**MSITM Unstructured Data Analytics Final Exam, Fall 2022**

**Total points: 100**

**Please submit your answers on Canvas by 3:15 p.m. on Dec 9, 2022 (late submissions will not be accepted)**

**Name \_\_\_Rithu Anand Krishnan \_\_\_**

**Please read all instructions carefully before answering the questions**

1. **All answers should be typed and not hand-written. The only exceptions involve questions where you may need to write math symbols or draw a diagram. In that case you can write (legibly) by hand, take a picture, and include it in the word or other file. Hand-drawn diagrams are fine as long as they are legible.**
2. **Unlike all other tasks in this course, where collaboration was encouraged, this exam is a strictly individual task. Do not discuss the questions and/or answers with a class- or team-mate (or anyone for that matter), for that would constitute a clear violation of the University honor code. Such cases are required to be reported to the Office of the Dean of Students. So once again, the exam format is open notes, Internet access, but with NO collaboration.**
3. **Please submit a single file – Word, pdf or Excel containing your answers and calculations. If you choose to submit an Excel file, create a worksheet for each question. Write your name inside the file for proper identification. If you are submitting a Word or pdf file and have Excel or Google sheets calculations, you can embed the Excel file or take a screenshot and paste in the Word file.**
4. **I have taken care to describe each problem in detail. I or the TA cannot provide any further guidance in solving the problems and will not answer any questions related to this exam. You have to interpret the questions and state any (reasonable) assumptions you make.**
5. **You can provide your answers in this Word file itself.**
6. **If you have any clarification questions, please post them here (but do check if a similar question has already been posted and answered):**

<https://docs.google.com/document/d/1Yv2tmnHtktUir2UwnZxjcPsSpCi1fR4xS7SzIV8pE84/edit?usp=sharing>

**I will check the document every 15 mins until noon.**

1. A common pre-processing technique in text analytics is to remove stopwords (e.g., “a”, “an”, “the”, etc.) from text. However, your friend, an expert in unstructured data analytics mentioned: “It is better to use TF-IDF scores instead of removing stopwords in a classification problem.” Do you agree with this statement? Justify your position with two examples. (15 points)

**Check one of the two choices below with X.**

**~~[ ] Yes, I agree~~ [ X ] No, I don’t agree**

**Justification:**

The correct pre-processing of any data is very specific to what we are trying to achieve.

This question is regarding classification. As we know, stop words occur very commonly and do not have the predictive power of separation. What makes a document different from the others are the words which are rare and unique to that particular document. On the other hand, TF-IDF provides us with more information and gives more value to words that are mentioned less often.

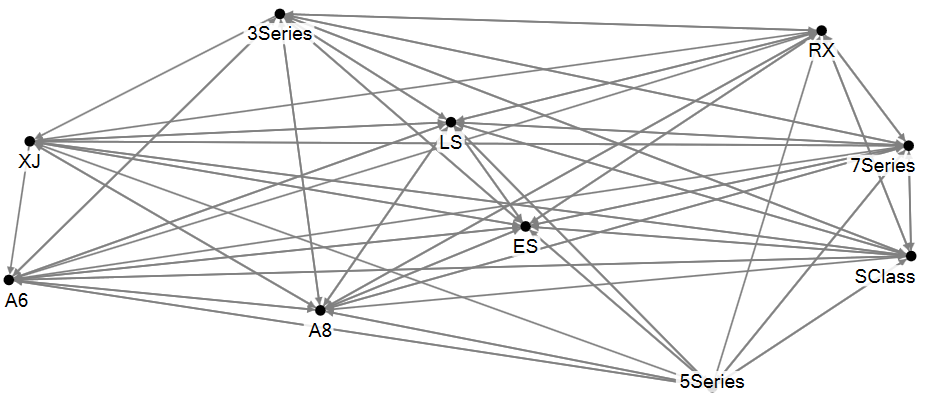
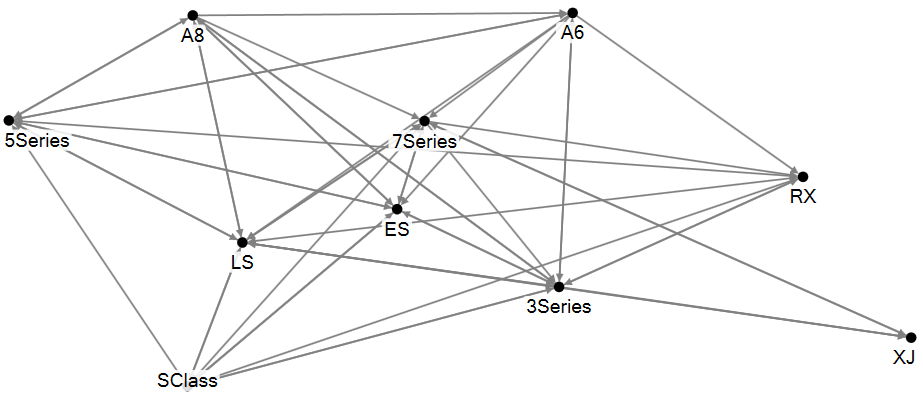
Let's dive into this deeper with an example:

Consider a hotel chain. Words such as room or hotel are likely to appear in every review compared to words such as "breathtaking" or "exceptional service". Similarly, when going through a document specific to one particular author, say Shakespeare, there is a mention of another author who may be less famous. In this case, this rare mention of the new author may go unnoticed with the word frequency approach because it doesn't appear immensely, but that one word would help us identify something unique about the document.

Further, let's note that sentiment analysis is more likely to be affected by stop words based on a high frequency. Luckily, classification problems usually don't need to stop words as you can still preserve the general concept of text even without stop words in it. Therefore, removing stop words is a better approach for classification problems.

Ultimately, one should only remove stop words if one thinks about the impact of these words on the problem they are trying to solve.

1. Consider two product preference networks shown below involving 10 products. If you calculated two sets of unweighted PageRank scores from the two networks A and B, which set would most likely show a higher correlation with sales data? **Why?** Do not actually calculate PageRank scores; instead answer this question conceptually. Unweighted PageRank scores ignore the weights on the arrows representing the strength of user preferences. (15 points)

Preference network A Preference network B

**Check one box below and justify your response:**

**[ X ] PageRanks from network A will have higher correlation with sales than those from network B.**

**~~[ ] PageRanks from network B will have higher correlation with sales than those from network A.~~**

**Justification:**

The diagram clearly depicts that network A is densely populated regarding edges between nodes compared to network B.

Let’s see how unweighted pageranks are calculated, based on the number of incoming relationships and the significance of the related source nodes, the PageRank algorithm calculates the importance of each node within the network.

If one were to calculate the unweighted PageRank score for both networks, it is likely that network A has more correlation with sales data. Higher the edges, the higher the sensitivity.As Network A is fully connected. At the same time, network B will produce low correlation values due to its uncomplicated structure. As we can see, it has fewer edges when compared to network A.

1. In a comparative analysis of smart watches, you extracted *N* messages from a smart watch forum where people discuss three products: Apple Watch, Fitbit Versa and Movado Connect (call this data set A). To boost the total amount of data, you also extracted an additional *N* messages posted on an Apple Watch forum, where every post mentions the Apple Watch, and where some (but not all) posts co-mention the other products (data set B). You want to calculate Lift(AppleWatch, battery) and Lift(MovadoConnect, battery) with data set A, and also with data set A+B (merging the two data sets). For simplicity assume (i) the proportion of posts mentioning **battery** is the same for data sets A and B, (ii) the proportion of MovadoConnect posts which also mention battery is the same for data sets A and B, (iii) Lift(AppleWatch, battery) > 1 for data set A.

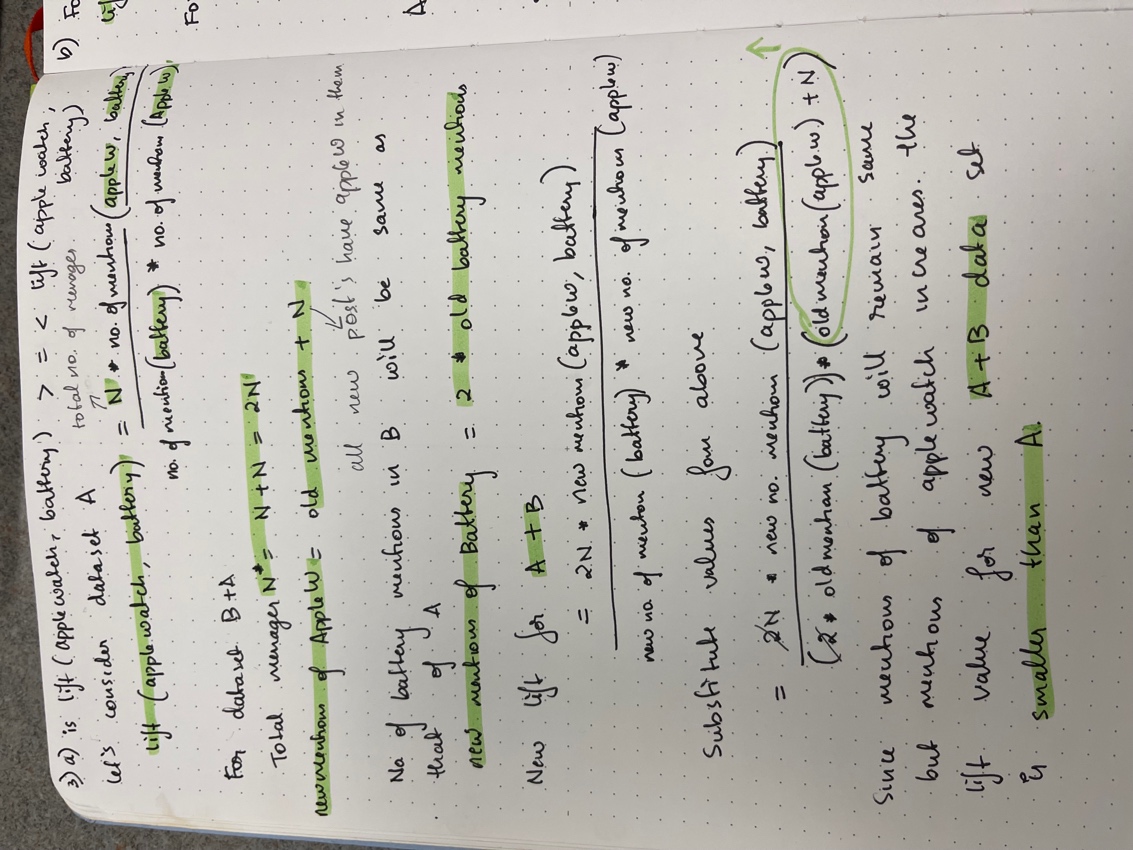
Is Lift(Applewatch, battery), calculated from data set A, **GREATER THAN, EQUAL TO, OR LESS THAN (Choose one)** Lift(Applewatch, battery) calculated using the combined data set A+B? Justify your response. You can use a numerical example, but must justify using logical arguments. (10 points)

**Make a choice below before you provide justification**

**Lift of Dataset A is GREATER THAN lift of Data set A + B**

**[ X ]** **GREATER THAN ~~[ ] EQUAL TO~~  ~~[ ] LESS THAN~~**

**Justification:**



Is Lift(Movadoconnect, battery), calculated from data set A, **GREATER THAN, EQUAL TO, OR LESS THAN (Circle one)** Lift(Movadoconnect, battery) using the combined data set A+B? Justify your response. You can use a numerical example, but must justify using logical arguments. (10 points)

**Make a choice below before you provide justification**

**~~[ ]~~****~~GREATER THAN~~ [ X] EQUAL TO ~~[ ] LESS THAN~~**

**Justification:**

Text, letter

Description automatically generated with medium confidence

1. “It is possible to guess reasonably accurately whether a post on social media is coming from a **spammer** from his/her network centrality metrics. “ Do you agree? Justify your response. (10 points)

~~Yes [ ]~~ No [ X ]

**Justification**

I don’t believe that it is easier to predict or guess whether a post is from a spammer or not solely based on his network centrality value.

Let’s look at different ways centrality value can be derived. One way would be degree centrality, a measurement of the degree of the vertex, which tells us how active a person is on social media or whether they are the most famous person based on high-degree vertices. This could mean that the person has various friends on social media or more.

A spammer is very clever, and he/she can create various fake accounts on social media to boost their degree centrality score by increasing or decreasing their network centrality metrics however needed. Plus not everyone of spammer friends may ping him but spammer is most likely to ping everyone. This falls into the in-degree and out-degree specifics. But then again, this information alone can not be used to identify the spammer.

**Answer EITHER 5a and 5b OR question 6 (Do not answer both)**

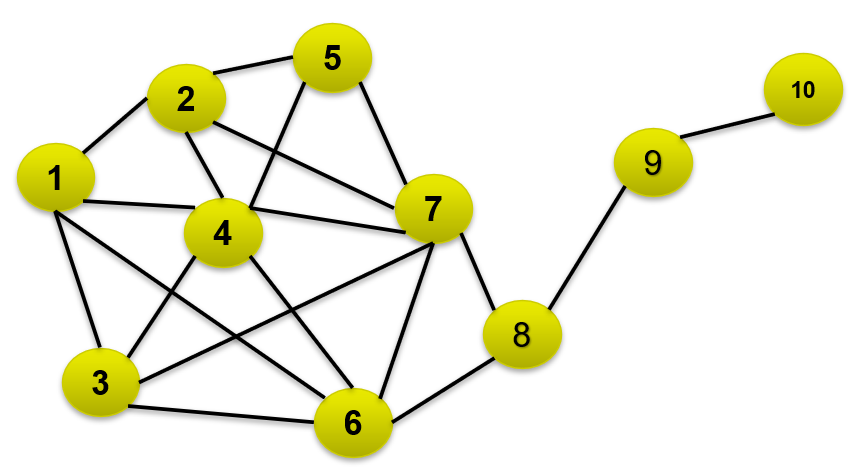
5a. In the network shown below, which link will be cut first in a community detection exercise using the Girvan-Newman algorithm? Justify your response. (6 points)

Let’s see how Girvan-Newman algorithm is different from others? In Girvan-newman algorithm, one must consider the betweenness centrality of the link instead of the node itself. This algorithm is used to cut down links based on the centrality value and form clusters or smaller networks, we keep repeating these steps until desired output is achieved. Based on this information, let’s focus on the network below. As we can see pictorially, the link between 8 and 9 seem to be the one with highest betweenness centrality. All the point on the right side of 9 will have to go via this link to reach all the points on the left side of 8, same goes for the other side. This shows that 8-9 link is the one with maximum betweenness centrality value.

For example:

If 1 should go to 10, then it must pass 8-9 link. Similarly, if 10 should go to 4, then the path it would take would be 10->9->8->7->4.

5b. Calculate the betweenness centrality of the link that will be cut first. Show your calculations. Hint: As noted in class, the betweenness of a link is calculated the same way as that of a node. (9 points)



The link is 8-9 and the centrality is 112.

Please find the calculations below.

Calendar

Description automatically generated with low confidence

A picture containing calendar

Description automatically generated

**If you have answered 5a and 5b, DO NOT answer question 6**

6. A hundred documents have to be divided into **three** clusters. There are **three** **classes** of documents -- say, positive, negative and neutral. Assume that the accuracy of classification of each class is important. Construct a numerical example that demonstrates the **superiority** of the *entropy* measure over *purity* for clustering. You can make any assumptions about the actual number of positive, negative and neutral documents (but they must add up to 100). Use the definitions of entropy and purity noted in the PowerPoint slides (where  refers to the number of documents in cluster *r* that belong to the most frequent or dominant class *i* in this cluster). The actual content (words, length, etc.) of the documents do not matter in this problem. Show detailed calculations of entropy and purity for the example you construct. (15 points)

**Answer ONE out of questions 7, 8, and 9 (do NOT answer more than one question).**

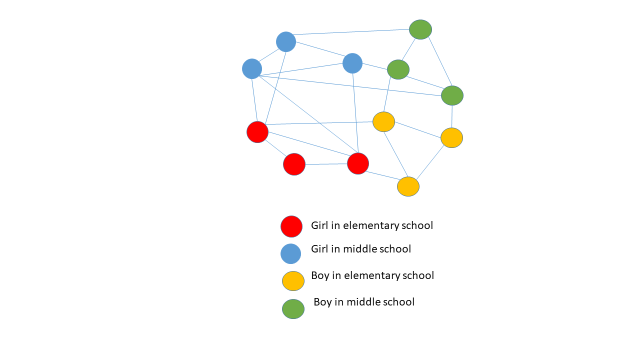
7. Consider the friendship network below (a separate ppt file with the network diagram is posted on Canvas for your convenience) between boys and girls in an elementary and a middle school (for your convenience, the PowerPoint file is posted in the Exam folder as well).

Red: Girl in an elementary school

Blue: Girl in a middle school

Orange: Boy in an elementary school

Green: Boy in a middle school



From this data, what test(s) of homophily can be performed? Show detailed calculations. (25 points).

8. Best Cruises (BC) recently ran into major problems with its ships. In a cruise forum, where folks discuss BC and its rival Royal Cruises (RC), a post may mention only BC, only RC or both. BC and RC were mentioned together in 8k posts. Further, BC was mentioned in 16k posts. RC found itself in 12k posts.

A post may express one of the following sentiments: (i) a positive sentiment about a cruise line, (ii) a negative sentiment about a cruise line, (iii) positive about both, (v) negative about both, (vi) no sentiment on either cruise line (e.g., just a fact like ticket price being mentioned). Assume for **simplicity** that there are NO posts that mentions one positively and the other negatively. BC got 7k negative posts. There were 5k negative posts that **only** mentioned BC. There were 2k negative posts that **only** mentioned RC. The two companies were mentioned together in a positive manner in 2k posts. 2k positive posts mentioned RC **only**. There were a total of 7k positive posts.

Based on the above numbers, extract **all** relevant information about BC and RC using **appropriate** **lifts**. What can you say about consumer perceptions of the two brands? Don’t just say “consumers think positively about x and negatively about y”; provide as much discussion and insights as possible, preferably in a table showing lifts and implications. Show all calculations. (25 points)

**Note:** All the data required for lift analysis are mentioned in the problem statement (i.e., nothing is missing here).

Please note that all the calculations are below.

Based on the lift calculations: It can be said that BC and RC don’t really have any lift values greater than 1. This goes to show that neither of these companies have a strong association with negative or positive posts. This could either be a good thing or a bad thing based on the company goals.

BC and RC have a lift value of 0.833, this goes to show that they don’t really seem to be mentioned together so often. Further, RC and positive posts lift has higher value (0.9523 almost 1) when compared to BC and positive posts(0.8928). This means that people seem to feel better about RC when compared to BC.

When it comes to negative posts and lift. The lift between BC and negative posts is (0.9722 almost close to 1), looks like not only is BC not getting any positive mentions but in fact it has higher negative mentions so higher lift. The management team of BC should really take a closer look into the posts and resolve issues if any.

The lift between RC and negative posts is 0.7407, it is relatively low. Even if some customers aren’t satisfied with RC, there doesn’t seem to be too many of them. Any business will have a few unsatisfied customers, looks like RC has a very few and a lot less than BC.

Calendar

Description automatically generated

A picture containing calendar

Description automatically generated

9. Consider two documents *d*1 and *d*2 represented by term weights as follows:

*di* = (*wi*,1, *wi*,2,…*wi*,*n*) where *i* ϵ {1, 2}. Now consider a document retrieval problem with a query expressed as a vector *r* = (*wr*,1, *wr*,2,…*wr*,*n*). Suppose *r* is closer to *d*1 than to *d*2 according to the Euclidean distance measure. Show that under a certain condition (which can always be achieved with **all** documents), cosine similarity will also lead to the same result (i.e., *r* will also be closer to *d*1 than to *d*2 according to cosine similarity). **Important:** You must show your analysis algebraically without assuming any numeric data. How can the above condition be achieved? Show and prove it algebraically. (25 points)