IT 350

Assignment 4

Name: Niraj Nandish

Roll no.: 191IT234

Batch no.: 7

Link to Colab notebook - https://colab.research.google.com/drive/1r7tHEf3k3quMNWdz1dZiAWKD6wwD5UWL?usp=sharing (https://colab.research.google.com/drive/1r7tHEf3k3quMNWdz1dZiAWKD6wwD5UWL?usp=sharing)

1. Find the clusters in the given dataset based on the content similarity and image similarity using k-means clustering and hierarchical clustering methods.

In [1]:

```
!pip install pytesseract
!sudo apt install tesseract-ocr
```

```
Requirement already satisfied: pytesseract in /usr/local/lib/python3.7/dist-packages (0.3.9)
Requirement already satisfied: Pillow>=8.0.0 in /usr/local/lib/python3.7/dist-packages (from pytesse
ract) (9.1.0)
Requirement already satisfied: packaging>=21.3 in /usr/local/lib/python3.7/dist-packages (from pytes
seract) (21.3)
Requirement already satisfied: pyparsing!=3.0.5,>=2.0.2 in /usr/local/lib/python3.7/dist-packages (f
rom packaging>=21.3->pytesseract) (3.0.7)
Reading package lists... Done
Building dependency tree
Reading state information... Done
tesseract-ocr is already the newest version (4.00~git2288-10f4998a-2).
0 upgraded, 0 newly installed, 0 to remove and 39 not upgraded.
```

In [2]:

```
import pytesseract
from glob import glob
import pandas as pd
from PIL import Image
from pprint import pprint

from google.colab import drive
drive.mount('/content/drive')

images = glob("/content/drive/MyDrive/Assignment 4 Clustering/Q2_Clusters/Cluster_1/*")
pprint(images)
data = pd.DataFrame({"FileName": images})
imageText = []
for img in images:
    text = pytesseract.image_to_string(Image.open(img))
    imageText.append(text.replace("\n", ""))
data["Text"] = imageText
pprint(imageText)
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/
drive", force_remount=True).
['/content/drive/MyDrive/Assignment 4 '
 'Clustering/Q2_Clusters/Cluster_1/191IT253.2.jpeg',
 '/content/drive/MyDrive/Assignment 4 '
 'Clustering/Q2_Clusters/Cluster_1/191IT219.2.jpeg',
 '/content/drive/MyDrive/Assignment 4 '
 'Clustering/Q2_Clusters/Cluster_1/191IT229.2.jpeg',
 '/content/drive/MyDrive/Assignment 4 '
 'Clustering/Q2_Clusters/Cluster_1/191IT247.2.jpeg',
 '/content/drive/MyDrive/Assignment 4 '
 'Clustering/Q2_Clusters/Cluster_1/191IT209.2.jpg',
 '/content/drive/MyDrive/Assignment 4 '
 'Clustering/Q2_Clusters/Cluster_1/191IT134.2.jpeg',
 '/content/drive/MyDrive/Assignment 4 '
 'Clustering/Q2_Clusters/Cluster_1/191IT127.2',
 '/content/drive/MyDrive/Assignment 4 '
 'Clustering/Q2_Clusters/Cluster_1/191IT235.2.jpeg',
 '/content/drive/MyDrive/Assignment 4 '
 'Clustering/Q2_Clusters/Cluster_1/191IT128.2.jpeg',
 '/content/drive/MyDrive/Assignment 4 '
 'Clustering/Q2_Clusters/Cluster_1/191IT113.2.jpg',
 '/content/drive/MyDrive/Assignment 4 '
 'Clustering/Q2_Clusters/Cluster_1/191IT103.2.jpeg',
 '/content/drive/MyDrive/Assignment 4 '
 'Clustering/Q2_Clusters/Cluster_1/191IT220.2.jpeg',
 '/content/drive/MyDrive/Assignment 4 '
 'Clustering/Q2_Clusters/Cluster_1/191it257.2.jpg']
[' \x0c',
 ' \x0c',
 'be thee Wah amahinWO) 2 Unt wod ¢WOW) F2uta wod 6 eNO) = n42 med 6 '
 '}BlementIN= (Moan) vod !¢Bet (30 42) pod tCVSS Kad gMwai Wauahiee  '
 'watrteMirai) Wa wah hrhil)vToute tia Wa tard) \x0c',
 '  1 a hash. jumdioo hE) tush Saksee   \x0c',
 ' auceraenog mls. at Gedo 6alate ane) med €ace teva) ed &oh me eePT Janaki te '
 'prompt: ee fpaading te dhe pared eth fare PesrostecionsGemeding topier heb '
 '4p. fpilesing hashing alppvithmin the pesiatentioniceFe the rinhaah, thefe '
 'coneidowed?, Fast methGace chee | the shingle mathe con9 Sigrtee sie '
 'ait      i yan in many comsane of muting of difieectdhingle ef decoments the '
 'samehost selec  \x0c',
 'wnee aneweyQorper   if te3Fist moth whae diateroles. b) Tre Taccarcd '
 'Sialooty SpeeSwihay=it Yaeosy‘Jz =0-23Hs= 06g2) =0-c6r2/2 = 0-6632]g= 0 '
 '<F              si natare similarMoth functions ten that. wm Jon. tere,ore '
```

```
'the. for7D, deve is a high chon fshingles bebveen docament)) belter hash '
'functioneeu) mes i J g 3nto) = (@x+b)N js nor Y total yewsa Vv ave yandom '
'2 \x0c',
'@@) Gfveur , Hirer har functtons ,AU) = A141 mod GAny = 3142 mod 6AL3) (A) = '
'5IA+Q Mode     L 0 Oo \\ |5 oO oO OoBulttal maka:5S$ 5. S3 Bon -aay Co Oo Oo '
'CA ;In | oo oO on /a0A 0 ~~ CA oNRevd O §, |S Se | SyAy, CO | OO |4, | 00 oa '
'oO ol.©Arey A¢| oO oO) “ilO° (©0 0 | {fRow 4Row 5 a a5      A2ABpars: '
'(Taccard senrlloxtly )Sentlaatty aunong re ARN   Sin (SarSy) | Sin(S5» Su) | '
'Bon(Sir$) | Comm (Sa ) | Month Ba| r Oo 2| 3 2| %.®vrsbuiich© Owe COM -o '
'Normallxs Hire data l¢fore chu tithstU det eu Hing chiar Wreundartes betwen '
'the cChutirsQ) Woe can ur hetevaschioal ee a@) Here, from te jaceard '
'Stwilarttter, we can Ace that(S,,S,) ; (S,,Sy) 5 iS, Si), (Sq) Ss) ale, Ae '
'StvQUar(® So, we can Praprove Hu fata waaiean spear " reductug threAMutlow '
'Plows ar tt oecwans Hee edad CF. \x0c',
'PapergrigDate: |                  \x0c',
'bP  \x0c',
'Pet | Gaatang Cnet neei ? oN a SShtite3 / ate _6) | /9ifTn yung a ;Nf Saas '
'ee _    ete = =                                     Qarcany Titer da '
'Voltvpfens. classmate. ‘(9) 17 173 _ C0 ! ee t Page ictu tk dees” WP a2 end '
'veeAd. ankles brik 4g - efader_¢Thas      \x0c',
' \x0c',
'\x0c',
'                                              oe.a [ofitiaits/e a           '
'| Nome - Vijoy ThiRoll Mo- 911708 FO32] Casares, Sepauabein—Isab.ls 1000 '
'drag 1a | Ate _oliscacte Custoenda— Gaia Lh| shaving —Chanaetesteog —Patt 2 '
'The obserby Ou betasI fae Qolucansfa DOrmW Nelial/ delat ais a     \x0c']
```

In [3]:

```
from sklearn.feature_extraction.text import CountVectorizer
vectorizer = CountVectorizer()
X = vectorizer.fit_transform(data["Text"])
```

In [4]:

```
# K-means clustering
from sklearn.cluster import KMeans
data["Kmeans"] = KMeans(n_clusters=5, random_state=0).fit_predict(X.toarray())
```

In [5]:
```
data
```

Out[5]:

	FileName	Text	Kmeans
0	/content/drive/MyDrive/Assignment 4 Clustering...		0
1	/content/drive/MyDrive/Assignment 4 Clustering...		0
2	/content/drive/MyDrive/Assignment 4 Clustering...	be thee Wah amahinWO) 2 Unt wod ¢WOW) F2uta wo...	0
3	/content/drive/MyDrive/Assignment 4 Clustering...	1 a hash. jumdioo hE) tush Saksee	0
4	/content/drive/MyDrive/Assignment 4 Clustering...	auceraenog mls. at Gedo 6alate ane) med €ace ...	2
5	/content/drive/MyDrive/Assignment 4 Clustering...	wnee aneweyQorper if te3Fist moth whae diate...	4
6	/content/drive/MyDrive/Assignment 4 Clustering...	@@) Gfveur , Hirer har functtons ,AU) = A141 m...	1
7	/content/drive/MyDrive/Assignment 4 Clustering...	PapergrigDate: |	0
8	/content/drive/MyDrive/Assignment 4 Clustering...	bP	0
9	/content/drive/MyDrive/Assignment 4 Clustering...	Pet | Gaatang Cnet neei ? oN a SShtite3 / ate ...	3
10	/content/drive/MyDrive/Assignment 4 Clustering...		0
11	/content/drive/MyDrive/Assignment 4 Clustering...		0
12	/content/drive/MyDrive/Assignment 4 Clustering...	...	0

In [6]:
```
# Agglomerative clustering
from sklearn.cluster import AgglomerativeClustering
data["Agglomerative"] = AgglomerativeClustering(n_clusters=4).fit_predict(X.toarray())
```

In [7]:

```
data
```

Out[7]:

	FileName	Text	Kmeans	Agglomerative
0	/content/drive/MyDrive/Assignment 4 Clustering...		0	0
1	/content/drive/MyDrive/Assignment 4 Clustering...		0	0
2	/content/drive/MyDrive/Assignment 4 Clustering...	be thee Wah amahinWO) 2 Unt wod ¢WOW) F2uta wo...	0	0
3	/content/drive/MyDrive/Assignment 4 Clustering...	1 a hash. jumdioo hE) tush Saksee	0	0
4	/content/drive/MyDrive/Assignment 4 Clustering...	auceraenog mls. at Gedo 6alate ane) med €ace ...	2	2
5	/content/drive/MyDrive/Assignment 4 Clustering...	wnee aneweyQorper if te3Fist moth whae diate...	4	3
6	/content/drive/MyDrive/Assignment 4 Clustering...	@@) Gfveur , Hirer har functtons ,AU) = A141 m...	1	1
7	/content/drive/MyDrive/Assignment 4 Clustering...	PapergrigDate: |	0	0
8	/content/drive/MyDrive/Assignment 4 Clustering...	bP	0	0
9	/content/drive/MyDrive/Assignment 4 Clustering...	Pet | Gaatang Cnet neei ? oN a SShtite3 / ate ...	3	0
10	/content/drive/MyDrive/Assignment 4 Clustering...		0	0
11	/content/drive/MyDrive/Assignment 4 Clustering...		0	0
12	/content/drive/MyDrive/Assignment 4 Clustering...	...	0	0

2. Plot t-SNE visualization for derived clusters.

In [8]:

```
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.manifold import TSNE
import numpy as np
X_embedded = TSNE(n_components=2, init='pca').fit_transform(X.toarray().astype("float"))
X_pd = pd.DataFrame(X_embedded)
```

```
/usr/local/lib/python3.7/dist-packages/sklearn/manifold/_t_sne.py:793: FutureWarning: The default le
arning rate in TSNE will change from 200.0 to 'auto' in 1.2.
  FutureWarning,
/usr/local/lib/python3.7/dist-packages/sklearn/manifold/_t_sne.py:986: FutureWarning: The PCA initia
lization in TSNE will change to have the standard deviation of PC1 equal to 1e-4 in 1.2. This will e
nsure better convergence.
  FutureWarning,
```

In [9]:

```
X_pd
```

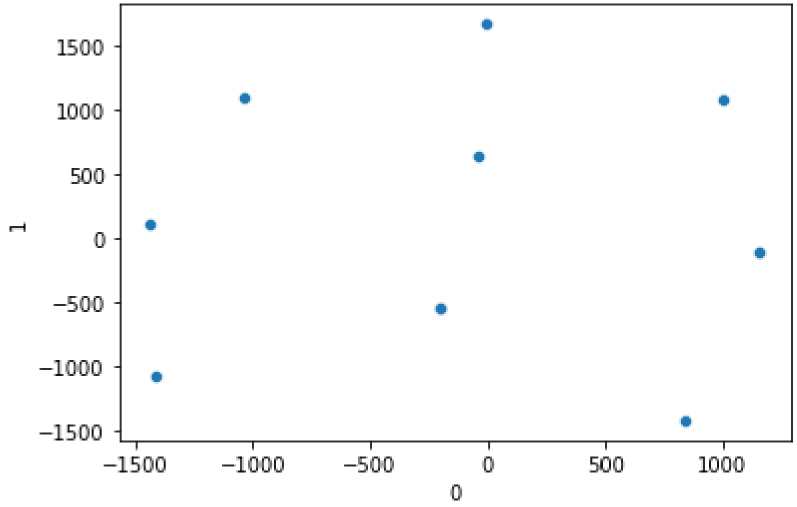
Out[9]:

	0	1
0	-204.667938	-552.070923
1	-204.667938	-552.070923
2	-1438.736450	118.308029
3	-1037.713257	1094.138794
4	-37.757725	650.830383
5	1003.618408	1080.078369
6	840.637329	-1423.135132
7	1159.793213	-102.525795
8	1159.793213	-102.525795
9	-8.582541	1675.383179
10	-204.667938	-552.070923
11	-204.667938	-552.070923
12	-1415.424561	-1086.230591

```
In [10]: sns.scatterplot(X_pd[0], X_pd[1])
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: Pass the following
variables as keyword args: x, y. From version 0.12, the only valid positional argument will be `data
`, and passing other arguments without an explicit keyword will result in an error or misinterpretat
ion.
  FutureWarning
```

```
Out[10]: <matplotlib.axes._subplots.AxesSubplot at 0x7febadd4c750>
```



3. Evaluate the clusters that are obtained using appropriate methods.

Performed clustering of answer sheet using pytesseract to recognize OCR text from images, and then used the extracted text to perform clustering. I used 2 types of clustering techniques K-means and Agglomerative clustering. Agglomerative clustering shows a wider distribution of papers into clusters than K-means. Since OCR detection is extremely poor and fails to identify text, I proceeded in checking similarity using image features.