

## NLP: Text Preparation

### Text Representation

- **Bag of Words:** Representing a document as a vector, where values indicate word frequency. “Mary had a little lamb, little lamb, little lamb” becomes

a	had	lamb	little	Mary
1	1	3	3	1

- **N-Grams:** all the combinations of  $n$  words. Common examples are bigrams and trigrams. “Mary had a little lamb” in bigrams:

(Mary had) (had a) (a little) (little lamb)

### TF-IDF

#### Term Frequency

Term frequency is how often a word appears in a document. It requires a word, and a document that contains the word to calculate.

$$\text{tf}(\text{word}, \text{doc}) = \frac{\# \text{ of times word occurs}}{\text{Total \# words in doc}}$$

#### Inverse Document Frequency

IDF tells how much information a word provides<sup>1</sup>.

$$\text{idf}(\text{word}, D) = \log \left( \frac{|D|}{|\{\text{doc} \in D, \text{word} \in \text{doc}\}|} \right)$$

The calculation for IDF requires a word, and a list of documents,  $D$ . The numerator is the length of  $D$ . The denominator is the length of the list of documents that contain the word, for every document in the list of documents.<sup>2</sup>

### TF-IDF

TF-IDF is simply the product of the previous two values:

$$\text{tf-idf}(\text{word}, \text{doc}, D) = \text{tf}(\text{word}, \text{doc}) \times \text{idf}(\text{word}, D)$$

Note that it doesn't make sense to talk about the tf-idf value for a single word without also talking about a specific document, as each combination of word and document will have a separate tf-idf value.

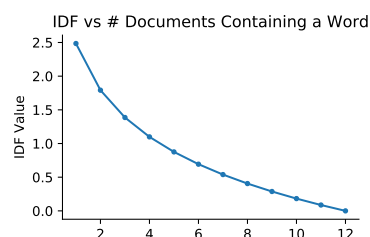
### Text Cleaning

- **Tokenization:** Breaking text down into discrete units, e.g. separating punctuation from words.
- **Stemming:** Finding the stem of the word; “chops off” the end of the word.
- **Lemmatization:** Finds the base form of the word. More computationally expensive than stemming.
- **Stopwords:** Words that are very common and are usually removed. For example, “the” or “and”.

There are several variations on term frequency:

- **Raw Count:** this is simply the count of the number of occurrences of each word.
- **Frequency:** the number of times each word appears divided by the total number of words.
- **Augmented Frequency:** the frequency of each word divided by the maximum frequency. This can help prevent bias towards larger documents.

<sup>1</sup> As the number of documents that a word appears in increases, the IDF value decreases. This can help us identify relatively important words.



<sup>2</sup> Some definitions of IDF will add some constant value to the denominator (usually 1) in order to allow for the case where one wishes to calculate the IDF for a word that doesn't appear in *any* of the documents. Without adding this term, the denominator would be 0.