Identification of Disease-Treatment Relationship in Bio-Science Text

A PROJECT REPORT

Submitted by

Kartheek Anumolu
Reg. No. 10MSE1027

in partial fulfillment for the award of the degree of

Master of Science

in

Software Engineering

VIT®
UNIVERSITY
(Estd. u/s 3 of UGC Act 1956)
www.vit.ac.in
Vellore • Chennai
CHENNAI CAMPUS
Vandalur - Kelambakkam Road, Chennai - 600127

School of Computing Science and Engineering
VIT University
Vandalur - Kelambakkam Road, Chennai - 600 127

May - 2015

VIT®
UNIVERSITY
(Estd. u/s 3 of UGC Act 1956)
www.vit.ac.in
Vellore ▪ Chennai
CHENNAI CAMPUS
Vandalur - Kelambakkam Road, Chennai - 600127

School of Computing Science and Engineering

DECLARATION

I hereby declare that the project entitled **Identification of Disease-Treatment Relationship in Bio-Science Text** submitted by me to the School of Computing Science and Engineering, VIT Chennai, 600 127 in partial fulfillment of the requirements of the award of the degree of **Master of Science** in **Software Engineering** is a bona-fide record of the work carried out by me under the supervision of **Prof. A Muralidhar**. I further declare that the work reported in this project, has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma of this institute or of any other institute or University.

Place: Chennai	Signature of Candidate
Date:	(Kartheek Anumolu)

VIT®
UNIVERSITY
(Estd. u/s 3 of UGC Act 1956)
www.vit.ac.in
Vellore ▪ Chennai
CHENNAI CAMPUS
Vandalur - Kelambakkam Road, Chennai - 600127

School of Computing Science and Engineering

CERTIFICATE

This is to certify that the report entitled **Identification of Disease-Treatment Relationship in Bio-Science Text** is prepared and submitted by **Kartheek Anumolu (Reg. No. 10MSE1027)** to VIT Chennai, in partial fulfullment of the requirement for the award of the degree of **Master of Science** in **Software Engineering** is a bona-fide record carried out under my guidance. The project fulfills the requirements as per the regulations of this University and in my opinion meets the necessary standards for submission. The contents of this report have not been submitted and will not be submitted either in part or in full, for the award of any other degree or diploma and the same is certified.

Guide/Supervisor

Name: Prof. A Muralidhar
Date:

Program Chair

Name: Dr. N.Maheswari
Date:

Examiner

Name:
Date:

Examiner

Name:
Date:

(Seal of SCSE)

Acknowledgement

I would like to extend my gratitude towards Prof. A Muralidhar (School of Computer Science and Engineering), it was under his guidance and supervision that I was able to complete my project with a satisfactory note. I would like to thank Prof. N. Maheswari (Program Manager, MS Software Engineering) and Dr.L.Jeganathan (Professor and Dean, School of Computer Science and Engineering) , VIT University, Chennai for their enormous support and advice which acted as a catalyst in successful completion of my project.

Kartheek Anumolu
Reg.No.10MSE1027

Abstract

People are very much cautious of their health and want to be well informed about the recent advancements in medical domain.There is an enormous increase in number of people searching disease related information over web. Health related information are very sensitive to trust and standardizing the data would increase the truthfulness of the results. There is a need to develop a tool that can assist people in finding the trusted information.One of the good ideas would be to integrate machine learning tools and Natural language processing techniques into medical domain to yield better results and user experience. My application can extract disease-treatment relationships from published Medline abstracts that are really sensible to trust. After extracting the data this application provides the users with an interface to select the disease and find its related treatment and also information about the treatment if the user has no idea regarding it

Contents

Declaration **i**

Certificate **i**

Acknowledgement **iv**

Abstract **v**

1 Introduction **1**

1.1 Background 1

1.2 Motivation 2

1.3 Objective 2

1.4 Challenges 2

1.5 Essence of the Approach 3

1.6 About MedLine 3

2 Overview / Literature Review **5**

2.1 Existing Methods 5

2.1.1 Bag of words 5

2.1.2 Support Vector Machines(SVM) 6

2.1.3 Decision based models(decision trees) 6

2.1.4 Attribute-value description 7

2.1.5 Data Chunking 7

2.2 My Methodology 8

3 System Design **9**

3.1 System Architecture 9

3.2 Process Flowchart 10

3.3 Data Flow Diagram 11
3.4 Usecase Diagram 12
3.5 Sequence Diagram 13
3.6 Requirements/Pre-requisites 14
3.6.1 Software: 14
3.6.2 Hardware: 14

4 Implementation of System/ Methodology 15
4.1 Module 1 16
4.2 Module 2 19
4.3 Module 3 22

5 Results and Discussions 25
5.1 Technical Results 25
5.2 Performance Analysis 29

6 Conclusion and Future Work 31

List of Figures

3.1 System architecture 9
3.2 Process Flowchart 10
3.3 Data Flow diagram 11
3.4 Usecase Diagram 12
3.5 Sequence Diagram 13

4.1 Disease related medical words 16
4.2 Sample abstract file 17
4.3 After Disease tagger 18
4.4 Treatment List 19
4.5 Input Abstract 20
4.6 After Treatment tagger 21
4.7 Lines with only Treatment 22
4.8 Lines with only Disease 23
4.9 Disease and its Treatment 24

5.1 Disease Tagger 26
5.2 Writing into the database 27
5.3 Database 28
5.4 Abstract Performance 29
5.5 Data insertion 30

Chapter 1

Introduction

1.1 Background

Machine learning is one of the domain that is being widely used at present. Integrating Machine learning in medical domain can result in finding a good application that can be used by general users and medical related people. This can also be used in Medical domain to yield better results. Machine learning is a subfield of artificial intelligence domain that concerns about study of systems that can learn from data. Supervised and unsupervised learning algorithms are used in many of applications like classifying emails and predicting the severity of a disease etc. Some machine learning systems attempt to eliminate the need for human intuition in data analysis, while others adopt a collaborative approach between human and machine. Because of this factor developing machine learning systems has increased in recent years.

Medline has health related journals that contain all the disease information and recent developments in their treatments. Using Medline one can retrieve all the documents related to a disease. But this retrieval lack reliable and faster access to required information because this retrieval displays user all the related documents that can be many in number and sometimes may display documents that are not what the user requires. This may be quite frustrating to the user when the user need only required information alone. Performing relation extraction and using learning algorithms can give the user required information about the query.

1.2 Motivation

All the health management systems are evidence based systems and not reliable. But we need better, faster and more reliable access to information. In the medical domain one of the richest and most used source of information is Medline, a database of extensive life science published articles. So we are using Medline data repository because it contains more standardized information about disease and treatment relationships and mostly used by clinicians and scholars. Our application need to extract the disease treatment relation between the from Medline using Pattern Recognition algorithm. Thus the user gets the required information alone which saves his time and improves the quality of the result.

One of the main use of this application is that it can also be used by doctors and medical persons to analyze the treatment for a particular disorder and to proof check his view on a particular disease disorder. It can help users to let them know about the disease they are most likely to suffer from and know about the treatment relations which can be used as first aid or precautionary measures they can take when the doctor is not available immediately. This application can be embedded in a e-commerce website related to health domain. Still there is no such reliable application that can provide standardized solution with the treatment information that user wants.

1.3 Objective

The main objective of this project is to prepare useful data from the medical abstracts that can be useful to the people. This data is the disease treatment relationship. The initial data are taken from the abstracts that are taken from Medline.

1.4 Challenges

One of the main challenges is here in my work a lot of data should be handled. My methodology invloves a series of data processing phases where the data should be properly managed. Even if a small part of data goes missing then it will in turn reflect the entire data preprocessing phase. Moreover we are handling a lot of data so the time taken to process this data should also be taken into consideration and thus we should use efficient technique for faster data processing.

1.5 Essence of the Approach

The approach that is being used in this work is my own idea which is mainly based on the concept of pattern matching and tagger classification. Here i have prepared my own set of data from the different resources and then used the above mentioned concepts to process the data obtained from the medical repository.

1.6 About MedLine

MEDLINE OR MEDLARS Online (Medical Literature Analysis and Retrieval System Online) is a bibliographic database of life sciences and biomedical information. It includes bibliographic information for articles from academic journals covering medicine, nursing,pharmacy, dentistry, veterinary medicine, and health care. MEDLINE also covers much of the literature in biology and biochemistry, as well as fields such as molecular evolution. It is compiled by the United States National Library of Medicine (NLM), MEDLINE is freely available on the Internet and searchable via PubMed and NLM's National Center for Biotechnology Information's Entrez system. MEDLARS (Medical Literature Analysis and Retrieval System) is a computerised biomedical bibliographic retrieval system. It was launched by the National Library of Medicine in 1964 and was the first large scale, computer based, retrospective search service available to the general public.

Since 1879, the National Library of Medicine had published Index Medicus, a monthly guide to medical articles in thousands of journals. The huge volume of bibliographic citations were manually compiled. By 1960 a detailed specification was prepared and by the spring of 1961 a request for proposals was sent out to many companies to develop the system. As a result a computer was developed to run MEDLARS was delivered to the NLM in March 1963. MEDLARS cost 3 million dollars to develop and at the time of its completion in 1964, no other publicly available, fully operational electronic storage and retrieval system of its magnitude existed. In late 1971, an online version called MEDLINE ("MEDLARS Online") became available as a way to do online searching of MEDLARS

from remote medical libraries. This early system covered 239 journals.
The database contains more than 21.6 million records from 5,639 selected publication covering Bio-medicine and health from 1950 to the present. Originally the database covered articles starting from 1965, but this has been enhanced, and records as far back as 1950/51 are now available within the main index. The database is freely accessible on the Internet via the PubMed interface and new citations are added Tuesday through Saturday. For citations added during 1995-2003: about 48 percent are for cited articles published in the U.S., about 88 percent are published in English, and about 76 percent have English abstracts written by authors of the articles. MEDLINE functions as an important resource for biomedical researchers and journal clubs from all over the world. MEDLINE facilitates evidence-based medicine. Most systematic review articles published presently build on extensive searches of MEDLINE to identify articles that might be useful in the review. MEDLINE influences researchers in their choice of journals in which to publish. More than 5,500 biomedical journals are indexed in MEDLINE. New journals are not included automatically or immediately. They are selected based on scientific scope and quality of a journal.

Chapter 2

Overview / Literature Review

The goal of machine learning is to build computer systems that can adapt and learn from their experience. Every machine learning algorithm has both a computational aspect how to compute the answer and a statistical aspect (how to ensure that future predictions are accurate). The ultimate aim of my study is to provide base of information technology model called as Health care Information Model (HIM) or frame work that helps the society to identify and disseminates health care information. There are various approaches (Rosario et al., 2004) (T.F. Michael Raj and S. Prasanna) of classifying the Medline abstracts http:structuredabstracts.nlm.nih.gov.

2.1 Existing Methods

2.1.1 Bag of words

The bag of words representation is commonly used for text classification tasks. It is a representation in which features are chosen among the words that are present in the training data. Because we deal with a feature appears more than once in a sentence, this means that it is important and the frequency value representation will capture thisthe features value will be greater than that of other features. We keep only the words that appeared at least three times in the training collection, contain at least one alphanumeric character, are not part of an English list of stop words [15] and are longer than three characters. Words that have length of two or one character are not considered as features because of two other reasons possible incorrect tokenization and problems with very short acronyms in the medical domain that could be highly ambiguous (could be an acronym or an abbreviation of a common word).

2.1.2 Support Vector Machines(SVM)

SVMs are a new learning method introduced by Vapnik (1995). A Support Vector Machine (SVM) is a concept in statistics and computer science for a set of related supervised learning methods that analyze data and recognize patterns, used for classification and regression analysis. The standard SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the input, making the SVM a non-probabilistic binary linear classifier. They are well founded in terms of computational learning theory and very open to theoretical understanding and analysis. Support vector machines are based on the Structural Risk Minimization principle from computational learning theory. The idea of structural risk minimization is to find a hypothesis h for which we can guarantee the lowest true error. The true error of his the probability that h will make an error on an unseen and randomly selected test example. An upper bound can be used to connect the true error of a hypothesis h with the error of h on the training set and the complexity of H (measured by VCDimension), the hypothesis space containing h (Vapnik, 1995). Support vector machines find the hypothesis h which (approximately) minimizes this bound on the true error by effectively and efficiently controlling the VC-Dimension of H.

2.1.3 Decision based models(decision trees)

Decision trees are powerful and popular tools for classification and prediction. Decision trees represent rules, which can be understood by humans and used in knowledge system such as database. Key requirement of this model is:

2.1.4 Attribute-value description

Object or case must be expressible in terms of a fixed collection of properties or attributes (e.g., hot, mild, cold).
Predefined classes:
The target function has discrete output values (bollean or multiclass)
Sufficient data:
Enough training cases should be provided to learn the model Decision trees classify instances or examples by starting at the root of the tree and moving through it until a leaf node. It is used in hand crafted models and Is suitable for short texts.
Decision tree is a classifier in the form of a tree structure:
Then a decision node which specifies a test on a single attribute and a Leaf node which indicates the value of the target attribute. There is an Arc/edge which splits an attribute and a path which is a disjunction of test to make the final decision

2.1.5 Data Chunking

This involves chunking the data step by step. The process involved here is

1)**Removal of Stop words:**
In general the sentences consists of words like (who, they, them) etc so here we try to remove all the stop words. A list of stop words are present thus we remove them

2)**Removal of Prepositions and Conjunctions:**
Now after the removal of stop words then we try to remove the prepositions that are present in the data. Thus this step involves the removal of all prepositions

3)**Removal of other words:**
Now after the step 2 the preprocessed data consists of proper nouns, adjectives and bio-medical words and other remaining data like dates numbers etc. Thus we are struck here in this step and it takes more number of processing methods to slowly extract the biomedical words. Thus this is a very long time taking process and also the efficiency is also very poor.

2.2 My Methodology

Pattern Matching and Taggers:

In my methodology i have taken the medical related words and then pattern matched them with the abstracts that are obtained from the MedLine. This pattern matching process is done in different steps. Thus the data is trained with a different set of medical related words and thus when ever a word is matched then its particular tag is inserted. Similarly for different data taggers are inserted. Then after the taggers are inserted then based on the taggers found then a particular taggers are inserted at the end of the line. Thus after steps of preprocessing we get the processed data.
Now after obtaining the processed data now using a java program i can extract the required data based on the line taggers and word taggers. Then the data obtained from the taggers is stored into a local database. Thus we get the required related disease treatment relationship.

Chapter 3

System Design

3.1 System Architecture

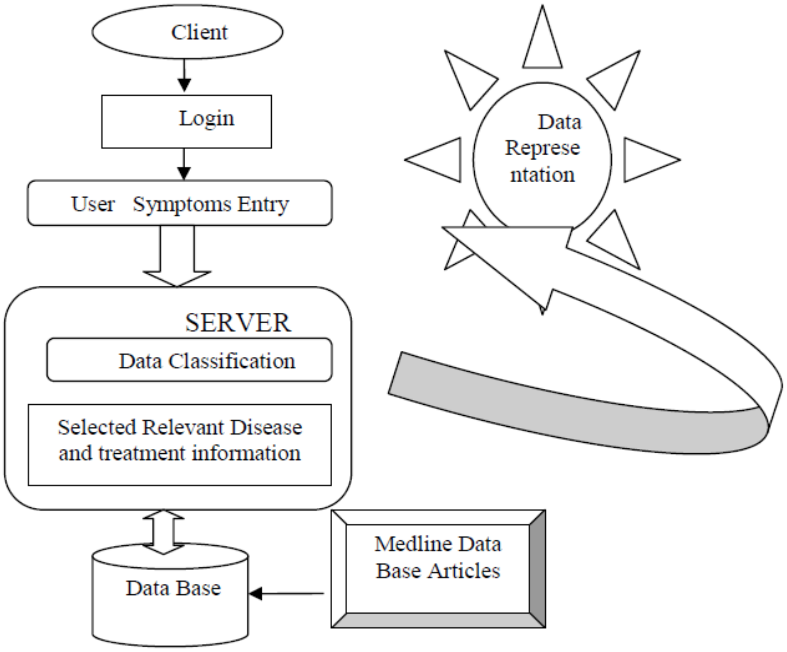


Figure 1.System Architecture

Figure 3.1: System architecture

3.2 Process Flowchart

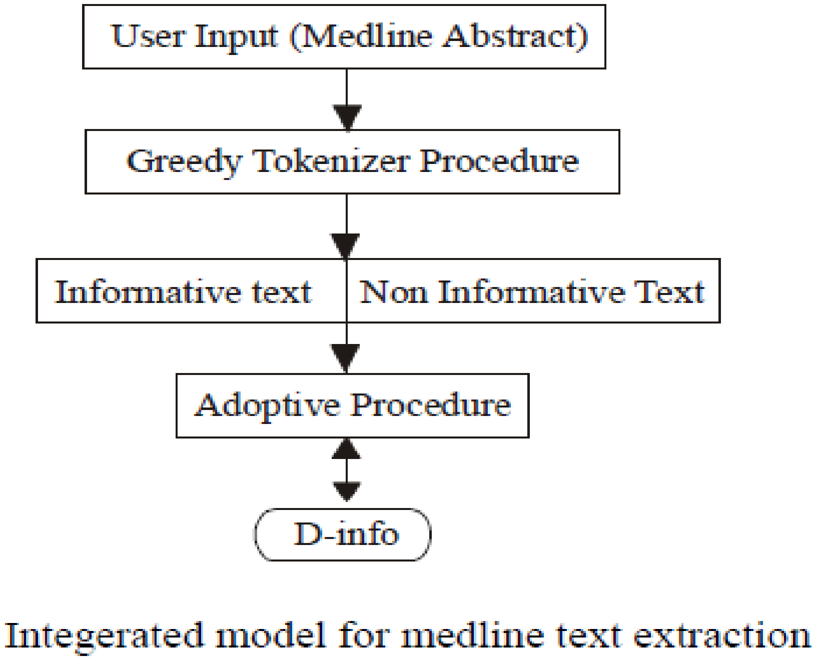

User Input (Medline Abstract)
Greedy Tokenizer Procedure
Informative text
Non Informative Text
Adoptive Procedure
D-info
Integerated model for medline text extraction

Figure 3.2: Process Flowchart

3.3 Data Flow Diagram

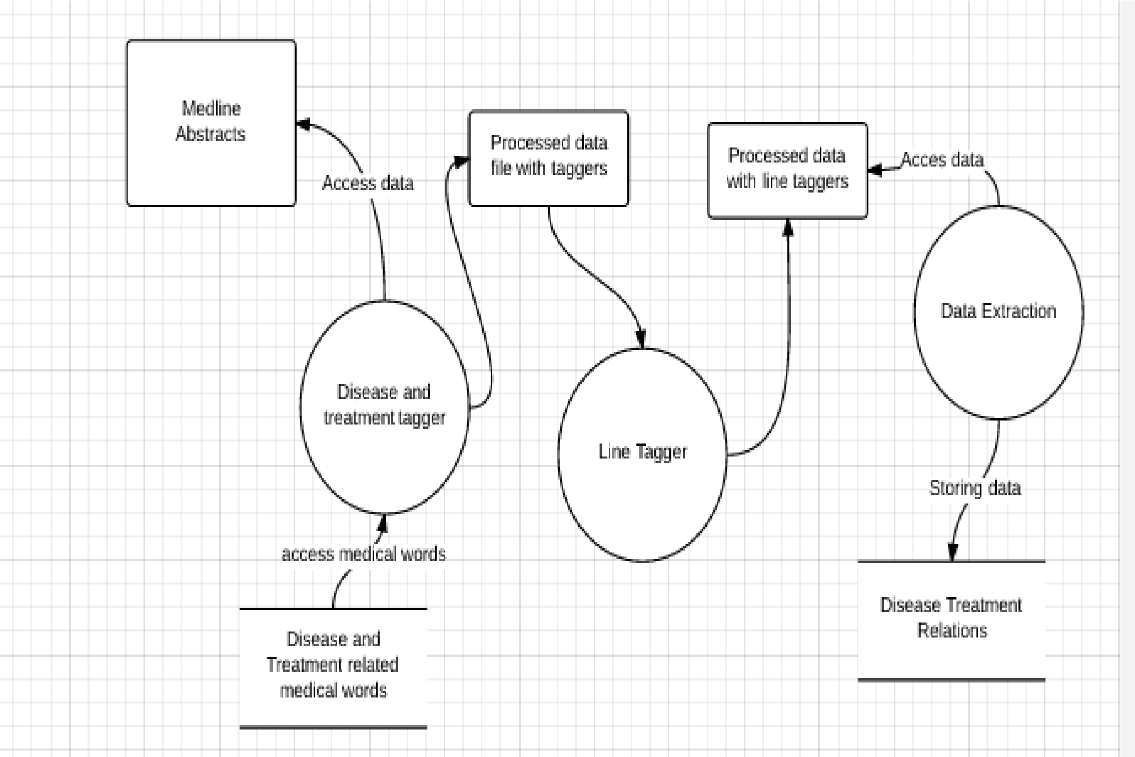

Medline Abstracts
Access data
Processed data file with taggers
Processed data with line taggers
Acces data
Data Extraction
Disease and treatment tagger
Line Tagger
Storing data
access medical words
Disease and Treatment related medical words
Disease Treatment Relations

Figure 3.3: Data Flow diagram

3.4 Usecase Diagram

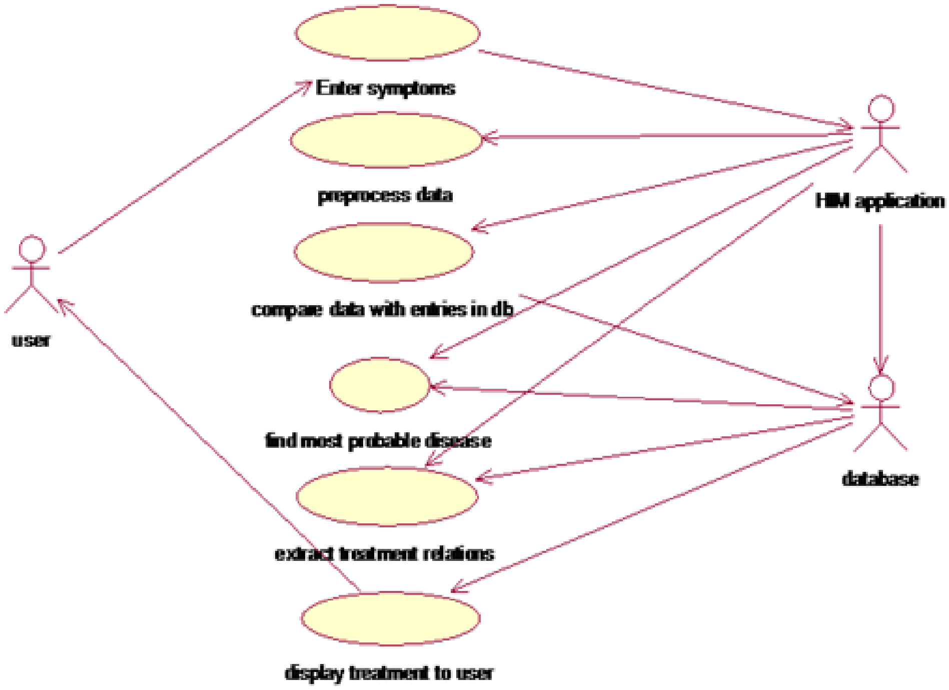

Enter symptoms
preprocess data
compare data with entries in db
find most probable disease
extract treatment relations
display treatment to user
user
HIM application
database

Figure 3.4: Usecase Diagram

3.5 Sequence Diagram

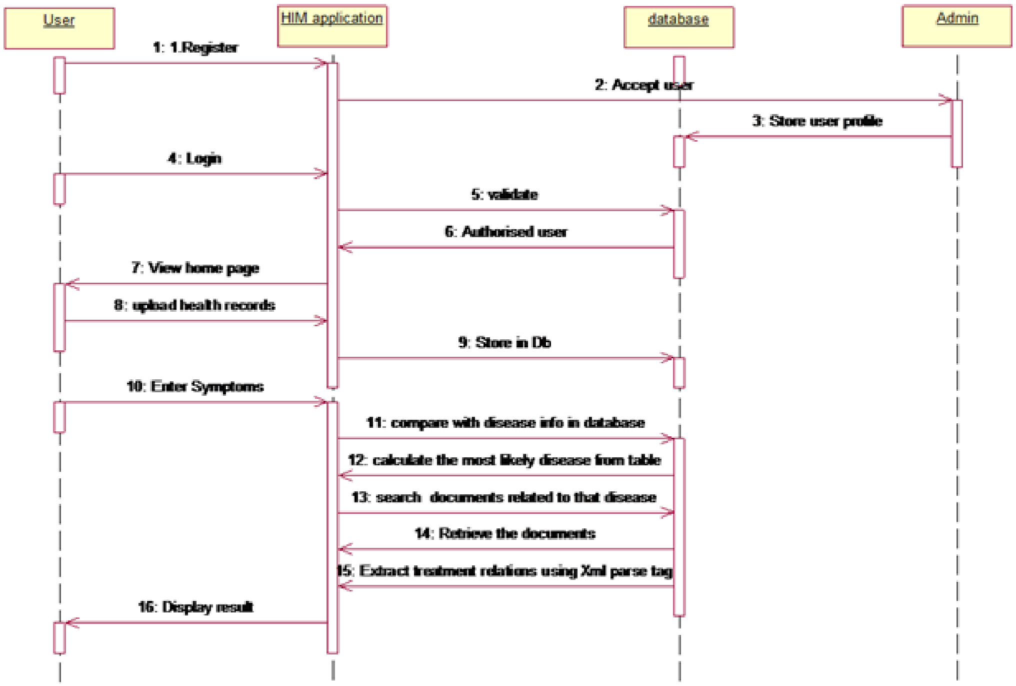

User
HIM application
database
Admin
1: 1.Register
2: Accept user
3: Store user profile
4: Login
5: validate
6: Authorised user
7: View home page
8: upload health records
9: Store in Db
10: Enter Symptoms
11: compare with disease info in database
12: calculate the most likely disease from table
13: search documents related to that disease
14: Retrieve the documents
15: Extract treatment relations using Xml parse tag
16: Display result

Figure 3.5: Sequence Diagram

3.6 Requirements/Pre-requisites

3.6.1 Software:

Languages: JAVA, php, SQL

Server: XXAMP

Database: phpmyadmin

Tools Used: Eclipse, Editplus

3.6.2 Hardware:

Processor: Intel

Hard Disk: More than 100GB

RAM: Greater than 512MB

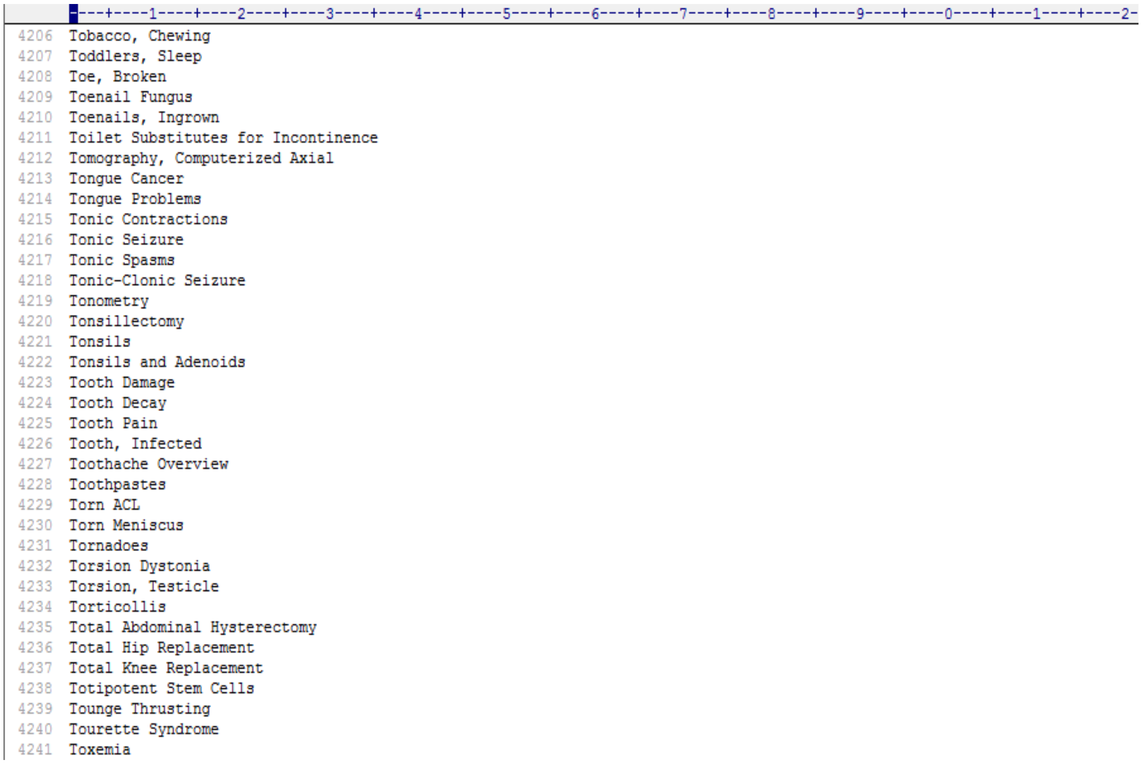
Chapter 4

Implementation of System/ Methodology

MedLine abstracts are in the form of paragraphs. These paragraphs consists information about a particular kind of disease or about the occurrence or something about a disease. But these paragraphs also consists of treatment to a particular disease. In some cases only diseases are mentioned but there may or may not be a treatment to that particular disease. So our aim is to identify the disease and its related treatment.
How do we identify that a particular word from a paragraph is a disease or some other bio-medical term. The methodology used by me is Pattern Matching and Taggers. Thus through a repeated process of data training the processed data is obtained. Thus the modules present in this are:

4.1 Module 1

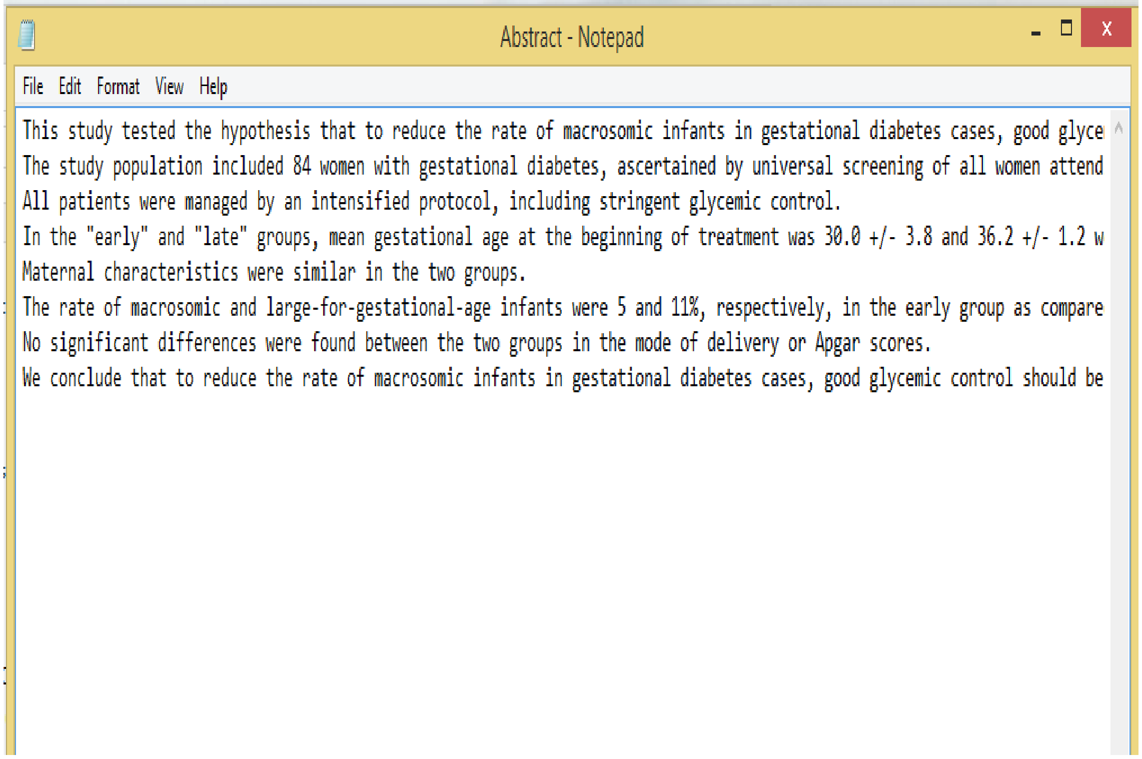
Here I have taken all the disease related words from the medical dictionary and put them into a file. There were nearly 5000 disease related words that are present



```
Tobacco, Chewing
Toddlers, Sleep
Toe, Broken
Toenail Fungus
Toenails, Ingrown
Toilet Substitutes for Incontinence
Tomography, Computerized Axial
Tongue Cancer
Tongue Problems
Tonic Contractions
Tonic Seizure
Tonic Spasms
Tonic-Clonic Seizure
Tonometry
Tonsillectomy
Tonsils
Tonsils and Adenoids
Tooth Damage
Tooth Decay
Tooth Pain
Tooth, Infected
Toothache Overview
Toothpastes
Torn ACL
Torn Meniscus
Tornadoes
Torsion Dystonia
Torsion, Testicle
Torticollis
Total Abdominal Hysterectomy
Total Hip Replacement
Total Knee Replacement
Totipotent Stem Cells
Tounge Thrusting
Tourette Syndrome
Toxemia
```

Figure 4.1: Disease related medical words

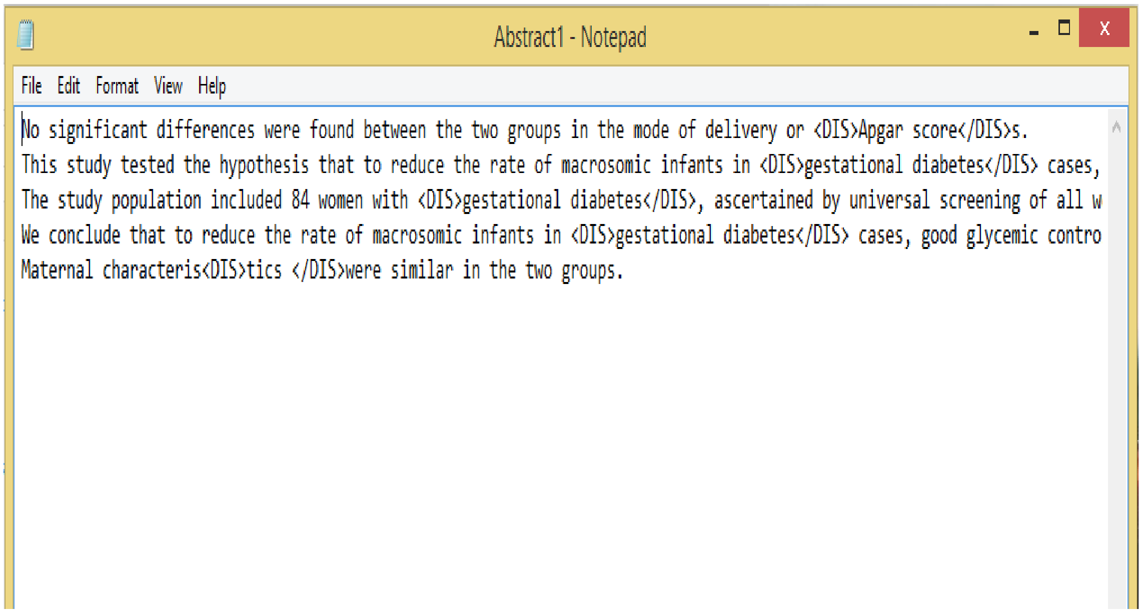
Now take an abstract file. This is a sample abstract file



```
This study tested the hypothesis that to reduce the rate of macrosomic infants in gestational diabetes cases, good glyce
The study population included 84 women with gestational diabetes, ascertained by universal screening of all women attend
All patients were managed by an intensified protocol, including stringent glycemic control.
In the "early" and "late" groups, mean gestational age at the beginning of treatment was 30.0 +/- 3.8 and 36.2 +/- 1.2 w
Maternal characteristics were similar in the two groups.
The rate of macrosomic and large-for-gestational-age infants were 5 and 11%, respectively, in the early group as compare
No significant differences were found between the two groups in the mode of delivery or Apgar scores.
We conclude that to reduce the rate of macrosomic infants in gestational diabetes cases, good glycemic control should be
```

Figure 4.2: Sample abstract file

Now I have taken each word and compared it each word in a single abstract and if the word is present then I have put a tag to it namely <DIS >. For example if the word in from our list was diabetes and if it was found in the abstract, then it would be made like this <DIS >diabetes </DIS >.

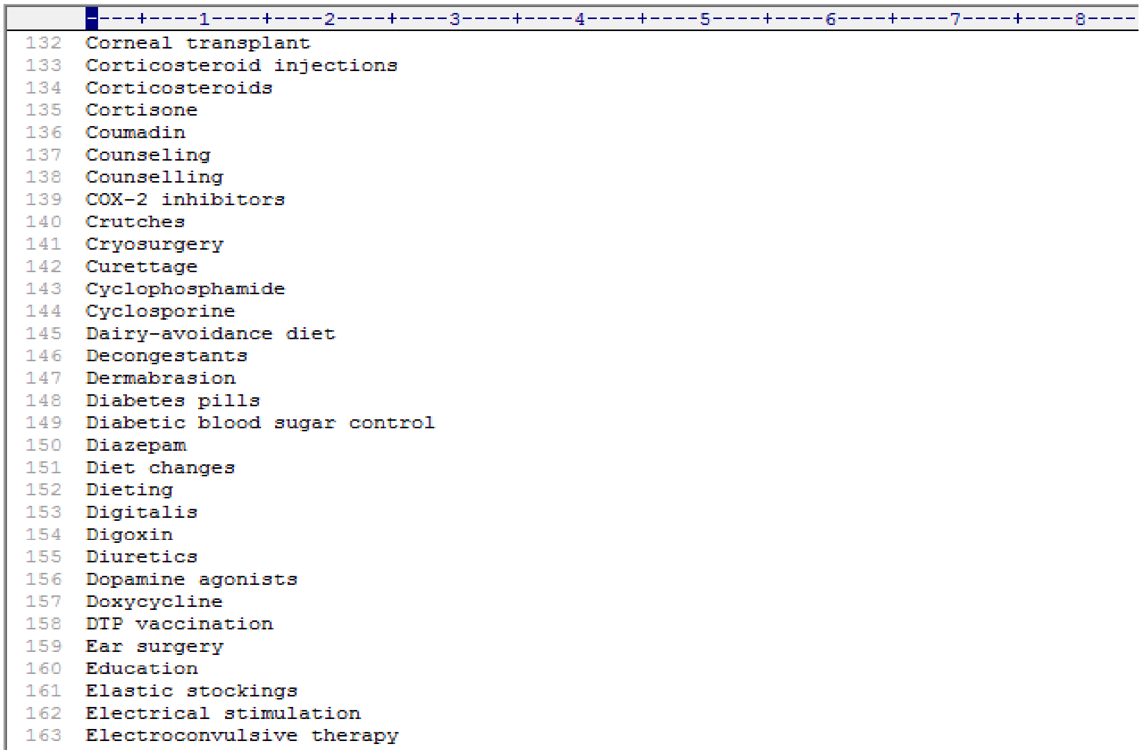


```
No significant differences were found between the two groups in the mode of delivery or <DIS>Apgar score</DIS>s.
This study tested the hypothesis that to reduce the rate of macrosomic infants in <DIS>gestational diabetes</DIS> cases,
The study population included 84 women with <DIS>gestational diabetes</DIS>, ascertained by universal screening of all w
We conclude that to reduce the rate of macrosomic infants in <DIS>gestational diabetes</DIS> cases, good glycemic contro
Maternal characteris<DIS>tics </DIS>were similar in the two groups.
```

Figure 4.3: After Disease tagger

4.2 Module 2

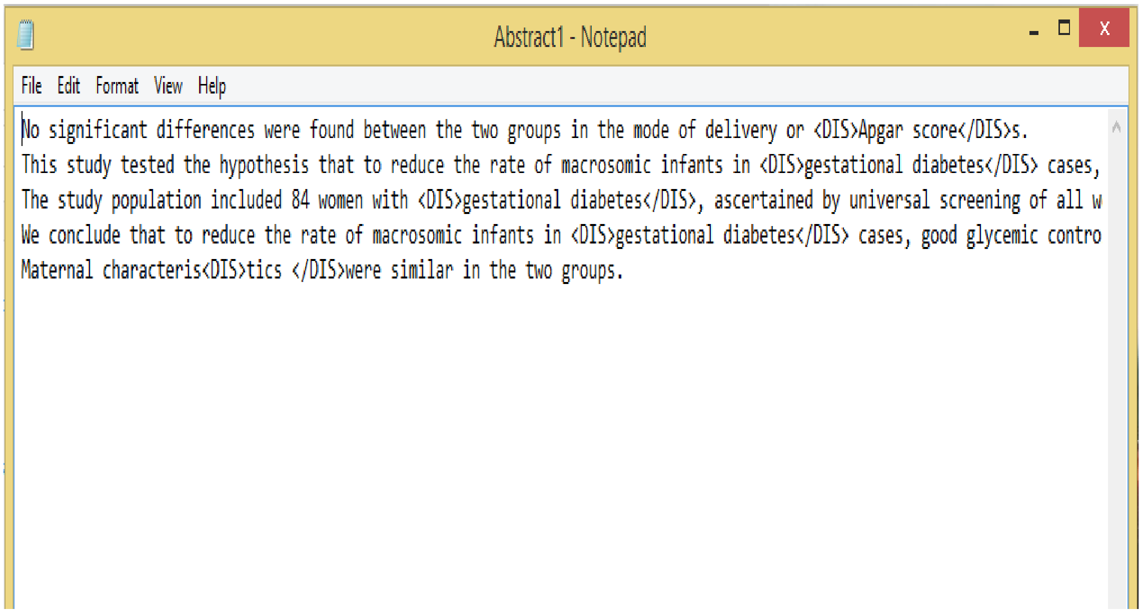
So similarly for the treatment also I have prepared a list of words that are taken from the medical dictionary and similarly the word is identified in the abstracts and then the tag is inserted as <TREAT >.
List of Treatments



```
Corneal transplant
Corticosteroid injections
Corticosteroids
Cortisone
Coumadin
Counseling
Counselling
COX-2 inhibitors
Crutches
Cryosurgery
Curettage
Cyclophosphamide
Cyclosporine
Dairy-avoidance diet
Decongestants
Dermabrasion
Diabetes pills
Diabetic blood sugar control
Diazepam
Diet changes
Dieting
Digitalis
Digoxin
Diuretics
Dopamine agonists
Doxycycline
DTP vaccination
Ear surgery
Education
Elastic stockings
Electrical stimulation
Electroconvulsive therapy
```

Figure 4.4: Treatment List

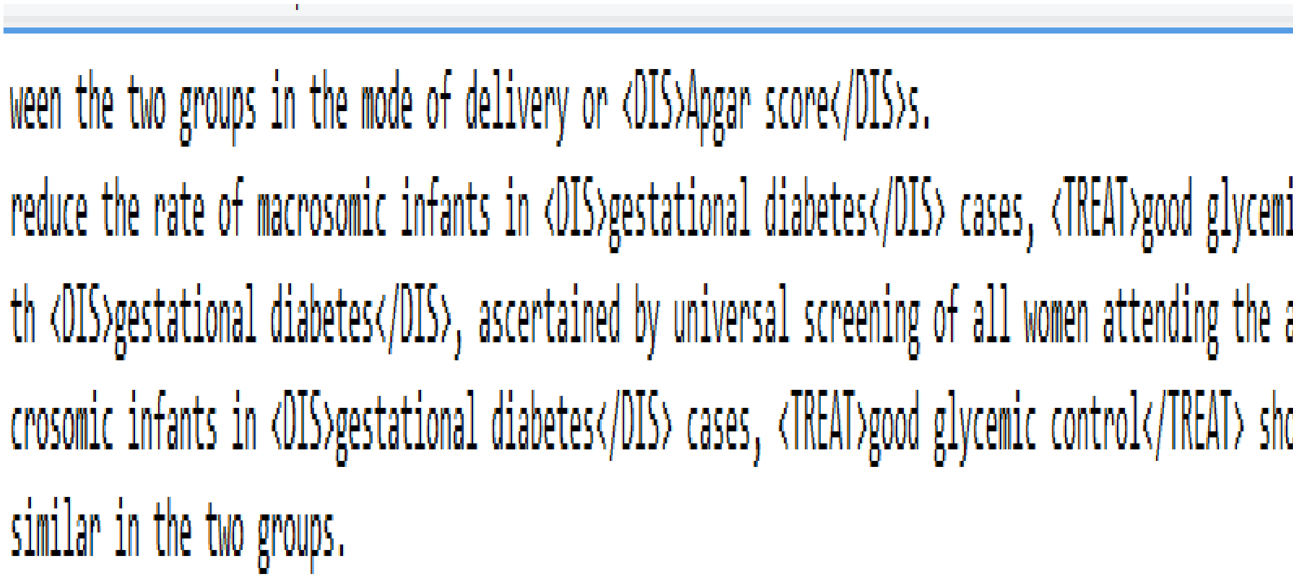
Input for the second module



```
No significant differences were found between the two groups in the mode of delivery or <DIS>Apgar score</DIS>s.
This study tested the hypothesis that to reduce the rate of macrosomic infants in <DIS>gestational diabetes</DIS> cases,
The study population included 84 women with <DIS>gestational diabetes</DIS>, ascertained by universal screening of all w
We conclude that to reduce the rate of macrosomic infants in <DIS>gestational diabetes</DIS> cases, good glycemic contro
Maternal characteris<DIS>tics </DIS>were similar in the two groups.
```

Figure 4.5: Input Abstract

Data after processing the treatment



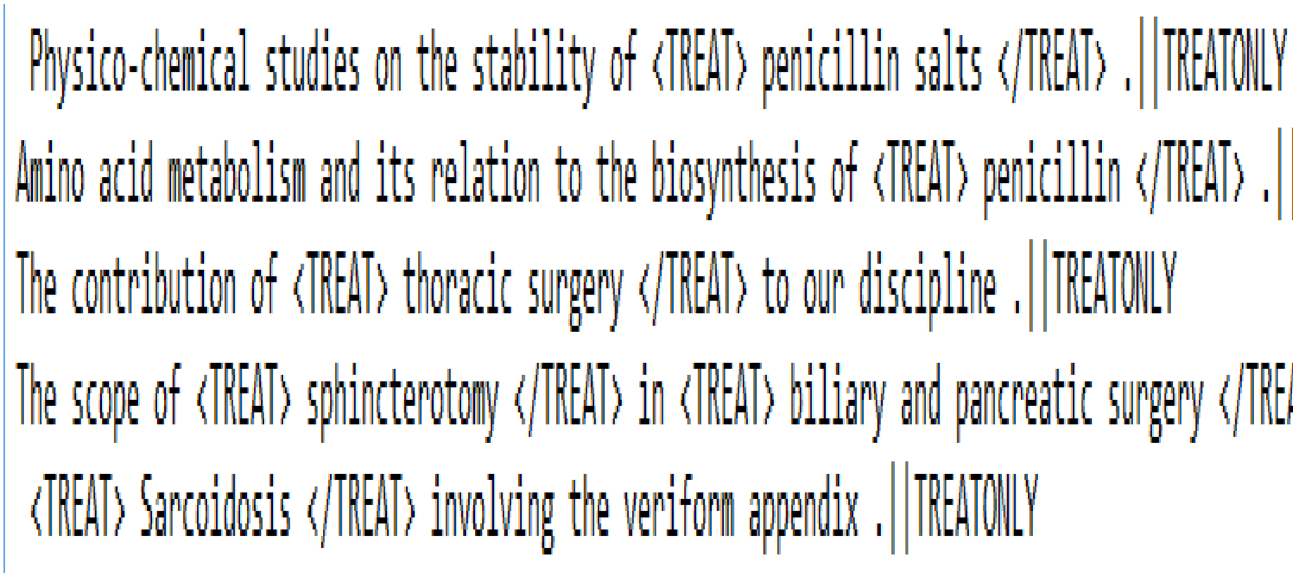
```
ween the two groups in the mode of delivery or <DIS>Apgar score</DIS>s.
reduce the rate of macrosomic infants in <DIS>gestational diabetes</DIS> cases, <TREAT>good glycemi
th <DIS>gestational diabetes</DIS>, ascertained by universal screening of all women attending the a
crosomic infants in <DIS>gestational diabetes</DIS> cases, <TREAT>good glycemic control</TREAT> sho
similar in the two groups.
```

Figure 4.6: After Treatment tagger

4.3 Module 3

We now take the consized data that is only the lines that consists of the tags. Now how to we know whether treatment is for a particular disease. Now we take the processed data that consists of tags and then read each line.

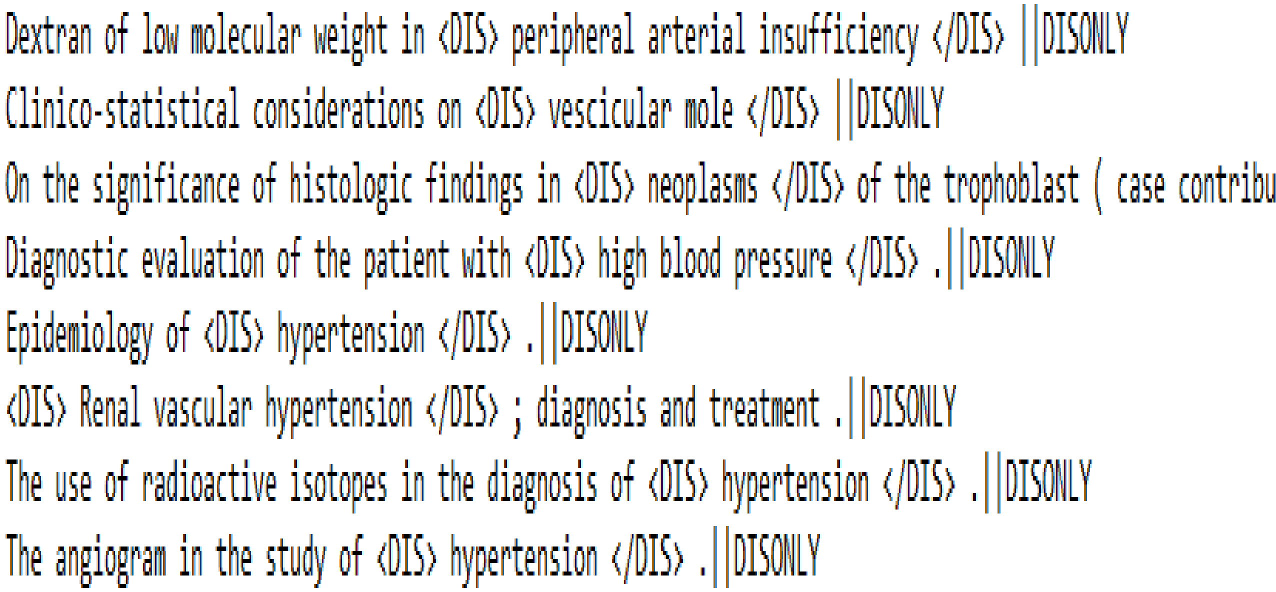
Now if a line contains only <TREAT >then at the end of the line we give it as TREATONLY



```
 Physico-chemical studies on the stability of <TREAT> penicillin salts </TREAT> .||TREATONLY
Amino acid metabolism and its relation to the biosynthesis of <TREAT> penicillin </TREAT> .|
The contribution of <TREAT> thoracic surgery </TREAT> to our discipline .||TREATONLY
The scope of <TREAT> sphincterotomy </TREAT> in <TREAT> biliary and pancreatic surgery </TREA
 <TREAT> Sarcoidosis </TREAT> involving the veriform appendix .||TREATONLY
```

Figure 4.7: Lines with only Treatment

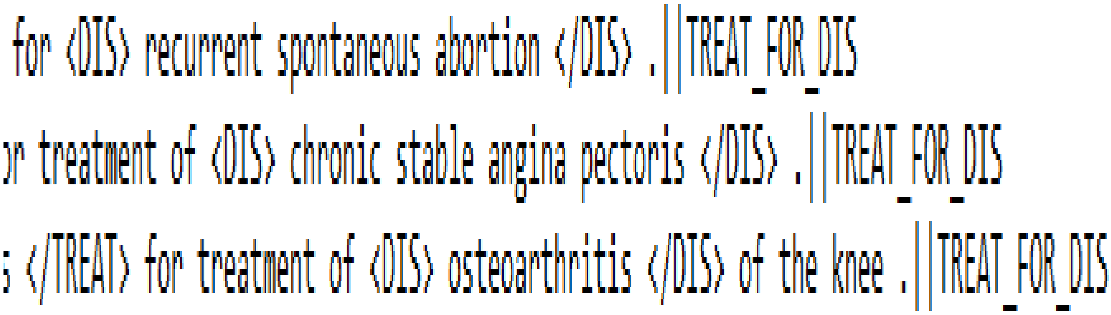
Similarly if the line contains only <DIS >then we give it as DIS$_{O}NLY$



```
Dextran of low molecular weight in <DIS> peripheral arterial insufficiency </DIS> ||DISONLY
Clinico-statistical considerations on <DIS> vescicular mole </DIS> ||DISONLY
On the significance of histologic findings in <DIS> neoplasms </DIS> of the trophoblast ( case contribu
Diagnostic evaluation of the patient with <DIS> high blood pressure </DIS> .||DISONLY
Epidemiology of <DIS> hypertension </DIS> .||DISONLY
<DIS> Renal vascular hypertension </DIS> ; diagnosis and treatment .||DISONLY
The use of radioactive isotopes in the diagnosis of <DIS> hypertension </DIS> .||DISONLY
The angiogram in the study of <DIS> hypertension </DIS> .||DISONLY
```

Figure 4.8: Lines with only Disease

Whereas if the line consists of both <DIS >and <TREAT >then at the end of the line we give it as TREATMENT



```
for <DIS> recurrent spontaneous abortion </DIS> .||TREAT_FOR_DIS
or treatment of <DIS> chronic stable angina pectoris </DIS> .||TREAT_FOR_DIS
s </TREAT> for treatment of <DIS> osteoarthritis </DIS> of the knee .||TREAT_FOR_DIS
```

Figure 4.9: Disease and its Treatment

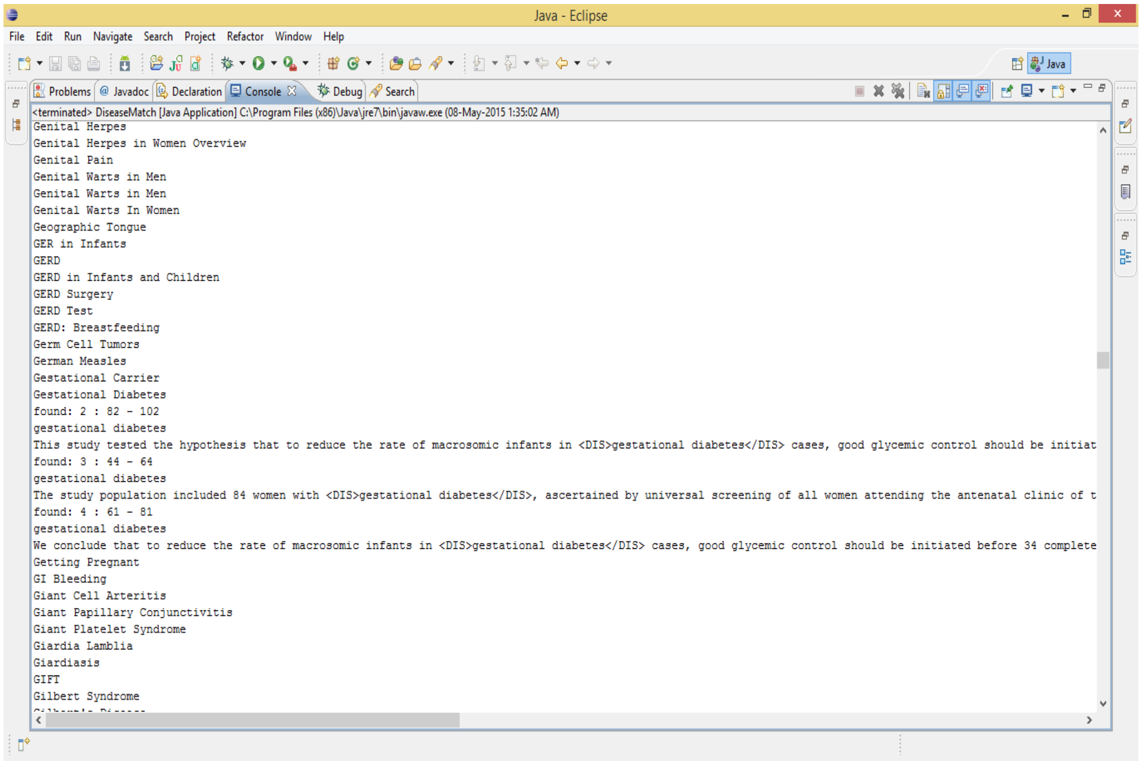
After preparing the dataset with tag elements and the classification we now extract the treatment to a particular disease by taking the line with ——TREATMENT and then extract the data between the tag parts and store them into a database. Later we use this database to display to the user a treatment to a particular disease After displaying the treatment we also provide the user with an idea or description of the treatment.

Chapter 5

Results and Discussions

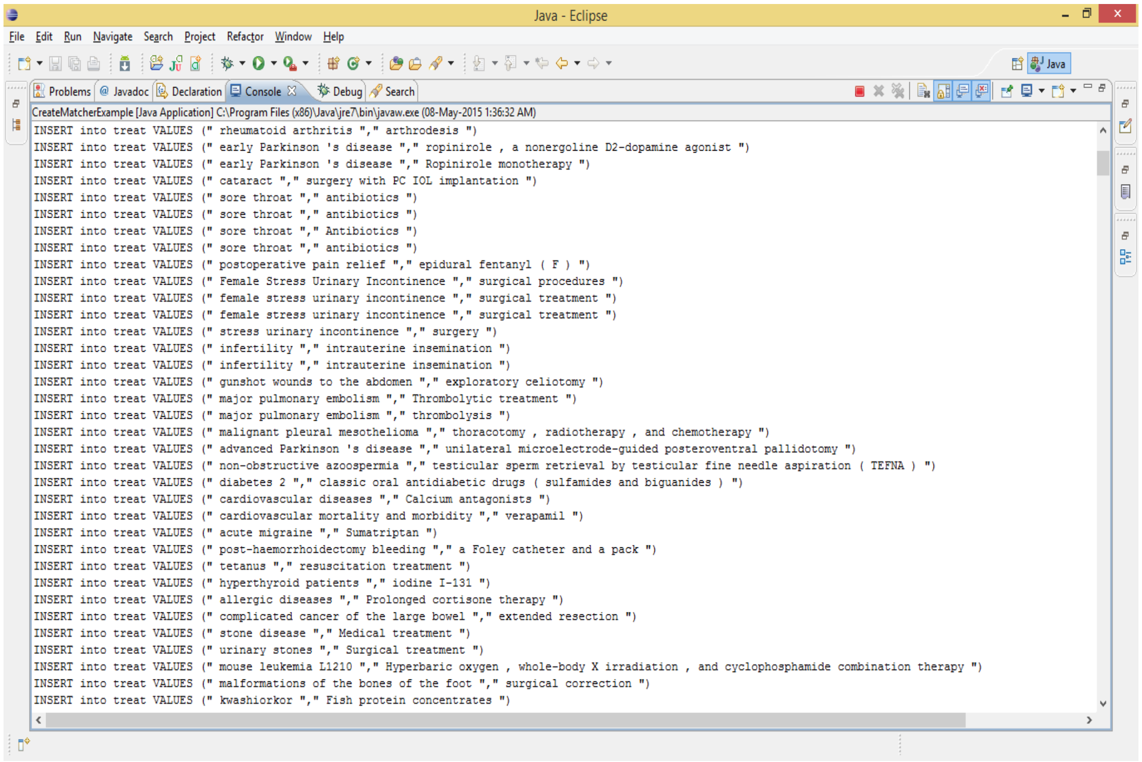
5.1 Technical Results

The step by step process involved here is pattern matching and the taggers part. First the data from the medical abstracts is taken and then compared with the diseases list and then taggers part is done. Similarly it is also compared with the treatment list. Then from that another line tagger part is done. Then a database is obtained from the data with in the tagger part and thus this data is stored in the database. Now i have linked the database to the interface part which is useful to the people where the people can select the disease and then get its related treatment.



```
<terminated> DiseaseMatch [Java Application] C:\Program Files (x86)\Java\jre7\bin\javaw.exe (08-May-2015 1:35:02 AM)
Genital Herpes
Genital Herpes in Women Overview
Genital Pain
Genital Warts in Men
Genital Warts in Men
Genital Warts In Women
Geographic Tongue
GER in Infants
GERD
GERD in Infants and Children
GERD Surgery
GERD Test
GERD: Breastfeeding
Germ Cell Tumors
German Measles
Gestational Carrier
Gestational Diabetes
found: 2 : 82 - 102
gestational diabetes
This study tested the hypothesis that to reduce the rate of macrosomic infants in <DIS>gestational diabetes</DIS> cases, good glycemic control should be initiat
found: 3 : 44 - 64
gestational diabetes
The study population included 84 women with <DIS>gestational diabetes</DIS>, ascertained by universal screening of all women attending the antenatal clinic of t
found: 4 : 61 - 81
gestational diabetes
We conclude that to reduce the rate of macrosomic infants in <DIS>gestational diabetes</DIS> cases, good glycemic control should be initiated before 34 complete
Getting Pregnant
GI Bleeding
Giant Cell Arteritis
Giant Papillary Conjunctivitis
Giant Platelet Syndrome
Giardia Lamblia
Giardiasis
GIFT
Gilbert Syndrome
Gilbert's Disease
```

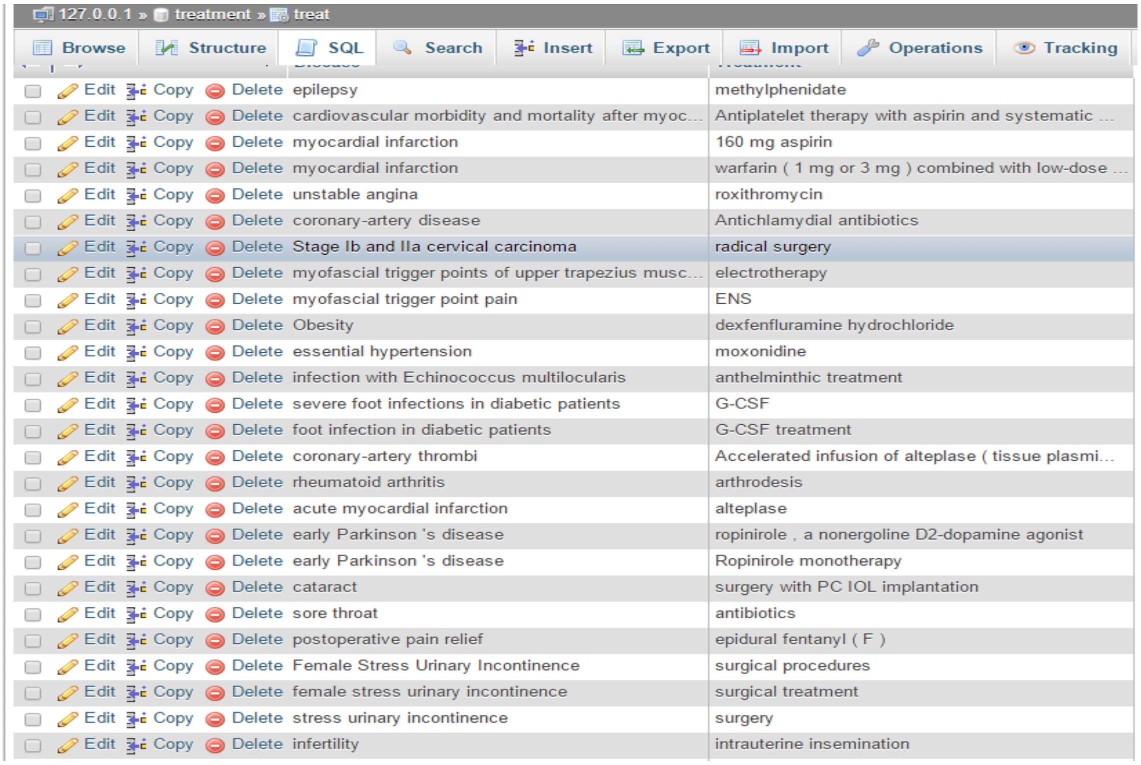
Figure 5.1: Disease Tagger


Java - Eclipse
File Edit Run Navigate Search Project Refactor Window Help
Java
Problems Javadoc Declaration Console Debug Search

CreateMatcherExample [Java Application] C:\Program Files (x86)\Java\jre7\bin\javaw.exe (08-May-2015 1:36:32 AM)

```
INSERT into treat VALUES (" rheumatoid arthritis "," arthrodesis ")
INSERT into treat VALUES (" early Parkinson 's disease "," ropinirole , a nonergoline D2-dopamine agonist ")
INSERT into treat VALUES (" early Parkinson 's disease "," Ropinirole monotherapy ")
INSERT into treat VALUES (" cataract "," surgery with PC IOL implantation ")
INSERT into treat VALUES (" sore throat "," antibiotics ")
INSERT into treat VALUES (" sore throat "," antibiotics ")
INSERT into treat VALUES (" sore throat "," Antibiotics ")
INSERT into treat VALUES (" sore throat "," antibiotics ")
INSERT into treat VALUES (" postoperative pain relief "," epidural fentanyl ( F ) ")
INSERT into treat VALUES (" Female Stress Urinary Incontinence "," surgical procedures ")
INSERT into treat VALUES (" female stress urinary incontinence "," surgical treatment ")
INSERT into treat VALUES (" female stress urinary incontinence "," surgical treatment ")
INSERT into treat VALUES (" stress urinary incontinence "," surgery ")
INSERT into treat VALUES (" infertility "," intrauterine insemination ")
INSERT into treat VALUES (" infertility "," intrauterine insemination ")
INSERT into treat VALUES (" gunshot wounds to the abdomen "," exploratory celiotomy ")
INSERT into treat VALUES (" major pulmonary embolism "," Thrombolytic treatment ")
INSERT into treat VALUES (" major pulmonary embolism "," thrombolysis ")
INSERT into treat VALUES (" malignant pleural mesothelioma "," thoracotomy , radiotherapy , and chemotherapy ")
INSERT into treat VALUES (" advanced Parkinson 's disease "," unilateral microelectrode-guided posteroventral pallidotomy ")
INSERT into treat VALUES (" non-obstructive azoospermia "," testicular sperm retrieval by testicular fine needle aspiration ( TEFNA ) ")
INSERT into treat VALUES (" diabetes 2 "," classic oral antidiabetic drugs ( sulfamides and biguanides ) ")
INSERT into treat VALUES (" cardiovascular diseases "," Calcium antagonists ")
INSERT into treat VALUES (" cardiovascular mortality and morbidity "," verapamil ")
INSERT into treat VALUES (" acute migraine "," Sumatriptan ")
INSERT into treat VALUES (" post-haemorrhoidectomy bleeding "," a Foley catheter and a pack ")
INSERT into treat VALUES (" tetanus "," resuscitation treatment ")
INSERT into treat VALUES (" hyperthyroid patients "," iodine I-131 ")
INSERT into treat VALUES (" allergic diseases "," Prolonged cortisone therapy ")
INSERT into treat VALUES (" complicated cancer of the large bowel "," extended resection ")
INSERT into treat VALUES (" stone disease "," Medical treatment ")
INSERT into treat VALUES (" urinary stones "," Surgical treatment ")
INSERT into treat VALUES (" mouse leukemia L1210 "," Hyperbaric oxygen , whole-body X irradiation , and cyclophosphamide combination therapy ")
INSERT into treat VALUES (" malformations of the bones of the foot "," surgical correction ")
INSERT into treat VALUES (" kwashiorkor "," Fish protein concentrates ")
```

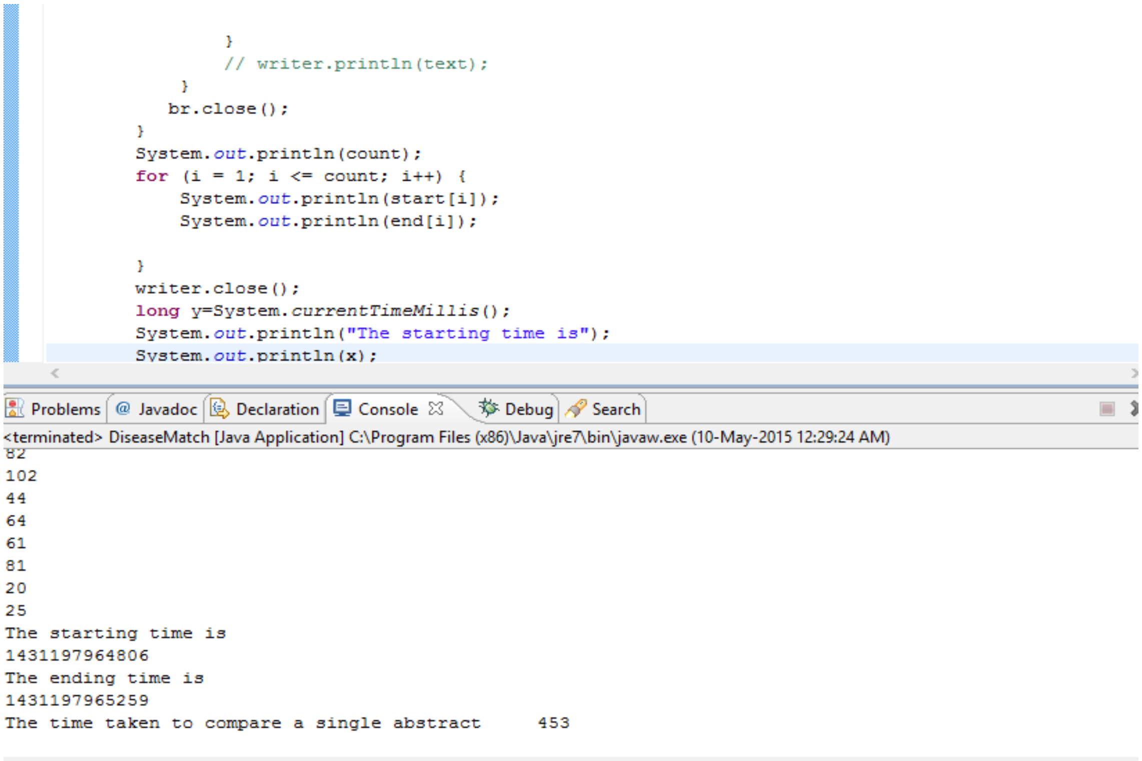
Figure 5.2: Writing into the database



□	Edit	Copy	Delete	epilepsy	methylphenidate
□	Edit	Copy	Delete	cardiovascular morbidity and mortality after myoc...	Antiplatelet therapy with aspirin and systematic ...
□	Edit	Copy	Delete	myocardial infarction	160 mg aspirin
□	Edit	Copy	Delete	myocardial infarction	warfarin (1 mg or 3 mg) combined with low-dose ...
□	Edit	Copy	Delete	unstable angina	roxithromycin
□	Edit	Copy	Delete	coronary-artery disease	Antichlamydial antibiotics
□	Edit	Copy	Delete	Stage Ib and IIa cervical carcinoma	radical surgery
□	Edit	Copy	Delete	myofascial trigger points of upper trapezius musc...	electrotherapy
□	Edit	Copy	Delete	myofascial trigger point pain	ENS
□	Edit	Copy	Delete	Obesity	dexfenfluramine hydrochloride
□	Edit	Copy	Delete	essential hypertension	moxonidine
□	Edit	Copy	Delete	infection with Echinococcus multilocularis	anthelminthic treatment
□	Edit	Copy	Delete	severe foot infections in diabetic patients	G-CSF
□	Edit	Copy	Delete	foot infection in diabetic patients	G-CSF treatment
□	Edit	Copy	Delete	coronary-artery thrombi	Accelerated infusion of alteplase (tissue plasmi...
□	Edit	Copy	Delete	rheumatoid arthritis	arthrodesis
□	Edit	Copy	Delete	acute myocardial infarction	alteplase
□	Edit	Copy	Delete	early Parkinson 's disease	ropinirole , a nonergoline D2-dopamine agonist
□	Edit	Copy	Delete	early Parkinson 's disease	Ropinirole monotherapy
□	Edit	Copy	Delete	cataract	surgery with PC IOL implantation
□	Edit	Copy	Delete	sore throat	antibiotics
□	Edit	Copy	Delete	postoperative pain relief	epidural fentanyl (F)
□	Edit	Copy	Delete	Female Stress Urinary Incontinence	surgical procedures
□	Edit	Copy	Delete	female stress urinary incontinence	surgical treatment
□	Edit	Copy	Delete	stress urinary incontinence	surgery
□	Edit	Copy	Delete	infertility	intrauterine insemination

Figure 5.3: Database

5.2 Performance Analysis

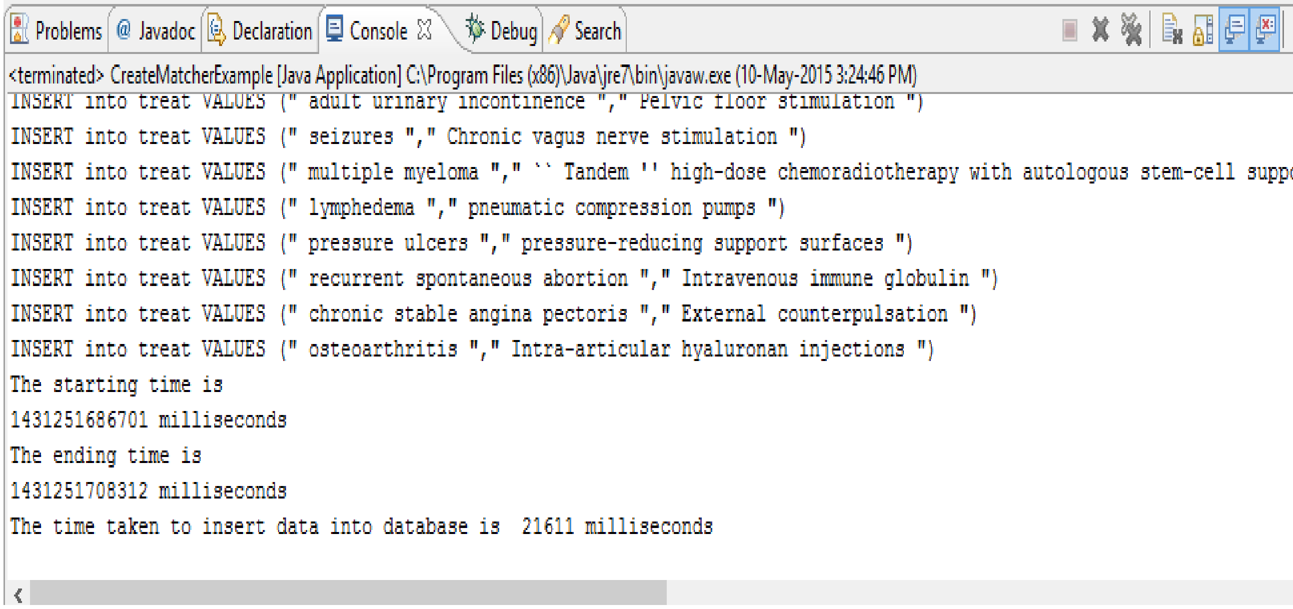


```
        }
        // writer.println(text);
    }
   br.close();
}
System.out.println(count);
for (i = 1; i <= count; i++) {
    System.out.println(start[i]);
    System.out.println(end[i]);

}
writer.close();
long y=System.currentTimeMillis();
System.out.println("The starting time is");
System.out.println(x);
```

```
<terminated> DiseaseMatch [Java Application] C:\Program Files (x86)\Java\jre7\bin\javaw.exe (10-May-2015 12:29:24 AM)
82
102
44
64
61
81
20
25
The starting time is
1431197964806
The ending time is
1431197965259
The time taken to compare a single abstract     453
```

Figure 5.4: Abstract Performance



```
<terminated> CreateMatcherExample [Java Application] C:\Program Files (x86)\Java\jre7\bin\javaw.exe (10-May-2015 3:24:46 PM)
INSERT into treat VALUES (" adult urinary incontinence "," Pelvic floor stimulation ")
INSERT into treat VALUES (" seizures "," Chronic vagus nerve stimulation ")
INSERT into treat VALUES (" multiple myeloma "," `` Tandem '' high-dose chemoradiotherapy with autologous stem-cell suppo
INSERT into treat VALUES (" lymphedema "," pneumatic compression pumps ")
INSERT into treat VALUES (" pressure ulcers "," pressure-reducing support surfaces ")
INSERT into treat VALUES (" recurrent spontaneous abortion "," Intravenous immune globulin ")
INSERT into treat VALUES (" chronic stable angina pectoris "," External counterpulsation ")
INSERT into treat VALUES (" osteoarthritis "," Intra-articular hyaluronan injections ")
The starting time is
1431251686701 milliseconds
The ending time is
1431251708312 milliseconds
The time taken to insert data into database is  21611 milliseconds
```

Figure 5.5: Data insertion

Chapter 6

Conclusion and Future Work

This project mainly emphasizes on the data preprocessing of the data that is taken from the MedLine and this is what the main objective of my work. This project also provides us with a database of the disease treatment relationships and thus we use this database in the website which is designed as an interface for the people. Here the people can select the particular disease and thus get its related treatment. The methodology used in this work is simple and easily understood by everyone and also the this methodology helped in obtaining the desired results. In this project i have also included the description or information about a particular treatment just to provide the users with some basic knowledge of what the treatment involves.

Coming to the future work, this is a medical based project and there is always scope for more improvement. In this project i have included only information for a finite number of treatments which can be more increased. Moreover as i have processed the disease treatment relationship we can also develop a new kind of system which takes the users symptoms and then finds out the occurrence of a disease and then its corresponding treatment. Moreover this work of mine includes extracting treatments but these may be a treatment or a prevention thus the next challenge would be to identify what it is whether a treatment or a prevention or a side effect.

Appendices

Sample Code

```
import java.io.*;
import java.util.*;
import java.util.regex.Matcher;
import java.util.regex.Pattern;
public class DiseaseMatch
public static void main(String[] args) throws Exception
String text = ””;
int[] start = new int[20];
int[] end = new int[20];
int i = 1;
int count = 0;
PrintWriter writer = new PrintWriter((”C:/Users/Kartheek/Desktop/review-1/pendrive/Abstract2.txt”)
BufferedReader brr = new BufferedReader(new FileReader ”C:/Users/Kartheek/Desktop/review-1/pendrive/DisList.txt”));
String dis = ””;
long x=System.currentTimeMillis();
System.out.println(System.currentTimeMillis());
while ((dis = brr.readLine()) != null)
System.out.println(dis);
BufferedReader br = new BufferedReader(new FileReader(”C:/Users/Kartheek/Desktop/review-1/pendrive/Abstract.txt”));
while ((text = br.readLine()) != null)
String patternString = dis;
Pattern pattern = Pattern.compile(patternString,Pattern.CASEINSENSITIVE);
Matcher matcher = pattern.matcher(text);
```

```
while (matcher.find())
count++;
System.out.println(”found: ” + count + ” : ”+ matcher.start() + ” - ” + matcher.end());
start[i] = matcher.start();
end[i] = matcher.end();
String sub1 = text.substring(matcher.start(), matcher.end());
System.out.println(sub1);
String newLine = text.substring(0, start[i]) + ”¡DIS¿”+ text.substring(start[i], end[i])
+ ”¡/DIS¿”+ text.substring(end[i], text.length());
String newLine2 = text.substring(end[i], text.length());
System.out.println(newLine);
writer.println(newLine);
i++;
System.out.println(count);
for (i = 1; i ¡= count; i++)
System.out.println(start[i]);
System.out.println(end[i]);
writer.close();
long y=System.currentTimeMillis();
System.out.println(System.currentTimeMillis());
System.out.println(y-x);
```

Appendices are provided to give supplementary information, which is not included in the main text may serve as a separate part contributing to main theme.

Bibliography

[1] Oana Frunza, Diana Inkpen and Thomas Tran
A Machine Learning Approach for Identifying Disease-Treatment Relations in Short Texts" vol. 23, 2011.

[2] B. Rosario and M.A. Hearst
Semantic Relations in Bioscience Text, Proc. 42nd Ann. Meeting on Assoc. for Computational Linguistics, vol. 430, 2004

[3] Mark Craven
Learning to Extract Relations from Medline, Carnegie Mellon University

[4] R. Bunescu and R. Mooney
A Shortest Path Dependency Kernel for Relation Extraction, Proc. Conf. Human Language Technology and Empirical Methods in EMNLP), pp. 724-731, 2005

[5] T.Sakthimurugan , S.Poonkuzhali
An Effective Retrieval of Medical Records using Data Mining Techniques

[6] A.M. Cohen and W.R. Hersh, and R.T. Bhupatiraju
Feature Generation, Feature Selection, Classifiers, and Conceptual Drift for Biomedical Document Triage, Proc. 13th Text Retrieval Conf. 2004

[7] http://www.biomedcentral.com/1471-2105/5/146

[8] http://www.nlm.nih.gov/bsd/licensee/2013stats/baselinemedfilecount.html