Natural Language Processing - CSE556 ReadMe

MONSOON SEMESTER 2020

ANUNAY YADAV - 2018021
HAMZAH AKHTAR - 2018051

Methodology:

- Read the BingLiu lexicon file and the english-gindi bilingual file and saved the mappings in a dictionary.
- Traversed over all the elements in the BingLiu dictionary. If a map for a word in BingLiu Dict was present in the english-hindi mapping then that english-hindi mapping was added to a list to create `L1.csv`.
- Read the `english.txt` and `hindi.txt` files and save its contents in a variable by splitting them on the newline character. Each sentence was then tokenized.
- The tokenized sentences list was passed to the word2vec for training a word embedding on the given corpus.
- Similarly the stanford glove script was used to train glove models on the given english and hindi corpus to create new word embedding models.
- The parameters for both word2vec and glove models were fine tuned to give the best results on the given corpuses.
- After training the two models we traverse over each word mapping in L1.csv file and find the top 5 similar english word and hindi for the corresponding english and hindi word using an inbuilt function. If and of those 25

combinations exist in the english-hindi bilingual dictionary, then that combination is added to the list of new additions to L1.csv.

- The above process is repeated on the new additions to the lexicon list until there are no more unique lexicons to be added.
- This process is done for both models word2vec and glove and the extended lexicon is reported for for both of them along with the new additions to the lexicons