Department of Computer Science and Engineering
Vimal Jyothi Engineering College
Chemperi

NLP-Powered Search Engine for the University Website

MEMBERS:

ALBERT TOM GEORGE(VML20CS025)
NAVEEN K MATHEW(VML20CS126)
SIDHARTH KESAV(VML20CS158)
SIDHARTH SHAM LAL(VML20CS161)

GUIDE:

Ms. Nisha P V

OUTLINE

AREA OF SELECTION

Machine Learning

Machine learning is a branch of artificial intelligence that teaches computers to learn from data and make decisions or predictions without being explicitly programmed for each task. It involves algorithms that improve themselves over time by identifying patterns in data, leading to applications in various fields such as image recognition, language understanding, and personalized recommendations.

ABSTRACT

- A specialized search engine for university websites, using advanced NLP for accurate and contextual results.
- Employs NLP to interpret complex academic queries, moving beyond traditional keyword searches.
- Utilizes semantic analysis to match user intent with relevant university website resources.
- Enhances user experience by delivering precise and contextually relevant search outcomes.
- Aims to streamline academic information access within the university community.

INTRODUCTION

- This project aims to optimize the accessibility and usability of university websites for improved information retrieval.
- Through natural language processing (NLP), a specialized search engine is being developed to handle intricate academic queries efficiently.
- Moving past basic keyword searches, the engine employs semantic analysis for precise and contextually relevant search results.
- By aligning user intent with available resources, this initiative focuses on user needs and seamless navigation on academic websites.
- The implementation of this NLP-driven search engine seeks to empower the academic community by facilitating quicker and more accurate information retrieval.

PROBLEM DEFINITION

- University websites struggle to provide efficient access to accurate information, hindering user experience.
- Implementing an advanced, user-centric search engine powered by natural language processing (NLP) is essential to address these information retrieval challenges within the academic community.

SCOPE OF THE SYSTEM

The NLP-powered search engine for university websites aims to revolutionize user experience by efficiently processing complex academic queries, providing precise results beyond traditional keyword searches. Using advanced natural language processing, it aligns user intent with relevant resources, ultimately streamlining information retrieval for the academic community.

OBJECTIVE

- The primary objective is to create a more user-friendly environment by improving the navigation and accessibility of information on university websites.
- To accurately handle complex academic queries, the system aims to utilize NLP techniques, delivering more relevant and accurate search results compared to traditional keyword-based searches.
- By employing advanced semantic analysis, the objective is to ensure the alignment of user intent with the most appropriate resources available on the university website.

LITERATURE SURVEY

The paper selected for the literature survey are:

- Search-Based Algorithm With Scatter Search Strategy for Automated Test Case Generation of NLP Toolkit
- Developing a Meta-Suggestion Engine for Search Queries
- A Unified Understanding of Deep NLP Models for Text Classification
- A Survey of Text Representation and Embedding Techniques in NLP

Search-Based Algorithm With Scatter Search Strategy for Automated Test Case Generation of NLP Toolkit

BRIEF SUMMARY

- NLP programs have numerous paths requiring specific input, making conventional algorithms struggle to cover all paths efficiently.
- The paper introduces a scatter search strategy to enhance test case generation in NLP programs, effectively covering specific-input-dependent paths.
- Implementing the scatter search strategy significantly reduces test cases and testing time, improving the efficiency of NLP program testing.

METHODOLOGIES

- Understanding path coverage challenges in NLP due to specific input requirements.
- Creating the scatter search method to efficiently explore input variables.
- Testing the strategy on NLP programs, measuring reduced test cases and time.
- Assessing strategy impact on covering specific-input-dependent paths.

Search-Based Algorithm With Scatter Search Strategy for Automated Test Case Generation of NLP Toolkit

ADVANTAGES

- Saves time and resources while ensuring all possible paths of NLP programs are covered.
- Helps reveal logic defects in NLP programs.
- Achieves better performance compared to other state-of-the-art algorithms.
- Reduces data silos and promotes collaboration.

DISADVANTAGES

- May not be suitable for all types of NLP programs.
- May require a large amount of computational resources.
- May generate redundant test cases.

BRIEF SUMMARY

- The paper proposes a new Meta-Suggestion Engine for improving query suggestions in search engines.
- The engine addresses the limitations of log-based suggestions and can be applied globally.
- It also discusses potential applications and future directions of the technology.

METHODOLOGY

- Meta-search to retrieve candidate queries from the target engine.
- Query ranking to rank the candidate queries based on their relevance to the initial query.
- Precision and recall to evaluate the effectiveness of the Meta-Suggestion Engine.

ADVANTAGES

- The proposed meta-suggestion engine provides more accurate and diverse query suggestions.
- The engine is versatile and can be applied to both global and domestic search engines.
- The engine is implemented as a browser extension, which allows for easy integration with existing search engines.

DISADVANTAGES

- The scalability of the proposed method is not discussed.
- The evaluation of the proposed method is limited to a specific set of queries.
- The potential impact of the proposed method on user privacy is not discussed.

BRIEF SUMMARY

- The paper introduces a visual analysis tool, DeepNLPVis, that provides a unified measure for explaining both low-level and high-level features of NLP models.
- DeepNLPVis enables users to smoothly navigate from the overall performance at the corpus-level to the detailed information at the sample- and word-level..
- The tool helps users identify potential problems caused by samples and model architectures, and provides enlightening information and deeper understanding of the models.

METHODOLOGY

- Attention Flows.
- Unified Method for Understanding NLP Models.
- DeepNLPVis.

ADVANTAGES

- Informative Visualization.
- Effective Communication.
- Efficient Analysis Process.
- Model Comparison.

DISADVANTAGES

- Limitations in Task Generalization.
- Limited Support for Other NLP Tasks.
- Learning Curve.

BRIEF SUMMARY

- Words as numerical vectors enable computation for NLP tasks using matrices.
- PLSI uses latent context variables to minimize word perplexity; LSI reduces dimensionality via SVD to handle word-related challenges.
- PLSI outperforms LSI in reducing perplexity by addressing polysemy explicitly.
- Both methods apply to NLP tasks like indexing, classification, and document similarity.

METHODOLOGY

- Hyperspace Analogue to Language
- Dimensionality Reduction Techniques
- ELMo (Embeddings from Language Models)
- ULMFiT
- Density Distribution Embedding
- Explicit Semantic Analysis (ESA)

ADVANTAGES

- Improved Performance.
- Pre-trained Models.
- Transfer Learning.
- Density Distribution Embedding.

DISADVANTAGES

- Curse of Dimensionality.
- Bias.
- Varying Performance and costs.
- Computational Power Limitations

COMPARISION TABLE

Paper 1:Search-Based Algorithm With Scatter Search Strategyfor Automated Test Case Generation of NLP Toolkit	Paper 2 Developing a Meta-Suggestion Enginefor Search Queries	Paper 3:A Unified Understanding of Deep NLPModels for Text Classification	Paper 4:A Survey of Text Representation and EmbeddingTechniques in NLP
Scatter search strategy streamlines NLP testing by efficiently generating specific-input-dependent test cases, reducing overall testing time	Global Meta-Suggestion Engine to overcome log-based limitations, offering enhanced query suggestions and exploring future applications	DeepNLPVis: a unified visual tool offering insights from corpus to word-level, aiding model evaluation and issue identification in NLP.	Word vectors aid NLP; PLSI reduces perplexity, excelling in polysemy; both assist in indexing, classification, and document similarity
Advantages:Enhances NLP coverage, detects flaws, outperforms algorithms, and fosters collaboration by minimizing data silos	Advantages:Introducing a versatile meta-suggestion engine via a browser extension, offering accurate, diverse query suggestions for global and domestic search engines.	Advantages:Facilitates informative visualization, effective communication, and an efficient analysis process for model comparison.	Advantages:Enhanced performance via pre-trained models, transfer learning, and density distribution embedding.
Disadvantages: Suitablity variation for NLP programs, high computational resource	Disadvantages:Limited discussion on scalability, evaluation based on specific queries, and	Disadvantages:Challenges in task generalization, limited	Disadvantages: dimensionality curse, bias, performance

REQUIREMENT SPECIFICATION

Functional requirements

1. Security
2. Integration
3. Storage capability

Hardware interfaces: There are no external hardware interface requirements for this system. System Hardware requirements:

1. CPU: 3+ Cores, 2.38+ Ghz
2. RAM: 4 GB or higher
3. Disk: 20 GB + free space

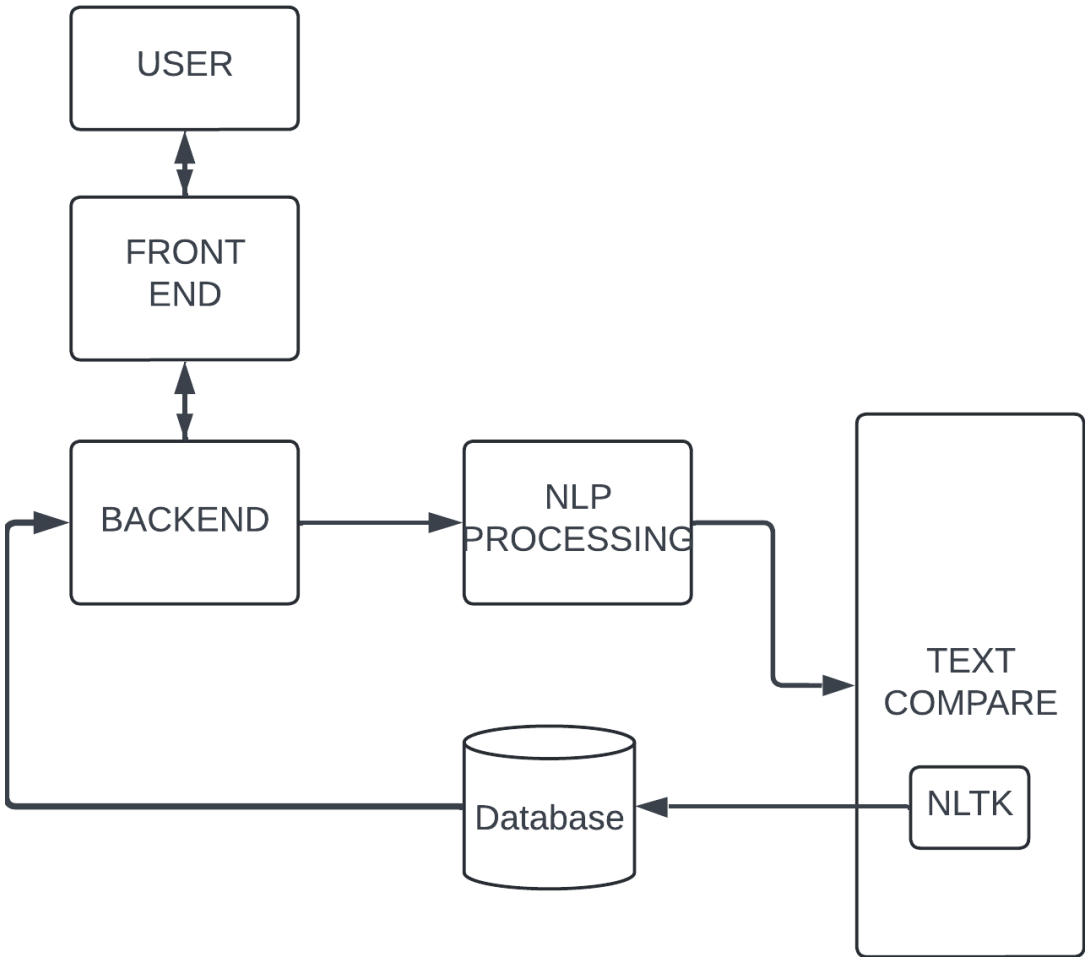
PROPOSED SYSTEM

Implementing an NLP-powered search engine for a university site is technically feasible due to available NLP tools like BERT, GPT, NLTK, and spaCy, abundant textual data, scalable cloud infrastructure, and accessible NLP model APIs, despite challenges like model accuracy and infrastructure integration.

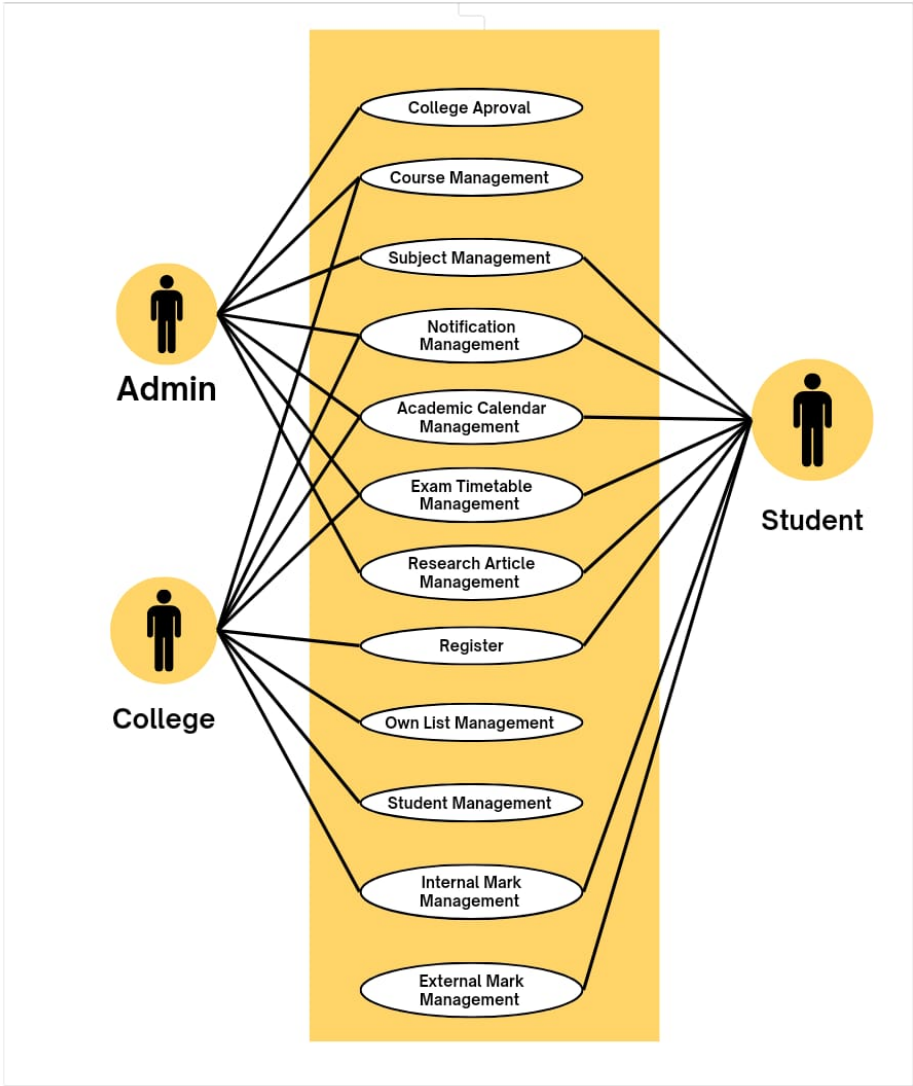
FEASIBILITY STUDY

- Technical Feasibility
  The technical feasibility of the NLP-based search engine seems strong, utilizing advanced language processing to understand complex academic queries and align user intent with relevant university resources. Challenges may arise in implementing and scaling NLP effectively for large data volumes. However, the project shows promise with its foundational setup.
- Economical Feasibility
  Since the project uses open source tools, the cost for development can be eliminated. Hence this project is economically feasible.
- Operational Feasibility
  Since there's no operational cost, the project is operationally feasible.

ARCHITECTURE DIAGRAM

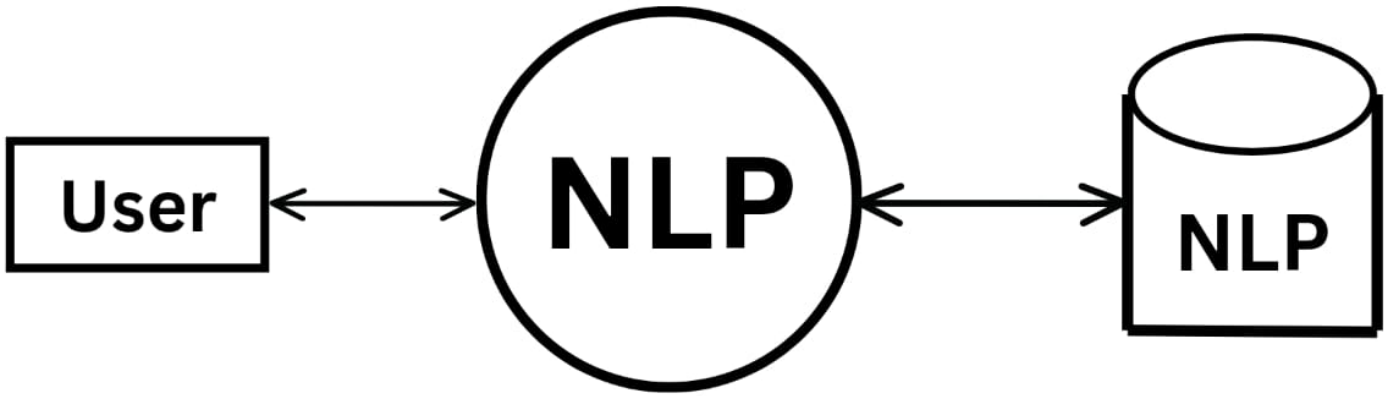


USECASE DIAGRAM



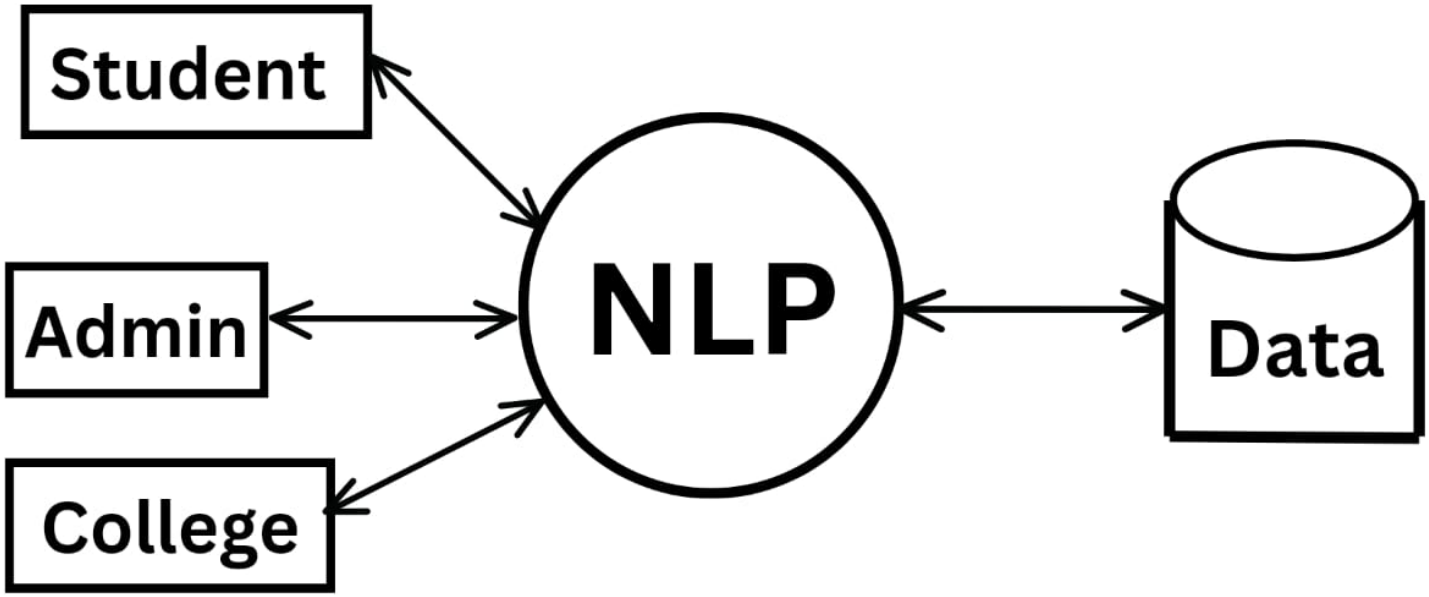
DATAFLOW DIAGRAM

- DFD Level 0



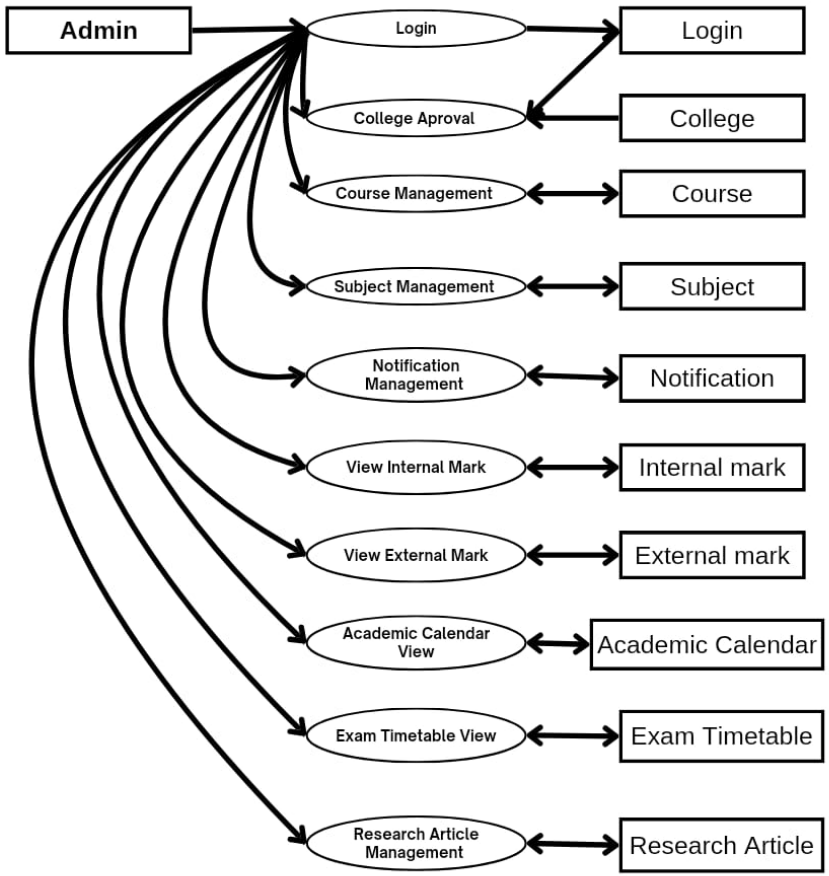
DATAFLOW DIAGRAM

- DFD Level 1



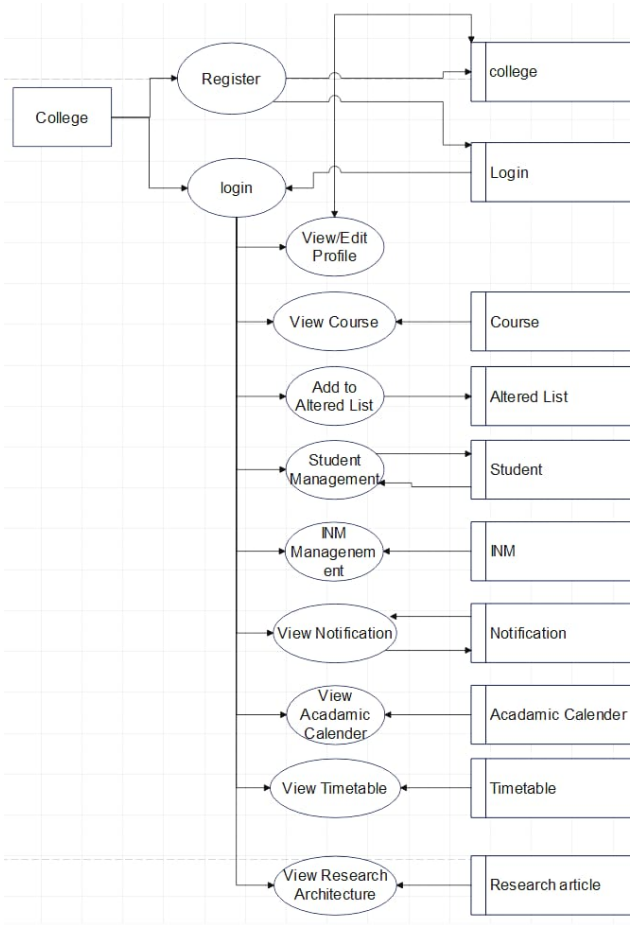
DATAFLOW DIAGRAM

- DFD Level 1.1



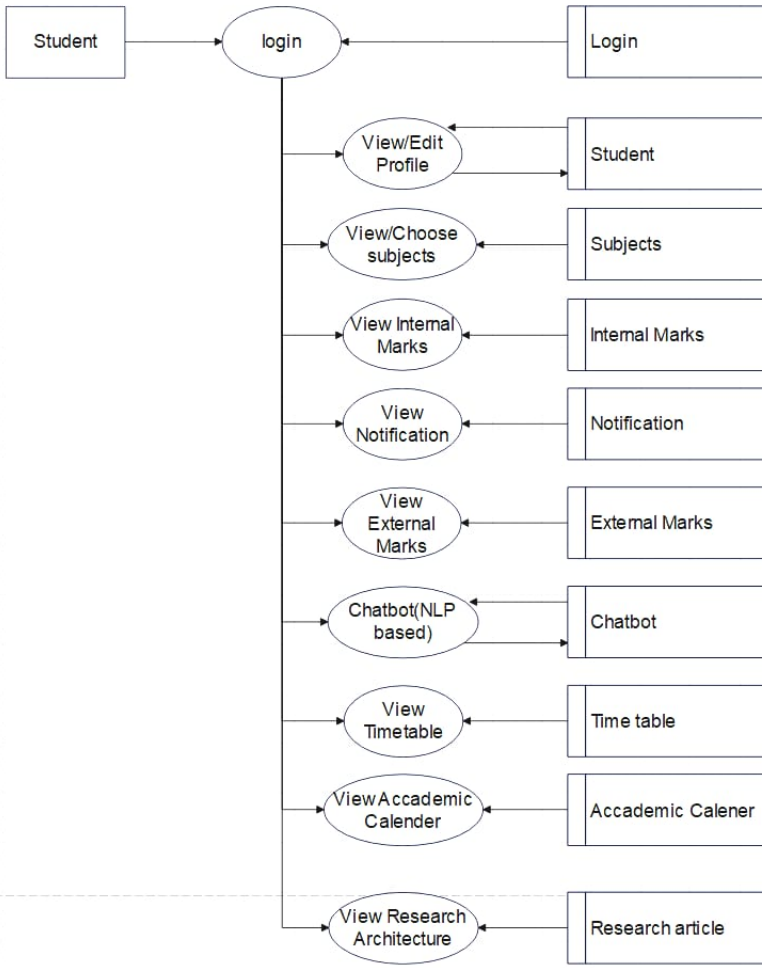
DATAFLOW DIAGRAM

- DFD Level 1.2

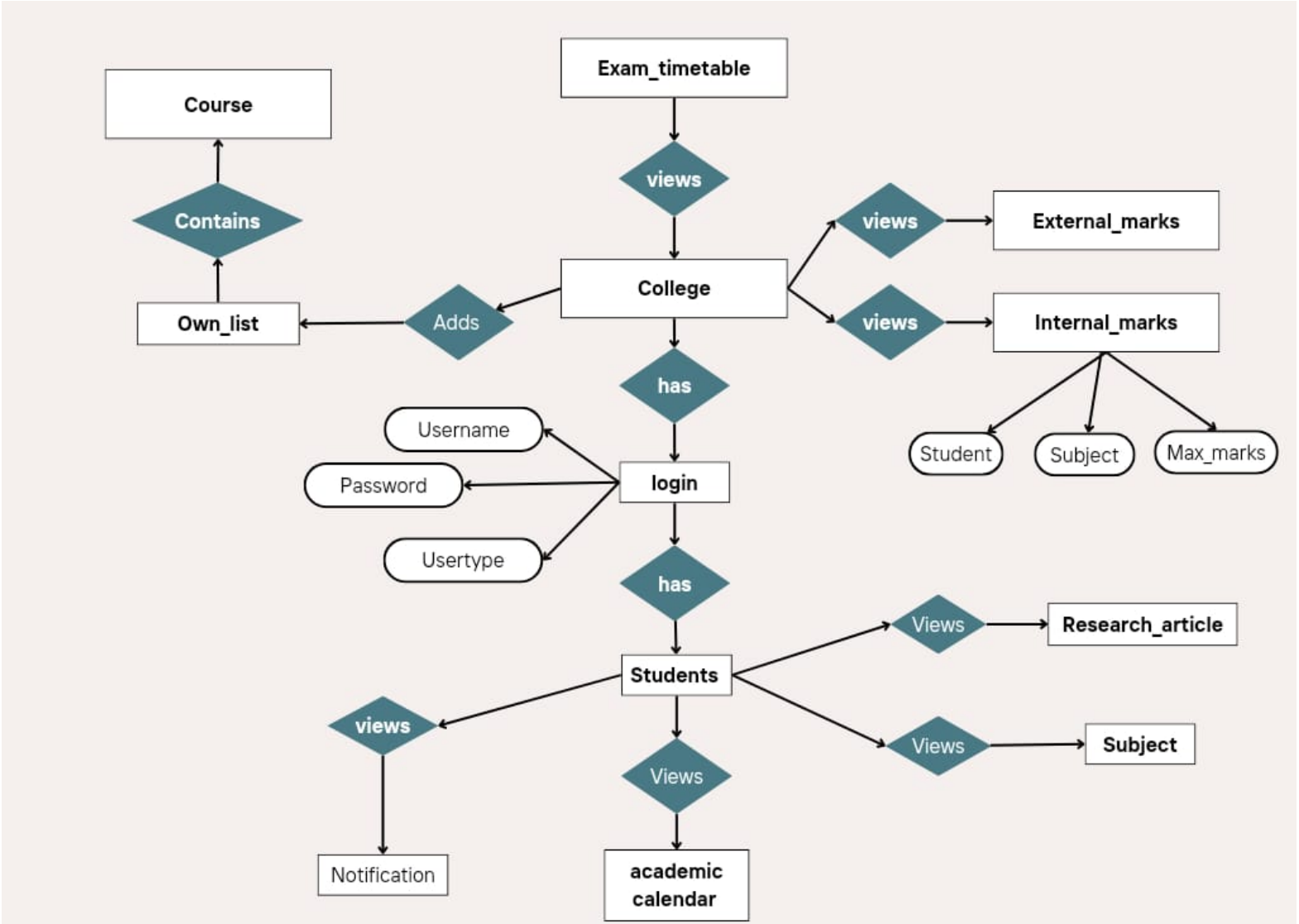


DATAFLOW DIAGRAM

- DFD Level 1.3



ER DIAGRAM


Exam_timetable
Course
views
Contains
views
External_marks
College
Own_list
Adds
views
Internal_marks
has
Username
Password
login
Usertype
Student
Subject
Max_marks
has
Views
Research_article
Students
views
Views
Views
Subject
Notification
academic calendar

METHOD AND TECHNIQUES

- Tokenization and Text Preprocessing
- Semantic Analysis and Understanding
- Named Entity Recognition (NER)
- Query Expansion and Correction
  - Synonym Mapping
- Machine Learning Algorithms
  - Classification and Ranking:

PROGRESS IN PROJECT

- Data Acquisition and Preprocessing
  1. Objective: Collect, clean, and preprocess data from various sources within the database.
  2. Tasks:
     - Identify and gather diverse datasets (research article,notification,course description,academic calendar etc.).
     - Clean the data to remove noise, inconsistencies, and irrelevant information.
     - Preprocess text data by tokenizing, removing stop words, performing stemming/lemmatization, and structuring it for NLP analysis.

PROGRESS IN PROJECT

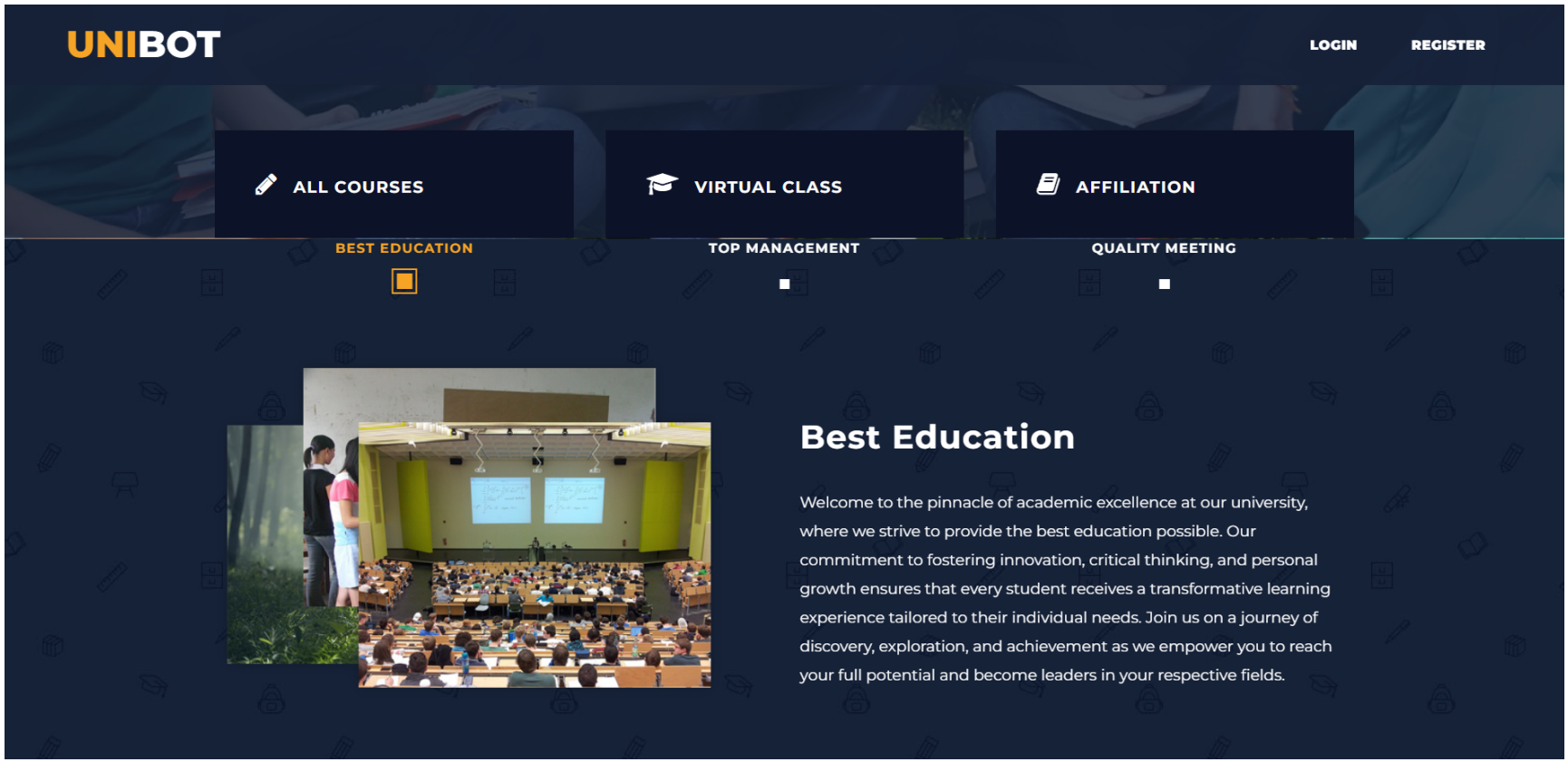
- User Interface and Experience Enhancement
  1. Objective: Create an intuitive and user-friendly interface for seamless interaction.
  2. Tasks:
     - Design and develop a user interface allowing natural language input and providing query suggestions.
     - Implement faceted search options and filters for refining search results.
     - Ensure responsiveness, accessibility, and ease of navigation in the interface.

PROGRESS IN PROJECT

- Deployment and Maintenance
  1. Objective: Deploy the developed search engine and ensure its ongoing maintenance and optimization.
  2. Tasks:
     - Monitor performance, address bugs, and make necessary updates or enhancements.
     - Develop documentation and training materials for users and administrators.

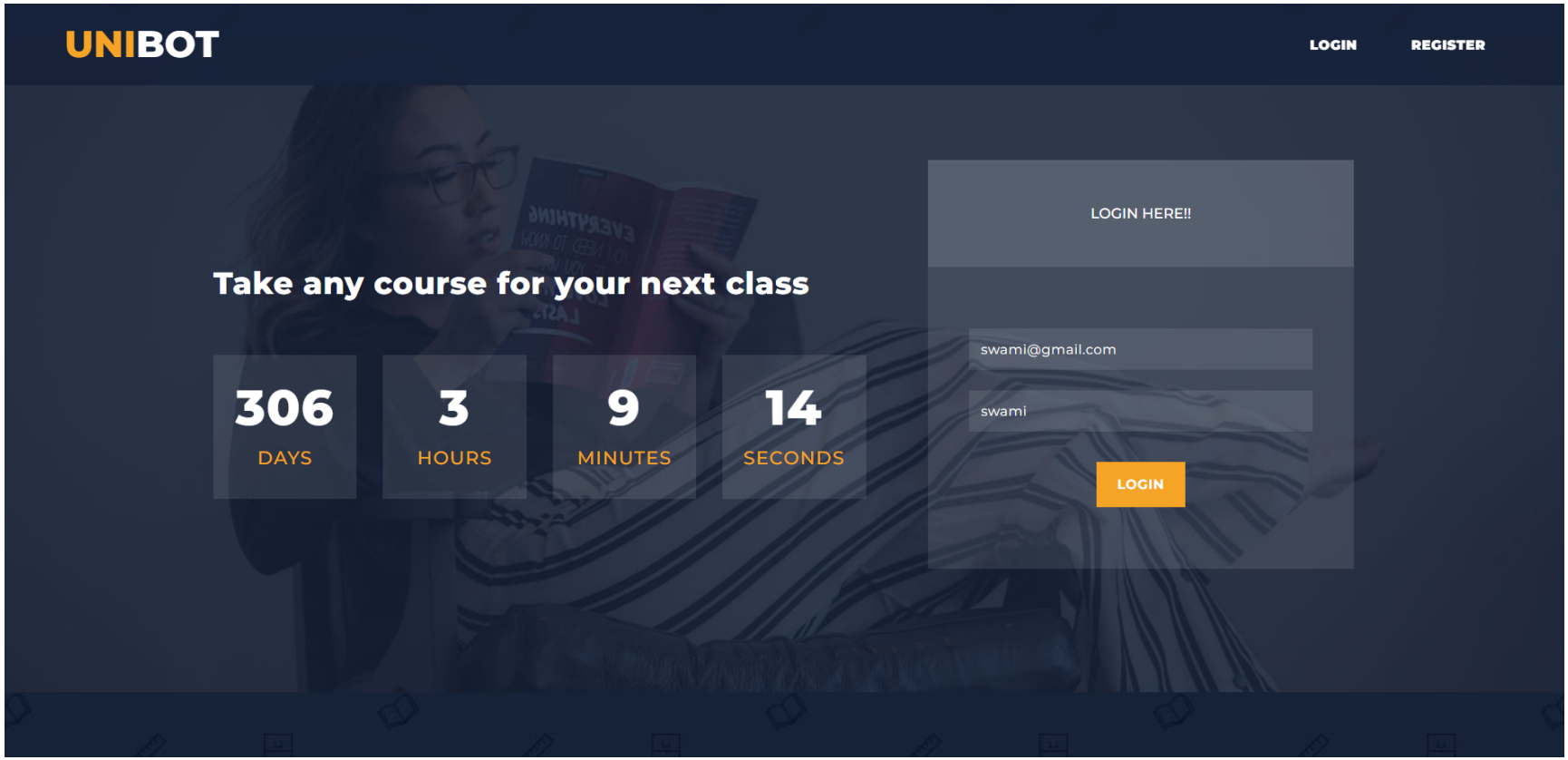
PROGRESS IN PROJECT

- Home Page



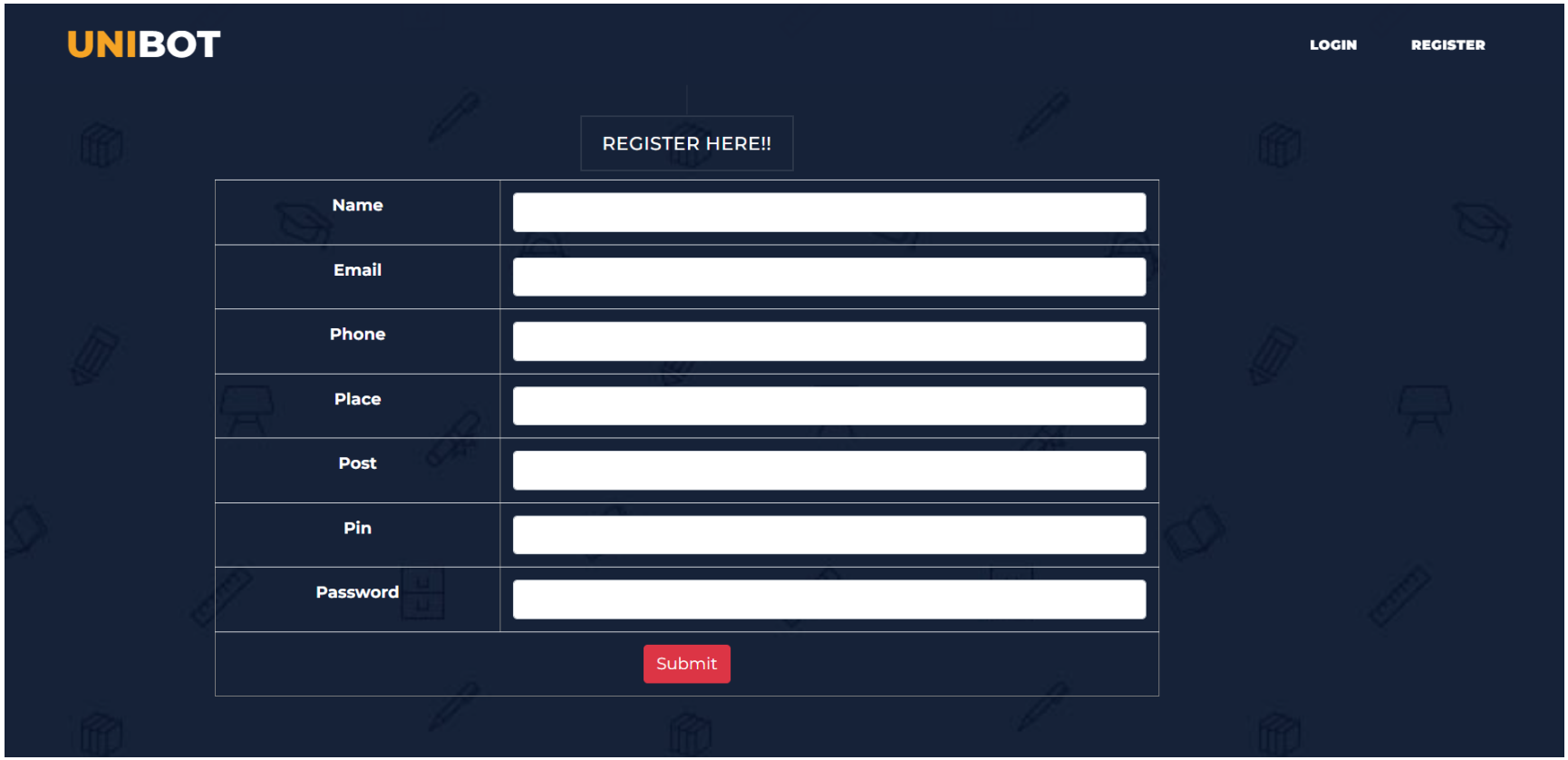
PROGRESS IN PROJECT

- Login Page



PROGRESS IN PROJECT

- College Registration

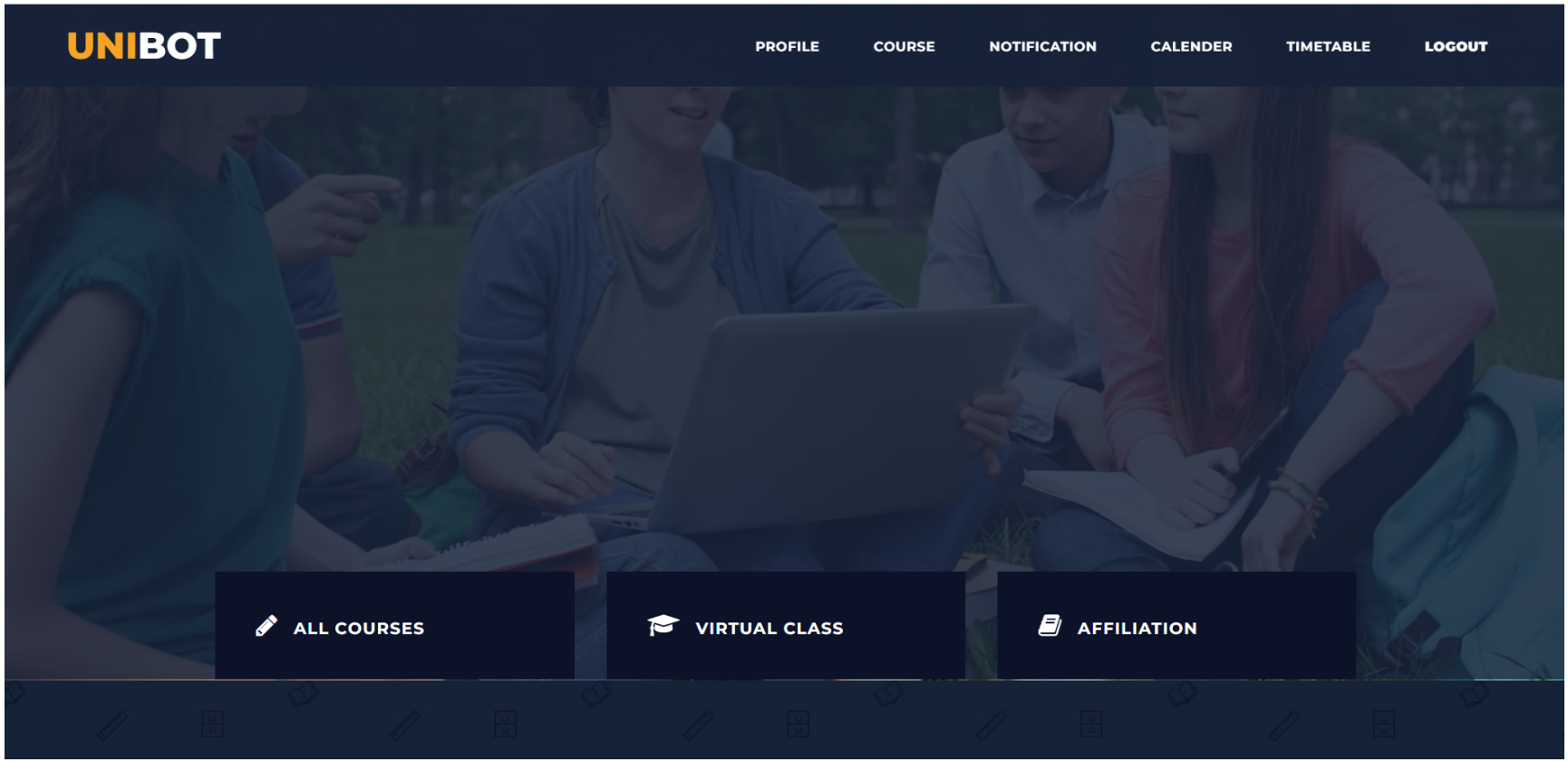


PROGRESS IN PROJECT

- Admin Login

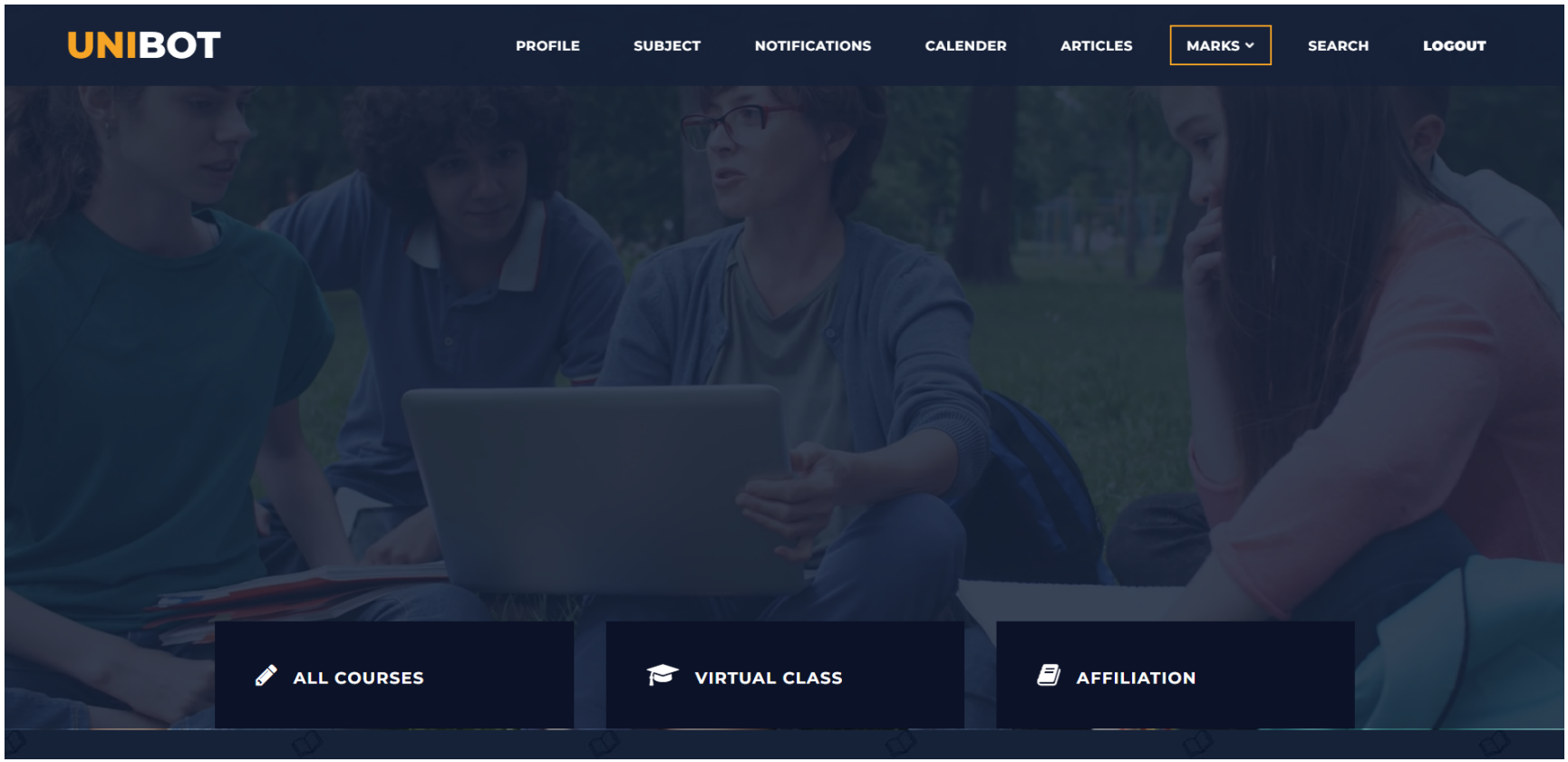
PROGRESS IN PROJECT

- College Login



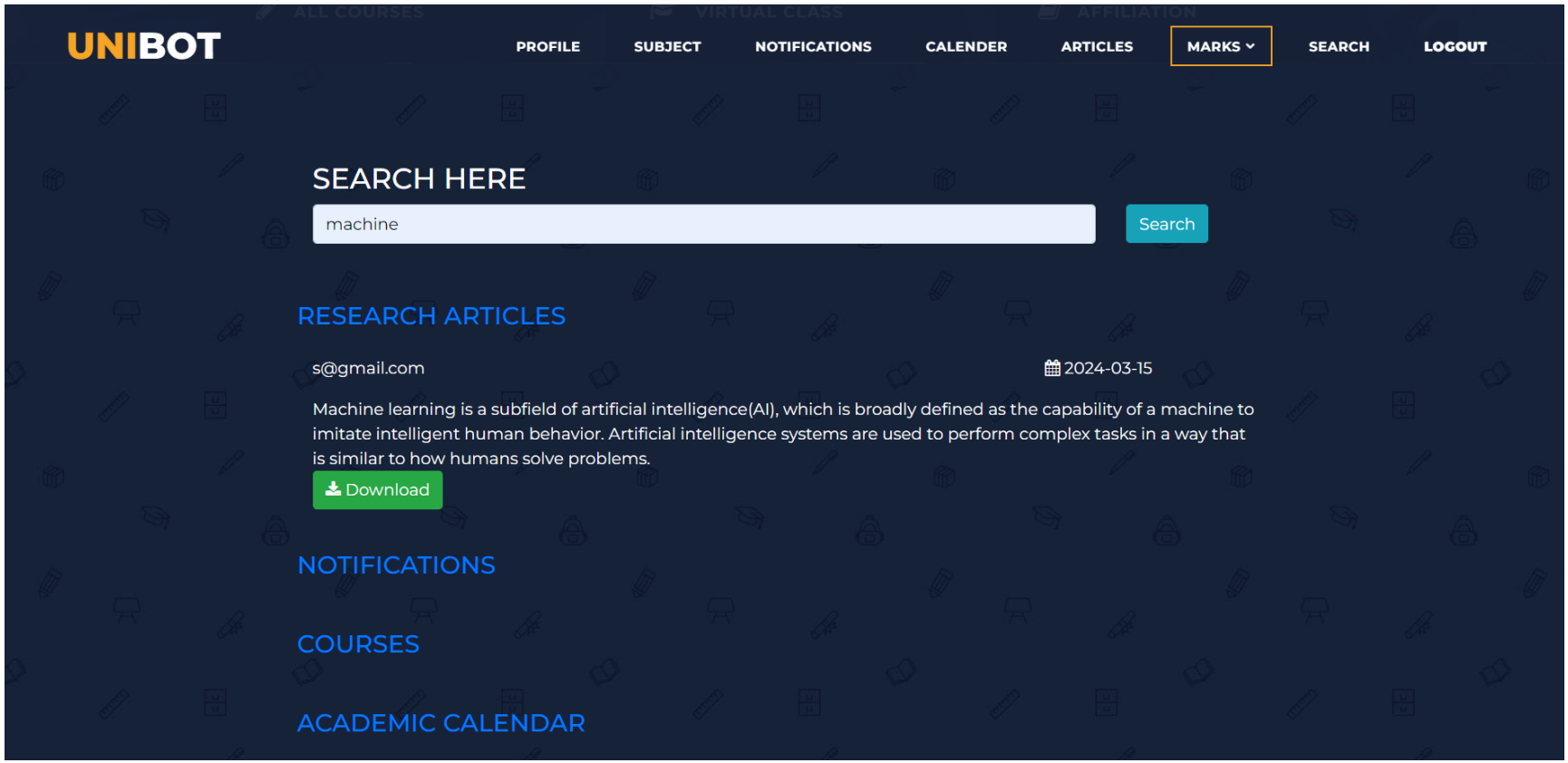
PROGRESS IN PROJECT

- Student Login



PROGRESS IN PROJECT

- Search Engine



GANTT CHART


AUGUST
SEPTEMBER
OCTOBER
NOVEMBER
DECEMBER
JANUARY
FEBRAUARY
Domain Selection
Problem Statement
Literature Review
Design Phase
20%
Implementation
Presentation
70%
Implementation

Paper Publication

ITC24CS114 | Abstract - SUBMISSION RECEIVED - submit full paper| ITECHCET 2024 Inbox

ITechCET 2... 4 days ago
to me

Dear Author **Sidharth Sham Lal**,

Greetings from Musaliar College of Engineering & Technology, Pathanamthitta, Kerala, India.

Hope this email finds you in good health and happiness!

We are pleased to inform you that your abstract of the paper entitled **"NLP Based University Search Engine"** with paper ID "ITC24CS114" has been received in the "International Conference on Research Advances in Engineering and Technology - ITechCET 2024" under Track 1.

To initiate the peer review process of the paper, the authors are requested to submit the full paper at the earliest as a reply to this mail.

Thank You
--
Dr Shan M Assis
Dr Ciby Jacob Cherian
Coordinator (s) | ITechCET 2024
Musaliar College of Engineering & Technology
Pathanamthitta, Kerala -689653
Ph: - +91 9539 328 242, +91 9496 330 997

CONCLUSION

- The NLP-powered search engine revolutionizes information retrieval on the university website, offering precise and relevant results beyond conventional search methods. Its user-centric design and adaptive nature align with user needs, providing an enhanced experience for navigating academic resources. This innovative solution ensures ongoing improvement, meeting the evolving demands of users seeking information on the university's website.

REFERENCE

[1] . Liu, H. Huang, Z. Yang, Z. Hao and J. Wang, "Search-Based Algorithm With Scatter Search Strategy for Automated Test Case Generation of NLP Toolkit," in IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 5, no. 3, pp. 491-503, June 2021, doi: 10.1109/TETCI.2019.2914280.

[2] . Kim, E. Na and S. B. Kim, "Developing a Meta-Suggestion Engine for Search Queries," in IEEE Access, vol. 10, pp. 68513-68520, 2022, doi: 10.1109/ACCESS.2022.3186096.

[3] . Li et al., "A Unified Understanding of Deep NLP Models for Text Classification," in IEEE Transactions on Visualization and Computer Graphics, vol. 28, no. 12, pp. 4980-4994, 1 Dec. 2022, doi: 10.1109/TVCG.2022.3184186.