IS2140 Midterm Exam Spring 2018

Name:_____ Examinee ID _____

First thing first, please save this file with your Examinee ID and your name as part of the file name.

You have **90 minutes** to complete this exam. Time will begin after we have read through the exam together.

- You may use your own personal copy of the course materials along with your class notes or the reference books that you have brought with you before the exam starts.
- You may NOT communicate with any other person during this exam (except the professor). You can use your laptop only for completing the exam questions, and cannot use it for any electronic communication during the exam. If it is found out, you will FAIL on this COURSE, not just the Exam.
- Write down your answers underneath each question.

As strategies for completing the exam, keep the following in mind:

- **Save your answers as often as you can!!!!!**
- If you find a question to be ambiguous, you may ask about it privately by coming to the front of the room. If the confusion is not resolved to your satisfaction, please explain your confusion along with your answer so that we can consider it during grading.
- You are more likely to get partial credit for a wrong answer if you show your work.
- Be careful not to get carried away and run over the time limit by spending too much time on one question. Plan ahead, and don't devote more time to a question than it is worth.

Score Summary (for use by grader)

Question	Possible points	Actual points
1	60	
2	40	
TOTAL	100	

1. Brief discussion and simple calculation (60%)

    a. (10%) Using your own words, describe what is known item search? Which evaluation measure is good for evaluating known item search, and why? **NOTE: must use your own words, cannot just copy and paste from the slides, must provide reason for the 2nd part of the question**

Your Answers:

For the definition of known item search, the answer needs to have

- **There is just one document in the collection satisfying the need**
- **User knows the existence of the document in the collection, the goal is to find it**

6% for the definition, each point worth 3%
If the student attempts to answer one point but fail to state it clearly, give it 1%

For the evaluation measure and the reason, the answer needs to have

- **The measure is MRR**
- **The reason is that this measure looks at the rank of the relevant document in the ranked list. Therefore, it consider one and only one relevant document in the evaluation, which makes it suitable for known item search**

4% for this part. 2% for identifying the measure, 2% for the reason.
If the student attempts to provide an explanation, but fail to state it clearly, give it 1% for the reason part of the score.

If any part of the answer is not in student's own words, no credits for that part.

    b. (10%) Use your own words, explain what is Zipf's law, and why it can be used to provide theoretic support for removing stop words in index? **NOTE: must use your own words, cannot just copy and paste from the slides**

Your Answers:

For the explanation of Zipf's law, the answer needs to have the following key points

- **If the unique words in a collection is ranked based on their frequency in the collection, we have the rank of the word R and the frequency of the word F**
- **The product of R*F is equal to a constant**

6% for the definition, each point worth 3%
If the student attempts to answer one point but fail to state it clearly, give it 1%

For the reason why it provides theoretic support for removing stopwords, the answer needs to have

- **A small number of words appear very frequently in the collection**
- **Studies show that these very frequent words are stopwords**
- **Remove them will greatly reduce the size of the index**

4% for this part. If all three points are clearly mentioned, give 4%. If missing each one – 1%.
If the student attempts to provide an explanation, but fail to state it clearly, -2%

If any part of the answer is not in student's own words, no credits for that part.

c. (10%) Using query likelihood model as the example, discuss why smoothing is important in statistical language models, and why Dirichlet smooth method is better than JM smooth. **NOTE: must use your own words, cannot just copy and paste from the slides. Must explain clearly why the advantages are advantages.**

Your Answers:
For why smoothing is important in statistical language models, the answer needs to have

- **Language model is estimated based on just one document, which does not provide enough sampling of the words that could appear in the language model**
- **There could be zero probability problem in performing retrieval using the language model**

5% for the smooth reasons, first point worth 3%, 2nd point worths 2%
If the student attempts to answer one point but fail to state it clearly, give it 1%

For the explanation of why Dirichlet is better than JM, the answer needs to have

- **JM has a parameter lambda which is trained before the search, and will be fixed during the search. The relative importance of the information from the document language model itself and that from the large collection is only affected by lambda.**
- **Dirichlet has also one parameber u, which is trained before the search, and will be fixed during the search. But the relative importance of the information from the document language model itself and that from the large collection is affected by u as well as by the length of the document itself.**
- **Therefore, Dirichlet is more accurate than JM in considering the document length for smoothing.**

5% for this part. If the student's answer somewhat capture the main points stated above but not all of them, -1%. If the answer attempts to provide an explanation, but fail to state clearly, give it 2%.

If any part of the answer is not in student's own words, no credits for that part.

d. (10%) We know that 10111100011111110011011 is encoded by Gamma code, and it corresponds to more than one binary codes. What are the binary codes? **NOTE: Must show the calculation steps.**

Your Answers:

There are three binary codes there

First one: length is 1, and offset is 1, so the binary code is 11
Second one: length is 3, which is 111, and offset is 001, so the binary code is 1001
Third one: length is 6, which 111111, and offset is 011011, so the binary code is 1011011

If can state that there are three binary codes, but fail to give more details +2%
If the length and the offset is correct, but fail to provide correct final binary codes, +6%
If just have the final code but no intermediary steps -7%

e. (10%) using your own words, explain the difference between implicit relevance feedback and pseudo relevance feedback. Then use an example from your own experience, explain why implicit relevance feedback could sometimes be unreliable. **NOTE: must use your own words.**

Your Answers:

For difference between implicit relevance feedback and pseudo relevance feedback, the answer needs to have

- **There is a user in implicit relevance feedback, although user is working on his/her own seeking of relevant information in the returned results**
- **There is no user at all in pseudo relevance feedback.**

5% for state clearly that the above difference
If the student attempts to answer but fail to state it clearly, give it 3%

For the example, the keys are that

- **The example should not be mentioned in the slides**
- **The example should be explained in own words**
- **The example should clearly indicate the unreliable in prediction**

If the student's answer violates any one of the above keys, -2%, any two of the above keys -4%.

f. (10%) use your own words, explain why Boolean model is called exact match model, then list at least two reasons why best match models like Vector Space model are viewed as better models for end web users who do not know much about the collection and the search. **NOTE: Must use your own words.**

Your Answers:

For why Boolean model is exact match, the answer needs to have

- **It assumes that the query imposes hard conditions that the returned documents have to satisfy fully.**
- **Documents are divided into two mutually exclusive parts. One group contains all documents fully satisfying the query. The other one contains all documents cannot full satisfying the query. There is no partial matching between the document and the query**

5% for state clearly the above two points
If the student attempts to answer but fail to state it clearly, give it 3%

For the two reasons that best match model is better

- **Best match model does not assume a perfect query from the user. So make it easier for end web users to formulate queries**
- **Best match model can handle partial match between the documents and the query**
- **Best match model considers the frequency of the words in a document, more suitable to model web pages with full content searchable**
- **Best match model generates a ranked list of returned documents, The example should not be mentioned in the slides**

5% for this part. Any two of the above four reasons would get 5%. If missing one -2%. If it is not own words, but copied from the slides, -3%.

2. Indexing, Retrieval Models and Relevance Feedback (40%)

Suppose we have a collection of 7 documents, and there are only 10 unique index terms after removing stopwords in the index we created (see below). The last row shows the original length of the documents including the stopwords.

Table 1: new index for language model, and the value in each cell is the raw count term frequency of the term in the document

Vocabulary	D1	D2	D3	D4	D5	D6	D7
amazon	0	0	0	0	5	9	0
compani	2	9	8	4	0	1	4
damag	3	1	1	1	3	1	1
electron	1	0	5	0	4	6	2
fallen	0	0	2	7	3	0	0
grown	1	3	0	4	10	1	8
hit	3	3	6	0	1	6	2
newton	2	0	0	5	5	1	1
product	1	3	3	0	1	1	0
reviv	0	0	0	8	0	1	0
Document Length	23	27	30	36	40	30	24

a. (8%) using the index words above as the examples, discuss why we need to consider both term frequency and inverse document frequency when modeling the importance of terms? **You must use the words in the index above as the example, you must use your own words to explain. You cannot just copy and paste from the slides**.

Your Answers:

Possible words used to discuss term frequency are amazon in D6, compani in D2 and D3, grown in D5 etc. the more frequent a term in a document, the higher chance the document is about the topic represented by the term

Possible words used to discuss inverse document frequency are compani appearing in 6 out of 7 documents, damag appearing in all 7 documents, amazon just appear in 2 out of 7 documents, reviv just appear in 2 out of 7 documents. A term appearing in too many documents in the collection would not have the power to differentiate one document from another in ranking. A term only appears in small number of documents would give strong differentiation power.

Each of the above discussion is 4%. If there is no example of words from the index -2%. If the explanation is not clear -1%. If does not use own words -2%

b. (12%) suppose the weight of term *i* in document *j* is calculated as $w_{ij}=TF_{ij}*(N/DF_i)$ where TF_{ij} is the raw count of term frequency of term *i* in document *j*, *N* is 7, and DF_i is the document frequency of term *i* in the collection, what is the vector of document D3? If we know that a query *q* contains four terms after stemming, which are "amazon, forest, damag, grown," and the query weight for each word is 2, 1, 2, 1, respectively, what is the query vector? What is the similarity between the query vector and the D3 vector? **You can ignore the length normalization for both the document vector and the query vector in the calculation. You must show as detail calculation as possible**.

Your Answers:

Vocabulary	D3	TF	DF	N/DF	W	Q	
amazon	0	0	2	3.5	0	2	
compani	8	8	6	1.167	9.336	0	
damag	1	1	7	1	1	2	
electron	5	5	5	1.4	7	0	
fallen	2	2	3	2.333	4.666	0	
grown	0	0	6	1.167	0	1	
hit	6	6	6	1.167	7.002	0	
newton	0	0	5	1.4	0	0	
product	3	3	5	1.4	4.2	0	
reviv	0	0	2	3.5	0	0	
Document Length							

So Vector for D3 is <0,9.336, 1, 7, 4.666, 0, 7.002, 0, 4.2, 0>

Query vector is <2,0,2,0,0,1,0,0,0,0>

Similarity = <0,9.336, 1, 7, 4.666, 0, 7.002, 0, 4.2, 0> * <2,0,2,0,0,1,0,0,0,0>
= 0*2+9.336*0+1*2+7*0+4.666*0+0*1+7.002*0+0*0+4.2*0+0*0
= 2

The key for grading is that if the answer is clear that students know what is TF, what is DF, and the calculation for TF, IDF and thus the weight is basically right, thus has a correct D3 vector +6%.
If the calculation is right, just a few numbers are not correct -1%

The query vector should clearly on the same set of words as the document vectors. If the query vector is just one the query words only -3%. The word "forest" should not appear in the query vector. If it appears -1%. Total for this part is 4%

The calculation of the similarity is 2%. If the calculation is correct, but the outcome is not -1%.

c. (10%) Suppose now we know that D5 is a relevant document specified by the user, and D3 is a non-relevant document. Using the Rocchio relevance feedback formula where α=1, β=0.7, and γ=0.2, calculate what is the new query vector after the relevance feedback. **You must show as detail calculation as possible**.

Your Answers:

Vocabulary	D5	TF	DF	N/DF	W		
amazon	5	5	2	3.5	17.5		
compani	0	0	6	1.167	0		
damag	3	3	7	1	3		
electron	4	4	5	1.4	5.6		
fallen	3	3	3	2.333	6.999		
grown	10	10	6	1.167	11.67		
hit	1	1	6	1.167	1.167		
newton	5	5	5	1.4	7		
product	1	1	5	1.4	1.4		
reviv	0	0	2	3.5	0		
Document Length							

So D5 vector is <17.5, 0, 3, 5.6, 6.999, 11.67, 1.167, 7, 1.4, 0>

QM = 1 * <2,0,2,0,0,1,0,0,0,0> + 0.7 * <17.5, 0, 3, 5.6, 6.999, 11.67, 1.167, 7, 1.4, 0> - 0.2 * <0,9.336, 1, 7, 4.666, 0, 7.002, 0, 4.2, 0> = <2,0,2,0,0,1,0,0,0,0> + <12.25, 0, 2.1, 3.92,4.899,8.169, 0.817,4.9,0.98,0> - <0, 1.867, 0.2, 1.4, 0.933, 0, 1.4, 0, 0.84, 0> = <14.25, -1.867, 3.9, 2.52, 3.966, 8.169, -0.583, 4.9, 0.14, 0>

Vocabulary	**Q**	D5	**0.7*D5**	D3	**0.2*D3**	**QM**	
amazon	**2**	17.5	**12.25**	0	**0**	**14.25**	
compani	**0**	0	**0**	9.336	**1.867**	**-1.867**	
damag	**2**	3	**2.1**	1	**0.2**	**3.9**	
electron	**0**	5.6	**3.92**	7	**1.4**	**2.52**	
fallen	**0**	6.999	**4.899**	4.666	**0.933**	**3.966**	
grown	**1**	11.67	**8.169**	0	**0**	**8.169**	
hit	**0**	1.167	**0.817**	7.002	**1.4**	**-0.583**	
newton	**0**	7	**4.9**	0	**0**	**4.9**	
product	**0**	1.4	**0.98**	4.2	**0.84**	**0.14**	
reviv	**0**	0	**0**	0	**0**	**0**	

Document Length							

As long as students demonstrate the understanding of calculation for Rochioo, we should try to give them score. If it is just careless on a few spots for miscalculation -1 or -2%.

If there is no steps for the calculation at all -6%

d. (10%) Suppose based on the ground truth, the precision and recall at each rank position for the query after the relevance feedback is shown below. How many relevance documents are there in the collection for the query? At which rank positions are the relevance documents? What is the average precision for this ranked list? **You must show as detail calculation as possible**.

Rank	Recall	Precision
1	0.2	1.000
2	0.2	0.500
3	0.4	0.667
4	0.6	0.75
5	0.6	0.600

Your Answers:

1) **With 1 relevant document, its recall value is 0.2, therefore it has 5 relevant documents. If this is wrong -3%.**

2) **The rank positions of relevant documents are 1, 3, 4. If missing each one -1%. Total 3%**

3) **Average precision = (1+0.667+0.75)/5 = 0.4834. If it is divided by 3, then -2%**