PDF Spell Checker

A Streamlit application designed to identify potential spelling errors within PDF documents by comparing words against customizable dictionary files. Users can upload a PDF, select dictionaries, view potential misspellings, and export the results either as a CSV file or as a new PDF with the selected words highlighted.

Features

- **PDF Upload:** Allows users to upload PDF files directly through the web interface.
- **Custom Dictionaries:** Uses `.txt` dictionary files located in a `default_dict` subdirectory. Multiple dictionaries can be selected and combined.
- **Word Cleaning:** Implements specific rules to clean words extracted from the PDF before checking against the dictionary (converts to uppercase, removes non-alphanumeric characters except hyphens, strips leading/trailing hyphens, ignores words containing digits).
- **Interactive Results:** Displays potential misspellings in an interactive table showing the original word, the number of instances, and the pages where it appears.
- **Selective Export:** Users can select which identified words they want to include in the output.
- **CSV Export:** Exports the selected misspelled words, their instance count, and page numbers to a CSV file.
- **Highlighted PDF Generation:** Creates a new PDF file where the selected misspelled words are highlighted with a semi-transparent yellow background.
- **Caching:** Uses Streamlit's caching (`st.cache_data`) to speed up dictionary loading.

Requirements

- Python >= 3.13
- pymupdf >= 1.25.3
- streamlit >= 1.42.2
- pandas (used for displaying results in `st.data_editor`)

Usage

1. **Clone the repository or download the script.**
2. **Double Click `Launch_SpellCheck.bat`**
3. **Upload PDF:** Use the file uploader in the web interface to select the PDF you want to check.
4. **Select Dictionaries:** Choose the dictionary files you want to use from the multiselect dropdown. The application will load and combine words from the selected files.
5. **Review Results:** Once processing is complete, a table will display potential misspellings.
6. **Select Words:** Check the "Include?" box next to the words you want to export or highlight.
7. **Export/Highlight:**
    - Click "Export to CSV" to download a CSV file containing the selected words.
    - Click "Generate Highlighted PDF" to download a new PDF with the selected words highlighted.

Project Structure

```
├── SpellCheck.py           # Main Streamlit application
├── CreateDict.py           # Dictionary creation utility
├── Launch_SpellCheck.bat   # Windows launcher script
├── Launch_CreateDict.bat   # Dictionary creator launcher
├── utils/                  # Utility modules
│   ├── dictionary.py       # Dictionary loading and processing
│   ├── html.py            # HTML and CSS templates
│   └── pdf.py             # PDF processing utilities
├── default_dict/           # Dictionary files
│   └── *.txt              # Dictionary text files
└── README.md              # This file
```

Dictionary Files

- Dictionary files must be plain text (`.txt`).
- Each word should be on a new line.
- Place these files in a subdirectory named `default_dict` located in the same directory as the script.
- The application reads these files, converts words to uppercase, and stores them in a set for efficient lookup.

Output Formats

1. **CSV File (`misspelled_words.csv`):**
   - Columns: `Word`, `Occurrences`, `Pages`
   - Contains only the words selected by the user in the interactive table.
2. **Highlighted PDF (`highlighted.pdf`):**
   - A copy of the original PDF.
   - Instances of the selected misspelled words are marked with a semi-transparent yellow highlight.

Contact


```
Arun Kishore
Structural E.I.T.
Associated Engineering (B.C.) Ltd.
#500 - 2889 East 12th Avenue, Vancouver, BC   V5M 4T5
Dir: 236.317.2201 | email: remaa@ae.ca
```

> **Version**: 1.0.0 (Updated: 06/11/2025)