FAI Competency Mapping Task

Similarity Score and Exact Search Matching

Scope of Work:

Utilizing the data provided by the client, create a sorted list of mapped course objectives relating to an individual competency. The use of various text analysis methods (Jaccard Distance and Cosine Similarity) to find the closest set of objectives to each competency will be used. Finally, the tool will utilize the given search terms provided by the client to ensure the objectives matched to each competency contain at least one of the search terms provided. The output returned to the user is a CSV file containing the objectives matched to each competency along with the rank, competency number, course and objective number, and similarity score.

Tasks:

The code utilized to perform this task is retrieved from a similar project; however, a number of changes are required:

1. Data Input and Cleaning

   Current code is not suitable for data sets with specific formatting; therefore, data will need to be cleaned and reformatted in order to function properly.

2. Similarity Score Matching and Ranking

   Current code searches a MongoDB and is capable of only one line item search per run. The current code should be manipulated to input two data sets: 1) objectives to match and 2) competency and exact search terms to match to the objectives.

3. Exact Term Search

   Once a list of similar objectives is calculated, the program will ensure at least one term in the search terms list can be found in each objective outputted.

4. Data Compilation and Cleaning

   Before returning the output to the user the program should compile all needed information before returning the output to the user.

5. Return Output

   Readable output returned to the user with associated information provided (rank, competency number, course and objective number, and similarity score).

Estimated Cost:

The estimated amount of time to complete the project is outlined below:

- 4 hours to update the current code in order to facilitate similarity and exact search matching simultaneously.
- 6 hours of data cleaning, compilation, and output. This value is dependent upon retrieving the initial dataset and identifying the changes needed for proper implementation.
- Estimated Cost: [redacted]