Spring B565 HW 4
Shardul Dabhane
Thursday, March 24, 2022

1. Textbook problems (30 points total)

(a) Tan et al v2, 5.10, Q18. (10 points):
Suppose we have market basket data consisting of 100 transactions and 20 items. Assume the support for item a is 25%, the support for item b is 90% and the support for itemset {a, b} is 20%. Let the support and confidence thresholds be 10% and 60%, respectively.

i. Compute the confidence of the association rule {a}→{b}. Is the rule interesting according to the confidence measure?
Answer:
Confidence ({a} → {b}) = Number of times b appear in transactions that contain a.
Confidence ({a} → {b})= Support({a} ∪ {b})/Support({a})
= 0.2/0.25=80%
The rule is interesting as it exceeds the confidence threshold of 60%.

ii. Compute the interest measure for the association pattern {a, b}. Describe the nature of the relationship between item a and item b in terms of the interest measure.
Answer:
Interest(a,b)=Support({a} ∪ {b})/(Support({a})*Support({b}))
= 0.2/(0.25 × 0.9) = 0.889
Items a and b are negatively correlated according to interest measure, as the interest is less than 1.

iii. What conclusions can you draw from the results of parts (a) and (b)?
Answer:
We can conclude that high confidence rules may not necessarily have high interest/ be interesting.

iv. Prove that if the confidence of the rule {a} → {b} is less than the support of {b}, then:

I. c({$\bar{a}$} → {b}) > c({a} → {b}),

II. c({$\bar{a}$} → {b}) > s({b})

where c(·) denote the rule confidence and s(·) denote the support of an itemset.

Answer:

Let c({a} → {b}) = P({a, b})/P({a}).

P({a, b})/P({a})< P({b}).

Therefore, P({a})P({b}) > P({a, b}).

Also,

c({$\bar{a}$} → {b})=P({a, b}) / P({$\bar{a}$})=P({b}) - P({a, b}) / 1 - P({a})

I. c({$\bar{a}$} → {b}) - c({a} → {b})

= (P({b}) - P({a, b}) / 1- P({a})) - (P({a, b}) / P({a}))

= P({a})P({b}) - P({a, b}) / P({a})(1 - P({a}))

This result will be positive as P({a})P({b}) > P({a, b}).

Therefore, c({a} → {b}) > c({a} → {b}).

II. c({a} → {b}) - s({b})

= (P({b}) - P({a, b}) / (1-P({a})) - (P({b}))

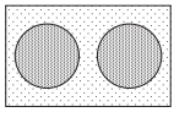
= P({a})P({b}) - P({a, b}) / 1 − P({a})

This result will be positive as P({a})P({b}) > P({a, b}).

Therefore, c({a} → {b}) > s({b}).

(b) Tan et al v2, 7.7, Q5, Q6 and Q7. (20 points)

Q5. Identify the clusters in Figure 7.36 using the center-, contiguity-, and density-based definitions. Also indicate the number of clusters for each case and give a brief indication of your reasoning. Note that darkness or the number of dots indicates density. If it helps, assume center-based means K-means, contiguity based means single link, and density-based means DBSCAN.

(a) 

i. **Center-based:** 2 clusters. The rectangular region in the given figure will be split into two equal parts. There will be noise in the two clusters.

ii. **Contiguity-based:** 1 cluster because the two circles in the figure will be joined by noise.

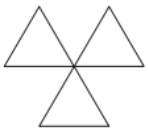
iii. **Density-based:** 2 clusters, one for each circle. Noise will be eliminated from the clusters.

(b) 

i. **Center-based:** 1 cluster as both the rings are concentric and will be in a single cluster.

ii. **Contiguity-based:** 2 clusters as the 2 circular rings will be joined by noise.

iii. **Density-based:** 2 clusters, one for each ring.

(c) 

i. **Center-based:** 1 cluster, with centroid at the intersection of these triangles.

ii. **Contiguity-based:** 1 cluster because all triangles are equilateral and their intersection point would be the center of the cluster.

iii. **Density-based:** 1 cluster, because the 3 triangles touch and the density at the point of intersection is higher than the interior regions of all triangles

.

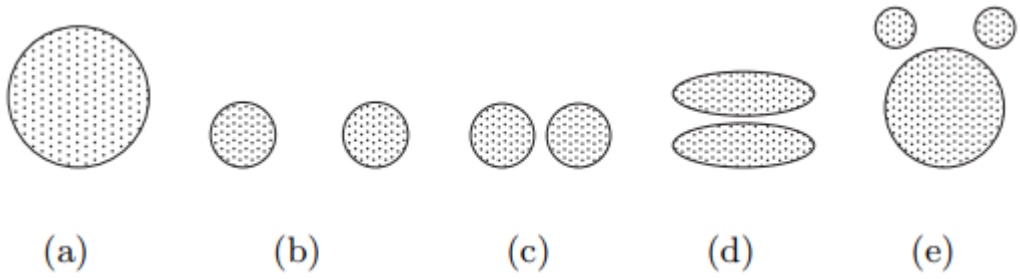
(d) 

i. **Center-based:** 2 clusters. According to the diagram, the points will be split into two.

ii. **Contiguity-based:** Overall 5 clusters. In first half of the image, there would be 3 clusters as two of the lines intertwine, and in the second half, there would 2 clusters as 3 lines Intertwine, giving us 5 clusters.

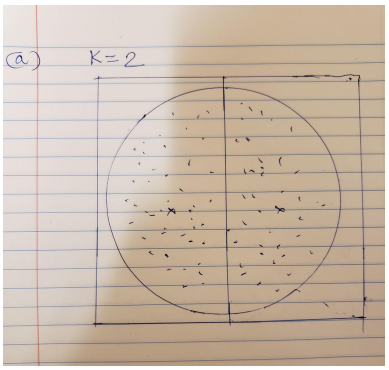
iii. **Density-based:** 2 clusters, as each region is highly dense and the gap between the regions is the low-density region.

Q6. For the following sets of two-dimensional points, (1) provide a sketch of how they would be split into clusters by K-means for the given number of clusters and (2) indicate approximately where the resulting centroids would be. Assume that we are using the squared error objective function. If you think that there is more than one possible solution, then please indicate whether each solution is a global or local minimum. Note that the label of each diagram in Figure 7.37 matches the corresponding part of this question, e.g., Figure 7.37(a) goes with part (a).



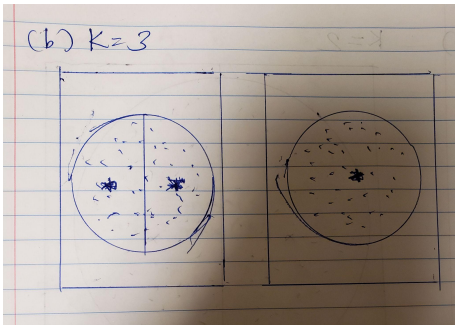
(a) K = 2. Assuming that the points are uniformly distributed in the circle, how many possible ways are there (in theory) to partition the points into two clusters? What can you say about the positions of the two centroids? (Again, you don't need to provide exact centroid locations, just a qualitative description.)

Answer:



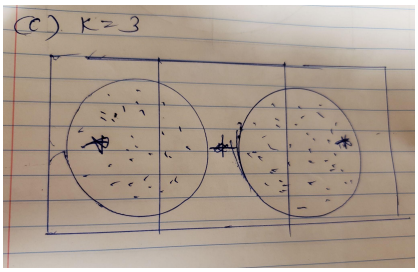
In theory, there are infinite ways to partition the points into 2 clusters. We can draw a diameter and it will divide the points into 2 clusters. The centroids will lie on the perpendicular bisector of the line dividing the circle into 2 clusters and they will be symmetrically positioned. **All the solutions will have the same global minimum.**

(b) K = 3. The distance between the edges of the circles is slightly greater than the radii of the circles.
Answer:



The first circle will have 2 clusters and the second circle will have the 3rd cluster. We will get this solution as the distance between the edges of the circles is greater than the radius of either circle. We could also have the other circle split and the partitioning line could be different in every case. **All solutions will have the same global minimum.**

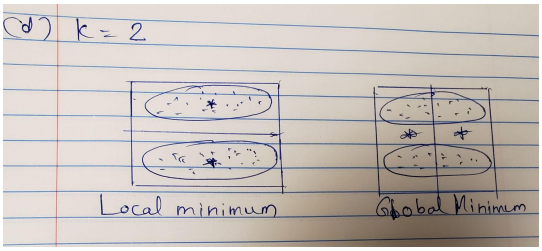
(c) K = 3. The distance between the edges of the circles is much less than the radii of the circles.
Answer:



As the distance between the circles is much less than the radii of the circles, we will have equal partitioning of the overall region. The centroid for the middle cluster will be equidistant from the centers of both circles. **The global minimum will be the same for all cases**.

(d) K=2.

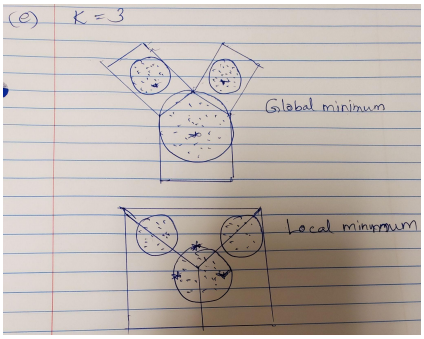
Answer:



There will be a separate local and global minimum in this case. In the local minimum case, the centroids will be located inside the elliptical regions as it is where the density of points is high. Each cluster will contain points from only one elliptical region.
For the global minimum solution, we will have centroids such that each cluster will have points from both elliptical regions.

(e) K = 3. Hint: Use the symmetry of the situation and remember that we are looking for a rough sketch of what the result would be.

Answer:



For the global minimum, the two top clusters are enclosed in 2 boxes, while the third cluster is enclosed by the regions defined by a triangle and a rectangle. The centroids will be located in each circular region with data points in them.
The local minimum would consist of one of the clusters containing parts of all 3 circular regions and the second and third clusters having centroids in the largest circular region.

Q7. Suppose that for a data set

- there are m points and K clusters,
- half the points and clusters are in “more dense” regions,
- half the points and clusters are in “less dense” regions, and
- the two regions are well-separated from each other.

For the given data set, which of the following should occur in order to minimize the squared error when finding K clusters:

(a) Centroids should be equally distributed between more dense and less dense regions.

(b) More centroids should be allocated to the less dense region.

(c) More centroids should be allocated to the denser region.

Note: Do not get distracted by special cases or bring in factors other than density. However, if you feel the true answer is different from any given above, justify your response.

Answer:

If (a) happens, the square error would be large as the less dense regions will create centroids that are further apart, increasing the squared error.

If (b) happens, again the minimum squared error will increase due to the lopsided distribution of the centroids, as more dense regions, will lack the centroids for minimization of square error.

The correct answer is (c). In the case of (c), by allocating more centroids to dense regions, we will overall get a low value for square error as the region with the most points will have a low square error value, and the region with less density will have high square error but it won’t contribute much to the overall square error as there are fewer points in less dense regions.

2. Apply single and complete link hierarchical clustering algorithms to cluster four coronavirus genomes (with their distances shown in the table below). Show your calculations (step by step) and the dendrogram of the clustering results. (20 points)

genome	A	B	C	D
A	0	20	7	10
B		0	15	8
C			0	6
D				0

(a) Single Link Hierarchical Clustering:

The shortest distance from the above graph = C → D = 6

The distance between CD and other genomes is:

CD → A = min(A → C, A → D) = min(7,10) = 7

CD → B = min(B → D, B → C) = min(15,8) = 8

The new matrix is as follows,

genome	A	B	CD
A	0	20	7
B		0	8
CD			0

Iterating again,

The shortest distance in this matrix is A→CD=7.

ACD→B distance is 8.

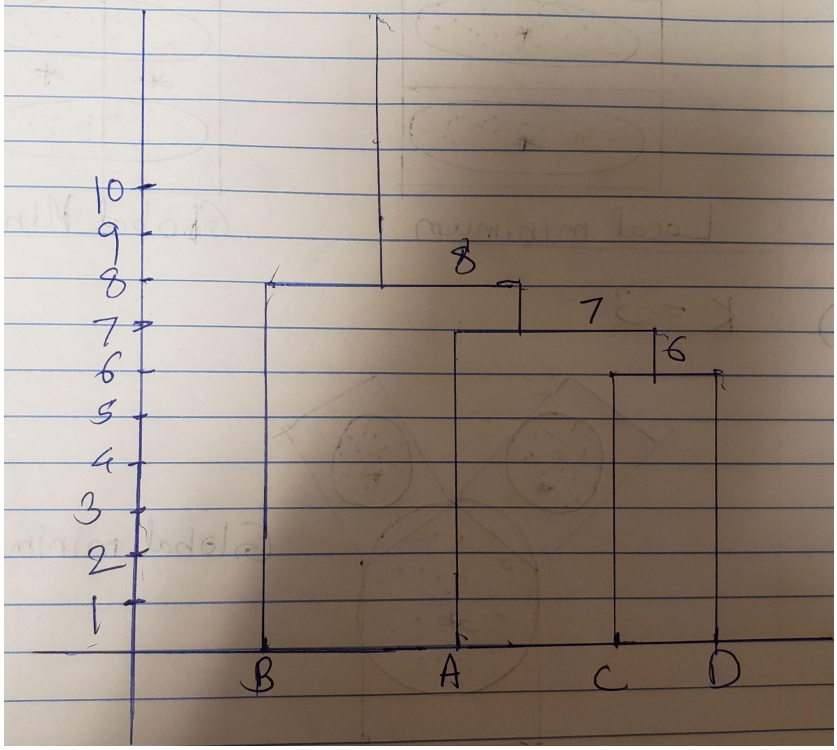
The new matrix generated would be,

genome	ACD	B
ACD	0	8
B		0

The final distance is 8.

The dendrogram for single-link hierarchical clustering for the genomes would be:

Single Link Hierarchical Clustering



(b) Complete Link Hierarchical Clustering:

The shortest distance from the above graph = C → D = 6

The distance between CD and other genomes is:

CD → A = max(A → C, A → D) = max(7,10) = 10

CD → B = max(B → D, B → C) = max(15,8) = 15

The new matrix would be,

genome	A	B	CD
A	0	20	10
B		0	15
CD			0

Iterating again,

The shortest distance from the above graph = A →CD = 10

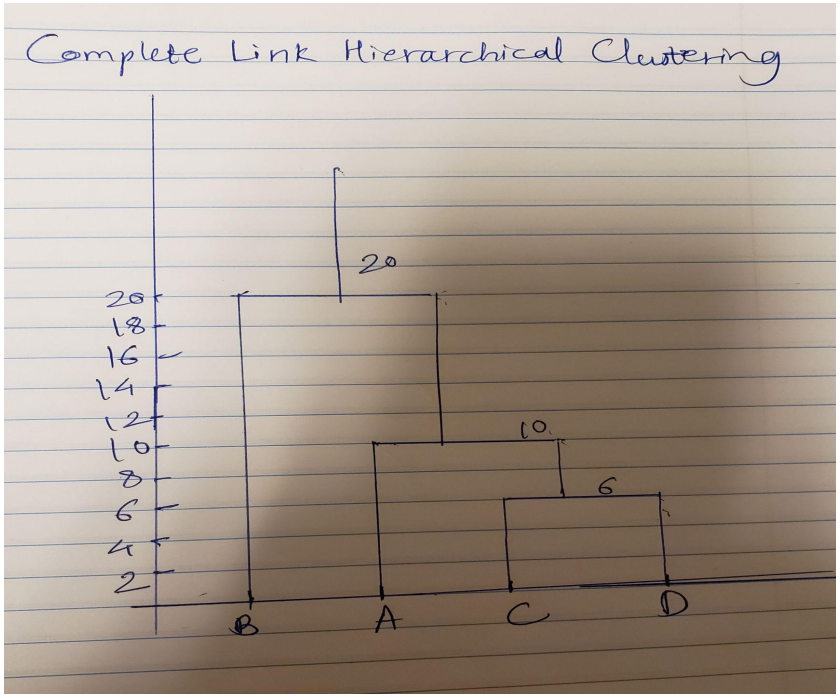
The distance between ACD and B is 20.

The new matrix would be,

genome	ACD	B
ACD	0	20
B		0

The final distance is 20.

The dendrogram for complete-link hierarchical clustering would be:



Question 3 of HW4

Shardul Dabhane

The professor enjoyed reading news that came out in late Feb on CNN, People, Science Daily and many other platforms, all about one study. The study (to be presented at the American Academy of Neurology's 74th Annual Meeting) found that long-term pet owners had higher cognitive scores than those in the same age group without pets. The professor (a) felt that some news read suspiciously similar to each other; and (b) worried that some of the news perhaps over interpreted the results from the study (e.g., one news mentioned that "Long-Term Pet Ownership Can Slow Cognitive Decline in Older Adults"). For this problem, you are going to do some reading and coding to help the professor out.

1. Collect related news and save them as text files (.txt). (10 pts)
2. Read the news and use the concepts that you have learned this semester to either help the professor back up her worry, or convince her that her worry is baseless. (10 pts)
3. Tokenize the news so each news can be represented as a binary vector (see ref code). You may try different values for the important parameters and see how that impact the downstream applications. Apply hierarchical clustering algorithms (min, average, and max) to cluster the news using their binary vectors. Summarize what you find about the relationship of the news based on the clustering results. Does your finding support the professor's claim that some news sound suspiciously similar to each other (just a qualitative description)? (30 points)

In [1]:

```
#import required modules

import numpy as np
import pandas as pd
import os
import matplotlib.pyplot as plt
from sklearn.feature_extraction.text import CountVectorizer
from scipy.cluster.hierarchy import dendrogram, linkage
%matplotlib inline
```

Collect related news and save them as text files (.txt). (10 pts)

Answer: For this problem, I have taken 14 articles about the study from publications like CNBC, Healthline, Insider, etc.

In [2]:

```
#Read all files from the folder

all_files = os.listdir("Folder_To_Read_Data/")
txt_files = filter(lambda x: x[-4:] == '.txt', all_files)
txt_files=list(txt_files)
print(txt_files)
```

```
['AAN.txt', 'CNBC.txt', 'CNN.txt', 'DailyPaws.txt', 'Healthline.txt', 'Insider.txt', 'MarthaStewart.
txt', 'MedPage.txt', 'MiamiHerald.txt', 'People.txt', 'ScienceDaily.txt', 'StudyFinds.txt', 'Univers
ityOfFlorida.txt', 'Yahoo.txt']
```

Read the news and use the concepts that you have learned this semester to either help the professor back up her worry, or convince her that her worry is baseless. (10 pts)

Answer:

The study was conducted by researchers at the University of Florida, University of Michigan and Virginia Commonwealth University. The study says that adults ages 50 or older who had owned any kind of pet for more than five years showed slower decline in verbal memory — being able to recall words, for example — over time compared to non-pet owners. While most publications ran with the headline, "Long-Term Pet Ownership Can Slow Cognitive Decline in Older Adults" or something similar, the University of Florida article about this study is more grounded and realistic.

(Link: https://news.ufl.edu/2022/02/pet-ownership-study/#:~:text=The%20researchers%20found%20that%20adults,compared%20to%20non%2Dpet%20owners (https://news.ufl.edu/2022/02/pet-ownership-study/#:~:text=The%20researchers%20found%20that%20adults,compared%20to%20non%2Dpet%20owners).)

I've included this article in my code as well. This article has a better headline:"Long-term pet ownership may help older adults retain cognitive skills". It also includes a quote from the lead author of the study, Jennifer Applebaum, a doctoral candidate in sociology and National Institutes of Health predoctoral Fellow at the University of Florida. The quote is, "We can't show that this is causal but it does show that pets could buffer or have a protective effect on older adults' cognition and we think it has to do with some of the mechanism related to stress buffering." This quote is only mentioned in 1 or 2 articles, leading to the authors of those articles concluding that owning pets will slow the cognitive decline process significantly, while the author mentions that there is no causal relationship between owning pets and slowing cognitive decline.

Another line in the article says "Applebaum said the researchers are not recommending pet ownership as a therapeutic intervention". This line is mentioned in only articles from CNN and MedPage, apart from the University of Florida article. Both these important lines indicate that further study needs to be done to conclude that owning pets will slower cognitive decline. Most of the articles do not contain this important information, so it is understandable that the professor worries that the news has overinterpreted the results.

We can use hierarchical clustering to show how pairs of articles are similar and include phrases similar to "Long-Term Pet Ownership Can Slow Cognitive Decline in Older Adults". From that, we will be able to see the similarity in the given articles.

Tokenize the news so each news can be represented as a binary vector (see ref code). You may try different values for the important parameters and see how that impact the downstream applications. Apply hierarchical clustering algorithms (min, average, and max) to cluster the news using their binary vectors. Summarize what you find about the relationship of the news based on the clustering results. Does your finding support the professor's claim that some news sound suspiciously similar to each other (just a qualitative description)? (30 points)

In [3]:

```
#Collect content from all news items into one array
corpus_for_content=[]
for txt in txt_files:
    file = open(txt, 'r',encoding="utf8")
    file_content=file.read()
    file_content=file_content.strip()
    corpus_for_content.append(file_content)
print(len(corpus_for_content))
```

```
14
```

In [4]:

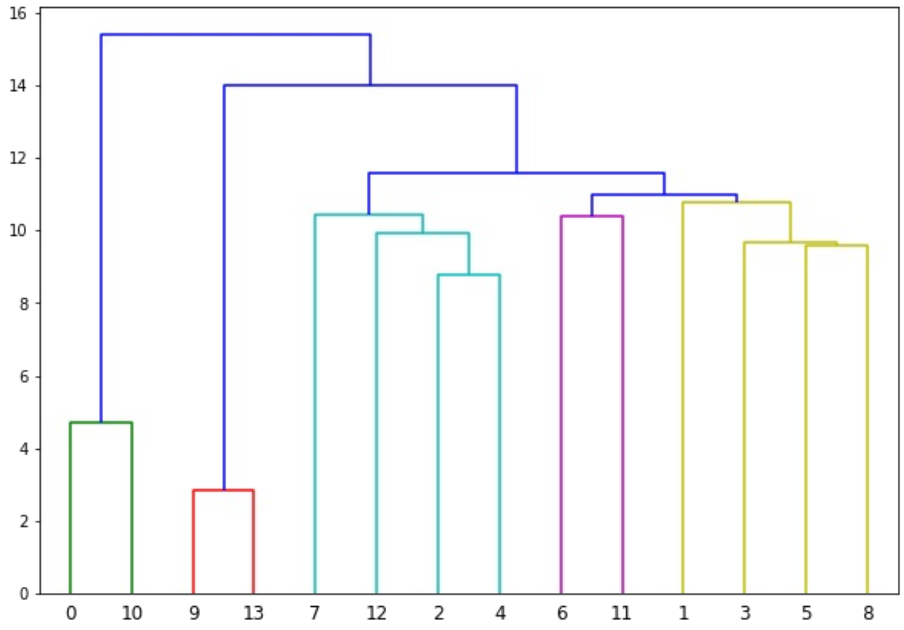
```
#Create binary vectors based on the data

vectorizer = CountVectorizer(stop_words='english', ngram_range=(3, 3), min_df=2)
X = vectorizer.fit_transform(corpus_for_content)
words = vectorizer.get_feature_names()
Y=X.toarray()
print(Y.shape)
print(Y)
```

```
(14, 602)
[[0 1 1 ... 1 1 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 ...
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [1 0 0 ... 1 0 1]]
```

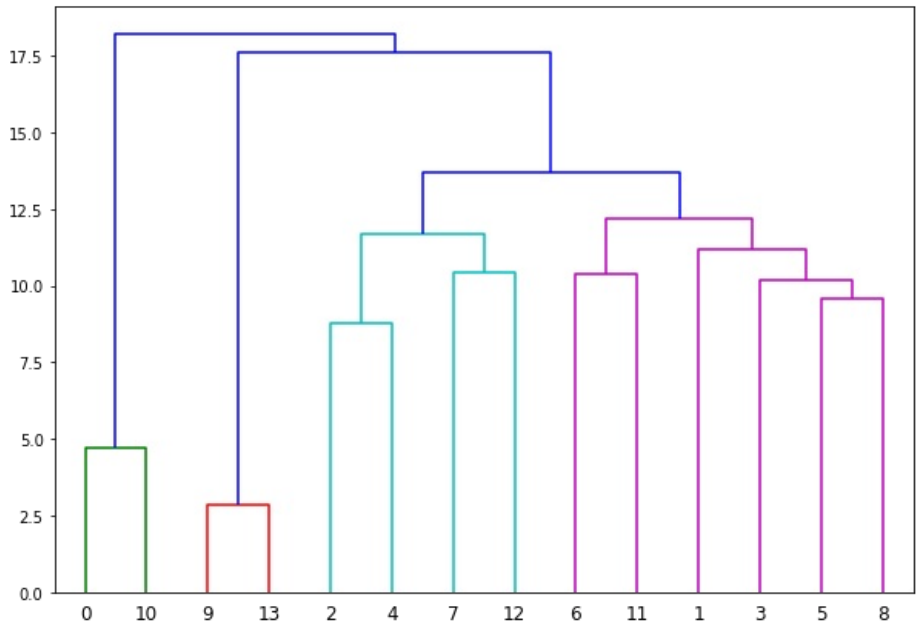
In [5]:

```
#min clustering algorithm
plt.figure(figsize=(10, 7))
dendrogram(linkage(Y, 'single'),orientation='top')
plt.show()
```



In [6]:

```
#average clustering algorithm
plt.figure(figsize=(10, 7))
dendrogram(linkage(Y, 'average'),orientation='top')
plt.show()
```



In [ ]:

```
#max clustering algorithm
plt.figure(figsize=(10, 7))
dendrogram(linkage(Y, 'complete'),orientation='top')
plt.show()
```

I ran the code with different values of ngram_range and min_df and the results were similar. Based on the above dendrograms, we can see that there is not much separation between the articles. Articles from People and Yahoo are really similar and have the least difference between them. The article on Yahoo is quoting the People article so it makes sense that these 2 will be clustered first. Then articles from AAN and Science Daily are the most similar. Then Martha Stewart and Study Finds are most similar, and so on. From the 3 dendrograms, we can see that article number 9 and 13 are the most similar, followed by articles 0 and 10. After them, articles 2 and 4 are the most similar. Articles 5 and 8, 6 and 11 and 7 and 12 are the remaining pairs. Articles 2, 4, 7 and 12 are fairly similar to each other. Articles 3,5,8,1,6,11 are similar to each other.Then articles 9,13,0 and 10 are similar to each other. The maximum distance between any 2 clusters is only just above 20. This indicates that the news is suspiciously similar. As mentioned in the answer for 2, the news is interpreting the results wrongly even though the actual study shows no causal link between slower cognitive decline and owning pets.The professor's findings are accurate in this case.

References:

[1] Introduction to Data Mining 2nd Edition By Tan, Steinbach, Kumar, Karpatne

[2] https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.complete.html (https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.complete.html)

[3] https://www.kaggle.com/code/morecoding/omicron-nlp (https://www.kaggle.com/code/morecoding/omicron-nlp)

[4] https://www.kaggle.com/code/morecoding/clustering-basic (https://www.kaggle.com/code/morecoding/clustering-basic)

[5] https://www.kaggle.com/code/morecoding/hierarchical-clustering/ (https://www.kaggle.com/code/morecoding/hierarchical-clustering/)

[6] https://stackoverflow.com/questions/35672809/how-to-read-a-list-of-txt-files-in-a-folder-in-python (https://stackoverflow.com/questions/35672809/how-to-read-a-list-of-txt-files-in-a-folder-in-python)