**Code**

**Program Details**

* This contains all the steps to compute the number of non english words .
* Each step takes the output of the previous step as the input .
* First step: take the song file with text.
* Final step will produce the output as number of english and non english words
* Data cleaning and preprocessing has been done in first three steps which are removing punctuation , converting upper to lowercase and converting the text to key value pairs

**Step 1: Remove Punctuation PREPROCESSING**

**Input : file.txt**

# define punctuation

Function remove\_punctuation (file.txt)

punctuations = '''!()-[]{};:'"\,<>./?@#$%^&\*\_~'''

my\_str = file.txt.content

no\_punct = ""

for char in my\_str:

if char not in punctuations:

no\_punct = no\_punct + char

print(no\_punct)

f.write(file.txt, no\_punct)

**Output: file.txt contains text with no puinctuation**

**Step 2 convert upper to lower case PREPROCESSING**

**Input : output of step 1 i.e. file.txt which contains text with no punctuation**

function(file.txt)

string = file.txt.content

Txt = string.lower()

f.write(file.txt, Txt)

**Output: file.txt contains all txt in lowercase**

**Step 3 Splitting the text in key value pairs(Record Reader)**

This is the first phase of MapReduce where the Record Reader reads every line from the input text file as text and yields output as key-value pairs.

**Input − Line by line text from the input file.**

**Output − Forms the key-value pairs.**

**Step 4 MAP PHASE:**

Mapper will generate output as key value pair where value 0 denotes E and 1 denotes NE

**Input : abc.txt content: ‘this is a ball, this is a pen yeh ek pen’**

from nltk.corpus import words

setofwords = set(words.words())

function mapper(file file.txt):

txt= file.txt

list1= txt.split(‘ ‘) //to create the list of words

arr =['E']

arr2 =['NE']

cnt=0;cnt2=0

for item in list1:

if(item in setofwords):

str1= ‘<’+ item + ‘0>’

arr.append(str1)

else:

str2= 1

arr.append(str2)

Return (arr)

Output:

First convert into sentence

<1,this is a ball in hindi yeh ek pen>

<2,this is a pen>

**Mapper output**

**<this , 0>, <yeh, 1>, <is ,0> <ball, 0>.......<word, 0>**

**STEP 5 COMBINER Phase**

**Input :[<w1,0>, <w2, 1>, <W3,1>, <W4,0>...........<wn,0>]**

function combiner(arr):

arr1=[‘English\_Words’]

arr2=[‘Non\_English’]

for item in arr: // <word1, E>

If item.value(‘0’):

arr1.append(0)

else:

arr2.append(1)

**Output: arr2 : [English\_Words, 0,0,0,0………………………]**

**arr1: [Non\_English, 1,1,1,1,1…………………………]**

**STEP 6: REDUCER PHASE**

**REDUCER:**

**Input :**

**arr2 : [English\_Words, 0,0,0,0………………………]**

**arr1: [Non\_English, 1,1,1,1,1…………………………]**

Cnt = 0;

Cnt2 =0 ;

Def reducer(arr2, arr1):

For item in arr1:

if (item ==0):

cnt= cnt + 1;

If (item ==1):

Cnt2 = cnt2 +1

print(‘English:’ + cnt)

print(‘Non\_English:’ + cnt2)

**Output:**

**Key value pair with key as english or non english and value as count**

**Eg: English: cnt**

**Non\_English: cnt2**