Documentation: Identifying and Marking Problematic Sentences in the Dataset

Objective:

To identify records in a dataset that have:

1. Spelling errors.
2. Incomplete sentences (either through lack of typical ending punctuation or truncated words).

Tools Used:

1. **Python** programming language.
2. **Pandas**: A popular data analysis library for Python.
3. **SpaCy**: A library for Natural Language Processing.
4. **PySpellChecker**: A pure Python spell checking library.

Procedure:

1. Setup:

   - Loaded the necessary Python libraries (pandas, spacy, and pyspellchecker).
   - Initialized the SpaCy model for English language processing.
   - Initialized the PySpellChecker for spelling verification.

2. Detecting Spelling Errors:

   - Used the unknown() method from PySpellChecker to identify words in the text that aren't recognized by its dictionary.

   - Created a function has_spelling_errors that:

     - Splits the given text into individual words.
     - Uses the unknown() method to check for misspelled words.
     - Returns True if there are misspelled words, otherwise False.

3. Detecting Incomplete Sentences:

   - Processed the given text with the SpaCy NLP model.
   - Checked the last token of the processed text. If it is not one of the typical sentence-ending punctuation marks (".", "?", "!"), marked the sentence as potentially incomplete.

- Further checked if the last word of the sentence is recognized by PySpellChecker. If not, it might indicate a word cut-off, marking the sentence as incomplete.
- Created a function is_incomplete_sentence that implements the above steps and returns True if the sentence is deemed incomplete, otherwise False.

4. Applying Checks to the Dataset:

- Loaded the dataset into a Pandas DataFrame.

- Applied the has_spelling_errors function to the Text column of the DataFrame to identify records with spelling errors.

- Applied the is_incomplete_sentence function to the Text column to identify records with incomplete sentences.

- Marked records with two new columns: Has_Spelling_Errors and Is_Incomplete.

5. Saving the Results:

- Filtered the DataFrame to only include records that have either spelling errors or are incomplete.

- Saved this filtered dataset to a new CSV file named data_with_errors.csv.

Result:

The resulting data_with_errors.csv file contains all the records from the original dataset that were identified to have spelling errors or incomplete sentences.