**SYNOPSIS**

|  |  |  |  |
| --- | --- | --- | --- |
| **NAME** | **USN** | **EMAIL** | **MOBILE NUM** |
| RAKSHITH ARYA | 4NI20CS075 | rakshitharya08@gmail.com | 8088506208 |
| SHASHANK K R | 4NI20CS094 | krshashank1118@gmail.com | 8660969038 |
| SHASHANK SHANDILYA | 4NI20CS096 | shashankshandiliya2002@gmail.com | 6361377960 |
| VISHAL M V | 4NI20CS124 | vishalvydhyanath21@gmail.com | 8088345115 |

**TITLE OF THE PROJECT**: Knowledge extraction from documents using LLM

**DESCRIPTION:** Extracting necessary information from any document has always been a very slow and manual process. It involves a person to read through each sentence and then arrive at conclusions of what the sentence meant. This is not feasible when there a lot of documents to read through in a short time span. Our final year project “Knowledge extraction from documents using LLM” is an approach to solve this problem by developing a model that uses word embedding and train it using LLM’s.

**PROJECT OBJECTIVES:**

* Data Collection: Gather a set of legal documents that are based on a specific niche and have different patterns.
* Preprocessing: Implement data cleaning, text tokenization, and various NLP techniques to prepare the data to be further processed.
* Abstractive Prediction: Develop a model using word embedding techniques and popular LLM models to predict the category that the input sentence belongs to. These predictions will be based on previous input and their patterns. Each sentence is then matched to a category.
* Evaluation: Evaluate the quality of the predictions using human evaluations and pre-trained models to ensure the effectiveness of the predictions.

**EXPECTED OUTCOMES:**

* An innovative, accurate and high-quality prediction by the model that categorizes all the sentences in a given document.

**HARDWARE REQUIREMENTS:**

* Computer (Quad Core CPU, 8GB RAM, GTX 1050 TI).

**SOFTWARE REQUIREMENTS:**

* Python
* Ubuntu
* Python libraries( pandas, tensorflow, torch)
* Llama 2

Guide Signature with Date