Biomedical Text Mining

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE DEGREE OF

BACHELOR OF TECHNOLOGY

(BioInformatics)

(SESSION 2007-2011)

Submitted by

Vishal Mahajan 071509

Arjun Bagai 071516

Under the Supervision of

Mr. Kapil Dev

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY

WAKNAGHAT

SOLAN, HIMACHAL PRADESH

INDIA

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY
WAKNNAGHAT
SOLAN, HIMACHAL PRADESH

Date: 25th May 2011

CERTIFICATE

This is to certify that the thesis entitled Biomedical text mining is submitted by Vishal Mahajan & Arjun Bagai in the partial fulfillment of the award of degree of Bachelor of Technology BioInformatics by **JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT.**

Signatures in full of Supervisor: 23/05/11

Names in Capital block letters: Mr. KAPIL DEV

Designation: Lecturer

ACKNOWLEDGEMENTS

We take this opportunity to express our profound sense of gratitude and respect to all those who helped us throughout the duration of our project.

This report acknowledges to the intense driving and technical competence of the entire individual that have contributed to it. It would have been almost impossible to produce fruitful results during the the project without the support of those people. We extend thanks and gratitude to our Project Guide Mr. Kapil Dev and Mr. Suman Saha who have helped us at every step. They spent their valuable time from their busy schedule to train us and provided their active and sincere support for our daily activities.

I would like to express my heartfelt thanks to Brig.(Mr.) S.P.Ghrera, H.O.D., Computer Science Department, Jaypee University of Information Technology, for his astute guidance and his constant encouragement and support throughout.

This report has been compiled by the sincere and active support from our guides who provided us proper guidance and direction regarding various problems. We have tried our best to summarize this report.

Vishal Mahajan (071509)
Arjun Bagai (071516)
B.Tech(BI)

Contents

Chapter 8 – Results And Conclusions

Chapter 9 – Future Work Of The Project

Abstract

A Vector Space Model, inverted index i.e an optimized structure, built primarily for retrieval, this basic structure inverts the text so that instead of the view obtained from scanning documents where a document is found and then its terms are seen (a list of documents each pointing to a list of terms it contains). The point of using an index is to increase the speed and efficiency of searches of the document collection. Without some sort of index, a user's query must sequentially scan the document collection, finding those documents containing the search terms. Users are not going to wait so many minutes for a response.

Hence, a sequential scan is simply not feasible. Given this situation a data structure called an inverted index is commonly used by search engines. The index is built which maps terms to documents (pretty much like the index found in the back of this book that maps terms to page numbers) , thus resulting in better retrieval of the documents according to the respective user query.

Chapter 1

Introduction

1.1 Problem Statement

The point of using an index is to increase the speed and efficiency of searches of the document collection. Without some sort of index, a user's query must sequentially scan the document collection, finding those documents containing the search terms. Consider the "Find" operation in Windows; a user search is initiated and a search starts through each file on the hard disk. When a directory is encountered, the search continues through each directory.

Users are not going to wait so many minutes for a response. Hence, a sequential scan is simply not feasible. Given this situation a data structure called an *inverted index* is commonly used by search engines

1.2 Objective and scope of the Project

Create a Vector Space Model, and an inverted index for better retrieval of the documents according to the respective user query. A Vector Space Model, inverted index i.e an optimized structure, built primarily for retrieval, this basic structure inverts the text so that instead of the view obtained from scanning documents where a document is found and then its terms are seen (a list of documents each pointing to a list of terms it contains), the index is built which maps terms to documents (pretty much like the index found in the back of this book that maps terms to page numbers).

1.3 Text mining

Text mining is defined as the automatic discovery of new, previously unknown, information from unstructured textual data. It uses NLP technique for discovery process and is often seen as comprising of three major tasks:

- Information retrieval (gathering relevant documents).
- Information extraction (extracting information of interest from these documents).
- Data mining (discovering new associations among the extracted pieces of information).

1.4 Approaches to text mining

- **Three basic types of approaches to text mining are:**

- **Co-occurrence–based methods** look for concepts that occur in the same unit of text—typically a sentence, but sometimes as large as an abstract—and posit a relationship between them.
- **Knowledge-based approaches** make use of some sort of knowledge like how language is structured, how biologically relevant facts are stated in the biomedical literature or what are the main fields of bioscientists interests and what kinds of relationships they can have with one another, or what variant forms by which they might be mentioned in the literature, or any subset or combination of these.
- **Statistical or machine-learning-based approaches** operate by building classifiers that may operate on any level, from labelling part of speech to choosing syntactic parse trees to classifying full sentences or documents.

1.4 Natural Language Processing

NLP – Natural Language Processing is a field of computer science and linguistics concerned with the interactions between computers and human (natural) languages. Natural language generation systems convert information from computer databases into readable human language.

1.5Text-mining in molecular biology, BioNLP

Text-mining in molecular biology is defined as the automatic extraction of information about genes, proteins and their functional relationships from text documents. Some applications of NLP techniques in biology (which is sometimes called Biomedical Natural Language Processing – BioNLP) can include:

- **bio-entity recognition, protein/gene normalization, interaction extraction and many others.**
- A major problems dealing with NLP tasks is Text segmentation.

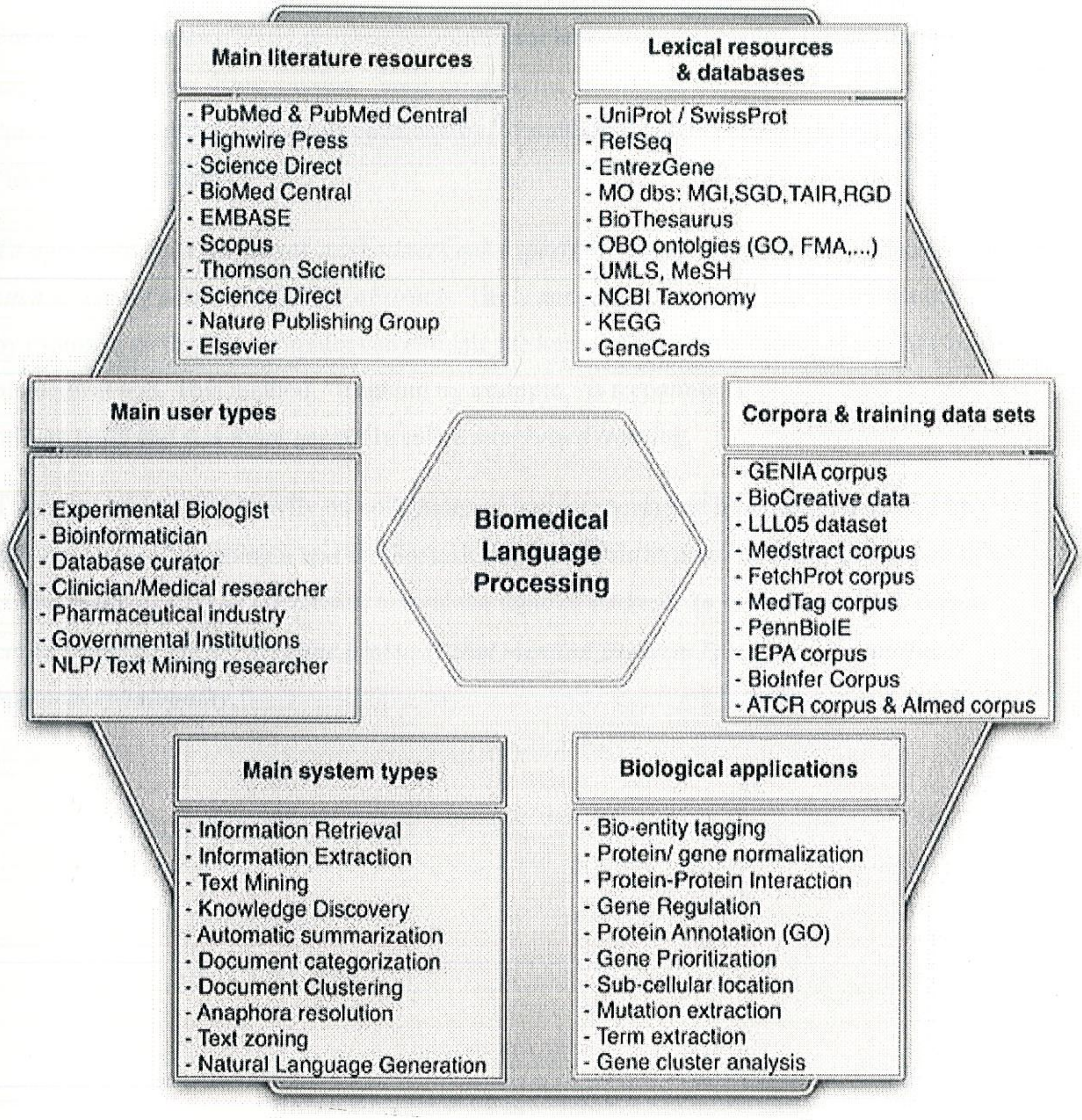
1.6 Natural Language Processing Technique

The **fundamental problem** is the fact, **that a particular meaning may be expressed using different but largely synonymous expressions,**

- For example, the information about the interaction between the proteins AKT and GSK-3 can be expressed in many ways, such as: AKT protein phosphorylates GSK-3 enzyme; the phosphorylation of GSK-3 by AKT and GSK-3 is inhibited by AKT.

So, The NLP techniques need to somehow 'decode' the information that is packaged in human language.

1.7 Literature Sources Diagram



Main literature resources

- PubMed & PubMed Central
- Highwire Press
- Science Direct
- BioMed Central
- EMBASE
- Scopus
- Thomson Scientific
- Science Direct
- Nature Publishing Group
- Elsevier

Lexical resources & databases

- UniProt / SwissProt
- RefSeq
- EntrezGene
- MO dbs: MGI,SGD,TAIR,RGD
- BioThesaurus
- OBO ontolgies (GO, FMA,...)
- UMLS, MeSH
- NCBI Taxonomy
- KEGG
- GeneCards

Main user types

- Experimental Biologist
- Bioinformatician
- Database curator
- Clinician/Medical researcher
- Pharmaceutical industry
- Governmental Institutions
- NLP/ Text Mining researcher

Biomedical Language Processing

Corpora & training data sets

- GENIA corpus
- BioCreative data
- LLL05 dataset
- Medstract corpus
- FetchProt corpus
- MedTag corpus
- PennBioIE
- IEPA corpus
- BioInfer Corpus
- ATCR corpus & Almed corpus

Main system types

- Information Retrieval
- Information Extraction
- Text Mining
- Knowledge Discovery
- Automatic summarization
- Document categorization
- Document Clustering
- Anaphora resolution
- Text zoning
- Natural Language Generation

Biological applications

- Bio-entity tagging
- Protein/ gene normalization
- Protein-Protein Interaction
- Gene Regulation
- Protein Annotation (GO)
- Gene Prioritization
- Sub-cellular location
- Mutation extraction
- Term extraction
- Gene cluster analysis

1.8 Biomedical Text Mining And Its Practical Applications

For example, the text "early progressive multifocal leukoencephalopathy" could possibly refer to any, or all, of these disease terms: "early progressive multifocal leukoencephalopathy," "progressive multifocal leukoencephalopathy," "multifocal leukoencephalopathy," and "leukoencephalopathy."

To overcome such dilemmas, text miners ask experts to identify terms within collections of text such as sets of selected Medline abstracts. These annotations are then used to train a computer by example, so that the computer can emulate the knowledge experts deploy when they read biomedical text. This method, "teaching by example," is a common approach used in many text mining tasks and it is more generally called ***supervised training.***

Thus, text miners rely heavily on collections of text (corpora) that have been annotated by experts. Before beginning a text mining task, it is advisable to limit the scope of the task to a corpus made of a set of documents around the topic of interest. ***In our case, a PML corpus could comprise all the Medline abstracts that mention the term "progressive multifocal leukoencephalopathy,"***

Chapter 2 –

Text Mining

2.1 What Text Miners are Interested In !

□ Terms

□ Relationships

□ Discovery

□ Quality

□ Comprehensiveness

2.1.1 Terminologies

Text miners are interested in terminologies that are more than just a flat list of terms. Some include term synonyms or relations between terms (taxonomies, ontologies).

For text miners, their usefulness comes from their ability to link to information. Once a text is mapped to one of these terminologies, a bridge is opened between the text and other resources been built manually.

This usefulness justifies efforts such as the National Library of Medicine's manual mapping of Medline abstracts to the Medical Subject Headings (MeSH) terminology. ***In our example, MeSH can be used to make the PML corpus more focused by restricting it only to abstracts with the MeSH term "leukoencephalopathy, progressive multifocal."***

2.1.2 Relationships

After recognizing terms, the natural next step is to look for relationships between terms.

The simplest method to identify relationships is using the *co-occurrence* assumption: terms that appear in the same texts tend to be related. For example, if a protein is mentioned often in the same abstracts as a disease, it is reasonable to hypothesize that the protein is involved in some aspect of the disease.

The degree of co-occurrence can be quantified statistically to rank and eliminate statistically weak co-occurrences. ***An example using GoGene _illustrate the use of simple co-occurrence, the query "leukoencephalopathy, progressive multifocal" in GoGene returns all the genes mentioned in Medline abstracts annotated for PML. The genes that appear most often are likely to be related to PML.***

2.1.3 Discovery

Besides finding relationships, text miners are also interested in *discovering* relationships. **Due to the size of the literature, scientists miss links between their work and other, related work.** Swanson called these links "undiscovered public knowledge."

One method to discover relationships is based on transitive inference. Simply stated, if A is linked to B, and B is linked to C, then there is a chance that A is linked to C.

2.1.4 Quality

The most common measure of output quality in text mining is the F-measure, which is the harmonic mean of two other measures, precision and recall. Both high precision and high recall are desirable, and a high F-measure reflects both because it is the harmonic mean.

Moreover, text mined applications may perform differently in different types of text and this may be reflected in lower F-measures than advertised. When the F-measure attainable is not high enough, one solution is to use text mining as a filter. Filtering with text mining is used as a preliminary step in databases such as MINT , DIP, and BIND.

2.1.5 Comprehensiveness

Doing comprehensive text mining means considering all sources of information—Medline and beyond.

The abstract conveys an article's main findings, but many other pieces of information are elsewhere in the full text, figures, tables, supplementary information, references, databases, Web sites, and multimedia files.

A search for "progressive multifocal leukoencephalopathy" in the Yale Image Finder yields only one figure, while a search for "PML" yields a large number of hits, most of them not relevant because PML is an ambiguous term.

Chapter 3 -

Vector Space Model

3.1 What is a Vector Space Model ?

Vector space model (or term vector model) is an algebraic model for representing text documents (and any objects, in general) as vectors of identifiers, such as, for example, index terms.

It is used in information filtering, information retrieval, indexing and relevancy rankings. Its first use was in the SMART Information Retrieval System.

3.2 Stages Of Vector Space Model

The vector space model procedure can be divided in to three stages -

- The **first stage** is the document indexing where content bearing terms are extracted from the document text.
- The **second stage** is the weighting of the indexed terms to enhance retrieval of document relevant to the user.
- The **last stage** ranks the document with respect to the query according to a similarity measure.

3.2.1 First Stage- Document Indexing

It is obvious that many **of the words in a document do not describe the content, words like *the, is.*** By using **automatic document indexing** those non significant words (function words) are **removed from the document vector, so the document will only be represented by content bearing words .**

This indexing can be based on term **frequency, where terms that have both high and low frequency within a document are considered to be function words** . **Use of a stop list which holds common words to remove high frequency words (stop words).**

In general, 40-50% of the total number of words in a document are removed with the help of a **stop list**. **Non linguistic methods** for indexing have also been implemented. **Probabilistic indexing** is based on the assumption that there is some statistical difference in the distribution of content bearing words, and function words.

Probabilistic indexing **ranks the terms in the collection w.r.t. the term frequency** in the whole collection. The **function words are modeled by a Poisson distribution** over all documents, as content bearing terms cannot be modeled.

3.2.2 Second Stage- Term Weighting

There are three main factors term weighting:
term frequency factor, collection frequency factor and length normalization factor.

These three factor are multiplied together to make the resulting term weight.

Collection frequency document : These assume that the importance of a term is proportional with the number of document the term appears in ,and leads to a more effective retrieval, i.e., an improvement in precision and recall.

The third possible weighting factor is a **document length normalization factor**. Long documents have usually a much larger term set than short documents, which makes long documents more likely to be retrieved than short documents.

3.2.3 Third Stage- Similarity Coefficients

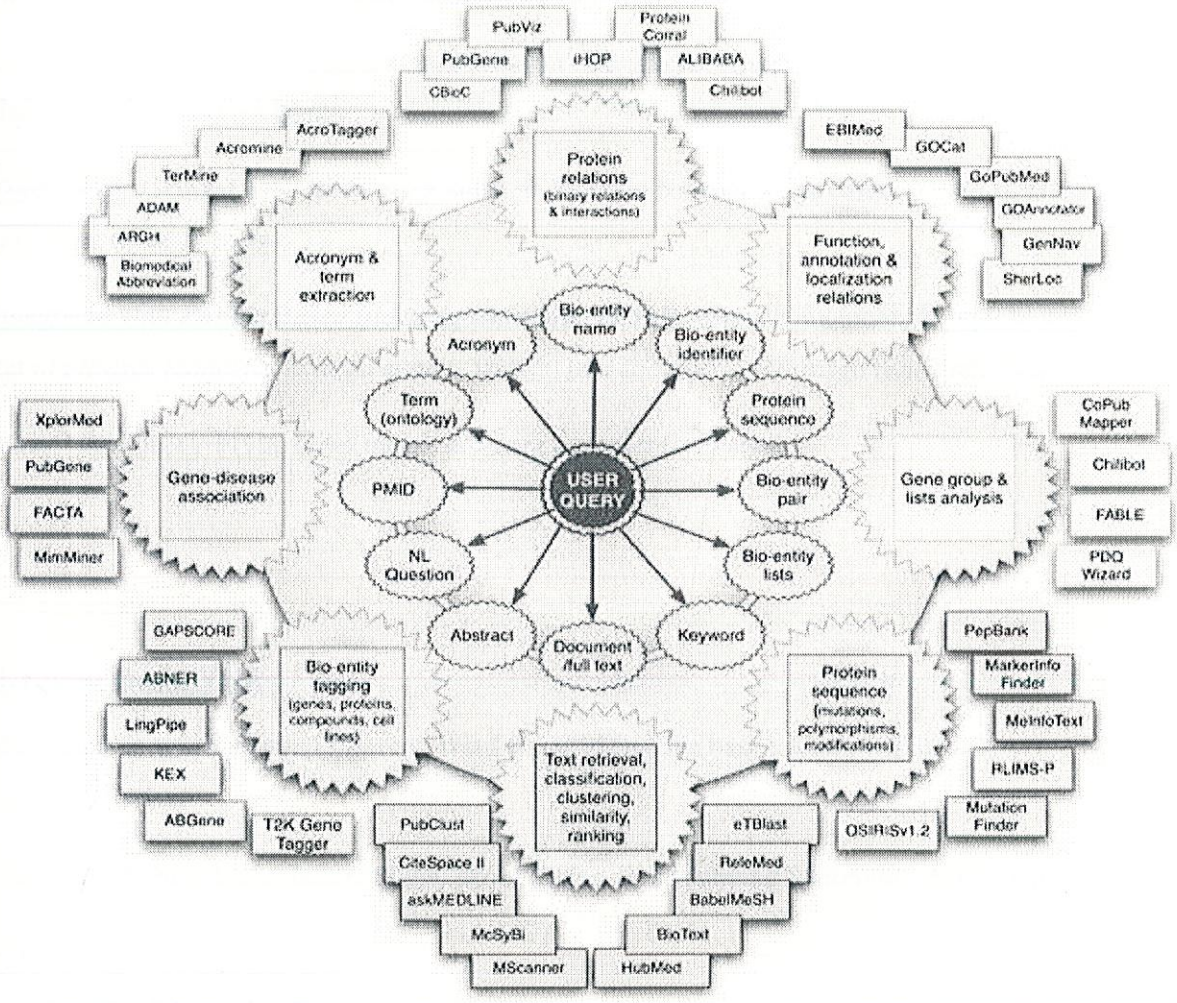
The **similarity in vector space models is determined by using associative coefficients** based on the inner product of the document vector and query vector, where word overlap indicates similarity. The inner product is usually normalized.

The most popular similarity measure is the **cosine coefficient**, which measures the angle between the a document vector and the query vector.

cosine [{(|v1|.|v2|)/(|v1|X|v2|}]

Other measures are e.g., Jaccard and Dice coefficients.

3.2.4 User Query Diagram



Chapter 4 –

Query Extension Technique and Vector Space Models

4.1 Query Extension Technique (knowledge based)-

Through Query Extension Technique Document Reteival easy/improved, because it matches the document Index terms better.

Query with, General concept terms→ subst.→ Specific concept terms (used in corpus, mined, co-occur, relevant to doc.'s).

Integrated with ***Phrase Based indexing technique (VSM),*** where a document is represented as a set of *phrases (concepts + word stems),* leading to a complete knowledge based retreival.

Thus, User Query→(via. Indexing)→ efficient document retreival.

4.2 Stem Based VSM-

It is based on **word stems**. Different word stems means the base vectors will be ORTHOGONAL.

The major ***Limitation*** of Indexing via. Word Stems is searcheing Synonyms, hypernyms, hyponyms in documents..

- ***Doc (VSM) → vector (of terms)***
- ***Wts. Of terms, ← importance of that term, in doc.***
- Stemming: {edema, edemas} → stem word.
- ***Word Stems in our query-***

  {hyperthermia},{leukocytosis}, {increased}, {intracranial}, {pressure}……..

The weight of a word stem u in a document a is determined by **W(a,u)**, where a is the document and, u is the word stem, number of times u appears in a is **term frequency**. Similarly, number of documents that contain u is **document frequency.**

- ***more often u appears in a, the more important u is in a.***
- ***more documents u belongs to, the less important it is.***

4.3 Concept Based VSM

Uses ***concepts*** *instead of single words or word stems, and* produces a VSM that better mimics the human thought processes .

Eg. - "increased intracranial pressure"

- **stems "increas," "intracran," and "pressur."**
- inappropriate fragmentation of concepts.
- concepts represented by, multi-word phrases such as "increased intracranial pressure."

Limitation of Concept Based Vector Space Model is incomplete/quality knowledge resources.

- {Conceptual similarities, derived from knowledge sources}
- ***Concepts based, → base v's. of concepts***, Orthogaonal

  →related, acute angle b/w them.

- Phrases→ *synonymous (repr. 1 concept)*

  →*polysemous (repr. >1 concept)*

The **cosine of the angle** between two concept vectors is defined as the ***conceptual similarity****, between the corresponding* concepts, (0-1). So, 0→ unrelated→ Orthogonal.

- Wt, W(a,Xi)→ *ith concept Xi in a* document a.

Longer Phrase means high weight, so more specific concepts. For Example, “increased intracranial pressure” and “hyperthermia” are identical, but the former concept would obtain a higher weight than the latter.

4.4 Phrase-based VSM

Previous Concept based VSM, treats “cerebral edema” and “cerebral lesion” as unrelated, so, is potentially harmful. Noticing their common component word “cerebral” in the above phrases, **phrase-based VSM** is there to remedy the incompleteness of the knowledge sources.

- In phrase-based VSM, a **document → represented→ set of phrases.**
- **phrase→ multiple concepts, and several word stems**.

Thus the Advantage of Phrase Based Vector Space Model is that it removes problem of incomplete knowledge.

- So, doc.→ parse into→ phrases (based on conceptual Terms)
- ***Similarity b/w doc.'s → sim. b/w concepts and common word stems.***

Similarity of two documents α and β, is the cosine of the angle between their document vectors. And Conceptual Similarity, i.e b/w i.th concept Xi in α, j.th con. Yj in β, is the Conceptual contribution to the similarity between two phrases calculation , i.e taking all possible concept pairs into account, where each pair consists of one concept from each phrase.

- S(Xi,Yj)→conc. sim.→dist. b/w X &Y concepts.
- S(X,Y)→ conc. sim. is inversely proportional to the *d(X,Y).*
- *or, **S(X,Y) → 1/ log(no. of descendants of the two).***

Example, a general concept like “disease” has much more descendants than a more specific concept like “hyperthermia” has.

Chapter 5 -

The Test Set, OHSUMED

Ohsumed, is the Test Collection comprising of -

- Reference→MEDLINE →title,abstract, author inf.
- Query→patient description.
- Judgement→relevant references.

The Knowledge Source-

- UMLS → medical lexical knowledge source.

  →lexical and semantic network→ VSM

5.1 *Phrase Detection-*

Given a set of documents (106 queries and 14K), detect any occurrences in a set of phrases. First, detects *all occurrences* of any phrase in a document, and ***only keep the longest, most specific phrase***.

Example- although both "edema" and "cerebral edema", are detected in the sample query, keep only the latter and ignore the former.

5.2 VSM models, accuracy-

- Retreival accuracy→calc.→Precision-Recall

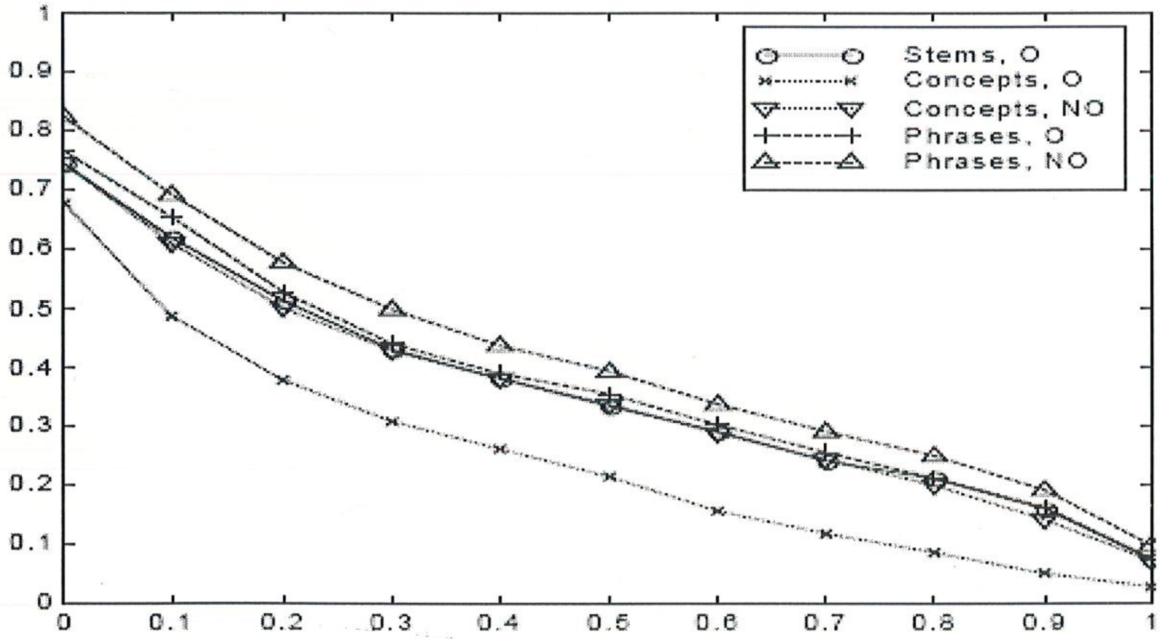
- VSM→ doc.'s → doc. similarity (14k doc.'s ,105 q.'s)

Rank the documents from the most to the least similar to the query.When a certain number of documents are retrieved, ***precision*** *is the percentage of* retrieved documents that are relevant; and ***recall*** *is the* percentage of the (relevant/irrelevant) documents that has been retrieved so far.

So, the evaluation of the retrieval accuracy is done by interpolating the precision values at eleven recall points. The overall effectiveness of different VSM can be compared by averaging over the performance of all the 105 queries.

5.2.1 Avg. precision recall comparison-

Effectiveness of VSM Diagram



Comparisons of VSM-

- **Stem based VSM-**

11 pts. Avg. Precision= 0.363

- **Concept based VSM-**

  Precision= 0.260

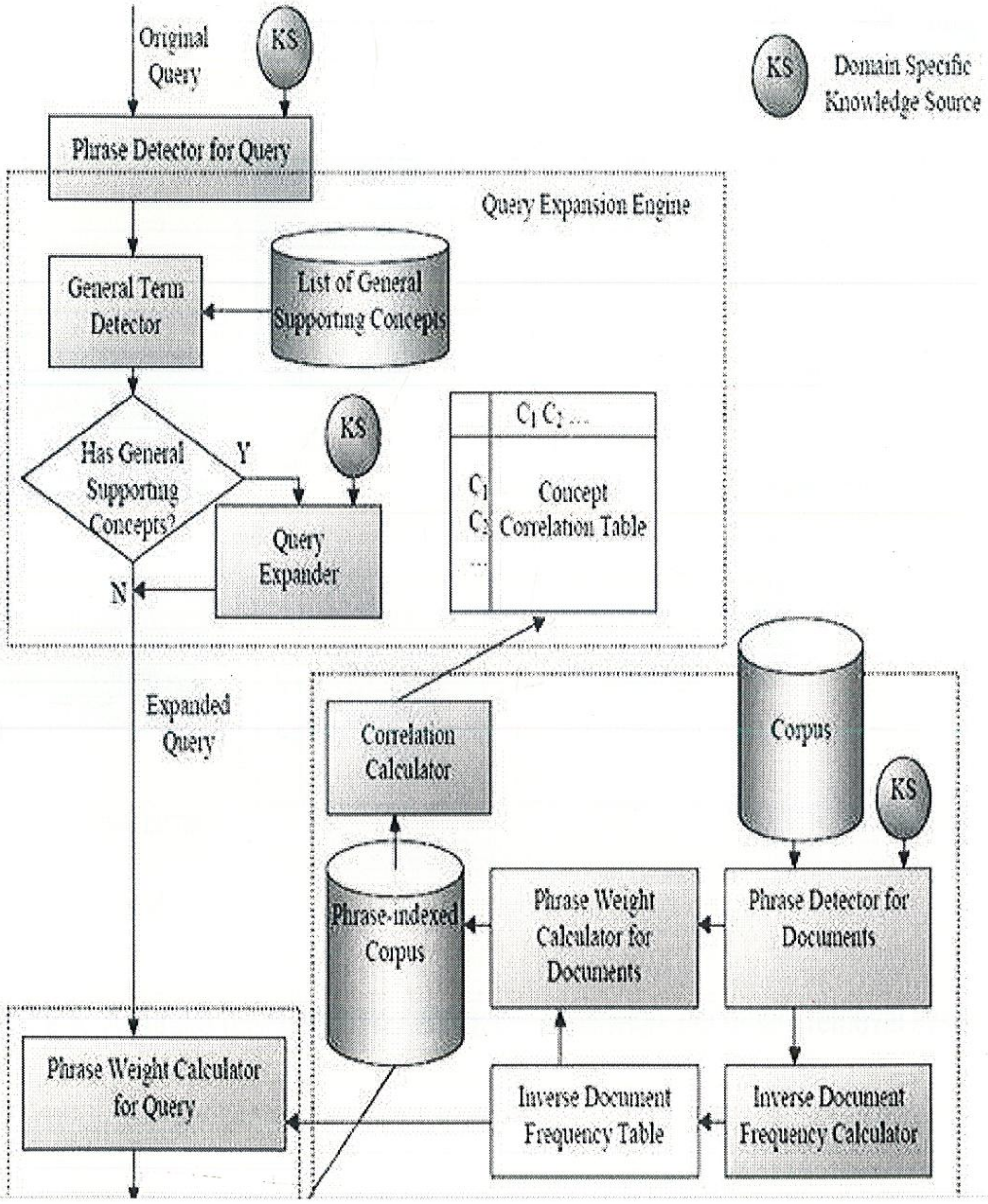
- **Phrase based VSM-** (concepts & word stems)

- Precision=0.420

So, there is 16% improvement in Retreival Accuracy, the best of the lot.

Chapter 6 –

The Flow Diagram-


Original Query
KS
KS Domain Specific Knowledge Source
Phrase Detector for Query
Query Expansion Engine
General Term Detector
List of General Supporting Concepts
Has General Supporting Concepts?
Y
N
KS
Query Expander
C_1 C_2 ...
C_1
C_2
...
Concept Correlation Table
Expanded Query
Corpus
Correlation Calculator
KS
Phrase-indexed Corpus
Phrase Weight Calculator for Documents
Phrase Detector for Documents
Phrase Weight Calculator for Query
Inverse Document Frequency Table
Inverse Document Frequency Calculator

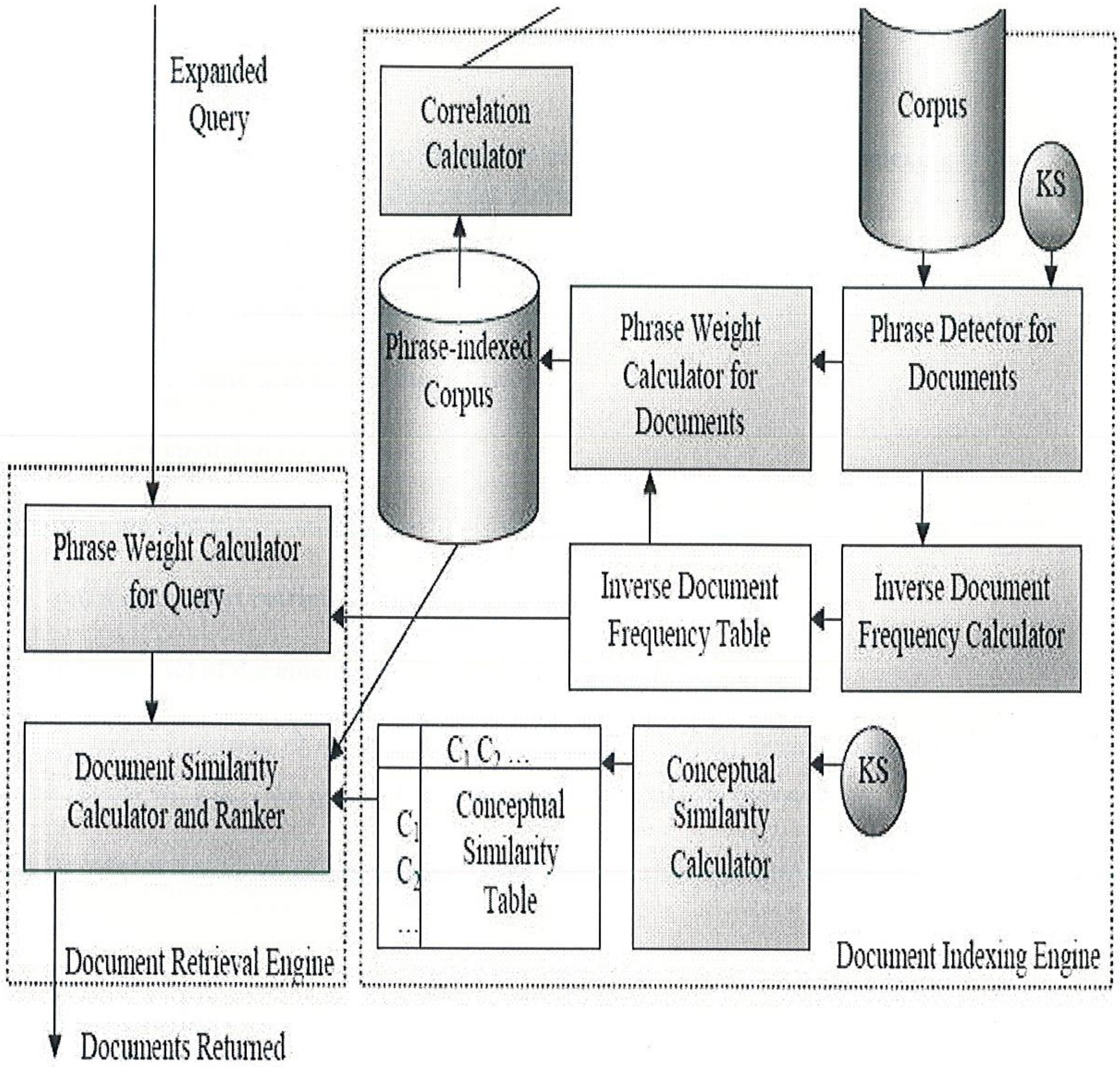


Figure 3. A phrase based indexing and query expansion document retrieval system.

6.1 Integration + Working-

Integration of knowledge based Query Extension technique and Phase based indexing is done because it gives the best accuracy of document Retreival.

It consists of three subsystems:

- **a document indexing engine,**

(knowl. based, create data str. for the nxt two)

- **a query expansion engine,**

(expands the query)

- **and a document retrieval engine,**
- (returns a set of documents relevant to the user)

Working- query→ phrases → (general c. terms-> specific c. terms) →small size query, retreival better → phrase wt. calc. , for query → ranking based on similarity measure

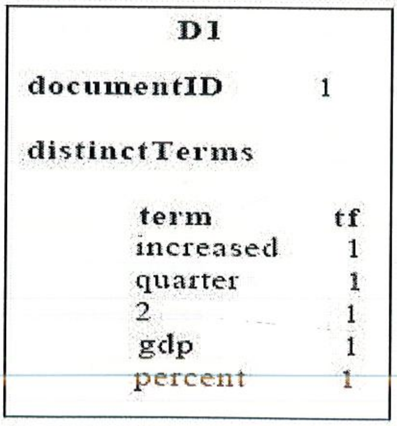
→ ***Document Retreival. (Phrase based VSM)***

6.2 Inverted Index-

Let's take the Doc.'s for Indexing..

- **D1**: The GDP increased 2 percent this quarter.
- **D2**: The spring economic slowdown continued to spring downwards this quarter.

- 2 - >[D1]
- continued ->[D2]
- downwards -> [D2]
- economic -> [D2]
- GDP -> [D1]
- increased -> [D1]
- percent -> [D1]
- quarter ->[D1] à [D2]
- slowdown -> [D2]
- spring -> [D2]
- the -> [D1] -> [D2]
- this ->[D1] -> [D2]
- to -> [D2]

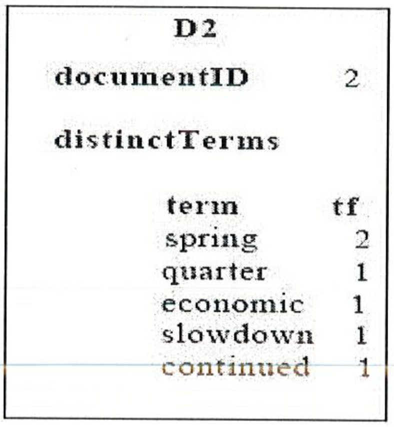


D1

documentID 1

distinctTerms

term	tf
increased	1
quarter	1
2	1
gdp	1
percent	1



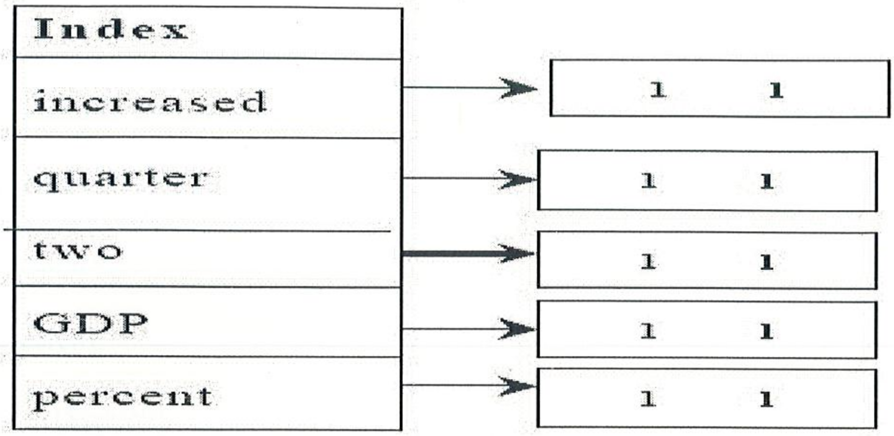
D2

documentID 2

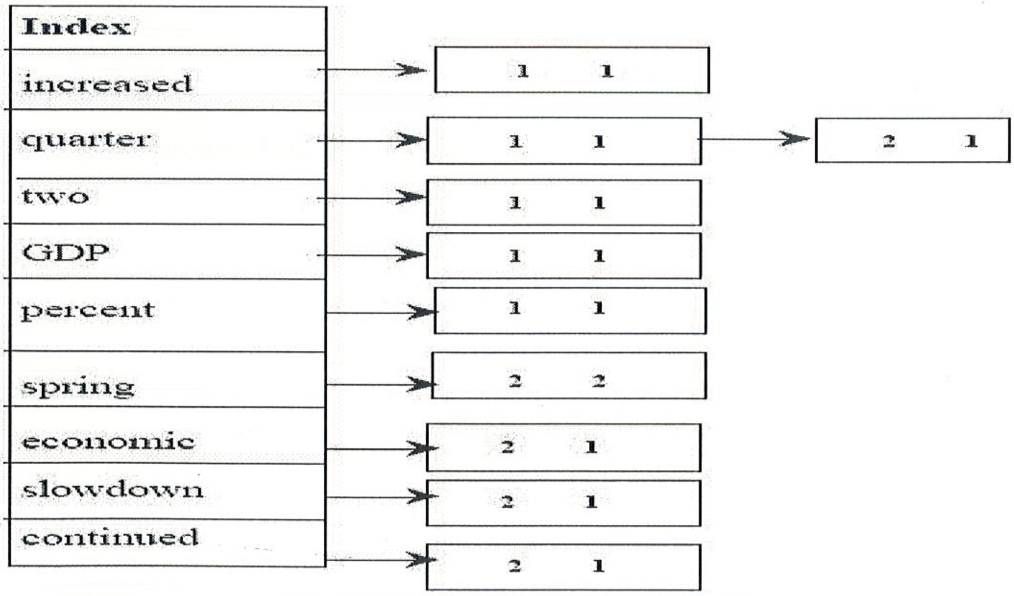
distinctTerms

term	tf
spring	2
quarter	1
economic	1
slowdown	1
continued	1

Inverted Index, Document 1 –



6.2.1 Inverted Index for added documnets



Chapter 7 –

Appendix

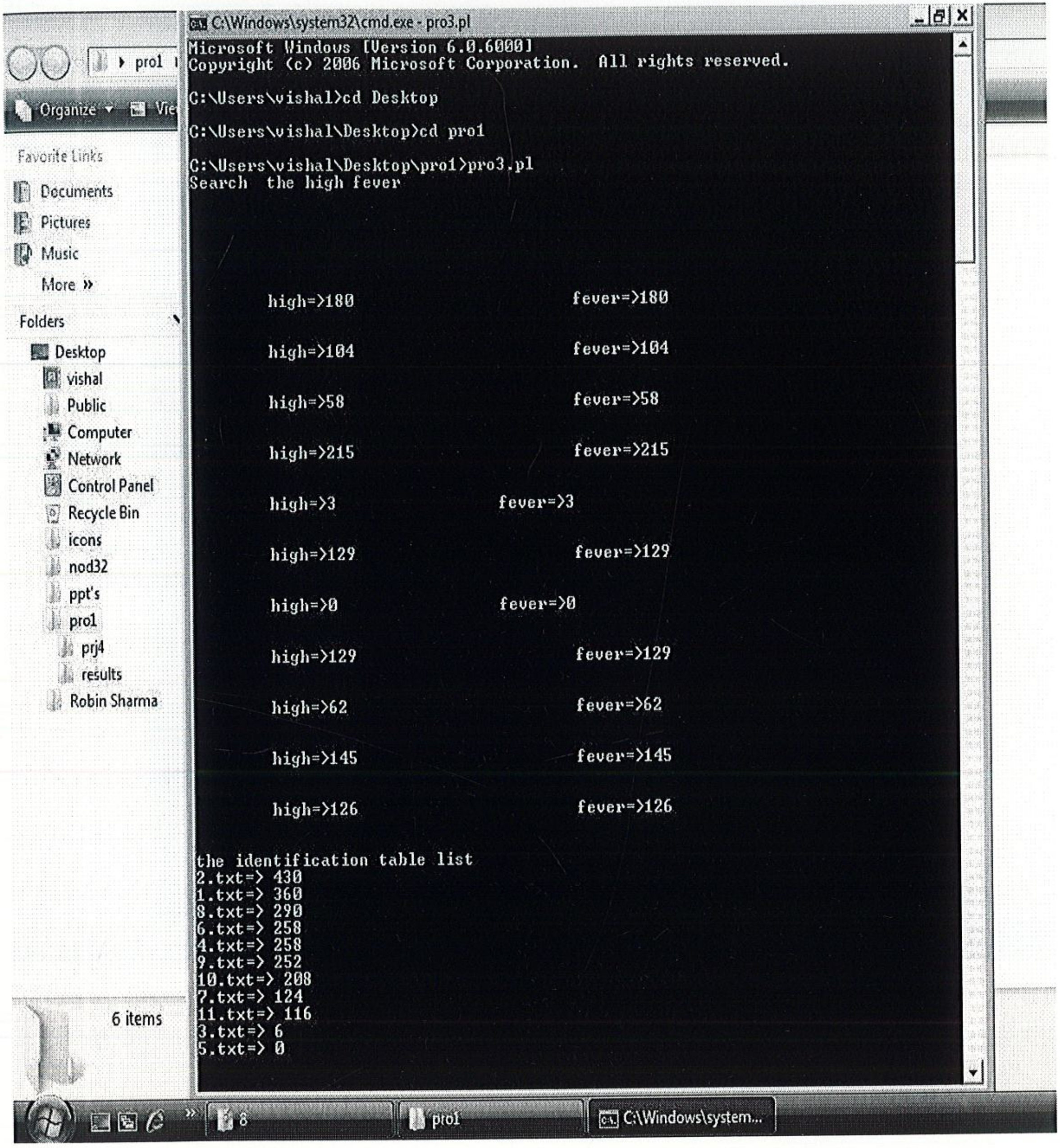
7.1 For building inverted index in Perl –

```
print "Search\t";

$search=<stdin>;

chomp($search);

#$search=uc($search);

@stoplist=qw(is an the with or between are not has had but can many what when who in it where
when why on);

@search=split('\s+',$search);

for($i=0;$i<scalar(@stoplist);$i++)

{

        for($j=0;$j<scalar(@search);$j++)

                {

                        if ($stoplist[$i] eq $search[$j])

                {

                        splice(@search,$j,1);

}}}

$dirtoget="prj4";

opendir(prj4, $dirtoget) || die("Cannot open directory");
```

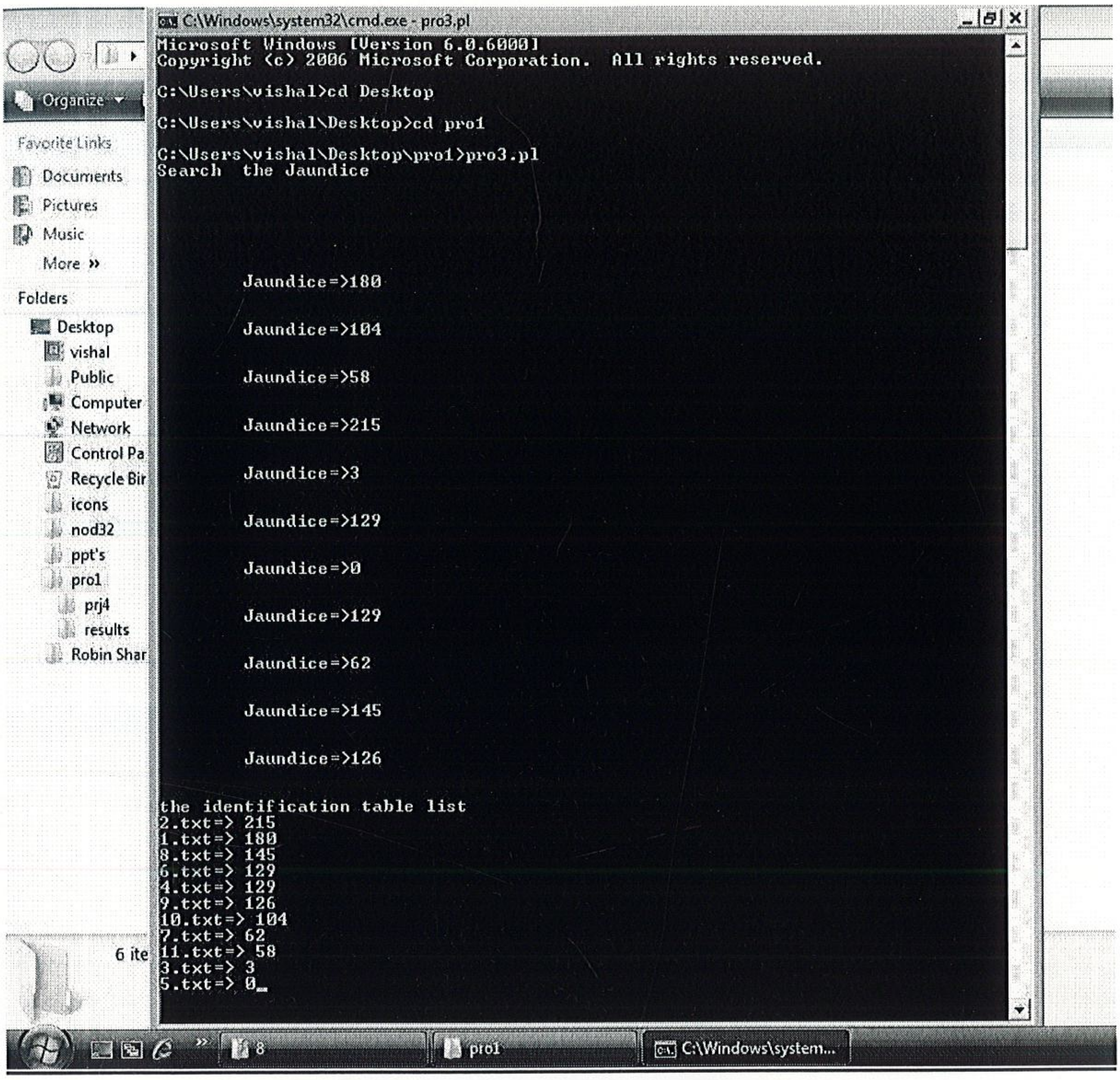
```
@thefiles= readdir(prj4);
closedir(prj4);
%it=();
foreach $f (@thefiles)
{
        unless ( ($f eq ".") || ($f eq "..") )
        {
                %l=NULL;
                open(F1,"./prj4/".$f);
                $data=<F1>;
                close (F1);
                open(F2,">./results/".$f);
                for($i=0;$i<scalar(@search);$i++)
                {
                        $l{$search[$i]}=0;
                        $l{$search[$i]}=($data=~tr/$search[$i]//);
                        if($l{$search[$i]} eq ")
                        {
                                $l{$search[$i]}=0;
                        }
                        print "\t$search[$i]=>$l{$search[$i]}\t\t";
```

```
                    print F2            "\t$search[$i]=>$l{$search[$i]}\t\n";
                }
                $total=0;
                close(F2);
                foreach $key(keys%l)
                {
                    $total=$total+$l{$key};
                }
                $it{$f}=$total;
                }
        print "\n\n\n"
}
print "the identification table list";
foreach my $key(sort {$it{$b} <=> $it{$a}} keys %it)
{
    print "\n".$key . "=> " . $it{$key};
}
<>;
```

7.1.1 Results in cmd –

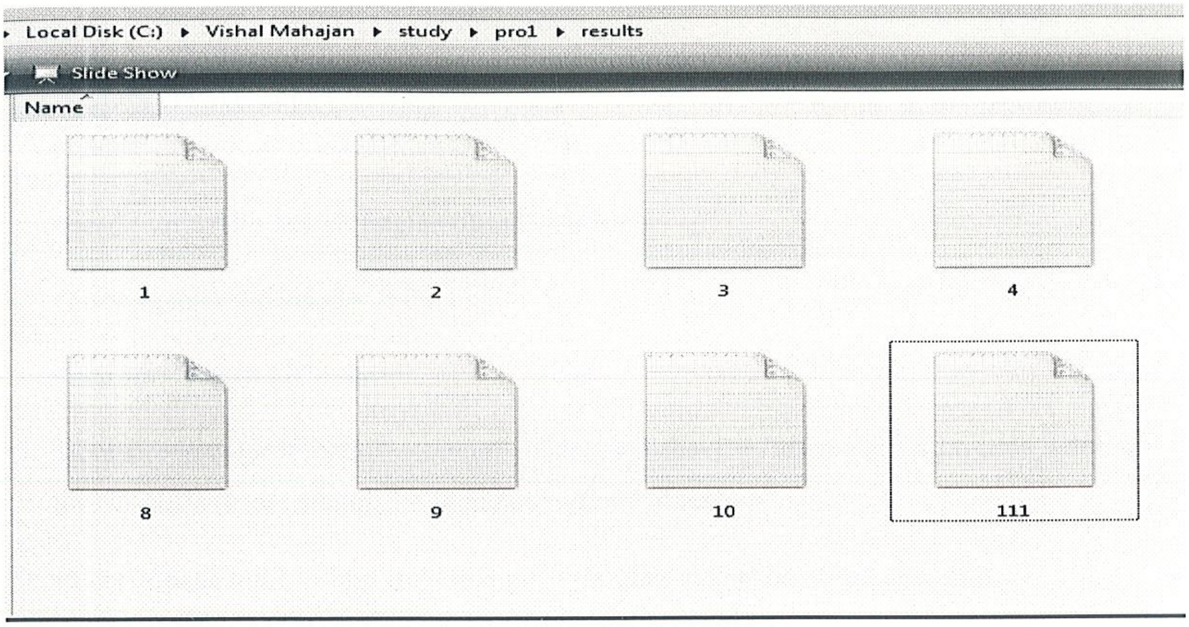


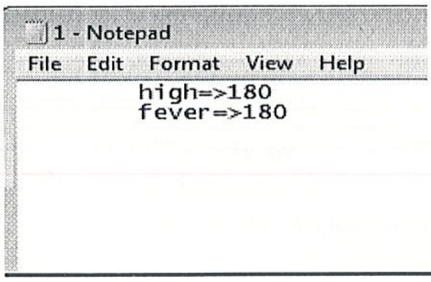
```
Microsoft Windows [Version 6.0.6000]
Copyright (c) 2006 Microsoft Corporation.  All rights reserved.

C:\Users\vishal>cd Desktop

C:\Users\vishal\Desktop>cd pro1

C:\Users\vishal\Desktop\pro1>pro3.pl
Search  the high fever




         high=>180                          fever=>180

         high=>104                          fever=>104

         high=>58                           fever=>58

         high=>215                          fever=>215

         high=>3                    fever=>3

         high=>129                          fever=>129

         high=>0                    fever=>0

         high=>129                          fever=>129

         high=>62                           fever=>62

         high=>145                          fever=>145

         high=>126                          fever=>126


the identification table list
2.txt=> 430
1.txt=> 360
8.txt=> 290
6.txt=> 258
4.txt=> 258
9.txt=> 252
10.txt=> 208
7.txt=> 124
11.txt=> 116
3.txt=> 6
5.txt=> 0
```



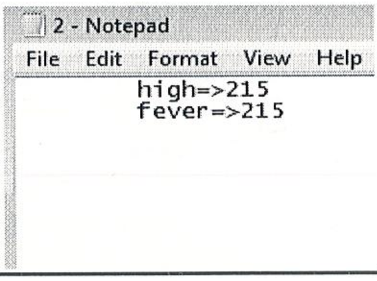
```
Microsoft Windows [Version 6.0.6000]
Copyright (c) 2006 Microsoft Corporation.  All rights reserved.

C:\Users\vishal>cd Desktop

C:\Users\vishal\Desktop>cd pro1

C:\Users\vishal\Desktop\pro1>pro3.pl
Search  the Jaundice




        Jaundice=>180

        Jaundice=>104

        Jaundice=>58

        Jaundice=>215

        Jaundice=>3

        Jaundice=>129

        Jaundice=>0

        Jaundice=>129

        Jaundice=>62

        Jaundice=>145

        Jaundice=>126

the identification table list
2.txt=> 215
1.txt=> 180
8.txt=> 145
6.txt=> 129
4.txt=> 129
9.txt=> 126
10.txt=> 104
7.txt=> 62
11.txt=> 58
3.txt=> 3
5.txt=> 0_
```

7.1.2 Results Directory –

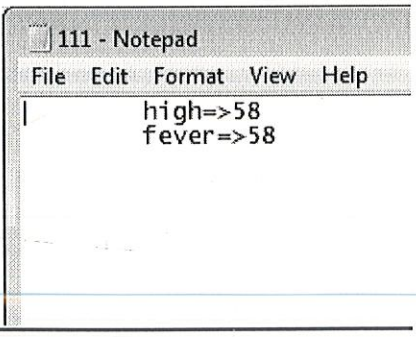




```
high=>180
fever=>180
```



```
high=>215
fever=>215
```



```
high=>58
fever=>58
```

7.2 Perl Programming with HTML –

7.2.1 Main HTML page –

```
<html>

<head><title>Bio Search Engine</head></title>

<body bgcolor=#666633>

<img src="search.gif"><br>

<font size=6 color=FF9900> <center> <marquee behavior="alternate">A inverted index tool
for bio-medical text mining....</marquee></font> </center>

<a href="page.html"><img src="srch.gif"></a>&nbsp&nbsp&nbsp&nbsp&nbsp

<a href=""><img src="defination.gif"></a>&nbsp&nbsp&nbsp&nbsp&nbsp

<a href=""><img src="line.gif"></a>&nbsp&nbsp&nbsp&nbsp&nbsp

<a href=""><img src="sym.gif"></a><br><br>

&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp

<form action="http://localhost/cgi-bin/pro_v.cgi" method="get">

<input type="text" name="text" size="90"/ ></p>

<input type="submit" value="search"/>

</form>

</body>

</html>
```

SEARCH PAGE –

```
<html>

<head><title>Bio Search Engine</head></title>

<body bgcolor=CC9900>

<marquee behavior="alternate"><img src="juit1.gif"></marquee><br>

<img src="search.gif"><br>

<font size=6 color=990000> <center> <marquee behavior="alternate">A inverted index tool for bio-medical text mining...</marquee></font> </center>

<a href="page1.html"><img src="srch.gif"></a>&nbsp

<a href="define.html"><img src="definition.gif"></a>&nbsp

<a href="line.html"><img src="line.gif"></a>&nbsp

<a href="sym.html"><img src="sym.gif"></a>&nbsp

<a href="contact.html"><img src="contact.gif"></a><br><br>

<form action="http://localhost/cgi-bin/vishal/pro_v.cgi" method="get">

<input type="text" name="text" size="90"/ ></p>

<input type="submit" value="search"/>

</form>

</body>

</html>
```

DEFINITION SEARCH –

```
<html>

<head><title>Bio Search Engine</head></title>

<body bgcolor=CC9900>

<marquee behavior="alternate"><img src="juit1.gif"></marquee><br>

<img src="search.gif"><br>

<font size=6 color=990000> <center> <marquee behavior="alternate">A inverted index tool for
bio-medical text mining...</marquee></font> </center>

<a href="page1.html"><img src="srch.gif"></a>&nbsp

<a href="define.html"><img src="definition.gif"></a>&nbsp

<a href="line.html"><img src="line.gif"></a>&nbsp

<a href="sym.html"><img src="sym.gif"></a>&nbsp

<a href="contact.html"><img src="contact.gif"></a><br><br>

<form action="http://localhost/cgi-bin/vishal/vis_define.cgi" method="get">

<p><font size=4 color=990000><b>D</b>efine</font><input type="text" name="text"
size="90"/ ></p>

<input type="submit" value="search"/>

</form>

</body>

</html>
```

SEARCH BY LINE –

```
<html>
<head><title>Bio Search Engine</head></title>
<body bgcolor=CC9900>
<marquee behavior="alternate"><img src="juit1.gif"></marquee><br>
<img src="search.gif"><br>
<font size=6 color=990000> <center> <marquee behavior="alternate">A inverted index tool for bio-medical text mining...</marquee></font> </center>
<a href="page1.html"><img src="srch.gif"></a>&nbsp
<a href="define.html"><img src="definition.gif"></a>&nbsp
<a href="line.html"><img src="line.gif"></a>&nbsp
<a href="sym.html"><img src="sym.gif"></a>&nbsp
<a href="contact.html"><img src="contact.gif"></a><br><br>
<form action="http://localhost/cgi-bin/vishal/vis_line.cgi" method="get">
<p><font size=4 color=990000><b>S</b>earch By Line</font><input type="text" name="text" size="90"/ ></p>
<input type="submit" value="search"/>
</form>
</body>
</html>
```

SEARCH BY SYMPTOMS-

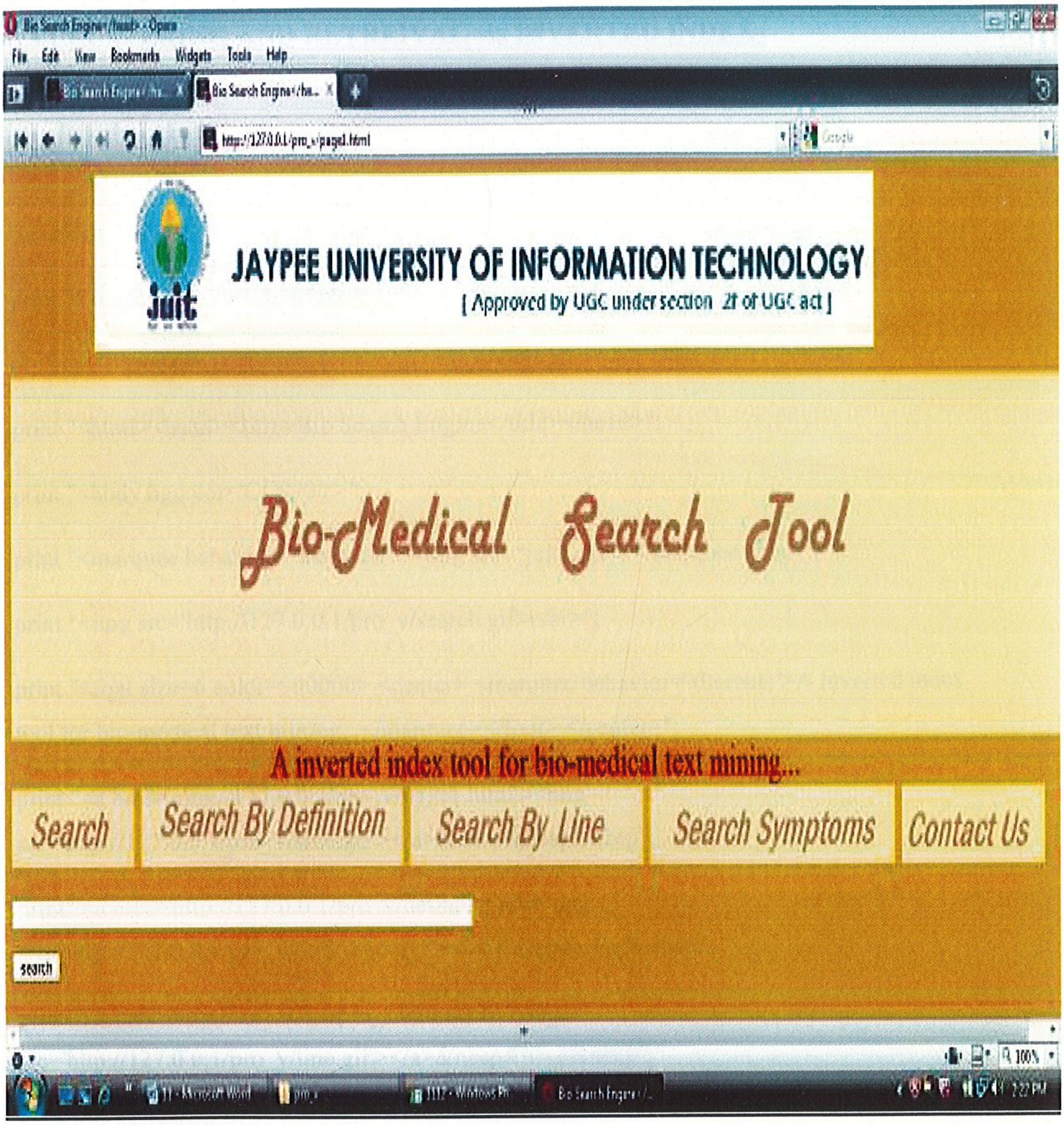
```
<html>

<head><title>Bio Search Engine</head></title>

<body bgcolor=CC9900>

<marquee behavior="alternate"><img src="juit1.gif"></marquee><br>

<img src="search.gif"><br>

<font size=6 color=990000> <center> <marquee behavior="alternate">A inverted index tool for
bio-medical text mining...</marquee></font> </center>

<a href="page1.html"><img src="srch.gif"></a>&nbsp

<a href="define.html"><img src="definition.gif"></a>&nbsp

<a href="line.html"><img src="line.gif"></a>&nbsp

<a href="sym.html"><img src="sym.gif"></a>&nbsp

<a href="contact.html"><img src="contact.gif"></a><br><br>

<form action="http://localhost/cgi-bin/vishal/vis_sym.cgi" method="get">

<p><font size=4 color=990000><b>S</b>ymptoms</font><input type="text" name="text"
size="90"/ ></p>

<input type="submit" value="search"/>

</form>

</body>

</html>
```

CONTACT US PAGE –

```
<html>

<head><title>Bio Search Engine</head></title>

<body bgcolor=CC9900>

<marquee behavior="alternate"><img src="juit1.gif"></marquee><br>

<img src="search.gif"><br>

<font size=6 color=990000> <center> <marquee behavior="alternate">A inverted index tool for
bio-medical text mining...</marquee></font> </center>

<a href="page1.html"><img src="srch.gif"></a>&nbsp

<a href="define.html"><img src="definition.gif"></a>&nbsp

<a href="line.html"><img src="line.gif"></a>&nbsp

<a href="sym.html"><img src="sym.gif"></a>&nbsp

<a href="contact.html"><img src="contact.gif"></a><br><br>

<p style="font-size:20px;color:990000;font-family:verdana">Name:<font color= white>Vishal
Mahajan </font> <br> Enrollment no: <font color= white>071509 </font><br> Emailid: <font
color= white><abbr title="world wide web"><i>
vishalmahajan06@gmail.com</i></font></abbr></p>

&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&

<img src="logo1.gif"> <br><br>

<p style="font-size:20px;color:990000;font-family:verdana">Name:<font color= white>Arjun
Bagai </font> <br> Enrollment no: <font color= white>071516 </font><br> Emailid: <font
color= white><abbr title="world wide web"><i>
arjun.bagai@gmail.com</font></i></abbr></p><br><br>
```

PAGE –

```
<html>

<head><title>Bio Search Engine</head></title>

<body bgcolor=CC9900>

<marquee behavior="alternate"><img src="juit1.gif"></marquee><br>

<img src="search.gif"><br>

<font size=6 color=990000> <center> <marquee behavior="alternate">A inverted index tool for bio-medical text mining...</marquee></font> </center>

<a href="page1.html"><img src="srch.gif"></a>&nbsp

<a href="define.html"><img src="definition.gif"></a>&nbsp

<a href="line.html"><img src="line.gif"></a>&nbsp

<a href="sym.html"><img src="sym.gif"></a>&nbsp

<a href="contact.html"><img src="contact.gif"></a><br><br>

<form action="http://localhost/cgi-bin/vishal/pro_v.cgi" method="get">

<input type="text" name="text" size="90"/ ></p>

<input type="submit" value="search"/>

</form>

</body>

</html>
```



JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY

[Approved by UGC under section 2f of UGC act]

Bio-Medical Search Tool

A inverted index tool for bio-medical text mining...

7.3 CLUSTERING AND RETREIVAL via INVERTED INDEX-

```
#!c:/Perl/bin/perl

use CGI;

use strict;

print "Content-type: text/html\n\n";

my $cgi=new CGI;

print "<html><head><title>Bio Search Engine</title></head>";

print "<body bgcolor='CC9900'>";

print "<marquee behavior="alternate"><img src="juit1.gif"></marquee><br>";

print "<img src='http://127.0.0.1/pro_v/search.gif'><br>";

print "<font size=6 color=990000> <center> <marquee behavior='alternate'>A inverted index
tool for bio-medical text mining...</marquee></font> </center>";

print "<a href='http://127.0.0.1/pro_v/page1.html'><img
src='http://127.0.0.1/pro_v/srch.gif'></a>&nbsp&nbsp&nbsp";

print "<a href='http://127.0.0.1/pro_v/define.html'><img
src='http://127.0.0.1/pro_v/definition.gif'></a>&nbsp&nbsp&nbsp";

print "<a href='http://127.0.0.1/pro_v/line.html'><img
src='http://127.0.0.1/pro_v/line.gif'></a>&nbsp&nbsp&nbsp";

print "<a href='http://127.0.0.1/pro_v/sym.html'><img
src='http://127.0.0.1/pro_v/sym.gif'></a>&nbsp&nbsp&nbsp";

print "<a href='http://127.0.0.1/pro_v/contact.html'><img
src='http://127.0.0.1/pro_v/contact.gif'></a><br><br><br><br>";
```

```
my $search=$cgi->param("text");

$search=lc($search);

#print "<p>here1 $search</p>";

my @stoplist=qw(is an the with or between are not has had but can many what when who in it
where when why on );

my @search=split('\s+',$search);

for(my $i=0;$i<scalar(@stoplist);$i++)

{

        for(my $j=0;$j<scalar(@search);$j++)

        {

                if ($stoplist[$i] eq $search[$j])

                {

                        splice(@search,$j,1);

}}}

#foreach my $x(@search)

#{

#print "<p>here2 $x</p>";

#}

my $dirtoget="C:/Users/vishal/Desktop/pro1/prj4";

opendir(prj4, $dirtoget) || die("Cannot open directory");

my @thefiles= readdir(prj4);
```

```
#foreach my $y(@thefiles)

#{

#print "<p>here3 $y</p>";

#}

closedir(prj4);

my %it=();

foreach my $f (@thefiles)

{

        unless ( ($f eq ".") || ($f eq "..") )

        {

                my %l=();

                open(F1,"C:/Users/vishal/Desktop/pro1/prj4/$f");

                my @data=<F1>;

                close (F;

                for(my $i=0;$i<scalar(@search);$i++)

                {

                        $l{$search[$i]}=0;

                        my $ref=0;

                        foreach my $data1(@data)

                        {

                                my @data1=split('\s+',$data1);
```

```
            foreach my $x(@data1)
                {
                $x=lc($x);
                #print "<p>here4 $x</p>";
                    if($x=~/$search[$i]/g)
                    {
                        #print "<p>here5</p>";
                        $ref++;
                }}}
            $l{$search[$i]}=$ref;
            if($l{$search[$i]} eq '')
            {
            #   print "<p>here 6</p>";
                $l{$search[$i]}=0;
            }
        #print "<p>&nbsp&nbsp&nbsp&nbsp
$search[$i]=>$l{$search[$i]}&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp
&nbsp&nbsp&nbsp&nbsp&nbsp&nbsp</p>";
        }
        my $total=0;
        foreach my $key(keys%l)
```

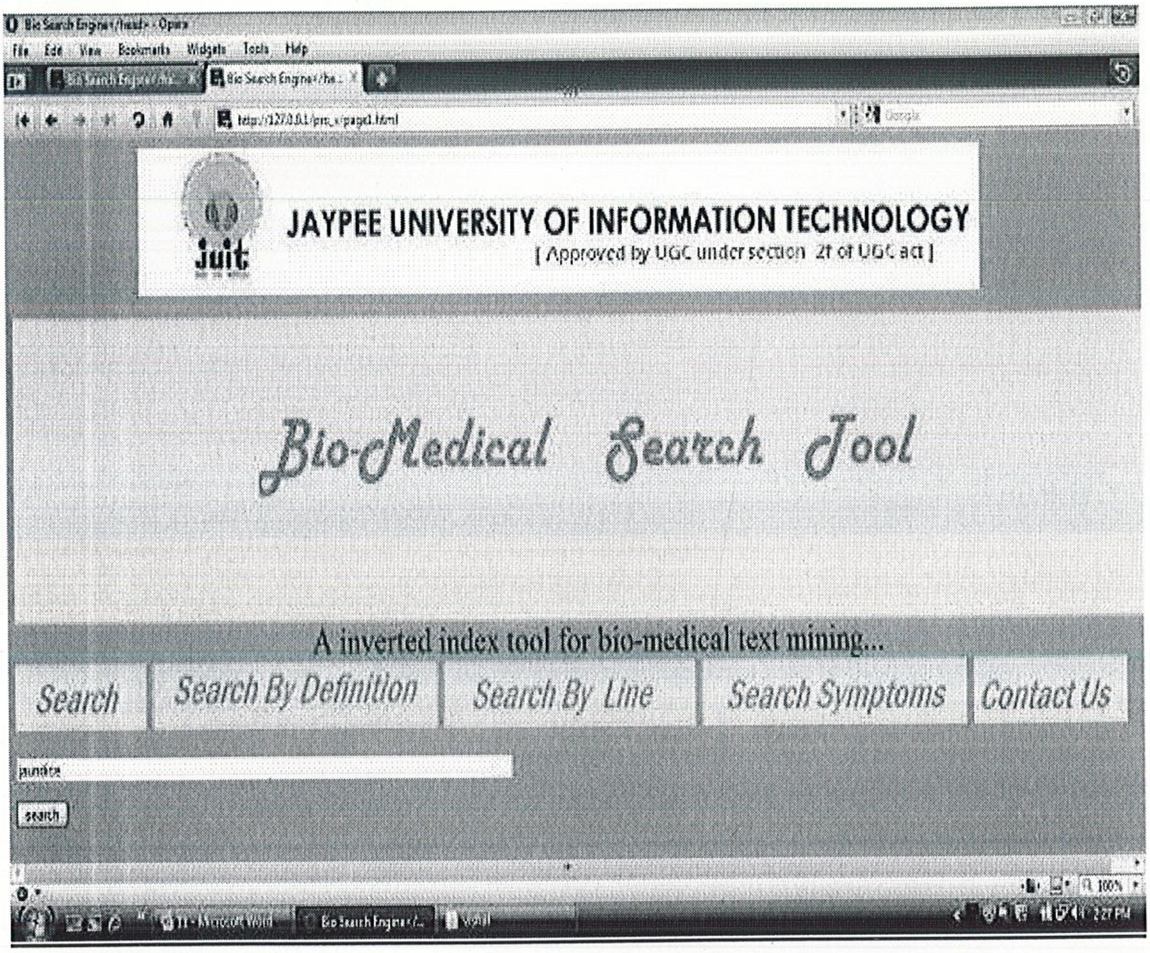
```
                {
                        $total=$total+$l{$key};
                }
                $it{$f}=$total;
                }
        print "<p></p><p></p><p></p><p></p>"
}
print "<table border=1 cellspacing=50 cellpadding=10>";
print "<tr align='center' valing='top'>";
print "<th><font color=990000><h2>Rank</font></h2></th>";
print "<th><font color=990000><h2>File Name</font></h2></th>";
print "<th><font color=990000><h2>weights</font></h2></th>";
print "</tr>";
my $rank=1;
foreach my $key(sort {$it{$b} <=> $it{$a}} keys %it)
{
        if($it{$key}!=0)
        {
        print "<tr align='center' valign='center'>";
        print "<td><font color=white>$rank</font></td>";
        print "<td><a href='http://127.0.0.1/pro_v/prj4/$key'>$key </a></td>";
```

```
          print "<td><font color=white>$it{$key}</font></td>";

          $rank++;

          }}

 print"</table>";

print "</body></html>";
```

Chapter 8 –

RESULTS AND CONCLUSIONS

8.1 Searching for a disease

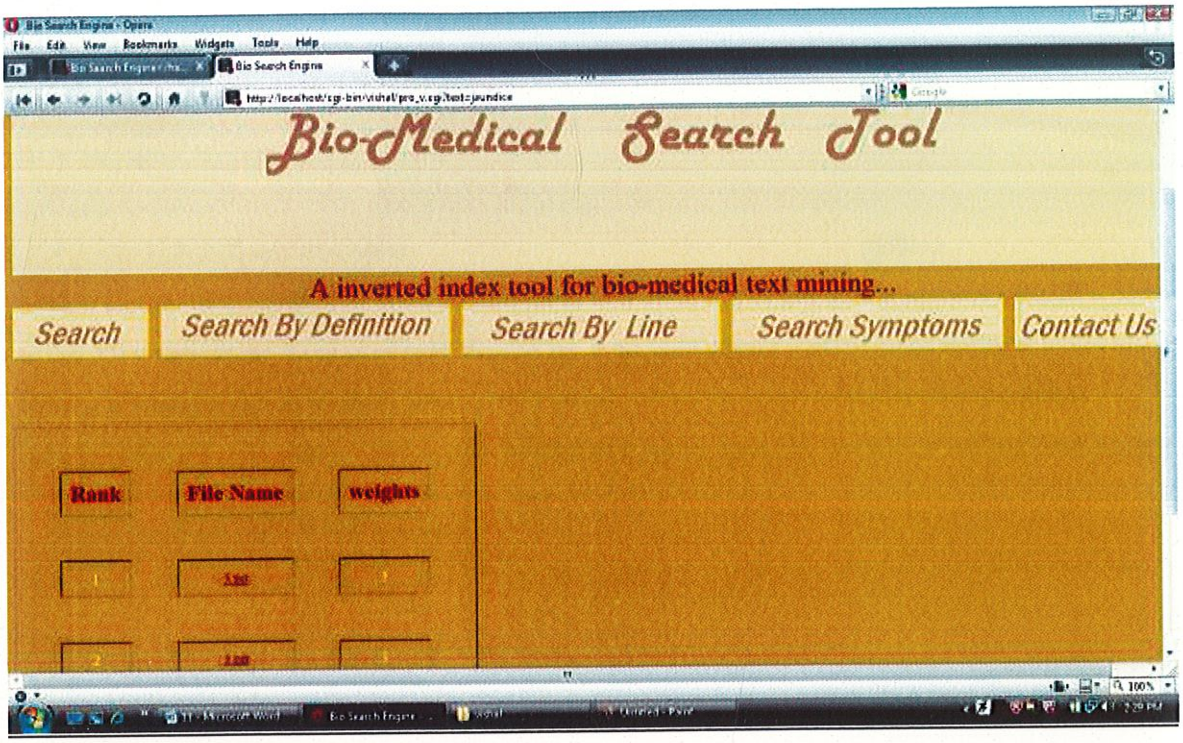


JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY

[Approved by UGC under section 2f of UGC act]

Bio-Medical Search Tool

A inverted index tool for bio-medical text mining...



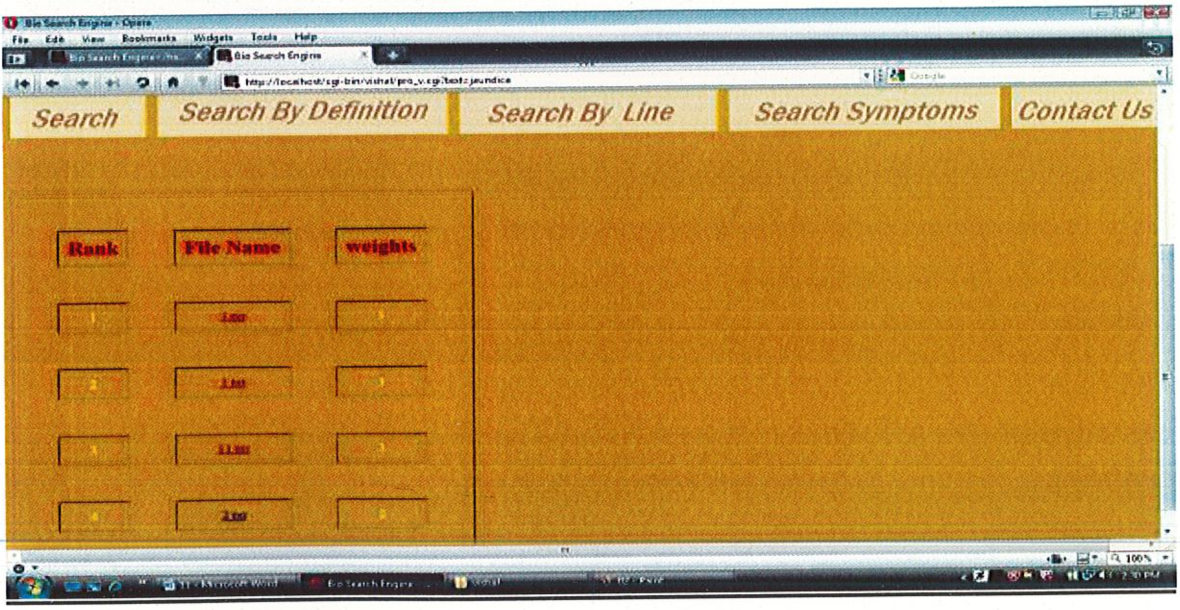
Bio Search Engine - Opera
File Edit View Bookmarks Widgets Tools Help
Bio Search Engine
http://localhost/cgi-bin/vishal/pro_v.cgi?text=jaundice
Google

Bio-Medical Search Tool

A inverted index tool for bio-medical text mining...

Search
Search By Definition
Search By Line
Search Symptoms
Contact Us
100%

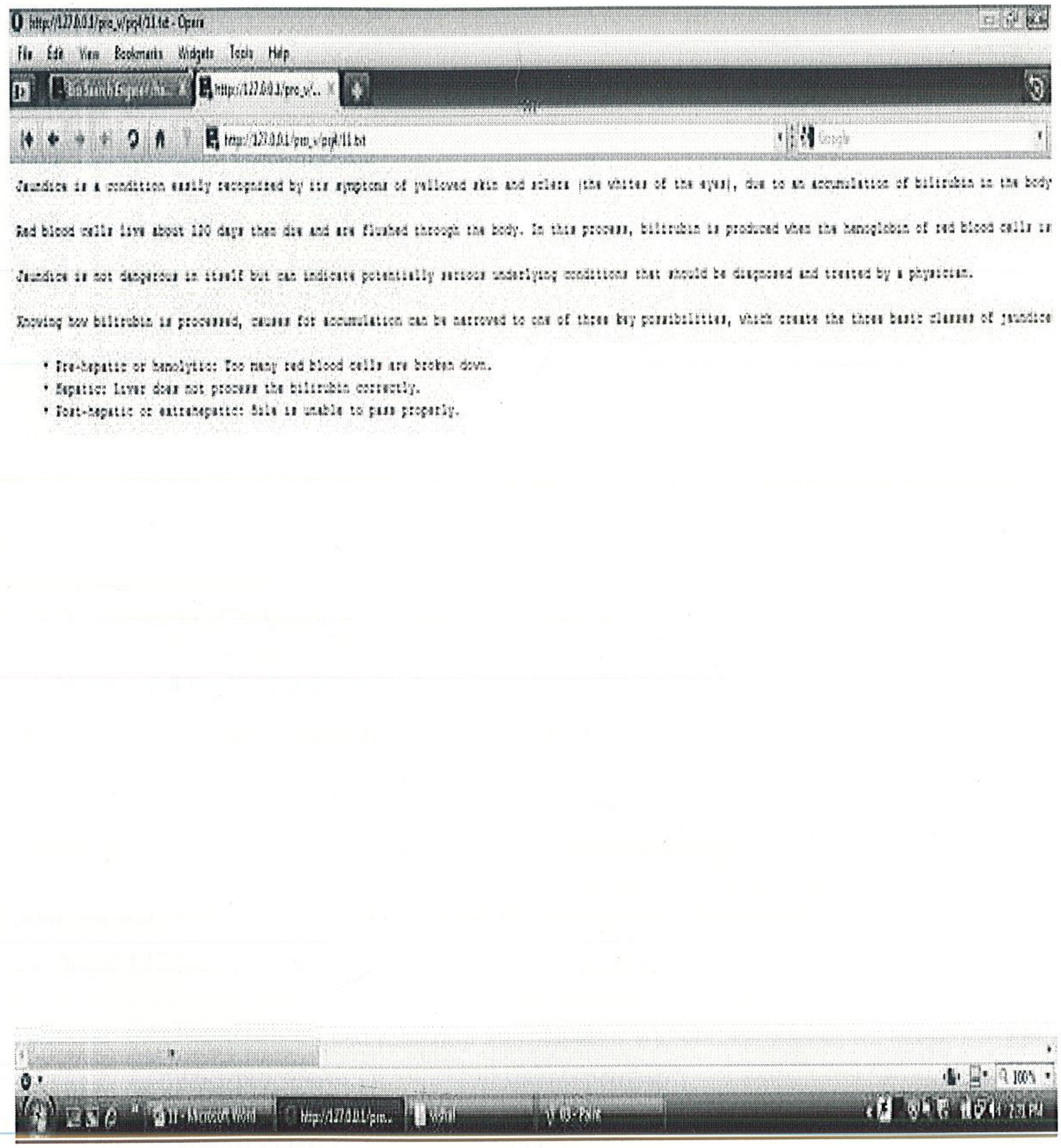
Rank	File Name	weights
1	3.txt	1
2	2.txt	1



Bio Search Engine - Opera
File Edit View Bookmarks Widgets Tools Help
Bio Search Engine
http://localhost/cgi-bin/vishal/pro_v.cgi?text=jaundice
Google
Search
Search By Definition
Search By Line
Search Symptoms
Contact Us
100%

Rank	File Name	weights
1	4.txt	1
2	3.txt	1
3	11.txt	1
4	2.txt	1

<u>So the best retrieved document is read, henceforth-</u>


http://127.0.0.1/pro_v/pq4/11.txt - Opera
File Edit View Bookmarks Widgets Tools Help
http://127.0.0.1/pro_v/...
http://127.0.0.1/pro_v/pq4/11.txt
Google
100%
http://127.0.0.1/pro...

Jaundice is a condition easily recognized by its symptoms of yellowed skin and sclera (the whites of the eyes), due to an accumulation of bilirubin in the body

Red blood cells live about 120 days then die and are flushed through the body. In this process, bilirubin is produced when the hemoglobin of red blood cells ar

Jaundice is not dangerous in itself but can indicate potentially serious underlying conditions that should be diagnosed and treated by a physician.

Knowing how bilirubin is processed, causes for accumulation can be narrowed to one of three key possibilities, which create the three basic classes of jaundice

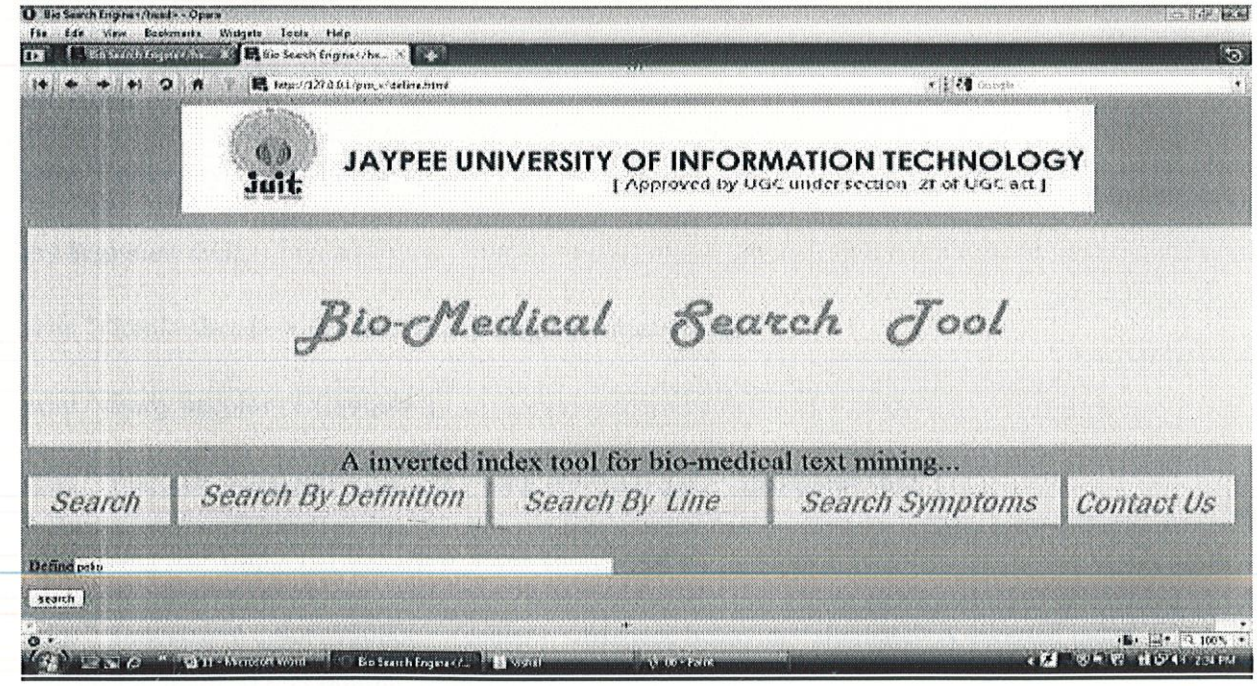
- Pre-hepatic or hemolytic: Too many red blood cells are broken down.
- Hepatic: Liver does not process the bilirubin correctly.
- Post-hepatic or extrahepatic: Bile is unable to pass properly.

SEARCH BY DEFINITON –

```
#!c:/Perl/bin/perl

use CGI;

use strict;

print "Content-type: text/html\n\n";

my $cgi=new CGI;

print "<html><head><title>Bio Search Engine</title></head>";

print "<body bgcolor='CC9900'>";

print "<marquee behavior='alternate'><img src='juit1.gif'></marquee><br>";

print "<img src='http://127.0.0.1/pro_v/search.gif'><br>";

print "<font size=6 color=990000> <center> <marquee behavior='alternate'>A inverted index
tool for bio-medical text mining...</marquee></font> </center>";

print "<a href='http://127.0.0.1/pro_v/page1.html'><img
src='http://127.0.0.1/pro_v/srch.gif'></a>&nbsp&nbsp&nbsp";

print "<a href='http://127.0.0.1/pro_v/define.html'><img
src='http://127.0.0.1/pro_v/definition.gif'></a>&nbsp&nbsp&nbsp";

print "<a href='http://127.0.0.1/pro_v/line.html'><img
src='http://127.0.0.1/pro_v/line.gif'></a>&nbsp&nbsp&nbsp";

print "<a href='http://127.0.0.1/pro_v/sym.html'><img
src='http://127.0.0.1/pro_v/sym.gif'></a>&nbsp&nbsp&nbsp";

print "<a href='http://127.0.0.1/pro_v/contact.html'><img
src='http://127.0.0.1/pro_v/contact.gif'></a><br><br><br><br>";
```

```
my $search=$cgi->param("text");

$search=lc($search);

$search=$search.".txt";

open(F1,"C:/Users/vishal/Desktop/pro1/definitions/$search");

my @data=<F1>;

foreach my $d(@data)

{         print "<p><font size=4 color=990000>$d</p>";

}

print "</body></html>";
```

8.2 Searching for the definition

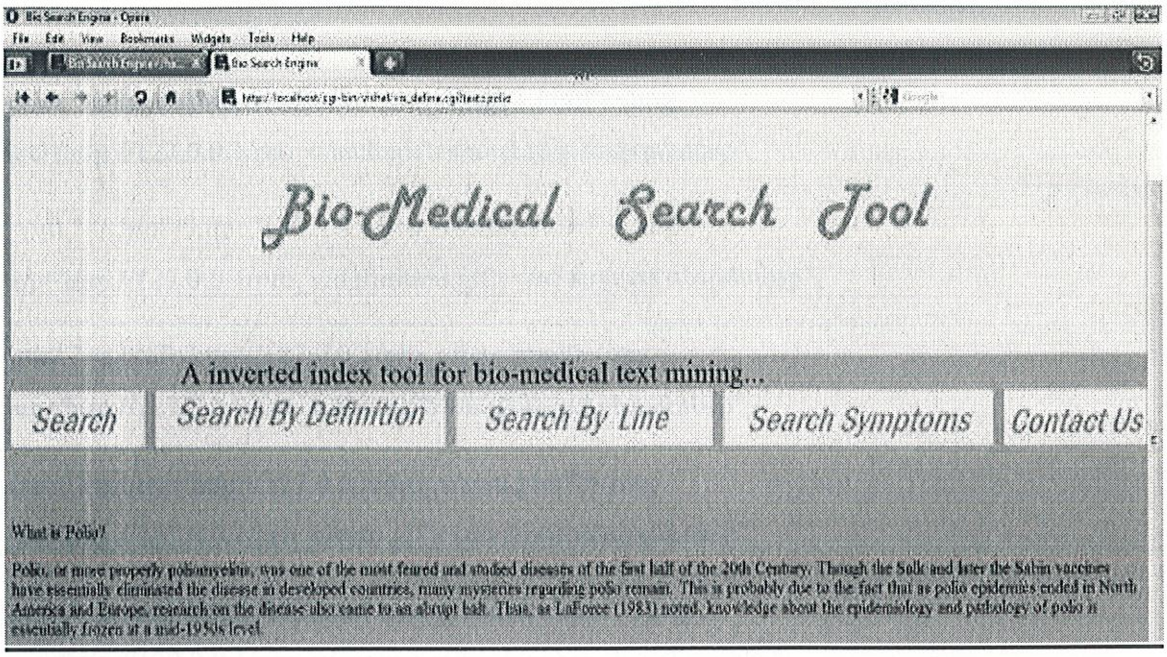


JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY

[Approved by UGC under section 2f of UGC act]

Bio-Medical Search Tool

A inverted index tool for bio-medical text mining...


Bio Search Engine - Opera
File Edit View Bookmarks Widgets Tools Help
Bio Search Engine
http://localhost/cgi-bin/vishal/vis_define.cgi?text=polio
Google
Search
Search By Definition
Search By Line
Search Symptoms
Contact Us

Bio-Medical Search Tool

A inverted index tool for bio-medical text mining...

What is Polio?

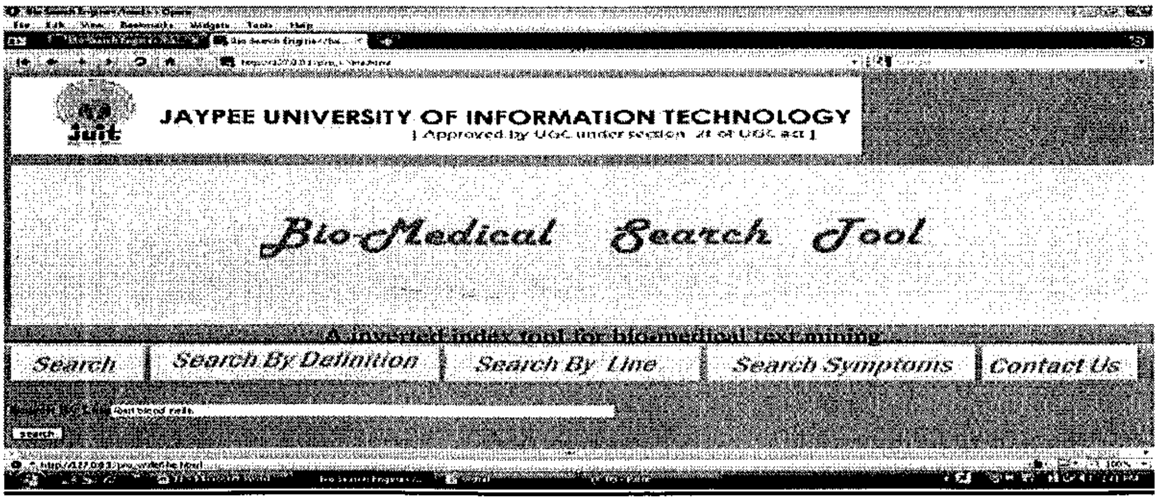
Polio, or more properly poliomyelitis, was one of the most feared and studied diseases of the first half of the 20th Century. Though the Salk and later the Sabin vaccines have essentially eliminated the disease in developed countries, many mysteries regarding polio remain. This is probably due to the fact that as polio epidemics ended in North America and Europe, research on the disease also came to an abrupt halt. Thus, as LaForce (1983) noted, knowledge about the epidemiology and pathology of polio is essentially frozen at a mid-1950s level.

SEARCH BY LINE –

```
#!c:/Perl/bin/perl

use CGI;

use strict;

print "Content-type: text/html\n\n";

my $cgi=new CGI;

print "<html><head><title>Bio Search Engine</title></head>";

print "<body bgcolor='CC9900'>";

print "<marquee behavior='alternate'><img src='juit1.gif'></marquee><br>";

print "<img src='http://127.0.0.1/pro_v/search.gif'><br>";

print "<font size=6 color=990000> <center> <marquee behavior='alternate'>A inverted index
tool for bio-medical text mining...</marquee></font> </center>";
```

```
print "<a href='http://127.0.0.1/pro_v/page1.html'><img
src='http://127.0.0.1/pro_v/srch.gif'></a>&nbsp&nbsp&nbsp";

print "<a href='http://127.0.0.1/pro_v/define.html'><img
src='http://127.0.0.1/pro_v/definition.gif'></a>&nbsp&nbsp&nbsp";

print "<a href='http://127.0.0.1/pro_v/line.html'><img
src='http://127.0.0.1/pro_v/line.gif'></a>&nbsp&nbsp&nbsp";

print "<a href='http://127.0.0.1/pro_v/sym.html'><img
src='http://127.0.0.1/pro_v/sym.gif'></a>&nbsp&nbsp&nbsp";

print "<a href='http://127.0.0.1/pro_v/contact.html'><img
src='http://127.0.0.1/pro_v/contact.gif'></a><br><br><br><br>";

my $search=$cgi->param("text");

my $dirtoget="C:/Users/vishal/Desktop/pro1/prj4";

opendir(prj4, $dirtoget) || die("Cannot open directory");

my @thefiles= readdir(prj4);

closedir(prj4);

my %it=();

foreach my $f (@thefiles)

{

        unless ( ($f eq ".") || ($f eq "..") )

        {

                open(F1,"C:/Users/vishal/Desktop/pro1/prj4/$f");
        my @data=<F1>;

        close (F1);

        my $ref=0;

        foreach my $data1(@data)

        {

            if($data1=~/$search/g)

                {$ref++;

            }

        }

        $it{$f}=$ref;

    }

}

print "<table border=1 cellspacing=50 cellpadding=10>";

print "<tr align='center' valing='top'>";

print "<th><font color=990000><h2>Rank</font></h2></th>";

print "<th><font color=990000><h2>File Name</font></h2></th>";

print "<th><font color=990000><h2>Weights</font></h2></th>";

print "</tr>";

my $rank=1;

foreach my $key(sort {$it{$b} <=> $it{$a}} keys %it)

{
        if($it{$key}!=0)

        {

        print "<tr align='center' valign='center'>";

        print "<td><font color=white>$rank</font></td>";

      print "<td><a href='http://127.0.0.1/pro_v/prj4/$key'>$key </a></td>";

        print "<td><font color=white>$it{$key}</font></td>";


        $rank++;

        }

}

 print"</table>";

print "</body></html>";
```

8.3 Searching by Line



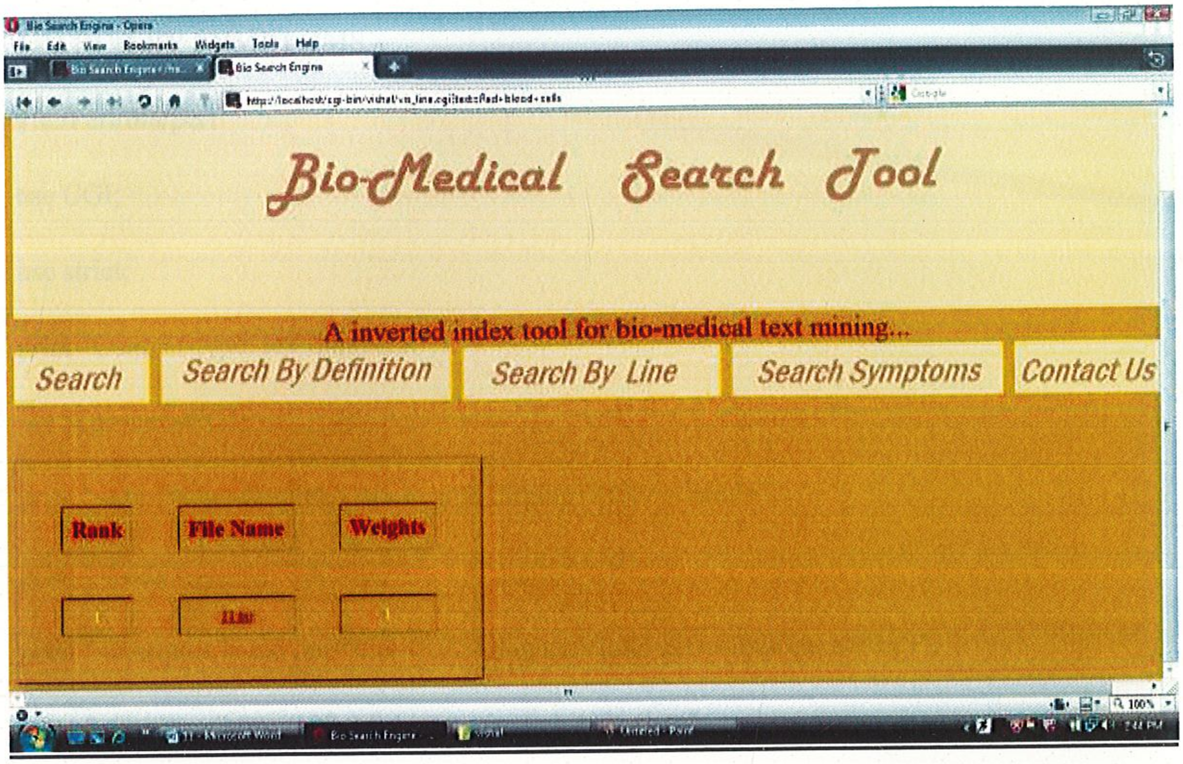
JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY

[Approved by UGC under section 2f of UGC act]

Bio-Medical Search Tool

A inverted index tool for biomedical text mining

Search By Line: Red blood cells

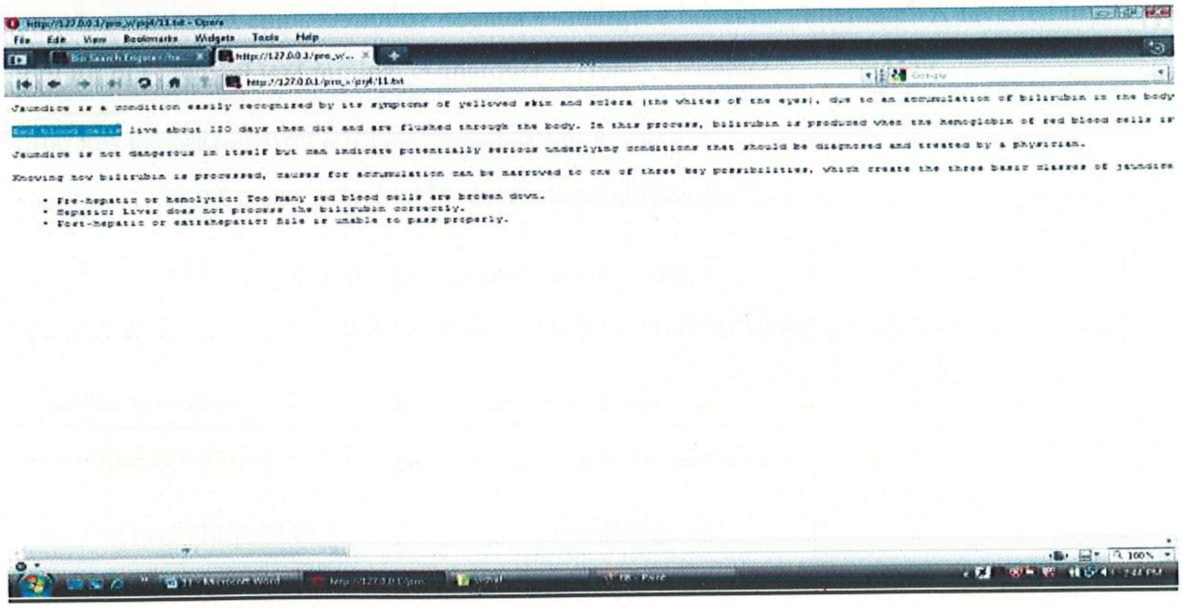


Bio Search Engine - Opera
File Edit View Bookmarks Widgets Tools Help
Bio Search Engine
http://localhost/cgi-bin/vishal/vis_line.cgi?text=Red+blood+cells
Search
Search By Definition
Search By Line
Search Symptoms
Contact Us
11 - Microsoft Word
Bio Search Engine
100%

Bio-Medical Search Tool

A inverted index tool for bio-medical text mining...

Rank	File Name	Weights
1	11.txt	1



http://127.0.0.1/pro_v/prj4/11.txt - Opera
File Edit View Bookmarks Widgets Tools Help
Bio Search Engine
http://127.0.0.1/pro_v/...
http://127.0.0.1/pro_v/prj4/11.txt
11 - Microsoft Word
100%

Jaundice is a condition easily recognized by its symptoms of yellowed skin and sclera (the whites of the eyes), due to an accumulation of bilirubin in the body

Red blood cells live about 120 days then die and are flushed through the body. In this process, bilirubin is produced when the hemoglobin of red blood cells is

Jaundice is not dangerous in itself but can indicate potentially serious underlying conditions that should be diagnosed and treated by a physician.

Knowing how bilirubin is processed, causes for accumulation can be narrowed to one of three key possibilities, which create the three basic classes of jaundice

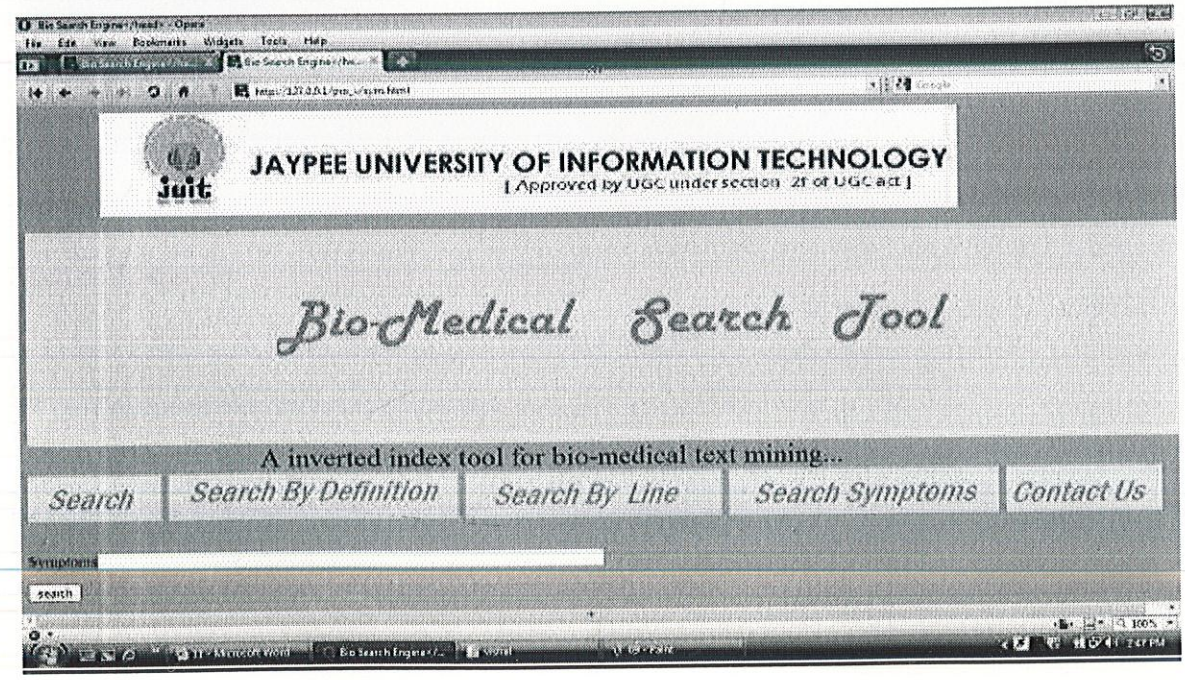
- Pre-hepatic or hemolytic: Too many red blood cells are broken down.
- Hepatic: Liver does not process the bilirubin correctly.
- Post-hepatic or extrahepatic: Bile is unable to pass properly.

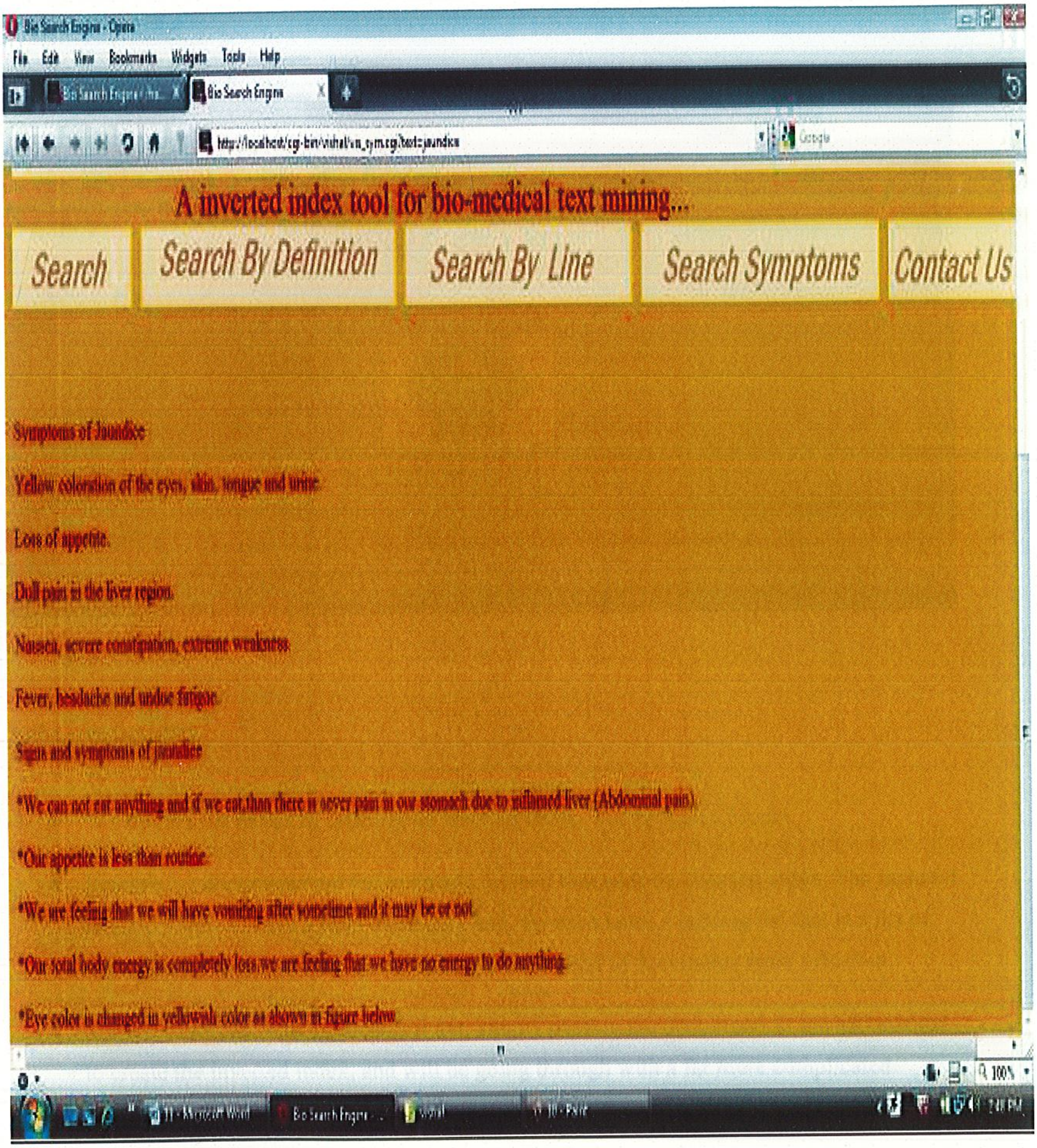
SEARCH BY SYMPTOMS –

```
#!c:/Perl/bin/perl

use CGI;

use strict;

print "Content-type: text/html\n\n";

my $cgi=new CGI;

print "<html><head><title>Bio Search Engine</title></head>";

print "<body bgcolor='CC9900'>";

print "<marquee behavior='alternate'><img src='juit1.gif'></marquee><br>";

print "<img src='http://127.0.0.1/pro_v/search.gif'><br>";

print "<font size=6 color=990000> <center> <marquee behavior='alternate'>A inverted index
tool for bio-medical text mining...</marquee></font> </center>";

print "<a href='http://127.0.0.1/pro_v/page1.html'><img
src='http://127.0.0.1/pro_v/srch.gif'></a>&nbsp&nbsp&nbsp";

print "<a href='http://127.0.0.1/pro_v/define.html'><img
src='http://127.0.0.1/pro_v/definition.gif'></a>&nbsp&nbsp&nbsp";

print "<a href='http://127.0.0.1/pro_v/line.html'><img
src='http://127.0.0.1/pro_v/line.gif'></a>&nbsp&nbsp&nbsp";

print "<a href='http://127.0.0.1/pro_v/sym.html'><img
src='http://127.0.0.1/pro_v/sym.gif'></a>&nbsp&nbsp&nbsp";

print "<a href='http://127.0.0.1/pro_v/contact.html'><img
src='http://127.0.0.1/pro_v/contact.gif'></a><br><br><br><br>";

my $search=$cgi->param("text");
```

```
$search=lc($search);

$search=$search.".txt";

open(F1,"C:/Users/vishal/Desktop/pro1/symptoms/$search");

my @data=<F1>;

foreach my $d(@data)

{

        print "<p><font size=4 color=990000>$d</p>";

}

print "</body></html>";
```

8.4 Searching for Symptoms


JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY
[Approved by UGC under section 2f of UGC act]
juit
Bio-Medical Search Tool
A inverted index tool for bio-medical text mining...
Search
Search By Definition
Search By Line
Search Symptoms
Contact Us
Symptoms
search



Bio Search Engine - Opera
File Edit View Bookmarks Widgets Tools Help
Bio Search Engine / ...
Bio Search Engine
http://localhost/cgi-bin/vishal/vs_sym.cgi?text=jaundice
Google
A inverted index tool for bio-medical text mining...
Search
Search By Definition
Search By Line
Search Symptoms
Contact Us
100%
11 - Microsoft Word
Bio Search Engine - ...
vishal
10 - Paint

Symptoms of Jaundice

Yellow coloration of the eyes, skin, tongue and urine.

Loss of appetite.

Dull pain in the liver region.

Nausea, severe constipation, extreme weakness.

Fever, headache and undue fatigue.

Signs and symptoms of jaundice

*We can not eat anything and if we eat,than there is sever pain in our stomach due to inflamed liver (Abdominal pain).

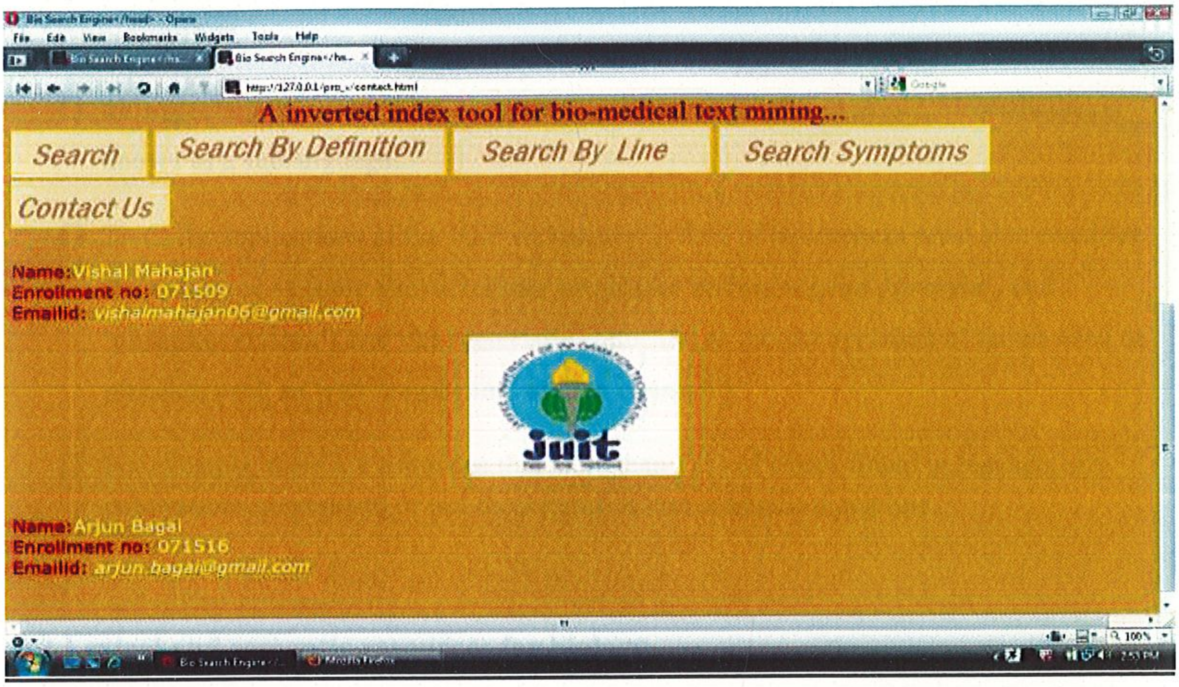
*Our appetite is less than routine.

*We are feeling that we will have vomiting after sometime and it may be or not.

*Our total body energy is completely loss,we are feeling that we have no energy to do anything.

*Eye color is changed in yellowish color as shown in figure below.

CONTACT US –


Bio Search Engine</head> - Opera
File Edit View Bookmarks Widgets Tools Help
http://127.0.0.1/pro_x/contact.html
Search
Search By Definition
Search By Line
Search Symptoms
Contact Us
juit

A inverted index tool for bio-medical text mining...

Name:Vishal Mahajan
Enrollment no: 071509
Emailid: vishalmahajan06@gmail.com

Name:Arjun Bagal
Enrollment no: 071516
Emailid: arjun.bagal@gmail.com

8.5 Conclusions

- We studied the basics of creating an inverted index, Vector Space Model.
- Essentially, documents are fed to an add method that adds an inverted index. The inverted index consists of both a term dictionary and, for each term, a posting list that is a list of documents that contain the term, making the retrieval of documents more effective.
- We went through with an IndexBuilder, simply passing the documents to the add method to create the inverted index, and will be going through with a far more complicated IndexBuilder in the time to come.

Chapter 9 – Future Work of the Project

9.1 Sublanguage

□ This specialized usage of language in a specific domain (for example, biomedicine) is known as a sublanguage.

□ Thus, the applications of the NLP techniques in field of biosciences must also **consider the existence of some characteristic terms (for instance,'gene','protein', and 'phosphorylation') and characteristic collocations (co-occurrences of terms used as phrases, such as 'cell membrane' or 'ion channel').**

□ For example, there are estimates that more than 12% of words found in biochemistry publications correspond to technical terms of that scientific discipline.

9.2 Lexical and semantic resources for biology

□ Functional descriptions of bio-entities, **relevant biological processes, or experimental techniques** are often **expressed in scientific papers using domain-specific technical terms.**

□ ***Terminological repositories*** *and dictionaries are important resources to* ***assist in the interpretation of scientific articles****, assisiingt authors in consistent use of domain specific terminology.*

9.3 Controlled vocabulary

□ So-called contrólled vocabulary is a set of predefined, authorized terms that have been preselected by the designer of the vocabulary. **One of the example is MeSH database, however one of the most popular and widely used controlled vocabulary is Gene Ontology database**, that describes gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner.

Chapter 10

BIBLIOGRAPHY

□ Zobel, Justin RMIT University, Australia; Moffat, Alistair The University of Melbourne, Australia (July 2006). "Inverted Files for Text Search Engines". *ACM Computing Surveys* (New York: Association for Computing Machinery) 38 (2): 6. doi:10.1145/1132956.1132959

□ Baeza-Yates, Ricardo; Ribeiro-Neto, Berthier (1999). *Modern information retrieval.* Reading, Massachusetts: Addison-Wesley Longman. p. 192. ISBN 0-201-39829-X

□ Luk, Robert; W. Lam (2007). "Efficient in-memory extensible inverted file". *Information Systems* 32 (5): 733–754. doi:10.1016/j.is.2006.06.001

□ G. Salton, A. Wong, and C. S. Yang (1975), "A Vector Space Model for Automatic Indexing," *Communications of the ACM*, vol. 18, nr. 11, pages 613–620.

□ Y.L. Lo, C.H. Lee, and C.H. Wang, "Creating multi-feature index structure for database", Journal of Information Sciences, Vol.179, pp.2662-2675, 2009.

□ A paper on Text-mining approaches in molecular biology and biomedicine.

□ Discovering hidden knowledge from biomedical literature.

□ A *text-mining* system for knowledge discovery from *biomedical* documents.

□ www.ncbi.nlm.nih.gov/

□ http://lib.bioinfo.pl/

□ www.biomedcentral.com

□ en.wikipedia.org/wiki/Biomedical_text_mining.