Analysis of Inbox for Fraudulent Activity

Project Report submitted in partial fulfillment of the requirement for

the degree of

Bachelor of Technology.

in

Information Technology

under the Supervision of

Mrs. Sanjana Singh

By

Rajat Jain (111456)

to

Jaypee University of Information and Technology Waknaghat,
Solan – 173234, Himachal Pradesh

Certificate

This is to certify that project report entitled "**ANALYSIS OF INBOX FOR FRAUDLENT ACTIVITY**", submitted by "**RAJAT JAIN** "in partial fulfillment for the award of degree of Bachelor of Technology in Information Technology to Jaypee University of Information Technology, Waknaghat, Solan has been carried out under my supervision.

This work has not been submitted partially or fully to any other University or Institute for the award of this or any other degree or diploma.

Date: **Supervisor's Name: Mrs. Sanjana Singh**

Designation: Assistant Professor

Acknowledgement

On the completion of our work scheduled for this phase of the project titled **"ANALYSIS OF INBOX FOR FRAUDLENT ACTIVITY"** I would like to thank to my supervisor **Mrs. Sanjana Singh** for his able guidance, valuable suggestions and constant encouragement. With her help I was able to complete the research work required for the completion of the project.

I am very grateful to **Mr. Amit Srivastva** (CSE Project Lab) for his assistance.

Date: Name of the student: Rajat Jain (111456)

Table of Content

Abbreviations and Symbols

SLA:	Service Level Agreement
TREC:	Text REtrivel Conference
IR:	Information Retrieval
IDE:	Integrated Development Environment
HTML:	Hyper Text Markup Language
CSS:	Cascading Style Sheet
CBM:	Concept Based Model
SVM:	Support Vector Machine
TFIDF:	Term Frequency Inverse Document Frequency
BM25:	Best Matching

List of Figures

List of Tables

Abstract

Many data mining techniques have been proposed for mining useful patterns in text documents. However, how to effectively use and update discovered patterns is still an open research issue, especially in the domain of text mining. Since most existing text mining methods adopted term-based approaches, they all suffer from the problems of polysemy and synonymy. Over the years, people have often held the hypothesis that pattern (or phrase)-based approaches should perform better than the term-based ones, but many experiments do not support this hypothesis. In the i try to implement pattern discovery technique which includes the processes of pattern deploying and pattern evolving, to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information. Many factors encourages towards its use due to huge potential in it and can give excellent results in terms of pattern finding which is highly effective and can put to test in various critical situations.

While seeking help for while planning a software project one faces various troubles and need for guidance the obvious choice for such situations is looking up blog sites that can give us some insight into our problem. As we know that the solution threads posted on the blogs are contributions of users and hence might give us a completely untrue result especially when we go beyond the basic coding. The blog threads also tend to deviate from the topic giving us no fruitful conclusions.

Therefore, we propose to provide solutions based on the developers interactions and suggestions available in the form mailing lists (as read only archives).

The mailing lists as such are not very comprehensive. They are also highly unstructured. So, direct search for a useful help through a mailing list is not usually a sought after solution. We plan to

extract the worthy data in the mailing list in a structured manner based on recognized keywords so that the data in the mailing list can be presented to the user in a readable format.

The motivation behind the exercise is the fact that mailing lists have conversation threads of the developers and hence the solution proposed in the mailing lists are authentic and usually the most efficient possible. Also, it is a possibility that the mailing list contains some solution idea that has never been implemented in real but could potentially be a highly yielding solution.

Apart from this we at times also encounter suggestion that if applied could improve the efficiency of programs by many folds.

CHAPTER 1 INTRODUCTION

1.1 General Introduction

1.1.1 Introduction to Text Mining

Today due to increase in use of analytics for various purposes data mining has gained important position because it helps in identification and transmission of useful information. Using mining one can increase the output of a process at any stage. Due to the rapid growth of digital data made available in recent years, knowledge discovery and data mining have attracted a great deal of attention with an imminent need for turning such data into useful information and knowledge. Many applications, such as market analysis and business management, can benefit by the use of the information and knowledge extracted from a large amount of data. Knowledge discovery can be viewed as the process of nontrivial extraction of information from large databases, information that is implicitly presented in the data, previously unknown and potentially useful for users.

Data mining is therefore an essential step in the process of knowledge discovery in databases. In the past decade, a significant number of data mining techniques have been presented in order to perform different knowledge tasks. These techniques include association rule mining, frequent item set mining, sequential pattern mining, maximum pattern mining, and closed pattern mining. Most of them are proposed for the purpose of developing efficient mining algorithms to find particular patterns within a reasonable and acceptable time frame. With a large number of patterns generated by using data mining approaches, how to effectively use and update these patterns is still an open research issue.

We focus on the development of a knowledge discovery model to effectively use and update the discovered patterns and apply it to the field of text mining. Text mining is the

discovery of interesting knowledge in text documents. It is a challenging issue to find accurate knowledge (or features) in text documents to help users to find what they want. In the beginning, Information Retrieval (IR) provided many term-based methods to solve this challenge, such as Rocchio and probabilistic models [1], rough set models [2], BM25 and support vector machine (SVM) [3] based filtering models. The advantages of term based methods include efficient computational performance as well as mature theories for term weighting, which have emerged over the last couple of decades from the IR and machine learning communities. However, term based methods suffer from the problems of polysemy and synonymy, where polysemy means a word has multiple meanings, and synonymy is multiple words having the same meaning. The semantic meaning of many discovered terms is uncertain for answering what users want.

Over the years, people have often held the hypothesis that phrase-based approaches could perform better than the term based ones, as phrases may carry more "semantics" like information. This hypothesis has not fared too well in the history of IR [4], [5], [6]. Although phrases are less ambiguous and more discriminative than individual terms, the likely reasons for the discouraging performance include: 1) phrases have inferior statistical properties to terms, 2) they have low frequency of occurrence, and 3) there are large numbers of redundant and noisy phrases among them [41]. In the presence of these setbacks, sequential patterns used in data mining community have turned out to be a promising alternative to phrases [7], [8] because sequential patterns enjoy good statistical properties like terms.

To overcome the disadvantages of phrase-based approaches, pattern mining-based approaches (or pattern taxonomy models (PTM) [8], [9]) have been proposed, which adopted the concept of closed sequential patterns, and pruned no closed patterns. These pattern mining-based approaches have shown certain extent improvements on the effectiveness There are two fundamental issues regarding the effectiveness of pattern-based approaches: low frequency and misinterpretation. Given a specified topic, a highly frequent pattern (normally a short pattern with large support) is usually a general pattern, or a specific pattern of low frequency. If we decrease the minimum support, a lot of noisy patterns would be discovered. Misinterpretation means the measures used in pattern mining

(e.g., “support” and “confidence”) turn out to be not suitable in using discovered patterns to answer what users want.

The difficult problem hence is how to use discovered patterns to accurately evaluate the weights of useful features (knowledge) in text documents. Over the years, IR has developed many mature techniques which demonstrated that terms were important features in text documents. However, many terms with larger weights (e.g., the term frequency and inverse document frequency (tf*idf) weighting scheme) are general terms because they can be frequently used in both relevant and irrelevant information. For example, term “LIB” may have larger weight than “JDK” in a certain of data collection; but we believe that term “JDK” is more specific than term “LIB” for describing “Java Programming Language”; and term “LIB” is more general than term “JDK” because term “LIB” is also frequently used in C and C++. Therefore, it is not adequate for evaluating the weights of the terms based on their distributions in documents for a given topic, although this evaluating method has been frequently used in developing IR models.

In order to solve the above paradox, this paper presents an effective pattern discovery technique, which first calculates discovered specificities of patterns and then evaluates term weights according to the distribution of terms in the discovered patterns rather than the distribution in documents for solving the misinterpretation problem. It also considers the influence of patterns from the negative training examples to find ambiguous (noisy) patterns and try to reduce their influence for the low-frequency problem. The process of updating ambiguous patterns can be referred as pattern evolution.

The proposed approach can improve the accuracy of evaluating term weights because discovered patterns are more specific than whole documents. We also conduct numerous experiments on the latest data collection, Reuters Corpus Volume 1 (RCV1) and Text Retrieval Conference (TREC) filtering topics, to evaluate the proposed technique. The results show that the proposed technique outperforms up-to-date data mining-based methods, concept-based models and the state-of-the-art term based methods.

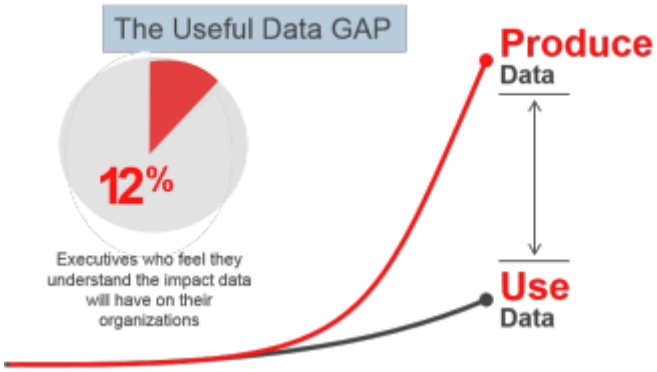
1.1.2 Introduction to analytics

Analytics is the discovery and communication of meaningful patterns in data (corporate, product, channel, and customer). It's not the data but the useful information in and across data. Insight, not hindsight is the essence of predictive analytics.

Analytics is changing expectations and business strategies. A decade ago, GE was in the mode of "the product breaks, we fix it," Today, GE has more than $100 billion in revenue tied to data-driven SLA contracts, whereby it gets paid based on a product — a power plant turbine, a jet engine, a locomotive — being in service. It needs predictive analytics software to help customers avoid downtime and thus make those contracts profitable. (source: InformationWeek)

It's a new world with new rules especially around man+machine interactions. Serving customers with "with few/isolated channels were enough" to now "seamlessly integrating multi-channels, screens, devices". However, change will be slower than people think as legacy systems have to replaced; it's impact greater than people envision.

As we can that the production of data is huge comparing with the use of data. There is a great need to simplify and utilize the data .Basically the data collected is in unstructured form and due to this people ignore the power of data which can be used in their business operations and in daily life.



(Source: Practical Analytics) fig:1

Data mining comes under descriptive analysis .Descriptive analysis is detailing the collection of information. From data scientists, business analysts, marketing executives and front line working adopting data mining directly or indirectly.

Analytics and Big Data will be highly disruptive to some industries, affecting not only revenue and cost structures but also shaking up the core business and operating models. The scope of Analytics is also expanding considerably as human behavior is modeled and expressed mathematically.

1.1.3 Choices Available in Data Mining

(Table 1)

Analysis Type	Real Time, Batch
Method	Text Analytics, Social Network Analysis, Speech Analysis
Data Source	Web, Transaction Data, Biometric Data, Machine Generated
Data Consumers	Human, Business Process, Data Repositories
Content Format	Images,Text,Videos,Documents,Audio
Hardware	Commodity Hardware, State of Art Hardware

1.1.4 Important Phase of Text Mining

1.1.4.1 Frequent and Closed Patterns

Given a termset X in document d,<X>is used to denote the covering set of d, which include all paragraphs dp ∈PS(d).Absolute support is the number of occurrences of X in PS(d).We can the list of paragraphs in the table below which consist of paragraph number and the respected terms and duplicate terms are removed. With min support =50% we can obtain 10 frequent patterns in table 2.

Parapgraph	*Terms*
dp_1	$t_1\ t_2$
dp_2	$t_3\ t_4\ t_6$
dp_3	$t_3\ t_4\ t_5\ t_6$
dp_4	$t_3\ t_4\ t_5\ t_6$
dp_5	$t_1\ t_2\ t_6\ t_7$
dp_6	$t_1\ t_2\ t_6\ t_7$

Table 2(A set of paragraphs)

Not all frequent patterns are useful. For example, pattern {t3; t4} always occurs with term t6 in paragraphs, i.e., the shorter pattern, {t3; t4}, is always a part of the larger pattern, {t3; t4, t6}, in all of the paragraphs. Hence, we believe that the shorter one, {t3; t4}, is a noise pattern and expect to keep the larger pattern, {t3, t4,t6}, only. Given a termset X, its covering set is a subset of paragraphs.

Similarly, given a set of paragraphs we can define its termset, which satisfies

$$\text{Termset}(Y)=\{t \mid \forall dp \in Y \Rightarrow t \in dp\}$$

Patterns can be structured into a taxonomy by using the is-a (or subset) relation. For the example of Table 2, where we have illustrated a set of paragraphs of a document, and the discovered 10 frequent patterns in Table 3 if assuming min sup =50%. There are, however, only three closed patterns in this example. They are {t3,t4, t6},{t1,t2}, and {t6}.

Frequent Pattern	*Covering Set*
$\{\mathbf{t_3}, \mathbf{t_4}, \mathbf{t_6}\}$	$\{dp_2, dp_3, dp_4\}$
$\{t_3, t_4\}$	$\{dp_2, dp_3, dp_4\}$
$\{t_3, t_6\}$	$\{dp_2, dp_3, dp_4\}$
$\{t_4, t_6\}$	$\{dp_2, dp_3, dp_4\}$
$\{t_3\}$	$\{dp_2, dp_3, dp_4\}$
$\{t_4\}$	$\{dp_2, dp_3, dp_4\}$
$\{\mathbf{t_1}, \mathbf{t_2}\}$	$\{dp_1, dp_5, dp_6\}$
$\{t_1\}$	$\{dp_1, dp_5, dp_6\}$
$\{t_2\}$	$\{dp_1, dp_5, dp_6\}$
$\{\mathbf{t_6}\}$	$\{dp_2, dp_3, dp_4, dp_5, dp_6\}$

Table 3(Frequent Pattern and Covering Sets)

1.2 Problem Statement

Analyzing a large dataset by applying techniques from text mining for fraudulent activities. The dataset is in the form emails (text) or the data can be in document form. By applying this the labor intensive manual mining approach can be eliminated which saves time and the activities can be easily judged to take further action on them.

1.3 Approach to Problem

The first step of the project is the identification of the keywords relevant to a particular language or platform. The keywords are identified dynamically by stop word elimination followed by stemming to remove all the common English words used with high frequency in conversations that would not make a desirable keyword for search as they are merely used for the formation of the query statement.

Once the above process is done we are ready to enter the actual piece of work, Patterns can be structured into a taxonomy by using the is-a (or subset) relation There are, however, only three After pruning, some direct "is-a" retaliations may be changed are usually more general because they could be used frequently in both positive and negative documents; and larger patterns, for example pattern ft3; t4; t6g, in the taxonomy are usually more specific since they may be used only in positive documents.

The semantic information will be used in the pattern taxonomy to improve the performance of using closed patterns in text mining .To improve the efficiency of the pattern taxonomy mining, an algorithm, SPMining, was proposed to find all closed sequential patterns, which used the well-known Apriori property in order to reduce the searching space.

The tools that would be involved in implementing the above would be:

- NetBeans IDE 7.4
- Rstudio (parallel work)

Technologies used would be:

- MySQL
- HTML
- Java

Java is used to implement the modules with MySQL for database connectivity .It is used because of its versatility and hassle free connection. It makes handling database easily. UI is made with java swing as of now.

CHAPTER 2 LITERATURE SURVEY

2.1. Summary of Papers

Title of Paper	Effective Pattern Discovery for Text Mining
Authors	Ning Zhong, Yuefeng Li, and Sheng-Tang Wu
Year of Publication	2012
Publishing Details	IEEE Transactions on Knowledge and Data Engineering,Vol.24 No.1
Summary	Many data mining techniques have been proposed for mining useful patterns in text documents. However, how to effectively use and update discovered patterns is still an open research issue, especially in the domain of text mining. Since most existing text mining methods adopted term-based approaches, they all suffer from the problems of polysemy and synonymy. Over the years, people have often held the hypothesis that pattern (or phrase)-based approaches should perform better than the term-based ones, but many experiments do not support this hypothesis. This paper presents an innovative and effective pattern discovery technique which includes the processes of pattern deploying and pattern evolving, to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information. Substantial experiments on RCV1 data collection and TREC topics demonstrate results successfully.

Title of Paper	Review on Text Mining Algorithms
Authors	Mrs. Sayantani Ghosh, Mr. Sudipta Roy, and Prof. Samir K. Bandyopadhyay
Year of Publication	2012
Publishing Details	International Journal of Advanced Research in Computer and Communication Engineering Vol.1,Issue 4
Summary	As we enter the third decade of the World Wide Web (WWW), the textual revolution has seen a tremendous change in the availability of online information. Finding information for just about any need has never been more automatic—just keystroke or mouse click away .It can be viewed as one of a class of nontraditional Information Retrieval (IR) strategies which attempt to treat entire text collections holistically, avoid the bias of human queries, objectify the IR process with principled algorithms, and "let the data speak for itself." These strategies share many techniques such as semantic parsing and statistical clustering, and the boundaries between them are fuzz. In this paper different existing Text Mining Algorithms i.e Classification Algorithm, Association Algorithm, Clustering Algorithm is briefly reviewed, stating the merits / demerits of the algorithms. In addition some alternate implementation of the algorithms is proposed. Finally the logic of these algorithms are , merged to generate an algorithm which will perform the task of Classification of a data set into some predefined classes, establish relationship between the classifier.

Title of Paper	Data Mining and Predictive Analytics in Public Safety and Security
Authors	Colleen McCue
Year of Publication	2006
Publishing Details	IEEE Computer Society
Summary	Used for many years in the business community, data mining and predictive analytics are finding new roles in areas outside business. Also referred to as *knowledge discovery* or *sense making* tools these analytical processes can help analysts, managers, and operational personnel identify actionable patterns and trends in data. Briefly, data mining is "[a]n information extraction activity whose goal is to discover hidden facts contained in the databases of many sites. In other words, data mining involves the systematic analysis of data using automated methods in an effort to identify meaningful or otherwise interesting patterns, trends, or relationships in the data.Crime and criminal behavior, including the most aberrant or heinous crimes, frequently can be categorized and modeled— a characteristic used successfully in the apprehension of serial killers and child predators, as well as drug dealers,robbers, and thieves. So it's no surprise that data mining and predictive analytics are rapidly gaining acceptance and use in the applied public safety, security,and intelligence .

Title of Paper	Next Step for Learning Analytics
Authors	Jinan Fiaidhi
Year of Publication	2014
Publishing Details	IEEE Computer Society
Summary	The paper presents description about Text analytics applies a variety of natural language processing analysis techniques along with linguistics, statistical, and data-mining techniques to extract concepts and patterns that can be applied to categorize and classify textual documents. It also attempts to transform the unstructured information into data that can be used with more traditional learning analytics techniques. Finally, it helps identify meaning and relationships in large volumes of information. However, there is no single method appropriate for all text analysis tasks. Learning analytics approaches must take several different perspectives and accommodate different data sources.The ideal vision for learning analytics is to integrate analytics for both structured and unstructured data (mainly of textual nature of a comprehensive learning analytics architecture.

Title of Paper	Predictive Analytics
Authors	Ravi Kalakota
Year of Publication	2014
Web link	http://practicalanalytics.wordpress.com/predictive-analytics-101/
Summary	This article presents the various techniques which are changing the world business and turning them into valuable and actionable information like summation, predictive, descriptive and prescriptive where descriptive is of our interest where data mining comes into action.It's a new world with new rules especially around man +machine interactions. "how companies find customers " to "how customers find companies today" is evolving. Serving customers with "with few/isolated channels, screens,devices".How demographic segmentation was enough to complex behavious segmentation to drive1:1personalization . The end goal of predictive analytics = [Better outcomes, smarter decisions, actionable insights, relevant information]. How you execute this varies by industry and information supply chain *(Raw Data -> Aggregated Data -> Contextual Intelligence -> Analytical Insights (reporting vs. prediction) -> Decisions (Human or Automated Downstream Actions)).*

2.2. Integrated Summary of Literature Studied

Literature survey of the research papers helped us to understand the need for finding the Effective Pattern with use in identifying fraudulent activity and the various techniques which can help make an accurate and effective system for analyzing inoxes for fraudulent activity. Inbox data is huge for a operating company and need to be structured for the extraction of useful pattern. Also there are techniques which aim to solve the problems described through other data mining techniques with varying results.

Studying various papers revealed the existence of various approaches to solve the above stated problem by using term weight approach .But there are certain issues which needs to be resolved .

Also Research papers have been published on Stop word elimination and Porter stemming and how they help in reduction of the search space by eliminating stop words which are general English words. Porters Stemming helps in suffix removal by converting a word into its root and improves the performance of an IR system will be improved if term groups are conflated into a single term. This may be done by removal of the various suffixes -ED, -ING, -ION, -IONS to leave the single term. In addition, the suffix stripping process will reduce the total number of terms in the IR system, and hence reduce the size and complexity of the data in the system, which is always advantageous.

We know that multiple data mining methods have been developed for finding useful patterns in contents like PDF files, text files. Current paper addresses the problem of making text mining results more effective to humanities scholars, journalists, intelligence analysts, and other researchers. To use effective and bring to up to date discovered patterns is still an open research task, especially in the domain of text mining. Text mining is the finding of very interesting knowledge (or features) in the text documents. It is a very difficult to find exact knowledge (or features) in text documents to help users what they actually want. A d-pattern mining technique is discovered. It evaluates specificities of patterns and then evaluates term-weights according to the distribution of terms in the discovered patterns. It solves Misinterpretation Problem.

The study of the associated literature about the above mentioned topics aided in developing an integrated approach as to how to develop an accurate system for mining. Effective pattern and present it to the user with the most appropriate answer where the user can decide the further action.

.

CHAPTER 3 ANALYSIS, DESIGN & MODELLING

3.1. Overall Description of Project

The primary aim of the project is to create a tool that analyze the inbox text for fraudulent activity. Inbox consist of various message in different context and processing the data to find an effective pattern form which the authorized person can see and take further action.

The dataset is loaded and the user is to retrieve one of the document and the document is given to next process and that process is called preprocessing. There are two processes done before next module start taking place. First one is the stop words removal and then text stemming. Stop words are words which are filtered out prior to, or after, processing of natural language data. Stemming is the process for reducing inflected (or sometimes derived) words to their stem base or root form. It generally a written word forms.

After the data is loaded and preprocessing is done pattern taxonomy model is applied where each paragraph is considered to be each document. In each document, the set of terms are extracted. The terms, which can be extracted from set of positive documents. After the pattern deploying and noise removal is done.

We focus on the development of a knowledge discovery model to effectively use and update the discovered patterns and apply it to the field of text mining. Although phrases are less ambiguous and more discriminative than individual terms, the likely reasons for the discouraging performance include: 1) phrases have inferior statistical properties to terms, 2) they have low frequency of occurrence, and 3) there are large numbers of redundant and noisy phrases among them. In the presence of these setbacks, sequential patterns used in data mining community have turned out to be a promising alternative to phrases because sequential patterns enjoy good statistical properties like terms.

For the purpose of front end java swing are used at initial and further provision of deploying it on the web is also a good option because of its dynamic nature. Rshiny application can be used in this respect where the work is done in Rstudio.

3.2. Specific Requirements

3.2.1 Functions

Module 1

Input Data

Input data	Loaded input dataset**.**
Functionality	To load the list of all document and user to retrieve one of the documents.

Module 2

Stopword Elimination

Input data	The dataset retrieved.
Output data	Stopword eliminated text.
Functionality	It removes the stopwords and save the text file.

Module 3

Text Stemming

Input data	Stopwords eliminated text file.
Output data	Text where stemming is done .
Functionality	Perform suffix stemming on the text file and store them .

Module 4

Pattern Taxonomy Process

Input data	Stemmed data.
Output data	Set of terms from positive document.
Functionality	To spilt the document in paragraphs then the set of terms are extracted.

Module 5

Pattern deploying

Input data	Set of terms.
Output data	Term support.
Functionality	The d-pattern algorithm is used to discover all pattern in the documents and term support is calculated.

Module 6

Pattern Evolving

Input data	Terms of pattern extracted
Output data	Corrected pattern terms.
Functionality	To remove the noise and shuffle if required.

3.2.2 Performance Requirements

Performance requirements deals with both the static and the dynamic numerical requirements placed on the software or on human interaction with the software as a whole. Static numerical requirements involve The number of terminals to be supported: There shall be a single terminal when the application would be tested on localhost and many when tested on the web (if possible).

3.2.3 Logical Database Requirements

This specifies the logical requirements for any information that is to be placed into a database.

3.2.3.1. Types of information used by various functions:
Strings and numbers.

3.2.3.2. Accessing capabilities:
Only the system shall access the logical database and all the results of various tests and the user has no rights and capabilities for accessing the database.

3.2.3.3. Integrity constraints:
Integrating the information of separate entities in a particular scenario shall be taken care of.

3.2.3.4. Data retention requirements:
Data retention of all the necessary information in the database for further processing and later access.

3.2.4. Design Constraints

I faced the following constraints in implementation of the implemented phase

- The data in the form of text file is difficult to convert from csv file working in R to working in Java environment.
- Storage requirements should be meet and storage space should be used efficiently.
- Integrating the modules.

3.2.5. Software Attributes

3.2.5.1. Reliability:

Java environment is very reliable and have its own unique attributes in terms of implementation. MySQL is also very reliable in terms of database .

3.2.5.2. Availability:

Several checkpoints shall be made in the development to ensure defined availability level of the system. These checkpoints shall occur after every main module. Our system insures check pointing as it follows a step wise approach.

3.2.5.3. Security

Restrict communications between some areas of the program such as restricting the database access from the user and allow him to perform only end user functions. Each main module, i.e. scrap, match have completely different functionalities.

3.2.5.4. Maintainability

This specifies to the attributes of the software that relate to the ease of maintenance of the software. In the project, modularity is taken into consideration and separate modules like stopword, stemming and patter taxonomy shall be implemented. upgradation and revising the current version an easy task with the modular structure defined.

3.2.5.5. Portability:

Current the tool is in implementation phase and not portable yet.

3.3 Design Diagrams

3.3.1 Use Case Diagram

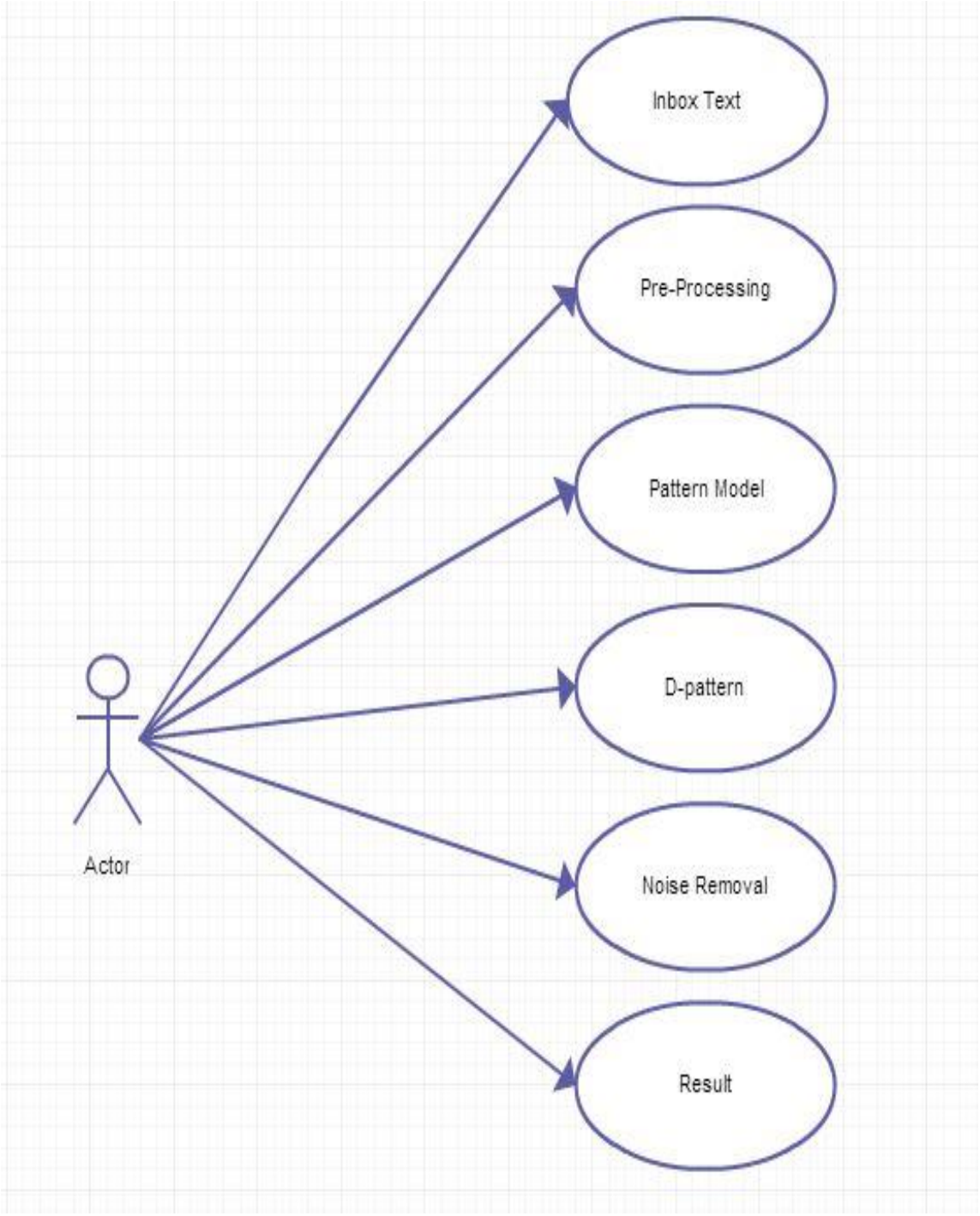


Fig: 2

3.3.2 Dataflow Diagram

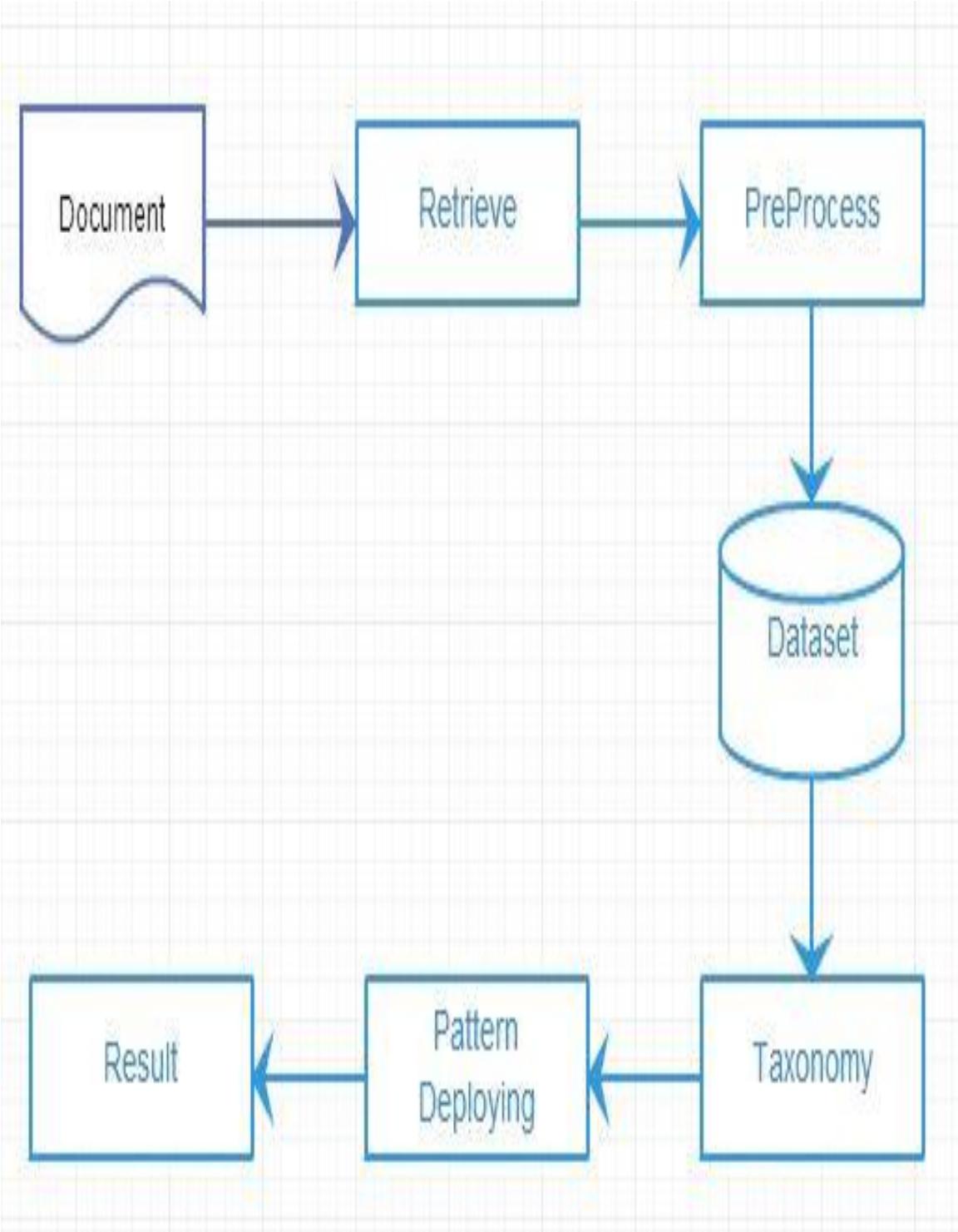


Fig:3

3.3.3 Sequence Diagram

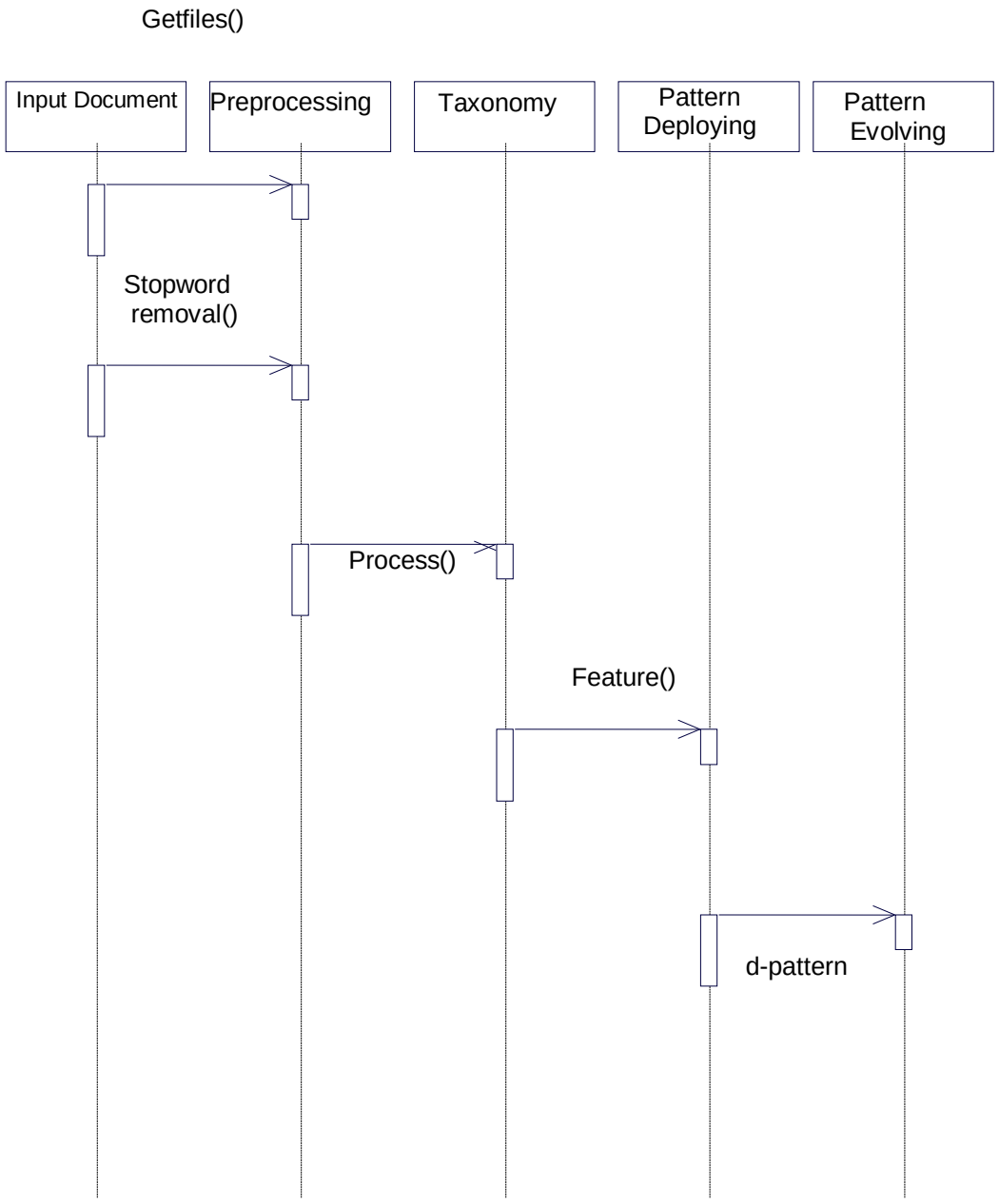


Fig:4

3.3.4 Activity Diagram

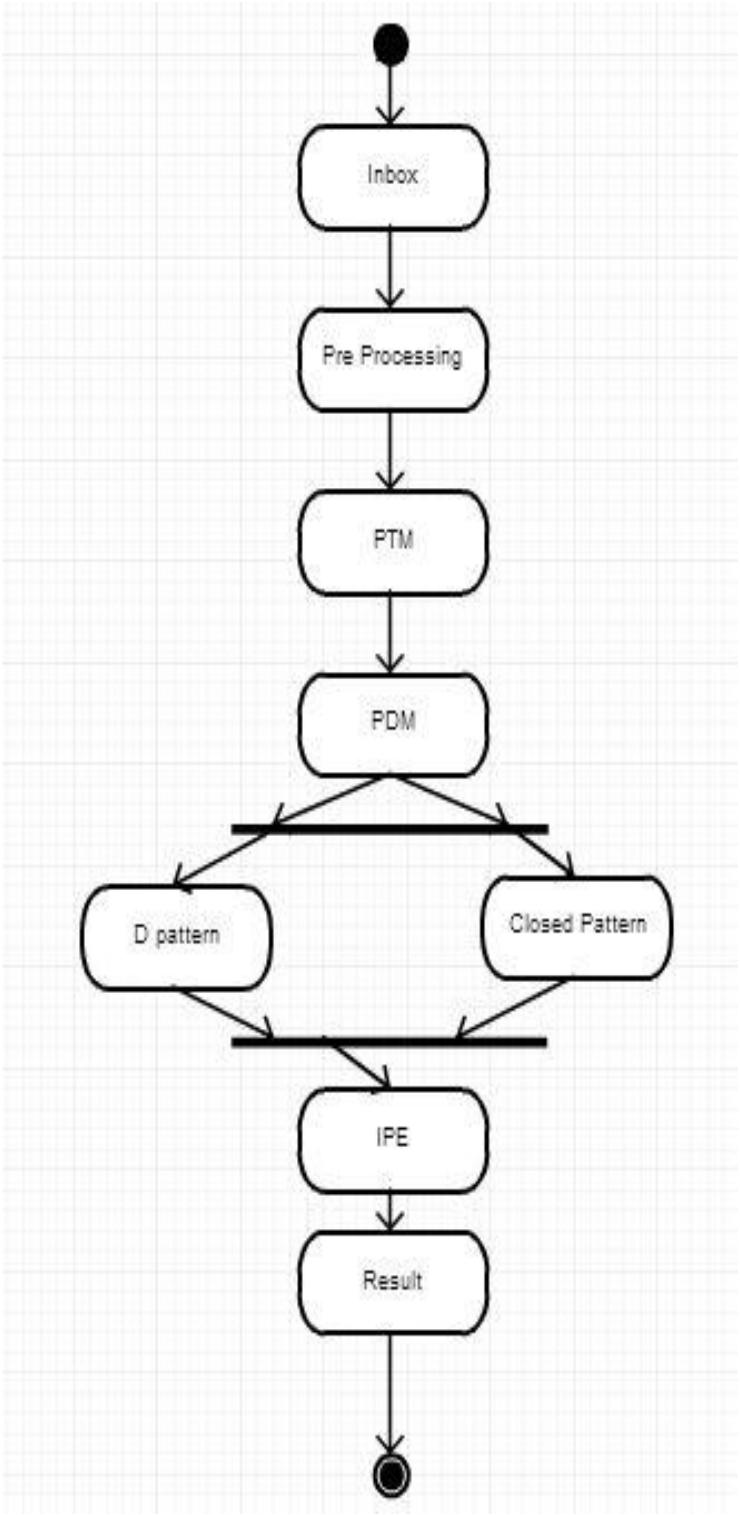


Fig:5

CHAPTER 4 IMPLEMENTATION

4.1 Implementation Details

fig:6

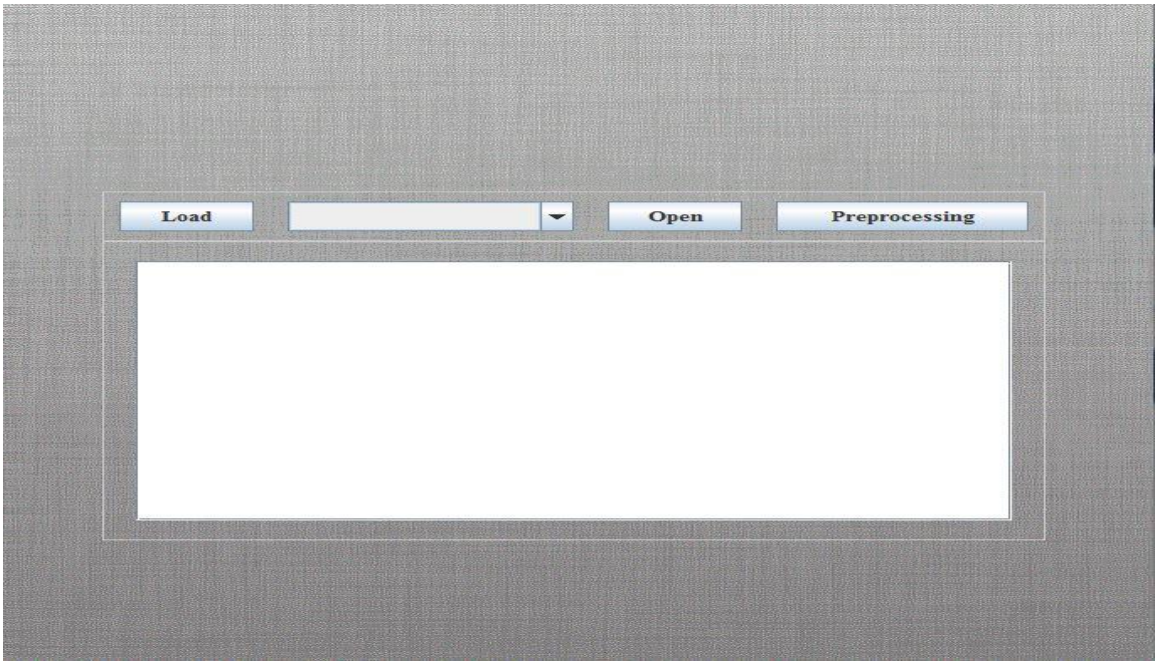


Fig:7

In the figure.6 we can see the home screen of the main.java file and after clicking the start button we come to the UI of retrieval module (figure.7) which consist of load, drop down menu for loading dataset, open button for showing the selected dataset and preprocessing button for navigating to preprocessing module.

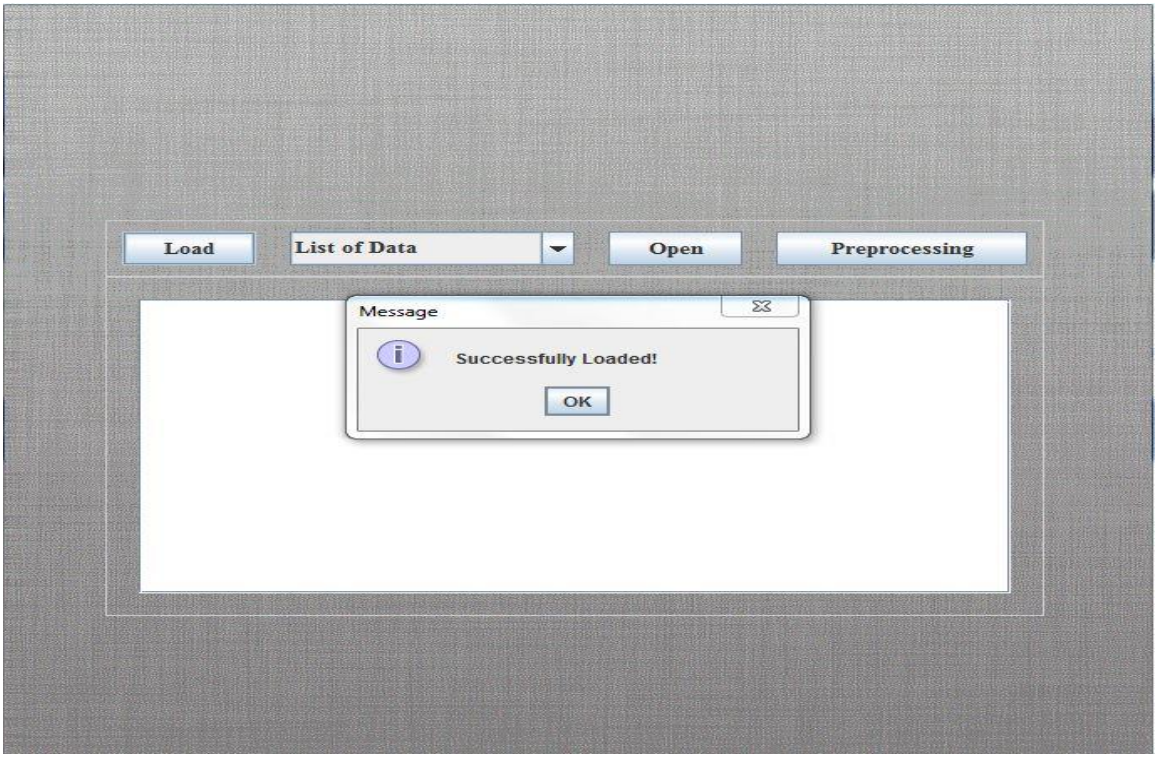


Fig:8

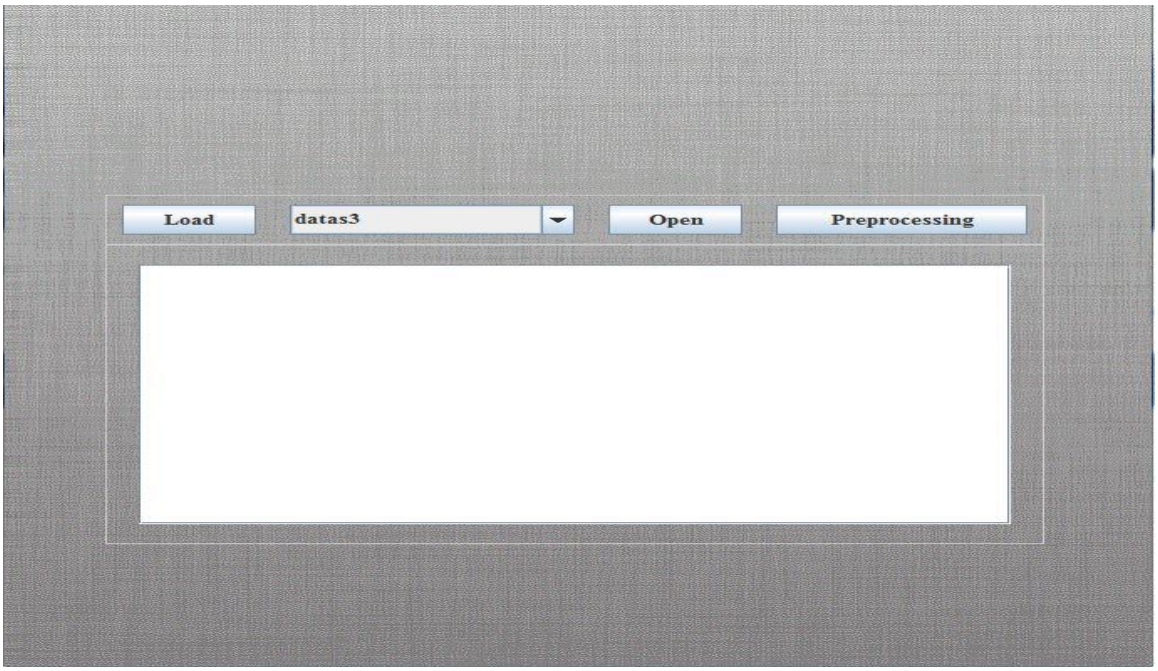


Fig:9

In figure.8 we can see that list of data is in the dropdown menu and notification for loading of list is generated and(fig.9) we have loaded a dataset 3 .

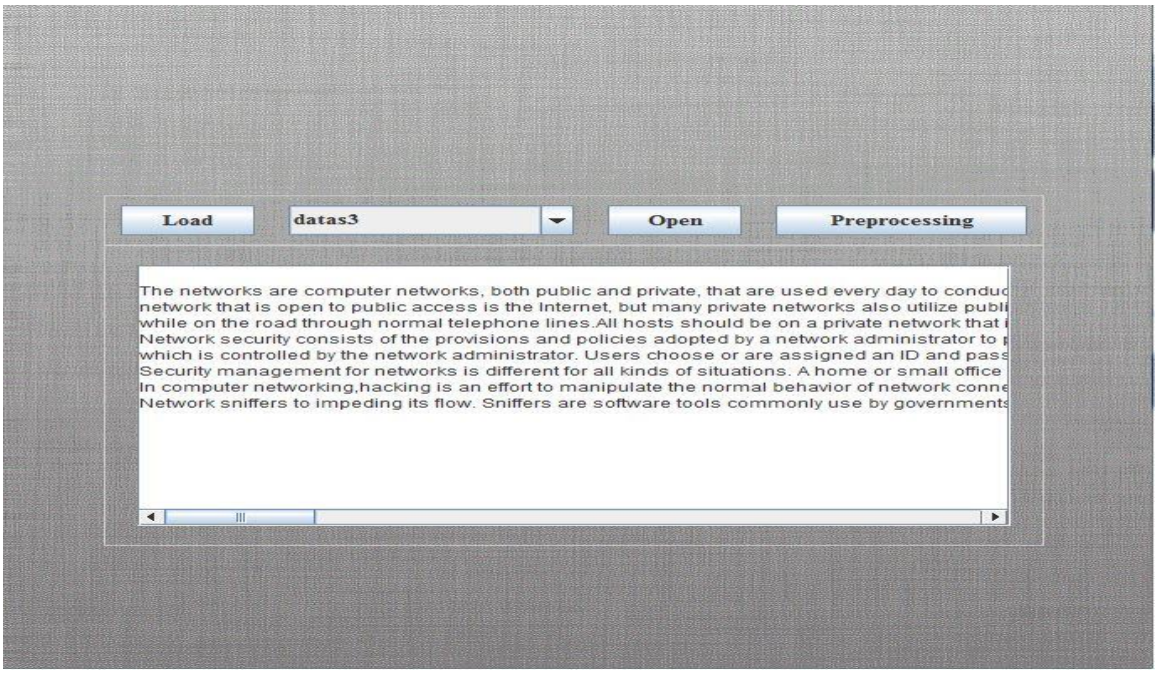

Load
datas3
Open
Preprocessing
The networks are computer networks, both public and private, that are used every day to conduc
network that is open to public access is the Internet, but many private networks also utilize publi
while on the road through normal telephone lines.All hosts should be on a private network that i
Network security consists of the provisions and policies adopted by a network administrator to p
which is controlled by the network administrator. Users choose or are assigned an ID and pass
Security management for networks is different for all kinds of situations. A home or small office
In computer networking,hacking is an effort to manipulate the normal behavior of network conne
Network sniffers to impeding its flow. Sniffers are software tools commonly use by governments

Fig:10

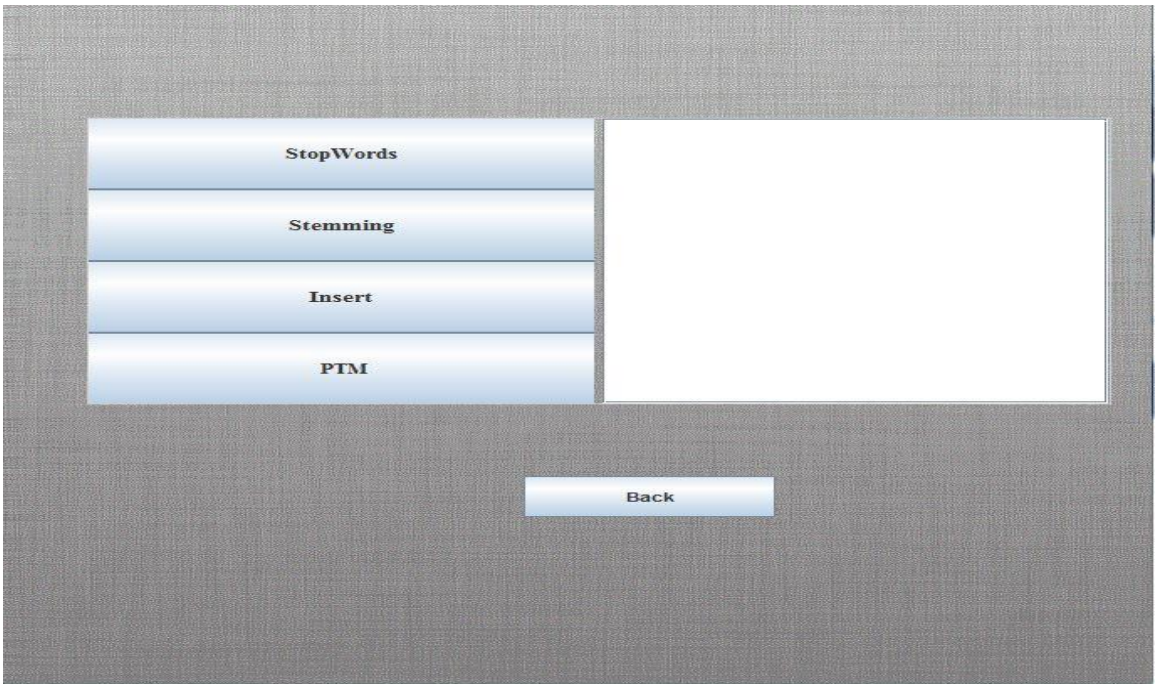

StopWords
Stemming
Insert
PTM
Back

Fig.11

In fig 10 the dataset is opened and can be seen in the text field and when we click on preprocessing it navigates to preprocessing module which consists of stopwords button, stemming button ,Insert button for the insertion of filtered data(Structured data)in the database for further (figure.11)processing and PTM button which Pattern Taxonomy Model button for navigating to the next module.

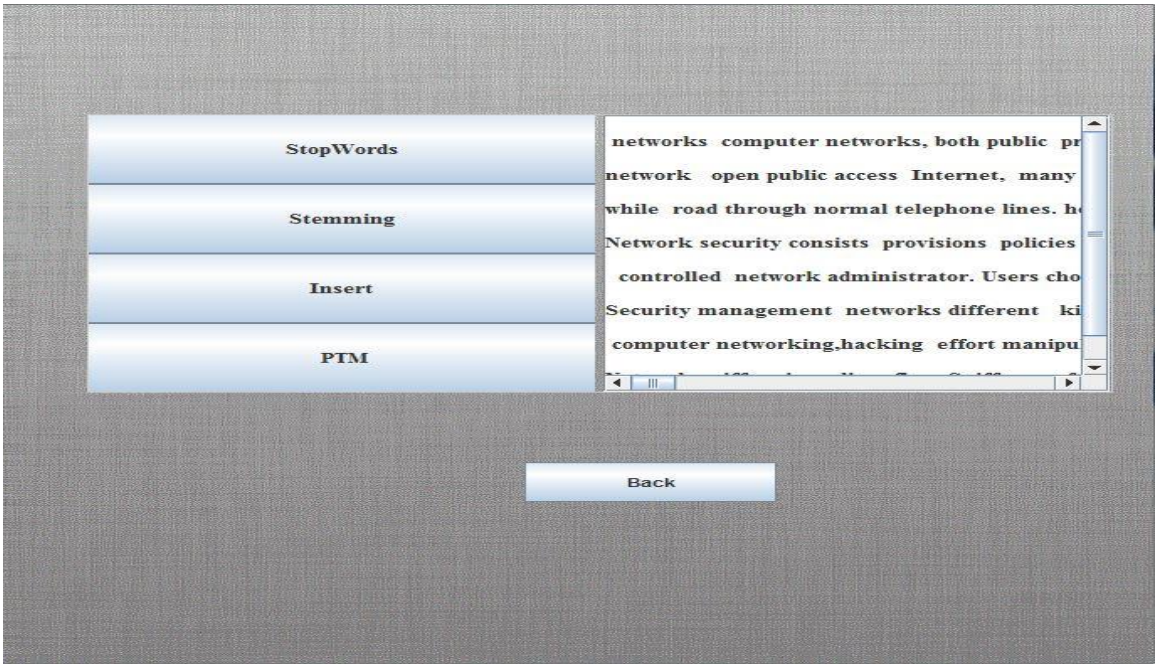


Fig.12

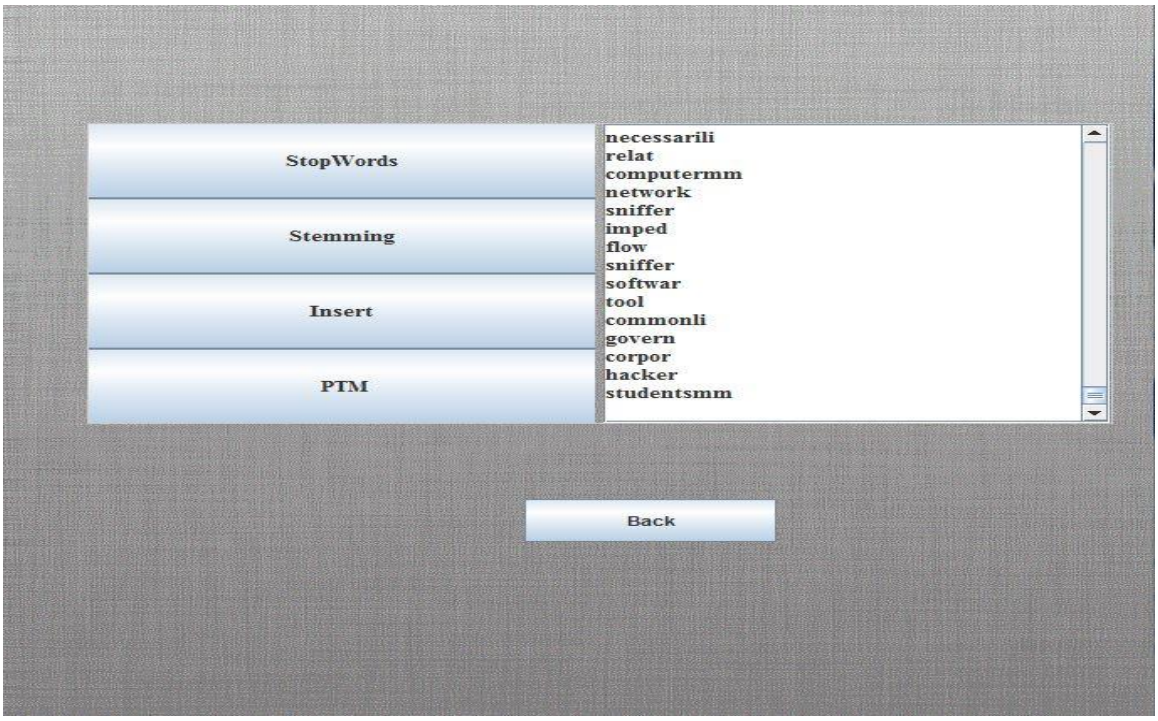

StopWords
Stemming
Insert
PTM
necessarili
relat
computermm
network
sniffer
imped
flow
sniffer
softwar
tool
commonli
govern
corpor
hacker
studentsmm
Back

Fig.13

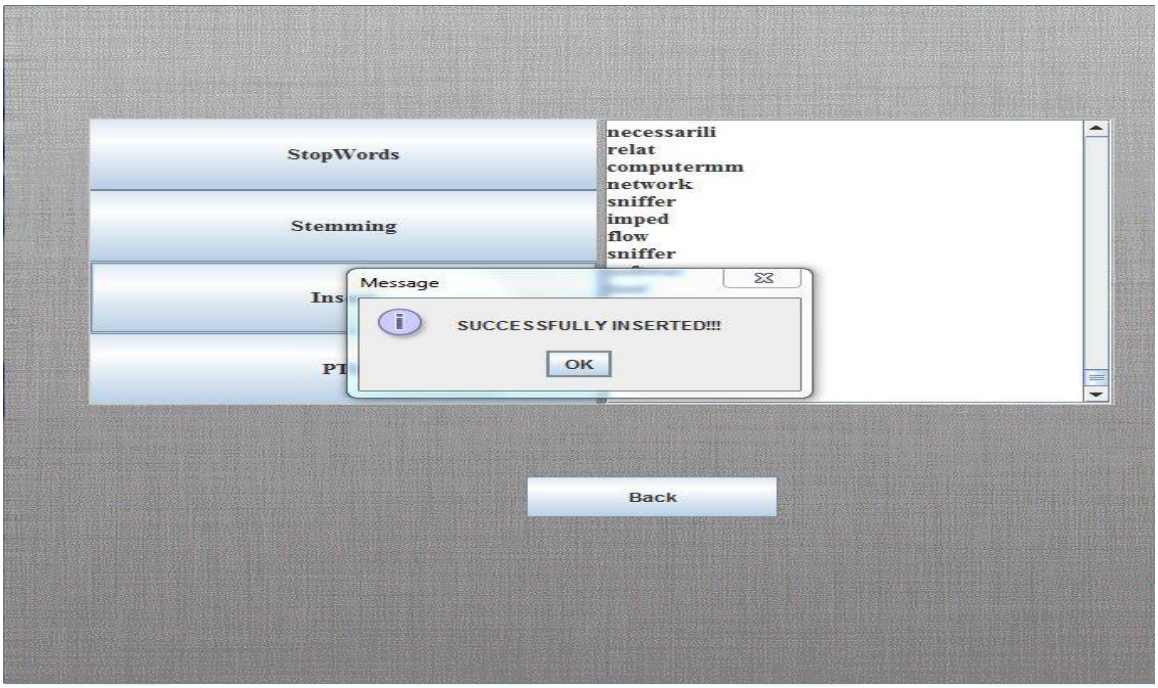

StopWords
Stemming
Ins
PT
necessarili
relat
computermm
network
sniffer
imped
flow
sniffer
Message
SUCCESSFULLY INSERTED!!!
OK
Back

Fig.14

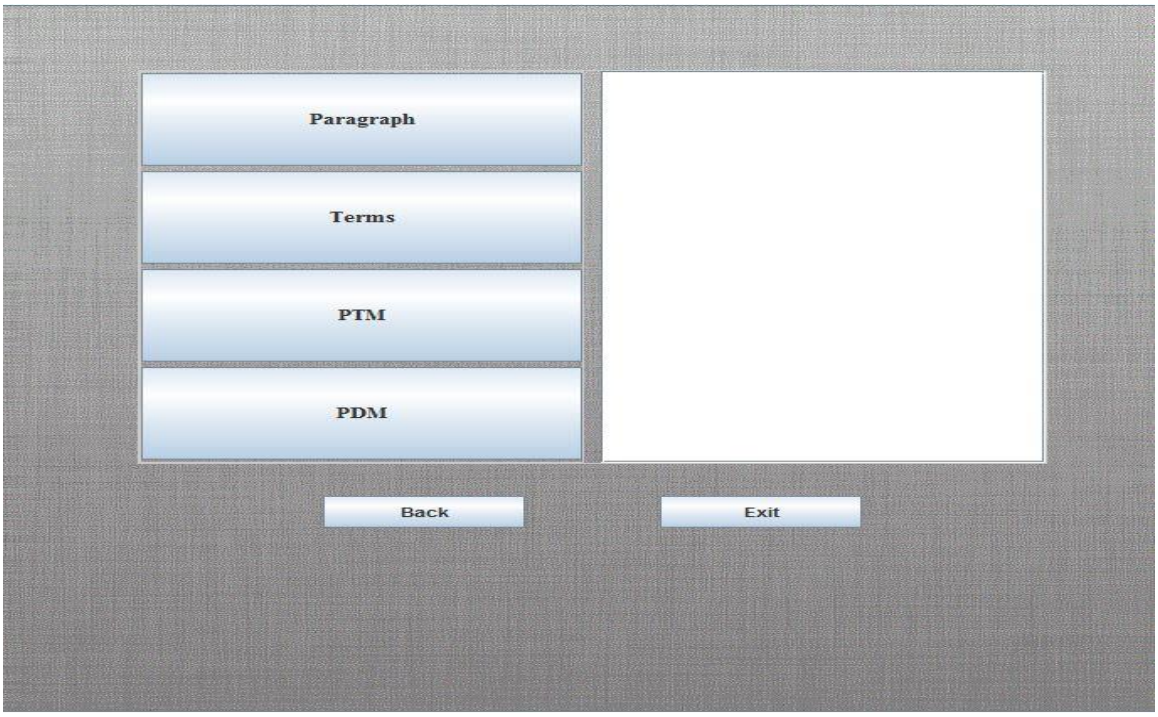
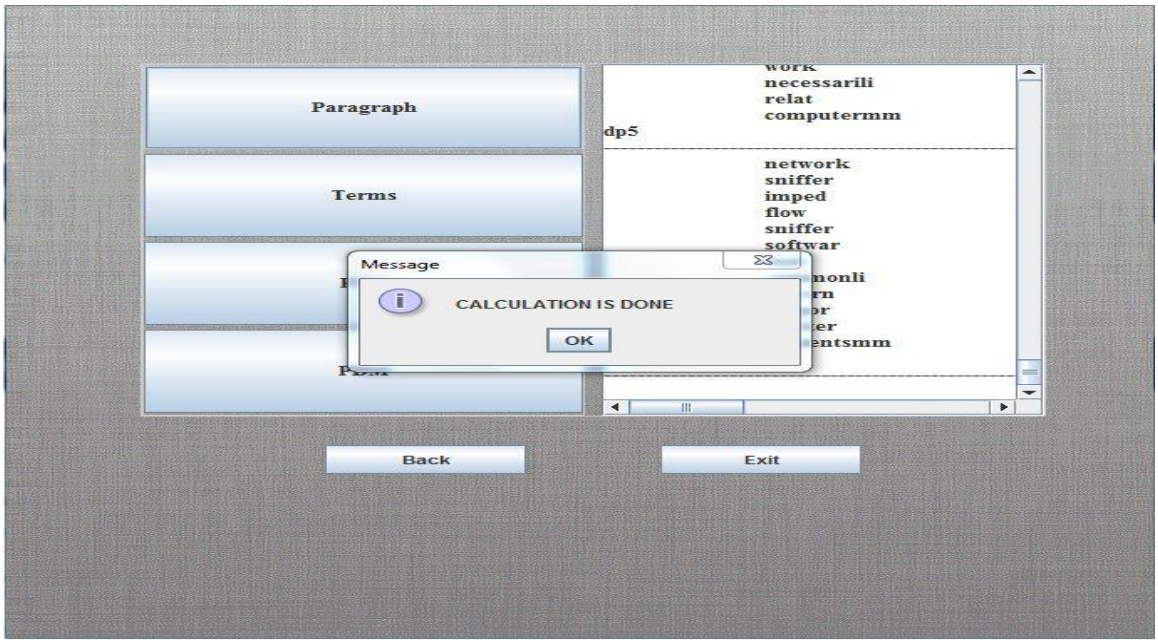
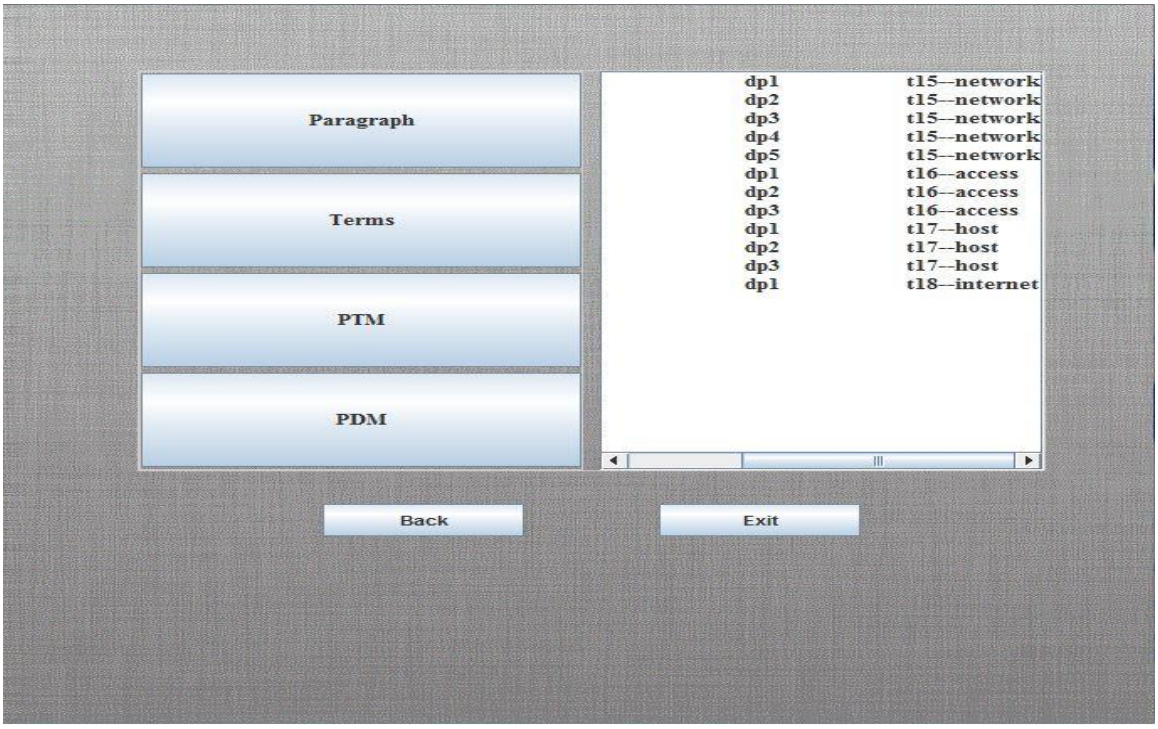


Fig.15

In figure.12 when the stopword button is pressed we can that the text comes filtered with stopwords, then after pressing the stemming button the data gets more filtered and words gets back to there natural form or we can it as root form.(figure.13).In the next action when insert is clicked, the data is inserted(fig.14) in the database and in the next step when clicked on PTM it navigate to the next UI(fig.15)


Paragraph
Terms
Message
CALCULATION IS DONE
OK
necessarili
relat
computermm
dp5
network
sniffer
imped
flow
sniffer
Back
Exit
Fig.16


Paragraph
Terms
PTM
PDM
dp1 t15--network
dp2 t15--network
dp3 t15--network
dp4 t15--network
dp5 t15--network
dp1 t16--access
dp2 t16--access
dp3 t16--access
dp1 t17--host
dp2 t17--host
dp3 t17--host
dp1 t18--internet
Back
Exit
Fig.17

In fig.16 when paragraph button is clicked the data which we have saved earlier is divided into paragraph as we can see dp1,dp2,etc and a notification is generated. The division is the main part of this phase because from here the positive documents can be applied to this data to find the presence of positive words in the data. Here positive words are the words whose presence will ensure the fraudulent activity. In fig.17 we can see that after term button is clicked it shows the paragraph number and term which is found in that.

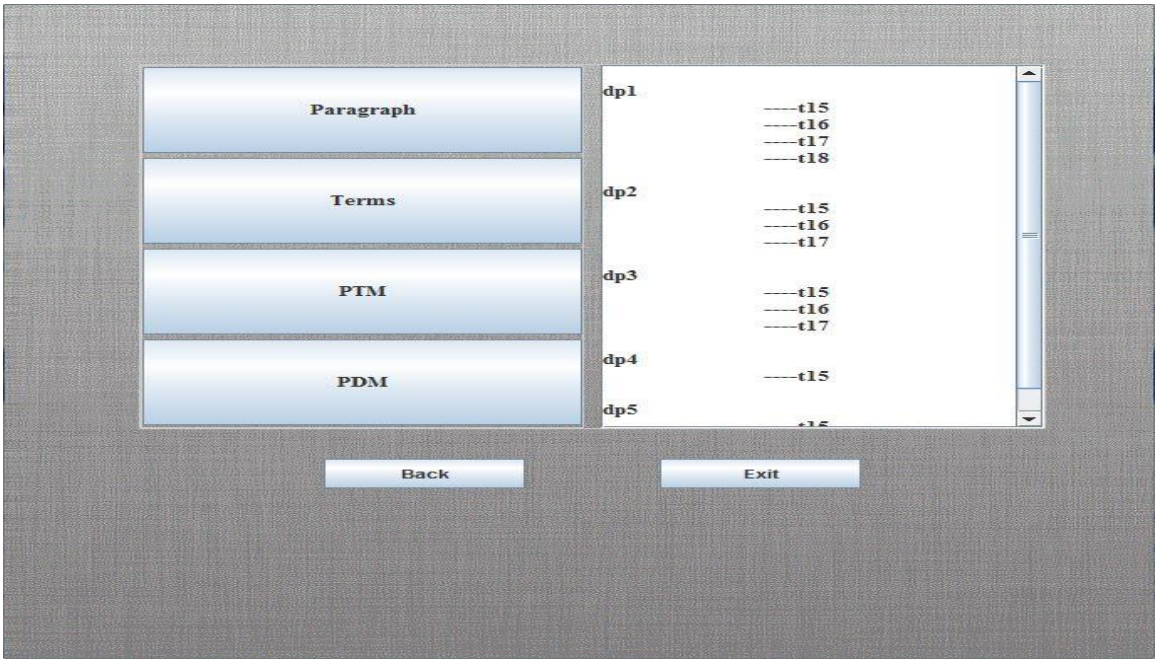


Fig.18

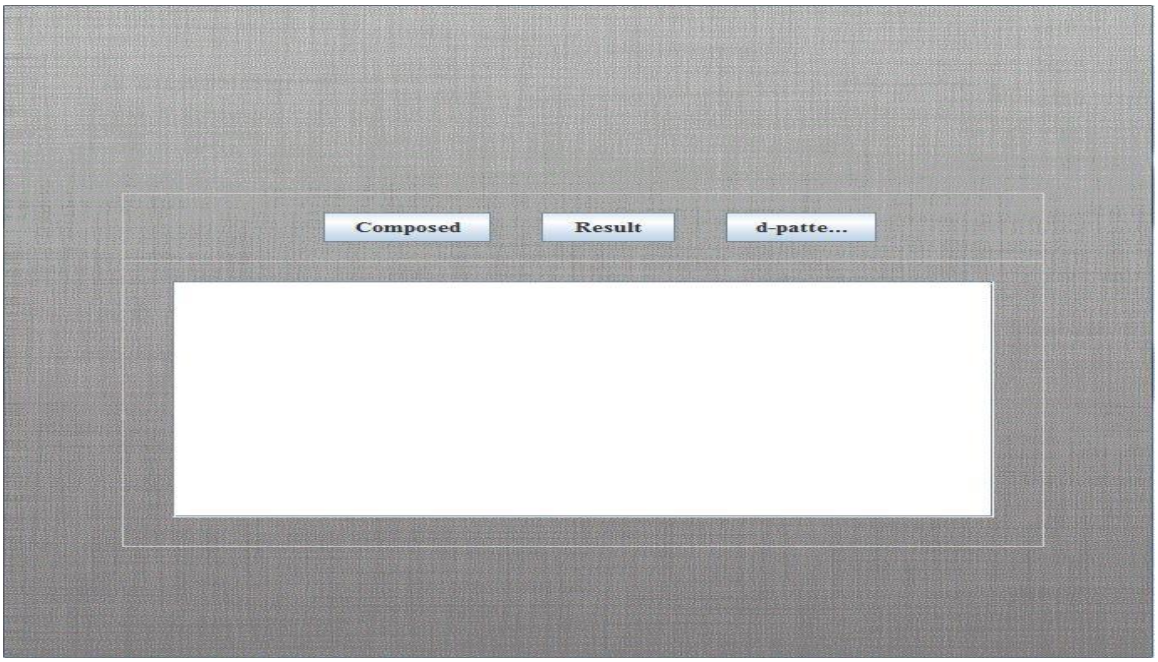
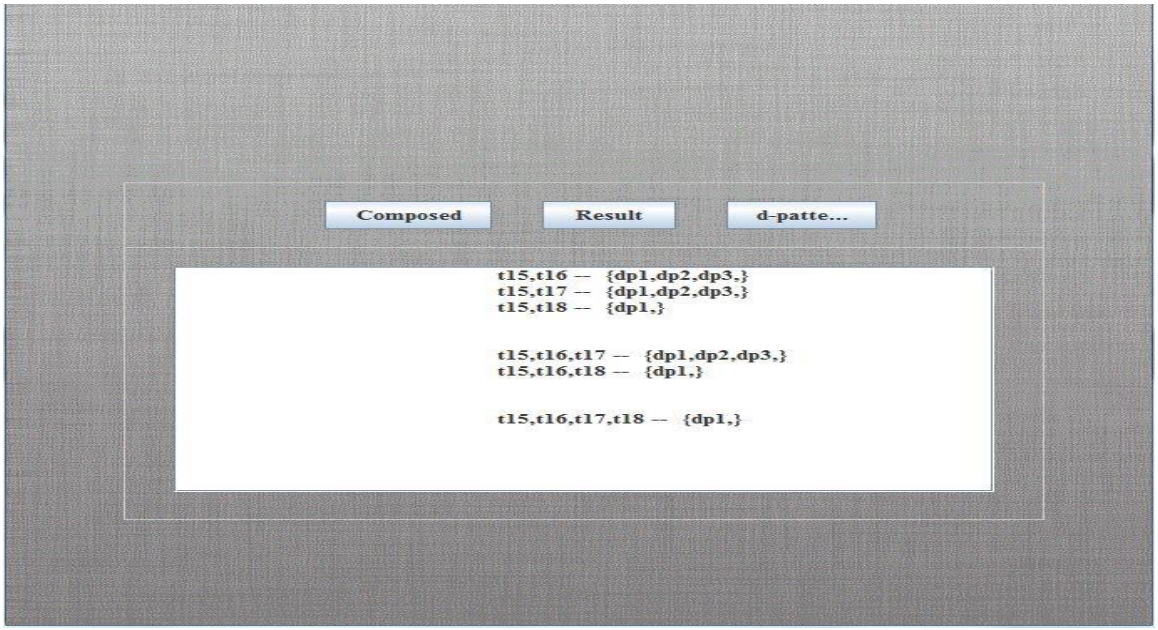
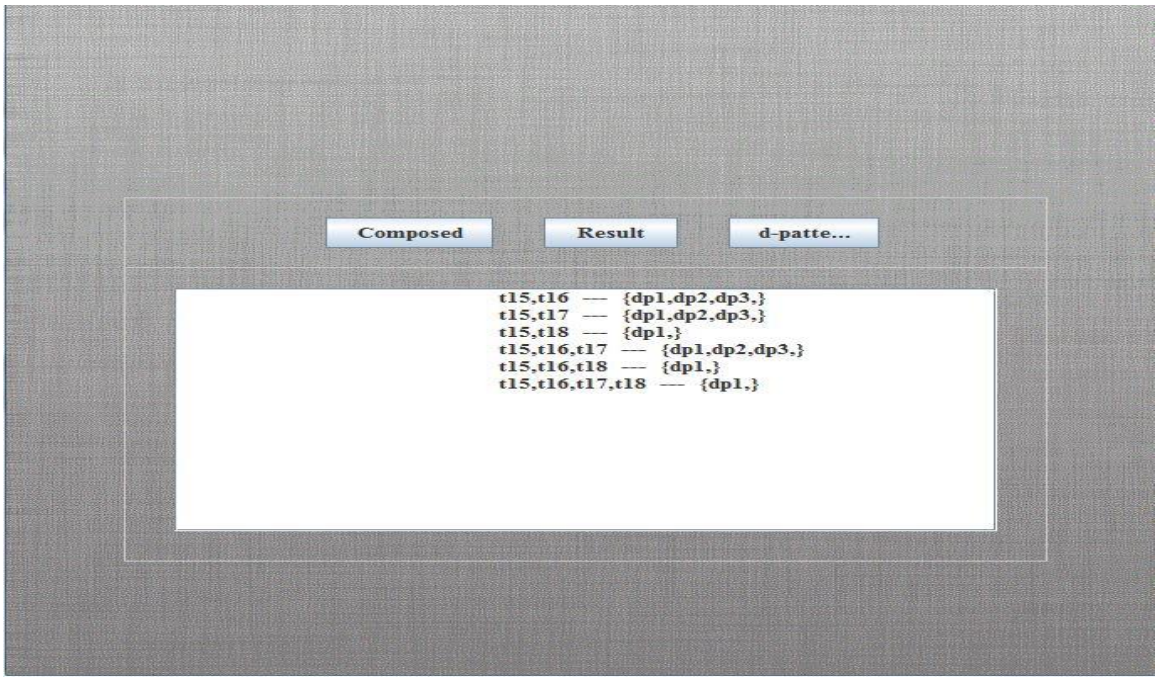


Fig.19

In fig.18 after clicking the PTM button we can see that paragraphs are listed with the terms they contain the set of positive terms and when PDM is clicked we navigate to next UI which consist of composed button, result button and d-pattern.


Composed
Result
d-patte...
t15,t16 -- {dp1,dp2,dp3,}
t15,t17 -- {dp1,dp2,dp3,}
t15,t18 -- {dp1,}
t15,t16,t17 -- {dp1,dp2,dp3,}
t15,t16,t18 -- {dp1,}
t15,t16,t17,t18 -- {dp1,}
Fig.20


Composed
Result
d-patte...
t15,t16 --- {dp1,dp2,dp3,}
t15,t17 --- {dp1,dp2,dp3,}
t15,t18 --- {dp1,}
t15,t16,t17 --- {dp1,dp2,dp3,}
t15,t16,t18 --- {dp1,}
t15,t16,t17,t18 --- {dp1,}
Fig.21

In fig.20 when clicked on composed the composition are calculated as shown in the figure which terms are in which paragraph after deploying d-pattern. After clicking Result button composition get (figure.21) filtered. Clicking d-pattern will take us to new UI.

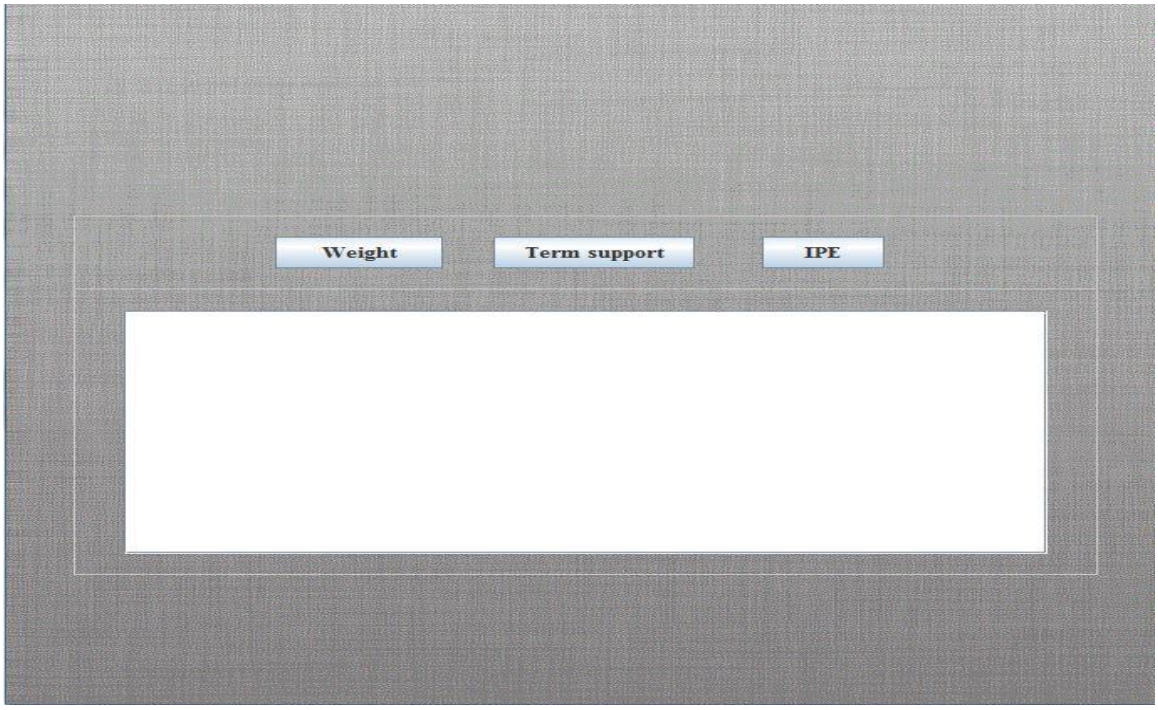


Fig.22

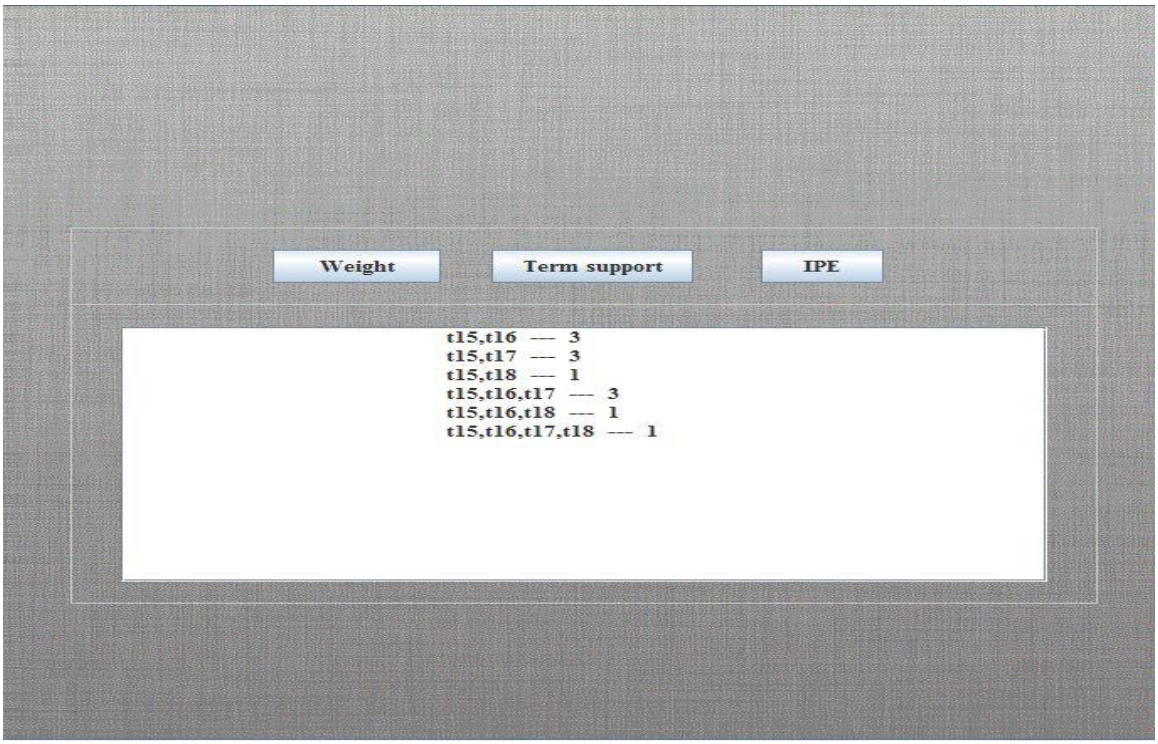
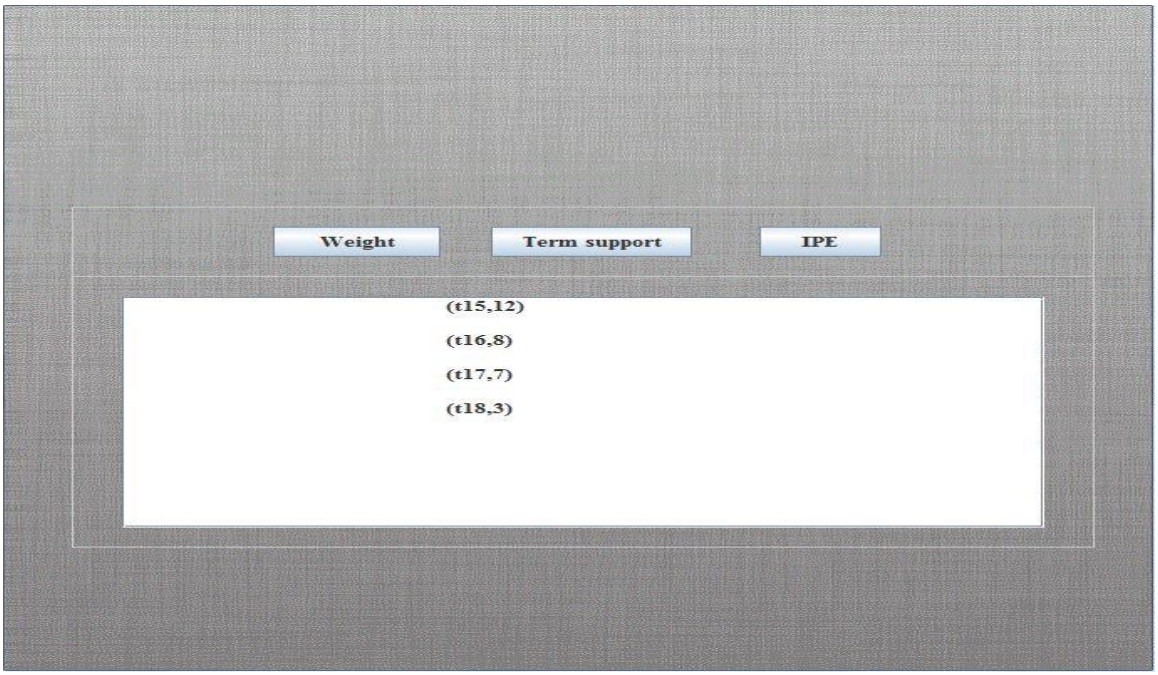
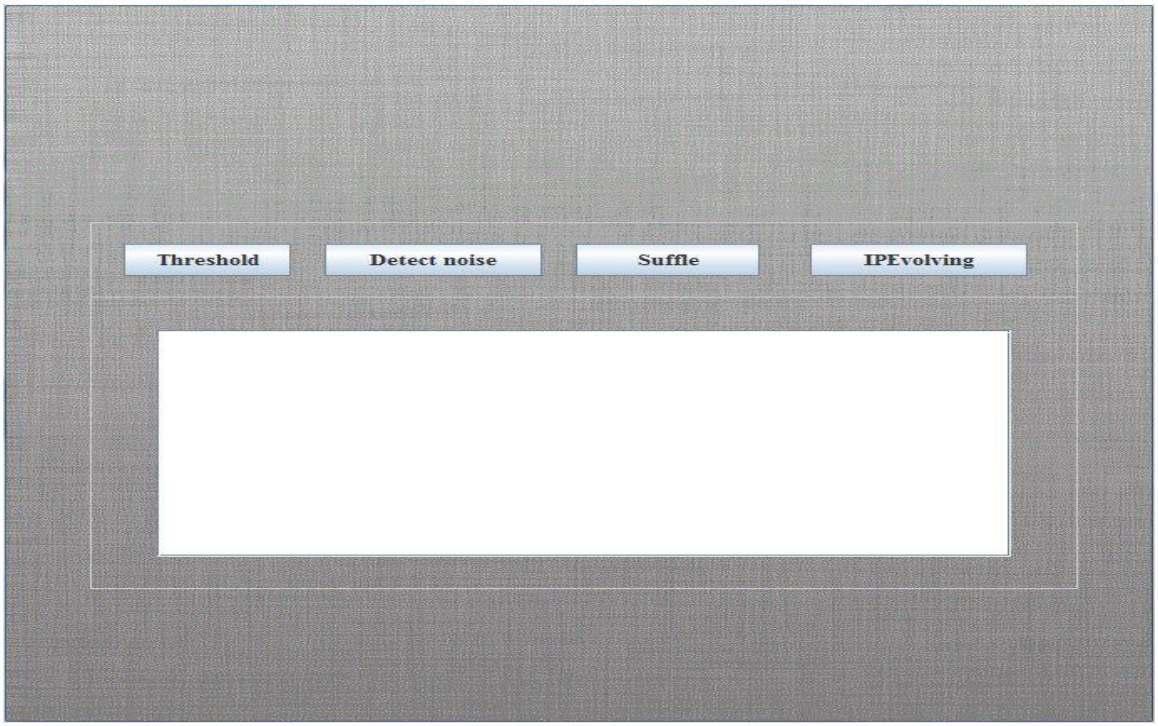


Fig.23

In fig.22 we have button for weight, term support and ipe (inner pattern evolution).In fig.23 when weight button is clicked terms are shown with their weight.


Weight
Term support
IPE
(t15,12)
(t16,8)
(t17,7)
(t18,3)
Fig.24


Threshold
Detect noise
Suffle
IPEvolving
Fig.25

In fig.24 after clicking term support which is calculated with previous weight calculated. We can see the term support in decreasing order. There may be some offender in the data which were identified as positive but were actually negative so there is a need of shuffling so that offender can be obtained and proper result can be retained which will showcase the analyzed pattern of (figure.25) fraudulent activity.

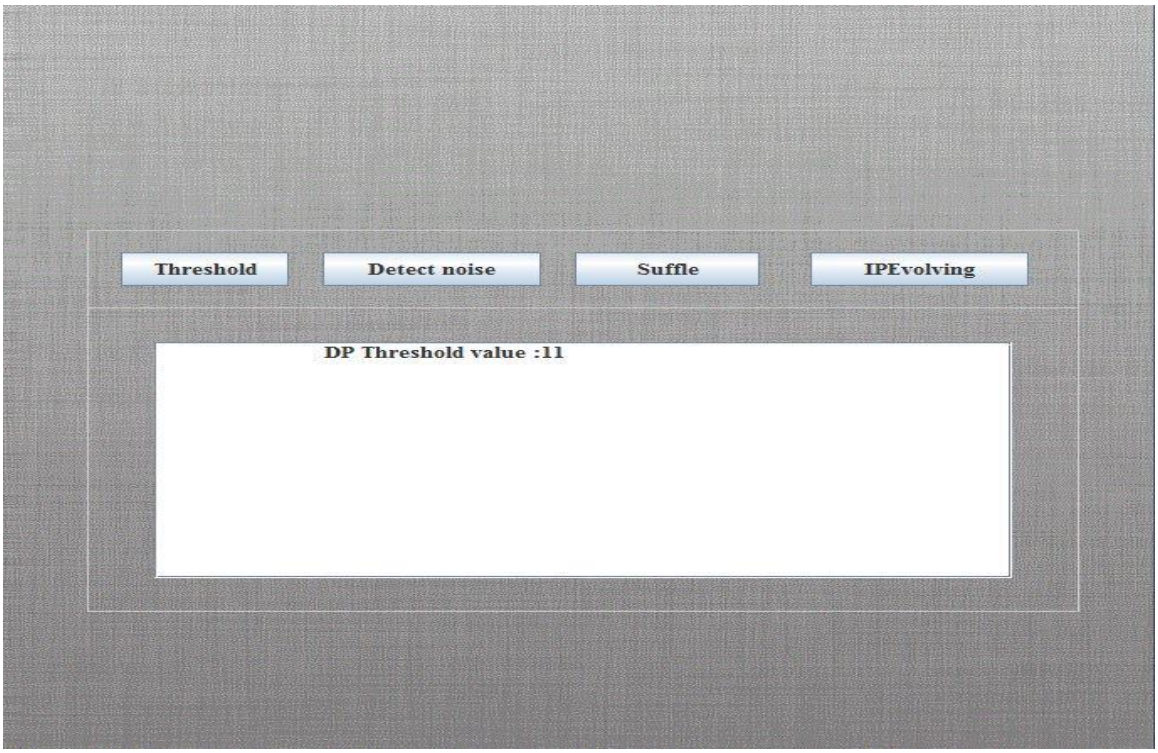


Fig.26

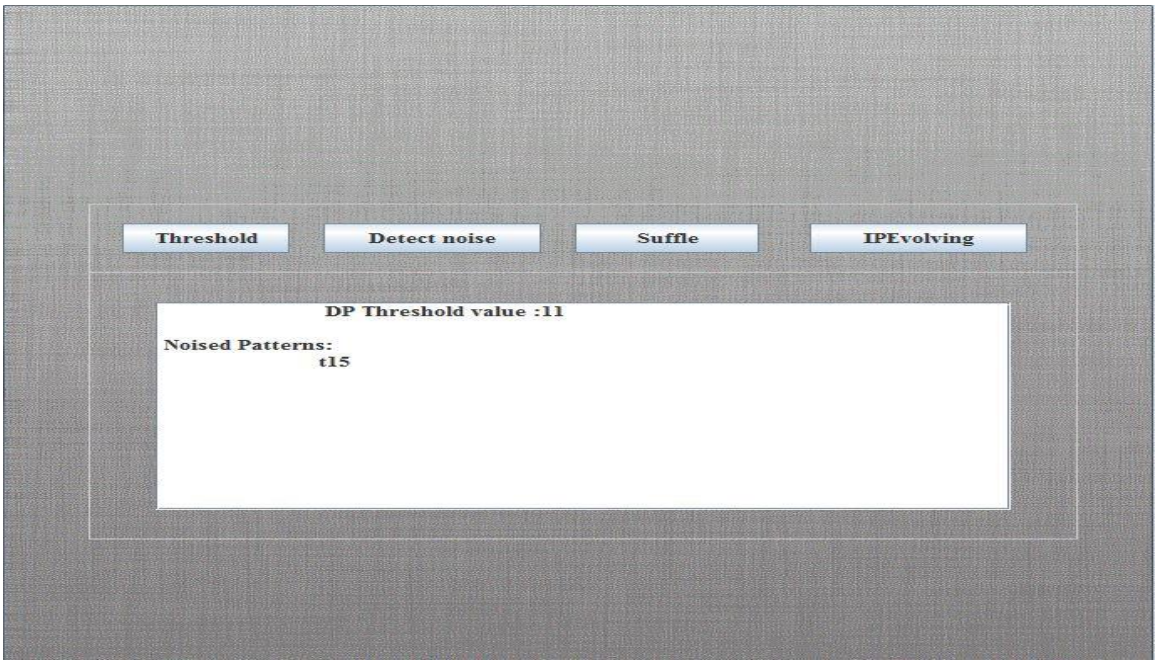


Fig.27

In fig.26 after clicking the IPE button we move to next UI screen which consist of Threshold button, Detect noise button, Shuffle button and IPEvolving button .Clicking the threshold button would give the threshold value according to which term are accepted or rejected. In fig.27 detect noise will calculate the noisy term which was accidentally made positive while it was negative.

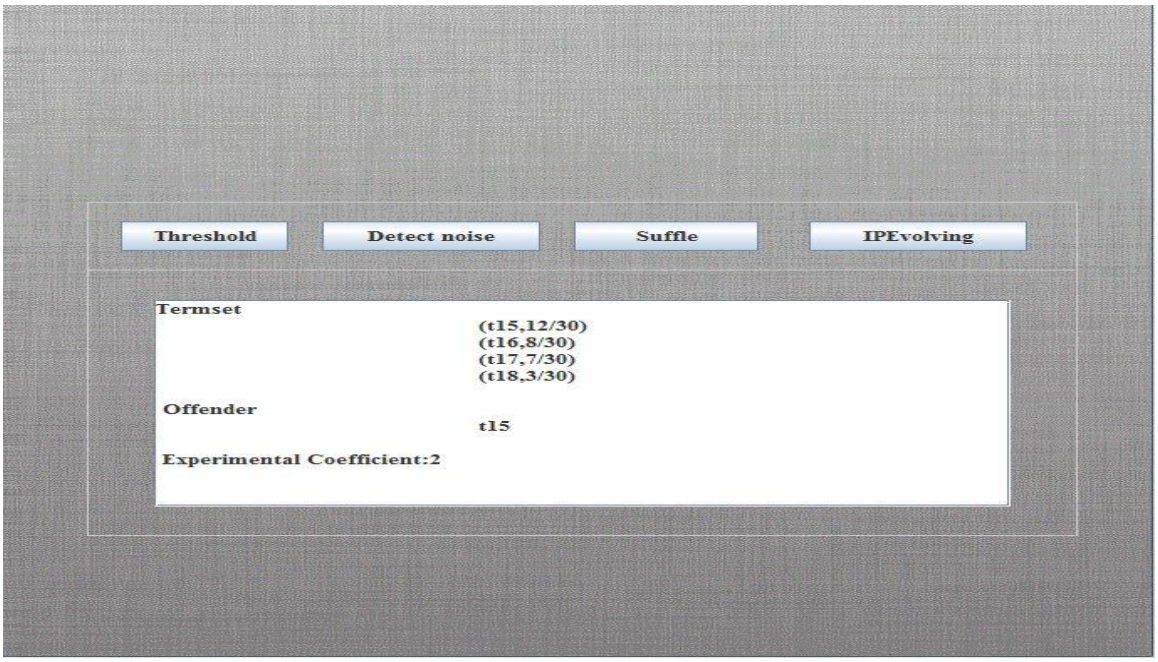

Threshold
Detect noise
Suffle
IPEvolving
Termset
(t15,12/30)
(t16,8/30)
(t17,7/30)
(t18,3/30)
Offender
t15
Experimental Coefficient:2

Fig.28

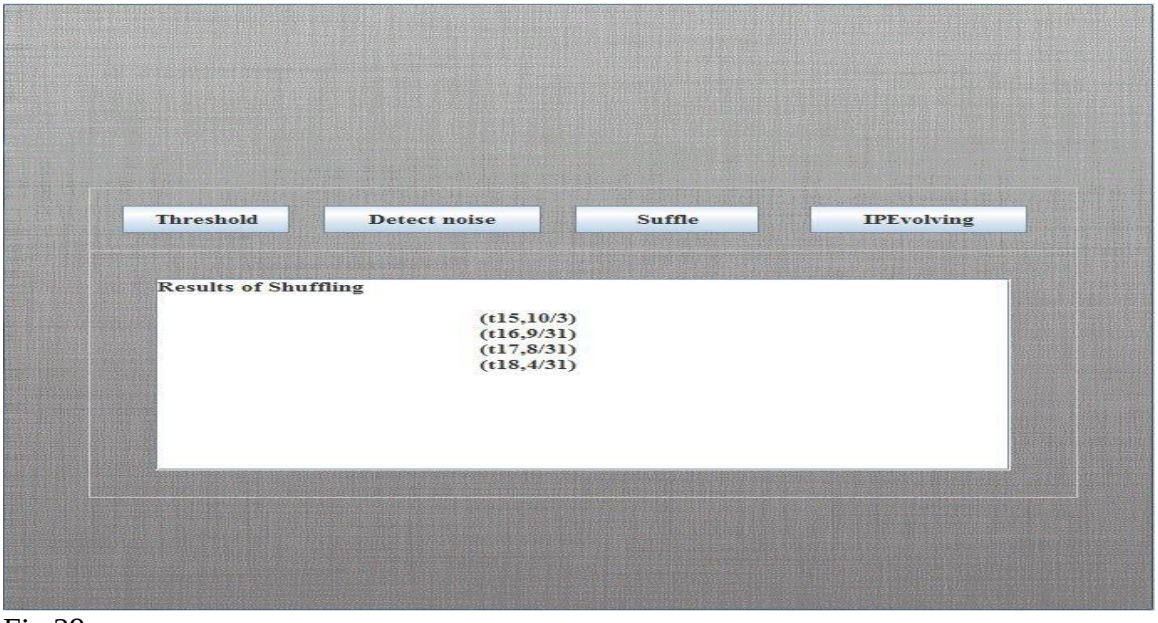

Threshold
Detect noise
Suffle
IPEvolving
Results of Shuffling
(t15,10/3)
(t16,9/31)
(t17,8/31)
(t18,4/31)

Fig.29

In fig.28 clicking shuffling button gave us termset, offender and experimental coefficient and clicking IPEvolving button will give the final termset.

CHAPTER 5 CONCLUSION AND FUTURE WORK

5.1 Conclusion

Many data mining techniques have been proposed in the last decade. These techniques include association rule mining, frequent item set mining, sequential pattern mining, maximum pattern mining, and closed pattern mining. The project is in partial phase where some modules are yet to get implemented. We began by using stopword elimination, Porter's Stemming algorithms for obtaining a filtered search space. We used MySQL for database. The fraudulent activity in a document is easy to find by manual search but difficult when required to do for thousands of documents. Though sometimes some text are straightforward but we require the most appropriate to come out.

However, using these discovered knowledge (or patterns) in the field of text mining is difficult and ineffective. The reason is that some useful long patterns with high specificity lack in support (i.e., the low-frequency problem). We argue that not all frequent short patterns are useful. Hence, misinterpretations of patterns derived from data mining techniques lead to the ineffective performance.

In this project, an effective pattern discovery technique has been proposed to shorten the time requires to analyze a fraudulent activity form a huge dataset . The proposed technique uses two processes, pattern deploying and pattern evolving, to refine the discovered patterns in text documents. The experimental results show that the proposed model outperforms not only other pure data mining-based methods and the concept based model.

5.2 Future Work

For future work the following can be taken up:

The project initially identifying the pattern for fraudulent activity but this can be used to analyze the market basket. In the following phase this can be used to compare with other models and showcase the efficiency of the PTM and D-pattern with respect to term based modelling and concept based modelling where concept based consist of CBM,CBM pattern matching and on the other hand term based consist of SVM,BM25,Rocchio,Prob,nGram,TFIDF.

The Rstudio is very good platform for data mining and with the help of Rshiny the application can be deployed on cloud. The R language can be used to implement all the modules (command line) and can be visualized easily.

The discovered patterns are summarized. The d-pattern algorithm is used to discover all patterns in positive documents are composed. The term supports are calculated by all terms in d-pattern. Term support means weight of the term is evaluated. In this module used to identify the noisy patterns in documents. Sometimes, system falsely identified negative document as a positive. So, noise is occurred in positive document. The noised pattern named as offender. This is the work done but carrying forward the efficiency can be further increased in terms of time to calculate the weight, offender and shuffling.

5.3 Limitations of the solution

Following are some of the limitations of the project

In the data retrieval module the dataset is divided into various parts to make the things understandable in the preprocessing phase .No large dataset is used though it can be used but situation will become complex.

No initial weights are allocated to be compared with the calculated weights.

The project is based on analysis of fraudulent activity it does not compare the pattern taxonomy model with the other models though it performs well then any other models whether term based or concept based models.

The application is static and cannot be run on web until php connectivity is done and interpreter is used to convert the java function into php functions.

References

[1] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. Addison Wesley, 1999.

[2] Y. Li, C. Zhang, and J.R. Swan, "An Information Filtering Model on the Web and Its Application in Jobagent," Knowledge-Based Systems, vol. 13, no. 5, pp. 285-296, 2000.

[3] S. Robertson and I. Soboroff, "The Trec 2002 Filtering Track Report," TREC, 2002, trec.nist.gov/pubs/trec11/papers/OVER.FILTERING.ps.gz.

[4] D.D. Lewis, "Feature Selection and Feature Extraction for Text Categorization," Proc. Workshop Speech and Natural Language, pp. 212-217, 1992.

[5] S. Scott and S. Matwin, "Feature Engineering for Text Classification," Proc. 16th Int'l Conf. Machine Learning (ICML '99), pp. 379-388, 1999.

[6] F. Sebastiani, "Machine Learning in Automated Text Categorization," ACM Computing Surveys, vol. 34, no. 1, pp. 1-47, 2002.

[7] N. Jindal and B. Liu, "Identifying Comparative Sentences in Text Documents," Proc. 29th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '06), pp. 244-251, 2006.

[8] S.-T. Wu, Y. Li, and Y. Xu, "Deploying Approaches for Pattern Refinement in Text Mining," Proc. IEEE Sixth Int'l Conf. Data Mining (ICDM '06), pp. 1157-1161, 2006.

[9] S.-T. Wu, Y. Li, Y. Xu, B. Pham, and P. Chen, "Automatic Pattern- Taxonomy Extraction for Web Mining," Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence (WI '04), pp. 242-248, 2004.

[10] N. Jovanovic, V. Milutinovic, and Z. Obradovic, Member IEEE, “Foundations of Predictive Data Mining” (2002).

[11] Margaret H.Dunham, “Data Mining- Introductory and Advanced Topics” Pearson Education, 2003, pages 106-112.

[12] Michael W. Berry and Malu Castellanos, Editors “Survey of Text Mining:Clustering, Classification, and Retrieval, Second Edition” Springe,September 30, 2007.

[13] Ying Zhao and George Karypis. Criterion Functions for Document Clustering: Experiments and Analysis. TR# 01-40, Department of Computer Science &Engineering, University of Minnesota, Minneapolis, 2000.

[14] Collen Mcque, Data mining and predictive analytics in Public safety and security,2006

[15]Zhong,Ning,Yuefeng Li, and Sheng-Tang Wu.”Effective Pattern Discovery for Text Mining.”IEEE Transactions on Knowlwdge and Data Engineering 24.1(2012):30-44.

APPENDIX A

Description of Tools

NetBeans

NetBeans refers to both a platform framework for Java desktop applications, and an integrated development environment (IDE) for developing with Java, JavaScript, PHP, Python (no longer supported after NetBeans 7), Groovy, C, C++, Scala , Clojure, and others. The NetBeans IDE 7.0 no longer supports Ruby and Ruby on Rails, but a third party has begun work on a separate plug-in. The NetBeans IDE is written in Java and can run on Windows, Mac OS, Linux, Solaris and other platforms supporting a compatible JVM. A pre-existing JVM or a JDK is not required. We are working with NetBeans 7.4.

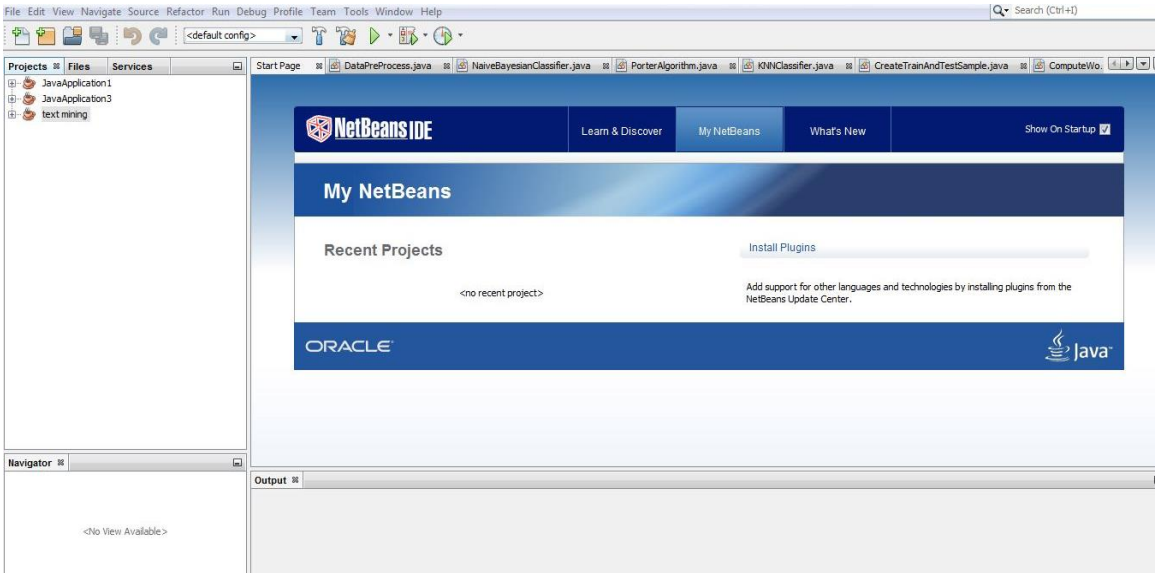


Fig:30

The NetBeans platform allows applications to be developed from a set of modular software components called modules. Applications based on the NetBeans platform (including the NetBeans IDE) can be extended by third party developers.

NetBeans IDE NetBeans IDE is an open-source integrated development environment. NetBeans IDE supports development of all Java application types (Java SE (includingJavaFX), Java ME, web, EJB and mobile applications) out of the box. Among other features are an Antbased project system, Maven support, refactorings,version control (supporting CVS, Subversion, Mercurial and Clear case).Modularity: All the functions of the IDE are provided by modules. Each module provides a well-defined function, such as support for the Java language, editing, or support for the CVS versioning system, and SVN. NetBeans contains all the modules needed for Java development in a single download, allowing the user to start working immediately.

Modules also allow NetBeans to be extended. New features, such as support for other programming languages, can be added by installing additional modules. For instance, Sun Studio, Sun Java Studio Enterprise, and Sun Java Studio Creator from Sun Microsystems are all based on the NetBeans IDE.

Rstudio

RStudio is an integrated development environment (IDE) for R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management.

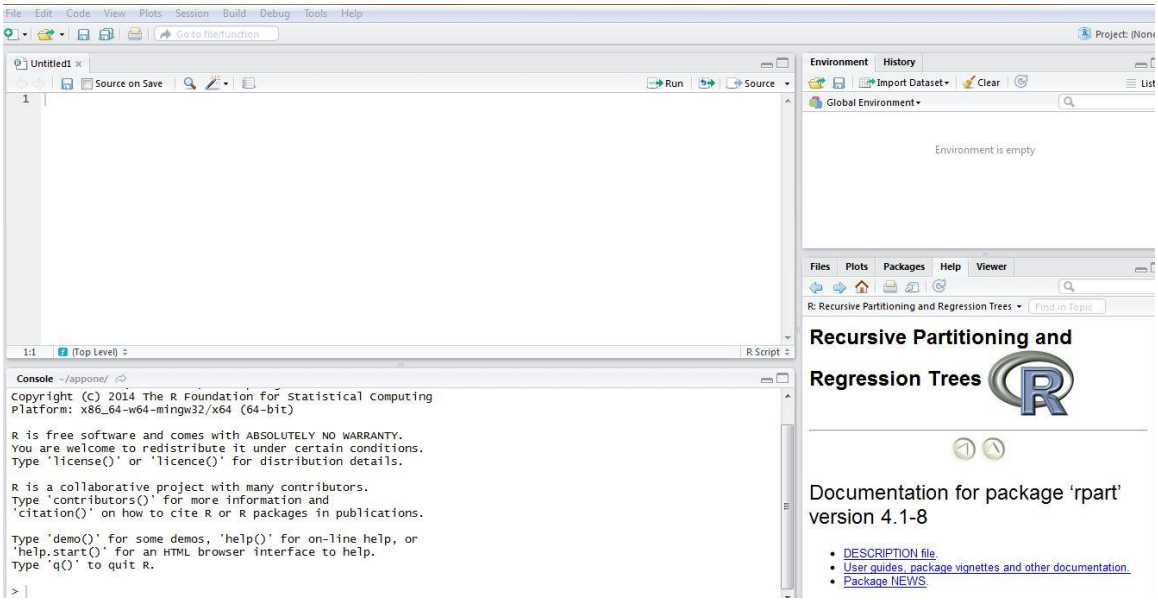


Fig:31

The frame in the upper right contains your workspace as well of a history of the commands that you've previously entered. Any plots that you generate will show up in the region in the lower right corner. The frame on the left is where the action happens. It's called the console. Every time you launch RStudio, it will have the same text at the top of the console telling you the version that you're running. Below that information is the prompt. As its name suggests, this prompt is really a request, a request for a command. Initially, interacting with R is all about typing commands and interpreting the output. These commands and their syntax have evolved over decades (literally) and now provide what many users feel is a fairly natural way to access data and organize, describe and invoke statistical computations.

Shiny package gives power to Rstudio .Shiny is an open source R package that provides an elegant and powerful web framework for building web applications using R. Shiny helps you turn your analyses into interactive web applications without requiring HTML, CSS, or JavaScript knowledge. Shiny applications have two components: a user-interface definition and a server script.

MySQL

MySQL is (as of March 2014) the world's second most widely used open-source relational database management system (RDBMS). It is named after co-founder Michael Wideness's daughter, My. The SQL phrase stands for Structured Query Language.

The MySQL development project has made its source code available under the terms of the GNU General Public License, as well as under a variety of proprietary agreements. MySQL was owned and sponsored by a single for-profit firm, the Swedish company MySQL AB, now owned by Oracle Corporation.

MySQL is a popular choice of database for use in web applications, and is a central component of the widely used LAMP open source web application software stack (and other 'AMP' stacks). LAMP is an acronym for "Linux, Apache, MySQL, Perl/PHP/Python." Free-software-open source projects that require a full-featured database management system often use MySQL.

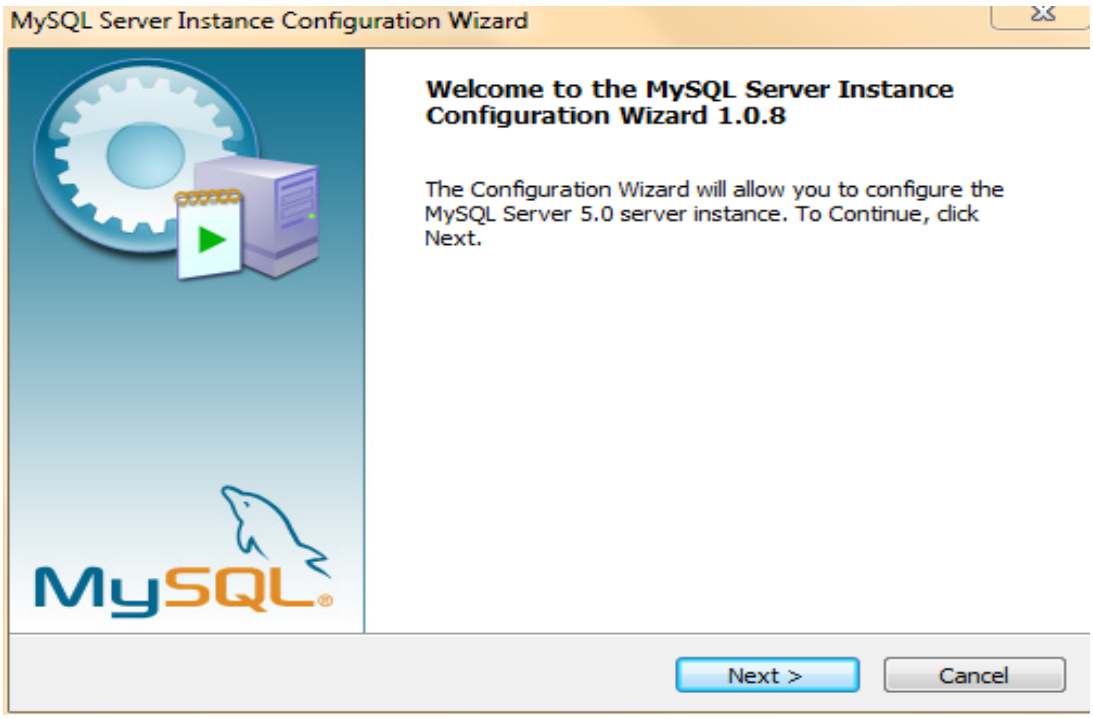


fig:32