Writing assistance documentation

Created	@November 4, 2024 10:21 PM
Class	JetBrains

Project Documentation

Datasets:

1. **Synthetic Misspelled Dataset**:
    - This dataset is generated programmatically by introducing common spelling errors into a base clean dataset. It includes intentional spelling mistakes that are representative of typical errors found in real-world scenarios.
    - **Purpose**: The synthetic misspelled dataset serves as the input for testing the spell-checking tools and models. It ensures a controlled environment where the effectiveness of each spell checker can be assessed reliably against known errors.
2. **Clean Dataset**:
    - This dataset consists of a large corpus of correctly spelled English words or phrases. It acts as the “ground truth,” providing the correct spelling for each word or phrase found in the synthetic misspelled dataset.
    - **Source**: http://mattmahoney.net/dc/text8.zip
    - **Purpose**: The clean dataset serves as the reference for evaluating the accuracy and performance of each spell-checking model.
3. **Misspelled Dataset**:
    - This dataset consists of commonly misspelled words paired with their correct versions. The dataset allows for direct assessment of spell-checking models against realistic, known spelling mistakes.

- Source: https://norvig.com/ngrams/spell-errors.txt
- **Purpose**: Provides a controlled benchmark for evaluating the effectiveness of each spell checker by offering a set of known spelling errors to be corrected.

Tools:

1. **Spell-Checking Libraries and Models**:
   - **PySpellChecker**:
     - A Python library that provides basic dictionary-based spell-checking capabilities.
     - **Strength**: Fast, lightweight, and simple to use, though limited in context-aware spell-checking.
   - **TextBlob**:
     - A natural language processing (NLP) library that includes spell-checking as part of its broader text processing toolkit.
     - **Strength**: Offers basic corrections and works well with simple typos, although it does not consider context as effectively as advanced models.
   - **Fine-Tuned Transformer Model**:
     - This model uses the `pszemraj/grammar-synthesis-small` transformer, fine-tuned for text correction and grammar synthesis tasks.
     - **Strength**: A powerful model for grammar correction and contextual spelling correction, though it requires more computational resources.
   - **GPT-3.5**:
     - A large language model developed by OpenAI, specifically used for its ability to understand and generate contextually appropriate text.
     - **Strength**: Highly context-aware and capable of understanding complex language patterns, making it effective for spell-checking in sentences or longer text.
2. **Libraries for Model Evaluation**:

- **Scikit-Learn**: Utilized for calculating performance metrics such as precision, recall, and F1 score.
- **Levenshtein**: A library for calculating the Levenshtein distance, which measures the similarity between two strings, useful in assessing minor differences between the correct word and the model output.

Metrics:

1. **Accuracy**:
   - **Definition**: The proportion of corrected words that exactly match the expected correct words.
   - **Purpose**: Provides a straightforward measure of how often each spell checker successfully corrects misspelled words.
2. **Levenshtein Distance**:
   - **Definition**: A string metric for measuring the number of single-character edits (insertions, deletions, or substitutions) required to change one word into another.
   - **Purpose**: Offers insight into the “closeness” of a correction when the spell checker fails to make an exact match, indicating the level of similarity between predicted and actual words.
3. **Precision**:
   - **Definition**: The proportion of true positive corrections out of all predicted corrections, indicating how many predicted corrections were actually correct.
   - **Purpose**: Helps measure the reliability of each spell checker, focusing on how many of its corrections were correct.
4. **Recall**:
   - **Definition**: The proportion of true positive corrections out of all actual correct words, showing how well the spell checker identifies correct options.
   - **Purpose**: Evaluates how often the spell checker successfully finds the correct correction among all possible errors.
5. **F1 Score**:

- **Definition**: The harmonic mean of precision and recall, balancing the two metrics to give a combined measure of accuracy.
- **Purpose**: Provides a more balanced view of performance, especially useful when there is a significant trade-off between precision and recall.

Each of these metrics provides a unique insight into the spell checkers' performances and helps identify the strengths and weaknesses of each approach.