**Sample questions for UGCA**

1. A survey was conducted within cohort of 65 students, which asked each student to indicate which 3 students in the class s/he corresponds with the most on Facebook. The Professor then used this data to draw the social graph and calculated the betweenness centrality of each student. The average betweenness centrality came out to be 200.

If the Professor had looked up each student on Facebook and created a social network between these 65 students using the actual friendship or contact list, would the average betweenness centrality be above 200? Justify your response.

2. “Unsupervised sentiment analysis is likely to be more accurate than supervised methods for a specific context (like movie or hotel reviews)” Do you agree with this statement? Justify your response.

3. “TF-IDF scores are more useful than wordclouds for classification purposes (e.g., predicting high and low engagement). “ Do you agree with this statement? Provide justification for your response.

4. Happy Cruises (HC) recently ran into major problems with its ships. Social media chatter shows HC was mentioned 10,000 times on a major cruise forum, and was mentioned along with negative words 6500 times. There were 5000 co-mentions of HC and positive words. Note that negative and positive words can appear in the same post in this problem. By contrast, during the same period, Regal Cruises (RC) was mentioned 4000 times, and was co-mentioned 1000 times with negative words, and 1500 times and with positive words . A third cruise line, Paradise Cruises (PC) got the rest of the mentions, with 60% positive and 40 percent negative. The forum had 20,000 total posts during the time period of interest. **Using appropriate lifts**, what can you say about consumer sentiments regarding HC and RC during the time period of the posts? Show all lift calculations. Assume for simplicity that HC and RC were never mentioned together.

5. Consider a short review of a smartphone: **“This phone has amazing features.”**

Shown below is a training data set with reviews labeled as positive or negative. **Use this data to classify the above review as positive or negative. Show your analysis and calculations.**

1. “Awesome phone, amazing features” - Positive
2. “Advertised as having amazing features, but it is actually awful to use.” -- Negative
3. “Amazing features, which I am enjoying a lot.” -- Positive
4. “Amazing features, gotta love it.” -- Positive
5. “This phone sucks” -- Negative
6. “How awful can be a phone be?” -- Negative

6. John, Sue, Liz and Ted follow each other on Twitter. A sample of their recent tweets is shown below.

John: XYZ sucks

Ted: Watching the Oscars

Liz: @Ted I’m glued to the TV

Sue: @Ted maybe I will watch too

Ted: @Sue enjoy the Oscars

John: @Ted I am watching and working as well

Sue: RT @John: XYZ sucks

John: @Sue pl tweet your thoughts

Sue: @John I will

Sue: @John Argo will win as @Liz predicted

Liz: Life of Pi looks good too.

Liz: @John why does XYZ suck?

John: @Liz XYZ is an idiot

Liz: @Sue told you Argo will be big

Ted: @Sue Looks like @Liz loves Argo

Ted: @Sue do you really think XYZ sucks?

Calculate the attention each person gets in this network. ***Do not*** calculate attention for retweets, @replies and @mentions ***separately***. Assume they are equally important, and create a si*ngle measure of attention for each person*. Show your calculations in detail. What can you conclude about the four people from your analysis?

***Note: Tweets without @ are not counted as they are not directed toward anybody in particular and therefore must not enter attention calculations.***

**Answers to select questions**

***1. The average betweenness centrality would be below 200 with actual Facebook data from the 65 students.*** *In the extremely unlikely event that each student has exactly three FB friends from the cohort, the betweenness centrality will be equal to 200. However, a much more plausible scenario is one where a student has many more than 3 FB friends from the cohort, in which case the actual network is denser than the one graphed by the Professor based on the data collected in class. Thus the average betweenness centrality will be lower than 200, since there are more alternative paths between any two nodes in a denser network, making each person less important for connectivity.*

***2.*** *I don’t agree. In unsupervised sentiment analysis, there are default weights for words which are independent of the domain and therefore not tailored for any specific application. In class we discussed examples of words like “scary” and “deadly” which would be very negative for review of cars, but which may be very positive for movies. By contrast, supervised methods check whether certain words are more likely to appear in positive or negative sentiments, and are therefore dependent on the domain (and hence generally more accurate). Note that when you tweak weights of reference words in unsupervised methods, they don’t remain unsupervised anymore.*

**3. Yes, I agree. Try the justification on your own – Hint:** The key difference between a wordcloud (a visualization of word frequency) and Term Frequency-Inverse Document Frequency (TF-IDF) is the IDF part. Focus on what the IDF does in a classification problem. Is wordfrequency able to do the same thing? Therein lies the justification.

**4.** *# mentions:*

*HC = 10k, HC- (negative) = 6.5k, HC+ = 5k*

*RC = 4k, RC- = 1k, RC+ = 1.5k*

*PC = 6k, PC- = 2.4k, PC+ = 3.6k*

*Total posts* (*T*) *= 20k*

*Lift(HC, positive) =* (*HC+/T*)*/*[(*HC/T*)*\**(*Positive mentions/T*)] *= (5\*20)/(10\*10.1) = .99*

*Lift(HC, negative) =* (*6.5\*20*)*/*(*10\*9.9*) *= 1.31*

*Obviously the recent problems have hurt HC, and it has a significant (though not super high) lift with negative comments. The lift with positive comments is just about 1, which means that it was probably quite positive before the problems. So with appropriate corrective action and communication with customers, HC should be able to get back on track.*

*Lift(RC, positive) =* (*1.5\*20*)*/*(*4\*10.1*) *= .74*

*Lift(RC, negative) =* (*1\*20*)*/*(*4\*9.9*) *= .51*

*Even though RC did not face any issues, customers are showing no significant enthusiasm for its cruises. On the bright side, they are not complaining either.*

***5. Try this on your own.***

**6.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Attention that John gets* | *Attention that Sue gets* | *Attention that Liz gets* | *Attention that Ted gets* |
| *John’s attention to* | *-* | *1/3* | *1/3* | *1/3* |
| *Sue’s attention to* | *3/5* | *-* | *1/5* | *1/5* |
| *Liz’s attention to* | *1/3* | *1/3* | *-* | *1/3* |
| *Ted’s attention to* | *0* | *¾* | *¼* | *-* |
| *Total attention* | *.933/3 = .311* | *1.417/3 = .4723* | *.783/3 = .261* | *.8667/3 = .2889* |
| *What can we say about each person?* |  | *Gets the highest overall attention, maximum from Ted, and gives John the max attention* | *Gets the highest attention from John, gives equal attention to all* | *Gets equal attention from John and Liz, but gives maximum attention to Sue* |

*Note that the four numbers add up to more than 1. That’s because for each person we are calculating the attention s/he got / the maximum attention s/he could get, which is 3 in this example. E.g., 3 would be the attention John would receive if all three folks (Sue, Liz and Ted) gave 100% of their attention to John.*

*Another valid way of calculating attention for each person would be to add up the retweets, replies, and mentions, and to divide by the total number minus the person’s own retweets/replies/mentions. E.g., for John, the attention would be (3+1)/12 = 1/3. Sue would get 5/10 = ½, Liz 3/12 = ¼ and Ted = 3/11*