**INFO 5731 - Computational Methods for Information Systems**

April 9, 2023

**Term Project First Deliverable**

**Project Title: Zero-shot learning for entity extraction using ChatGPT**

**Group six:** Sravika Reddy Danda, Mahitha Poduvu, Eshwitha Naini, Manisha Lokasani, Aliya Yasmin.

**Introduction:**

Zero-shot learning is the model which is not pretrained and is expected to generate an answer without additional information. Entity extraction is the technique that identifies key elements from the given data, then classifies them into predefined categories. ChatGPT can perform a variety of tasks of NLP including the parsing text, summarization, translation, classification, responding to questions given in the prompt.

Since entity extraction using ChatGPT is a useful tool particularly for those involving the Natural Language Processing and Text Analytics we planned to work on a project to develop the model with zero shot learning for entity extraction using ChatGPT. Here we have chosen healthcare dataset and wanted to extract the entity like symptoms, treatment for disease etc.

**Significance:** This model will recognize and classify all entities without prior training. Using ChatGPT, a powerful language model which is trained on large corpus text data will improve the accuracy of the model by leveraging its language understanding capabilities.

**Research questions for the project:**

* How is the performance of the ChatGPT on zero shot learning of entity extraction when compared with other traditional supervised learning approaches in terms of accuracy and efficiency.
* Is there any impact on performance of zero shot entity extraction in specific domains such as health, finances, sports and on size and diversity?
* Does the incorporation of additional data impact the performance of model.
* How to improve and develop the model performance?

Here we will use OpenAI python library to create the model. To improve the performance of model topic modelling techniques are used such as LDA, LSA, Bertopic.

**Objectives:**

1. The primary objective of the project is to develop a model for zero-shot entity extraction that can recognize and classify entities that it has not been explicitly trained on. The model would use descriptions of entities' attributes to extract the relevant information from text. The project could also aim to evaluate the model's performance using appropriate metrics and compare it with existing approaches.

2. Another objective of the project is to finetune the model and implement techniques such as data augmentation to improve its accuracy. Additionally, it could aim to deploy the model as a web service or integrate it into an existing application. Finally, it could also include monitoring the model's performance over time and providing regular updates to improve its accuracy.

3. Overall, the objectives of the project would be to develop a robust and accurate zero-shot entity extraction model using Chat GPT, contribute to the field of NLP and machine learning research, and potentially have real-world applications in various industries.

## **Methodology:**

1. Define the model:

Collect the healthcare dataset containing relevant reports, documents, or web pages from which entities can be extracted. The dataset should include a variety of text samples to ensure the model's robustness.

1. Data Collection and Preprocessing:

Collect the healthcare dataset that contains relevant reports, documents, or web pages from which to extract entities. To ensure the model's robustness, the dataset should contain a variety of text samples. Python libraries such as NLTK, Pandas, and Scikit-learn are used to clean and preprocess the dataset. To obtain the root form of words, remove stop words, punctuation, and perform lemmatization. Tokenize and vectorize the text to convert it to a machine-readable format.

1. Model Training:

Using OpenAI’s ChatGPT model we are performing zero-shot entity extraction on the healthcare dataset with no fine-tuning or training. Use the language understanding capabilities of the model to identify and categorize entities such as treatments, symptoms, and diseases.

1. Evaluation:

In terms of accuracy and efficiency, compare the ChatGPT model's performance to that of traditional supervised learning approaches such as NER and POS tagging. Examine the effect of domain-specific text on model performance by testing it on various domains such as finance, sports, and entertainment. Examine the impact of dataset size and diversity on model performance.

1. Incorporating Additional Data:

To improve the model's performance, add more healthcare data to it. To analyze the data and identify relevant topics, use topic modeling techniques such as LDA, LSA, and Bertopic. To improve performance, fine-tune the ChatGPT model using the new data.

1. Model Improvement:

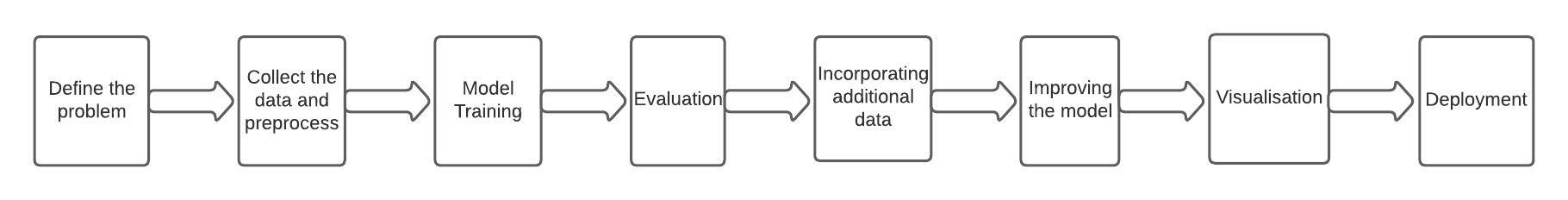
Improving the ChatGPT model's performance by fine-tuning it with the healthcare dataset. To optimize the model, use techniques such as hyperparameter tuning, regularization, and early stopping. Pre-train the model on a large corpus of text data before fine-tuning it on the healthcare dataset using transfer learning.

1. Visualization:

To gain insights and improve the model, visualize the extracted entities with Python libraries such as Spacy and Matplotlib.

1. Deployment:

To enable real-time entity extraction in healthcare applications, deploy the model as a RESTful API. Improve the functionality of other healthcare systems by integrating the model.



**Data Collection and Cleaning:**

The processes of data gathering and cleaning are crucial in every machine learning project, including zero-shot learning for entity extraction. The following are some guidelines for compiling and scrubbing data: Establish your entity types, get information from many sources, Utilize data augmentation methods.

Data cleaning: After gathering the data, you must clean it to eliminate noise and unnecessary information. Finally, the data needs to be labeled to show which words or phrases are entities and which ones are not. The training of the zero-shot entity extraction model requires this, which can be a time-consuming operation.

**Experiment and Data analysis plan:**

Here we try to filter out the existing entities from the healthcare data and then we remove the table, figures and non-alphabetical characters. Then we try extract each entity from healthcare data and which then transformed into entity graphs. To visualize the results and for better understanding we use charts and graphs. The evaluation metrics can be accuracy, precision, recall, or F1-score. Implement the model to extract the entities.

**References:**

1. [**https://stackoverflow.com/questions/74616546/zero-shot-named-entity-recognition**](https://stackoverflow.com/questions/74616546/zero-shot-named-entity-recognition)
2. [**https://datascience.stackexchange.com/?tags=machine-learning**](https://datascience.stackexchange.com/?tags=machine-learning)
3. [**https://arxiv.org/pdf/2111.07734.pdf**](https://arxiv.org/pdf/2111.07734.pdf)
4. **https://github.com/shmulvad/zero-for-ner**