

# Carnegie Mellon University in Qatar

AI for Medicine

15-182 - Spring 2023

## Assignment 2

Name: \_\_\_\_\_

Andrew ID: \_\_\_\_\_

**Due on:** February 15, 2023 by midnight

### Instructions:

- This assignment consists of four problems. Answer them all.
- Submit your answers through Gradescope.

Question	Points	Score
Reviewing Recall, Precision and F1	15	
Understanding ROC and AUC	30	
TF-IDF Playground	20	
Getting your hands dirty with TF-IDF	35	
Total:	100	

## Problem 1: Reviewing Recall, Precision and F1 (15 Points)

(i) The following list of  $R$ s and  $N$ s represents relevant ( $R$ ) and non-relevant ( $N$ ) returned documents in a ranked list of 20 documents retrieved in response to a query from a collection of 10,000 documents. The top of the ranked list (the document the system thinks is most likely to be relevant) is on the left of the list. This list shows 6 relevant documents. Assume that there are 8 relevant documents in total in the collection.

**RRNNNNNNRRNRNNNRNNNNR**

2pts

(a) What is the precision of the system on the top 20?

5pts

(b) What is the F1 on the top 20?

(ii) Below is a table showing how two doctors, namely, MD1 and MD2, classified the atrial fibrillation in a set of 12 patients based on their ECG readings (0 = normal sinus rhythm, 1 = atrial fibrillation). Let us assume that you've designed an IR system that when asked to return the set of patients that exhibit abnormal sinus rhythm, it gives back the patients with IDs 4, 5, 6, 7, 8.

PatientID	MD1	MD2
1	0	0
2	0	0
3	1	1
4	1	1
5	1	0
6	1	0
7	1	0
8	1	0
9	0	1
10	0	1
11	0	1
12	0	1

4pts

(a) Calculate precision, recall, and F1 of your system if an ID is considered relevant only if the two doctors agree.

4pts
------

- (b) Calculate precision, recall, and F1 of your system if an ID is considered relevant if either of the two doctors agrees.

--

*Assignment continues on the next page(s)*

## Problem 2: Understanding ROC and AUC (30 Points)

You are trying to write an AI model to detect the presence of cancerous cells in your test subjects. As this process is somewhat involved, you decided to keep track of the different models you produce in every step of your parameter tuning. Due to a disk corruption, your carefully picked labels for your tests were compromised, and you lost track of which test was which. However, you roughly remembered that in one of the tests your threshold was too high, and in one of them it was too low, and in the subsequent tests you were getting better accuracy. As a student in 15-182 you know that this file corruption is an easy task to solve!

2pts

(a) Which of the trials had the lowest threshold?

2pts

(b) Which of the trials had the highest threshold?

6pts

(c) How do the rest of the trials compare to each other in terms of average recall and precision?

15pts

(d) Compute the AUC and ROC scores for the models?

5pts

(e) Moving forward, what might you want to try out to make an even better model?

### Problem 3: TF-IDF Playground (20 Points)

You learned that TF-IDF is a way to gauge the relevance of a word or a query to a collection of documents. The aim of this question is to demystify the elements of this metric and to help you think critically about how different modifications can be made to it.

2pts

- (a) What does it mean for an IDF score to be zero?

2pts

- (b) What does it mean for a term to have a high term frequency in a document?

2pts

- (c) How does doubling the count of a term in a document affect that term's IDF?

4pts

- (d) How does duplicating the corpus of documents change the TF and IDF scores of terms?

5pts

- (e) How does the quality of the ranking change by substituting  $\log(\cdot)$  to  $\log^2(\cdot)$  in computing the IDF score? How about using  $\log(\log(\cdot))$  instead?

5pts

- (f) You realize that the definition of term frequency might be a little too harsh, as the length of the documents could vary too much. As such, you choose to define the TF score as  $tf(f_{i_k}) = \frac{1}{2} + \frac{1}{2} \frac{f_{i_k}}{\max D_i}$  where  $f_{i_k}$  denotes the frequency that term  $T_k$  appears in document  $D_i$  and  $\max D_i$  denotes the frequency of the most common term in  $D_i$ . Discuss why you might want to consider this score instead.

### Problem 4: Getting your hands dirty with TF-IDF (35 Points)

6pts

- (a) Consider the table of term frequencies for 3 documents denoted as Doc1, Doc2, and Doc3 in the table below. Compute the tf-idf weights for the terms *cancer*, *diagnosis*, *insurance*, and *doctor* for each document, using the idf values from the table on the right.

	Doc1	Doc2	Doc3
cancer	27	4	24
diagnosis	3	33	0
insurance	0	33	29
doctor	14	0	17

	$idf_t$
cancer	1.65
diagnosis	2.08
insurance	1.62
doctor	1.5

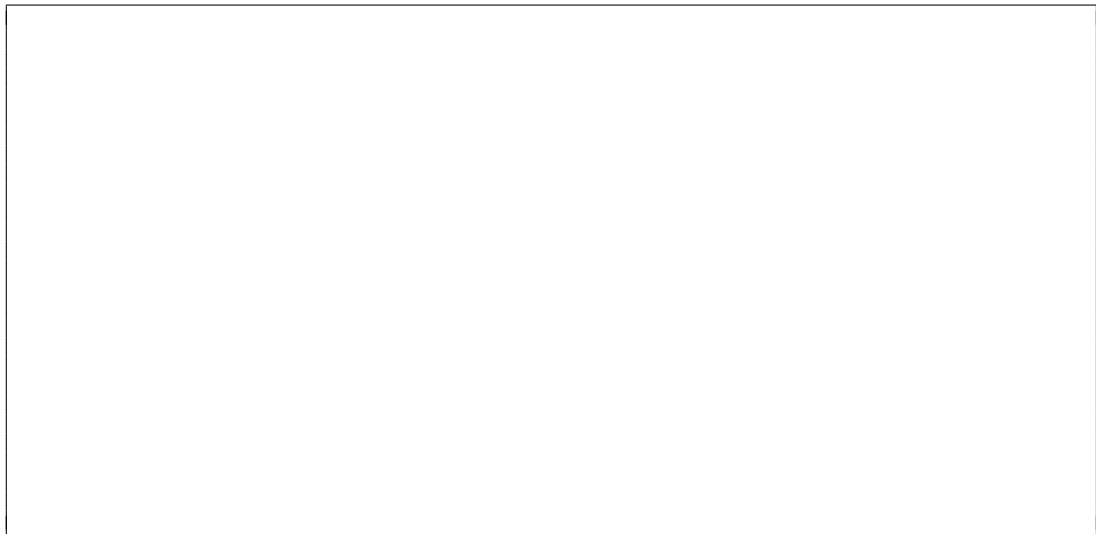
9pts

- (b) Rank the documents according to the computed tf-idf scores for the following three queries:

$$\begin{pmatrix} 2 \\ 0 \\ 3 \\ 10 \end{pmatrix}, \begin{pmatrix} 10 \\ 3 \\ 2 \\ 5 \end{pmatrix}, \begin{pmatrix} 10 \\ 10 \\ 4 \\ 0 \end{pmatrix}.$$

14pts

- (c) Normalize the vectors above and use cosine similarity to rank the documents for the given queries accordingly.



6pts

- (d) Was there a difference in ranking between parts (b) and (c)? If yes or no, explain why.

