Page No.
Date

Experiment No :- 01 (Group - A)

Problem Statement :-

- Implementation of Conflation Algorithm to generate documents representative of a text file.

Objective :-

To study:

1. The various concepts and components of information retrieve
2. Conflation Algorithm
3. The role of clustering information retrival
4. Indexing structures of information retrieval.

Outcomes :-

At the end of the assignment the students will be able to

1. Understand the concept of Information retrieval and to apply clustering in information retrieval.

Scope :-

Removal of stop words

Suffix Stripping (Any five Grammer Rules)

Frequency occurences of keywords (Weight calculation).

Theory

Introduction of Information Retrieval :-

- In today's information explosion era, increase in demand for quicker dissemination of information, from contents stored in variety of forms, requires speedy search and timely retrival. The values of documents are measured according to the information it contains but they are proved useless

until the stored information is brought out for use by the readers. This may be either by subject analysis or representation of the terms through symbols. It has always been the need of the scholar and the lingering turmoil in the minds of the library organizers, to suitably facilitate the extraction of the contents expeditiously and exhaustively that has brought forward of the concept of information retrieval.

Meaning & Definition :-

- Calvin Mooers coined the term information retrieval in 1950. In the context of library and information science, we mean to get back information, which is, in a way, hidden, from normal sight or vision.

Functions :-

- The major functions that constitute an information retrival system, comprise of Acquisition, Analysis, Representatn of information, organisation of the indexes matching.

Components of Information Retrival system.

1. The document selection subsystem
2. The indexing subsystem
3. The vocabulary subsystem
4. The searching subsystem
5. The user-system interface
6. The marching subsystem.

Page No.
Date

Document Representive:-

Document in a collection are frequently represented through a set of index terms or keywords. Such keywords might be extracted directly from the text of the document or might be specified by a human subject.

The full text is clearly the most complete logical view of a document but its usage usually implies higher computational costs. A small set of categories provides the most concise logical view of a document but its usage might lead to retrieval of poor quality.

Conflation Algorithm:-

Ultimately one would like to develop a text processing system which by means of computable methods with the minimum of human intervention will generate from the input text a document representative adequate for use in an automatic retrieval system. This is a tall order and can only be partially met. A document will be indexed by a name if one of its significant words occurs as a member of that class.

Such a system will usually consist of three parts:-

1) Removal of high frequency words.
2) Suffix stripping
3) detecting equivalent stems.

Page No.
Date

Luhn's ideas :-

In one of Luhn's early papers he states :' It is here proposed that the frequency of word occurence in an article furnishers a useful measurement of word significance. It is further proposed that the relative position within a sentence of words having given values of significance furnish a useful measurement of determining the significance of sentences. The significance factor of a sentence will therefore be based on a combination of these two measurements.' This quote fairly summaries Luhn's contribution to automatic text analysis.

The removal of frequency words, stop words or fluff words is one way of implementing Luhn's upper cut-off.

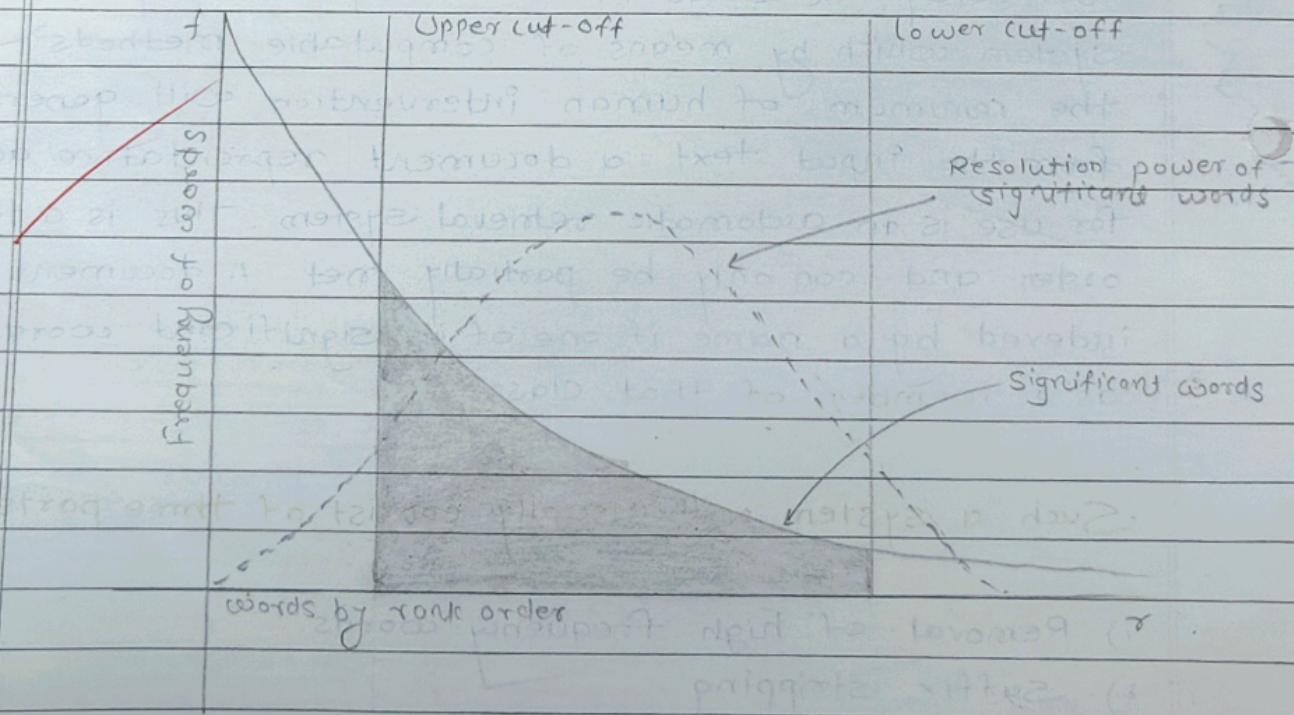


Fig: A plot of Hyperbolic curve relating frequency of words 'f' vs words by rank order 'r'.

1. Let f be the frequency of occurrence of various word types in given position of text and r their rank order, that is, the order of their frequency of occurrence, then a plot relating f and r yeilds a curve similar to the hyperbolic curve in fig. This is in fact a curve demonstrating that the product of the frequency of cuse of words and the rank order is approximately constand.

2. Suffix Stripping:

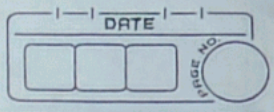
- The second stage, suffix stripping, is more complicated. A standard approach is to have a complete list of suffixes and to remove the longest possible one. For example we may well wont UAL removed from FACTUAL but not from EQUAL. To avoid erroneously removing suffixes, context rules are devised so that a suffix will be removed only if the context is right. 'Right' may mean a no. of things:

1> the length of remaining stem exceeds a given no. the default is usually 2;

2) the stem-ending satisfies a certain condition, eg. does not end with Q.

3> Detecting equivalent stems:

Many words, which are equivalent in the above sense, map to one morphological form by removing their suffixes. The simplest method of dealing with it is to construct a list of equivalent stem-endings. For two stems to be equivalent they must except for their engings, which themselves must appear in the list as equivalent. For example, stems such as ABSORB- and ABSORPT- are conflated because there is an entry in the list defining B and PT as equivalent



stem-endings if the preceding character match.
A document representative then becomes a list of class name
These are often referred to as the document's index terms
or keywords.

Input:
1> A text file containing stop words
2> A document which is searched and index according to frequency of words.

Output:
Document containing frequently appearing words without stop words and removing stemming.

Conclusion:
Thus, we have implemented the Conflation Algorithm to generate document representative of a text file.