Information Retrieval – Assignment 17

# Code

# IR20A.py CS5154/6054 cheng 2022  
# HAC with four different linkage modes  
# display confusion matrix and NMI between the clusterings  
# Usage: python IR20A.py  
  
import numpy as np  
from sklearn.feature\_extraction.text import TfidfVectorizer  
from sklearn.cluster import AgglomerativeClustering  
from sklearn.metrics import confusion\_matrix, ConfusionMatrixDisplay, normalized\_mutual\_info\_score  
import matplotlib.pyplot as plt  
  
f = open("bible.txt", "r")  
docs = f.readlines()  
f.close()  
N = 1000  
firstk = docs[0:N]  
  
cv = TfidfVectorizer(max\_df=0.4, min\_df=3)  
X = cv.fit\_transform(firstk).toarray()  
  
single = AgglomerativeClustering(n\_clusters=4, linkage='single')  
single.fit\_predict(X)  
  
complete = AgglomerativeClustering(n\_clusters=4, linkage='complete')  
complete.fit\_predict(X)  
  
ward = AgglomerativeClustering(n\_clusters=4, linkage='ward')  
ward.fit\_predict(X)  
  
average = AgglomerativeClustering(n\_clusters=4, linkage='average')  
average.fit\_predict(X)  
  
cm\_single\_complete = confusion\_matrix(single.labels\_, complete.labels\_)  
disp = ConfusionMatrixDisplay(cm\_single\_complete)  
disp.plot()  
plt.show()  
print(f'NMI Score for single and complete = {normalized\_mutual\_info\_score(single.labels\_, complete.labels\_)}')  
  
cm\_single\_ward = confusion\_matrix(single.labels\_, ward.labels\_)  
disp = ConfusionMatrixDisplay(cm\_single\_complete)  
disp.plot()  
plt.show()  
print(f'NMI Score for single and ward = {normalized\_mutual\_info\_score(single.labels\_, ward.labels\_)}')  
  
cm\_single\_average = confusion\_matrix(single.labels\_, average.labels\_)  
disp = ConfusionMatrixDisplay(cm\_single\_complete)  
disp.plot()  
plt.show()  
print(f'NMI Score for single and average = {normalized\_mutual\_info\_score(single.labels\_, average.labels\_)}')  
  
cm\_complete\_ward = confusion\_matrix(complete.labels\_, ward.labels\_)  
disp = ConfusionMatrixDisplay(cm\_complete\_ward)  
disp.plot()  
plt.show()  
print(f'NMI Score for complete and ward = {normalized\_mutual\_info\_score(complete.labels\_, ward.labels\_)}')  
  
cm\_complete\_average = confusion\_matrix(complete.labels\_, average.labels\_)  
disp = ConfusionMatrixDisplay(cm\_complete\_average)  
disp.plot()  
plt.show()  
print(f'NMI Score for complete and average = {normalized\_mutual\_info\_score(complete.labels\_, average.labels\_)}')  
  
cm\_ward\_average = confusion\_matrix(ward.labels\_, average.labels\_)  
disp = ConfusionMatrixDisplay(cm\_ward\_average)  
disp.plot()  
plt.show()  
print(f'NMI Score for ward and average = {normalized\_mutual\_info\_score(ward.labels\_, average.labels\_)}')

# Results

## Single and Complete

Chart

Description automatically generated

## Single and Ward

Chart

Description automatically generated

## Single and Average

Chart

Description automatically generated

## Complete and Ward

Chart, treemap chart

Description automatically generated

## Complete and Average

Chart

Description automatically generated

## Ward and Average

Chart

Description automatically generated

## NMI Scores

Text

Description automatically generated