Megha Sundriyal PhD20009, Pragya Sethi 2018067

Assignment 4
Multilingual extension of BingLiu lexicon

How to run?

1. Open the *bingliu_extension.ipynb* file and run all the cells. Running on colab is preferred. Hindi symbols are not formed on command prompt.
2. You can run binglu_extensions.py directly. Hindi symbols will not show on command prompt.

Methodology

1. For creating multilingual extensions of BingLiu lexicon, the bilingual English-Hindi dictionary was utilized. For each term in the present English Bing Liu, we retrieved it's Hindi translation. There were terms for which Hindi translation did not exist and thus such English terms were discarded while rest were used to form the extension of Bing Liu lexicon. (L1)

2. For training word embedding models, the text data was first cleaned employing following text-preprocessing methods:
    a. Remove punctuations
    b. Remove stop words
    c. Remove unwanted special characters like ( ) | # @ $ etc.
    d. Remove digits
    e. Lower-case English text
    f. Each word is then tokenized using:
        i. nltk.word_ tokenizer for English text
        ii. Indic_tokenize.trivial_tokenize for Hindi text

3. Once the data was cleaned for both English and Hindi text, we trained word2vec and glove models and hereby obtained two types of monolingual word embeddings for both English and Hindi data.

4. For getting an extended lexicon list, we picked the first item from L1. Then for each word we found the closest 5 words. All the combinations of these two lists were taken and added to L1 if the combination was found in English-Hindi dictionary.

Megha Sundriyal PhD20009, Pragya Sethi 2018067

Outputs

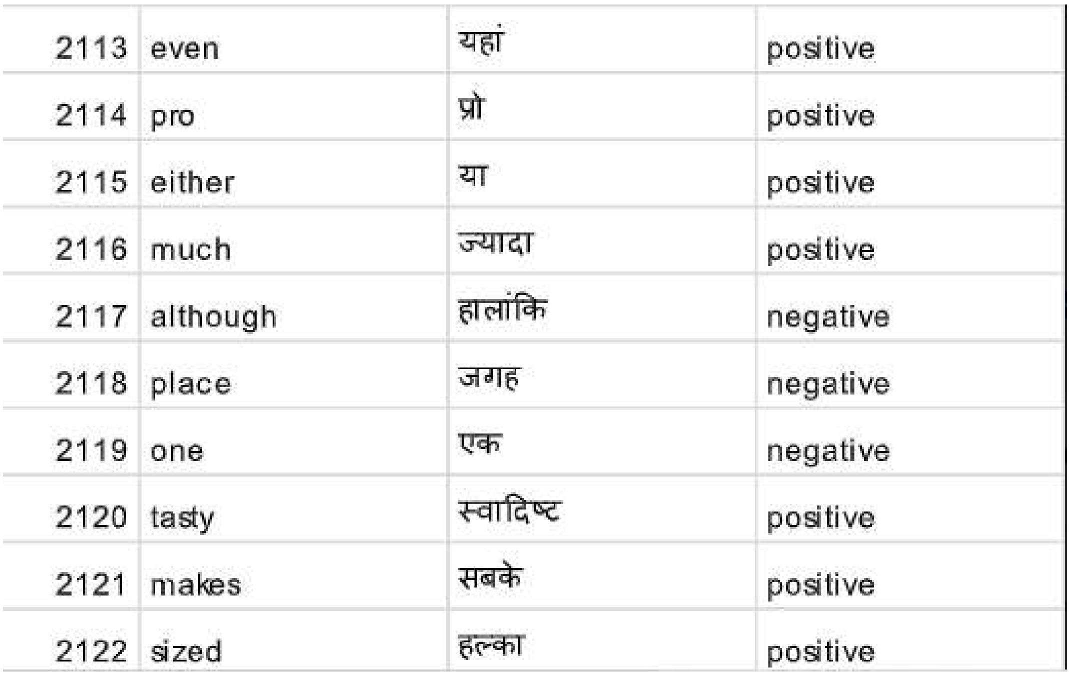
The extended lexicon using:

1. word2vec word embedding - 7

```
{('although', 'हालांकि'): 'negative',
 ('either', 'या'): 'positive',
 ('even', 'यहां'): 'positive',
 ('much', 'ज्यादा'): 'positive',
 ('one', 'एक'): 'negative',
 ('place', 'जगह'): 'negative',
 ('pro', 'प्रो'): 'positive'}
```

2. glove word embedding - 3

   a. `{('makes', 'सबके'): 'positive',`
   b. `('sized', 'हल्का'): 'positive',`
   c. `('tasty', 'स्वादिष्ट'): 'positive'}`



2113	even	यहां	positive
2114	pro	प्रो	positive
2115	either	या	positive
2116	much	ज्यादा	positive
2117	although	हालांकि	negative
2118	place	जगह	negative
2119	one	एक	negative
2120	tasty	स्वादिष्ट	positive
2121	makes	सबके	positive
2122	sized	हल्का	positive