**Unstructured Data Analytics, F2022**

**Sample Questions & Answers for final exam (text analytics part) – Part I**

**Network Analytics questions will be posted later**

**Sample Questions**

The exam will have both conceptual and calculation type questions. The questions below are indicative of the complexity of the exam, but do not cover all topics discussed in class. The exam, however, is comprehensive, and not limited to the topics below.

1. “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.

No. Supervised will be more accurate because the classification or prediction model will pick up the context (e.g., scary is good in the context of a movie) from the training data. In unsupervised methods, the default sentiment values for words in the lexicon do not have a context.

1. A friend who has not taken this course has just discovered the wordcloud, which shows words that appear with the highest frequencies. Explain to your friend the problems that one may face using frequencies alone for classification or prediction. Also explain in plain English (your friend is not into math ☺) a better way to find words in a collection of documents that are “important”.

Regardless of frequency, if a word appears in most documents, it can’t indicate the class of the document. Words that appear frequently in documents of one class, but not in documents of other classes, are the best predictors of the class of a document. To find important words, check how commonly each word can be found in the documents. Words that are relatively less common are the important words.

1. In text clustering, under what conditions would you choose purity as a measure of external validity over entropy?

When only the dominant class is important for an application. Say there are people who love, like and dislike your products. If you only care about people who will love your product, a cluster where 70% of customers love the product, 15% like, and 15% dislike is no different from a cluster where the % are 70, 25 and 5.

1. “If I double the number of messages in my data set (assume that the additional data came from the same online forum), all lift values will get doubled.” Do you agree with this statement? Justify your response.

I don’t agree. When you bring additional data from the same source, the number of mentions and co-mentions will go up proportionately, and therefore the lift values should not be affected.

1. In MDS visualization, why is it advisable to keep the number of brands and attributes to a minimum?

In a typical 2-d MDS visualization, with n brands, the algorithm must try to represent nC2 values on a 2-d plane, which is impossible to accomplish without significant errors. The smaller the n is, the better the representation.

1. True / False “If you use cosine similarity and spacy similarity on the same data, your similarity scores will be higher in the former case.” T / F (but justify your response)

F. Spacy will give higher similarity because it will find similarity between words (e.g., awesome and great will be treated as being similar), while cosine similarity using the bag-of-words will find the words to have no similarity.

1. How would you manually determine the suitable number of topics in topic modeling?

To find if there are too many topics: If the same words show high (and comparable) weights for multiple topics, then we need to reduce the number of topics.

To find if we have too few topics: If totally different words (and this is subjective) start loading heavily on one topic, we need to increase the number of topics.

**Calculations: A large chunk of the exam involves calculations. Use a calculator and/or Excel.**

1. Assume that a feature vector uses binary representation of different words (i.e., just presence or absence of a word, do not use frequencies or TF-IDF here). Make up 4 documents (i.e., make up some words – exclude stopwords, but don’t make the maximum size too large) to illustrate how cosine similarity may be better than Euclidean distance for clustering or classification under certain conditions. A part of this question involves constructing the documents in a smart way that highlights the difference between the two approaches.

**Hint:** Make (D1, D2) and (D3, D4) similar in content (e.g., D1, D2 could be about movies, and D3, D4 could be hotel reviews – of course, some words should be common across the two domains, but make D1 and D3 short, and D2 and D4 longer. Show that the Euclidean distance between the movie documents (one short and one long) is higher than the distance between a movie and a hotel doc of the same size. Then use cosine similarity to show that this can be avoided – i.e., in spite of length difference, the movie docs are seen as more similar than a movie and a hotel doc of the same size.

D1: Awesome movie

D2: Awesome hotel

D3. Awesome movie, awesome movie, awesome movie.

D4. Awesome hotel, awesome hotel, awesome hotel

Note that each review vector will have 3 dimensions: awesome, movie, hotel

The vector representations (using TF) will be:

D1: (1,1,0)

D2: (1,0,1)

D3: (3,3,0)

D4: (3,0,3)

Cosine similarity between D1 and D3 = cosine similarity between D2 and D4 = 1 (angle is 0 degree)

Euclidean distance between D1 (movie review) and D3 (longer movie review) and that between D2 and D4 (hotel reviews) = (8)^.5

Euclidean distance between D1 (movie review) and D2 (hotel review) = 2^.5

That is, Euclidean distance thinks that D1 and D3 (both hotel reviews) are more dissimilar than D1 (movie review) and D2 (D2 is a hotel review). But cosine similarity does not make this mistake. There will be 0 degrees between D1 and D3 and between D2 and D4, and a non zero angle between any movie and hotel review!

1. Consider the reviews below (some stopwords have been removed):

|  |  |
| --- | --- |
| Review | Label |
| Did not like, spend time well elsewhere. | Negative |
| Not 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 gem, wastage | Negative |
| Did not spend good time | Negative |
| Gem, glad did not spend time elsewhere | Positive |
| Glad was not flop, liked | Positive |
| Good time, spent well, liked | Positive |

Using the k-nearest neighbors (use k=1) approach discussed in class for sentiment classification, find the sentiment (Positive or negative) of the two following reviews. Do not use TF-IDF weights, but use the following approach. If a word appears x% of the time in the training data, a match with this word should get 1/x weight (e.g., 1/.3 for 30% occurrence).

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

Partial answer (try it on your own):

The general approach should be as follows (see slide son k-nearest neighbors in sentiment analysis)

Find the important words in the two new reviews after removing stopwords. E.g., for the first new review the important words are like, good, and time.

Find the % of time these words appear in the training data. E.g., like appears in 50% of the docs. Now for a new review and each review in the training dataset, calculate a matching or similarity score. Between this new review and the first review in the training data set, there are two matches: like and time. Time appears in 8 out of 10 reviews. So the matching score would be:

log(1/.5) + log(1/.8)

The higher the score, the more similar the two reviews. Now do this for the remaining 9 reviews in the training data set (Excel is best), and find the nearest neighbor (highest similarity score). The sentiment of that review will be the sentiment of the new review.

Do NOT use synonyms or remove stopwords. Tense should be ignored: E.g., like = liked, spend = spent, etc. Further, waste = wasted = wastage.

Show calculations that help you classify the sentiment of the two reviews.

1. 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 the cruise lines were never mentioned together.

*Answer: 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.*