Tf means term-frequency. Term frequency defines by tf(t,d). The easiest way is to use the actual count of a term t in a document for example, if the number of times that a term t occurs in document d and if we denote the actual count by (ft,d), then the simplest term frequency scheme is:

$$\text{tf(t,d)} = (\text{f}_{\text{t,d}})$$

2 Inverse Document Frequency

Idf means inverse document-frequency. It denotes number of times of a term t that contains in a given document is multiplied with idf. There are several formula to calculate idf. They are slightly different from each other. Below is one of them:

$$\text{idf(t)}=\log \frac{n_{\text{d}}}{1+df(d,t)}$$

Here total number of documents is denoted by n_d and df(d,t) is denoting the number of documents that contains the term t. The tf-idf vectors are then normalized by the below Euclidean norm:

$$\text{v}_{norm} = \frac{v}{||v||^2} = \frac{v}{\sqrt{{v_1}^2+{v_2}^2+...+{v_\text{n}}^2}}$$

3 Term Frequency Inverse Document Frequency

As we defined before the definition of tf-idf. To calculate tf-idf of a given corpus we need to calculate tf and idf individually and multiplied both. Simple formula of tf-idf calculation is:

$$\text{tf-idf(t,d)=tf(t,d)} \times idf(t)$$

We can calculate idf in several way though they are slightly different from each other as stated before. TfidfTransformer and TfidfVectorizer. We used TfidfVectorizer in our experiment.

1 Term Frequency