

- **Term Frequency-inverse document frequency (or TF-idf):**
  - Very high level summary:
    - if a term (word) appears frequently in a document, that term is probably important in that document.
    - but, if a term appears frequently in many documents, that term is probably less important generally.
  - Consider a document containing 100 words wherein the word *cat* appears 3 times. The term frequency (i.e., tf) for *cat* is then  $(3 / 100) = 0.03$ . Now, assume we have 10 million documents and the word *cat* appears in one thousand of these. Then, the inverse document frequency (i.e., idf) is calculated as  $\log(10,000,000 / 1,000) = 4$ . Thus, the Tf-idf weight is the product of these quantities:  $0.03 * 4 = 0.12$ .
  - Typically, the tf-idf weight is composed by two terms: the first computes the normalized Term Frequency (TF), aka. the number of times a word appears in a document, divided by the total number of words in that document; the second term is the Inverse Document Frequency (IDF), computed as the logarithm of the number of the documents in the corpus divided by the number of documents where the specific term appears.