explain the progress done so far on the project. **Let each member of the group write about their contribution to the project.**

Since this is a group project, all students should submit the same report.