**Text Analytics Exam (F2019)**

**Date: Dec 10, 2019 Time: 19:00 - 22:30 hours (submit to Canvas)**

**Maximum Points: 100**

**Name: \_\_\_\_Ryan Harty\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

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

1. **All answers should be typed and not hand-written. If you have math symbols or equations, you can write (legibly) by hand, take pictures and include them in your answer 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 group-mate (or anyone for that matter), for that would constitute a clear violation of the University honor code. Such cases will be reported to the Office of the Dean of Students.**
3. **Please submit a single file – Word, pdf or Excel file containing your answers and main results of 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 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 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. A common pre-processing technique in text analytics is to remove stopwords from text. However, is it better to use TF-IDF scores instead of removing stopwords in a classification problem? Justify your response. (10 points).

**[X] TF-IDF is better [ ] TF-IDF is NOT better (check one)**

**Justification: When you use tf-idf it automatically decreases words with high frequencies. This makes it so you do not need to remove stop words as any words with high frequencies will have low weights and will therefore be less important than words with low frequencies.**

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? You must show the result mathematically and not using a numeric example. (10 points)

**[ ] GREATER THAN [ ] EQUAL TO [ X] LESS THAN (check one)**

**Lift = P(A,B) / (P(A)\*P(B)) = (N \* #(A,B)) / (#A \* #B)**

**Since the mentions of battery is the same in data set B, the numerator will be twice as large in the A+B data set opposed to the A data set. But the mentions of Apple Watch in data set A+B will be more than twice as large as that of data set A. As a result the lift denominator for data set A+B when looking at apple watches will grow faster than the numerator and the lift will be less than that of data set A for apple watch.**

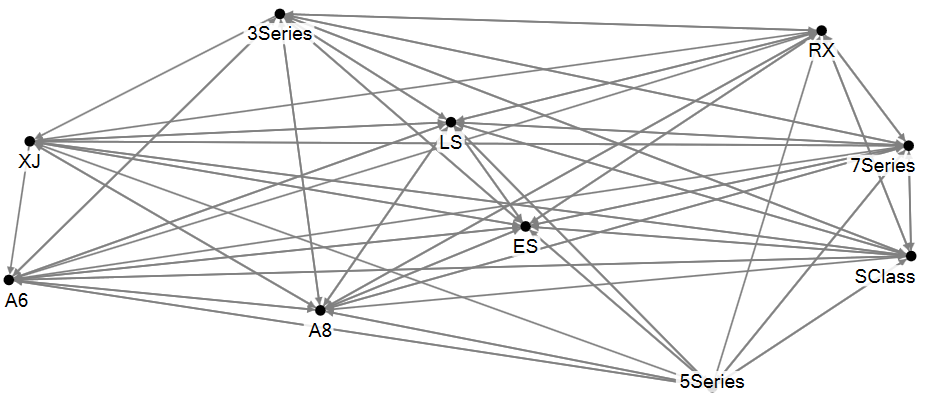
Is Lift(MovadoConnect, battery), calculated from data set A, **GREATER THAN, EQUAL TO, OR LESS THAN (Choose one)** Lift(MovadoConnect, battery) using the combined data set A+B? You must show the result mathematically and not using a numeric example. (10 points)

**[ ] GREATER THAN [ X] EQUAL TO [ ] LESS THAN (check one)**

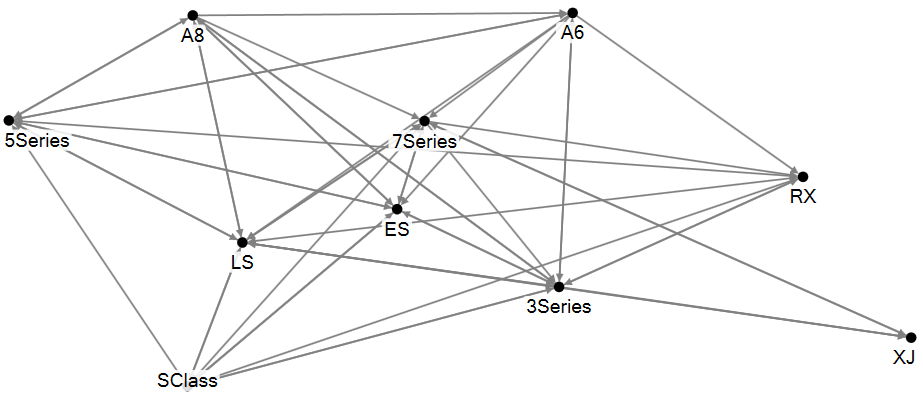
Since the proportion of MovadoConnect posts which also mention battery is the same for data sets A and B, the lift will be the same for both data set A and data set A+B.

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. (10 points)

**PageRank scores are higher when they are dense. When referring pages are referred to by many other pages, the effect is stronger and the score is higher. As a result the PageRanks froms network A will have a higher correlation with sales than that of network B.**



**Preference network A**



**Preference network B**

4. 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. (20 points)

**Documents will be closer in Euclidean distance when they are of similar size. Documents will be closer in cosine similarity when they have similar content. As a result, document r will be close to d1 when the size and contents are more similar than that of d2.**

**Euclidean distance: d(d1,r) ((w1,1-wr,1)^2 … (w1,n-wr,n)^2)^1/2 > d(d2,r) ((w2,1-wr,1)^2 … (w2,n-wr,n)^2)^1/2**

**&**

**Cosine: cos(d1,r) < cos(d2,r)**

5. 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). You will need two sets of clusters, say, set A with clusters A1, A2, A3, and set B with clusters B1, B2, B3, where A1 + A2 + A3 = B1 + B2 + B3 = 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. (20 points)

6. 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. (20 points)

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

