TF-IDF

### TF-IDF stands for term frequency-inverse document frequency and it is a measure, used in the fields of information retrieval (IR) and machine learning, that can quantify the importance or relevance of string representations (words, phrases, lemmas, etc) in a document amongst a collection of documents (also known as a corpus).

# Overview

TF-IDF can be broken down into two parts TF (term frequency) and IDF (inverse document frequency).

### TF (term frequency)

Term frequency works by looking at the frequency of a particular term you are concerned with relative to the document. There are multiple measures, or ways, of defining frequency:

Number of times the word appears in a document (raw count).

Term frequency adjusted for the length of the document (raw count of occurences divided by number of words in the document).

Logarithmically scaled frequency (e.g. log(1 + raw count)).

Boolean frequency (e.g. 1 if the term occurs, or 0 if the term does not occur, in the document).

### IDF (inverse document frequency)

Inverse document frequency looks at how common (or uncommon) a word is amongst the corpus. IDF is calculated as follows where t is the term (word) we are looking to measure the commonness of and N is the number of documents (d) in the corpus (D).. The denominator is simply the number of documents in which the term, t, appears in.

# Using TF-IDF

* Information retrieval, with one common example being search engines.
* Text summarization & keyword extraction
* Vectors & Word Embeddings