

## B565 (Spring 2022) - DATA MINING HOMEWORK 4

Name: Ameya Dalvi

Email: [abdalvi@iu.edu](mailto:abdalvi@iu.edu)

### 1. Textbook problems:

1. Suppose we have market basket data consisting of 100 transactions and 20 items. If 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.

- (a) Compute the confidence of the association rule  $\{a\} \rightarrow \{b\}$ . Is the rule interesting according to the confidence measure?

Confidence for association rule  $\{a\} \rightarrow \{b\}$  can be given as:

Confidence in a rule is calculated by dividing the probability of the items occurring together by the probability of the occurrence of the antecedent.

Antecedent:  $\{a\}$

Confidence for association rule  $\{a\} \rightarrow \{b\}$ :

$$\frac{P(\{a, b\})}{P(\{a\})}$$

$$\frac{P(\{a, b\})}{P(\{a\})} = \frac{\text{sup}(\{a, b\})}{\text{sup}(\{a\})} = \frac{20\%}{25\%} = \frac{0.2}{0.25} = 0.8 = 80\%$$

This rule is interesting as confidence measure for this rule is greater than the threshold.

**(b) 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.**

Interest measure or Interest Factor is also known as 'Lift'. Lift indicates the strength of a rule over the random co-occurrence of the antecedent and the consequent, given their individual support. It provides information about the improvement, the increase in the probability of the consequent given the antecedent. Ref: [Apriori](#)

Antecedent: {a}

Precedent: {b}

$$Lift(\{a, b\}) = \frac{sup(\{a, b\})}{sup(\{a\}) * sup(\{b\})}$$

$$\frac{sup(\{a, b\})}{sup(\{a\}) * sup(\{b\})} = \frac{20\%}{25\% * 90\%} = \frac{0.2}{0.25 * 0.9} = 0.89$$

As Lift for the association pattern {a, b} is less than 1, The items are said to be negatively correlated.

**(c) What conclusions can you draw from the results of parts (a) and (b)?**

From part (a) we can see that The rule/ itemset has a high confidence thus making it interesting, but in part (b) we identified the itemset to be negatively correlated. Hence by these observations, we can conclude that confidence metric alone cannot be used to determine if a rule is interesting or not. High confidence need not necessarily mean rule is interesting.

(d) Prove that if the confidence of the rule  $\{a\} \rightarrow \{b\}$  is less than the support of  $\{b\}$ , then:

- i.  $c(\{a'\} \rightarrow \{b\}) > c(\{a\} \rightarrow \{b\})$ ,
- ii.  $c(\{a'\} \rightarrow \{b\}) > s(\{b\})$ ,

where  $c(\cdot)$  denote the rule confidence and  $s(\cdot)$  denote the support of an itemset.

**Given:** confidence of the rule  $\{a\} \rightarrow \{b\}$  is less than the support of  $\{b\}$

$$\text{We know, } c(\{a, b\}) = c(\{a\} \rightarrow \{b\}) = \frac{P(\{a, b\})}{P(\{a\})}$$

$$\frac{P(\{a, b\})}{P(\{a\})} < \text{sup}(\{b\}) = \frac{P(\{a, b\})}{P(\{a\})} < P(\{b\})$$

$$\therefore P(\{a, b\}) < P(\{a\}) * P(\{b\}) \dots\dots (1)$$

$$\text{Now, } c(\{a', b\}) = c(\{a'\} \rightarrow \{b\}) = \frac{P(\{b\}) - P(\{a, b\})}{1 - P(\{a\})}$$

**To prove:**  $c(\{a'\} \rightarrow \{b\}) > c(\{a\} \rightarrow \{b\})$

$$\therefore c(\{a'\} \rightarrow \{b\}) - c(\{a\} \rightarrow \{b\}) > 0$$

$$\therefore \frac{P(\{b\}) - P(\{a, b\})}{1 - P(\{a\})} - \frac{P(\{a, b\})}{P(\{a\})} > 0$$

$$\therefore \frac{P(\{b\})P(\{a\}) - P(\{a, b\})P(\{a\}) - P(\{a, b\}) + P(\{a, b\})P(\{a\})}{(1 - P(\{a\}))P(\{a\})} > 0$$

$$\therefore \frac{P(\{b\})P(\{a\}) - P(\{a, b\})}{(1 - P(\{a\}))P(\{a\})} > 0$$

$$\therefore P(\{b\})P(\{a\}) - P(\{a, b\}) > 0 \text{ which is true as given in (1)}$$

**To prove:**  $c(\{a'\} \rightarrow \{b\}) > s(\{b\})$

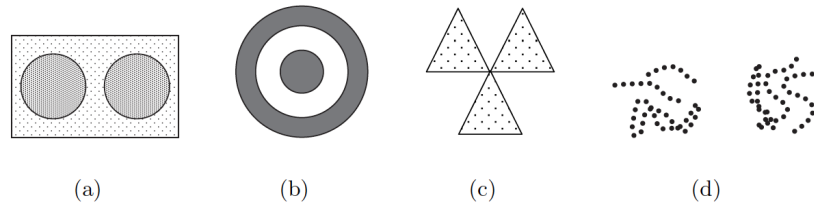
$$\therefore c(\{a'\} \rightarrow \{b\}) - s(\{b\}) > 0$$

$$\therefore \frac{P(\{b\}) - P(\{a, b\})}{1 - P(\{a\})} - P(\{b\}) > 0$$

$$\therefore \frac{P(\{b\}) - P(\{a, b\}) - P(\{b\}) + P(\{b\})P(\{a\})}{1 - P(\{a\})} > 0$$

$$\therefore P(\{b\})P(\{a\}) - P(\{a, b\}) > 0 \text{ which is true as given in (1)}$$

2. Identify the clusters in Figure 8.3 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) Center-based:**

If we use centroids, then the rectangular area can be split into two parts to form two clusters having two circles each.

**Contiguity-based:**

As the distance between the circle edges doesn't look greater than the radii of the two circles, 1 cluster could be formed joining the two circles.

**Density-based:**

2 clusters could be formed as cluster lines are separated by a region low density .

**(b) Center-based:**

One centroid at the center of the two concentric circles would create one single cluster including both the circles.

**Contiguity-based:**

One cluster can be formed joining the inner circle with the outer ring.

**Density-based:**

2 clusters could be formed as the two concentric circles have a higher density than the low-density region in between them.

**(c) Center-based:**

We could have 3 centroids for each of the triangular regions forming 3 clusters.

**Contiguity-based:**

One cluster could be formed as the three triangular regions all three join at one single point.

**Density-based:**

3 clusters could be formed as the three triangular regions seem to have equal density.

**(d) Center-based:**

We can have two centroids essentially splitting the two recognizable group of lines forming two clusters.

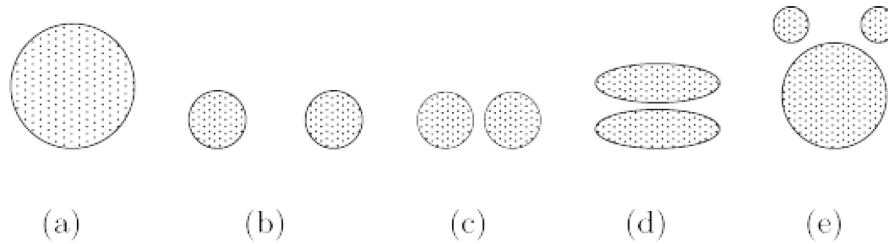
**Contiguity-based:**

We can split the set of lines into 4 clusters

**Density-based:**

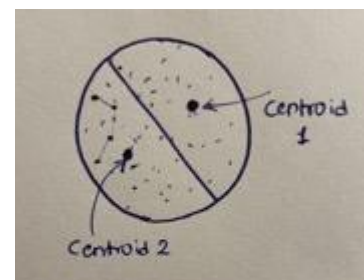
3 clusters could be formed as the three triangular regions seem to have equal density.

3. 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 8.4 matches the corresponding part of this question, e.g., Figure 8.4(a) goes with part (a).



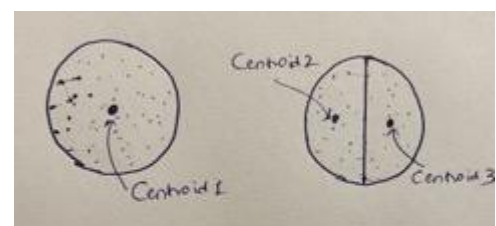
**(a)  $K = 2$**

As we are asked to form two clusters, this can only be done by splitting the circle into two halves (semi-circles) and selecting the centroids as the average point of all the points in those two semi-circles.



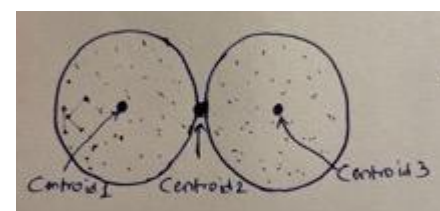
**(b)  $K = 3$**

As the distance between the edges of the circle is greater than radii of the circles, centroids could be formed as two centroids splitting one circle into two clusters and the other circle being one whole cluster. These cluster formations are possible only because the edges of the circle is greater than the radii and there's a very less chance of any centroid from either of the circles moving towards the other one.



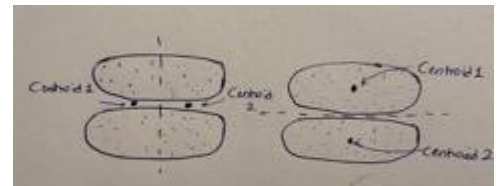
**(c)  $K = 3$**

The distance between the edges of the circles is much less than the radii of the circles. Hence clusters formed by the three centroids would also thus be equidistant.



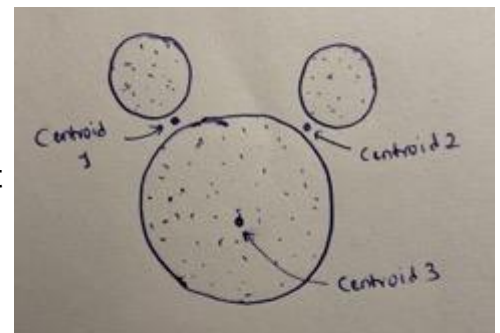
**(d)  $K = 2$**

There are many ways to split the two ellipses into two clusters, horizontally, vertically, diagonally etc. Im showing two possible cluster formations, horizontal and vertical. For Horizontal division, the cluster centroids would lie at the center of the two ellipses, for vertical division, the cluster centroids would lie in between the two halves of the two ellipses as shown in the figure.



**(e)  $K = 3$**

The data mass is more concentrated in the larger circle as compared to the other two smaller circles, hence the 3 cluster centroids could be arranged as given in the figure where two centroid would right outside the two small circles, leaning towards the data mass that is concentrated in the bigger circle, and the third centroid being the center of the bigger circle.



**4. 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.

**A: (c)** If the squared error is to be reduced, less dense regions require more centroids.



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.

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

### 1. Single Link Hierarchical Clustering

**Q-2) SINGLE LINK HIERARCHICAL CLUSTERING**

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

minimum distance = 6 = dist(C,D)

Cluster 1 = {C,D}

To update the genome table -

$$\min(\text{dist}((C,D), A))$$

$$= \min(\text{dist}(C,A), \text{dist}(D,A))$$

$$= \min(7, 10)$$

$$= 7$$

$$\min(\text{dist}((C,D), B))$$


$$= \min(\text{dist}(C,B), \text{dist}(D,B))$$

$$= \min(15, 8)$$

$$= 8$$

Updated Table -

genome	A	B	C,D
A	0	20	7
B		0	8
C,D			0



genome	A	B	C,D
A	0	20	7
B		0	8
C,D			0

minimum distance = 7 = dist((C,D), A)

Cluster 2 = {C,D, A}

To update the genome table -

$$\min(\text{dist}((C,D,A), B))$$

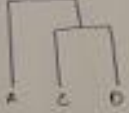
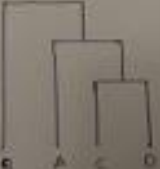
$$= \min(\text{dist}((C,D), B), \text{dist}(A,B))$$

$$= \min(8, 20)$$

$$= 8$$

Updated Table -

genome	B	C,D, A
B	0	8
C,D, A		0

## 2. Complete Link Hierarchical Clustering

**COMPLETE LINK HIERARCHICAL CLUSTERING**

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

minimum distance = 6 = dist (C,D)

Cluster 1 = (C,D)

To update the genome table:-

$\max(\text{dist}(C,D), A)$

$= \max(\text{dist}(C,A), \text{dist}(D,A))$

$= \max(7, 10)$

$= 10$

$\max(\text{dist}(C,D), B)$


$= \max(\text{dist}(C,B), \text{dist}(D,B))$

$= \max(15, 8)$

$= 15$

Updated Table:-

genome	A	B	C,D
A	0	20	10
B		0	15
C,D			0



genome	A	B	C,D
A	0	20	10
B		0	15
C,D			0

minimum distance = 10 = dist ((C,D), A)

Cluster 2 = C,D,A

To update the genome table:-

$\max(\text{dist}((C,D), A), B)$

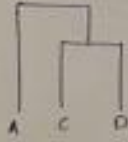
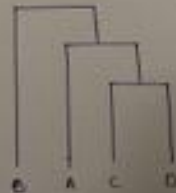
$= \max(\text{dist}((C,D), B), \text{dist}(A,B))$

$= \max(15, 20)$

$= 20$

Updated Table:-

genome	A	B	C,D,A
A	0	20	
B		0	15
C,D,A			0

**3. 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).**

I have used 8 new articles, from these websites:

People.com: [Long-Term Pet Ownership Can Slow Cognitive Decline in Older Adults, New Study Finds](#)

CNN: [Pets can boost your brain power, study says](#)

ScienceDaily: [Do pets have a positive effect on your brain health?](#)

Yahoo: [Long-Term Pet Ownership Can Slow Cognitive Decline in Older Adults, New Study Finds](#)

abc7.com: [Pets can boost your brain power, study says](#)

AARP.com: [Pet Ownership May Delay Cognitive Decline in Older Adults](#)

KSL.com: [Pets can boost your brain power, study says](#)

Audacy: [Pets can boost your brain power, study says](#)

HOLA.com: [WHY OWNING A PET HAS A POSITIVE EFFECT ON YOUR BRAIN HEALTH, INCLUDING MEMORY LOSS AND COGNITIVE SKILLS](#)

**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.**

After reading the news I noticed that many identical and similar articles exists online, this claim can be backed by the experimental analysis that is done in the below sections. Apart from that, for web scraping the news articles, I have made use of the [newspaper](#) python module that reads the webpage and converts the content of the webpage into sections like Title, article text and keywords. The keywords section can also be used to determine the similarity between these documents, if we apply association rule mining on these keywords, we can determine which keywords exists as a group of keywords and if those same keywords are identified by the newspaper module while scraping. Hence professor's worry isn't baseless.

## Article scraping example text file using newspaper:

Title:

Pet Ownership May Delay Cognitive Decline in Older Adults

Article Text:

Getty Images

Science has long declared that pets help people de-stress and stick to a healthy

A new study suggests that pet ownership is even better for older people than previous studies. The new data will show that pet ownership may delay cognitive decline in adults over 65. The new data will

"Prior studies have suggested that the human-animal bond may have health benefits," says Dr. David Reis, a professor of psychology at the University of Michigan in Ann Arbor, who oversaw the study, in a press release. "Our results suggest pet

Using data from the Health and Retirement Study, an examination of 1,369 Medicare beneficiaries, researchers measured cognitive skills at the start of the study. A total of 53 percent owned pets, and

For that Medicare study, researchers measured cognitive function through various tests ranging from 0 to 27 based on how well they performed.

Over six years, cognitive scores decreased at a slower rate in pet owners, especially those with dogs. The study found that pet owners had a cognitive score that was 1.2 points higher compared to non-pet owners. Those cognitive drop

"I definitely think having a pet makes a difference," says caregiver Lorie Martai

Article Keywords:

decline  
cognitive  
adults  
data  
delay  
university  
longterm  
study  
ownership  
owners  
older  
pet

### 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)?

Tokenizing each news article using CountVectorizer.

```
from sklearn.feature_extraction.text import CountVectorizer
#corpus: 8 news articles
corpus = ["The University of Michigan's Health and Retirement Study found adults around the age of 65 that owned pets for five or longer had higher cognitive scores than those in the same age group who did not own a pet. The study found that owning a pet, like a dog or cat, especially for five years or longer, may be linked to slower cognitive decline in older adults, according to a preliminary study released by the university. A new study reveals that owning pets for more than five years can slow cognitive decline in older people. A team at the University of Michigan studied over 1,300 people over the age of 65. The University of Michigan's Health and Retirement Study found adults around the age of 65 that owned pets for five or longer had higher cognitive scores than those in the same age group who did not own a pet. Science has long declared that pets help people de-stress and stick to a healthy routine. They can also decrease depression and physical pain and, in many cases, provide a sense of purpose. Having a long-term pet companion may delay memory loss and other kinds of cognitive decline, a new study has found. Pet ownership was especially beneficial for working adults. Turns out having a pet companion may delay cognitive decline or memory loss according to a new study, CNN reports. To our knowledge, our study is the first to consider the impact of pet ownership on cognitive decline. Did you know that having a furry friend by your side can improve your mental health and memory loss? Researchers at the University of Michigan revealed that having pets can help with cognitive decline." ]

vectorizer = CountVectorizer()
X = vectorizer.fit_transform(corpus)
words = vectorizer.get_feature_names()
print(f"#words {len(words)} {words}")

#words 575 ['1369', '2022', '23', '24', '26', '27', '300', '32', '369', '50', '53', '60', '60s', '65', '74th', '88', 'abolishing', 'about', 'academy', 'according', 'account', 'activity', 'added', 'adding', 'addition', 'additionally', 'address', 'adult', 'adults', 'affect', 'against', 'age', 'aging', 'all', 'also', 'although', 'alzheimer', 'american', 'americans', 'among', 'an', 'analysis', 'analyzed', 'and', 'animal', 'animals', 'ann', 'annual', 'another', 'answer', 'any', 'anyone', 'anything', 'applebaum', 'april', 'arbor', 'are', 'aren', 'arund', 'as', 'assessed', 'associate', 'associations', 'at', 'atlantic', 'author', 'average', 'bad', 'based', 'be', 'because', 'been', 'beginning', 'being', 'believe', 'beneficial', 'beneficiaries', 'benefit', 'benefited', 'benefits', 'best', 'better', 'between', 'biased', 'birds', 'black', 'blood', 'boarding', 'bond', 'bonded', 'boost', 'bover', 'brain', 'braley', 'buffering', 'by', 'can', 'candidate', 'care', 'cared', 'caregiver', 'cases', 'cat', 'category', 'cats', 'cause', 'center', 'chronic', 'clinic', 'clinical', 'cnn', 'cognition', 'cognitive', 'college', 'color', 'combination', 'combines', 'common', 'communities', 'community', 'companion', 'companionship', 'compared', 'components', 'composite', 'concluded', 'conditions', 'confirm', 'consider', 'considering', 'consisted', 'continues', 'core', 'cortisol', 'cost', 'could', 'counting', 'crisis', 'data', 'de', 'declared', 'de line', 'decrease', 'decreased', 'decreasing', 'defined', 'definitely', 'delay', 'delayed', 'delaying', 'delays', 'dementia', 'depressed', 'depression', 'devastating', 'develop', 'did', 'difference', 'direct', 'director', 'disease', 'do', 'doctoral', 'doctors', 'does', 'dog', 'dogs', 'dr', 'drops', 'due', 'duration', 'duty', 'each', 'early', 'educated', 'education', 'effect', 'effects', 'email', 'engagement', 'especially', 'estimated', 'ethnicity', 'even', 'evidence', 'exactly', 'examination', 'excluded', 'experts', 'explain', 'explained', 'explaining', 'explore', 'factors', 'february', 'fees', 'fell', 'fellow', 'final', 'findings', 'finds', 'first', 'fish', 'five', 'florida', 'followed', 'for', 'former', 'foster', 'found', 'f
```

Converting the tokenized news vectors to binary vectors

```
+ Code + Markdown
X[X>0]=1 #Converting the tokenized news vectors to binary vectors

print(f"vector for first news\n{X.toarray()[0]}")

vector for first news
[1 0 0 0 0 1 1 1 0 0 1 0 0 1 1 1 0 0 1 0 0 1 1 0 0 0 0 0 1 1 1 1 1 1 1 0 0
 1 0 1 1 0 0 1 1 0 1 1 1 0 0 0 0 0 1 1 1 0 1 1 0 0 1 1 1 1 0 1 1 0 0 1 0 0
 1 0 0 0 0 1 0 1 0 0 0 0 0 1 1 0 1 0 0 0 1 1 1 0 1 0 0 0 0 0 0 0 0 0 1 0 0
 0 0 0 1 1 0 0 0 0 0 0 1 0 1 0 1 0 0 1 0 1 0 0 0 0 0 1 1 0 1 0 0 1 0 1 1 1
 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 1 0 1 0 0 1 0 1 0 0 1 0 0 0 1
 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 1 1 0 0 1 0 0 1 0 0 1 0 0 1 1 0 1 0 0 0
 0 0 0 1 1 0 0 0 1 0 1 0 0 1 0 0 0 1 1 0 0 1 1 0 0 1 0 0 1 0 1 0 0 1 0
 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 1 1 0 1
 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 1 0 0 1 0 0 0 0 1 0 1 0 1 0 0 0 0
 0 0 0 1 0 1 0 1 1 1 0 0 0 0 1 1 1 1 0 0 0 0 0 1 1 1 0 0 1 0 0 1 0 1 1 1 0
 0 0 0 1 0 0 0 1 1 1 1 0 1 0 0 0 1 0 0 1 1 1 0 1 0 0 0 0 0 0 0 0 1 0 0
 0 0 1 1 0 0 0 0 0 1 1 1 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 0 1 0 0 0
 0 1 1 0 0 0 0 0 0 1 1 1 0 0 0 0 0 0 1 0 0 1 0 0 1 1 1 1 0 1 0 0 0 0 0 0
 0 1 0 0 0 1 0 1 1 1 1 1 0 0 0 1 0 0 0 0 1 0 0 1 0 1 1 1 1 0 0 1 0 0 1
 0 0 0 1 1 0 1 0 1 0 1 0 0 0 0 0 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1
 0 1 0 1 1 1 0 0 0 1 1 0 1 0 0 0 0 1 1 0 0 0 1]
```

**In the image below, we can see documents 3 and 0 are exactly identical:**

```
print(X.toarray()[3])
```

[illegible]

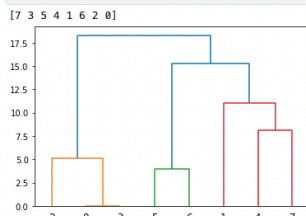
+ Code + Markdown

```
print(X.toarray()[3] - X.toarray()[0]) #Articles 0 and 3 are exactly identical, indicating similarity
```

[illegible]

## Hierarchical Clustering:

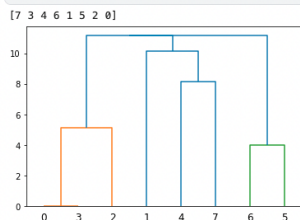
```
clustering2 = AgglomerativeClustering(distance_threshold=0, n_clusters=None, linkage='complete', affinity='euclidean')
clustering2.fit(X)
print(clustering2.labels_)
plot_dendrogram(clustering2, truncate_mode='level', p=3)
```



+ Code + Markdown

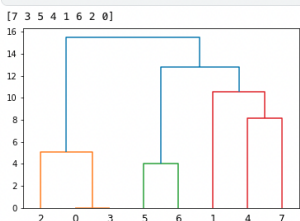
Max Hierarchical Clustering using Euclidean Distance Metric

```
clustering3 = AgglomerativeClustering(distance_threshold=0, n_clusters=None, linkage='single', affinity='euclidean')
clustering3.fit(X)
print(clustering3.labels_)
plot_dendrogram(clustering3, truncate_mode='level', p=3)
```



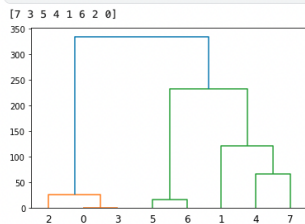
### Min Hierarchical Clustering using Euclidean Distance Metric

```
clustering4 = AgglomerativeClustering(distance_threshold=0, n_clusters=None, linkage='average', affinity='euclidean')
clustering4.fit(X)
print(clustering4.labels_)
plot_dendrogram(clustering4, truncate_mode='level', p=3)
```



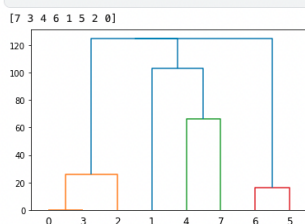
Average Hierarchical Clustering using Euclidean Distance Metric

```
clustering5 = AgglomerativeClustering(distance_threshold=0, n_clusters=None, linkage='complete', affinity='manhattan')
clustering5.fit(X)
print(clustering5.labels_)
plot_dendrogram(clustering5, truncate_mode='level', p=3)
```



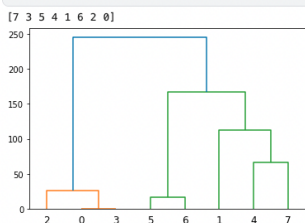
Max Hierarchical Clustering using Manhattan Distance Metric

```
clustering6 = AgglomerativeClustering(distance_threshold=0, n_clusters=None, linkage='single', affinity='manhattan')
clustering6.fit(X)
print(clustering6.labels_)
plot_dendrogram(clustering6, truncate_mode='level', p=3)
```



Min Hierarchical Clustering using Manhattan Distance Metric

```
clustering7 = AgglomerativeClustering(distance_threshold=0, n_clusters=None, linkage='average', affinity='manhattan')
clustering7.fit(X)
print(clustering7.labels_)
plot_dendrogram(clustering7, truncate_mode='level', p=3)
```



Average Hierarchical Clustering using Manhattan Distance Metric

## Summary:

The hierarchical clustering of news documents reveals a lot about how distinct news pieces are connected to one another, whether by repeating words, forming identical phrases, or having fully identical articles.

For example, articles 0 and 3 from People and Yahoo are exactly same, therefore we can see that they are in the same cluster in the clustering dendrograms and the distance between them is 0. Articles 5 and 6 also change somewhat since the distance between them is 4.

As a result, we can see that there are many publications on the internet with the same or almost identical information, indicating that my findings back up the professor's allegation.