

O'REILLY®

Data Chunking

Bert Gollnick





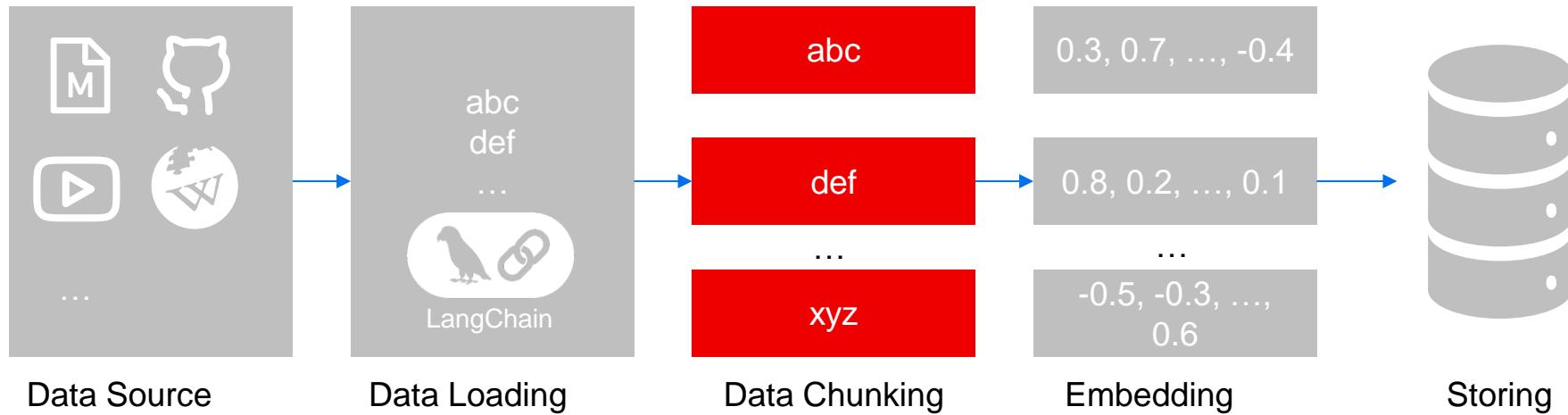
Learning Objectives

By the end of this module, you will:

- Know how to split the data
- Why it is important to split the data
- Know what tokens and sequence lengths are



Data Ingestion Process



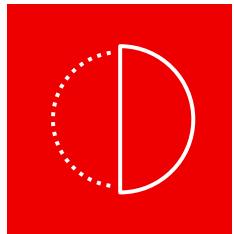


What is „Chunking“?

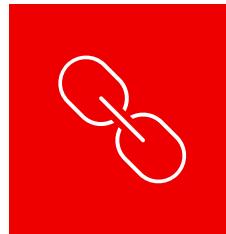
- Dividing larger pieces of information into smaller, manageable units
- These units called „chunks“
- Required to fit model context window
- Chunks should be:
 - Small
 - Semantically meaningful



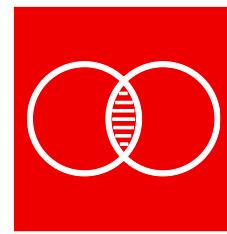
Process



Split



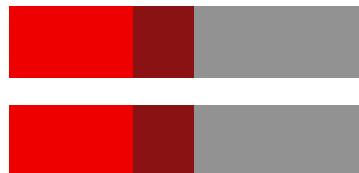
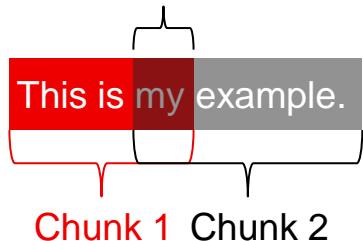
Combine



Overlap

Chunking Approaches

Overlap

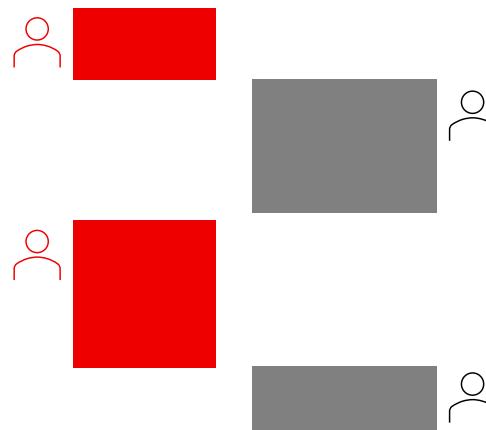


Fixed Chunk-Sizes

- Identical
- pre-defined

Structure-Based Chunk-Sizes

- e.g. chat messages should be consistent, no mix of users and chunks



Sentence 1

The cat chased the mouse. The mouse

Sentence 2

was afraid. Trees provide oxygen for

us to breath.

us to breath.

- Sentence 1 and 2 are very similar → same chunk
- Sentence 3 different → new chunk

Semantic Chunking

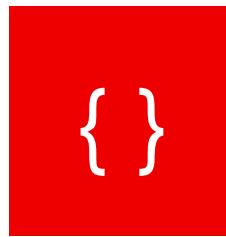
- based on semantic similarity
- e.g. when semantic break is observed



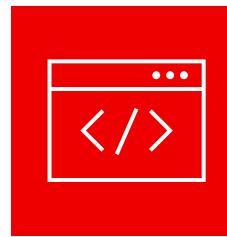
Splitter-Types



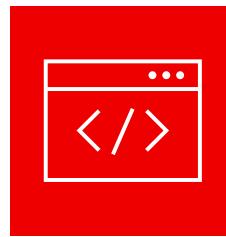
Text



JSON



HTML



Code

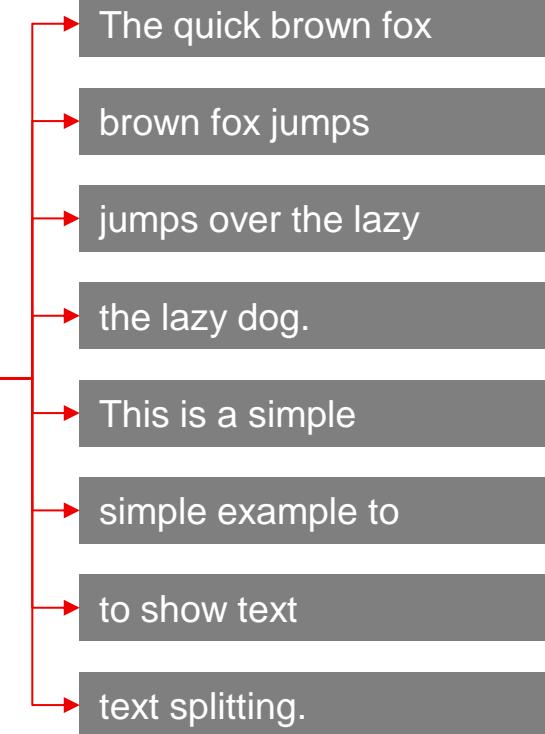


Recursive Splitting

- chunk_size...defines maximum size of chunks [characters]
- chunk_overlap...possible overlap of max 5 characters

The quick brown fox jumps
over the lazy dog.\n This is a
simple example to show text
splitting.\n.

```
RecursiveCharacterTextSplitter(  
    chunk_size=20,  
    chunk_overlap=5  
    separators=[“\n”, “ “, “”]  
)
```





Tokenization

The quick brown fox jumps over the lazy dog.

The quick brown fox jumps over the lazy dog .



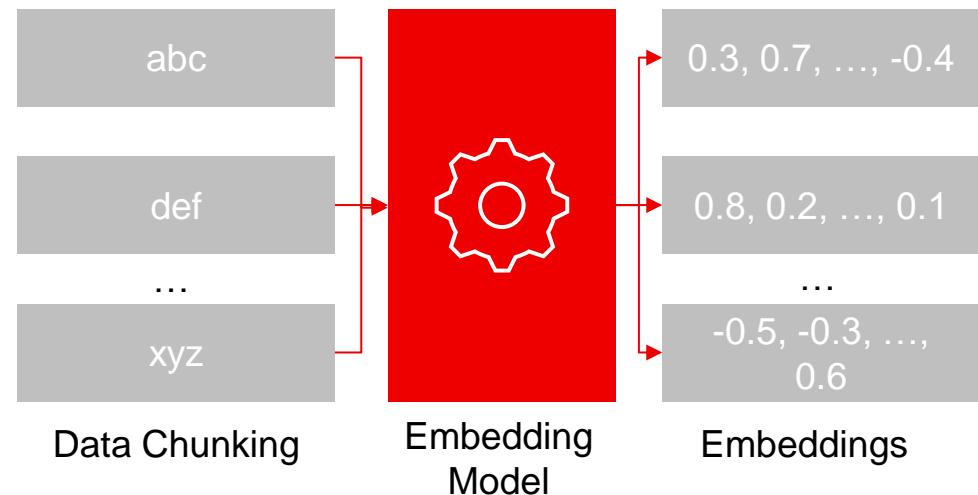
Tokenization

791 4062 14198 39935 35308 927 279 16053 5679 13



Context Window

- Embedding model works with tokens, NOT words
- Model can cover only specific sequence lengths
- Too long text (longer than context window) will be truncated



The background features a vibrant red-to-yellow gradient. Overlaid on this gradient are several semi-transparent, overlapping circles in shades of red, orange, and yellow, creating a dynamic, layered effect.

O'REILLY®