REVIEW

LSA is a means of getting the contextual meaning of words from documents or passages by using mathematical and statistical methods. It can be used to show various relationships between word to word, word to passage and passage to passage. LSA can be used to simulate human cognitive phenomena such as developmental acquisition of recognition vocabulary to word-categorization, sentence-word, semantic priming and is closely identical to the way a human might categorize various documents.

LSA has been defined as a fully automatic mathematical/statistical technique for extracting and inferring relations of expected contextual usage of words in passages of discourse. It is not a traditional natural language processing or artificial intelligence program; it uses no humanly constructed dictionaries, knowledge bases, semantic networks, grammars, syntactic parsers, or morphologies, or the like, and takes as its input only raw text parsed into words defined as unique character strings and separated into meaningful passages or samples such as sentences or paragraphs.[1]

It involves the use of Singular Value Decomposition. This in mathematics means the factorization of a matrix using 3 matrices. It can be shown that for any m x n matrix M, it can be expressed as a product of three vectors

**M = UEVT**

where ,

M = the matrix,

U = an m x m matrix

VT = transpose of V of order n x n

IMPLEMENTATION

A simple implementation can be done as follows:

* First you express the texts as a matrix. The rows will consist of unique keywords while the columns will represent the documents from which they were extracted
* Next, the cells in the matrix will hold the counts of the keywords in their documents (ie columns)
* The counts are usually modified so that so that rare words are weighted more heavily than common words. A popular weighing technique is the TFIDF (ie Term Frequency – Inverse Domain Frequency)
* Next, SVD is applied to the generated matrix using an appropriate dimension