**Measuring Text Similarity: TF-IDF and Cosine Similarity**

One established method for computing the similarity between two bodies of texts is to measure their cosine similarity. In order to measure the cosine similarity between two bodies of text, both bodies of text must first be represented in a quantitative form. By using a TF-IDF analysis, the bodies of text are vectorized so that their cosine similarity can be calculated.

**TF-IDF**

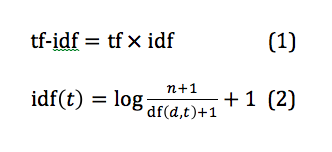
TF-IDF, short for Text Frequency – Inverse Document Frequency, is a statistical measure that evaluates how relevant a word is to a document in a collection of documents. TF-IDF originates as a method for document search and information retrieval. This value of statistical significance is obtained by multiplying two metrics: how many times a word appears in a document, and the inverse document frequency of the word across a set of documents.

The TF-IDF weighting of a term in a collection of documents increases proportionally to the number of times a term appears in a document. However, this value is offset by the number of documents that contain that specific term. This is so that common words that appear across multiple documents, such as “this”, “what”, and “if”, are ranked lower despite their high frequency of appearance. Likewise, terms that are unique and appear sparsely in the collection of documents are weighted as more relevant.

The term frequency (TF) of a term in a document refers to the number of occurrences of that term inside a document. The simplest calculation of term frequency is a raw numerical count of the instances of that term. This frequency can be adjusted several different ways, one of which is by accounting for the length of a document.

The inverse document frequency (IDF) of a term in a document refers to the rarity of a word in the entire collection of documents. The value of an IDF ranges from 0 to 1. Values that trend towards 0 are more common, while values that trend towards 1 are rarer. This is calculated by taking the total number of documents, dividing it by the number of documents that contain the term, and calculating the logarithm of that value.

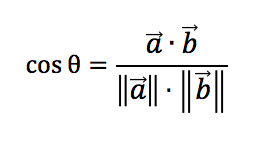
The TF-IDF score for a term is calculated by multiplying these two values (TF and IDF). The higher the score, the more relevant that term is in a particular document. The equations below are what is used in our calculations to calculate the TF-IDF.



In a pure IDF formula, the 1 is not added to the numerator, denominator, or the final logarithmic value. In our calculations the 1 is added to prevent division by 0 and negative values, which are harder to interpret.

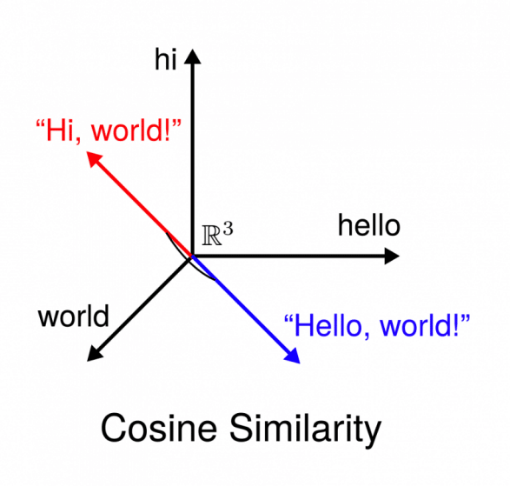
**Cosine Similarity**

With a TF-IDF analysis, we can determine how relevant a term is to a document. However, to determine the similarity between documents, we need additional metrics and a way to compare the similarity of the terms between documents as a whole. Cosine similarity is a metric used to measure how similar documents are by measuring the cosine of the angle between two vectors. In our case, the two vectors represent two bodies of text. Cosine similarity is more accurate than other methods of determining similarity (i.e. counting the number of similar terms) because even if two similar documents are far apart due to term frequency or length of text, calculating the cosine of the angle between the two will still give an accurate representation of similarity.



The formula above represents how the cosine similarity of the two vectors is calculated, where *a* and *b* are the vectors of the bodies of text. The cosine of the angle is the dot product of the two vectors divided by the product of their Euclidean norms. As the value of θ increases, the value of cos θ decreases. Two texts that are identical will have a cos θ value of 1. Two texts that are completely different will have a cos θ of 0. Thus, larger cos θ values will indicate that two texts are more similar.

When vectors are plotted on a multi-dimensional space, the cosine of the angle between vectors determines how far apart they are in similarity. The diagram below shows a graphical example of two sentences. There are a total of 3 terms across this collection of documents: “hello”, “hi”, and “world”.



**Examples**

Let’s take a look at some examples. We want to find out how similar these two sentences are:

1. Elise ate a salad for lunch on Friday.
2. John skipped lunch and went for a walk instead.

The first step in our process is to determine how significant each term is in the collection of documents. In this example, there are only two documents. We start by vectorizing the terms through a TF-IDF analysis.

**Using the Cosine Similarity Values**