In this clustering analysis, we aim to identify inherent patterns within the dataset of comments from CommonCrawl docx files. The hierarchical classification of comments into levels (level\_0 to level\_4) serves as the primary outcome, and we will explore how key predictors, such as comment length, contribute to the clustering results.

**1. Data Preprocessing:**

* **Exclusion of Irrelevant Observations:** Any comments deemed irrelevant or outside the study's scope were excluded.
* **Handling Missing Values:** Steps were taken to address missing values in relevant fields to ensure dataset integrity.

**2. Key Predictors:**

**a. Comment Length:**

* Predictor: comment\_length
* Description: Length of comments, measured in characters or words, is considered a key predictor, potentially influencing hierarchical classification.

**3. Clustering Model:**

* **Model Choice:** K-Means clustering was chosen as it efficiently identifies natural groupings in the data based on similarities.

**4. Visualizations:**

**a. Scatter Plot for Comment Length:**

* Create a scatter plot of comment length against the hierarchical classification levels.
* Color code points based on their assigned clusters to observe if there's any relationship between comment length and clusters.

**b. Silhouette Plot:**

* Generate a silhouette plot for different cluster numbers to assess the quality of clustering.
* Observe how well-defined the clusters are and choose the optimal number of clusters (K).

**c. Cluster Centers Visualization:**

* Plot the cluster centers on the scatter plot to visualize the central points of each cluster.
* Explore how comment length varies within each cluster.

**d. Hierarchical Tree Diagram:**

* Construct a hierarchical tree diagram or chart to represent the outcome structure, showcasing the relationships between levels.

**e. Bar or Pie Charts for Level Distribution:**

* Use bar charts or pie charts to visualize the distribution of comments across different hierarchical levels.
* Understand the prevalence and importance of each category.

**5. Domain-specific Insights:**

* **Interpretation using Levels:** Explore the distribution of comments across levels (level\_0 to level\_4) to uncover patterns and trends in the main intent.
* **Language Inference:** Consider exploring inferred language (English) to understand linguistic diversity within the dataset.

**6. Model Evaluation:**

* Assess the silhouette score to gauge the quality of clustering and determine the optimal number of clusters for meaningful interpretation.

This clustering analysis aims to provide a comprehensive understanding of the inherent structures within the dataset, emphasizing the relationship between comment length and hierarchical categorization levels. Visualizations will aid in interpreting the results and gaining insights for further analysis.

A line graph with blue dots

Description automatically generated

A graph of a bar chart

Description automatically generated with medium confidence

A white graph with green and blue dots

Description automatically generatedA diagram with green and blue dots

Description automatically generated