**User Generated Content Analytics (MIS381) Take-home Final, Fall 2017**

**Handed out: Dec 7, 2017, Date due: On Canvas by 11:59 p.m. on Dec 11. 2017**

Name: AKSHAT PARAKH

**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.**

**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 on all files for proper identification.**

**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.**

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 discussions from 19th September (check slides).

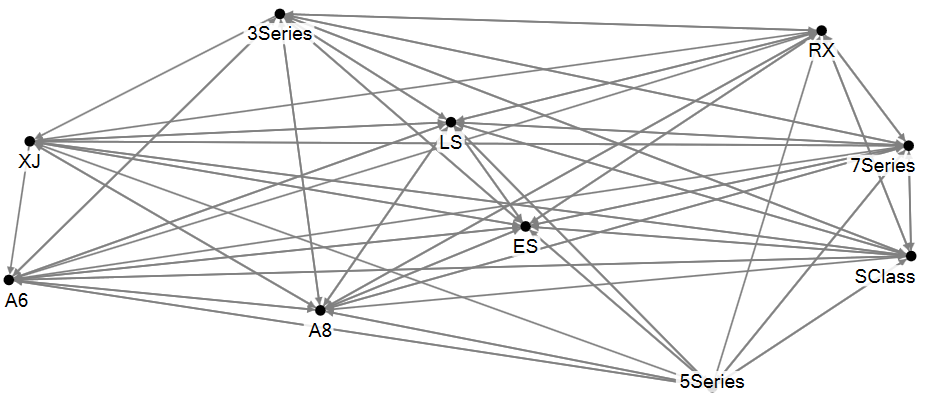
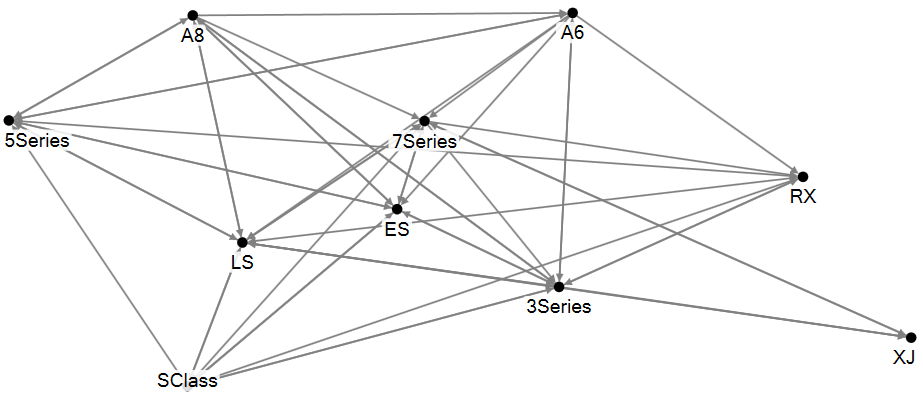
[ • ] Yes, I agree [ ] No, I don’t agree

Justification:

IDF stands for Inverse document frequency. TF-IDF is better suited for classification. IDF = 1/DF. DF tells what percentage of document in the data set have that word.

TF-IDF = term frequency\*log(1/DF). This formula gets rid of words that occur, everywhere and cannot be used to distinguish between high or low engagement. Hence, it is unnecessary to remove stopwords such as “a”, “an” as they will not be given considerable value when IDF is calculated.

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)

Preference network A Preference network B

Check one box below:

[ ] 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.

Justify your response.

In network A, all the brands are connected to each other and as the network is unweighted, PageRank has low correlation with sales data. In network B, it seems that a threshold value has been applied, so the edge that is below that value is not displayed. It can be said that only important connections will be displayed, so the PageRank from this network will have more correlation with sales data. When we put a threshold value, the connections having a value greater than that are shown. As all the brands are connected to each other, threshold helps in removing the unimportant connections thus making rest of the connections important.

3. 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 retention 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.

ANS:

From the confusion matrix, 15 and 70 are the correct predictions, hence

Accuracy(T) = (70+15)/100 = 17/20

CNLV (A)= CLV of A + CIV of A (for B, C and D)

CNLV (A) =(Margin (A) – Retention(A) – Acquisition(A))  + (T) (Margin(B)+ Acquisition saved(B)– Retention(B)) + (T) (Acquisition(C and D))

CNLV for (A) = (100-20-5) + (17/20) \* (100+20-5) + (17/20)\* (2) \*(20)

= (75) + (17/20) \*(115) +(34)

**CNLV (A) =$ 206.75**

\*CIV = Customer Influence Value

4. 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. (20 points)

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

ANS:

**Google: DIRECT WOM = 20\*300,000/100 = 60,000**

CTR = 3\*10,000,000/100 = 300,000 CTR = 6,000

TCR = 2\*300,000/100 = 6000 TCR = 3\*6000/100 = 180

TOTAL TCR = 6000+180 = 6180

COST = 300,000\*1 = $300,000

COST(WOM) = 6000\*1 = 6,000

REVENUE = 6180\*500 = $3,090,000

**FACEBOOK: DIRECT**

CTR = 10\*3,000,000/100 = 300,000

TCR = 1\*300,000/100 = 3,000

COST =300,000\*0.75 = $225,000

**WOM:**

# PEOPLE WHO RECEIVED NOTIFICATION = 300,000 \*5 = 1,500,000

# PEOPLE WHO ANSWERED = 30\*1,500,000/100 = 450,000

TCR = 1\*450,000/100 = 4,500

MILES = # PEOPLE WHO ANSWERED \* 1000 + # PEOPLE WHO ANSWERED \* 500 = 450,000\*(1000+500) = 675,000,000

SPENT = 675,000,000\*5/25,000 =$135,000

SPENT(CTR) = 450,000 \* 0.75 = $337,500

TOTAL COST =225,000 + 135,000+337,500= $697,500

REVENUE = (3,000 +4,500) \*500 = $3,750,000

(i) Return on advertising for Facebook / Return on advertising on Google

((3,750,000-697,500)/697,500) / ((3,090,000-306,000)/306,000) = 0.4810

(ii) Return on advertising for WOM traffic / Return on advertising for “direct” traffic = ((2,340,000 – 478,500)/478,500) / ((4,500,000-525,000)/525,000)) = 0.513

(iii) profit for Facebook WOM traffic / profit for Google WOM traffic = ((2,250,000- 472,500))/(90,000-6,000)) = 21.16

Overall ROA for google is more, which is understood because facebook is spending more than 2 times the amount that google is spending and the revenue generated by Facebook is $660,000 more than that of google. Facebook is spending more money by giving free miles for the users who verify and those who answer questions, this increases the cost. Google on the other hand does not provide benefits for WOM.

Facebook has a very high TCR rate for WOM, but Google has only 180 TCR by WOM so the ratio of ROA between WOM and Direct is less than one. Facebook has more TCR for WOM because they are offering free miles.

When WOM profit for Facebook and Google is compared, it is obvious that the ratio is so high because Facebook is pulling in more traffic by their free miles offer.

ROA is the factor to be considered when advertising, hence google is performing better. Facebook has increased their expenditure by including the free miles offer and the revenue generated does not justify the expense.

5. Best Cruises (BC) recently ran into major problems with its ships. In a cruise forum online, where folks discuss BC and its rival Royal Cruises (RC), there were 20,000 posts on these two companies (in one month following the incidents). A post may mention only BC, only RC or both. BC and RC were mentioned together in 8,000 posts. Further, BC was mentioned in 16,000 posts. RC found itself in 12,000 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) a positive about one and negative about the other, (iv) positive about both or (v) negative about both. Assume that there are NO neutral posts.

BC got 8,000 negative posts. Further, there were 5,000 negative posts that only mentioned BC. Note: This means that there were 3000 posts where (i) BC was mentioned negatively **and** where (ii) RC was also mentioned either positively or negatively.

RC was negatively viewed in 6,000 posts. Further, there were 2,000 negative posts that only mentioned RC. The two companies were mentioned together in a positive manner in 2,000 posts, and in a negative manner in 1,000 posts. **Note that in calculating the total number or positive (or negative) posts regardless of the cruise line, if a post mentions both companies, then at least one company must be mentioned positively (or negatively) for the post to count as a positive (or negative).**

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. As always, show all calculations. (20 points)

BC (-ve mentions) = 8000

BC (+ve mentions) = 8000

RC (-ve mentions) = 6000

RC (+ve mentions) = 6000

Only BC(-ve) = 5000

Only BC(+ve) =8000-5000 = 3000

Only RC(-ve) = 2000

Only RC(+ve) = 4000-2000 = 2000

Negative posts that mentioned BC (-ve) and RC (+ve) = (Posts that mention BC (-ve) and RC (-ve or +ve) – Number of posts where both are -ve) = 3000-1000 = 2000.

Negative posts that mentioned BC (+ve) and RC (-ve) = (Posts that mention RC (-ve) and BC (-ve or +ve) – Number of posts where both are -ve) = 4000-1000 = 3000.

The above two numbers add up to 5000 which can be verified.

Number of post in which one is +ve and one is –ve = (# post in which both are mentioned) – (# post in which both are –ve) - (# post in which both are +ve) = 8000-1000-2000= 5000.

L (RC, +ve) =(20k\*6k)/(12k\*12k) = 5/6

L (BC, +ve) =(20k\*8k)/(16k\*12k) = 5/6

L (BC, -ve) = (20\*8k)/(16k\*13k) = 10/13

L (RC, -ve) = (20k\*6k)/(16k\*13k) = 10/13

L(BC(-ve) , RC(+ve)) = (20k\*2k) / (8k\*6k) = 5/6

L(BC(+ve), RC(-ve)) = (20k\*3k) / (8k\*6k) = 5/4

ANALYSIS:

L(BC(+ve), RC(-ve)) has a value of 1.25, which states that when RC and BC are compared, BC is being talked about in positive manner and RC in negative manner. RC will need to find out why they are being mentioned negatively even though problems were with BC’s ships. They will have to promote the positive aspects of themselves and also analyze the factors responsible for negative image.

6. State if the following statements are True (T) or False (F). You must **justify** your response for full credit. (10 points)

6a. 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

From the network centrality metrics of a person, one can judge if that person is connected to them directly or via a network of friends. If there is no network pathway leading from that person to you and you are receiving a post from them, it is likely that, that person is a spammer. Common connections can be used a factor to determine the credibility of post from someone you don’t know.

6b. **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 contained), 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.

With the increase in number of words in a mention the Euclidean distance between two mentions will increase as the similarity will decrease because of an increase in number of words. But the angle between them is going to remain same, so Cosine similarity is a better way to assess similarity between documents.

1. 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 k-nearest neighbors approach discussed in class for sentiment classification, find the sentiment (Positive or negative) of the two following reviews.

|  |  |
| --- | --- |
| Review | Label (sentiment) |
| Liked it, had a good time | ? |
| Did not like it, did not have a good time | ? |

Do NOT use synonyms, but remove **only** the following words 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 two reviews. Look at the slides (19th Sep) for a similar problem we solved in class. (15 points)

ANS:

The positive and negative probability, based on occurrence in training data, of the words that are present in reviews whose sentiments are to be calculated are:

|  |  |  |
| --- | --- | --- |
| Words (mentions) | Positive (probability) | Negative (probability) |
| Like (3(+ve) , 2(-ve)) | 3/5 | 2/5 |
| Good (1(+ve) , 2(-ve)) | 1/2 | 1/2 |
| Time (4(+ve) , 4(-ve)) | 4/8 | 4/8 |
| Did (2(+ve) , 4(-ve)) | 2/6 | 4/6 |
| Not (3(+ve) , 5(-ve)) | 3/8 | 5/8 |

|  |  |  |
| --- | --- | --- |
| Review | Similarity score for 1st review | Similarity score for 2nd review |
| Did not like, spend time well elsewhere. | Log (5/2) +Log (2) = 0.698 | Log (3/2) +Log (3/2) + Log (8/5) +Log (8/5) + Log (5/2) +Log (2) = 1.458 |
| Not a gem, glad I did not waste time | Log (2) = 0.301 | Log (3/2) +Log (3/2) + Log (8/5) +Log (8/5) +Log (2) = 1.601 |
| Not a wastage of my time, liked it | Log (5/3) +Log (2) = 0.522 | Log (8/3) +Log (8/3) + Log (5/3) +Log (2) = 1.3738 |
| I did spend a gem of a time | Log (2) =0.301 | Log (3) +Log (3) + Log (2) = 1.2552 |
| Did not like, a wastage, a flop | Log (5/2) =0.397 | Log (3/2) +Log (3/2) + Log (8/5) +Log (8/5) + Log (5/2) =1.191 |
| Spend time elsewhere for a gem, wastage | Log (2) = 0.301 | Log (2) = 0.301 |
| Did not spend a good time | Log (2) +Log (2) = 0.602 | Log (3/2) +Log (3/2) + Log (8/5) +Log (8/5) + Log (2) +Log (2) = 1.362 |
| Gem, glad I did not spend time elsewhere | Log (2) = 0.301 | Log (3) +Log (3) + Log (8/3) +Log (8/3) +Log (2) = 2.0107 |
| Glad it was not a flop, liked it | Log (5/3) = 0.221 | Log (8/3) +Log (8/3) + Log (5/3) =1.0728 |
| Good time, spent well, liked it | Log (5/3) = 0.823 | Log (2) +Log (2) + Log (5/3) = 0.823 |

Review 1: positive sentiment, the highest similarity score is with a review that is positive.

Review 2: positive sentiment, the highest similarity score is with a review that is positive.

