import nltk

import math

import io

import gensim

from gensim import \*

from nltk.tokenize import word\_tokenize, sent\_tokenize, RegexpTokenizer

from nltk.corpus import stopwords

from nltk.stem import PorterStemmer

from nltk.stem import WordNetLemmatizer

final\_words=[]

'''UTILITY FUNCTIONS'''

def tokenizeWords(fcontent):

tokenizer=RegexpTokenizer(r'\w+')

word\_tokens=[]

word\_tokens=tokenizer.tokenize(fcontent)

return word\_tokens

def removeStopWords(word\_tokens):

stop\_words=set(stopwords.words('english'))

filtered\_words=[]

for w in word\_tokens:

if w not in stop\_words:

filtered\_words.append(w)

return filtered\_words

def lemmatize(filtered\_words):

lem\_words=[]

lemmatizer=WordNetLemmatizer()

for w in filtered\_words:

word=(lemmatizer.lemmatize(w))

lem\_words.append(word.lower())

return lem\_words

def copyfinal(lem\_words):

final\_words.append(lem\_words)

def newBigramList(lem\_words):

length=len(lem\_words)

new\_list=[]

for i in range(0,length-1):

new\_list.append(lem\_words[i]+lem\_words[i+1])

return new\_list

def distance(a,b):

dist=0

for i in range(0,4):

dist=dist+((a[i]-b[i])\*\*2)

dist=math.sqrt(dist)

return dist

#driver code

doc\_count=0

docwords=[]

choice=1

while(choice==1):

#accepting file and reading its content

name=input('Enter filename: \n')

fp=open(name,'r')

fcontent=fp.read()

doc\_count+=1

#tokenizing

word\_tokens=tokenizeWords(fcontent)

fp.close()

#removing stop words

filtered\_words=removeStopWords(word\_tokens)

#lemmatizing

lem\_words=lemmatize(filtered\_words)

copyfinal(lem\_words)

lem\_words=newBigramList(lem\_words)

docwords.append(lem\_words)

choice=int(input("Add another file? (0 or 1): "))

print("Documents are given")

print(docwords)

#Word2Vec

import gensim

model=gensim.models.Word2Vec(docwords,min\_count=1,size=4)

print(choice)

X= model[model.wv.vocab]

print(X)

# KMEANSCLUSTER

from nltk.cluster import KMeansClusterer

NUM\_CLUSTERS=100

kclusterer = KMeansClusterer(NUM\_CLUSTERS, distance=nltk.cluster.util.euclidean\_distance, repeats=25)

assigned\_clusters = kclusterer.cluster(X, assign\_clusters=True)

#Wordcluster is a dictionary with cluster no as keys and list of words as values

for i,word in enumerate(words):

wordcluster[assigned\_clusters[i]].append(word)

total\_words=0

for i in range(0,100):

total\_words=total\_words+len(wordcluster[i])

print(str(i)+" :"+str(len(wordcluster[i])))

print("Total words "+str(total\_words))

document\_clusters={}

for i in range(1,doc\_count+1):

document\_clusters.update({i:[0 for i in range(0,100)]})

for i in range(0,doc\_count):

for j in range(0,100):

for w in wordcluster[j]:

if w in docwords[i]:

document\_clusters[i+1][j]=document\_clusters[i+1][j]+1

distancelist=[]

for i in range(1,doc\_count+1):

distancelist.append(document\_clusters[i])

import numpy as np

myarray = np.asarray(distancelist)

NUM\_CLUSTERS=5

kclusterer = KMeansClusterer(NUM\_CLUSTERS, distance=nltk.cluster.util.euclidean\_distance, repeats=15)

assigned\_clusters = kclusterer.cluster(myarray, assign\_clusters=True)

for i in range(0,100):

print("Document "+str(i+1)+" : Cluster "+str(assigned\_clusters[i]))

**OUTPUT**

0 :216

1 :168

2 :176

3 :204

4 :148

5 :174

6 :177

7 :156

8 :136

9 :178

10 :183

11 :137

12 :178

13 :173

14 :180

15 :152

16 :141

17 :134

18 :193

19 :173

20 :161

21 :148

22 :168

23 :153

24 :178

25 :193

26 :208

27 :118

28 :187

29 :187

30 :145

31 :145

32 :139

33 :158

34 :133

35 :166

36 :153

37 :157

38 :157

39 :157

40 :146

41 :179

42 :148

43 :168

44 :179

45 :166

46 :151

47 :175

48 :122

49 :141

50 :149

51 :141

52 :147

53 :150

54 :142

55 :148

56 :157

57 :150

58 :141

59 :131

60 :134

61 :152

62 :188

63 :115

64 :117

65 :198

66 :151

67 :135

68 :143

69 :141

70 :134

71 :123

72 :159

73 :179

74 :131

75 :135

76 :167

77 :128

78 :139

79 :149

80 :116

81 :144

82 :126

83 :122

84 :157

85 :162

86 :161

87 :165

88 :147

89 :130

90 :143

91 :129

92 :145

93 :147

94 :168

95 :215

96 :141

97 :135

98 :132

99 :149

Total words 15471

Document 1 : Cluster 1

Document 2 : Cluster 2

Document 3 : Cluster 0

Document 4 : Cluster 0

Document 5 : Cluster 0

Document 6 : Cluster 0

Document 7 : Cluster 0

Document 8 : Cluster 2

Document 9 : Cluster 2

Document 10 : Cluster 1

Document 11 : Cluster 0

Document 12 : Cluster 1

Document 13 : Cluster 1

Document 14 : Cluster 2

Document 15 : Cluster 2

Document 16 : Cluster 1

Document 17 : Cluster 0

Document 18 : Cluster 0

Document 19 : Cluster 3

Document 20 : Cluster 0

Document 21 : Cluster 1

Document 22 : Cluster 1

Document 23 : Cluster 1

Document 24 : Cluster 0

Document 25 : Cluster 0

Document 26 : Cluster 0

Document 27 : Cluster 4

Document 28 : Cluster 0

Document 29 : Cluster 2

Document 30 : Cluster 1

Document 31 : Cluster 1

Document 32 : Cluster 0

Document 33 : Cluster 0

Document 34 : Cluster 0

Document 35 : Cluster 1

Document 36 : Cluster 1

Document 37 : Cluster 1

Document 38 : Cluster 2

Document 39 : Cluster 2

Document 40 : Cluster 0

Document 41 : Cluster 3

Document 42 : Cluster 1

Document 43 : Cluster 1

Document 44 : Cluster 0

Document 45 : Cluster 1

Document 46 : Cluster 2

Document 47 : Cluster 0

Document 48 : Cluster 0

Document 49 : Cluster 1

Document 50 : Cluster 1

Document 51 : Cluster 3

Document 52 : Cluster 3

Document 53 : Cluster 0

Document 54 : Cluster 0

Document 55 : Cluster 1

Document 56 : Cluster 1

Document 57 : Cluster 1

Document 58 : Cluster 2

Document 59 : Cluster 0

Document 60 : Cluster 2

Document 61 : Cluster 1

Document 62 : Cluster 2

Document 63 : Cluster 1

Document 64 : Cluster 2

Document 65 : Cluster 0

Document 66 : Cluster 0

Document 67 : Cluster 1

Document 68 : Cluster 1

Document 69 : Cluster 1

Document 70 : Cluster 0

Document 71 : Cluster 2

Document 72 : Cluster 0

Document 73 : Cluster 0

Document 74 : Cluster 1

Document 75 : Cluster 1

Document 76 : Cluster 1

Document 77 : Cluster 1

Document 78 : Cluster 2

Document 79 : Cluster 0

Document 80 : Cluster 0

Document 81 : Cluster 2

Document 82 : Cluster 1

Document 83 : Cluster 1

Document 84 : Cluster 3

Document 85 : Cluster 1

Document 86 : Cluster 2

Document 87 : Cluster 0

Document 88 : Cluster 0

Document 89 : Cluster 2

Document 90 : Cluster 1

Document 91 : Cluster 2

Document 92 : Cluster 0

Document 93 : Cluster 4

Document 94 : Cluster 1

Document 95 : Cluster 0

Document 96 : Cluster 0

Document 97 : Cluster 1

Document 98 : Cluster 1

Document 99 : Cluster 1

Document 100: Cluster 0