PROJECT-2 GUIDELINES

This will be a 3-membered team project. For each topic, we provided a few papers to read and a dataset. 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 on July 2nd. Project will be allocated on an FCFS basis, so be ready with your choices before we release the form.

OUTCOMES OF THE PROJECT-2:

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”.

TOPICS:

1. **Sentiment Analysis:**

   Identifying the polarity of the sentences as positive, negative or neutral.

   Paper to read:

   a. ML:https://nlp.stanford.edu/courses/cs224n/2012/reports/WuJean_PaoYuanyuan_224nReport.pdf
   b. Rule-based: https://sci-hub.do/https://ieeexplore.ieee.org/document/7379415

   Languages : Telugu, Hindi.

2. **Hate Speech Detection:**

   Hate speech is any form of expression through which speakers intend to vilify, humiliate, or incite hatred against a group or a class of persons on the basis of race, religion, skin color, sexual identity, gender identity, ethnicity, disability, or national origin. This task involves identifying sentences containing such instances.

   Papers to read:

   a. Rule-based: https://gvpress.com/journals/IJMUE/vol10_no4/21.pdf

b. Dictionary based: https://www.aclweb.org/anthology/N16-2013.pdf

Languages : Hindi

3. **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)

Papers to read :

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

Languages : Any Indian Language. Dataset collection will be a part of the project.

4. **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
b. ML based: https://arxiv.org/pdf/1508.01991.pdf
a. Paper of your choice.

Datasets :

Languages : Telugu, Hindi.

5. **Document Classification**

Choose 3-4 topics. Scrap data for these corresponding topics . Use Gazetteers list, Frequency Analysis and Using Inverse frequency methods to obtain the result.

Papers to read:

a. Rule-based:https://www.researchgate.net/publication/342170927_A_Rule-Based_Approach_to_Embedding_Techniques_for_Text_Document_Classification
b. Paper of your choice.

Language : Any Indian Language. Dataset collection will be a part of the project.

6. **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.

   Papers to read:
   a. Rule-based : https://arxiv.org/pdf/1906.08570.pdf

   Datasets : Any Indian Language. Requires Data preparation

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

   Papers to read:
   a. Rule-based : https://www.researchgate.net/publication/341446275_Resolving_Actor_Coreferences_in_Hindi_Narrative_Text

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

8. **Clickbait Identification:**
   Clickbait is a text/title/link designed to attract the user's attention. In this task, use different rule-based approaches to Identify whether a text is a clickbait or not.

   Papers to read:
   https://www.researchgate.net/publication/323156528_Believe_it_or_not_identifying_bizarre_news_in_online_news_media

   Language : Hindi, Telugu