Mini Project Report on

# Generating SQL query from Natural Language using NLP

Submitted by

|  |  |  |
| --- | --- | --- |
| Name of student | Class | Roll No |
| Rakesh Chunilal Choudhary | TE3 | 08 |
| Satyam Subhash Gupta | TE3 | 14 |
| Yogesh Sudhakar Hole | TE3 | 15 |

Under the guidance of

Prof. Deepshikha Chaturvedi



DEPARTMENT OF COMPUTER ENGINEERING

SHAH AND ANCHOR KUTCHHI ENGINEERING COLLEGE CHEMBUR, MUMBAI - 400088.

2020-2021

# 

Certificate

# *This is to certify that the report of the project*

# Generating SQL query from Natural Language using NLP

is a bonafide work of

|  |  |  |
| --- | --- | --- |
| Name of student | Class | Roll No |
| Rakesh Chunilal Choudhary | TE3 | 08 |
| Satyam Subhash Gupta | TE3 | 14 |
| Yogesh Sudhakar Hole | TE3 | 15 |

submitted to the

**UNIVERSITY OF MUMBAI** during semester VI

in

## COMPUTER ENGINEERING DEPARTMENT

---------------------------- ------------------------------------

(Prof. Deepshikha Chaturvedi) (Prof. Uday Bhave)

Guide I/c Head of Department

## Approval for Mini Project Report for T. E. Semester VI

This mini project report entitled “**Generating SQL query from Natural Language using NLP**” by Rakesh Chunilal Choudhary, Satyam Subhash Gupta, and Yogesh Sudhakar Hole is approved for the partial fulfillment of the requirement for the completion of Semester VI.

Name and Sign of Internal Examiner \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Name and Sign of External Examiner \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Date: 24/05/2021

Place: Mumbai

## Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We 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.

|  |  |  |  |
| --- | --- | --- | --- |
| Name of student | Class | Roll No | Signature |
| Rakesh Chunilal Choudhary | TE3 | 08 |  |
| Satyam Subhash Gupta | TE3 | 14 |  |
| Yogesh Sudhakar Hole | TE3 | 15 |  |

Date: 24/05/2021

Place: Mumbai

## Attendance Certificate

Date: 24/05/2021

To,

The Principal

Shah and Anchor Kutchhi Engineering College,

Chembur, Mumbai-88

Subject: Confirmation of Attendance

Respected Sir,

This is to certify that Third year (TE) students

(Rakesh Chunilal Choudhary, Satyam Subhash Gupta, Yogesh Sudhakar Hole)

have duly attended the sessions on the day allotted to them during the period from 8/01/2021 to 15/04/2021 for performing the Mini Project titled “Generating SQL query from Natural Language using NLP”.

They were punctual and regular in their attendance. Following is the detailed record of the student’s attendance.

Attendance Record:

|  |  |  |  |
| --- | --- | --- | --- |
| Date | Rakesh Choudhary | Satyam Gupta | Yogesh Hole |
|  |  |  |  |
| 27/01/2021 | **Present** | **Present** | **Present** |
| 3/02/2021 | **Present** | **Present** | **Present** |
| 10/02/2021 | **Present** | **Present** | **Present** |
| 17/02/2021 | **Present** | **Present** | **Present** |
| 24/02/2021 | **Present** | **Present** | **Present** |
| 03/03/2021 | **Present** | **Present** | **Present** |
| 18/03/2021 | **Present** | **Present** | **Present** |
| 25/03/2021 | **Present** | **Present** | **Present** |
| 01/04/2021 | **Present** | **Present** | **Present** |
| 15/04/2021 | **Present** | **Present** | **Present** |

Signature and Name of Internal Guide

**Abstract**

This paper describes a method for effectively automating the conversion of Natural Language Query to SQL Query. SQL is a potent tool for managing data in a relational databases. To retrieve or process data, the user must enter the appropriate SQL Query. However, users who are unfamiliar with SQL are unable to retrieve relevant data. To address this, we suggested a model in Natural Language Processing for converting Natural Language Query to SQL query. This allows a naive user to obtain the required data without having to know any complex SQL details. This system is also capable of handling complex queries. This system is intended for users who work with student databases but are unfamiliar with SQL.

The system will accept text input. This natural language query will be converted into a SQL query. The query will be executed by the system, and the user will receive the results.

**Table of contents:-**

[**Abstract**](#_Toc72246250) vi

**List of figures** vii

[**List of tables**](#_Toc72246253) viii

1. [Introduction](#_Toc72246254) 1
2. [Literature Survey](#_Toc72246255) 2

2.1 Survey of existing systems 2

2.2 Limitations of existing systems and research gaps 4

2.3 Objectives 5

2.4 Literature Survey Table 5

1. Problem Statement 7
2. [Project Design](#_Toc72246257) 8

4.1 System Block Diagram 8

4.2 Flowchart 9

4.3 Algorithm 10

**5. Implementation Details 12**

5.1 Module Description 12

5.2 Snapshot **13**

**6. Results and Analysis 15**

**7. Conclusion and Future scope 16**

**8. References 17**

**Acknowledgement 18**

**List of figures:-**

***Figure.4.1:*** *System Block Diagram*

***Figure.4.2:*** *Flowchart*

***Figure.5.1:*** *Example 1 input and output*

***Figure.5.2:*** *Example 1 retrieved data from database*

***Figure.5.3:*** *Example 2 input and output*

***Figure.5.4:*** *Example 2 retrieved data from database*

**List of Tables:-**

**Table.2.1:**Literature Survey Table

**Chapter 1**

**Introduction**

Natural Language Processing is a branch of Artificial Intelligence that is used to create intelligent computers that can interact with humans in the same way that humans do. It connects the human-machine gap. The primary goal of Natural Language Query Processing is to have a computer interpret English Sentence structures. Despite these problems, it is commonly used for research purposes. Natural Language Processing is used to access the database by asking Natural Language questions and receiving the necessary results. Asking questions via natural language to databases is a very convenient and easy method of data access, particularly among users who are unfamiliar with complex database query languages such as SQL (Structured Query Language). This system is intended for users who deal with student databases but lack SQL expertise. The system suggests an algorithm for handling the English Query fired by the user in order to obtain a SQL query using feedback as text or expression. Tokenization, syntactic, and semantic analysis, as well as the use of dictionaries and grammars used for such analysis, can be performed using the natural language to SQL query conversion tool.

A query can be entered in natural language by the user. When the user enters the query in English, it is translated to a SQL query. There are many difficulties in converting natural language queries to SQL queries, such as complexity, which implies that a single term may have several meanings. In this case, a single word may be mapped to several meanings. Another difficulty is the development of complex SQL queries.

**Chapter 2**

**Literature Survey**

We conducted research by reading the research papers and publications on earlier work.

* 1. **Survey of Existing Systems:**

1. **Formation of SQL from Natural Language Query using NLP**

This approach uses natural language processing (NLP), with a structured natural language question as input and a SQL query as output, to quickly access information from the railway reservation database. Tokenization, lemmatization, parts of speech tagging, parsing, and mapping are all stages in this process. There are 2880 formal natural language queries on train fares and available seats in the dataset used by the proposed scheme. They reached a precision of 98.89 percent.

**Methodology**:

They have implemented the method in two stages:

1. NLP Phase

-Tokenization  
-Lemmatization  
-Syntactical Analysis  
-Semantic Analysis

1. Mapping Phase

- Attribute Identification  
-SQL Query Formation

**Objective:**

Formation of SQL query using Natural Language Processing on train seats and fare datasets.

**Limitations:**

They have used just one table and have taken input in text format.

1. **Conversion of Natural Language Query to SQL Query**

This paper describes a method for converting Natural Language Query to SQL Query that is automated. The user can use speech to enter the question in this system. Speech would be converted to text by the system. This query is going to be converted to a SQL query. The system will run the question and return the results to the customer.

**Methodology:**

They used two dictionaries for their analysis: one for lexical analysis and the other for syntax or semantic analysis. Tokenization, Lexical Analysis, Syntactic Analysis, and Semantic Analysis are all part of the process.

**Objective:**

To make it easier for users to retrieve and handle student data from a database.

**Limitations:**

The findings of this method are not quite accurate.

1. **MyNLIDB : A Natural Language Interface to Database**

In this article, they present a method called MyNLIDB that performs well in terms of keyword mapping. They used a Schema-Graph constructed from the underlying database, a Stanford part-of-speech parser, and a dependency parser to convert NL Query to SQL using pipeline analysis. MyNLIDB is an information management system that is domain and database independent. They were able to achieve a high level of precision with this method.

**Methodology:**

Input🡪Pre-Processing🡪POStagger🡪Node generator🡪SQL generator🡪Database🡪Result

**Objective:**

To build a database with a natural language interface. (i.e., the interface used to access data from a database).

**Limitation:**

It is made only for simple queries

1. **A Question Answering on Structured Data using NLIDB Approach**

In this paper, they used an intermediate query approach to implement a Natural Language Interface to Database (NLIDB) scheme. Their process is shown with a chatbot for the Movie domain, but it can also be extended to other domains. The architecture has performed admirably and can manage the vast majority of user requests pertaining to the database in question.

**Methodology:**

User🡪NL Query (English)🡪Tokenizer🡪POS Tagger🡪

Dependency Parsing🡪Build SQL query🡪Execute Query From database🡪Result

**Objective:**

Develop NLID for alumni database using CFG base system.

**Limitations:**

The accuracy is very poor, at just 47

* 1. **Limitations of existing systems and research gaps**

In the all system discussed in research papers, they have used simple database with only one table which can solve simple queries. Also, every person can speak different wordings but having same meaning which is tedious to get accurate query.

* 1. **Objectives**

Objective of this project is to generate SQL query from the Natural Language using Natural Language Processing(NLP).

* 1. **Literature Survey Table:**

Table 2.1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Author | Title | Methodology | Objective | Limitation |
| M. Uma, V. Sneha, G. Sneha,  J. Bhuvana and B. Bharathi, 2019 | Formation of SQL from Natural Language Query using NLP | NLP Phase: -Tokenization -Lemmatization -Syntactical Analysis -Semantic Analysis Mapping Phase: - Attribute Identification -Sql Query Formation | Formation of SQL query using Natural Language Processing on train seats and fare datasets. | They have taken input in text form and they have make use of only one table. |
| Alaka Das, Rakesh Chandra Balabantaray, 2019 | MyNLIDB: A Natural Language Interface to Database | Input🡪Preprocessing 🡪POStagger🡪 🡪Node generator🡪 SQL generator🡪 Database🡪 Result | To make Natural Language Interface to Database.(i.e. Interface through which we retrieve data from database) | It is made only for simple queries |
| A. kate, S. Kamble, A. Bodhke, M. Joshi, 2018 | Conversion of Natural Language Query to SQL Query. | Tokenization , Lexical Analysis , Syntactic Analysis , Semantic Analysis. | To help T&P officer to easily retrieve and manage student data from database. | This system does not provide high accuracy in results. |
| Tanzim Mahmud, K. M. Azharul Hasan,  Mahtab Ahmed,  Thwoi Hla Ching Chak, 2015 | A Rule Based Approach for NLP Based Query Processing | Design context free grammars  word check,removing excess words, tokenization and mapping to CFG rules. | Develop NLID for alumni database using CFG base system | Accuracy is very low i.e 47% |
| N. Sangeeth ,  R. Rejimoan , 2015 | An Intelligent System For Information Extraction From Relational Database Using HMM | Two modules used: Linguistic module, database module. | develop NLIDB system based on HMM using GEOQUERY database. | Accuracy and Performance can be increased, audio input is not taken |
| Vishal Wudaru, Aruneswara Reddy, Radhika Mamid,  2019 | Question Answering on Structured Data using NLIDB Approach | Intermediate query approach Syntactic approach 1.Tokenizer & POS Tagging 2.Stop word remover 3.Dependency parsing. | Proposed system able to handle most queries related to movie domain through database | Focused on limited number of database attributes |

**Chapter 3**

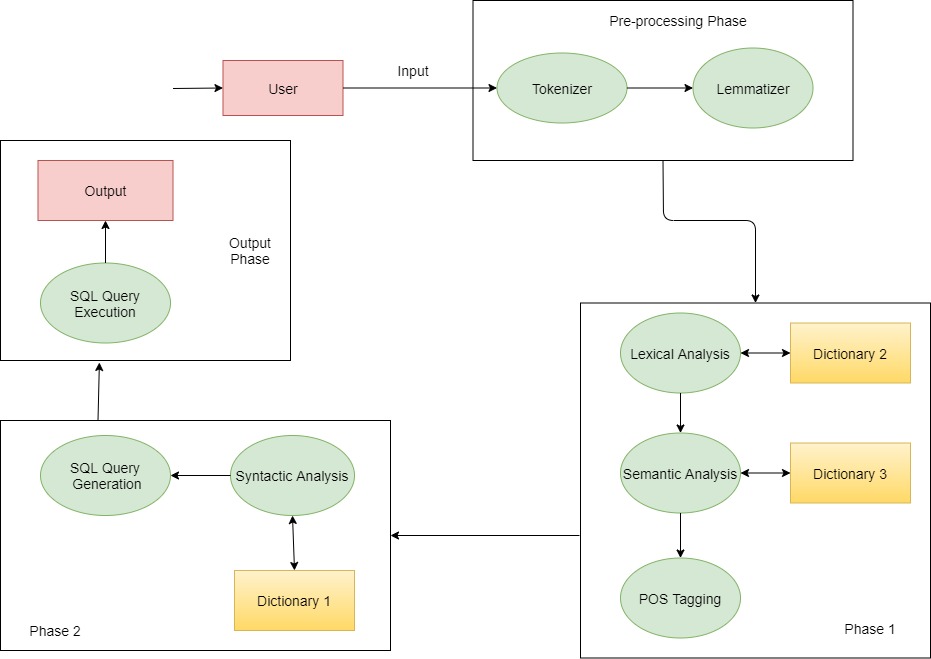
**Problem Statement**

To create a system which converts Natural Language To SQL query by taking input as text format, process it and generate SQL query and execute it to provide output.

**Chapter 4**

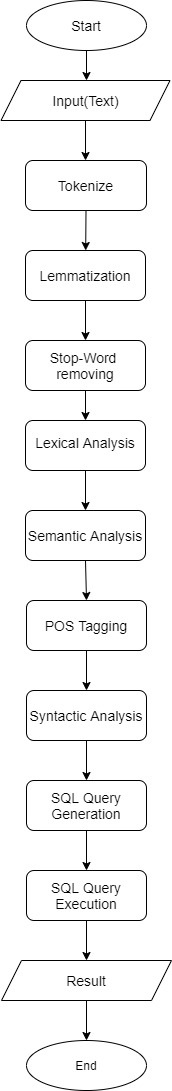
**Project Design**

* **System Block Diagram**

****

*Fig.4.1*

* **Flow Chart**

****

*Fig.4.2*

* **Algorithm**

**Step 1:** Query entered by the user (in text) in Natural Language is stored in a string.

Example: fetch all the information of the students who have scored more than 70 in 10th.

**Step 2:** Considering the spaces in a sentence, it is split into

individual words (tokens). These tokens are stored in a separate list.

These are the tokens for above query.

['fetch', 'all', 'the', 'information', 'of', 'the', 'students', 'who', 'have', 'scored', 'more', 'than', '70', 'in', '10th’]

**Step 3:** We will use lemmatization to get the lemma of the tokens generated.

['fetch', 'all',’the’, 'information', 'of', 'the', 'student', 'who', 'have', 'scored', 'more', 'than', '70', 'in', '10th’]

**Step 4:** Tokens are compared with the ignore list, only important

tokens are considered for further processing.

['fetch', 'all', 'student', 'who', 'have', 'more', '70', '10th’]

**Step 5:** Selected tokens are mapped with database words. These

tokens gets replaced by their synonyms.

['SELECT', '\*', 'FROM student', 'WHERE', 'more', '70', 'ssc’]

**Step 6:** Identify the conditions or values specified by the user in

his/her query, if any. Those conditions are mapped with the

database words. (For example, Less than is replaced with " < "

symbol)

['SELECT', '\*', 'FROM student', 'WHERE', '>', '70', 'ssc’]

**Step 7:** Parts of Speech tagging of tokens is done here. The tag in case of is a part-of-speech tag, and signifies whether the word is a noun, adjective, verb, and so on.

[('SELECT', 'NNP'), ('\*', 'NNP'), ('FROM student', 'NNP'), ('WHERE', 'NNP'), ('>', 'VBZ'), ('70', 'CD'), ('ssc', 'NN’)]

**Step 8:** Syntax analysis checks the text for meaningfulness comparing to the rules of formal grammar.

['SELECT'] ['\*'] ['FROM student'] ['ssc'] ['>'] ['70'] []

**Step 9:** SQL Query generation

SELECT \* FROM student WHERE ssc > 70

**Step 10:** SQL Query will be executed on the database and user will be provided with output.

**Chapter 5**

**Implementation Details**

**5.1 Module Description**

The system proposed in this paper is explained in following modules as Tokenize module, lemmatize and stop-word module, Lexical module, Semantic module, POS\_tagging module, Syntactic module, SQL query generation module.

**Tokenize module**

System will perform tokenization on the entered query by separating it into single words. Each word represents a token. Then these words will be stored in a separate list and passed to Lemmatized and Stop-Word module.

**Lemmatized and Stop-Word module**

System will perform lemmatization on the output of the tokenized module and also stop word module removes all the unwanted words from the list using ignore\_list. Then this will stored in separate list and pass to the lexical module.

**Lexical module**

The lemmatized list will be mapped with the dictionary. These words will get replaced by the database words from the dictionary and passed to syntactic analysis.

**Semantic module**

System will find words which represent conditions or symbols and that word will get mapped with the dictionary. (For Example: If there is "less than or equal to" in the query, it will get mapped with the symbol "<=").

**POS\_tagging module**

Parts of Speech tagging of tokens is done here. The tag in case of is a part-of-speech tag, and signifies whether the word is a noun, adjective, verb, and so on.

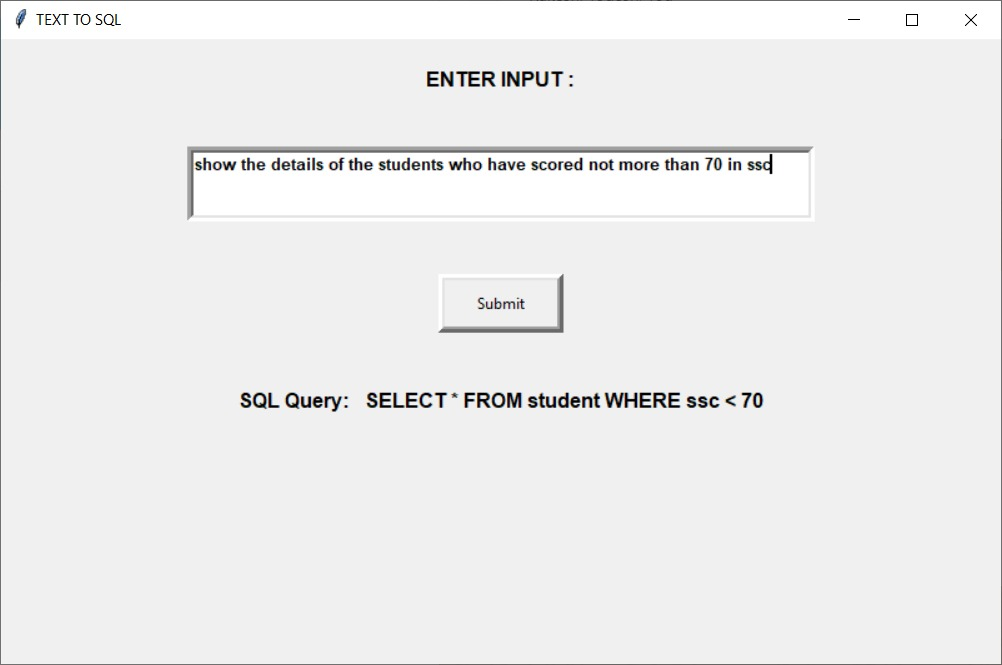
**Syntactic module**

In this step dictionary of table names, attributes and keywords are maintained. Each tokenized word gets mapped with attributes in the dictionary.

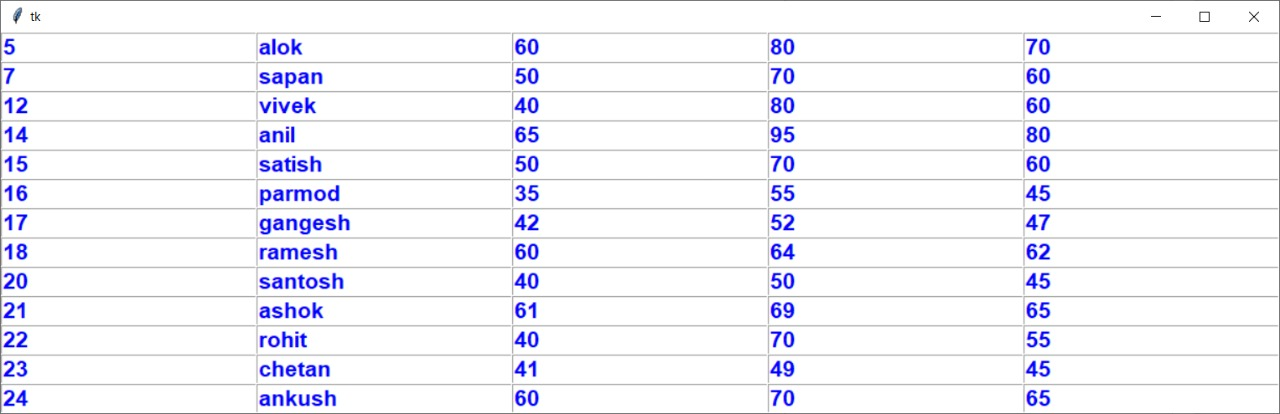
**SQL query generation module**

In this module, SQL query will be generated using the output of the syntactic module.

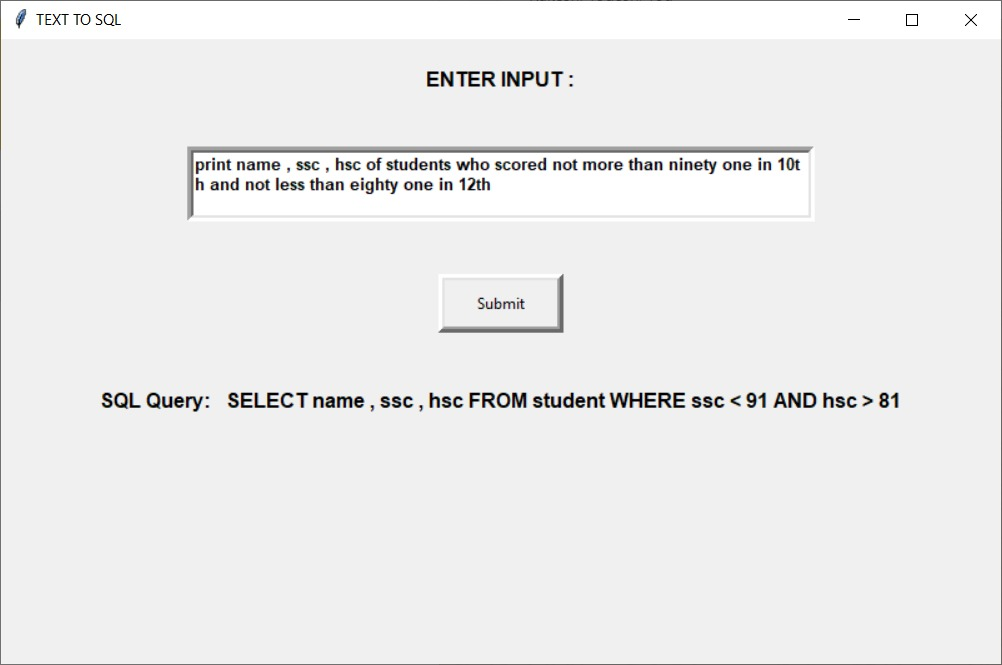
**5.2 Snapshot**



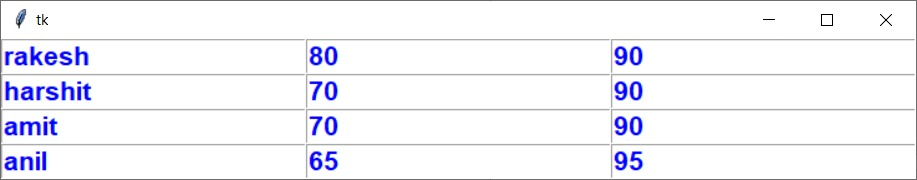
*Figure.5.1*



*Figure.5.2*



*Figure.5.3*



*Figure.5.4*

**Chapter 6**

**Results and Analysis**

This system converts Natural Language Query to SQL Query efficiently and effectively. Tokenization, syntactic, and semantic analysis, as well as the use of dictionaries and grammar is used for such analysis is performed using the natural language to SQL query conversion tool.

Libraries used – MySQL, Tkinter, word\_tokenize, wordNetLemmatizer, pos\_tag by nltk. System has achieved accurate results for many natural language queries.

**Chapter 7**

**Conclusion and Future Scope**

**Conclusion:**

Any human being can benefit from using Natural Language. Using plain English, this system will assist users in retrieving and managing data from the student database. The user does not need to learn SQL or any other complicated query language. The system is user-friendly due to its ability to receive input in text format. Our system will convert natural language queries into SQL language queries and supply users with the necessary information from the student database.

**Future Scope:**

The following are some potential enhancements that might be implemented: The input can be in the form of audio, which may then be transformed to text. The SQL query might be more complicated. In terms of attributes and tuples, the database might be bigger. There may also be many tables with relevant data that may be accessed with the JOIN keyword. To make the system more interactive, the output may be transformed into a sentence and then into audio format. This work could also be expanded to other languages.

**References**

[1] A. kate, S. Kamble, A. Bodhke, M. Joshi, “Conversion of Natural Language Query to SQL Query,” 2018 IEEE Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India,2018.

[2] M. Uma, V. Sneha, G. Sneha, J. Bhuvana and B. Bharathi, "Formation of SQL from Natural Language Query using NLP," 2019 IEEE International Conference on Computational Intelligence in Data Science (ICCIDS), Chennai, India, 2019.

[3] Alaka Das, Rakesh Chandra Balabantaray, “MyNLIDB: A Natural Language Interface to Database”, 2019 IEEE International Conference on Information Technology (ICIT), Bhubaneswar, India,2019.

[4] Tanzim Mahmud, K. M. Azharul Hasan, Mahtab Ahmed, Thwoi Hla Ching Chak, “A Rule Based Approach for NLP Based Query Processing ,“ 2015 Khulna University of Engineering & Technology (KUET) Khulna – 9203, Bangladesh,2015 IEEE.

[5] Vishal Wudaru, Nikhil Koditala, Aruneswara Reddy, Radhika Mamid, “Question Answering on Structured Data using NLIDB Approach”, 2018 International Institute of Information Technology, Hyderabad, India, 2018 IEEE.

[6] <https://www.geeksforgeeks.org/introduction-to-natural-language-processing/>

[7] https://www.ibm.com/cloud/learn/natural-language-processing

**Acknowledgement**

We wish to express our profound gratitude to our principal Dr. Bhavesh Patel for allowing us to go ahead with this project and giving us this opportunity to explore this domain. We would also like to thank our head of department Prof. Uday Bhave for his constant encouragement and support towards achieving this goal.

We would like to thank Review Committee for their invaluable suggestion and feedback without which our work would have been very difficult.

We take this opportunity to express our profound gratitude and deep regards to our guide Ms. Deepshikha Chaturvedi for this exemplary guidance, monitoring and constant encouragement throughout the course of this project. We owe special acknowledgement to her for giving us a lot of her time during the period of preparing this project. We could never have done it without her support technical advice and constant suggestion to improve our work. No project is ever complete without the guidelines of these experts who have already established milestones on this path before and have master of it. So, we would like to this opportunity to thank all those who have helped us in implement this project.

Rakesh Choudhary: \_\_\_\_\_\_\_\_\_\_\_

Satyam Gupta: \_\_\_\_\_\_\_\_\_\_\_

Yogesh Hole: \_\_\_\_\_\_\_\_\_\_\_

Date: 24/05/2021