Analytics for Unstructured Data

Final Exam (F2021)

Open notes and Internet access (but see B. below)

Date: October 21, 2021
Time: 19:00 - 22:00 hours 10/21 (submit to Canvas). You can take an extra 30 minutes and submit by 22:30 without penalty (I have set the cutoff for submission to 22:30 hours) .

Maximum Points: 100

Name: **Ankita Kundra**

Please read all instructions carefully before answering the questions

A. **Answers should preferably be typed. Hand-written answers are acceptable, but they must be completely legible. 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.**

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 are required to be reported to the Office of the Dean of Students.**

C. **You can submit a maximum of two files, e.g., a Word or pdf, and an Excel file. If you choose to submit a single Excel file, create a worksheet for each question. Write your name inside all files for proper identification.**

D. **If you have clarification questions, please post them here (please check if a similar question has already been answered:**
https://docs.google.com/document/d/1mrQ8B9CZW1lTGa4FUxdkBf3M_XCft0NEaJ-bsReT7Fs/edit?usp=sharing
I will check the document every 10 mins.

Answer questions 1, 2, 3, 4 (10x4 = 40 points)

1. 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).

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

Justification:

Stopwords are words like 'a', 'the', 'and' which have a high frequency in a corpus or they do not carry much meaning and context.

On the other hand, TF-IDF scores can be used to give more weightage to rare words instead of repetitive words. TF-IDF assigns a value to a term according to its importance in a document scaled to its importance across all documents in the corpus.

Tf = Term frequency of t, no of times t occurs in a document.
IDF = Inverse of no of time the term occurs in all the documents
A commonly occurring word will have very low TF-IDF scores as compared to rare words
This **may be helpful in case of prediction** since a particularly repeating word may give no additional benefit in prediction. But incase of classification, a word which is repeatedly used across many documents may also help to classify the document correctly. This information about the importance of a repeatedly used variable will be lost if we use TF-IDF. And in this case of classification, removing stop-words may be a better approach. For example, 'movies' repeated acrossmany documents will help us to classify the document correctly as movies.

2. "In principle, the semantic orientation (PMI) approach to sentiment analysis is more accurate than other unsupervised methods like VADER." Do you agree with this statement? Justify your response. (10 points)

[ ] Yes **[ ] No**

Semantic orientation (PMI) approach to sentiment analysis is more accurate than other unsupervised methods like VADER. VADER lacks context of the data on which it is being used. In

VADER, every word is pre-assigned a positive sentiment or a negative sentiment irrespective of what is the context of the data. For example: 'scary', 'horror' might be assigned a negative score in VADER but it conveys a positive meaning if it is used in context of a horror movie.

Semantic orientation (PMI) on the other hand finds mutual information between positive words as defined by the user and a particular comment. In PMI, we define scores based on the context of the corpus making use of log of lift values. We calculate lift of non stop-words present in each document with positive words and negative words. Hence, we do not use pre-defined global values and evaluate sentiment based on provided corpus of docs

Thus, semantic orientation is a better approach to sentimental analysis.

3. Consider the allocation of topics to words in LDA (topic modeling) using the Collapsed Gibbs Sampling in the three documents below. In the diagram, what is shown in the current allocation of two topics to different words. Calculate the probability of allocating the topic of **War** to the word **Fight** in document 2 after one more iteration. Show your calculations. (10 points)

Topic	Words in document 1
War	Fight
Romance	Kill
Romance	Love
War	Kill
War	Love

Topic	Words in document 2
Romance	Fight
War	Love
War	Soldier
Romance	Love
Romance	Love

Topic	Words in document 3
War	Kill
War	Fight
Romance	Soldier
War	Kill
Romance	Love

Solution:

Words

Words	War	Romance
Fight	2	1
Kill	3	1
Love	2	4
Soldier	1	1

Topics	War	Romance

Doc 1	3	2
Doc 2	2	3
Doc 3	3	2

For each document

Doc1→

Words	P(Wi|War) *P(War|Doc1)	P(Wi|Romance)*P(romance|Doc1)	New topic
Fight	(2/8)*(3/5)=3/20	(1/7)*(2/5)=2/35	WAR
Kill	(3/8)*(3/5)=9/40	(1/7)*(2/5)=2/35	WAR
Love	(2/8)*(3/5)=6/40	(4/7)*(2/5)=8/35	Romance

DOC2→

Words	P(Wi|War) *P(War|Doc2)	P(Wi|Romance)*P(romance|Doc2)	New topic
Fight	2/8*2/5	1/7*3/5	WAR
Soldier	1/8*2/5	1/7*3/5	Romance
Love	2/8*2/5	4/7*3/5	Romance

DOC 3→

Words	P(Wi|War) *P(War|Doc3)	P(Wi|Romance)*P(romance|Doc3)	New topic
Kill	3/8*3/5	1/7*2/5	War
Fight	2/8*3/5	1/7*2/5	War
Soldier	1/8*3/5	1/7*2/5	War
Love	2/8*3/5	4/7*2/5	Romance

Updates docs with new topics

Doc 1-

War	Fight
War	Kill
Romance	Love
War	Kill
War	Love

War -4 R-1

Fight war-1

Doc 2-

War	Fight
Romance	Love
Romace	Soldier
Romance	Love
Romance	Love

War -1

War fight-1

Doc 3→

War	Kill
War	Fight
War	Soldier
War	Kill
Romance	Love

War fight-1

We need to computer P(Fight|War)*(War|doc2)

war = 4+1(don't consider the instance we wish to compute) +4=9

Fight and war=1+1=2

War in doc 2= 1/5

P(fight,war,doc2)=(2/9)*(1/5)=2/45

4. Consider the (i) labels from image analytics and (ii) captions for a few Instagram images:

	Image analytics labels	**Caption from Instagram**
Image 1	Tiger Bengal Siberian deer forest trees blood	A tiger and her prey
Image 2	Elephant, tusk, playful, tourist, car, jeep, safari	A memorable safari

Image 3	Temple village people architecture ancient	Scenes from a village
Image 4	River water boats shore blue sky fishermen	Boats on the river

If you combine the image labels and the captions, and run a model predicting engagement, will you will get higher accuracy than with a model with the image labels alone? Justify your response. (10 points)

~~[ ] Higher accuracy with image labels + captions~~ [ ] Higher accuracy with only image labels

Correct choice has been striked out

We will get higher accuracy on combining the labels captured from image analytics with captions. The captions provide the gist of what the images are trying to convey and thus complements information to that provided by the image labels alone. These labels often refer to specific area or places, activities, people, scenery etc. that appear on these images. However, these labels often are not able to capture the sentiment behind the image and what the author of the image is trying to convey.

Case in point:

For image 1, the caption indicates the sentiment of the image which is tiger preying on a dear. While the image analytics captures “Bengal”, “Siberian”, “forest”, “blood”, without the context of tiger preying, we wouldn’t have captured the sentiment of the image. On the same note, if we consider Image 2, we see that the labels are telling us that the image has “elephant”, “tusk”, “playful”, “tourist”, “car”, “jeep” and “safari” in it where the different words could be interpreted differently. However, only upon considering the caption we are able to capture the sentiment behind the image which is described by the word “memorable”.

Hence, higher accuracy can be achieved for a model with image labels and captions as compared to a model with labels alone.

Answer any THREE questions from 5, 6, 7, 8, and 9 (3x20 = 60 points). If you answer more than 3, I will only grade the first 3 answers.

5. 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** **[ ] LESS THAN (check one)**

For dataset A,

Lift(AppleWatch, battery) = N*n(AppleWatch, battery)/n(battery)/n(AppleWatch)

For dataset A and B combined,

Total reviews N' = 2N

n'(AppleWatch) = n(AppleWatch) + N (since all the new posts mention AppleWatch)

The proportion of posts mentioning **battery** is the same for data sets A and B

Since, the number of reviews are same for dataset A and B, this means that number of comments with battery is same in dataset A and B

Thus, n'(battery) = 2*n(battery)

Lift'(AppleWatch, battery) = N'*n'(AppleWatch, battery)/n'(battery)/n'(AppleWatch)

= 2N* n'(AppleWatch, battery)/2n(battery)/(n(AppleWatch) + N)

= N* n'(AppleWatch, battery)/n(battery)/(n(AppleWatch) + N)

Lift'(AppleWatch, battery)/ Lift(AppleWatch, battery)

= n'(AppleWatch, battery)* n(AppleWatch)/ (n(AppleWatch) + N) *n(AppleWatch, battery)

= [n(AppleWatch, battery) + n(battery)])* n(AppleWatch)/ (n(AppleWatch) + N) *n(AppleWatch, battery)

Since n(battery) < N

Thus, Lift(AppleWatch, battery) for dataset A + B will be less than the lift for dataset A

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 **~~[ ] EQUAL TO~~** **[ ] LESS THAN (check one)**

For dataset A,

Lift(MovadoConnect, battery) = N*n(MovadoConnect, battery)/n(battery)/n(MovadoConnect)

For dataset A and B combined,

Total reviews N' = 2N

The proportion of posts mentioning **battery** is the same for data sets A and B

Since, the number of reviews are same for dataset A and B, this means that number of comments with battery is same in dataset A and B

Thus, n'(battery) = 2*n(battery)

The proportion of MovadoConnect posts which also mention battery is the same for data sets A and B

This implies = n'(MovadoConnect, battery) = 2*n(MovadoConnect, battery)

This, also implies that n'(MovadoConnect) = 2*n(MovadoConnect)

Lift'(MovadoConnect, battery) = N'*n'(MovadoConnect, battery)/n'(battery)/n'(MovadoConnect)

= 2N* 2*n(MovadoConnect, battery)/2n(battery)/n'(MovadoConnect)

=2*N*n(MovadoConnect, battery)/n(battery)/n'(MovadoConnect)

Lift'(MovadoConnect, battery)/ Lift(MovadoConnect, battery)

= 2* n(MovadoConnect)/ n'(MovadoConnect)

= 2* n(MovadoConnect)/ 2*n(MovadoConnect)

=1

Thus, Lift(MovadoConnect, battery) for dataset A + B will be equal to the lift for dataset

6. Consider two documents d_1 and d_2 represented by term weights as follows:

$d_i = (w_{i,1}, w_{i,2}, \ldots w_{i,n})$ where $i \in \{1, 2\}$. Now consider a document retrieval problem with a query expressed as a vector $r = (w_{r,1}, w_{r,2}, \ldots w_{r,n})$. 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)

7. 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 $\max_i n_r^i$ 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)

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. (20 points)

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

Solution:

Total posts = BC mentioned + RC mentioned – posts where both were mentioned

= 16,000 + 12,000 – 8,000 = 20,000

Total of 7000 positive posts. 2000 positive posts with both RC and BC. 2000 positive posts with RC only. Thus, 3000 posts with BC only

Thus, total comentions of BC and positive words = 5000

Total comentions of RC and positive words = 4000

We know that

Lifts(A,B) = N* #n(A and B)/#n(A)*#n(B)

Lift between BC and positive posts = 20k*5k/(16k*7k) **= 0.89**

Lift between RC and positive posts = 20k*4k/(12k*7k) **= 0.95**

For negative posts,

BC got 7000 negative posts. 5000 posts that mentioned BC only. There were 2000 negative posts that **only** mentioned RC

Thus, total comentions of BC and negative words = 7000

Total comentions of RC and negative words = 2000+(7000-5000) = 4000

Total negative posts = 5000+2000+(7000-5000) = 9000

Lift between BC and negative posts = 20k*7k/(16k*9k) **= 0.972**

Lift between RC and negative posts = 20k*4k/(12k*9k) **= 0.74**

Lift between BC and RC = 20k*8k/16k/12k **= 0.833**

Conclusion:

The lift between BC and RC is 0.833 which indicates they are not mentioned together very often. But still people, do make comparisons between them.

Lift between BC and negative posts is more than lift between RC and negative posts which indicates RC is viewed more negatively than BC. Lift between RC and positive posts is more than lift between BC and positive posts which indicates which indicates that RC is talked about more positively than BC.

Thus, it can be concluded that:

While association between positive and RC is not very high, people talk about RC more positively than BC and talk negatively about RC less than BC

9. Shown in an attached Excel file are 100-dimensional embeddings for the words “obama”, “usa”, “france”, “sarkozy”, and “gandhi”

From these embeddings, find the word that is most similar to

obama – usa + france (Hint: the answer is either sarkozy or gandhi)

Briefly explain the steps you will use to get the answer. Show all calculations. You can use the Excel file I have shared to show your calculations. If you don’t have Excel, you can use Google sheets instead. (20 points)

Solution:

Word embeddings involve representing every word as a vector of real numbers.

Hence, in excel sheet: Obama, USA, France, Sarkozy and Gandhi have been represented as a vector of real numbers.

We first calculate the real number vector corresponding to obama – usa + france.

We then compare the vector corresponding to (obama – usa + france) with other given vectors. We can find similarity in various ways. Two of the ways could be:

a) Using Euclidian distance
b) Using Cosine similarity

On the excel sheet attached, we find that the similarity on the basis of Cosine Similarity is maximum for Sarkozy.