**ADDITIONAL LABELS**

get\_additional\_labels() Function give us additional dataset labels

The function take **extract\_results, existing\_labels,** asparametersand **topic\_words**, **stop\_words**, and **df\_thres**, which are optional parameters.

The function then checks whether a dataset label meets the following conditions:

1. It appears in at least **df\_thres** training data entries.
2. It contains at least one word from the **topic\_words** set.
3. It does not contain any words from the **stop\_words** set.

If a dataset label meets these conditions, and if its Jaccard similarity to any of the existing dataset labels in **existing\_labels** is less than 0.5, it is added to the **addtional\_labels** set.

The additional labels are obtained by comparing the cleaned existing labels (labels that have already been identified and cleaned) with the extracted labels from the text data. Any label that is not already in the existing set of labels is added to the additional set.

**existing\_labels** are set of all the unique dataset labels found in the training data after cleaning

**extract\_results** is a variable that holds the result of a JSON file that contains the extracted information from the documents in the dataset.

**Purpose**

The purpose of the additional labels in these codes is to expand the set of labels for the Coleridge Initiative Show US the Data competition.

This helps to improve the performance of machine learning models by providing more diverse training data.