

# Document Splitters and Chunkers

## Instructor

Dipanjan Sarkar

Head of Community & Principal AI Scientist at Analytics Vidhya

Google Developer Expert - ML & Cloud Champion Innovator

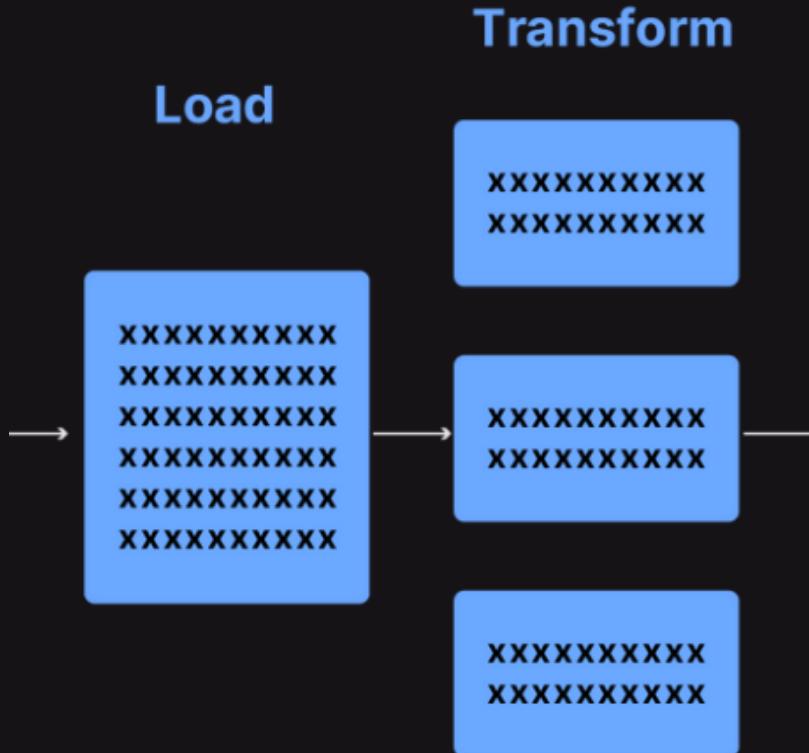
Published Author



# Outline

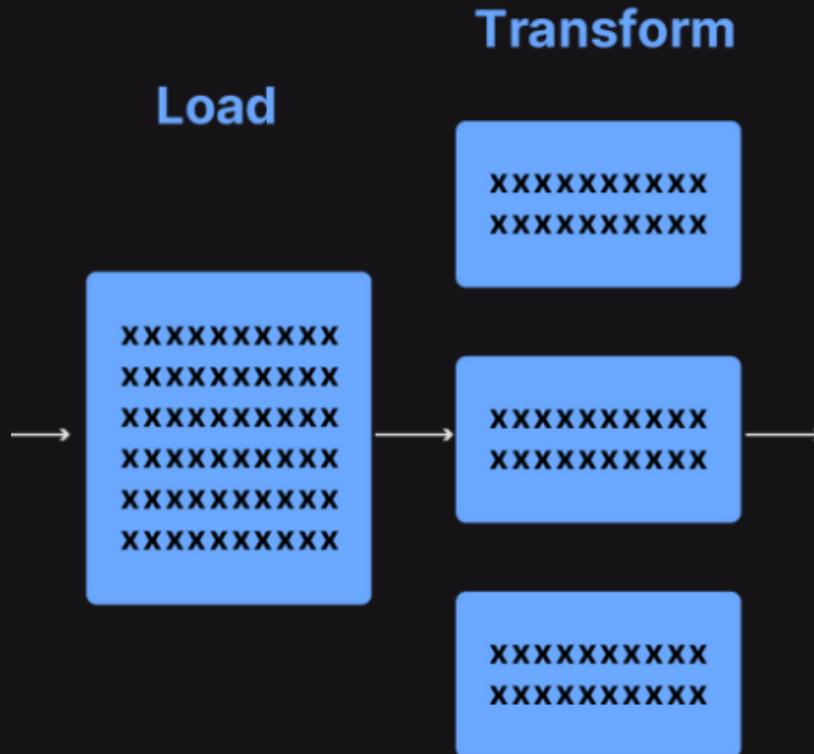
- Document Splitters and Chunkers
- How does a Splitter work?
- Popular Built-in Document Splitters in LangChain

# Document Splitters and Chunkers



- LangChain supports various document splitting and chunking mechanisms for transforming documents
- Splitting documents into smaller chunks or paragraphs enables them to fit into your LLM's context window
- Documents can be split based on various methods such:
  - sections
  - character count
  - token counts

# How does a Splitter work?



- At a high level, text splitters work as follows:
  - Split the text into small, semantically meaningful chunks
  - Combine these small chunks into a larger chunk until you reach a certain size based on character count or token count
  - Once you reach that size, make that chunk its own piece of text and then start creating a new chunk of text with or without chunk overlaps
- Key aspects of splitting:
  - Splitting strategy - Character, Tokens, Semantic, Sectional etc.
  - Chunk size measurement - Character or Token counts

# Popular Built-in Document Splitters in LangChain

Splitter Type	Description
<a href="#"><u>RecursiveCharacterTextSplitter</u></a>	Recursively splits text into larger chunks based on several defined characters. Tries to keep related pieces of text next to each other. LangChain's recommended way to start splitting text
<a href="#"><u>CharacterTextSplitter</u></a>	Splits text based on a user defined character. One of the simpler text splitters
<a href="#"><u>tiktoken</u></a>	Splits text based on tokens using trained LLM tokenizers like GPT-4
<a href="#"><u>spaCy</u></a>	Splits text using the tokenizer from the popular NLP library - spaCy
<a href="#"><u>SentenceTransformers</u></a>	Splits text based on tokens using trained open LLM tokenizers available from the popular sentence-transformers library
<a href="#"><u>unstructured.io</u></a>	The unstructured library allows various splitting and chunking strategies including splitting text based on key sections and titles

# Thank You

---