# IS2140 Midterm Exam Fall 2018

Name:\_\_\_\_\_\_\_\_\_\_Junxi Huang\_\_\_\_\_\_\_\_\_\_ Examinee ID \_\_\_\_\_\_\_34\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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

You have **100 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 TAs in the room). 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%)
   1. (10%) Using your own words, describe what is topical relevance? Which evaluation framework is designed to perform evaluation based on topical relevance idea, 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:**

* + 1. It is object which is relative with document content. It is mainly used in connect with incoming links. Websites which get similar content have topical relevance.

1. Cranfield evaluation, it is good in s set of search request, which means in one document
   1. (10%) Use your own words, explain what is inverse document frequency, and why it is important in vector space model to consider inverse document frequency? **NOTE: must use your own words, cannot just copy and paste from the slides**

**Your Answers:**

* Document frequency means how often a term show in documents, which means how many documents get that term in all documents. Inverse means use 1 to modify document frequency. Big inverse document frequency means that term is more important and get more information.
* Big inverse document frequency means that term is more important and get more information.
  1. (10%) Our task is to use Gamma code to encode three binary codes. The first one is 110111, the second one is 111, and the third one is 1000001. What is the final one long code that contains all the Gamma codes of the three binary codes? **NOTE: Must show the calculation steps.**

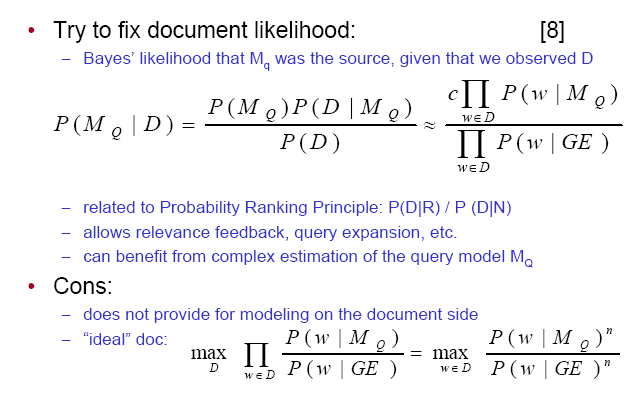
**Your Answers:**

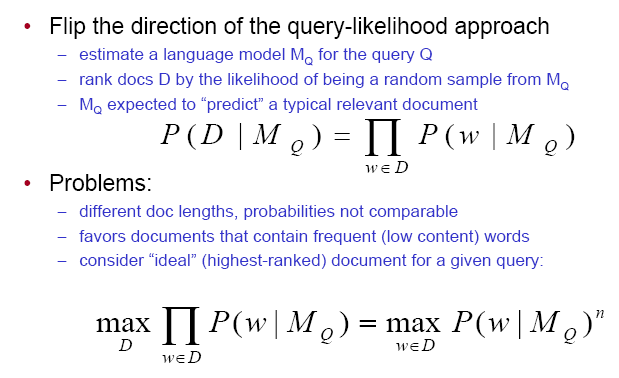
The first one:110111->10111, 5->1001->10010 ans:1001010111

The second one:111->11, 2->10->100 ans:10011

The third one: 1000001->000001, 6->1010->10100 ans:10100000001

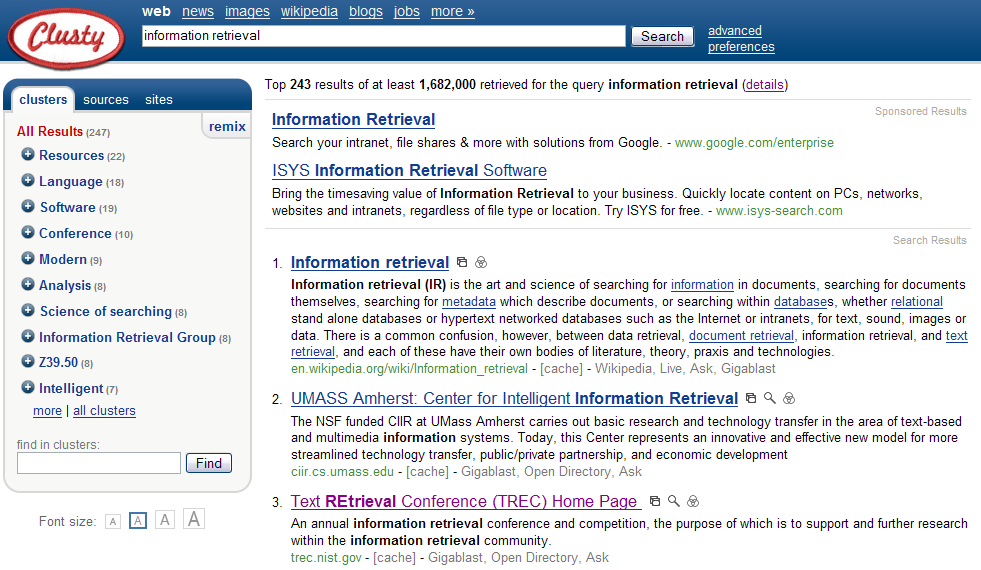
* 1. (10%) Using your own words, discuss two benefits of using formula (1) over (2) in retrieving documents using language model. **NOTE: must use your own words, cannot just copy and paste from the slides. Must explain clearly why they are the benefits.**

**** (1)

**** (2)

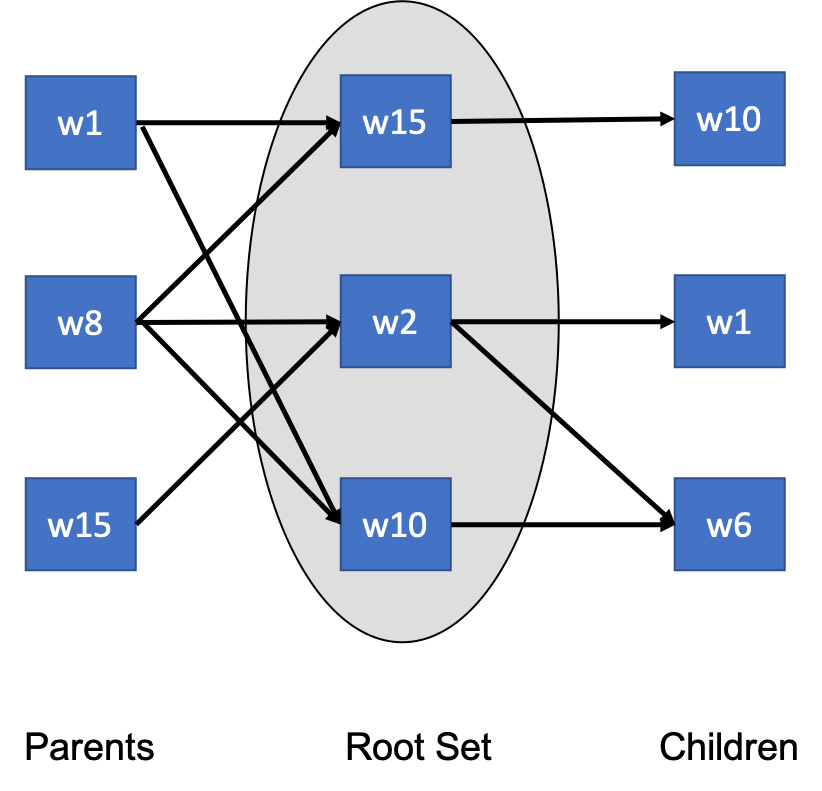
**Your Answers:**

* **It is easy to get relevance feedback**
* **It is easy to make a query expansion**
* **It will be easy in complex estimation of query model**
* **Got more flexible than query-likehood and document-likehood** 
  1. (10%) using your own words, explain one advantage and one limitation of the search interface below in comparison to a standard Google Search interface. Then propose one possible improvement on this interface, make an argument why your idea is an improvement. **NOTE:** **must use your own words.**



**Your Answers:**

* **Advantage: it shows the result in many different domain, which is more easy to get the result which is customer wanted.**
* **disadvantage: the results are shown are too small and it should show more results.**
* **improvement: add a year range which customer can select. For example, customer can choose only show result in 2014 to 2016**
  1. (10%) Below is the webpages needed for calculating HITS. Each blue square represents a webpage, and the text string in it is the id of the webpage. The root set contains three webpages: w15, w2, w10, and the parent part of the root set also contains three webpages: w1, w8, w15, where w15 also appears in the root set. The children part contains w10, w1, w6. Based on this information, write down the L matrix of this subset of the web. **Tip: you may want to draw the graph on a draft paper first, but you do not have to show your draft graph in this document.**



**Your Answers:**

The L Matrix is

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | W1 | W2 | W6 | W8 | W10 | W15 |
| W1 | 0 | 0 | 0 | 0 | 1 | 1 |
| W2 | 1 | 0 | 1 | 0 | 0 | 0 |
| W6 | 0 | 0 | 0 | 0 | 0 | 0 |
| W8 | 0 | 1 | 0 | 0 | 1 | 1 |
| W10 | 0 | 0 | 1 | 0 | 0 | 0 |
| W15 | 0 | 1 | 0 | 0 | 1 | 0 |

1. 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 : 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 |

* 1. (8%) suppose we use maximum likelihood estimation to estimate a language model using D5, what are the top 3 words according to the probability that it is generated from the language model of D5? **Note:** **You must show the calculation steps and explain your answer**

**Your Answers:**

PMLE(amazon|D5) = #(5) / |40|=0.125

PMLE(compani|D5) = #(0) / |40|=0

PMLE(damag|D5) = #(3) / |40|=0.075

PMLE(electron|D5) = #(4) / |40|=0.1

PMLE(fallen|D5) = #(3 )/ |40|=0.075

PMLE(grown|D5) = #(10) / |40|=0.25

PMLE(hit|D5) = #(1) / |40|=0.025

PMLE(newton|D5) = #(5) / |40|=0.125

PMLE(procut|D5) = #(1) / |40|=0.025

PMLE(reviv|D5) = #(0) / |40|=0

The top three is grown,amazon,newton

* 1. (12%) suppose a query *q* contains five terms after stemming, which are “amazon, forest, damag, grown, reviv” and the query weight for each word is 2, 1, 2, 1,1, respectively, if the language model calculation uses query likelihood model, and the smoothing is JM smoothing with alpha=0.8, what is the probability of the query generated from the language model of D5? **Note:** **You must show as detail calculation as possible**.

**Your Answers:**

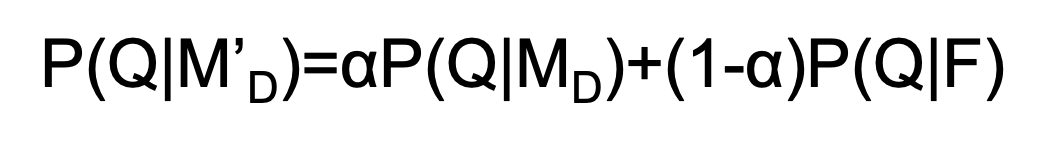
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Vocabulary | D5 | TF | DF | N/DF | W | Q |
| amazon | 5 | 5 | 2 | 3.5 | 16.5 | 2 |
| compani | 0 | 0 | 6 | 1.167 | 0 | 0 |
| damag | 3 | 3 | 7 | 1 | 3 | 2 |
| electron | 4 | 4 | 5 | 1.4 | 6.4 | 0 |
| fallen | 3 | 3 | 3 | 2.333 | 6.999 | 0 |
| grown | 10 | 10 | 6 | 1.167 | 11.67 | 1 |
| hit | 1 | 1 | 6 | 1.167 | 1.167 | 0 |
| newton | 5 | 5 | 5 | 1.4 | 7 | 0 |
| product | 1 | 1 | 5 | 1.4 | 1.4 | 0 |
| reviv | 0 | 0 | 2 | 3.5 | 0 | 1 |
| Document Length | 40 |  |  |  |  |  |

p(q|MD)=0.125\*0.25\*0.075=0.00234

p(q|Mc)=2/7\*2/7\*1/7\*1/7=0.00166

p(t|MD) = λ\*p(t|D)+(1-λ)\*p(t|C)

=0.8\*0.00234+0.2\*0.00166=0.002205

* 1. (10%) Suppose now we know that D6 is a relevant document specified by the user. Using formulafor relevance feedback where this α=0.6, calculate what is the new probability that the query is generated from the language model of D5 after the relevance feedback. **Note: You must show as detail calculation as possible. For simple calculation here, you can only use MLE to estimate the language model for D6, even though in reality you would smooth D6 too.**

**Your Answers:**

* 1. (10%) Suppose based on the ground truth, a query has 3 relevant documents, they are D1, D4 and D7. A search engine returns a ranked list with 5 documents, and the ranking are D1, D2, D3, D4 and D5. What is precision@5 for this result? What is R-precision for this result? What is the average precision for this ranked list? If by choosing a different relevance judgment method, we now know that the relevance score for the five returned documents is 3, 0, 0, 2, 0, respectively. What is the DCG@5 for this ranked list? **You must show as detail calculation as possible**.

**Your Answers:**

|  |  |  |
| --- | --- | --- |
|  | relevant | Not-relevant |
| retrieved | 2 | 3 |
| Not retrieved | 1 |  |

1. Precision: 2/(2+3)=0.4
2. Recall: 2/(2+1)=0.666
3. Average precision: (1+0.5)/2=0.75
4. Dcg= (2^3-1)/log(1+1) +(2^2-1)/log(1+4)=7+1.29=8.29