# Vector space representation and similarity

Normalisation

Let’s have an n-dimensional space called Vector Space. This model is given from a collections of vectors, fundamental for operations in information retrieval, document classification, clustering and several others NLP applications. This vector space may be a representation of a document collection.

Premises: tf-idf is the weighting scheme of choice for this system but this space model may apply to any different scheme. Anyway here the assumption is the adoption of tf-idf.

This space has one axis for each term. The ordering of the terms is lost in a Bag of Words representation.

One way to measure the similarity between two vectors is to consider the difference in magnitude of two vectors. That can have a huge drawback in case two vectors, although extremely similar, have a considering difference in length. That may cause one to have a much greater absolute term frequency.

To deal with the effects caused by the above situation, one common method to calculate the similarity is to obtain the *Cosine Similarity* between the two vectors:

Description: Numerator being the *Dot Product* of the tow vectors, denominator being the product of *Euclidean Lengths* defined as , where being the vector representation of a document and *M* components.

The *Euclidean Lengths* has the effect to normalise the length of two vectors in relation to each other.

 ϴ