■ Chunkify – AI-Ready Document Splitter

Project Overview:
Chunkify is a specialized tool built using **Streamlit** that enables users to convert large text or PDF files into smaller, manageable sections called *chunks*. These chunks are essential when processing large documents for Natural Language Processing (NLP), Machine Learning, or AI model training, as many models perform better with smaller, contextually connected segments of text.

By automating the process of splitting and structuring data, Chunkify reduces manual preprocessing time and ensures data is ready for downstream tasks such as embeddings creation, semantic search, summarization, or chatbot integration.

How It Works:
1. **Upload Phase:** Users upload a document in either .txt or .pdf format. The app uses *tempfile* to temporarily store the file.
2. **Document Loading:** Based on file type, LangChain's *TextLoader* or *PyPDFLoader* extracts the content while preserving structure and text encoding.
3. **Splitting Mechanism:** The *RecursiveCharacterTextSplitter* algorithm breaks the text into chunks based on a user-defined chunk size (number of characters) and overlap (to maintain context between chunks).
4. **Visualization:** Progress bars and informational messages guide the user through the process in real-time.
5. **Export Options:** The chunked data is presented in a table and can be downloaded in CSV or JSON format for integration into other systems.
6. **Cleanup:** Temporary files are deleted after processing to optimize memory usage and maintain privacy.

Features:
- Multi-format support: Works with both text and PDF documents.
- Adjustable chunking logic: Customize chunk size and overlap for specific NLP requirements.
- Interactive UI: Built with Streamlit for instant feedback and progress tracking.
- Data Preview: Review the first few chunks before downloading.
- Dual download formats: Export chunked data as CSV or JSON.
- Lightweight & Portable: Can run locally or be deployed on the cloud.
- Privacy-friendly: Temporary file storage ensures no persistent data is left behind after processing.
- Optimized for AI workflows: Perfect for preparing data for LLMs like GPT, BERT, or other AI/ML models.

Technologies Used:
- **Python**: Core programming language for logic and integrations.
- **Streamlit**: Framework for creating the interactive web application.
- **LangChain**: Provides document loading and text splitting utilities.
- **Pandas**: Handles tabular data, CSV/JSON export, and chunk previews.
- **ReportLab**: Generates PDF reports and documentation.
- **Tempfile** & **OS**: Manage temporary file storage and cleanup.
- **Time**: Adds simulated delays for smooth progress bar animations.

Usage Instructions:
1. Install dependencies:

```
pip install streamlit langchain_community pandas reportlab pypdf
```

2. Save the Streamlit code as app.py.
3. Run the application:

```
streamlit run app.py
```

4. In the browser interface:
- Upload your document (.txt or .pdf).
- Adjust chunk size and overlap sliders.
- Monitor progress until the chunking process completes.
- Preview initial chunks.
- Download full chunks as CSV or JSON.

5. Use the downloaded chunks in your NLP/AI pipeline.

Real-World Use Cases:
- Preparing large documents for semantic search systems.
- Splitting eBooks into smaller sections for AI summarization.
- Creating training datasets for chatbots.
- Segmenting technical manuals for question-answering systems.
- Preprocessing research papers for embedding generation.
- Chunking transcripts for video/audio summarization AI tools.

Example Output (Cloud Computing Training Document):
Processed with chunk size 1000 and overlap 200:
- Total Chunks: 36
- Average Chunk Length: 795.86 characters
- Shortest Chunk Length: 192 characters
- Longest Chunk Length: 998 characters

Conclusion:
Chunkify simplifies and accelerates the document preprocessing stage for AI and NLP projects. Its flexibility, user-friendliness, and output versatility make it an indispensable tool for developers, data scientists, and researchers who work with large volumes of unstructured text data.