Comparing texts

Data Science in a Box

datasciencebox.org

What is a document about?

- Term frequency
- Inverse document frequency

$$idf(\text{term}) = \ln\left(\frac{n_{\text{documents}}}{n_{\text{documents containing term}}}\right)$$

tf-idf is about comparing **documents** within a **collection**