M22CL3.202: Computational Linguistics -II
Project
Mid-submission

Suyash Sethia 2021114010

Jainit Bafna 2021114003

IIIT Hyderabad — October 15, 2022

Note

All the Progress, research work, and related projects for reference can be viewed on Repository-link

Progress

1. The data sets were made for different domains like arts, math, science, engineering, biology, appearance, strength. These domains will be looked upon for gender bias.
2. The data was also made for entities that are male /female by definition to act as a reference for the word embeddings to be made.
3. The Hindi data was collected from news articles by web scraping and also was cleaned removing stop words and foreign words and unwanted characters were removed (the corpus cleaning is done carefully to remove only those stop words which do not possess any information on gender)(check file corpusformation.py.
4. For our final project, we will create a word embedding system based on Neural Network. For which we created unique ids for all words (after removing stopwords) (check file uniqueID.py).
5. Then We created the word embedding matrix using one-hot encoding check file matrixformation.py

Example

1. We took a sentence
   हिमाचल प्रदेश में विधानसभा चुनाव का एलान हो गया है

2. After the removal of stopwords.
   The words left are :-[हिमाचल ,प्रदेश ,विधानसभा ,चुनाव ,का ,एलान ,गया]

3. We will provide every word a unique id
   unique-word-dict =
   “हिमाचल”:1 ,
   “प्रदेश ”:2 ,
   “विधानसभा”:3 ,
   “चुनाव”:4 ,
   “का”:5 ,
   “एलान”:6 ,
   “गया ”

4. Then We create the word embedding Matrix.
   For eg.
   a = [’नीला’, ’आकाश’, ’नीला’, ’गाड़ी’]

   ’नीला’ = [1, 0, 0]
   आकाश= [0, 1, 0]
   ‘गाड़ी’ = [0, 0, 1]

5. The list can be converted into a matrix:

   A = [(1, 0, 0) , (0, 0, 1) , (1, 0, 0) , (0, 1, 0)]