**Project Description and List of Projects**

This will be a three-membered team project. For each topic, we provided a reference paper to read and the dataset will be a small dataset in your native language (if a team is of size three then you can work on three, two or one native language of your choice). These papers are just to give you an idea regarding the project. You are free to use any other rule-based approaches. Each project will have some amount of Data preparation, Data refinement,etc. A preference form will be released soon. Project will be allocated on an FCFS basis, so be ready with your choices before we release the form.

**OUTCOMES OF THE PROJECT:**

1. Original code. Plagiarism will be penalised.
2. Work distribution for each team member should be equal.
3. Each team should submit one report. Please do not submit multiple reports from the same team.
4. . Zip the pdf of the paper selected by the team, report PDF, code files in a folder, and submit it with the team name.
5. The submission zip file should be “teamname.zip”.

**PROJECT EVALUATION:**

1**.** Phase1 - The team should complete data crawling and preprocessing.

A brief presentation about the aim and objectives of the project, data description etc. is expected.

2. Phase2 - The team should be ready with the methods they want to use and show the results for baseline.

3. Phase3 - The team should be ready with the whole project code, results and the final presentation,

**List of Projects with Description:**

1. Timex Identification: This task involves identifying time expressions in the text following the TIMEX3 rules. Data Annotation will be a part of the project.( BIO tagging)

Rule-based Paper: <https://www.researchgate.net/publication/248737128_TimeML_Annotation_Guidelines_Version_121> (Section 2.2)

Data : Any Indian Language. Dataset collection will be a part of the project (atleast 500 sentences).

1. Question Generation from stories : Given fables written in a very simple language you need to generate questions based on the fable, whose answers are available in the fable itself.

Rule-based Paper:

<https://arxiv.org/pdf/1906.08570.pdf>

Data : Any Indian Language of your choice and dataset preparation will be a part of the project.

1. Coreference resolution : Identify the different mentions of a Entities in Narrative texts. Here for this project the team needs to implement the given paper.

Rule-based Paper:

<https://www.researchgate.net/publication/341446275_Resolving_Actor_Coreferences_in_Hindi_Narrative_Text>

Data : Any Indian language. Pick scripts from Wikipedia for movies/short films.

1. Sentiment Analysis: Identifying the polarity of the sentences as positive, negative or neutral.

Rule-based Paper:

[https://sci-hub.do/https://ieeexplore.ieee.org/document/7379415](https://sci-hub.do/https:/ieeexplore.ieee.org/document/7379415)

Data : Any Indian language (atleast 200 sentences)

1. Named Entity Recognition: The task is to identify named entities (NE) in the annotated data. Identify patterns for different named entity categories, and test them on unseen data.

Rule based NERs employ Gazetteers list, containing a list of entities from

different classes to find NEs. Domain specific NER can also be implemented.

Papers to read :

a. Rule-based: <https://www.aclweb.org/anthology/C96-1071.pdf>

Datasets : Any Indian language (atleast 200 sentences)

1. Generate Wikipedia articles given factual triplets. Data will have to be extracted and rules for templates have to be made.
2. Fine Grained Categorisation of Wikipedia Labels over 320 labels using statistical methods.

A very similar [task](http://shinra-project.info/shinra2020ml/?lang=en#:~:text=Task%20Description&text=SHINRA2020%2DML%20is%20the%20first,of%20the%20NTCIR%2D15%20tasks)

1. Word Similarity: Given a set of test words, find the next n words which are similar to the given word in multilingual space.

Dataset - <https://aclanthology.org/W17-0811/>

1. A rule-based chatbot for a specific domain.