"NLP To Create SQL Query"

Project Report

Submitted in partial fulfillment of the requirements for the degree of

Bachelor of Engineering

by

Shaikh Shagufta Shezad Zubeda (13CO69)

Momin Ummiya Salim Rehana (13CO68)

Shaikh Sharmeen Salim Shahina (12CO13)

Shaikh Rameeza Mohammad Yaqoob Shehnaz Begum (12CO07)

Supervisor

Prof. Tabrez Khan

Co-Supervisor

Prof. Rehaal Qureshi

Department of Computer Engineering,
School of Engineering and Technology
Anjuman-I-Islam's Kalsekar Technical Campus
Plot No. 2 3, Sector -16, Near Thana Naka, Khanda Gaon,
New Panvel, Navi Mumbai. 410206

Academic Year : 2015-2016

CERTIFICATE

Department of Computer Engineering,

School of Engineering and Technology,
Anjuman-I-Islam's Kalsekar Technical Campus
Khanda Gaon,New Panvel, Navi Mumbai. 410206

This is to certify that the project entitled *"NLP To Create SQL Query "* is a bonafide work of **Shaikh Shagufta Shezad Zubeda (13CO69), Momin Ummiya Salim Rehana (13CO68), Shaikh Sharmeen Salim Shahina (12CO13), Shaikh Rameeza Mohammad Yaqoob Shehnaz (12CO07).** submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of **"Bachelor of Engineering"** in **Department of Computer Engineering**.

Prof. Tabrez Khan **Prof.Rehaal Qureshi**

Supervisor Co-Supervisor

Prof. Tabrez Khan **Dr. Abdul Razak Honnutagi**

Head of Department Director

Project Approval for Bachelor of Engineering

This project entitled ***NLP to Create SQL Query*** by ***Shaikh Shagufta Shezad Zubeda(13CO69) ,Momin Ummiya Salim Rehana(13CO68) ,Shaikh Sharmeen Salim Shahina(12CO13) ,Shaikh Rameeza Mohammad Yaqoob(12CO07)*** is approved for the degree of ***Bachelor of Engineering in Department of Computer Engineering.***

Examiners

1. ....
2. ....

Supervisors

1. ....
2. ....

Chairman

....

Declaration

I declare that this written submission represents my ideas in my own words and where others ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Shaikh Shagufta Shezad Zubeda(13CO69)

Momin Ummiya Salim Rehana(13CO68)

Shaikh Sharmeen Salim Shahina(12CO13)

Shaikh Rameeza Mohammad Yaqoob(12CO07).

Abstract

Tittle: NLP To Create SQL Query

NLP to create sql query , this intelligent system converts the human language into the structure query language.Every year,every minute thousands of data is generated and managed.To use this data or retreiving information database intercation is important and for this purpose database experties are required.The problem is that it restricts the interaction between the naive user and the database. Only few people who have knowledge of formal database language can retrieve the desired information from the database. To overcome such a problem this proposed system will help a normal educated person having no knowledge of query language to easily interact with the database.

Shaikh Shagufta Shezad Zubeda (13CO69)

Momin Ummiya Salim Rehana (13CO68))

Shaikh Sharmeen Salim Shahina (12CO13)

Shaikh Rameeza Mohammad Yaqoob Shehnaz(12CO07)

B.E. (Computer Engineering)
University of Mumbai.

Contents

List of Figures

List of Tables

Keywords And Glossary

Keywords :

NLP , IQR , Morphoms , SQL , DBMS , NLIDB , DDL , DML .

Glossary :

A

- **Attribute :**
  Column of the table which describes property of the table.

- **Aggregations :**
  Data aggregation is a type of data and information mining process where data is searched, gathered and presented in a report-base.

- **Ambiguity:**Ambiguity is a word, phrase, or statement which contains more than one meaning.

C

- **Comparison Predicate :**A comparison predicate uses arithmetic operators to compare column data to a literal value.

- **Conjunctions:**A word used to connect clauses or sentences or to coordinate words in the same clause.

D

- **DDL:**Data Definition Language used to create the structure of the databaes.

- **DML:** Data Manipulation Language,used to insert,modifyand delete the data from the database.

- **DBMS:**Database Management System,is a computer software application that interacts with the user, other applications, and the database itself to capture and analyze and is used to store data and information.

- **DOB:**Date of Birth.

F

- **Free Grammer:**Set of rules and grammer used analysing of programming languages,parsing and manage the stucture of the documnet.

G

- **GUI:** Graphical User Interface,is a type of interface that allows users to interact with electronic devices through graphical icons and visual indicators such as secondary notation, as opposed to text-based interfaces, typed command labels or text navigation.

- **Geoquery:**Data for parsing English queries about a simple U.S. geography database.

H

- **History log:** A snapshot of file history is just a click away with the File History Log.

I

- **IQR:**Intermediate Query Representation,System's internal representation of natural langusge.

M

- **Metaphor:** A figure of speech in which a word or phrase is applied to an object or action to which it is not literally applicable.

- **Metonymy:** The substitution of the name of an attribute or adjunct for that of the thing meant.

- **Morphoms:**It is a basic unit of meaning.

- **MySQL:**SQL language.

N

- **NLIDB:**Natural Language Interface for Database,interface to interact with the database system.

- **NLTK:** Natural Language Toolkit,python library used to process the natural language.

- **NLP:**Natural Language Processing,field of Artificial Intelligence used to process human natural language.

p

- **Presuppositions:** Presuppositions in general are beliefs underlying a system. The presuppositions of NLP are beliefs that guide and have guided the development of NLP.

- **Php:**Hypertext Preprocessor ,is a server-side scripting language designed for web development but also used as a general-purpose programming language.

- **Pattern:** Web mining module for Python, with tools for scraping, natural language processing, machine learning, network analysis and visualization.

- **Phonology :**The system of contrastive relationships among the speech sounds that constitute the fundamental components of a language.

- **Parser:**A parser is a program, usually part of a compiler.

Q

- **Quantifications:** Use of an indicator of quantity in linguistic term.

S

- **Speech tagging:** Process of marking up a word in a text (corpus) as corresponding to a particular part of speech.

- **SQL:** Structured Query Language is a special-purpose programming language designed for managing data held in a database management system .

- **Shallow:** Things that aren't very deep,thus Deep linguistic processing is a natural language processing framework.

- **Software as Service (SaaS):** SaaS is a software distribution model in which applications are hosted by a vendor or service provider and made available to customers over a network.

T

- **Token:** Each seperate word of the senetence.

- **TK:** Tk/Tcl has long been an integral part of Python. It provides a robust and platform independent windowing toolkit, that is available to Python programmers using the tkinter package.

Chapter 1

Introduction

1.1 Statement of Project

1.1.1 Need Of SQL Query Generation

In the present computing world, computer based information technologies have been extensively used to help many organizations, private companies, academic and education institutions to manage their processes and information systems.
Information systems are used to manage data. The information management system that is capable of managing several kinds of data, stored in the database systems is known as Database Management System (DBMS).
Databases are comprehensive element in private and public information systems which are essential in number of application areas. Databases are built with the objective of facilitating the activities of data management in information systems. Due to the progress and in deep applications of computer querying system. It is due to the fact that the technology in several areas to be accurate, databases have become the repositories of huge volumes of data .In relational databases, to retrieve information from a database, one needs to formulate a query in such way that the computer will understand and produce the desired output.

1.1.2 Problem And Solution

The limitation of such programs is that it restricts the interaction between user and database to predefined set of queries. Only few people who have knowledge of database structure and formal database language such as Structured Query Language (SQL) can retrieve the desired information from database. A novice user having no knowledge of database structure and format database query language cannot retrieve desired information if it is not supported by well thought application. Hence, it is a need to improve human computer interface that allows people to interact with the database in their natural language (such as English).

1.1.3 Project Architecture

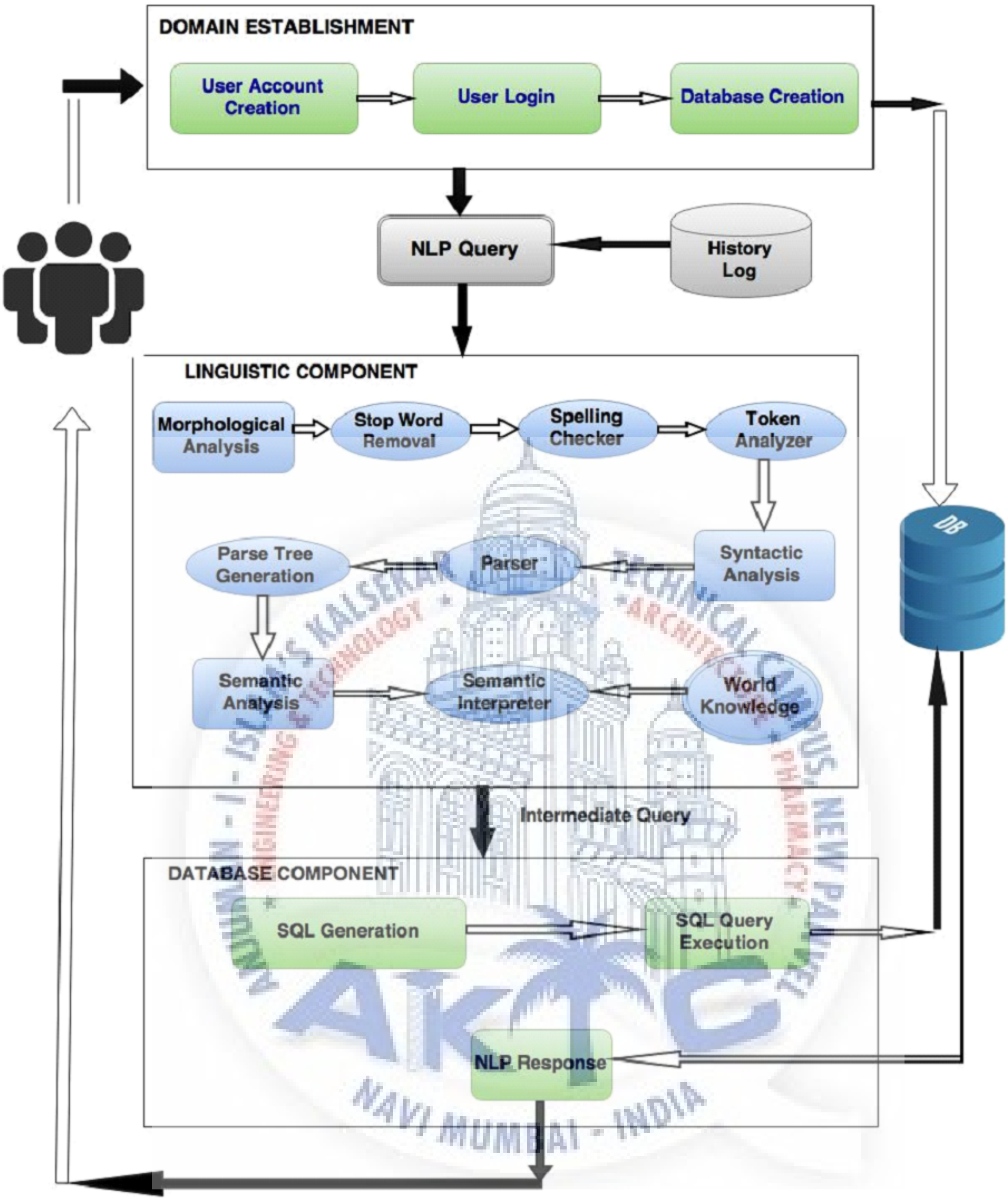


Figure 1.1: System Architecture

The System Architecture consists of three modules:

- Domain Establishment
- Linguistic Components
- Database Components

Domain Establishment This module is responsible for creating user accounts and database creation as the proposed system is domain independent and would be used by multiple users.

Linguistic Components This module is responsible for translating natural language input into a logical query. In this, the sentence is syntactically and semantically analyzed and processed, and an intermediate query is generated by the following steps.

1. Morphological Analysis.

2. Syntactic Analysis.

3. Semantic Analysis.

Morphological Analysis:- Morphology in linguistics is the study and description of how words are formed in natural language. In this phase the sentence is broken down into tokens- smallest unit of words, and determine the basic structure of the word.
For instance, unusually can be thought of as composed of a prefix un-, astem usual, and an affix -ly. composed is compose plus the inflectional affix -ed: a spelling rule means we end up with composed rather than composed.

- *Stop word removal:*
  Stop words are non context bearing words, also known as noisy words which are to be excluded from the input sentence to speed up the process.

- *Spelling check:*
  Three most popular method -

Table 1.1: Spelling Check

Correct Token	Error Token	Spelling Checking Operation
Student	Dtudent	Substitution
Student	Studnt	Insertion
Student	Studeent	Deletion

- *Token analyzer:*

  Each identified tokens can be represented as attribute token, value token, core token, multi-token, continuous token, etc.

  - Attribute token- using metadata.

  - Core Token-first, all capital letters.

  - Numeric Token-digits , digits separated by decimal point.

- Sentence Ending Markers-(. ? !).
- Value Token-(M.C.A, âœmcaâ, â˜mcaâ™).
- Continuous Token â“ (â˜@â™, apostrophe (â˜)).
- Multi-token- emp_no or e-no.
- Abbreviated Token-CE for Computer Engineering.

Syntactic Analysis:-

The objective of the syntactic analysis is to find the syntactic structure of the sentence.It is also called Hierarchical analysis/Parsing, used to recognize a sentence, to allocate token groups into grammatical phrases and to assign a syntactic structure to it.

- Parse tree:-
  Parser generates a parse tree with the help of syntactic analysis. A parse tree or parsing tree is an ordered, rooted tree that represents the syntactic structure of a string according to some context free grammer.

  Example:
  list of student live in mumbai.

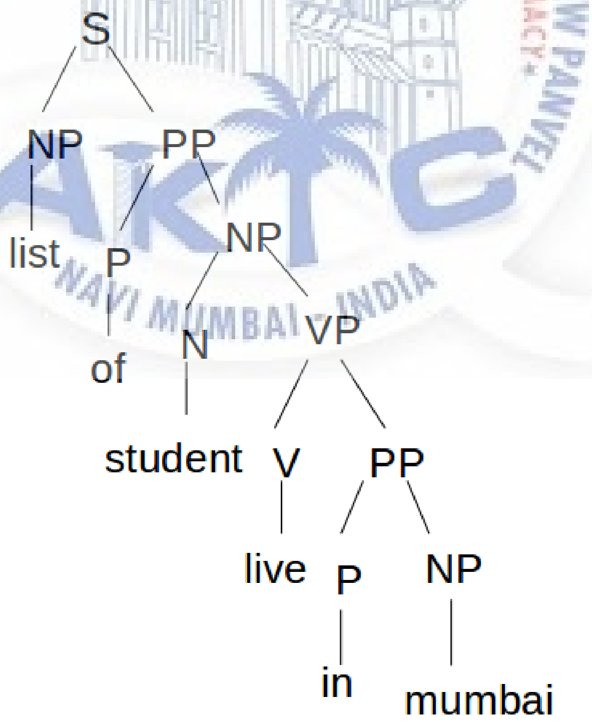


Figure 1.2: Parse Tree

- **Semantic analysis:** Semantic Analysis is related to create the representations for meaning of linguistics inputs.It deals with how to determine the meaning of the sentence from the meaning of its parts. So, it generates a logical query which is the input of Database Query Generator. It is another form of representation for user tokens and user input symbols in the form of semantic word.

- **Intermediate Query Represntation:** It is very difficult to directly map the syntax tree as well as semantic meaning of the sentence directly to the sql query, an intermediate query is generated from the semantic analysis then this logical query is going to be converted into sql query.

Database Components

SQL query generation:
It consists of SQL query generation where intermediate query is going to map to the sql query.
SQL query execution:
Generated sql query is going to be executed here and desired data is extracted from the database.
Response of NLP:
Extracted result is going to be displayed to the user.

1.1.4 Motivation

Nowadays the field of education is increasing greatly and so are the number of students. As such the students have many questions related to their education such as admission details,exam dates,best available colleges in the city or across the country and so on.With so many questions the avaiable facilities for answering all these questions of students are quite limited and less effeciency. The answer to these question of students are present in multiple databases in a distributed manner. However the students can't query the database to get the required information. At such a time having a system that can interpret the question posed by the student in their natural language and answer these queries of students in a quick and comprehensive manner will prove to be very beneficial for the students.

1.2 Objective and Scope

1.2.1 Objective

The main objective of the proposed system is to built such an intelligent system that takes the natural language input, understands and learns its meaning, convert it in some internal form and give result back to the user regardless of the structure and format of the sentences and words. Since normal educated user can't access the database, a person having complete knowledge of database management system, all the required syntaxes, commands must be known to the person. To overcome this problem there is a need to translate the normal user language such as English into SQL query in order to get the result from the database,here user need not to learn

anything related to database and can interact directly to the database in his/her known language. Every day every minute thousands of data is generated, there is the need to store this; data storage of data can been done using many database managment software but the question arises who is going to interact with this system the answer is none other than one who is having the knowledge of SQL syntax.
The main objective of the project is to create a system for user so that they can easily interact with the system not necessary the user should know specific database ,the user can easily interact in his/her language the most common language is none other than English language the user need not to know the complex command of database language. He/She will just need to write the normal English sentence the machine will try to understand and decide the meaning of sentence and accordingly it will convert the English sentence in SQL form for machine understanding and the output is again converted into the normal English language.

1.2.2 Scope

Proposed system is domain independent therefore can be used in any application. Where there is need to store the data and retrieving of data is done by non technical person or naive user, this system is useful.

- **School and college :**The system can be used for many purpose in School and College .there is a need of storage of student,teacher,cleark data .the detail about student fees, result, document etc this data is usefull to student, teacher as well as to the administration department . Thus this data need to be stored in data base and native user find difficult to interact with the SQl software thus the system can be used in School and Colleges.

- **Hospital :** In Hospital the patient, medicane are the important attribute there is the need that detail of this attribute need to be maintained so that one can get the information as required per their need this data are important for the user the day when the doctor are availabe the detail of the medicane present in the hosptial ,the day when the operation is to bee done are the important document for the user as well hosptial staff thus this data need to be stored in to the database thus interaction with this data form database need SQl command thus creating a system that can easily used to communicate with normal user in normal language that is English so that help to communicate with the system with ease.

- **Transport :**The system can be used in Railway ,Air reservation purpose the detail of the user who have booked thier ticket the day the trail and flight will leave the place and the day it will reach the destation . The number of the customer which are in the wating list so on this detail can be maintained using the system and normal user can easily check for the detail like wise the system can bee used for below purpose.

- Bank,Government Office, Service Sector, Navy, Manufacturing database, Sensus system,Agriculture,Chat system,Business organization,Chemical Industies.

Chapter 2

Literature Review

2.1 An Overview of NLIDB Approaches and Implementation for Airline Reservation System.

This system was developed for Flight Reservation. A combination of Syntax Analysis and Intermediate Query approach is used for this system. Syntax Analysis performs syntactic processing and breaks the input sentence into its constituent parts and identifies the relations between the concepts. Intermediate Query approach allows to easily perform the mapping of concepts to an intermediate representation. The intermediate representation can be used even in case of database portability i.e.even if database is ported to another database. The system was developed using Java, Spring framework, Hibernate ORM and MySQL database. The database has 2 tables for flights and bookings. Stanford Core NLP is used.The software is tested by running various natural language queries. The results are evaluated. The output is checked with the actual records in the database and verified for correctness.[1]

2.1.1 Pros

- The Aircraft time and date is shown and the system can easily check the present detail as per schedule. Thus the NLIDB system can be used for search the user query.

2.1.2 Cons

- In this system the user is forced only the specificc words which are related to datadases.
- The user is forced to enter the words in proper order only .
- It is standalone system.

- It is not implemented for complex queries like nested,join,group,order, queries,etc.
- History log is not maintain

2.2 Accessing Database Using NLP.

The goal of accessing database by natural language processor is to make dataset access easier for the common people. While natural language may be the easiest symbol system to learn and use, it has proved to be the hardest to a computer to master. To access database a user must have the knowledge of Structured Query Language(SQL). In India, there are many people who know English but are not fluent enough to formulate queries in it. With the help of this interface an end user can query the system in natural languages like English, Hindi and Marathi etc., and can see the result in the same language.
They developed an application that took the Database queries in the form natural language and then processes it and gives the result.This includes many sub components like Language Analyzer, Query Builder and Viewer.The system first parses the query in natural language and finds the major parts in the stringThen first it looks for the table name and then it parses the string for the where clause and then for the order by clause.After parsing constructs the query string based on the data available. The generated SQL query is posted to the database to fetch the results.[2]

2.2.1 Pros

- The user is able to interact with database in English,Hindi and in Marathi language. The GUI of the system is similar to forms.

2.2.2 Cons

- The interaction is in the not in the form of NL sentence.
- It is standalone system.
- History log is not maintain

2.3 SQL Generation And Execution From Natural Language Processing.

This paper work uses NLP for interfacing with the Database using natural language. In this work only English language is used as a mean for providing inputs. In this system a database ORACLE and a default table is used which is properly normalized. A system is developed that eliminates the problem of normal user to interact with database with rigid language SQL. The users are able to access informationâ^{TM}s by issuing query in simple English language. The system is developed in JAVA language.This syetem is based on employee information system.[4] The system architecture for the this system is as followes:

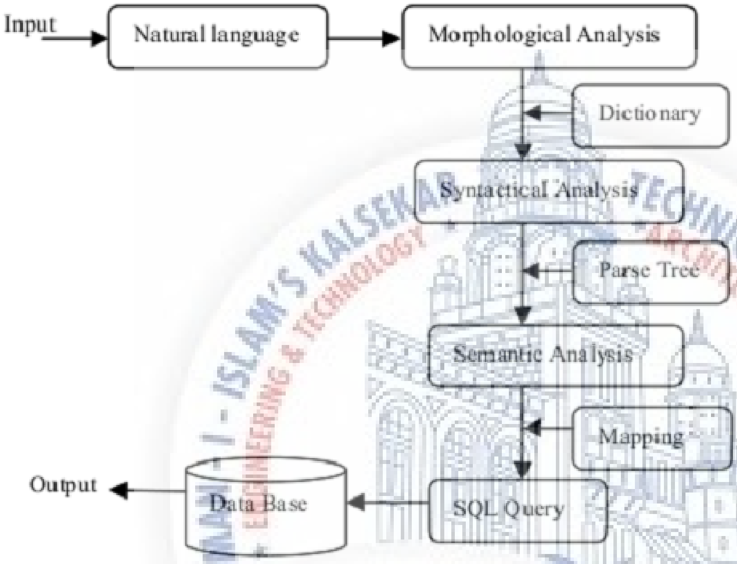


Figure 2.1: SQL Generation And Execution From Natural Language Processing.

2.3.1 Pros

- This system is able to excute DDL and DML statements .
- Ambiguity is also handled by the system

2.3.2 Cons

- The system is not able to handle the query in the form of question.

- Oracle database is used which is not open source.

2.4 Challenges and Implementation Steps of Natural Language Interface for Information Extraction from Database.

This paper converts the English query into sql query.This system accepts the natural query in the form of questions such as who, what, when etc.It is developed for Employee database.GATE standford parser are used.This system is developed to deal various natural language processing challenges such as reducing abmiguity, Metaphor and metonymy in question, Spelling mistakes in question, Handling vagueness in sentence ,Testing over all possible form of natural language ,Mapping the meaning of query ,User assumes intelligence and Insufficient information in query .[3]

System Architecture of this system as follows:

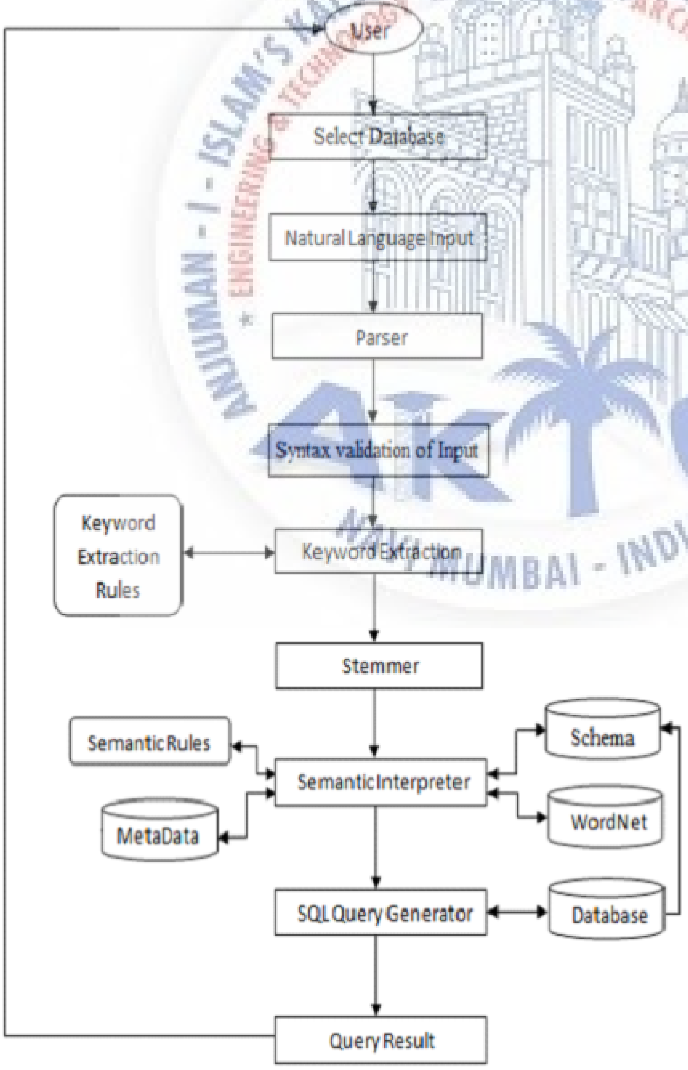


Figure 2.2: Challenges and Implementation Steps of Natural Language Interface for Information Extraction from Database

2.4.1 Pros

The GUI of the system is similar to forms.

2.4.2 Cons

- The system is not able to able to create complex query.
- The GUI is not user friendly.
- History log is not maintain
- It is standalone system.

2.5 How to Overcome

- Proposed system is developed for DDL as well as DML statement.
- It would have the ability to deal with complex queries.
- The system is able to maintain history log.
- The system is Web based application.
- It provide user friendly UI.
- The software is developed using open source technology.

Chapter 3

Requirement Analysis

3.1 Platform Requirement :

3.1.1 Supportive Operating Systems :

The supported Operating Systems for client include:

- Windows xp onwards
- Linux any flavour.

Windows and Linux are two of the operating systems that will support comparative website. Since Linux is an open source operating system, This system which is will use in this project is developed on the Linux platform but is made compatible with windows too.The comparative website will be tested on both Linux and windows.

The supported Operating Systems for server include: The supported Operating Systems For server include Linux. Linux is used as server operating system. For web server we are using apache 2.0

3.2 Software Requirement :

The Software Requirements in this project include:

- Python
- Flask

- Mysql

In this project, python is used for creating backbone structure. Python is intended to be a highly readable language. It is designed to have an uncluttered visual layout, it uses whitespace indentation, rather than curly braces or keywords. Python has a large standard library, commonly cited as one of Python's greatest strengths. NLTK is a suite of libraries and programs for symbolic and statistical natural language processing (NLP) for the python programming language. It is a leading platform for building Python programs. It provides easy-to-use interfaces of text processing libraries. In the project it is being used for tokenization, spelling check, stemming, tagging, parsing, and semantic reasoning. Flask is a micro web framework written in Python. It is used for frontend where the user is going to interact from.

Python is fundamental language being used in the development of the project. MySQL is used as a database. 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. MySQL is easy to use, yet extremely powerful, secure, and scalable. And because of its small size and speed, it is the ideal database solution for Web applications.

3.3 Hardware Requirement :

3.3.1 Hardware Required For Project Development:

- 1 GB Ram.

- 40 GB Hard Disk Minimum.

- Intel Core i3-3xxx

Chapter 4

Project Design

4.1 Design Approach

Design is the first step in the development phase for any techniques and principles for the purpose of defining a device, a process or system in sufficient detail to permit its physical realization. Once the software requirements have been analyzed and specified the software design involves three technical activities design, coding, implementation and testing that are required to build and verify the software. The design activities are of main importance in this phase, because in this activity, decisions ultimately affecting the success of the software implementation and its ease of maintenance are made. These decisions have the final bearing upon reliability and maintainability of the system. Design is the only way to accurately translate the customer requirements into finished software or a system. Design is the place where quality is fostered in development. Software design is a process through which requirements are translated into a representation of software. Software design is conducted in two steps. Preliminary design is concerned with the transformation of requirements into data.

4.2 Software Architectural Designs

Our system follows the three tier architecture . First tier consist of GUI, Linguistic component and the Database.

1. GUI: The GUI(Graphical User Interface) in our project deals with the interface for the user where the user enters the query(information to be rereived) in english language. The GUI provides a platform for the user to communicate with the database. It acts as a connector as well as communicator which connects the database and helps in transfer of data between the GUI and the database.

2. Linguistic component: The linguistic component is the block where the actual processing of our project is done.This module is responsible for translating natural language input into a logical query. In this, the sentence is syntactically and semantically analyzed and processed, and an intermediate query is generated.

3. Database: Database tier is the tier used for the storage of data. This tier contains all the data that is need for the processing of the whole project. The data in this tier is related to the student information.

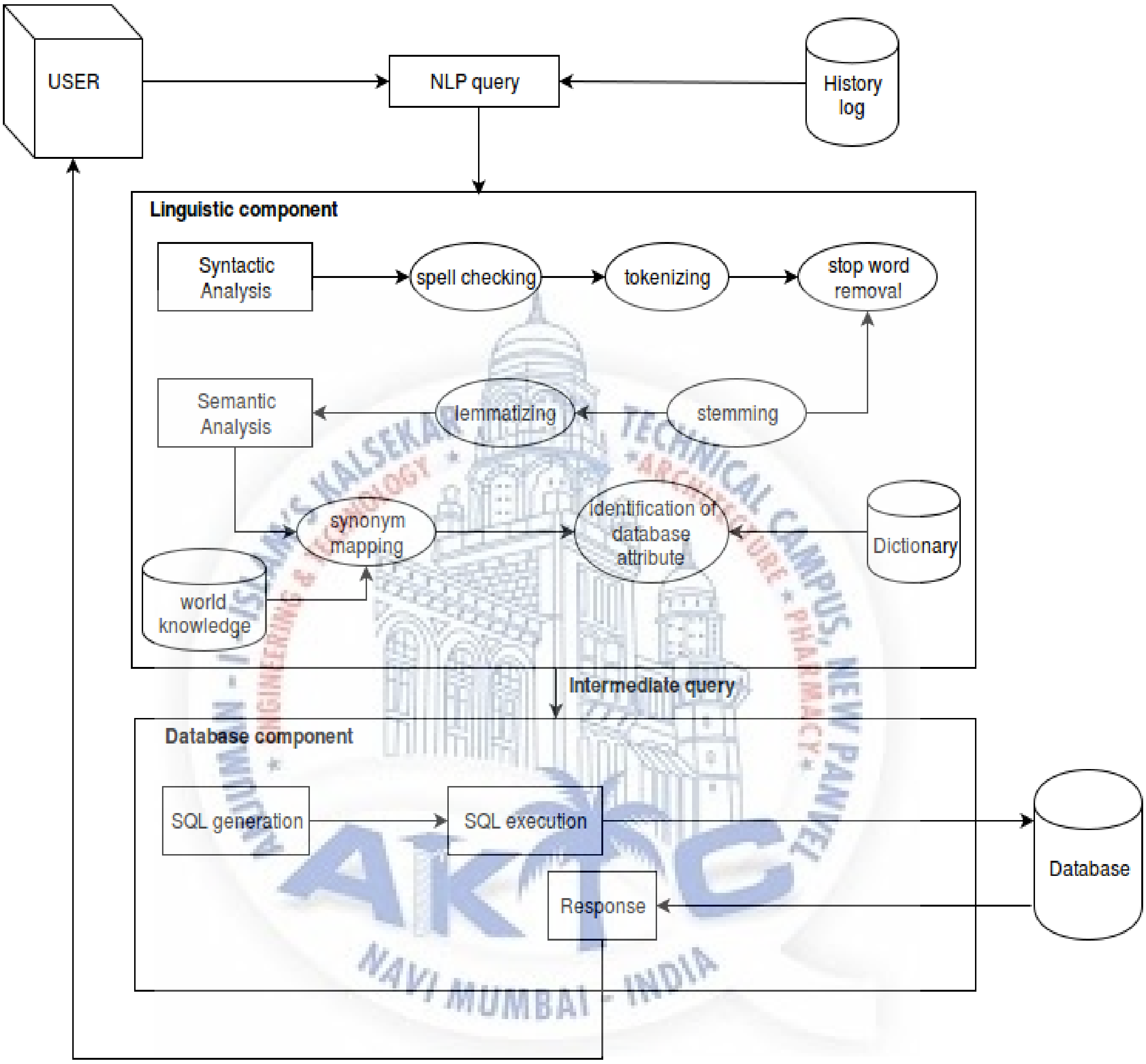


Figure 4.1: Software architecture Design

4.2.1 Front End Designs

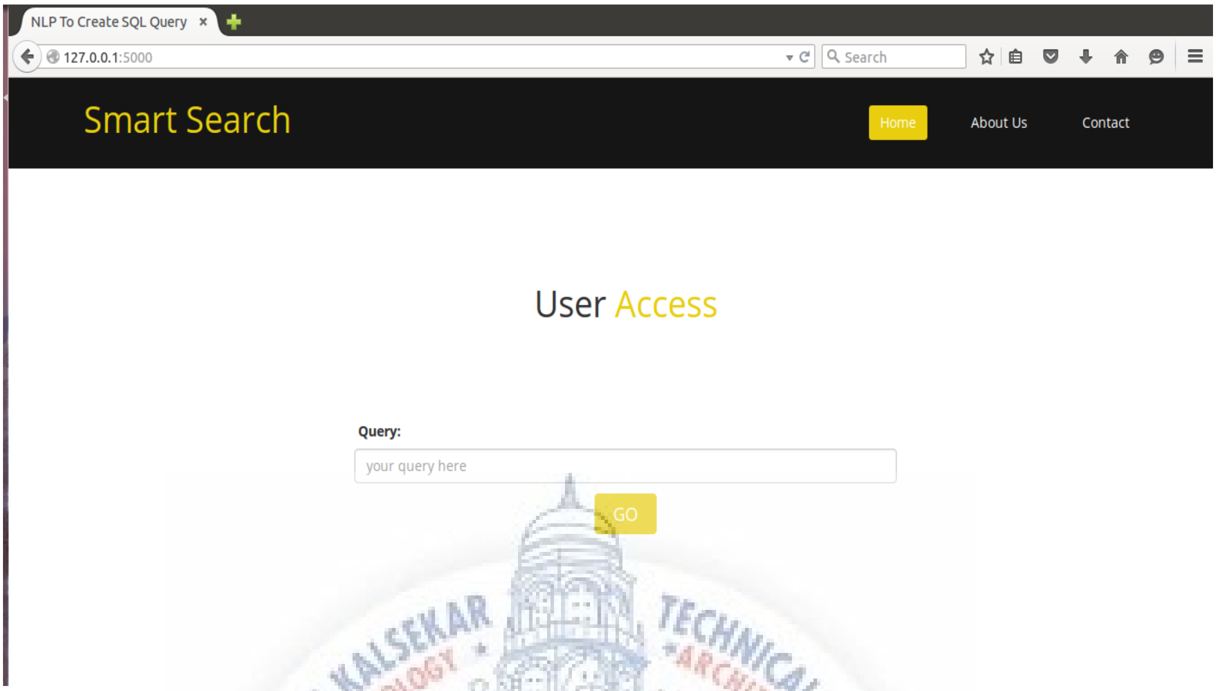

NLP To Create SQL Query
127.0.0.1:5000

Figure 4.2: Front End Design

4.2.2 Component Diagram

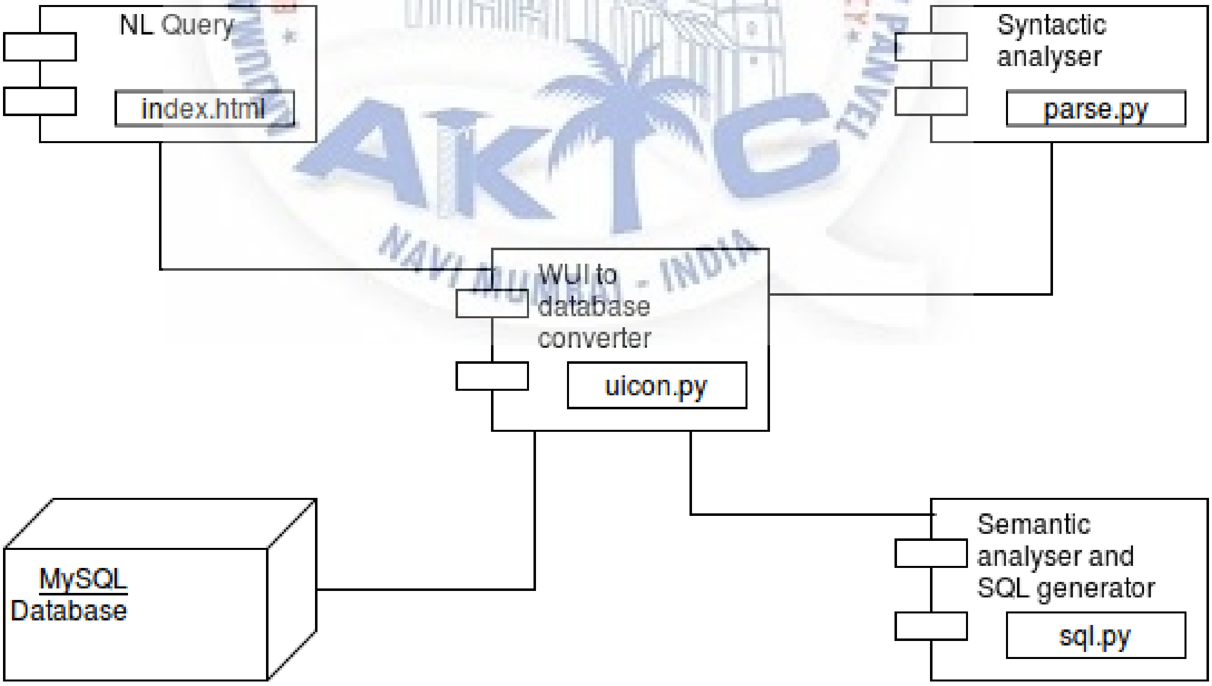


Figure 4.3: Component Diagram of NLP to sl query

4.2.3 Deployment Diagram

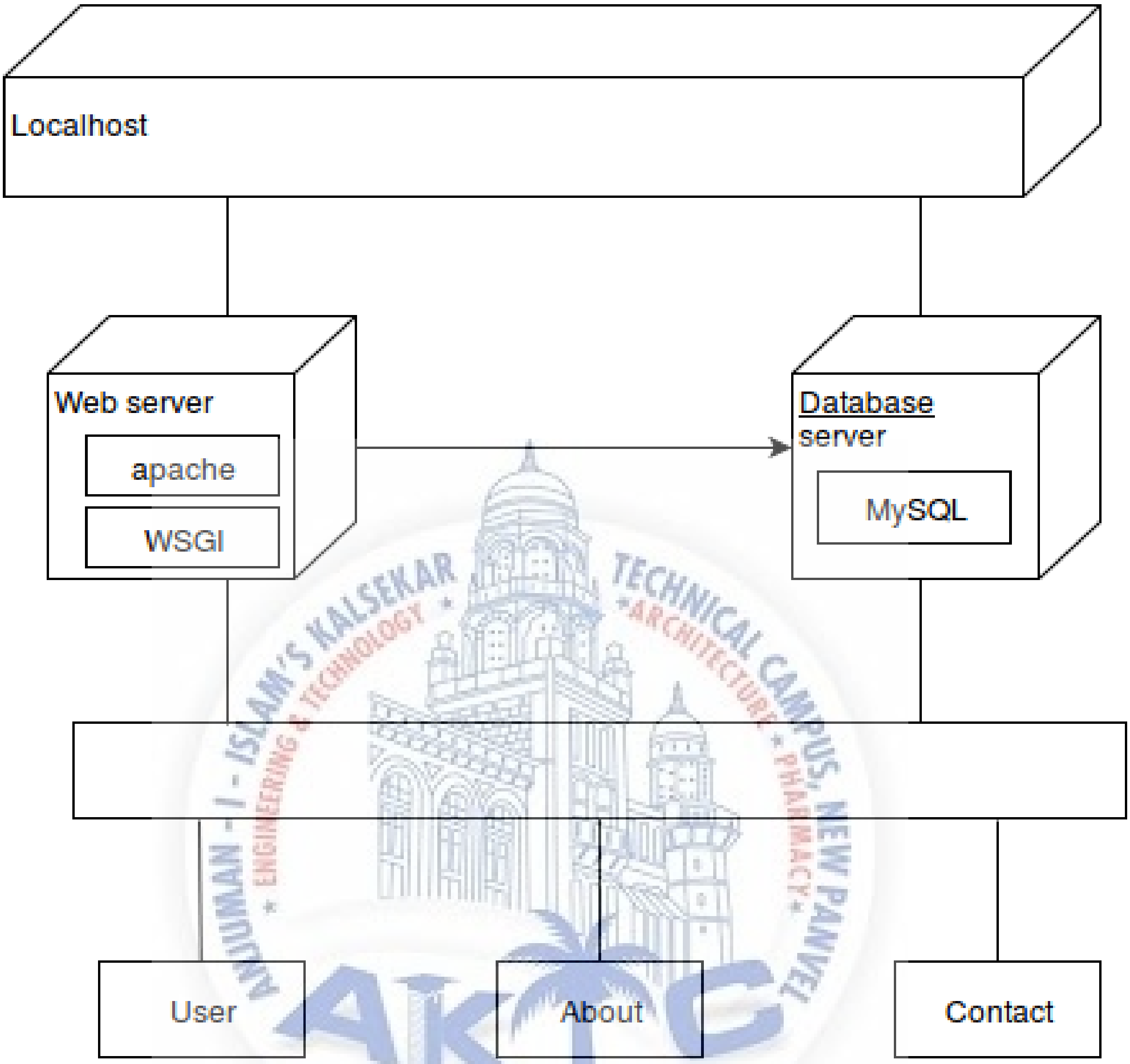


Figure 4.4: Deployment Diagram of NLP to sl query

4.3 Database Design

4.3.1 E-R Diagram

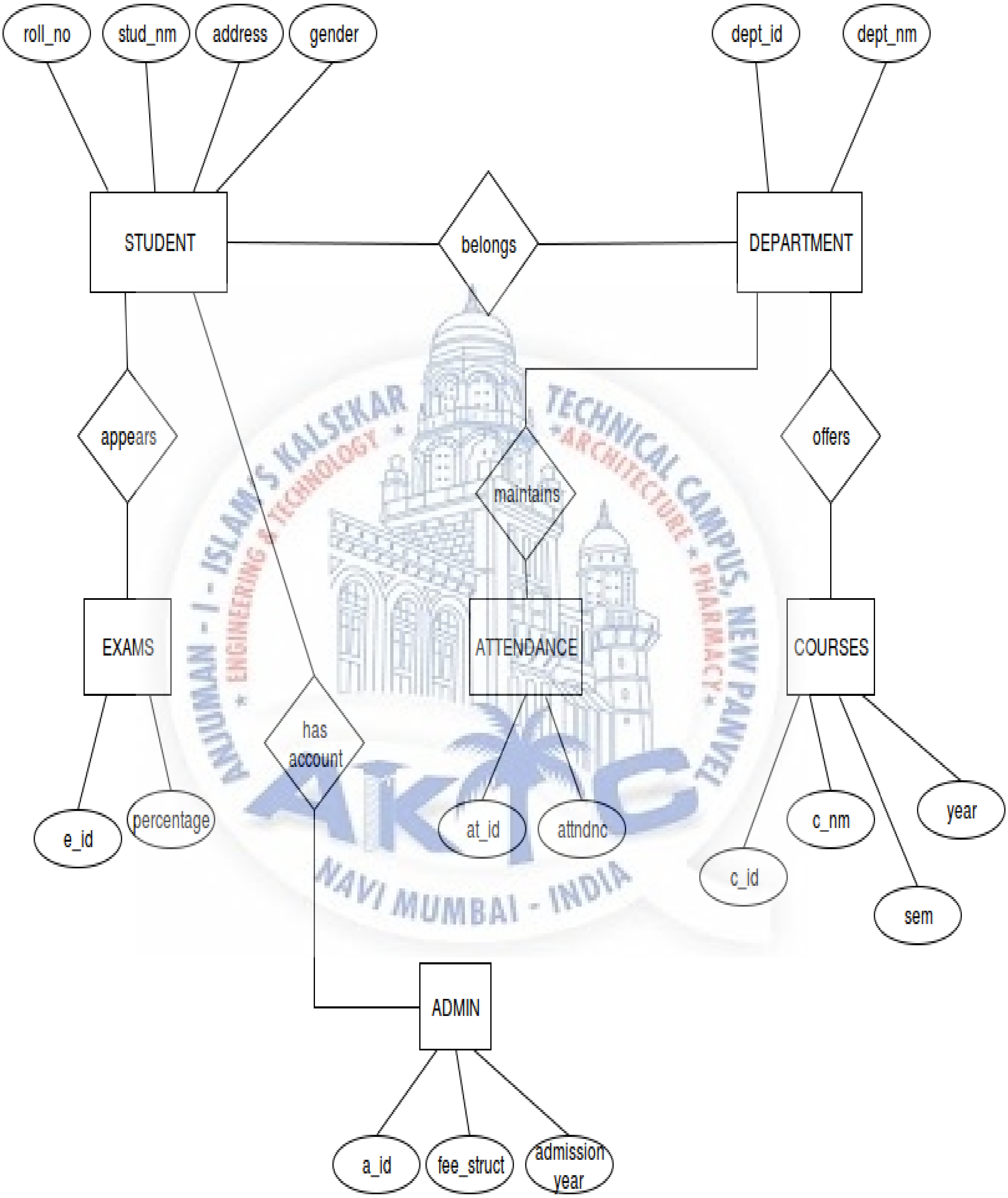


Figure 4.5: E-R Diagram of NLP to sl query

4.4 Work-flow Design

4.4.1 Flow Diagram

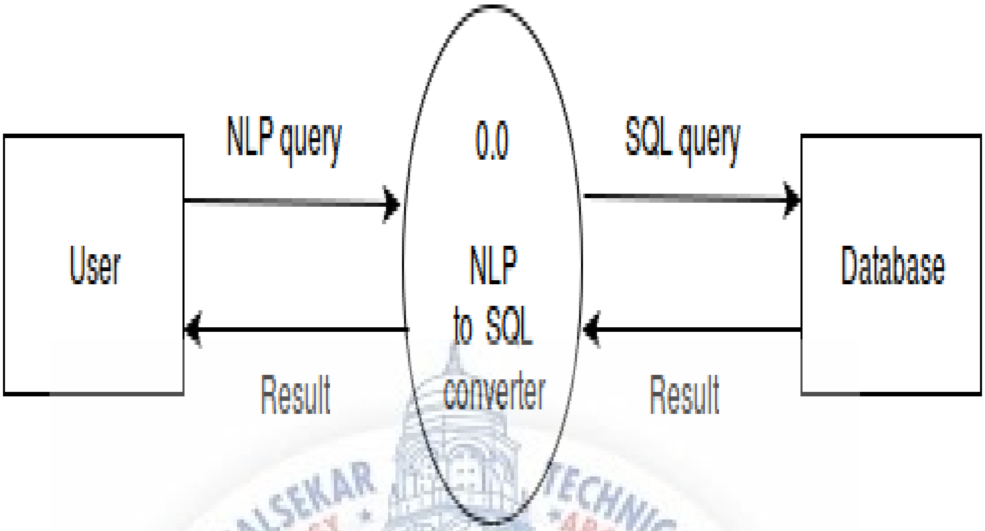


Figure 4.6: Level 0 DFD of NLP to create sql query

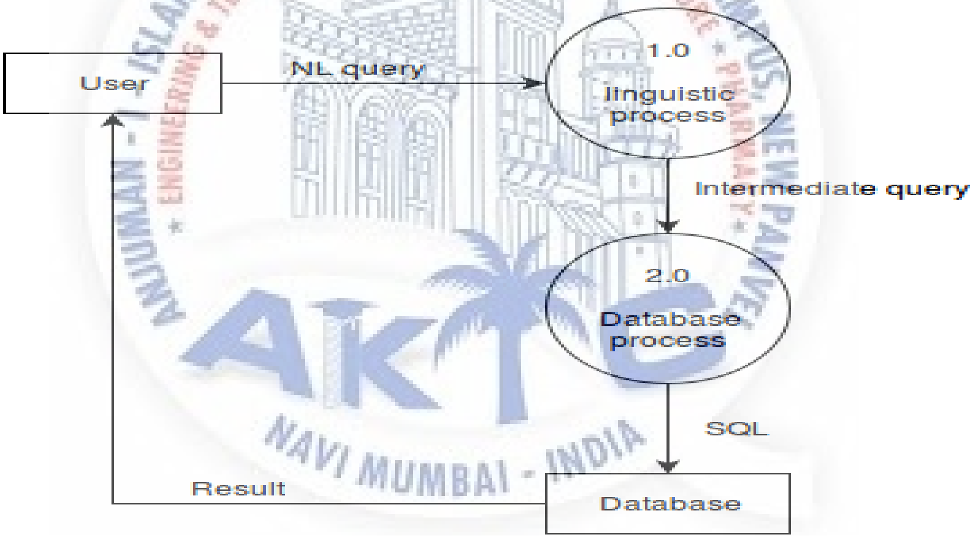


Figure 4.7: Level 1 DFD of NLP to create sql query

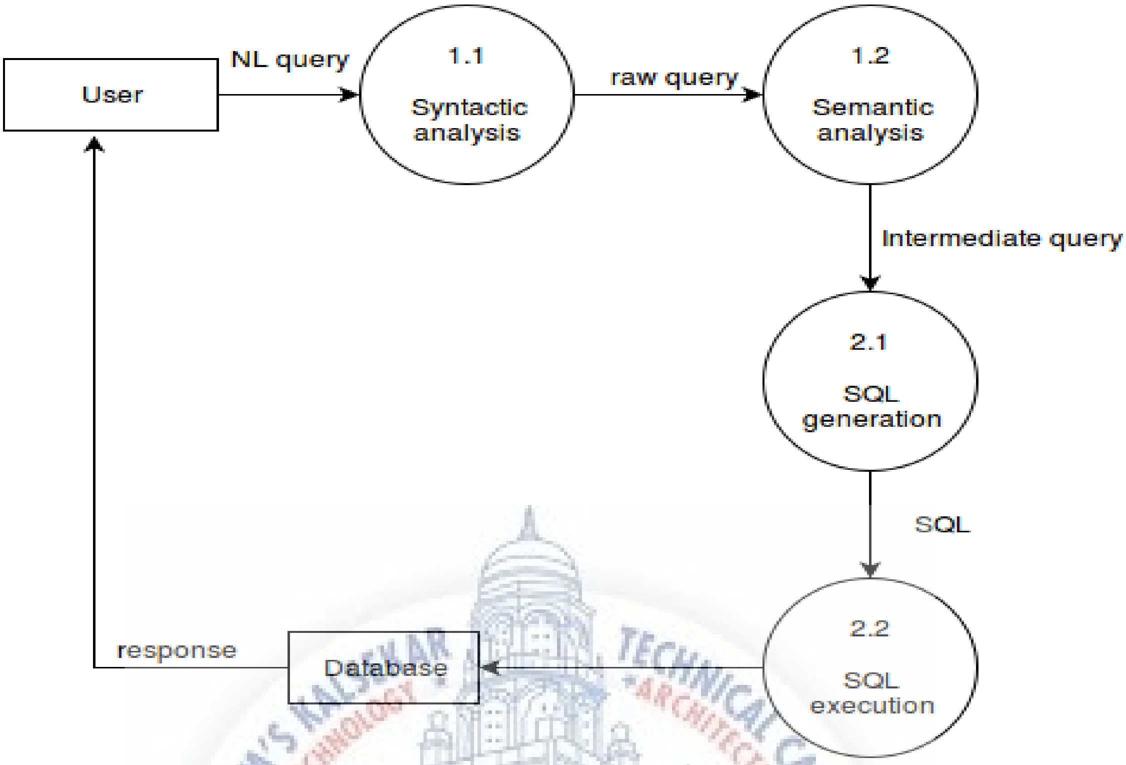


Figure 4.8: Level 2 DFD of NLP to create sql query

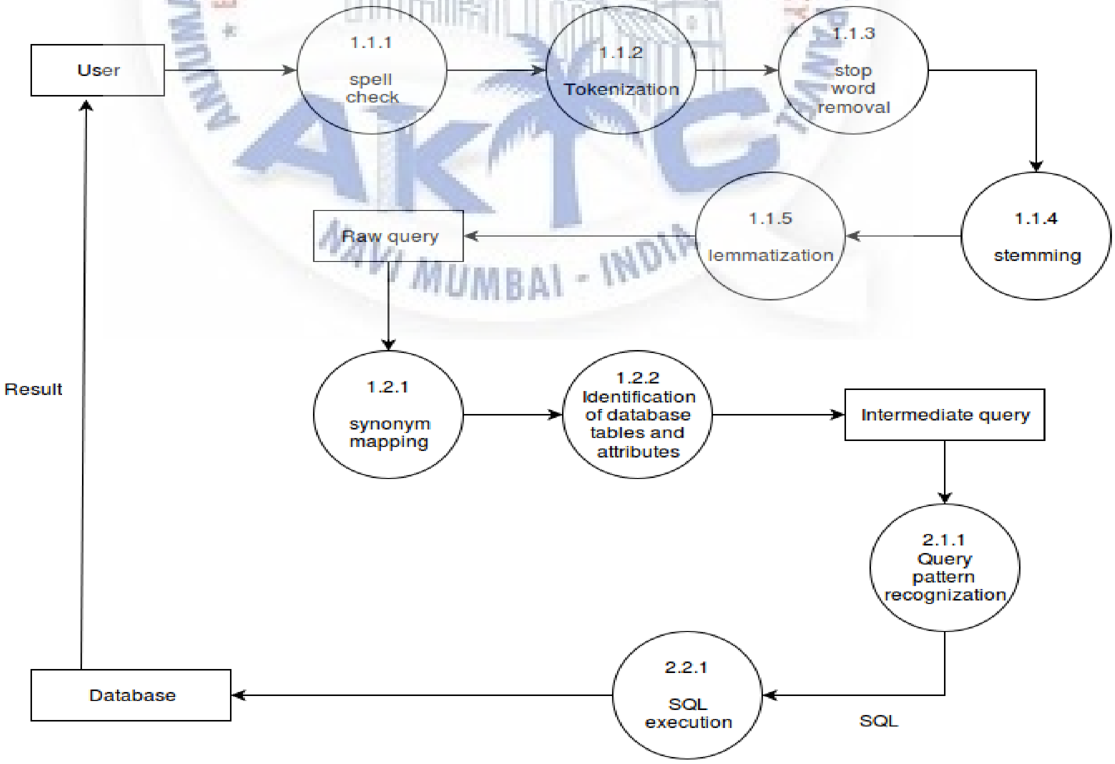


Figure 4.9: Level 3 DFD of NLP to create sql query

Chapter 5

Implementation Details

5.1 Assumptions And Dependencies

5.1.1 Assumptions

The following Assumption was taken into consideration:

- The script converts the natural language input from the user to the corresponding sql query. The script is to spell check the input, perform the stop word removal, lemmatization and tokenization. Then with the help of dictionaries it is assumed to correctly map the keys and values and convert it to the accurate sql query form. Therefore for proper query generation the correct patterns need to be entered while taking input. Once the query is formed the data is retreived from the database and user will get the output in tabular form.

- As soon as the user enters the search the script is assumed to convert it to sql query and fetch the required results.The user interface should be simple and clear that allows soothing effect to the user.

5.1.2 Dependencies

For backend processing we used MySQL database. Python script is generated which is doing the syntactic analysis. Different nltk packages are being used for stop word removal, lemmetization, stemming, tokenization. Textblob package for spelling check. Metadata for mapping process which is part of semantic analysis. From WUI the natural language query is passed to python script where the actual syntax of structured query is formed and result is fetched from database.

5.2 Implementation Methodologies

Different modules or components created are domain establishment,linguistic component and database component. Linguistic component is the heart of the system which is created using python nltk libraries. This component does the morphological analysis, syntactic analysis, semantic analysis and generates the intermediate query. The database component is then given the sql query to return the results.

5.2.1 Modular Description of Project

5.3 Detailed Analysis and Description of Project

Domain establishment: This module is responsible for creating user accounts and database creation as the proposed system is domain independent and would be used by multiple users.

Linguistic component: This module is responsible for translating natural language input into a logical query. In this, the sentence is syntactically and semantically analyzed and processed, and an intermediate query is generated by the following steps.

Morphological Analysis: Morphology in linguistics is the study and description of how words are formed in natural language. In this phase the sentence is broken down into tokens- smallest unit of words, and determine the basic structure of the word.

Syntactic Analysis: The objective of the syntactic analysis is to find the syntactic structure of the sentence.It is also called Hierarchical analysis/Parsing, used to recognize a sentence, to allocate token groups into grammatical phrases and to assign a syntactic structure to it.

Semantic analysis: Semantic Analysis is related to create the representations presentations for meaning of linguistics inputs. It deals with how to determine the meaning of the sentence from the meaning of its parts. So, it generates a logical query which is the input of Database Query Generator. It is another form of representation for user tokens and user input symbols in the form of semantic word.

5.4 Usecase Diagram

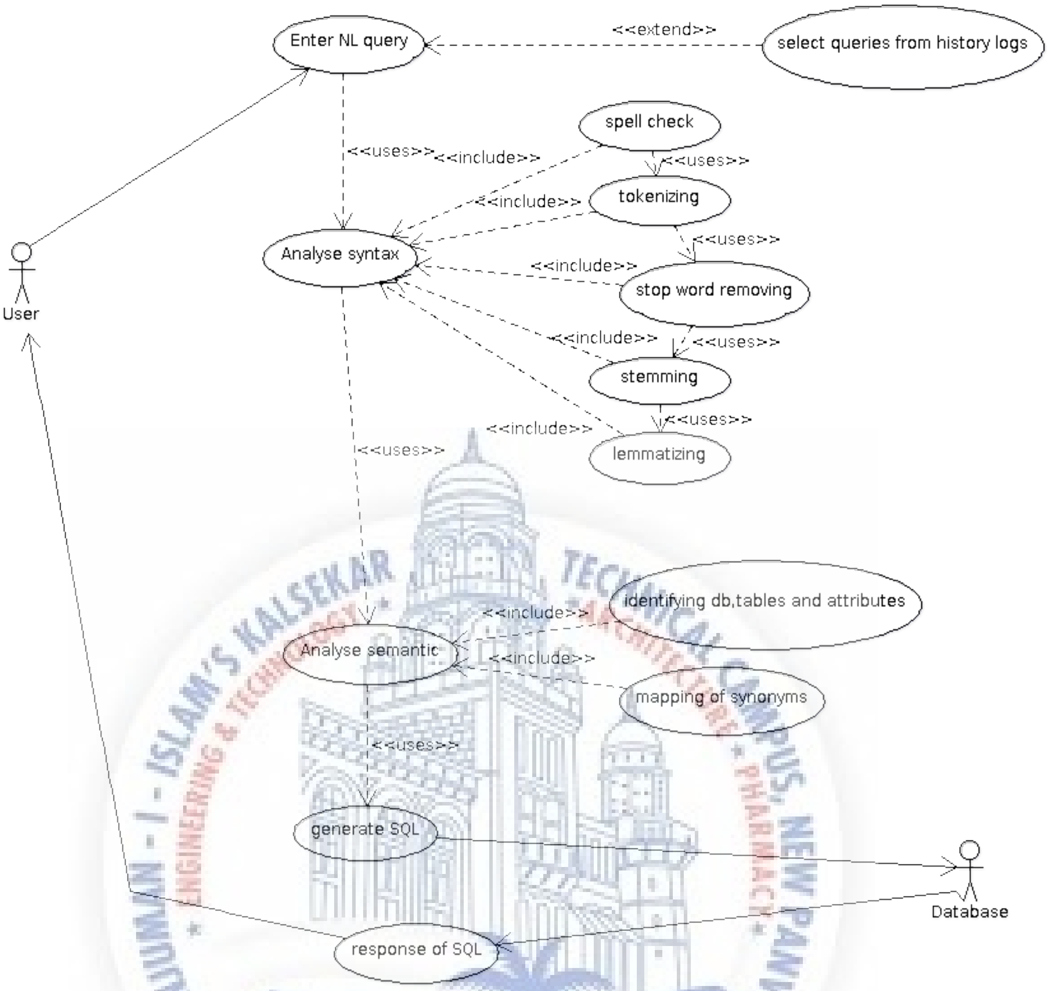


Figure 5.1: Usecase Diagram

5.4.1 Usecase Report

Title:	NLP to create sql query
Description:	Every year,every minute thousands of data is generated and managed.To use this data or retreiving information database interaction is important and for this purpose database experties are required. It restricts the interaction between the naive user and the database. NLP to create sql query helps a normal educated person having no knowledge of query language to easily interact with the database.
Primary Actor:	Normal User having no knowledge of query language. OR User
Preconditions:	User must know english language
Post conditions:	User enters in natural language i.e English, the information he/she is desired to get from database
Main Success Scenario:	1. User enters the information to be retrievd. 2. The script do the spell check,tokenization and generates the corresponding sql query. 3. It is then fired to database and the result is being displayed to user along with the sql query.
Frequency of Use:	User can retrieve information any no. of time.
System Requirement:	Normal, no specific requirement.

Table 5.1: Usecase Report

5.5 Class Diagram

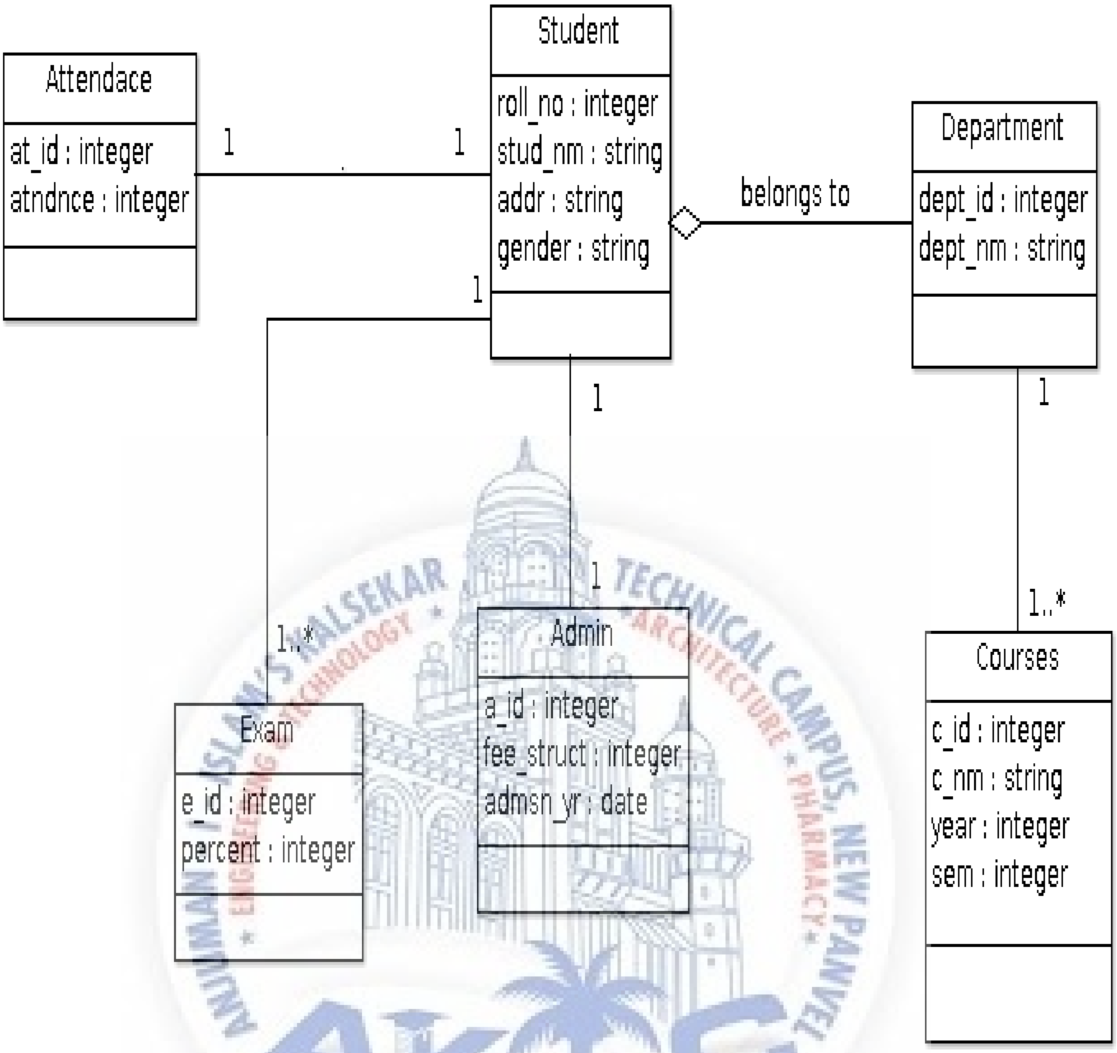


Figure 5.2: Class Diagram

5.5.1 Class Diagram Report

Title:	NLP to create sql query
Description:	Every year,every minute thousands of data is generated and managed.To use this data or retreiving information database interaction is important and for this purpose database experties are required. It restricts the interaction between the naive user and the database. NLP to create sql query helps a normal educated person having no knowledge of query language to easily interact with the database.
Primary Actor:	Normal User having no knowledge of query language. OR User
Preconditions:	CUser must know english language.
Post conditions:	User enters in natural language i.e English, the information he/she is desired to get from database
Web application OR Flask:	The entered search is first passed to python script wherein it is translated to sql query and result is fetched from database.
Python script:	1. It gets the input from web and it apply spell check, lemmatization, tokenization. 2. Mapping the keys and values to the appropriate dictionaries i.e created using tables and attributes, the sql query is generated.
Database:	Databse is used for retrieving the result from. It is the main component.

Table 5.2: Class Diagram Report

Chapter 6

Results and Discussion

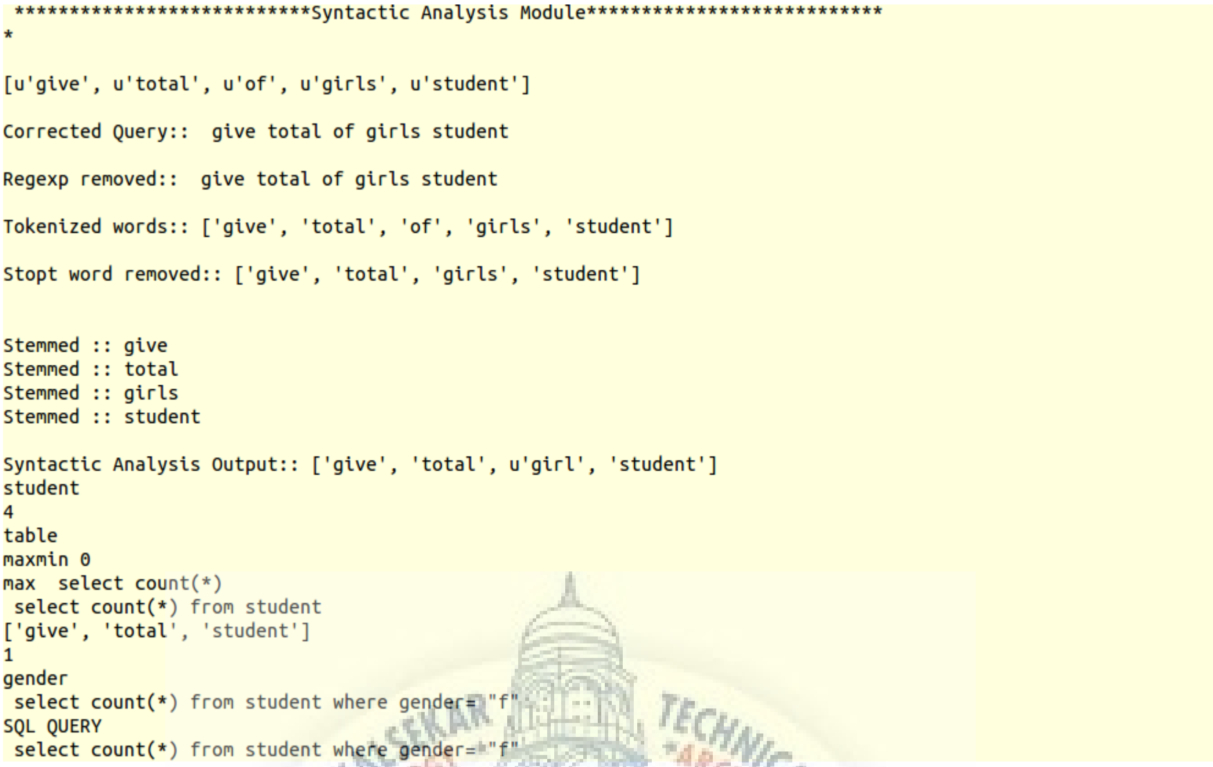
6.1 Test cases and Result

When the user enters the information he/she needs in natural language, the result is fetched from the database and displayed to the user in tabular format. We have tested our web application by considering following test cases:

6.1.1 Unit Testing

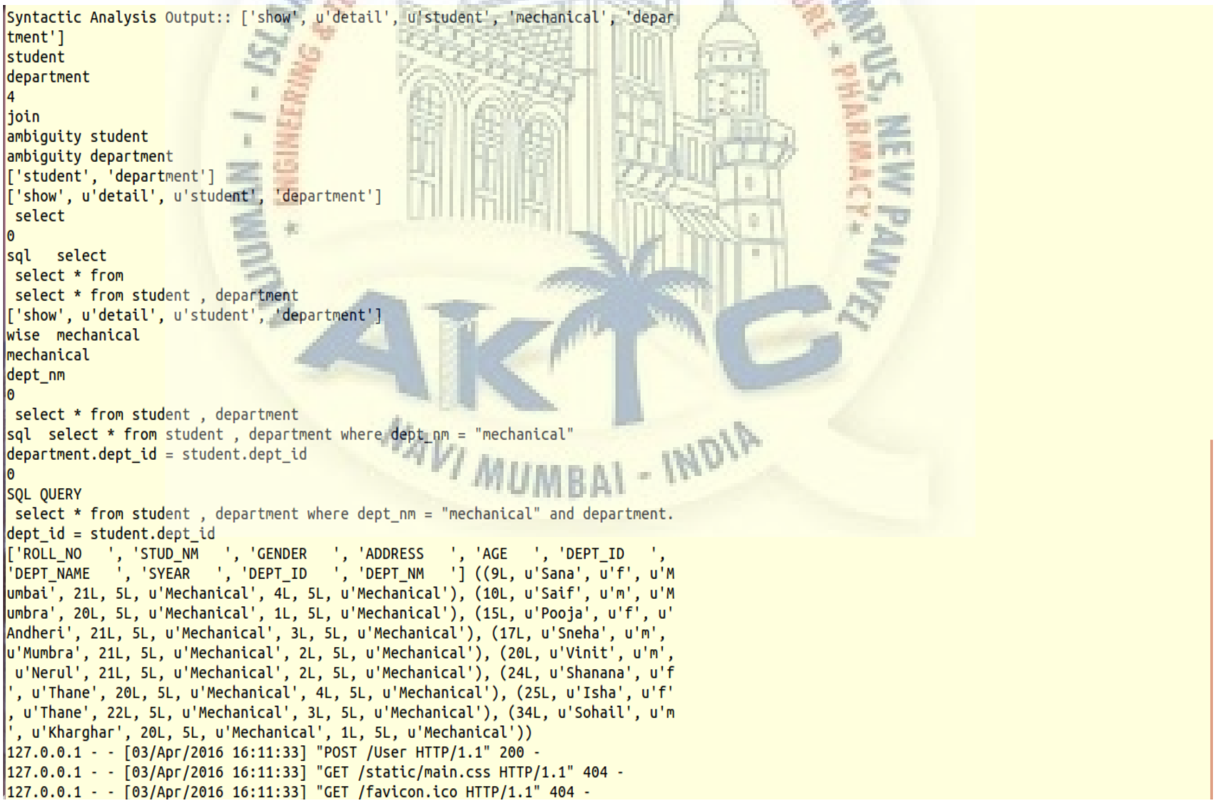
We are firstly performing the spell check, stop word removal followed by lemmatization and tokenization. Example below figures shows how the spelling check is done and regular expressions are being eliminated. It also displays the output of syntactic analysis. These all are performed through the python script.

Example(Figure 6.1): Natural Language Query (give total of girls student):-
Firstly user natural query will be given to sytactic analysis which correct wrong spelled words, tokenize, eliminate stop words, stem and lemmetize and gave the output :['give' , 'total' , 'girl' , 'student'], output of the syntactic analysis is given to the semantic analyzer which identifies the database tables names and attributes names and map the synonyms of them and form the intermediate query :['select' , 'count(*)' , 'girl' , 'student''], since in this output there is a attribute value in the intermediate query so it will form the where clause and the final sql query will be formed : select count(*) from student where gender = 'f '



```
 ***************************Syntactic Analysis Module***************************
*

[u'give', u'total', u'of', u'girls', u'student']

Corrected Query::  give total of girls student

Regexp removed::  give total of girls student

Tokenized words:: ['give', 'total', 'of', 'girls', 'student']

Stopt word removed:: ['give', 'total', 'girls', 'student']


Stemmed :: give
Stemmed :: total
Stemmed :: girls
Stemmed :: student

Syntactic Analysis Output:: ['give', 'total', u'girl', 'student']
student
4
table
maxmin 0
max  select count(*)
 select count(*) from student
['give', 'total', 'student']
1
gender
 select count(*) from student where gender= "f"
SQL QUERY
 select count(*) from student where gender= "f"
```

Figure 6.1: Syntactic analysis module for simple query.



```
Syntactic Analysis Output:: ['show', u'detail', u'student', 'mechanical', 'depar
tment']
student
department
4
join
ambiguity student
ambiguity department
['student', 'department']
['show', u'detail', u'student', 'department']
 select
0
sql   select
 select * from
 select * from student , department
['show', u'detail', u'student', 'department']
wise  mechanical
mechanical
dept_nm
0
 select * from student , department
sql  select * from student , department where dept_nm = "mechanical"
department.dept_id = student.dept_id
0
SQL QUERY
 select * from student , department where dept_nm = "mechanical" and department.
dept_id = student.dept_id
['ROLL_NO   ', 'STUD_NM   ', 'GENDER   ', 'ADDRESS   ', 'AGE   ', 'DEPT_ID   ',
'DEPT_NAME   ', 'SYEAR   ', 'DEPT_ID   ', 'DEPT_NM   '] ((9L, u'Sana', u'f', u'M
umbai', 21L, 5L, u'Mechanical', 4L, 5L, u'Mechanical'), (10L, u'Saif', u'm', u'M
umbra', 20L, 5L, u'Mechanical', 1L, 5L, u'Mechanical'), (15L, u'Pooja', u'f', u'
Andheri', 21L, 5L, u'Mechanical', 3L, 5L, u'Mechanical'), (17L, u'Sneha', u'm',
u'Mumbra', 21L, 5L, u'Mechanical', 2L, 5L, u'Mechanical'), (20L, u'Vinit', u'm',
 u'Nerul', 21L, 5L, u'Mechanical', 2L, 5L, u'Mechanical'), (24L, u'Shanana', u'f
', u'Thane', 20L, 5L, u'Mechanical', 4L, 5L, u'Mechanical'), (25L, u'Isha', u'f'
, u'Thane', 22L, 5L, u'Mechanical', 3L, 5L, u'Mechanical'), (34L, u'Sohail', u'm
', u'Kharghar', 20L, 5L, u'Mechanical', 1L, 5L, u'Mechanical'))
127.0.0.1 - - [03/Apr/2016 16:11:33] "POST /User HTTP/1.1" 200 -
127.0.0.1 - - [03/Apr/2016 16:11:33] "GET /static/main.css HTTP/1.1" 404 -
127.0.0.1 - - [03/Apr/2016 16:11:33] "GET /favicon.ico HTTP/1.1" 404 -
```

Figure 6.2: Syntactic analysis for join query

6.1.2 Functional Testing

We have tested our web application on the localhost server by integrating all the units. In this testing we focus on the output as it is what is required or not which is as follows: On the home page the user need to enter the search and click on the Go button. This will take the user input in natural language form and display the result from database in table format as well as the generated SQL query.

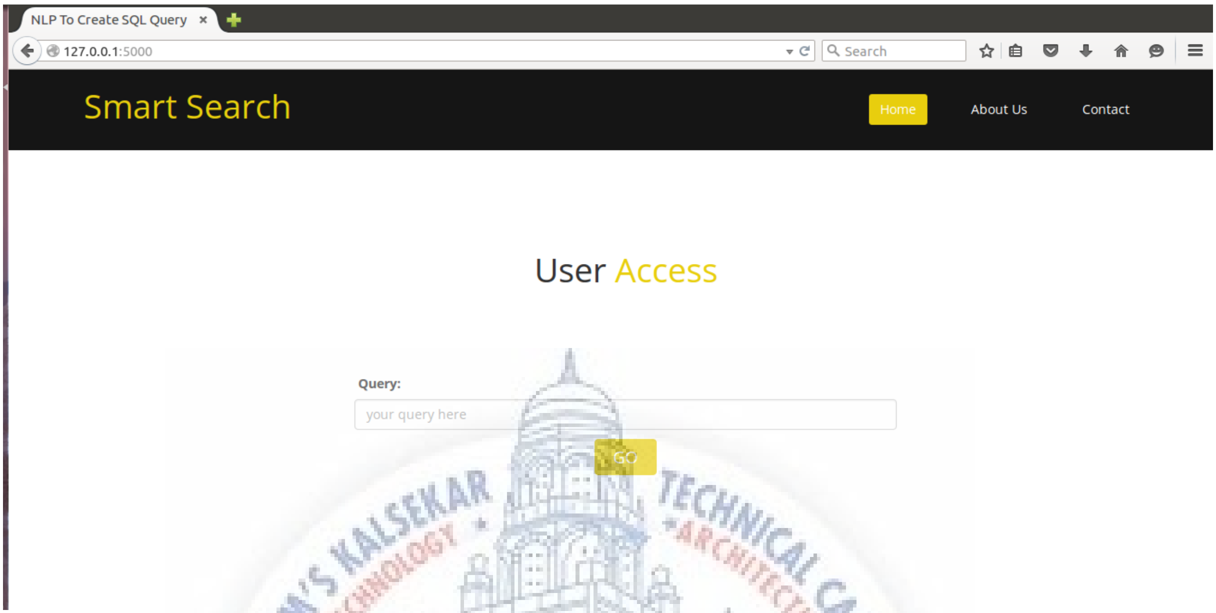
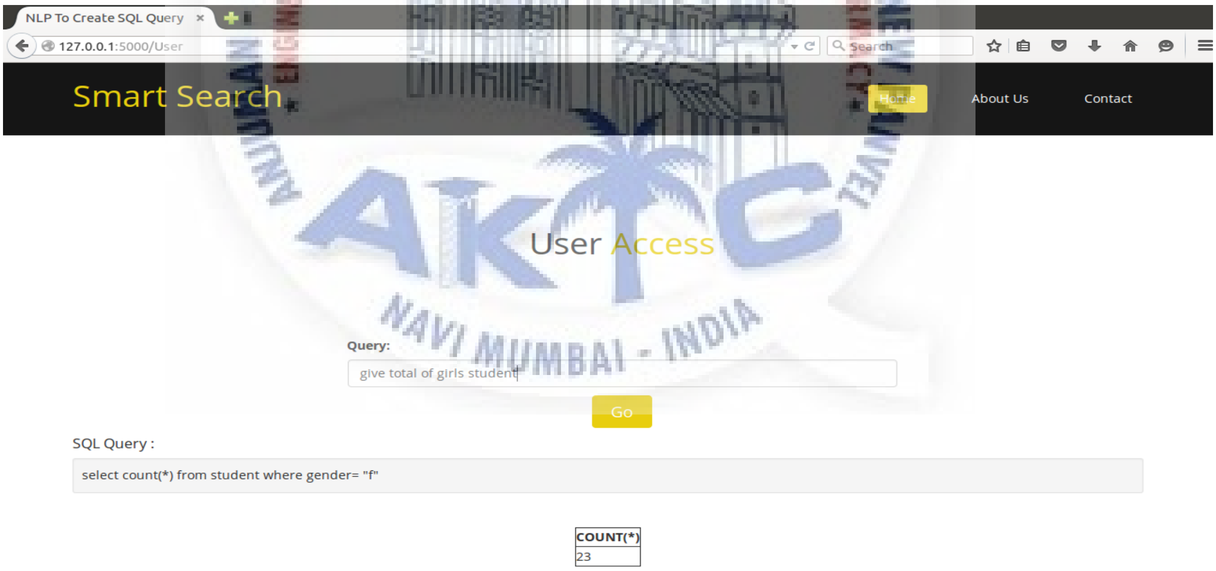


Figure 6.3: Home Page

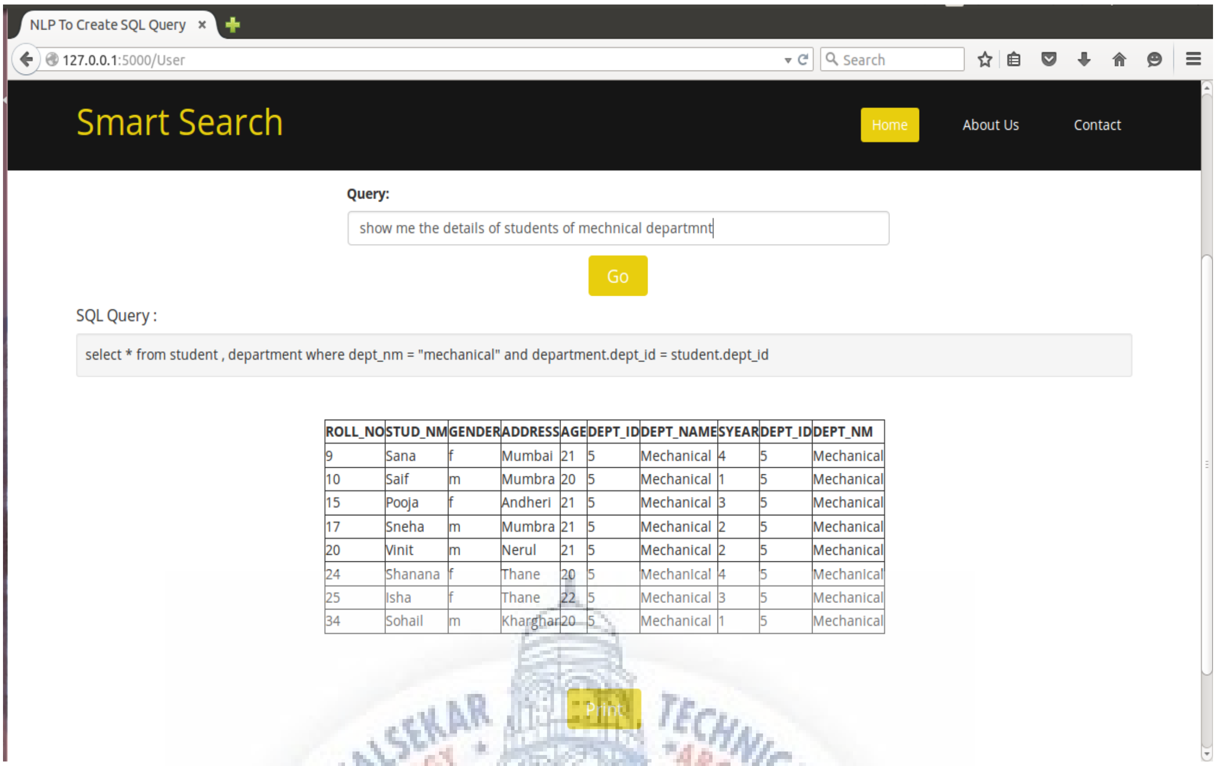


Query: give total of girls student

SQL Query : select count(*) from student where gender= "f"

COUNT(*)
23

Figure 6.4: Simple Query



Query:

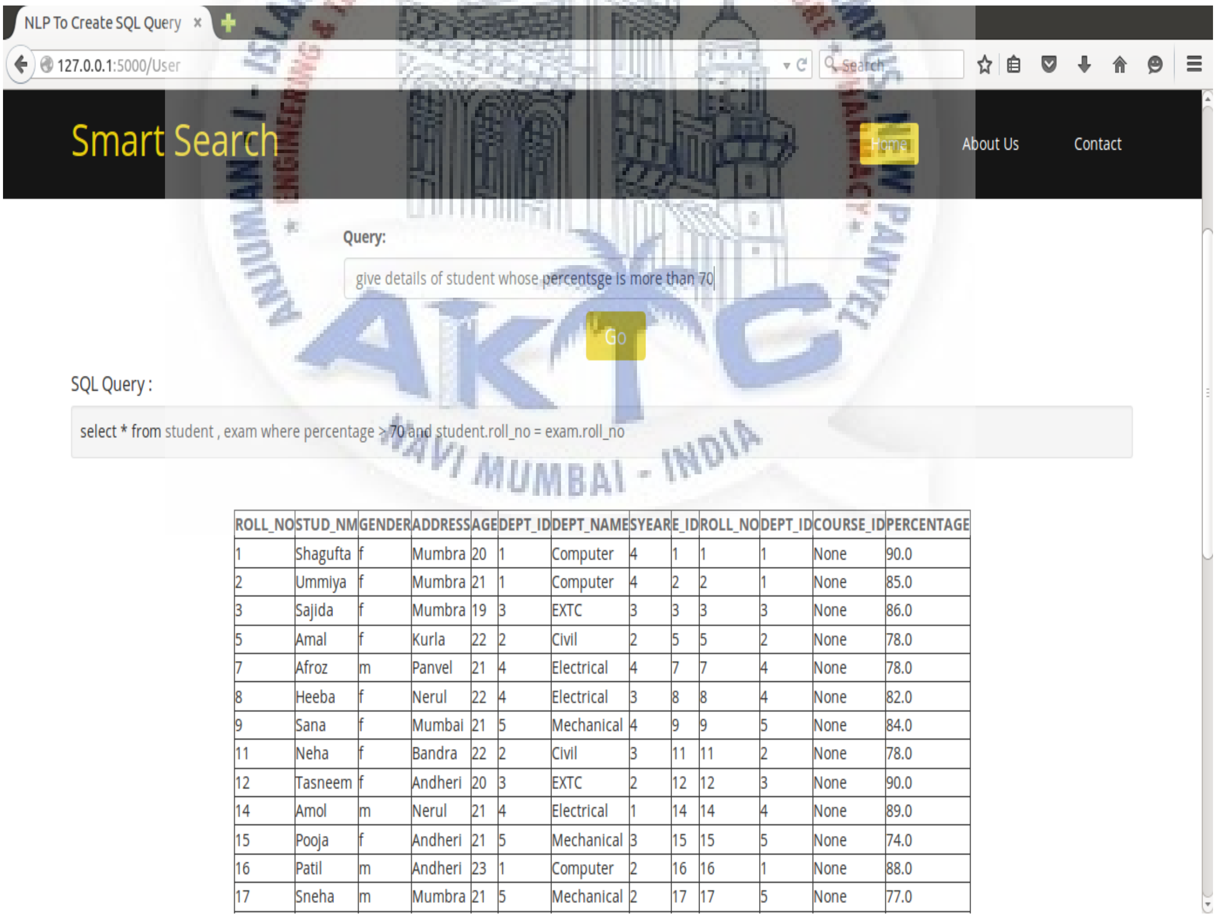
show me the details of students of mechnical departmnt

SQL Query :

select * from student , department where dept_nm = "mechanical" and department.dept_id = student.dept_id

ROLL_NO	STUD_NM	GENDER	ADDRESS	AGE	DEPT_ID	DEPT_NAME	SYEAR	DEPT_ID	DEPT_NM
9	Sana	f	Mumbai	21	5	Mechanical	4	5	Mechanical
10	Saif	m	Mumbra	20	5	Mechanical	1	5	Mechanical
15	Pooja	f	Andheri	21	5	Mechanical	3	5	Mechanical
17	Sneha	m	Mumbra	21	5	Mechanical	2	5	Mechanical
20	Vinit	m	Nerul	21	5	Mechanical	2	5	Mechanical
24	Shanana	f	Thane	20	5	Mechanical	4	5	Mechanical
25	Isha	f	Thane	22	5	Mechanical	3	5	Mechanical
34	Sohail	m	Kharghar	20	5	Mechanical	1	5	Mechanical

Figure 6.5: Join Query



Query:

give details of student whose percentsge is more than 70

SQL Query :

select * from student , exam where percentage > 70 and student.roll_no = exam.roll_no

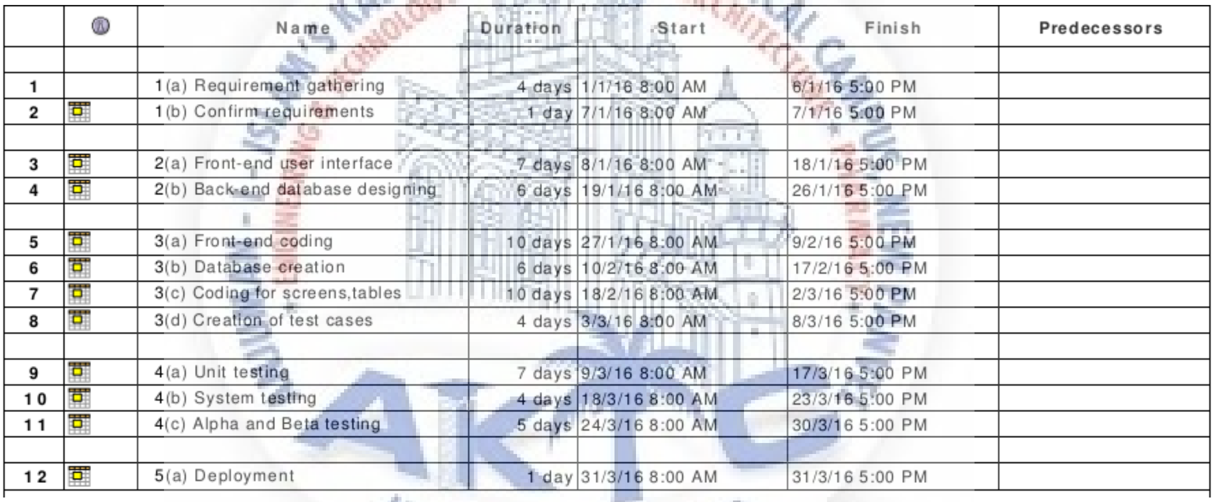
ROLL_NO	STUD_NM	GENDER	ADDRESS	AGE	DEPT_ID	DEPT_NAME	SYEAR	E_ID	ROLL_NO	DEPT_ID	COURSE_ID	PERCENTAGE
1	Shagufta	f	Mumbra	20	1	Computer	4	1	1	1	None	90.0
2	Ummiya	f	Mumbra	21	1	Computer	4	2	2	1	None	85.0
3	Sajida	f	Mumbra	19	3	EXTC	3	3	3	3	None	86.0
5	Amal	f	Kurla	22	2	Civil	2	5	5	2	None	78.0
7	Afroz	m	Panvel	21	4	Electrical	4	7	7	4	None	78.0
8	Heeba	f	Nerul	22	4	Electrical	3	8	8	4	None	82.0
9	Sana	f	Mumbai	21	5	Mechanical	4	9	9	5	None	84.0
11	Neha	f	Bandra	22	2	Civil	3	11	11	2	None	78.0
12	Tasneem	f	Andheri	20	3	EXTC	2	12	12	3	None	90.0
14	Amol	m	Nerul	21	4	Electrical	1	14	14	4	None	89.0
15	Pooja	f	Andheri	21	5	Mechanical	3	15	15	5	None	74.0
16	Patil	m	Andheri	23	1	Computer	2	16	16	1	None	88.0
17	Sneha	m	Mumbra	21	5	Mechanical	2	17	17	5	None	77.0

Figure 6.6: Join Query

Chapter 7

Project Time Line

7.1 Project Time Line Matrix



	ⓘ	Name	Duration	Start	Finish	Predecessors
1		1(a) Requirement gathering	4 days	1/1/16 8:00 AM	6/1/16 5:00 PM	
2	📅	1(b) Confirm requirements	1 day	7/1/16 8:00 AM	7/1/16 5:00 PM	
3	📅	2(a) Front-end user interface	7 days	8/1/16 8:00 AM	18/1/16 5:00 PM	
4	📅	2(b) Back-end database designing	6 days	19/1/16 8:00 AM	26/1/16 5:00 PM	
5	📅	3(a) Front-end coding	10 days	27/1/16 8:00 AM	9/2/16 5:00 PM	
6	📅	3(b) Database creation	6 days	10/2/16 8:00 AM	17/2/16 5:00 PM	
7	📅	3(c) Coding for screens,tables	10 days	18/2/16 8:00 AM	2/3/16 5:00 PM	
8	📅	3(d) Creation of test cases	4 days	3/3/16 8:00 AM	8/3/16 5:00 PM	
9	📅	4(a) Unit testing	7 days	9/3/16 8:00 AM	17/3/16 5:00 PM	
10	📅	4(b) System testing	4 days	18/3/16 8:00 AM	23/3/16 5:00 PM	
11	📅	4(c) Alpha and Beta testing	5 days	24/3/16 8:00 AM	30/3/16 5:00 PM	
12	📅	5(a) Deployment	1 day	31/3/16 8:00 AM	31/3/16 5:00 PM	

Figure 7.1: Time Line Matrix

7.2 Project Time Line Chart

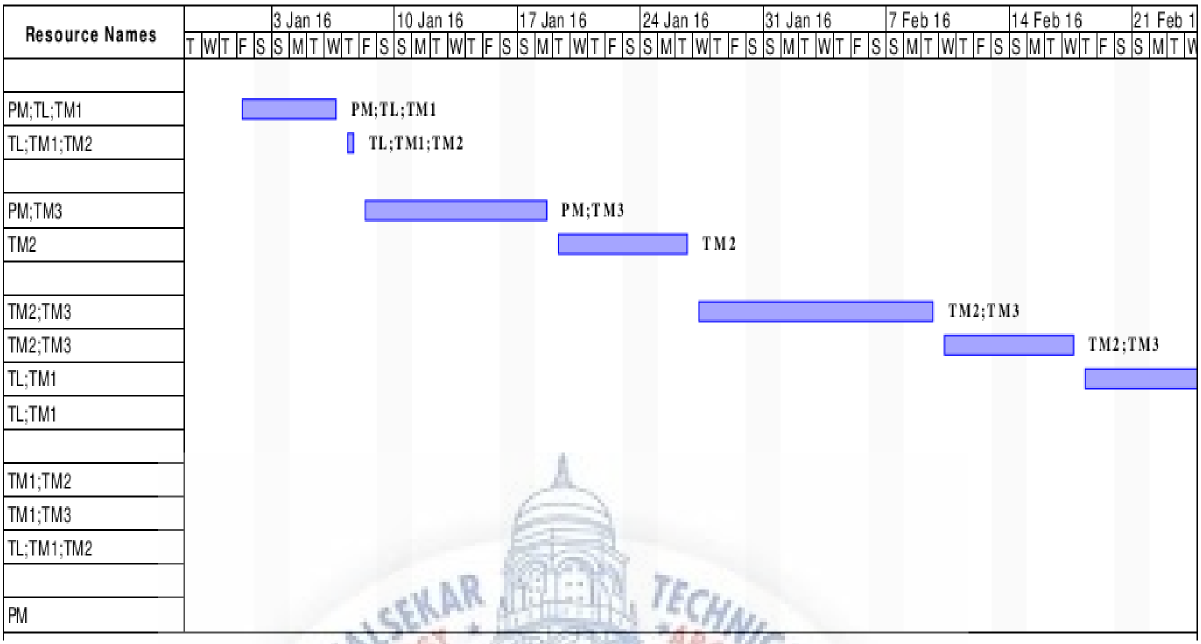


Figure 7.2: Time Line Chart

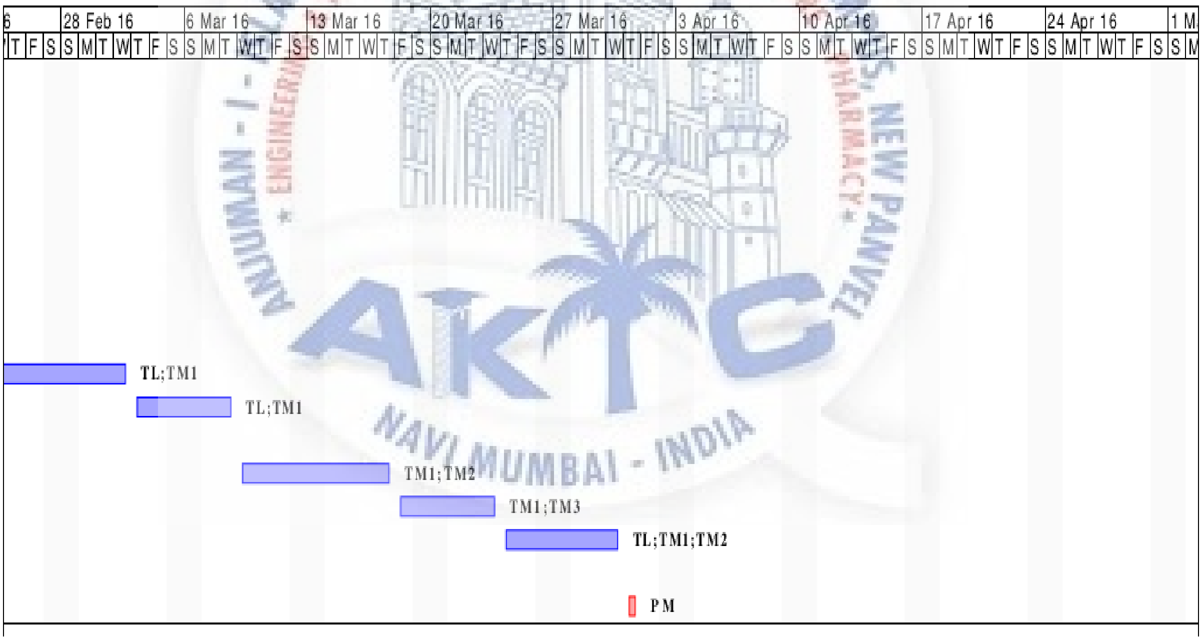


Figure 7.3: Time Line Chart

Chapter 8

Task Distribution

8.1 Distribution of Workload

8.1.1 Scheduled Working Activities

Activity	Time Period	Comment
Requirement Gathering	04 Days	Requirement gathering was to be done through searching on internet and taking the ideas, sharing the views among group members.
Planning	04 Days	Planning was done by reviewing of literature of IEEE papers and by taking the walkthrough.
Design	04 Days	Designing was accomplished by creating UML diagram, charts.
Implementation	70 days	Implementation was started with creating the backend, script and then frontend.
Testing	5 days	Testing has been done by performing unit testing,alpha nad beta testing, integrated testing and system testing.
Deployment	4 days	Deployment phase has been done by installing project on the server.

Table 8.1: Scheduled Working Activities

8.1.2 Members actvities or task

Member	Activity	Time Period	Start Date	End Date	Comment
M1, M2, M3, M4	Requirement Gathering	4 Days	01/01/16	04/01/16	M1 and M2 has perfomed the seaching for project requirement on the internet by reviewing the related literature and by anlysing the related prject which is already available in the market. Regularly inform to the other member of team.
M1, M2, M3, M4	Analysing of the requirement	3 Days	05/01/16	07/01/16	M1, M2, M3, M4 done the requirement analysing of project by sharing the ideas, and by discussing on related information which is gather by the M1, And M2. M3 and M4 has created the list of requirement after every meeting
M1, M2, M3, M4	Finalizing the requirement	1 Day	08/01/16	08/01/16	Whole team finalized the requirement. M1 and M3 has created a list of finalise requirement.
M1, M2, M3, M4	Planning	4 Days	09/01/16	12/01/16	Planning was done by walk-through and by analysing the available applications. M2 and M3 creates a list of funtion which will be implement in the project. Each and every module were discuss in every group meeting and M1 and M2 created a blue print for project .

M3, M4	Front End design	4 Days	13/01/16	16/01/16	M3 and M4 created the UML diagram for frontend of the system and data flow diagrams and informed to the whole team respectively.
M1, M2	Back End design	4 Days	13/01/16	16/01/16	M1 and M2 created the UML diagram for backend of the system and data flow diagrams and informed to the whole team reapectively.
M3, M4	Installation of tools and technology for front end	4 Days	16/1/16	19/1/16	M3 and M4 installed the all the require tools and packages which is used for frontend design.
M1, M2	Installation of tools and technology for back end	4 Days	16/1/16	16/1/16	M1 and M2 installed the all the require tools and packages which is used for backend design.
M3, M4	Implementation of GUI	6 Days	20/1/16	25/1/16	M3 and M4 creates the GUI of the project and informed to other member.
M1	Implementation of script for spelling check, stop word removal	6 days	24/1/16	29/1/16	M1 implemented the script for spell check and stop word removal using nltk packages and explained the code to other team members.
M2	Implementa-tion of script for lemmati-zation and tokenization	6 days	30/1/16	4/2/16	MM2 implemented script for lemmatization and tokenization using nltk packages and explained the code to other team members.
M3,M4	Implementa-tion of dictionaries	7 Days	26/1/16	2/2/16	M3 and M4 implemented the dictionaries using database tables and attributes, synonyms,etc and informed to other team members.

M1,M2	Implementa-tion of script for mapping with dictionaries	25 days	5/2/16	29/2/16	M1 and M2 coded the script for mapping the keys and values to the appropriate dictionaries of tables and attributes and explained the code to other team members.
M3,M4	Connectivity of GUI with script	8 days	4/2/16	11/2/16	M3 and M4 did the connectivity of script with Flask.
M4	Database connectivity	3 days	12/2/16	14/2/16	M4 has done the database connectivity with script and flask.
M1, M2	GUI connectivity	3 days	15/02/16	25/0216	M3 and M4 created the connectivity GUI with database.
M3	Data gathering into database	3 days	18/2/16	20/216	M3 gathered the data required in the database with respect to the domain selected.
M1, M2	Connectivity of Comparison program	8 Days	26/02/15	04/03/15	M1 and M2 created the connectivity of comparison program with scraper, crwaler, and Database. They expalin the code to other member of team.
M1,M2,M3,M4	Integration of all modules	15 days	5/3/16	19/3/16	M1, M2,M3 and M4 integrated all the module. Implemented whole system properly.
M1,M2	Connectivity of indexing with database	2 Days	05/03/15	07/03/15	M4 makes the indexing program connectivity with database. And informed to the other member of team.
M3,M4	Unit testing	4 days	26/3/16	29/3/16	M3 and M4 performed the functional testing and noted down results and discussed with other members of team.
M1,M2,M3,M4	Deployment	1 day	30/3/16	30/3/16	M1 and M2 makes the connectivity of searching algorithm program with GUI.

Table 8.2: Member Activities and Task

Chapter 9

Conclusion and Future Scope

9.1 Conclusion

NLP for sql generation is very crucial aspect for naive and non technical person to interact with the database system and this proposed system fulfills the requriment of the user to handle the database system.
System is designed that converts the English language to the sql in order to retrive the data from the database.Proposed system is domain independent that is it can be used any database application not restricted to particular application.Complex database queries can be evaluated which are asked in natural language.Queries include order queries,join queries,nested queries,range queries,comparison predicates, conjunctions, quantifications, multi-level aggregations etc.It is designed for DDL and DML statements as well.System is based on Intermeduate Representation technique which is internal represntation query to and it is the combination of sysntactic and semantic based system.System is also designed to deal with query logs in order to reduce the duplicate search to the system and help the user to interact with the system easily.

9.2 Future Scope

1. To accept queries in vernacular languages.
2. To include question based on prediction in case of Information Retrieval system. For example, user can ask question like: âœwhen the student puja will complete the final year of her studies?â, âœwhat will happen if a student failâ, etc.
3. To support multimedia data such as image, sound and graphics can be attempted.
4. To include computational phonology and text-to-speech.

References

[1] Manju Mony ,Jyothi M. Rao ,Manish M. Potey *An Overview of NLIDB Approaches and Implementation for Airline Reservation System* International Journal of Computer Applications (0975 â“ 8887) Volume 107 â“ No 5, December 2014

[2] Saravjeet kaur,Rashmeet Singh Bali Generation and Excution From Natural Language Processing,International Journal of Computing and Business Research, ISSN 2229-6166

[3] Anuradha Mohite, Varunakshi Bhojane *Challenges and Implementation Steps of Natural Language Interface for Information Extraction from Database,*International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-3, Issue-1, March 2014

[4] F.Siasar djahantighi1, M.Norouzifard1, S.H.Davarpanah, M.H.Shenassa *Using Natural Language Processing in Order to Create SQL Queries* Proceedings of the International Conference on Computer and Communication Engineering 2008 May 13-15, 2008 Kuala Lumpur, Malaysia

[5] F.Siasar djahantighi1, M.Norouzifard1, S.H.Davarpanah, M.H.Shenassa *Using Natural Language Processing in Order to Create SQL Queries* Proceedings of the International Conference on Computer and Communication Engineering 2008 May 13-15, 2008 Kuala Lumpur, Malaysia

[6] Dr. Paresh Virparia ,Amisha Shingala *Design and Development of Natural Language Query Interface for Relational Databases*

[7] I. Androutsopoulos G.D. Ritchie P. Thanisch*Natural Language Interfaces to Databases â“ An Introduction*University of Edinburgh 80 South Bridge, Edinburgh EH1 1HN, Scotland, U.K

[8] Pooja A.Dhomne1, Sheetal R.Gajbhiye, Tejaswini S.Warambhe Vaishali B.Bhagat*ACCESSING DATABASE USING NLP* IJRET: International Journal of Research in Engineering and Technology eISSN: 2319-1163 | pISSN: 2321-7308

Own Publications

Shaikh Shagufta Shezad Zubeda ,Momin Ummiya Salim Rehana,Shaikh Sharmeen Salim Shahina ,Shaikh Rameeza Mohammad Yaqoob *NLP To Create SQL Query* IJSRD International Journal for Scientific Research Development. Volume-03 ,Issue-09,2015

NLP TO Create SQL Query

Prof.Khan Tabrez[1] Shaikh Shagufta[2] Shaikh Sharmeen[3] Momin Ummiya[4] Shaikh Rameeza[5]
[1,2,3,4,5]Department of Computer Engineering
[1,2,3,4,5]Anjuman-I-Islam Kalsekar Technical Campus School of Engineering

Abstract—NLP to create sql query, this intelligent system converts the human language into the structured query language. Every year, every minute thousands of data is generated and managed. To use this data or retrieving information, database interaction is important and for this purpose expertise are required. The problem is that it restricts the interaction between the naive user and the database. Only few people who have knowledge of formal database language can retrieve the desired information from the database. To overcome such a problem this proposed system would have the capability to analyze the user statements written in different ways and accordingly it gives response to the user. In this way it help a normal educated person having no knowledge of query language to easily interact with the system.

Key words: NLP, Morphoms

I. INTRODUCTION

In the present computing world, computer based information technologies have been extensively used to help many, private companies, academic and education institutions to manage their processes and information systems. Information systems are used to manage data. The information management system that is capable of managing several kinds of data, stored in the database systems. is known as Database Management System (DBMS). Databases are comprehensive element in private and public information systems which are essential in number of application areas. Databases are built with the objective of facilitating the activities of data management in information systems. Due to the progress and in deep applications of computer querying system. It is due to the fact that the technology in several areas to be accurate, databases have become the repositories of huge volumes of data .In relational databases, to retrieve information from a database, one needs to formulate a query in such way that the computer will understand and produce the desired output.

II. KEYWORDS

NLP (Natural Language Processing, IQR (Intermediate Query Representation), Morphoms (individual word, token (broken words), SQL (Structured Query Language).

III. PROBLEM AND SOLUTION

The limitation of such programs is that it restricts the interaction between user and database to predefined set of queries. Only few people who have knowledge of database structure and formal database language (such as Structured Query Language (SQL) can retrieve the desired information from database. A novice user having no knowledge of database structure and formal database query language cannot retrieve desired information if it is not supported by well thought application. Hence, it is a need to improve human computer interface that allows people to interact with the database in their natural language (such as English).

IV. OBJECTIVE AND SCOPE

The objective of the proposed system is the interaction of the naive user and intelligent system. The proposed system would have the capability to understand the natural language, learn the meaning of the query and process it and give the desired result. Since normal educated user can't access the database to overcome this problem there is a need to translate the normal user language such as English into SQL query in order to get the result from the database, here user need not to learn anything related to database and can interact directly to the database in his/her known language.

Proposed system can be used in any application. Where there is need to store the data and retrieving of data is done by non-technical person or naive user, this system is useful:

School, Colleges, Hospital, Transport, Bank ,Government Office, Service Sector, Navy, Manufacturing database, Sensus system, Agriculture, Chat system, Business organization, Chemical Industies.

V. LITERATURE REVIEW

An Overview of NLIDB Approaches and Implementation for Airline Reservation System: [1]

This system was developed for Flight Reservation. A combination of Syntax Analysis and Intermediate Query approach is used for this system. Syntax Analysis performs syntactic processing and breaks the input sentence into its constituent parts and identifies the relations between the concepts. Intermediate Query approach allows to easily perform the mapping of concepts to an intermediate representation. The intermediate representation can be used even in case of database portability i.e. even if database is ported to another database.

– Weaknesses
– This system is not developed for DDL (Data Definition Language).
– It is not implemented for complex queries like nested, join, group, order, queries, etc.
– Modification and Deletion to the database in this system is not done by natural language.
– It is restricted to single user only.
– It is also domain dependent i.e limited for airline reservation system.

A. *Solution*

1) Proposed system is designed for DDL as well as DML statement.
2) It would have the ability to deal with complex queries.
3) It would deal with the ambiguity in the query.
4) It would be accessible by multiple users.
5) It would be domain independent i.e. applicable to any database applications.
6) There will se security for admin and normal user so

that data can be more secured.

VI. PROPOSED SYSTEM ARCHITECTURE

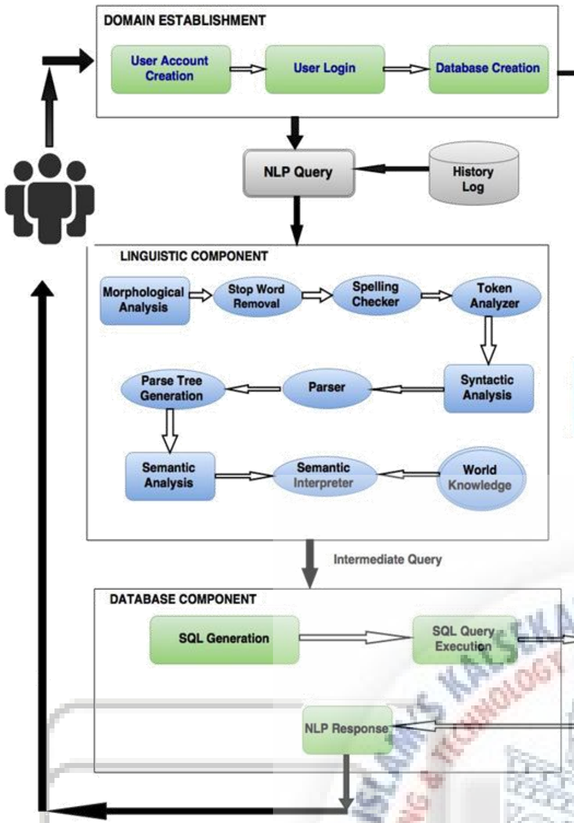


Fig. 1: System Architecture

The System Architecture consists of three modules:

- DOMAIN ESTABLISHMENT
- LINGUISTIC COMPONENTS
- DATABASE COMPONENTS

A. *Domain Establishment:*

This module is responsible for creating user accounts and database creation as the proposed system is would be used by any database application and by multiple users.

B. *Linguistic Component:*

This module is responsible for translating natural language input into a logical query. In this, the sentence is syntactically and semantically analyzed and processed, and an intermediate query is generated by the following steps.

- Morphological Analysis
- Syntactic Analysis
- Semantic Analysis

1) *Morphological Analysis*

Morphology in linguistics is the study and description of how words are formed in natural language. In this phase the sentence is broken down into tokens- smallest unit of words, and determine the basic structure of the word.[6]

For instance, unusually can be thought of as composed of a prefix un-, a stem usual, and an affix -ly. Composed is compose plus the inflectional affix -ed: a spelling rule means we end up with composed rather than composed.

2) *Stop Word Removal:*

Stop words are non-context bearing words, also known as noisy words which are to be excluded from the input sentence to speed up the process. For example again, already, amongst, etc. [6]

3) *Spelling check:*

Three methods for spelling check are as follows: [6]

Correct Token	Error Token	Spelling Checking Operation
student	dudent	Substitution
student	studnt	Insertion
student	studeent	Deletion

4) *Token Analyzer: [6]*

Each identified tokens can be represented as attribute token, value token, core token, multi-token, continuous token, etc.

- Attribute token- using metadata
- Core Token-first, all capital letters
- Numeric Token-digits , digits separated by decimal point
- Sentence Ending Markers- (. ? !)
- Value Token- (M.C.A, “mca”, ‘mca’)
- Continuous Token – (‘@’, apostrophe (‘) , ‘$’)
- Multi-token- emp_no or e-no
- Abbreviated Token- CE for Computer Engineering

5) *Syntactic Analysis:*

The objective of the syntactic analysis is to find the syntactic structure of the sentence.It is also called Hierarchical analysis/Parsing, used to recognize a sentence, to allocate token groups into grammatical phrases and to assign a syntactic structure to it.[2]

6) *Parse Tree: [5]*

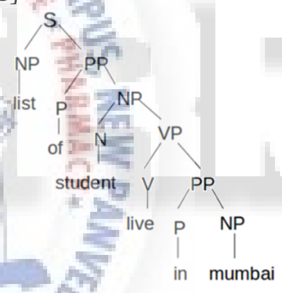


Fig. 2: Parse Tree

Parser generates a parse tree with the help of syntactic analysis. A parse tree or parsing tree is an tree in the ordered and structured form that represents the syntactic structure of a string according to some context free grammar.

Example: list of student live in Mumbai

7) *Semantic Analysis:*

Semantic Analysis is related to represent the meaning of linguistics sentences. It concerns with how to determine and understand the meaning of the each word. So, it is responsible for creating the logical query which acts as the input query to Database Query Generator. Hence this is another form of presenting the user tokens in the form of semantic word.[2]

8) *IQR generation:*

It becomes very difficult to map the syntax tree and semantic tree of the sentence directly to the sql query, an intermediate query is generated from the semantic analysis then this logical query is going to be converted into sql query. The logical query expresses the meaning of the user’s question in terms of user world knowledges, which are independent of the structure of the database. The logical query is then translated to database’s query language expression, and generates the user understandable result. [4]

C. Database Component:

1) SQL query generation:

It consists of SQL query generation where intermediate query is going to map to the sql query.

2) SQL query execution:

Generated sql query is going to be executed here and desired data is extracted from the database.

3) Response of NLP:

Extracted result is going to be displayed to the user.

VII. METHODOLOGY

Proposed system is based on Intermediate Representation Technique which is a combination of syntactic based system and semantic based system. System follows the following steps:

A. Domain Creation:

- User create the account
- User login to the system
- User creates the respective database
- User request the data from the database in English language

B. Query Analysis:

- This phase consists of morphological syntactic, semantic analysis of the sentence
- English query is split up into words.
- Morphons are recognized from the sentence.
- Stop words are removed from the sentence.
- Error detection and correction of words will be processed
- Parser is responsible for the parse tree generation using NLTK library of Python.
- Part of speech tagging, sementic meaning of senetence takes place using Pattern library of Python.

C. Intermediate Query Generation [4]

Intermediate query is generated from above steps.For example: the question “What is the capital of each country bordering UP?” would be mapped to the following logical query.

answer([Capital, Country]):-
iscountry(Country),
border(Country, UP),
capitalof(Capital, Country).

D. Query Mapping: [4]

Mapping of words to sql attributes is done using Database Dictionary and semantic rules. Example of mapping and sql query generation for the above intermediate query. The mapping information could link the predicate is country to the Sql query. Example of above IQ. iscountry(Country),

SQL >> SELECT country FROM countries_table;

E. Response Generation:

- SQL query is generated and executed.
- Result is displayed to the user.

Example of the above query.
>>countries table;

Country
India
USA
Africa
Australia
Canada

VIII. COMPETITIVE ADVANTAGE OF PROJECT

There are many softwares are developed for processing the natural language to generate sql query and to extract data from database which are as follows:

A. Lunar:

It was involved in a system that answered questions about rock samples brought back from the moon.[5]

B. Liffer/Ladder:

It was designed interface of natural language to a database of information about US Navy ships.[5]

C. Chat-80:

It is designed to know the location of ocean, rivers, cities and countries. Semantic grammar technique is used in this system.[5]

D. Precise:

It is developed for Air Travel Information System & GEOQUERY. Lexical analysis and semantic constrains approach is used.

All the above systems are domain dependent system that means used for particular database application. These are based on shallow parsing, semantic grammar-based system, syntactic based system and it is very difficult to convert the natural language sentence to directly to the database query and the proposed system is domain independent.

Proposed system is based on IQR Technique which is internal logical query so it becomes easy to convert the logical query to the sql query. Another benefit is that existing systems does not maintain the history log of the query abd proposed system does. Besides these another benefit of the proposed system over existing system is that it can be used by many users.

IX. CONCLUSION

NLP for sql generation is very crucial aspect for naive and non-technical person to interact with the database system and this proposed system fulfills the requirements of the user to handle the database system.

System is designed that translate the english language to the sql in order to retrieve the data from the database. Propoesd system is domain independent that is it can be used any database application not restricted to particular application. Complex database queries can be evalutaed which are asked in natural language. Queries include order queries, join queries, nested queries, range queries, comparison predicates, conjunctions, quantifications, multi-level aggregations etc. It is designed for DDL and DML statements as well. System is based on IRQ technique which is internal representation query to and it is the combination of syntactic and semantic based system. System is also designed to deal with query logs in to reduce the duplicate interaction to the system to help the user to interact with the system.

X. Future Scope

- To accept queries in vernacular languages.
- To include question based on prediction in case of Information Retrieval system. For example, user can ask question like: "when the student puja will complete the final year of her studies?", "what will happen if a student fail", etc.
- To support multimedia data such as image, sound and graphics can be attempted.
- To include computational phonology and text-to-speech.

Acknowledgment

We would like to extend our deepest gratitude to our HOD Prof.Tabrez Khan for extending his valuable guidance in the research. His mentoring gave way a direction to our work.

References

[1] Manju Mony ,Jyothi M. Rao ,Manish M. Potey ,An Overview of NLIDB Approaches and Implementation for Airline Reservation System ,International Journal of Computer Applications (0975 – 8887) Volume 107 – No 5, December 2014

[2] F.Siasar djahantighi1, M.Norouzifard1, S.H.Davarpanah, M.H.Shenassa,Using Natural Language Processing in Order to Create SQL Queries ,Proceedings of the International Conference on Computer and Communication Engineering 2008 May 13-15, 2008 Kuala Lumpur, Malaysia .

[3] Dr. Paresh Virparia ,Amisha Shingala ,Design and Development of Natural Language Query Interface for Relational Databases.

[4] I. Androutsopoulos G.D. Ritchie P. Thanisch ,Natural Language Interfaces to Databases – An Introduction ,University of Edinburgh 80 South Bridge, Edinburgh EH1 1HN, Scotland, U.K.

[5] Pooja A.Dhomne1, Sheetal R.Gajbhiye, Tejaswini S.Warambhe Vaishali B.Bhagat ACCESSING DATABASE USING NLP ,IJRET: International Journal of Research in Engineering and Technology eISSN: 2319-1163 | pISSN: 2321-7308

[6] Ms. Amisha H. Shingala and Dr. Paresh V. Virparia, Research paper on Intelligent Natural Language Processor, (under press) in National Journal of Systems and Information Technology (NJSIT), ISSN:0974-3308.

DJ ASCII

ISBI: 978-93-85777-43-1

PAPER ID: AI25

m

University Classification and Prediction using Data Mining

Shaikh Shagufta Shezad , Shaikh Sharmeen Salim, Momin Ummiya Salim,Shaikh Rameeza Mohammad Yaqoob and Prof. Khan Tabrez Mohd.Tahirs

B.E. Student, Dept. of Computer Engineering,
AIKTC, New Panvel, Mumbai University, India
H.O.D Department of Computer, AIKTC,
New Panvel, Mumbai University, India

NLP to create sql query , this intelligent system converts the human language into the structured query language Every year,every minute thousands of data is generated and managed.There is a need to interact with the in database using SQL syntax to retrive information interaction is important and for this purpose database experties are required.The problem is that it restricts the interaction between the naive user and the database. Only few people who have knowledge of formal database language can retrieve the desired information from the database. To overcome such a problem this proposed system will help a normal educated person having no knowledge of query language to easily interact with the database and the system will provide the instant search for the desired result the user is looking for. The user need not to look into each file for the data .The system will help to featch the instant result where there is need for more computation power and where there is a need for Quick decision from the output .

Chapter 10

Appendix I

10.1 Flask

Flask is a micro web application framework written in Python and based on the Werkzeug toolkit and Jinja2 template engine. It is BSD licensed. Examples of applications that make use of the Flask framework are Pinterest, LinkedIn, as well as the community web page for Flask itself.

Flask is called a microframework because it does not presume or force a developer to use a particular tool or library. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions. However, Flask supports extensions, that can add application features as if they were implemented in Flask itself. Extensions exist for object-relational mappers, form validation, upload handling, various open authentication technologies and several common framework related tools.

10.1.1 Features of Flask

- Contains development server and debugger
- Integrated support for unit testing
- RESTful request dispatching
- Uses Jinja2 templating
- Support for secure cookies (client side sessions)
- 100% WSGI 1.0 compliant

- Unicode-based
- Extensive documentation
- Google App Engine Compatibility
- Extensions available to enhance features desired.

ACKNOWLEDGMENT

I would like to take the opportunity to express my sincere thanks to my guide **Prof. Tabrez Khan**, Assistant Professor, Department of Computer Engineering, AIKTC, School of Engineering, Panvel for his invaluable support and guidance throughout my project research work. Without his kind guidance & support this was not possible.

I am grateful to him/her for his timely feedback which helped me track and schedule the process effectively. His/her time, ideas and encouragement that he gave me have helped me to complete my project efficiently.

I would also like to thank **Dr. Abdul Razak Honnutagi**, AIKTC, Panvel, for his encouragement and for providing an outstanding academic environment, also for providing the adequate facilities

I am thankful to **Prof. Tabrez Khan**, HOD, Department of Computer Engineering, AIKTC, School of Engineering, Panvel and all my B.E. teachers for providing advice and valuable guidance.

I also extend my sincere thanks to all the faculty members and the non-teaching staff and friends for their cooperation.

Last but not the least, I am thankful to all my family members whose constant support and encouragement in every aspect helped me to complete my project.

Shaikh Shagufta Shezad Zubeda (13CO69)

Momin Ummiya Salim Rehana (13CO68)

Shaikh Sharmeen Salim Shahina (12CO13)

Shaikh Rameeza Mohammad Yaqoob (12CO07)
(Department of Computer Engineering)
University of Mumbai.