**Course: Advance Bio Informatics**

**Module Title: Feature Extraction Methods**

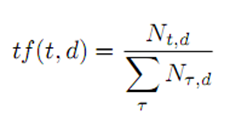
**Module No: 99**

**Feature Extraction Methods**

* A boy is playing football
* A boy is smiling
* Classes: Close-up, Sports
* Finding Useful features from text
* Occurrences of words
* Nouns, verb, adjectives …

**Term Frequency**

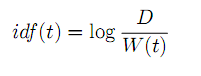
tfidf(t, d) = tf(t, d) · idf(t)



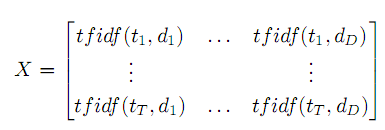
Nt,d no of occurrences of a term t in a document d

Denominator: sum of occurrences of all terms in that document d

Tf·idf

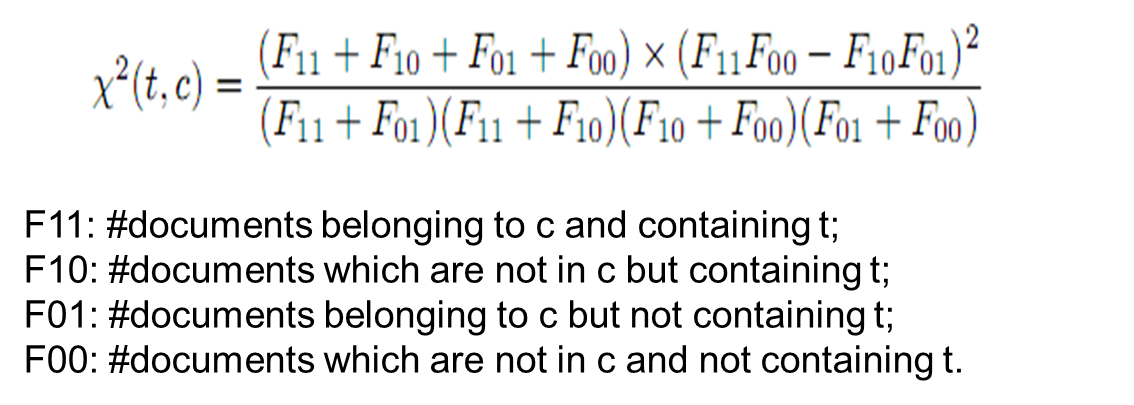


W(t) is the number of documents containing the term t.



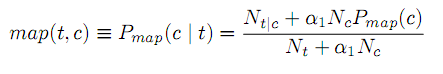
**Chi- Square Statistics**

Used to evaluate the independence between two events. The relevance of a term t in a class c can be estimated.

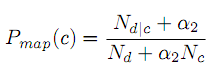
****

**MAP Estimates**

Measure importance of term t for a class c.



Nt|c and Nt are the numbers of term t in the class c and in the entire corpus, respectively. Nc is the number of distinct classes.



Nd|c is the number of documents in the scene class c, and Nd is the entire number of documents. Note that α1 and α2 are the smoothing parameters that are typically determined empirically.