AI-Ready Document Splitter

Project Overview

Chunkify is a user-friendly tool developed with Streamlit that allows users to break down large text or PDF documents into smaller, manageable sections called chunks. This chunking process is crucial in preparing documents for Natural Language Processing (NLP), Machine Learning, and AI model training, where smaller and contextually linked segments yield better performance.

The tool automates the process of document splitting, reducing manual preprocessing efforts and making the data ready for tasks like embedding generation, semantic search, summarization, or chatbot integration.

How It Works

1. File Upload
   Users upload a document in either .txt or .pdf format. The file is temporarily stored using the tempfile module.
2. Content Extraction
   Based on the file type, the app uses LangChain's TextLoader or PyPDFLoader to extract text while preserving the structure and encoding.
3. Text Splitting
   The RecursiveCharacterTextSplitter algorithm divides the document into chunks, using a user-defined chunk size and overlap. This ensures that each chunk maintains context with its neighboring sections.
4. User Interface and Feedback
   The application offers real-time progress tracking with progress bars and informative messages throughout the chunking process.
5. Output and Export
   Once processed, the chunked data is displayed in a preview table. Users can download the results in either CSV or JSON format for further use.
6. Temporary File Cleanup
   To maintain privacy and reduce memory usage, all temporary files are automatically deleted after processing.

Key Features

- Supports both PDF and plain text files
- Customizable chunk size and overlap for precise control
- Clean and responsive interface with instant visual feedback
- Preview of processed chunks before exporting
- Downloadable results in CSV or JSON formats
- Lightweight design suitable for both local and cloud deployment
- Ensures user privacy through temporary file handling
- Designed to fit seamlessly into AI/NLP pipelines

Technologies Used

- Python for core application logic
- Streamlit for building the web interface
- LangChain for document loading and splitting
- Pandas for managing data and export functions
- ReportLab for optional PDF documentation
- Tempfile and OS modules for managing file storage and cleanup
- Time module for simulating loading animations

How to Use

1. Install required libraries using pip:
   pip install streamlit langchain_community pandas reportlab pypdf
2. Save the code as app.py
3. Run the app:
   streamlit run app.py
4. Use the web interface to:
   - Upload a .txt or .pdf file
   - Adjust the chunk size and overlap settings
   - Track progress of the chunking process
   - Preview the initial chunks
   - Download the results in CSV or JSON format
5. Use the downloaded chunks in your NLP or AI applications.

Real-World Applications

- Preparing documents for semantic search engines

- Splitting eBooks for AI-based summarization
- Generating training datasets for chatbots
- Preprocessing manuals for question-answering systems
- Creating embeddings from research papers
- Segmenting transcripts for audio/video summarization tools

Sample Output

For a cloud computing training document processed with a chunk size of 1000 and an overlap of 200:

- Total Chunks: 36
- Average Chunk Length: 795.86 characters
- Shortest Chunk Length: 192 characters
- Longest Chunk Length: 998 characters

Conclusion

Chunkify is a powerful and efficient solution for anyone working with large volumes of text. Its customizable settings, lightweight architecture, and AI-ready output format make it an ideal preprocessing tool for data scientists, researchers, and developers involved in NLP and AI model training.