TERMS & CONCEPTS

Corpus

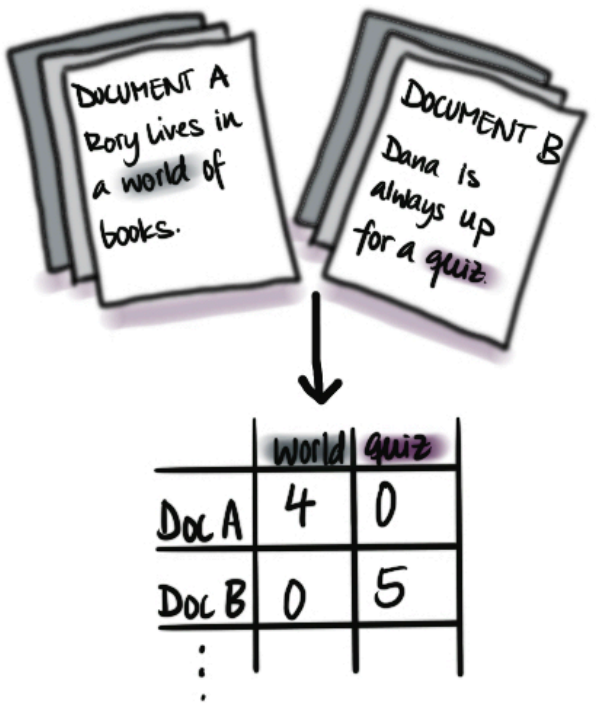
Collection of documents

Tokens

Each individual word in a text (but it could also be a sentence, paragraph, or character)

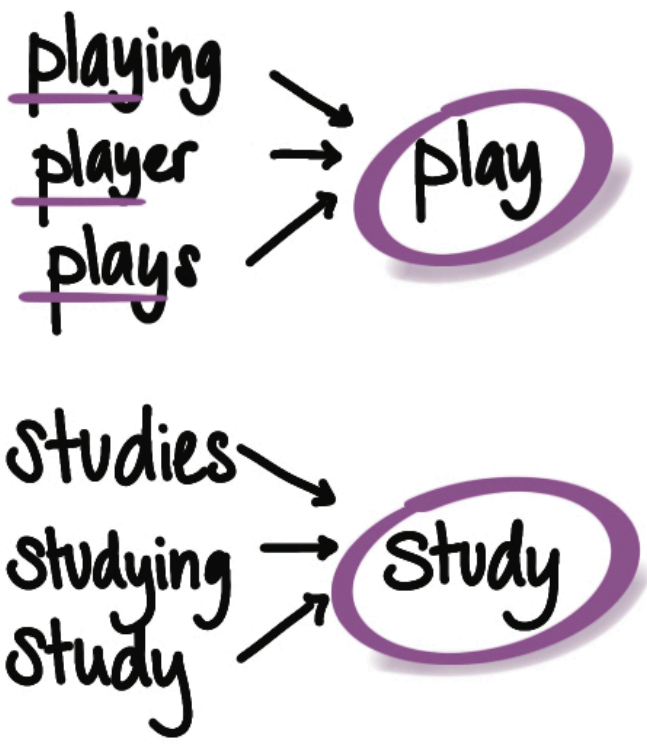
Tokenization

Creating a **bag of words**



Document-feature matrix (DFM)

First split the text into its single terms (tokens), then count how frequently each token occurs in each document



Stemming

Getting the stem of the word

Lemmatization

Getting the meaningful stem of the word

Cosima_meyer