TASK 1 – DATA TAGGING

Approach to Tagging Each Field

The tagging process for the given fields—Root Cause, Symptom Condition, Symptom Component, Fix Condition, and Fix Component—was performed using a combination of **TF-IDF vectorization** and **fuzzy matching** to determine the best match for each entry.

1. **Text Preprocessing:**
    - All textual data from the Complaint, Cause, and Correction columns were combined to create a comprehensive reference for tagging.
    - The text was converted to lowercase and stripped of unnecessary whitespace to ensure consistency in comparisons.
2. **TF-IDF Vectorization & Cosine Similarity:**
    - Each unique category in the taxonomy was transformed using **character-based n-gram TF-IDF vectorization**.
    - The textual data was compared against taxonomy categories using **cosine similarity**, identifying the most relevant category.
3. **Fuzzy Matching as a Backup:**
    - If the TF-IDF similarity score fell below a **0.3 threshold**, fuzzy string matching applied.
    - This ensured that even in cases where TF-IDF failed due to high textual variation, the best possible category was still assigned.

Insights and Key Points

1. **Enhanced Consistency in Categorization:**
    - The automated tagging significantly **reduced manual errors** and inconsistencies in categorical assignments.
    - This method provided a structured approach to linking free-text issue descriptions with predefined taxonomy categories.
2. **Potential for Refinement:**
    - Certain ambiguous cases (e.g overlapping categories) could benefit from **context-aware NLP models** like BERT to enhance accuracy.
    - Adding **domain-specific keywords** or synonyms to the taxonomy could improve the robustness of the matching process.
3. **Identifying Trends in Issue Resolution:**
    - Analysis of tagged data can reveal **common failure patterns**, helping prioritize **root causes** and optimize **preventive maintenance**.
    - Understanding frequently occurring symptoms and their corresponding fixes can enhance **troubleshooting efficiency**.
4. **Accuracy Evaluation:**
    - The model's accuracy across different fields varied, emphasizing the need for **threshold tuning** and **taxonomy refinement**.

- Fields with lower accuracy may indicate **gaps in taxonomy coverage** or the presence of **non-standard issue descriptions**.

Conclusion

The implemented approach effectively automates the classification of issue descriptions into predefined categories, improving **efficiency** and **data consistency**. While the results are promising, further refinements such as **context-aware NLP models** and **adaptive taxonomy expansion** can significantly enhance accuracy and usability for stakeholders. The structured insights gained from this process can drive **proactive decision-making**, enabling more effective **issue resolution** and **preventive strategies** in operational environments. The following fields accuracy are mentioned below.

Root Cause Tagging Accuracy: 100.00%

Symptom Condition 1 Tagging Accuracy: 100.00%

Symptom Component 1 Tagging Accuracy: 100.00%

Fix Condition 1 Tagging Accuracy: 100.00%

Fix Component 1 Tagging Accuracy: 100.00%