6.2.2 Vector Space Model

This model is perhaps the best known and most widely used IR model.

Document Representation

A document in the vector space model is represented as a weight vector, in which each component weight is computed based on some variation of TF or TF-IDF scheme. The weight w_{ij} of term t_i in document $\mathbf{d}_j$ is no longer in $\{0, 1\}$ as in the Boolean model, but can be any number.

Term Frequency (TF) Scheme: In this method, the weight of a term t_i in document $\mathbf{d}_j$ is the number of times that t_i appears in document $\mathbf{d}_j$, denoted by f_{ij}. Normalization may also be applied (see Equation (2)).

The shortcoming of the TF scheme is that it does not consider the situation where a term appears in many documents of the collection. Such a term may not be discriminative.

TF-IDF Scheme: This is the most well known weighting scheme, where TF still stands for the **term frequency** and IDF the **inverse document frequency**. There are several variations of this scheme. Here we only give the most basic one.

Let N be the total number of documents in the system or the collection and df_i be the number of documents in which term t_i appears at least once. Let f_{ij} be the raw frequency count of term t_i in document $\mathbf{d}_j$. Then, the **normalized term frequency** (denoted by tf_{ij}) of t_i in $\mathbf{d}_j$ is given by

$$tf_{ij} = \frac{f_{ij}}{\max\{f_{1j}, f_{2j}, ..., f_{|V|j}\}}, \tag{2}$$

where the maximum is computed over all terms that appear in document $\mathbf{d}_j$. If term t_i does not appear in $\mathbf{d}_j$ then $tf_{ij} = 0$. Recall that $|V|$ is the vocabulary size of the collection.

The inverse document frequency (denoted by idf_i) of term t_i is given by:

$$idf_i = \log \frac{N}{df_i}. \tag{3}$$

The intuition here is that if a term appears in a large number of documents in the collection, it is probably not important or not discriminative. The final TF-IDF term weight is given by:

$$w_{ij} = tf_{ij} \times idf_i. \tag{4}$$