МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
“Київський політехнічний інститут ім. Ігоря Сікорського”
ФАКУЛЬТЕТ Інформатики та Обчислювальної Техніки
КАФЕДРА Інформаційних систем та технологій

Звіт до лабораторної роботи №3

з предмету: Обробка та Аналіз текстових даних на мові Python

Перевірила:
Тимофєєва Ю.С.

Виконли:
студенти групи ІК-01
Філоненко І. Р.
Гацан С. Ю.

КИЇВ - 2023

Тема: Моделі текстових даних

Мета: Ознайомитись з основними текстовими моделями та їх створення за допомогою scikit-learn та gensim.

Варіант: 11

Завдання:

Зчитати файл doc11. Вважати кожен рядок окремим документом корпусу. Виконати попередню обробку корпусу.

1. Представити корпус як модель “Сумка слів”. Вивести вектор для слова MARINER.
2. Представити корпус як модель TD-IDF. Спробувати кластеризувати документи за лопомогою ієрархічної агломераційної кластеризації.
3. Представити корпус як модель Word2Vec. Знайти подібні слова до слів mobile, athens.

Код програми

```
#!/usr/bin/env python3

import re
import numpy as np
import pandas as pd

from matplotlib import pyplot as plt
from sklearn.feature_extraction.text import CountVectorizer, TfidfTransformer
from sklearn.metrics.pairwise import cosine_similarity
from scipy.cluster.hierarchy import dendrogram, linkage, fcluster
import nltk
from nltk.tokenize import WordPunctTokenizer
from gensim.models import word2vec

nltk.data.path.append("../Lab2/nltk_data")

file = open("doc11.txt", 'r').read().split('\n')
file = [sentence for sentence in file if len(sentence) != 0]

corpus = [re.sub(r'[^a-zA-Z\s]', '', sentence, re.I | re.A).lower() for sentence
in file]

print("Підготовлений корпус: ")
print(corpus)

# Task 1
cv = CountVectorizer(min_df=0., max_df=1.)
cv_matrix = cv.fit_transform(corpus)
vocab = cv.get_feature_names_out()

array = pd.DataFrame(cv_matrix.toarray(), columns=vocab)

print(array['mariner'])
# --------
```

```
# Task 2
# Отримуємо TD-IDF матрицю
tt = TfidfTransformer(norm='l2', use_idf=True)
tt_matrix = tt.fit_transform(cv_matrix)

array2 = pd.DataFrame(tt_matrix.toarray(), columns=vocab)
print("TD-IDF: ")
print(array2)
# Отримуємо матрицю подібності
similarity_matrix = cosine_similarity(tt_matrix)
array3 = pd.DataFrame(similarity_matrix)
print("\n")
print("Similarity matrix: ")
print(array3)
print('\n')

# Creating links
links = linkage(similarity_matrix, 'ward')

# Creating plot
print("Dendrogram: ")
plt.figure(figsize=(8, 3))
plt.title('Dendrogram')
plt.xlabel('Documents')
plt.ylabel('Length')
dendrogram(links)
plt.show()

# From dendrogram we can choose max distance = 1.5
max_dist = 1.5

cluster_labels = fcluster(links, max_dist, criterion='distance')

array4 = pd.DataFrame(corpus, columns=["Sentences"])
array5 = pd.DataFrame(cluster_labels, columns=["Cluster Labels ID"])
array6 = pd.concat([array4, array5], axis=1)
print(array6)
print("\n\n")
# --------
# Task 3
wpt = nltk.WordPunctTokenizer()
tokenized_corpus = [wpt.tokenize(sentence) for sentence in corpus]
vector_size = 100
window = 30
min_count = 1
sample = 1e-3
w2v_model = word2vec.Word2Vec(tokenized_corpus,
                              vector_size=vector_size,
                              window=window,
                              min_count=min_count,
                              sample=sample)
# Similar words
sim_words = {search_term: [
                            item[0]
                            for item
                            in w2v_model.wv.most_similar(
                                                         [search_term],
                                                         topn=5
                                                        )
                            ]
             for search_term in ["mobile", "athens"]}

print(sim_words)
# --------
```

Результат роботи програми

Підготовлений корпус:

['mariner and were sent to mars in and expanded upon the work done by mariner four years earlier', 'back under the spotlight of public scrutiny she will attempt to erase the double disappointment of the athens olympics', 'mariner ended up in the atlantic ocean in when the rocket launcher autopilot failed', 'more than million people own a mobile in the uk but mobile operators are keen to encourage people to move onto more sophisticated handsets that can do more', 'mariner the sister probe to mariner did reach mars in ', 'a further recent surv ey said that only of mobile owners were thinking of upgrading to g phones', 'radcliffe concedes she will probably learn a lot from her bad experiences in athens in time']

0   2
1   0
2   1
3   0
4   2
5   0
6   0
Name: mariner, dtype: int64

TD-IDF:

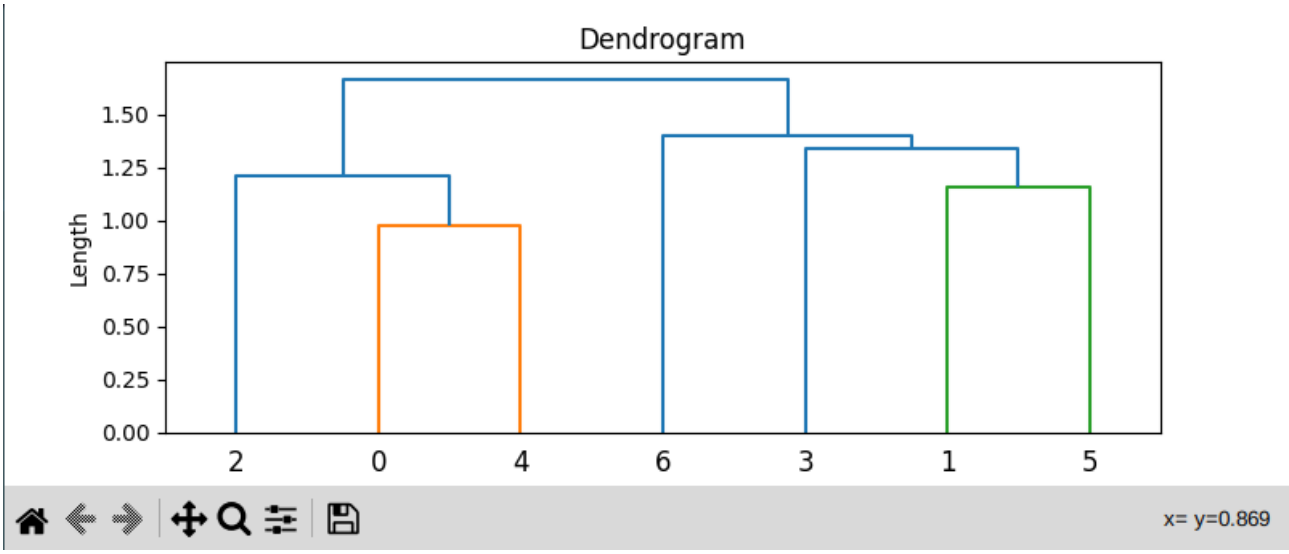
	and	are	athens	atlantic	attempt	autopilot	...	upon	were	when	will	work	years
0	0.481329	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.240665	0.199772	0.000000	0.000000	0.240665	0.240665
1	0.000000	0.000000	0.197109	0.000000	0.237456	0.000000	...	0.000000	0.000000	0.000000	0.197109	0.000000	0.000000
2	0.000000	0.000000	0.000000	0.290706	0.000000	0.290706	...	0.000000	0.000000	0.290706	0.000000	0.000000	0.000000
3	0.000000	0.173573	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
4	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
5	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.220948	0.000000	0.000000	0.000000	0.000000
6	0.000000	0.000000	0.228198	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.228198	0.000000	0.000000

[7 rows x 81 columns]

Similarity matrix:

	0	1	2	3	4	5	6
0	1.000000	0.066562	0.151931	0.048654	0.312694	0.062792	0.038530
1	0.066562	1.000000	0.120603	0.060007	0.100481	0.192607	0.134940
2	0.151931	0.120603	1.000000	0.058771	0.229354	0.000000	0.093084
3	0.048654	0.060007	0.058771	1.000000	0.073448	0.122409	0.027789
4	0.312694	0.100481	0.229354	0.073448	1.000000	0.028158	0.058165
5	0.062792	0.192607	0.000000	0.122409	0.028158	1.000000	0.000000
6	0.038530	0.134940	0.093084	0.027789	0.058165	0.000000	1.000000

Dendrogram:



	Sentences	Cluster Labels ID
0	mariner and were sent to mars in and expand...	1
1	back under the spotlight of public scrutiny sh...	2
2	mariner ended up in the atlantic ocean in wh...	1
3	more than million people own a mobile in the ...	2
4	mariner the sister probe to mariner did reac...	1
5	a further recent survey said that only of mob...	2
6	radcliffe concedes she will probably learn a l...	2

{'mobile': ['earlier', 'erase', 'survey', 'encourage', 'disappointment'],
'athens': ['scrutiny', 'move', 'under', 'did', 'people']}