UK to US Dialect Converter

Overview

This notebook performs the conversion of UK English text to US English. It uses a combination of rule-based text processing and a pre-trained transformer model to handle the dialect conversion. The goal is to convert words and phrases like "theatre" to "theater" and "favourite" to "favorite."

How to Run the Notebook

1. **Install Dependencies**: Install the required libraries using the following command:

!pip install transformers torch emoji scikit-learn pandas

2. **Run the Cells**: After installing the dependencies, execute each cell in the notebook. The notebook will:
    - Load the model.
    - Process input sentences.
    - Display the converted output.
3. **Test Cases**: The notebook includes several example sentences that demonstrate how UK English is converted into US English. Simply modify or add your own sentences to see the conversion.

Dependencies

The following Python packages are required:

- **transformers**: For loading and utilizing the pre-trained T5 model.
- **torch**: PyTorch is required to run the transformer model.
- **emoji**: For handling emoji characters and converting them to placeholders.
- **scikit-learn**: Required for any data preprocessing (currently unused but can be helpful for future improvements).
- **pandas**: For managing and manipulating the input text data.

To install all dependencies, run:

!pip install transformers torch emoji scikit-learn pandas

Key Findings

- **Dialect Conversion**: The rule-based approach works for many common UK to US English conversions (e.g., "colour" to "color"), but the model might generate redundant or imperfect results for certain phrases.
- **Model Behavior**: The pre-trained T5 model struggles with consistent and accurate dialect conversion. It sometimes outputs unnecessary translations like "Translate the following UK English to US English."
- **Handling Emojis**: Emojis are effectively converted to text placeholders like <emoji>:) </emoji>, which can be further refined for better representation.

Potential Improvements

- **Expand UK to US Mappings**: Currently, the dictionary contains only a small set of word mappings. Expanding this list will improve the rule-based conversion.
- **Refine Model Output**: The T5 model's translations need improvements to make the output more accurate and less verbose. Trying different models like GPT-3 or fine-tuning the T5 model on specific UK-to-US text would help.
- **Emoji Handling**: Instead of converting emojis to placeholders, consider replacing them with more contextually appropriate words or symbols.
- **Extend Dialect Support**: Add functionality to handle more dialects beyond UK-to-US conversion, such as Australian or Canadian English.

Known Limitations

- **Model Output**: The model may generate redundant phrases or incorrect translations for certain sentences.
- **Limited Mappings**: The UK-to-US word mapping is incomplete, and there are many words that might not be covered.
- **Emoji Conversion**: Emojis are currently replaced with placeholder text. Future versions could try to incorporate a more sophisticated handling of emojis.

Time-Saving Tips

- **Use Rule-Based Conversion**: If time is tight and you don't need the model's complexity, use just the rule-based UK-to-US word conversion.
- **Test with Small Input**: Use a small batch of test sentences to quickly see how the system works, especially for common words.