**Exploratory Data Analysis (EDA) Summary**

The unit of analysis in this dataset is individual comments in documents from CommonCrawl docx files, spanning the period from 2013 to 2020. Each comment has been labeled hierarchically with categories and subcategories based on the main intent, utilizing fields such as level\_0, level\_1, level\_2, level\_3, and level\_4. The dataset includes various contextual fields like file\_id, comment\_date, anonymized\_nickname, document\_paragraph\_text, document\_selected\_text, and more.

Summary of the Data Set:

Total Observations: The dataset comprises comments from CommonCrawl docx files, with a total of 5200 observations.

Unique Observations: There are 5200 unique comments in the dataset.

Time Period Covered: The data spans from 2013 to 2020.

Data Cleaning Steps:

Exclusion of Irrelevant Observations:

Any observations that were deemed irrelevant or outside the scope of the study were excluded.

Handling Missing Values: Steps were taken to address missing values in relevant fields, ensuring the integrity of the dataset.

Outcome Description:

Hierarchical Classification Problem: The primary outcome is a hierarchical classification of comments, with levels of categorization denoted by fields level\_0, level\_1, level\_2, level\_3, and level\_4.

Visualization Technique: A tree diagram or hierarchical chart can effectively represent the outcome structure, showcasing the levels and their relationships.

Key Predictors Description:

Selection of Predictors: For this summary, we focus on predictors deemed most important, guided by domain knowledge or preliminary EDA.

1.Comment Length:

Predictor: comment\_length

Description: The length of a comment, measured in terms of the number of characters or words, can be a key predictor. Longer comments might indicate more detailed or complex thoughts, potentially influencing the hierarchical classification.

2.Hierarchical Classification

level\_0: Main category  
level\_1: First level category  
level\_2: Second level category  
level\_3: Third level category  
level\_4: Fourth level category

Visualization Techniques: Bar charts or pie charts can be employed to visualize the distribution of comments across different levels, providing insights into the importance and prevalence of each category.

Domain-specific Insights:

Utilizing Levels for Interpretation: Levels such as level\_0 to level\_4 provide a structured way to interpret the main intent of comments. Exploring the distribution of comments across these levels can uncover patterns and trends.

Language Inference: The inferred language -English- of the comment text can be explored to understand the linguistic diversity within the dataset.