Project Proposal

Suyash Sethia (2021114010)
Jainit Bafna (2021114003)

September 30, 2022

Problem Statement

Creation of word embedding system for Hindi data set scraped through news articles and twitter. Detection of gender-bias using different similarities like cosine similarity in word embeddings vector. The bias will be analyzed for gender-bias in different domains separately like family relations, career, arts, strength.

Dataset

The data would be scraped from

- Various news sites like BBC hindi, The Hindu, etc ..
- Twitter

Then it will tested to find bias in various domains

- Career vs Family
- Maths vs Arts
- Science vs Arts
- Intelligence vs Appearance
- Strength vs Weakness

Architecture

So the project is primarily divided into 5 parts

1. Scraping data from news articles and twitter
2. Cleaning the corpus
3. Creation of data set of different domains
4. Creation of word embeddings
5. Analyzing the data set for gender - bias using similarities functions

Timeline

10th October	Scraping data and corpus
25th October	Creation of data set of different domains
5th November	Creation of word embedding
13th November	Analyzing data set for gendee
15th November	Final submission

References

[1] Measuring Gender Bias in Word Embeddings across Domains and Discovering New Gender Bias Word Categories(https://aclanthology.org/W19-3804) (Chaloner & Maldonado, 2019)