**Relationship Extraction (REL) Model Delivery Using SpaCy**

**Problem Statement**: To build NLP Relation Extraction (REL) model using SpaCy, train the model with improved accuracy and serve the REL model using FastAPI.

**Dataset**: I will work on one of these Clinical Trials datasets from Clinical research website. (<https://markusstrasser.org/biomedical-datasets/>, <https://n2c2.dbmi.hms.harvard.edu/data-sets> )

**Strategy**:

1. Checking the annotations of the text which displays all the relations present in the data.
2. Perform the exploratory data analysis steps like data cleaning, data preprocessing, etc.
3. Performing the pretraining task to extract the relations between textual entities without any labels.
4. Showcasing the complete NLP (Natural Language Processing) model for clinical dataset using Relation Extraction (REL).
5. Benchmarking the performances of each transformer used in this REL model.

**Approaches**:

**Approach 1**:

1. Validating all the relations annotated using python packages (displacy visualizer) and prodigy tool.
2. Working with only top 5 relations present in the dataset.
3. Training the REL model and using tok2vec, transformer pipelines to check for the performance.
4. Serving the trained REL model on FastAPI and identifying all the relations.

**Approach 2**:

1. Exploring the biomedical clinical trial dataset and performing data preprocessing on it.
2. Working with only top 5 relations present in the dataset.
3. Training the REL using Pytorch models to check for the performance.
4. Serving the trained REL model on FastAPI.

**Approach 3**:

1. Perform the REL model training using only BERT transformer to check for the performance.
2. Increasing the number of relations at the time of data pre-processing.
3. Continuously improving the performance metrics (precision, F1-score) of the REL model and increasing it to 75-80%.
4. Serving the trained REL model on Fast API.

**Technologies to be used**: SpaCy, Python, Google Colaboratory, Python packages, FastAPI, Transformers, Tok2vec, Spacy explosions REL component.

**Timelines**: The following is the rough timeline of what needs to be achieved/delivered for next 5 weeks.

1. **Monday 10th October 2022**: Start of the REL model task. Identifying the problem statement and carrying out the experimentation on the clinical trial dataset.
2. **Monday 17th October 2022**: Performing the data cleaning, preprocessing on clinical trials dataset, and finding out the total relationships present in the dataset. Finishing 30% of the task. Starting with Approach 1.
3. **Thursday 27th October 2022**: Performing data preprocessing on the new biomedical dataset. Training the REL on PyTorch model, improving the performance metrics (like precision, accuracy, etc.). Working on Approach 2. Finishing 50% of the task.
4. **Monday 31st October 2022**: Improving the performance metrics of the transformer model (precision, accuracy, etc.) by increasing the relations and fine-tuning the parameters. Working on Approach 3. Finishing 75% of the task.
5. **Monday 7th November 2022**: Benchmarking the performances of each transformer models used in REL.

**Challenges**:

1. Sometimes the performance metrics like precision, accuracy may not increase in the REL model due to the nature of the dataset.
2. During increasing the number of relations, the performance metrics of the model may change or may decrease.
3. Time of execution of the REL model when using tok2vec pipeline.