Web Mining – Lab 9 (Virtual Lab)

Cosine Similarity Clustering

By Abhijeet Ambadekar (16BCE1156)

Problem Statement:
To clustering documents using the cosine similarity in between them

Program Code:

```
import numpy as np
from pprint import pprint
from math import inf, sqrt


def get_word_count(words, text):
	if type(words) == str:
		return text.count(words)
	elif type(words) == list:
		return sum([text.count(wordi) for wordi in words])


class CosineSimilarityClusterer():
	def __init__(self, threshold=0.25):
		self.threshold = threshold
		self.wordlist = []
		self.doclist = {}
		self.doc_vecs = {}

		self.clusters = {}

	def get_text_data(self, docfile):
		with open(docfile) as fp:
			content = list(map(str.lower, fp.readlines()))

		for textline in content:
			docname, doctext = textline.split(" : ")
			self.doclist[docname] = doctext

	def get_word_freq(self):
		for doc in self.doclist:
			self.doc_vecs[doc] = np.array([get_word_count(word,
self.doclist[doc]) for word in self.wordlist])

	def get_dot_prod(self, doc1, doc2):
		return self.doc_vecs[doc1].T.dot(self.doc_vecs[doc2])

	def get_doc_length(self, doc):
		return sqrt(self.get_dot_prod(doc,doc))

	def get_cos_angle(self, doc1, doc2):
		return self.get_dot_prod(doc1, doc2) / (self.get_doc_length(doc1) *
self.get_doc_length(doc2))
```

```
	def get_nearest_doc(self, i, doc):
		return sorted(self.doc_vecs.keys())[max(range(len(self.doc_vecs)),
key=lambda k: self.cos_mat[i][k])]

	def compute_cosine_matrix(self):
		for i, doci in enumerate(sorted(self.doc_vecs.keys())):
			for j, docj in enumerate(sorted(self.doc_vecs.keys())):
				if i == j:
					self.cos_mat[i][j] = -inf # Similar docs will have cos
value as 1 and will be redundant in the calculation
				else:
					self.cos_mat[i][j] = round(self.get_cos_angle(doci,
docj), 2)


	def fit(self, wordlist, docfile):
		self.wordlist = wordlist
		self.get_text_data(docfile)
		self.get_word_freq()

		self.cos_mat = [[0 for i in range(len(self.doc_vecs))] for j in
range(len(self.doc_vecs))]
		self.compute_cosine_matrix()

	def cluster(self):
		clusters = []
		for i, doc in enumerate(sorted(self.doc_vecs.keys())):
			nearest_doc = self.get_nearest_doc(i, doc)
			clusters.append({doc, nearest_doc})

		end_clusters = []

		while clusters:
			flag_merged = True
			first = clusters.pop(0)
			while flag_merged:
				flag_merged = False
				for i in range(len(clusters)):
					if first.intersection(clusters[i]):
						first.update(clusters[i])
						clusters[i] = set()
						flag_merged = True

			clusters = [j for j in clusters if len(j) != 0]
			end_clusters.append(first)

		self.clusters = {"Cluster "+str(i):end_clusters[i] for i in
range(len(end_clusters))}


	def print_details(self):
		print("The document vectors look as follows:")
		pprint(self.doc_vecs)

		print("\nThe cosine similarity matrix looks as follows: ")
		pprint(self.cos_mat)
```

```
            print("\nThe documents are clustered using the Nearest Neighbour method as
follows:")
            pprint(self.clusters)



wordlist = ["automotive", "car", "motorcycle", "self-drive", "iot", "hire", "dhoni"]
docfile = "docfile.txt"

csc = CosineSimilarityClusterer()
csc.fit(wordlist, docfile)
csc.cluster()
csc.print_details()
```

Output:

```
/media/anonymous/Work/Vit/Semester 5/WM/Lab/L9_CosineSimilarity  echo 16BCE1156
16BCE1156
/media/anonymous/Work/Vit/Semester 5/WM/Lab/L9_CosineSimilarity  python3 L9_CosineSimCluster.py
The document vectors look as follows:
{'doc 1': array([1, 0, 0, 0, 0, 0, 0]),
 'doc 2': array([1, 1, 0, 0, 0, 0, 0]),
 'doc 3': array([0, 0, 1, 0, 0, 0, 0]),
 'doc 4': array([0, 1, 0, 1, 1, 0, 0]),
 'doc 5': array([1, 0, 0, 0, 0, 1, 0]),
 'doc 6': array([0, 1, 0, 0, 0, 0, 1]),
 'doc 7': array([0, 0, 0, 0, 0, 0, 1]),
 'doc 8': array([0, 1, 0, 0, 1, 0, 0]),
 'doc 9': array([0, 0, 0, 0, 0, 0, 1])}

The cosine similarity matrix looks as follows:
[[-inf, 0.71, 0.0, 0.0, 0.71, 0.0, 0.0, 0.0, 0.0],
 [0.71, -inf, 0.0, 0.41, 0.5, 0.5, 0.0, 0.5, 0.0],
 [0.0, 0.0, -inf, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0],
 [0.0, 0.41, 0.0, -inf, 0.0, 0.41, 0.0, 0.82, 0.0],
 [0.71, 0.5, 0.0, 0.0, -inf, 0.0, 0.0, 0.0, 0.0],
 [0.0, 0.5, 0.0, 0.41, 0.0, -inf, 0.71, 0.5, 0.71],
 [0.0, 0.0, 0.0, 0.0, 0.0, 0.71, -inf, 0.0, 1.0],
 [0.0, 0.5, 0.0, 0.82, 0.0, 0.5, 0.0, -inf, 0.0],
 [0.0, 0.0, 0.0, 0.0, 0.0, 0.71, 1.0, 0.0, -inf]]

The documents are clustered using the Nearest Neighbour method as follows:
{'Cluster 0': {'doc 5', 'doc 2', 'doc 1', 'doc 3'},
 'Cluster 1': {'doc 8', 'doc 4'},
 'Cluster 2': {'doc 6', 'doc 7', 'doc 9'}}
/media/anonymous/Work/Vit/Semester 5/WM/Lab/L9_CosineSimilarity
```