CS172: Information Retrieval

UC Riverside

Midterm 1 Review Fall 2019

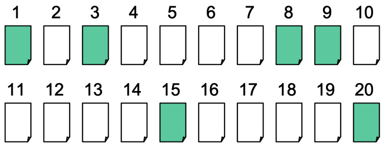
This is meant to be a study guide for Midterm 1. Please note, you should use this to practice concepts but don't limit yourself to only studying these questions and be sure to also be familiar with the content covered in class and in the chapter readings.

- know how text retrieval is different from database retrieval
- know the basic architecture of a text retrieval system
- know what is tokenization, stemming, what is a stop word
- know what is an inverted index and how to score documents quickly using index
- know how to measure relevance using binary relevance and vector space models
- know the major term weighting heuristics (TF, IDF, and document length normalization). Remember how TF-IDF is computed
- understand probabilistic ranking principle
- know Bayes rule and chain rule in probability
- know what is a statistical language model, what is a unigram/bigram language model
- know why smoothing is necessary when estimating a language model and know the general idea of smoothing
- be familiar with formulas for Add-1 smoothing, Dirichlet prior smoothing, and linear interpolation smoothing and their similarities and differences
- know how to compute the basic retrieval evaluation measures (e.g., precision, recall, and mean average precision, MRR, F1). You need to remember the formulas
- know Zipf's and Heap's law
- know evaluation metrics and why one metric is better than another.

Note, you should also study the handouts in class!

1 Evaluation

1. An information retrieval system returns the following ranked list for a particular query:



Colored blocks represent relevant documents; white blocks represent irrelevant documents. There are **eight** relevant documents in total.

(b) What is the Precision at 10, i.e. the precision after the 10th document is retrieved?

(c) What is the Recall at 16, i.e. the recall after the 16th document is retrieved?

(d) What is the harmonic mean?

(e) Why is accuracy a misleading measure in ranking (i.e. why is precision and recall more useful)?

2 Boolean and Vector Space Retrieval

2. Assume a Vector Space Retrieval System that uses tf.idf weighting scheme with vector length normalization. Suppose your collection contained the document 'The rain in Spain ...' You later discovered that some person had written rain twice, which you corrected. To fix the index, you need to recalculate all the vectors of every document in the collection. TRUE or FALSE?

3. We have a search engine with a corpus containing only three documents. Assume stop-words are NOT removed.

   - Document ID 1: 'A time to plant and a time to reap'
   - Document ID 2: 'Time for you and time for me'
   - Document ID 3: 'Time flies'

   - Given that the stemmed version of the word 'flies' is the term 'fly', what is the TF-IDF of 'fly' in document 3?

   - What is the TF-IDF of 'time' in document 1?

3 Language Models

4. Define a statistical language model. A statistical language model (LM) is a distribution over word sequences. Intuitively, it gives us a probability for any sequence of words, thus allows us to compare two sequences of words to see which has a higher probability. In general, LMs help capture the uncertainties associated with the use of natural language. For example, in general, non-grammatical sentences would have much smaller probabilities than grammatical sentences. Specialized language models can be used to answer many interesting questions that are directly related to many information management tasks. It has applications in speech recognition, and machine translation.

5. The simplest kind of Language Model is the unigram language model. Explain.

   This model corresponds to a multinomial distribution over words. According to this model, a piece of text is 'generated' by generating each word independently. As a result, the joint probability of generating all the words in a document $D = w_1 w_2 \ldots w_n$ is simply the product of generating each individual word, i.e., $p(D) = p(w_1)p(w_2)(w_n)$.

   With a bigram Language Model, we'd have $p(D) = p(w_1)p(w_2|w_1)p(w_3|w_2)\ldots p(w_n|w_{n-1})$, which would capture local dependency between two adjacent words.

6. Smoothing is necessary because otherwise the model would assign a zero probability to queries that contain terms not present in the original document(from which the model was built). TRUE or FALSE

7. Give a reason why Dirichlet Prior smoothing is preferred than Add-1 smoothing.

8. Dirichlet prior smoothing and retrieval. Suppose we have a document collection with an extremely small vocabulary with only 6 words $w_1, \ldots, w_6$. The following table shows the estimated reference language model $p(w\|REF)$ using the whole collection of documents (2nd column) and the word counts for document d1 (3rd column) and d2 (4th column), where c(w, d_i) is the count of word w in document di . Let Q = $w_1 w_2$ be a query.

Word	$p(w\|REF)$	$c(w, d_1)$	$c(w, d_2)$
$w1$	0.8	2	7
w_2	0.1	3	1
w_3	0.025	2	1
w_4	0.025	2	1
w_5	0.025	1	0
w_6	0.025	0	0
SUM	1.0	10	10

(a) Suppose we do not smooth the language model for d_1 and d_2. Compute the likelihood of the query for both d_1 and d_2, i.e., $p(Q|d_1)$ and $p(Q|d_2)$. Show your calculations. Which document would be ranked higher?

(b) Suppose we now smooth the language model for d1 and d2 using Dirichlet prior smoothing method with $\mu = 10$. Recompute the likelihood of the query for both d_1 and d_2, i.e., $p(Q|d_1)$ and $p(Q|d_2)$. Note, the Dirichlet prior is given as $\frac{N}{N+\mu} * p(Q|D) + \frac{\mu}{N+\mu} * p(Q|REF)$, where N is the number of documents. Show your calculations. Which document would be ranked higher this time?

4 Similarity Measures

9. Define cosine similarity.

10. This problem deals with the vector space model (using TF-IDF weights) as applied to the following collection of 4 documents:

Doc 1 : Information Retrieval Systems
Doc 2 : Information Storage
Doc 3 : Digital Speech Synthesis Systems
Doc 4 : Speech Filtering, Speech Retrieval

(i) Compute all non-zero entries in the normalized vector for Doc 1.

	IDF	TF DOC_1	TF DOC_2	TF DOC_3	TF DOC_4
Information					
Retrieval					
Systems					
Digital					
Storage					
Speech					
Synthesis					
Filtering					

(ii) Rank all the documents in the collection for the query ‘Speech Systems’ using Cosine Similarity?

5 Probabilistic Ranking

11. What is the Probability Ranking Principle ? Why is it difficult to achieve its promise?

12. What is the Bayes rule?

13. Given the following data which describes three features (words) in a document (W_1, W_2, W_3) and the document outcome (whether its relevant R=1 or non-relevant NR=0)

	W_1	W_2	W_3	R or NR
Doc1	1	1	1	0
Doc2	1	1	0	0
Doc3	0	0	0	0
Doc4	0	1	0	1
Doc5	1	0	1	1
Doc6	0	1	1	1

Please apply the Naive Bayes classifier to this data set and compute $P(R = 1|W)$ for W = (1,0,0), i.e. assume that you have a document that contains W_1 but does not contain W_2 or W_3 and you would like to compute the probability that this document is relevant. Recall that the Naive Bayes classifier assumes the features are independent. Show your work, but you can leave your solution as fractions

Hint. Use Bayes rule to compute $P(R = 1|W_1)$, $P(R = 1|W_2)$, and $P(R = 1|W_3)$. Bayes rule is $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

6 Misc

In class, we discussed some simplifying assumptions that are made by modern document retrieval systems. For each assumption listed, briefly provide an example why it is not true in real-world information seeking environments. Each example should not take more than one sentence. You can give examples to make your point.

(a) Binary relevance

(b) Document independence

(c) Term independence

Write the formula for Zipfs Law (define your terms). Give one practical example of its use (i.e., a situation where it would be useful)