MIS 381 User Generated Content Analytics Take-home Final, Fall 2018

Handed out: Dec 7, 2018, Date due: On Canvas by 11:59 p.m. on Dec 11. 2018

Name: Sanchit Singhal (ss84657)

Please read all instructions carefully before answering the questions

A. **All answers should be typed and not hand-written. Hand-drawn diagrams are fine as long as they are legible.**

B. **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.**

C. **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.**

D. **I have taken care to describe each problem in detail, and have also provided hints where appropriate. 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.**

E. **You can provide your answers in this Word file itself.**

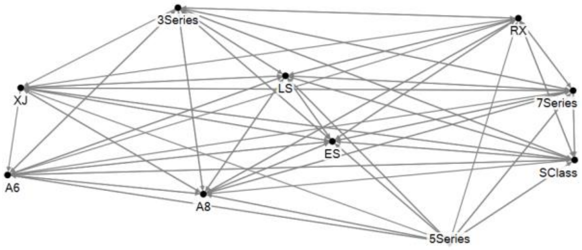
1. **A common pre-processing technique in text analytics is to remove stopwords (e.g., "a", "an", "the", etc.) from text. However, your friend, an expert in user generated content analytics mentioned: "It is better to use TF-IDF scores instead of removing stopwords in a classification problem." Do you agree with this statement? Justify your position. Note: A classification is what you did with the image analytics assignment, e.g., predict high or low engagement using text as independent variables. (10 points) Hint: Focus on what the IDF part of the TF-IDF does based on our class discussions (check slides).**

[x] Yes, I agree **[ ] No, I don't agree**

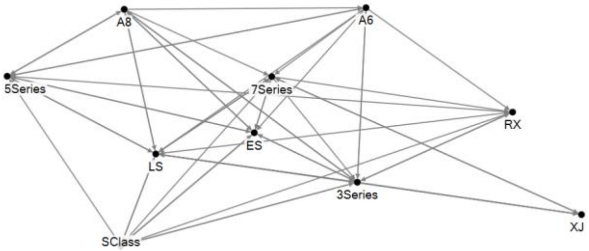
Justification:

Removing stopwords from the text for a classification problem is a good idea – although it becomes unnecessary when using TF-IDF scores instead. This metric helps show the importance of a term to a document in a corpus via the following calculations. TF, which stands for term frequency, defines the number of times a word has appeared in document. IDF, which stands for inverse document frequency, is calculated by the following formula: log(1/DF). DF, or document frequency, of a term tells us what percentage of documents in our dataset contain that term. By putting this altogether, we calculate the TF-IDF scores: TF*log(1/DF). Through this formula, we can distinguish between important and unimportant words in the context of classification. A high TF-IDF score means that this term is important to distinguish between classes of our outcome (for example to predict high or low engagement). Conversely, a low TF-IDF scores indicate that either the term does not appear much or that it appears everywhere – in both scenarios, it's useless for classification. Stop words, such as "a" and "an" should appear everywhere and therefore will get low TF-IDF scores. Consequently, it is better to simply use TF-IDF scores instead of removing stopwords for a classification problem.

2. **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. (10 points)**



Preference network A



Preference network B

Check one box below and justify your response:

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

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

Justification:

Using an unweighted page rank, product preference network B will have a higher correlation with sales data because it is a less dense network as compared to A. Product preference network A is extremely dense, where almost all nodes are connected to each other. Although the degree centrality would be high (there are a large amount of both in and out degrees), the betweenness centrality is extremely low(since there are many alternative paths between any two nodes). Therefore since almost all nodes are connected to each other in network A, it would be difficult for an algorithm to determine correlation with sales data. If we were to however use a weighted page rank, product preference network A would have an higher correlation with sales data since the links would automatically adjust for importance and therefore having a more comprehensive dataset would provide better results.

3. **In a comparative analysis of smart watches, you extracted conversations 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 messages posted on an Apple Watch forum, where every post mentions the Apple Watch, and where only a small % of posts co-mention the other two 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 (concatenating the two data sets).**

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. (8 points)

GREATER THAN (lift from dataset A > lift from dataset A+B)
Lift (Applewatch, battery) = N*#(Applewatch,battery) / #(Applewatch) * #(battery)
Since every post from the new data will mention Applewatch, both the terms N and #(Applewatch) will increase the same amount. By the same logic, since every post contains Applewatch, every new datapoint that mentions battery will also contain Applewatch and therefore the terms #(Applewatch,battery) and #(battery) should increase by the same amount. If we rewrite the lift equation with both these sets of terms rearranged:

Lift (Applewatch, battery) = (N / #(Applewatch) * #(Applewatch,battery)/ #(battery)

As mentioned above, if in both sets of fractions we increase the numerator and denominator by the same amount, then the first fraction should get smaller (since N is always bigger) as we increase data from the new set: (2/1 > 3/2 > 4/3 > 5/4) but the second fraction should get bigger (since #(battery) is always bigger) : (1/2 < 2/3 < 3/4 < 4/5). Assuming that the number of mentions of battery is always proportional to the total number of posts, the first fraction (with N) should decrease by more than what the second fraction (with #(battery)) will increase by and therefore the lift value should fall. For example (both with #battery/N = 0.5):

<u>Dataset A</u>

Lift (Applewatch, battery) = (N / #(Applewatch) * #(Applewatch,battery)/ #(battery)

Lift (Applewatch, battery) = (100/30) * (20/50) = 1.3

<u>Dataset A+B (adding 50 new posts with 25 mentioning battery)</u>

Lift (Applewatch, battery) = (N / #(Applewatch) * #(Applewatch,battery)/ #(battery)

Lift (Applewatch, battery) = (150/80) * (45/75) = 1.125

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? Justify your response. You can use a numerical example, but must justify using logical arguments. (8 points)

LESS THAN (lift from dataset A < lift from dataset A+B)
Lift (Movadoconnect, battery) = N*#(Movadoconnect,battery) / #(Movadoconnect) * #(battery)
Since a very small number from the new data will mention Movadoconnect, the N will increase far more than #(Movadoconnect). By the same token, #(battery) will increase far more than #(Movadoconnect,battery).

If we rewrite the lift equation with both these sets of terms rearranged:

Lift (Movadoconnect, battery) = (N / #(Movadoconnect) * #(Movadoconnect,battery)/ #(battery)

As mentioned above, the first fraction will get bigger with more data while the second fraction will get smaller with more number of posts. Assuming that the number of mentions of battery is always proportional to the total number of posts, the first fraction (with N) should increase by more than what the second fraction (with #(battery)) will decrease by and therefore the lift value should rise. For example (both with #battery/N = 0.5):

<u>Dataset A</u>

Lift (Movadoconnect, battery) = (N / #(Movadoconnect) * #(Movadoconnect,battery)/ #(battery)

Lift (Movadoconnect, battery) = (100/30) * (20/50) = 1.3

<u>Dataset A+B (adding 50 new posts with 25 mentioning battery)</u>

Lift (Movadoconnect, battery) = (N / #(Movadoconnect) * #(Movadoconnect,battery)/ #(battery)

Lift (Movadoconnect, battery) = (150/32) * (22/75) = 1.375

4. State if the following statements are True (T) or False (F). You must justify your response for full credit. (14 points)

4a. It is possible to guess reasonably accurately whether a post on social media is coming from a spammer from his/her network centrality metrics. T / F

True.

It is possible to predict whether a post on social media is coming from a spammer by analyzing his/her network centrality metrics. In a directed network (for example Twitter or Email), the ratio between in-degree and out-degree can help: a network with a high number of out-degrees but very low in-degrees would a flag for a spammer. It would make sense to assume that spammers would be reaching out a lot but getting very few, if any, responses back. Further, a very low normalized degree centrality score would indicate that this network is not connected to much of the larger social media network. In the case of undirected networks (such as Facebook), it makes sense to assume that if there is very long, or no, network path that connects an account to the spammer's account then it is most likely a spammer. A variant of this assumption is used almost daily when deciding whether or not to accept someone as a friends by looking at mutual friends.

4b. Cosine similarity is a better way to assess the similarity between documents than Euclidean distance when the social mentions come from diverse sources like Twitter, blogs, forum posts, etc. T / F

Note: When documents are represented as vectors (for example, of frequencies of words), one can calculate the Euclidean distance between the tips of the vectors as a measure of similarity between the documents (the smaller the distance, the greater the similarity).

True.

When the social mentions come from a diverse set of sources, Cosine similarity is a better way to assess the similarity between documents than the Euclidean distance. Although both metrics are driven by counts (the frequencies of words), the Euclidean distance assumes that the number of words in each document are equal. With Cosine similarity, by considering an n-dimensional space containing n terms, the angle between the two documents determines similarity and therefore accounts for the number of words in the document when determining similarity. When collecting information from various sources that could have varying lengths of data, it is important to make this consideration because an increase or decrease in the length of a document might change the Euclidean distance but not the Cosine similarity which makes sense – the similarity should not depend on the number of words but rather the terms themselves and how similar or dissimilar they are to each other.

5. 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 <u>all</u> relevant information about BC and RC using <u>appropriate</u> <u>lifts</u>. 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: Pay special attention to statements mentioning only one brand. All the data required for lift analysis are mentioned in the problem statement (i.e., nothing is missing here).

<u>Extract information from question:</u>

# (Both) = 8k

# (BC) = 16k

# (RC) = 12k

BC (-) = 7k

BC only (-) = 5k

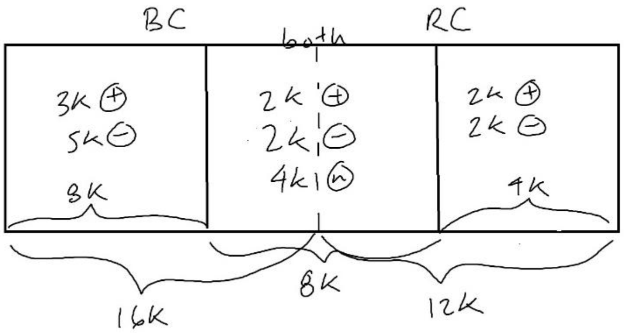
RC only (-) = 2k

Both (+) = 2k

RC only (+) = 2k

Total (+) = 2k

Infer data to get totals:



N = 20k

#(BC) = 16k, #(RC) = 12k

#(+) = 7k, #(-) = 9k

#(BC,+) = 5k, #(BC,-) = 7k

#(RC,+) = 4k, #(RC,-) = 4k

Lift Calculations:

L(BC,+) = [N * #(BC,+)] / [#(BC) * #(+)] = [20k*5k]/[16k*7k] = 0.893

L(BC,-) = [N * #(BC,-)] / [#(BC) * #(-)] = [20k*7k]/[16k *9k] = 0.972

L(RC,+) = [N * #(RC,+)] / [#(RC) * #(+)] = [20k*4k]/[12k *7k] = 0.952

L(RC,-) = [N * #(RC,-)] / [#(RC) * #(-)] = [20k*4k]/[12k *9k] = 0.740

Analysis

The recent problems of BC have certainly seemed to have a negative impact on brand perception but the lift for negative comments is below 1 – indicating that it is non-significant. Because the lift with positive comments is only slightly below 1, it is safe to assume that it was quite positive before the problems and therefore with good communication and right corrective measures, BC should be able to overcome this problem and hopefully bring back the positive comment lift to above 1 while, bringing down the negative comment lift further down.

RC's brand perception seems to have improvement a little with a positive comment lift value of almost 1 but it is not very significant. On the positive side, they are not displaying any complaints either.

6. **Consider the following movie reviews.**

Review	**Label (sentiment)**
Did not like, spend time well elsewhere.	**Negative**
Not a gem, glad I did not waste time	**Negative**
Not a wastage of my time, liked it	**Positive**
I did spend a gem of a time	**Positive**
Did not like, a wastage, a flop	**Negative**
Spend time elsewhere for a gem, wastage	**Negative**
Did not spend a good time	**Negative**
Gem, glad I did not spend time elsewhere	**Positive**
Glad it was not a flop, liked it	**Positive**
Good time, spent well, liked it	**Positive**

Using the training data above, use the k-nearest neighbors approach (with k=1) discussed in class for sentiment classification, find the sentiment (Positive or negative) of the following review.

Review	**Label (sentiment)**
Liked it, had a good time	**?**

Do NOT use synonyms. Remove only the following stopwords from your analysis: a, I, of, my, it, for, had, have. Also tense should be ignored: like = liked, spend = spent, etc. Further, waste = wasted = wastage. Show every calculation that helps you classify the sentiment of the review. Look at the slides for a similar problem we solved in class. Use Excel for the calculations. (15 points)

Removing stop words:

Review	Label (sentiment)
Did not like, spend time well elsewhere.	Negative
Not a gem, glad did not waste time	Negative
Not wastage time, liked	Positive
did spend gem time	Positive
Did not like, wastage, flop	Negative
Spend time elsewhere for gem, wastage	Negative
Did not spend a good time	Negative
Gem, glad did not spend time elsewhere	Positive
Glad was not a flop, liked it	Positive
Good time, spent well, liked it	Positive

Document frequency:

Term	Frequency
Like	5/10 = 0.5
Not	7/10 = 0.7
Did	6/10 = 0.6
Spend	6/10 = 0.6
Time	8/10 = 0.8
Well	1/10 = 0.1
Elsewhere	3/10 = 0.3
Gem	4/10 = 0.4
Glad	3/10 = 0.3
Like	5/10 = 0.5
Good	2/10 = 0.2

Calculating similarity to each training review:

Removing stopwords from new review becomes: liked, good time

Review #	Terms included	Similarity
1	Liked, time	log(1/0.5) + log(1/0.8) = 0.398
2	Time	log(1/0.8) =0.097
3	Liked, time	log(1/0.5) + log(1/0.8) = 0.398
4	Time	log(1/0.8) = 0.097
5	Liked	log(1/0.5) = 0.301
6	Time	log(1/0.8) = 0.097
7	Good, time	log(1/0.2) + log(1/0.8) =0.796
8	Time	log(1/0.8) = 0.097
9	Like	log(1/0.5) = 0.301
10	Liked, good, time	log(1/0.2) + log(1/0.8) + log(1/0.5) = 1.097

Classify new review:

Since k=1, we are only looking the most similar review which is review 10. Because it was classified as positive in our training set, we can classify the new review as positive as well.

Answer <u>any one</u> out of questions 7 and 8 (do not answer both).

7. Deals Galore (DG), an online retailer, is trying to calculate the customer network lifetime value (CNLV) of its customer base. By using Google Analytics and its FB fan page, DG finds that a certain influential customer, A, helped acquire three new customers, B, C and D, in one year. Assume that all customers buy one unit of DG's product, giving DG a margin of $100 per transaction. The average acquisition cost of a new customer is $20 for DG, while the annual retenti7on cost is $5. DG does not offer any incentive for referrals or advocacy.

Using analytics, DG is able to predict (classify) who may buy a product after its promotion campaign (i.e., become a customer). DG predicts B to be the type who will not respond to its campaign, while C and D are predicted to be types who will respond. Of course this prediction process is not perfect. For example, the table below shows the accuracy of predicting the type of people with a sample size of 100. Out of 20 people who would actually buy after a promotion, DG could predict 15 correctly as a would-be-buyer.

	Predicted by analytics to respond to promotion and buy a product	**Predicted by analytics to ignore the promotion**
Actually responded to a promotion by DG and bought a product	**15 people**	**5 people**
Did not respond to DG's promotion (i.e., would not buy)	**10 people**	**70 people**

What is the one-year Customer Network Lifetime Value (CNLV) of customer A from the above data? Do not use any discounting. Show all calculations. (15 points) Note: The confusion matrix shown in this problem is similar to the ones you got in the image analytics assignment.

<u>Calculate difference in values:</u>

Acquiring a customer that buys anyways = Value with A – Value without A

=($100-5) – ($100-$20-$5) = $95 – $75 = $20

Acquiring a customer that wouldn't buy without A = Value with A – Value without A

=($100-$5) – (-$20) = $95 + $20 = $115

Calculate CNLV:

Using confusion matrix,

Value = P[right prediction]*($ at right prediction) + P[wrong prediction]*($ at wrong prediction)

Value of influencing B = (70/75)*($115) + (5/75)*($20) = $108.66

Value of influencing C = (15/25)*($20) + (10/25)*($115) = $ 58

Value of influencing D = (15/25)*($20) + (10/25)*($115) = $ 58

Therefore,

Customer Influence value of A = $108.66 + $ 58 + $ 58 = $224.66

Customer Lifetime value of A = $100 - $20 - $5 = $75

Customer Network Lifetime Value (CNLV) of A = $224.66 + $75 = $299.66

8. In a paid search campaign on Google to increase visits to its new website with a set of promotional ticket prices, All American Airways (AAA) paid Google $1 per click. Google showed the ad to 10 million people. The click through rate (CTR) was 3%, and the transaction conversion rate (TCR), which is the % of people who buy upon visiting the website, of direct traffic was 2%. 20% of the people who clicked on the Google paid search link shared the information with a friend through email or other channels (who had not seen the Google ad), and the "word-of-mouth" (WOM) CTR and TCR were 10% and 3% respectively.

AAA also advertised on Facebook (at $0.75 per click) with a guessing game, where an individual, say, John, had to guess what his friends (up to 5 friends, must be members of the AAA Frequent Flyer Program) liked the most about AAA. Assume that the attention given by the friends to John is .3; that is, there is a 30% chance that they would visit the AAA website to verify John's responses after knowing that John had answered some questions about them. If friends verify, John get 1000 miles per friend, while each friend gets 500 miles for verification. Both John and his friends may buy air tickets as well during their visit to the website. The Facebook ad was seen by 3 million users, and had a CTR of 10% and a TCR of 1%. Assume that each Facebook user who clicked on the ad answered questions regarding five friends and notified them through Facebook messaging. 1% of friends who visited the AAA page to verify also bought tickets.

Folks who answered questions got it right 100% of the time (maybe they took the trouble to tell their friends what they had answered about them, but AAA didn't care; the objective was to get the friends to visit the site in the name of verification!). AAA guesstimated its cost to be $5 for every 25,000 free miles it gave. Assume that the average ticket price was the same (=$500) regardless of the type of ad or traffic.

Calculate the following ratios (i) Return on advertising for Facebook / Return on advertising on Google, (ii) Return on advertising for WOM traffic / Return on advertising for "direct" traffic, and (iii) profit for Facebook WOM traffic / profit for Google WOM traffic. What can you conclude based on the analysis? Show all calculations. (15 points)

Note: Return on Advertising is calculated the same way as return on investment using benefits and costs: ROA = (Revenue – Cost) / Cost