# CSE556: Natural Language Processing

## Assignment 04 Report

Naman Jain
2018347

Navneet Agarwal
2018348

---

**Assignment 4: Multilingual extension of BingLiu lexicon**

# Preprocessing

- For english.txt
    - Removed all the digits and multiple spaces between words
    - Convert into Lower case
    - Removed the punctuations.
    - Tokenize into sentences and each sentence into words using nltk.RegexpTokenizer(r"\w+") function.
    - Removed all the stop words from these sentences.
- For hindi.txt
    - Removed multiple spaces
    - Removed the digits
    - Removed the punctuations

# Methodology

### Procedure to Construct L1 (Bilingual Lexicon):

- Read BingLiu.csv , english-hindi-dictionary.txt and stored them in a dataframe and a list respectively.
- For every word in Bing_Liu lexicon :
    - Check if the word is in the english-hindi-dictionary list.
        - If yes append the original word, translated word and it's polarity into L1 list(Bilingual Lexicon)

### GloVe embedding extraction:

The official repository was cloned from GitHub and CORPUS name was changed in demo.sh file to train GloVe algorithm on english.txt and hindi.txt. Obtained vectors txt files were then moved to files directory and imported into the notebook.

### Word2Vec embedding extraction

The preprocessing was done on english.txt and hindi.txt according to the preprocessing steps mentioned above. These were then fed as input to the Word2Vec model under the gensim library to train on the given corpus. These trained models could be further used to extract top 5 words for any input word.

### Procedure for extending L1 using mapped pairs:

- Eng -> Hindi and Hindi -> Eng dictionary were stored initially.
- Then models were trained for word2vec and GloVe separately.
- L1 was iterated and closest 5/10 words were found for both english and hindi words using the corresponding trained model (word2vec/GloVe).
- Now all the possible unique pairs using 5 english & 5 hindi were created and checked if present in the eng-hindi dictionary.
    - If any pair mapped then it was added into L1 with respective polarity.
- The above steps were continued until there was no new addition to L1.

## Outputs

### Bilingual Lexicon L1 :

```
L1
```

```
[['abundant', 'हुस्न', 'positive'],
 ['accessible', 'सुलभ', 'positive'],
 ['accomplish', 'पूरा', 'positive'],
 ['accomplished', 'पूरा हुआ', 'positive'],
 ['accomplishment', 'उपलब्धि', 'positive'],
 ['accomplishments', 'उपलब्धियों', 'positive'],
 ['accurate', 'सटीक', 'positive'],
 ['accurately', 'यथासंभव', 'positive'],
 ['achievement', 'उपलब्धि', 'positive'],
 ['achievements', 'उपलब्धियाँ', 'positive'],
 ['acumen', 'कौशल', 'positive'],
 ['adjustable', 'समायोजनीय', 'positive'],
 ['admirable', 'प्रशंसनीय', 'positive'],
 ['admire', 'प्रशंसा', 'positive'],
 ['admirer', 'प्रशंसक', 'positive'],
 ['adorable', 'प्यारे', 'positive'],
```

**GloVe outputs:**

```
Extended Lexicons for closest 5 words:
('simple', 'सरल', 'positive')
('disappointed', 'निराश', 'positive')
('impressed', 'प्रभावित', 'negative')
```

Total 3 additions were there in the above case.

```
Extended Lexicons for closest 10 words:
('more', 'अधिक', 'positive')
('simple', 'सरल', 'positive')
('disappointed', 'निराश', 'positive')
('impressed', 'प्रभावित', 'negative')
('much', 'ज्यादा', 'positive')
```

Total 5 additions were there in the above case.

**Word2Vec outputs：**

- Top 5 words considered (1 addition was there)

```
['usage', 'उपयोग', 'negative']
```

- Top 5 words considered (5 additions were there)

```
['one', 'एक', 'positive']
['something', 'कुछ', 'positive']
['usage', 'उपयोग', 'negative']
['much', 'ज्यादा', 'negative']
['time', 'समय', 'positive']
```

# Contribution

Naman : All the work related to GloVe
Navneet : All the work related to word2vec