Exercises

The following list of exercises is a selection from textbooks and previous years' exams.

Parsing and evaluation

1. Are the following statements true or false?
    a. In a Boolean retrieval system, stemming never lowers precision.
    b. In a Boolean retrieval system, stemming never lowers recall.
    c. Stemming increases the size of the vocabulary.
    d. Stemming should be invoked at indexing time but not while processing a query.

2. Consider the tf-idf term weighting.
    a. What is the *idf* of a term that occurs in every document? Compare this with the use of stop word lists.
    b. Can the *tf-idf* weight of a term in a document exceed 1?

3. Assume a biword index. Give an example of a document which will be returned for a query of *New York University* but is actually a false positive which should not be returned.

4. Consider an information need for which there are 4 relevant documents in the collection. Contrast two systems run on this collection. Their top 15 results are judged for relevance as follows (the leftmost item is the top ranked search result):

System 1	R N R N N N N N R N N N N N R
System 2	N R N N R N N N R N N R N N N

    a. What is the MAP of each system? Which has a higher MAP?
    b. Does this result intuitively make sense? What does it say about what is important in getting a good MAP score?
    c. What is the Recall after 10 retrieved documents of each system?
    d. Plot the precision-recall curve for both systems.

5. Below is a table showing how two human judges rated the relevance of a set of 12 documents to a particular information need (0 = nonrelevant, 1 = relevant). Let us assume that you've written an IR system that for this query returns the set of documents {4, 5, 6, 7, 8}.

docID	Judge 1	Judge 2
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

   a. Calculate the kappa measure between the two judges.

   b. Calculate precision, recall, and F1 of your system if a document is considered relevant only if the two judges agree.

   c. Calculate precision, recall, and F1 of your system if a document is considered relevant if either judge thinks it is relevant.

6. Consider two ranking algorithms that for the same query produced the two following ranks:

S1: d4 d3 d5 d8 d2 *and* *S2: d3 d9 d5 d6 d1*

   a. Assuming that the relevant documents are d9, d1, d3 and d4, compute the precision and recall values of each system.

   b. Assuming that the multi-value relevance judgments of documents are d9=1, d1=1, d3=3 and d4=2, assess and compare the two ranks with the appropriate metric.

   c. Assume no relevance judgments and compare the two systems.

Retrieval models

7. Consider the table of term frequencies for 3 documents denoted Doc1, Doc2, Doc3. The collection contains 750,000 documents in total.

	docFrequency	Doc1	Doc2	Doc3
car	18,165	27	4	24
auto	6,723	3	33	0
insurance	19,241	0	33	29
best	25,235	14	0	17

   a. Compute the tf-idf weights for the terms car, auto, insurance, best, for each document.

   b. Compute the rank of the three documents for the query "auto insurance" on the vector space model.

8. Consider a retrieval system with TF-IDF weighting and cosine ranking. The repository has a total of 1000 documents.

   a. Compute the similarity between the two documents.

> *A: "This update is designed to reduce rankings for low-quality sites—sites which are low-value add for users, copy content from other websites or sites that are just not very useful."*
>
> *B: "We can't make a major improvement without affecting rankings for many sites. It has to be that some sites will go up and some will go down."*
>
> Source: http://googleblog.blogspot.com/2011/02/finding-more-high-quality-sites-in.html

   b. Which document is the most relevant for the query "ranking sites"?

   c. The TF-IDF weighting is composed by two parts. Explain the motivation for each part and detail how they are combined.

9. Consider the two following documents:

> *d1 : Jackson was one of the most talented entertainers of all time*
>
> *d2: Michael Jackson anointed himself King of Pop*

   a. Using a BM25 retrieval model determine which document is more relevant to the query q= "Michael Jackson" (consider b = 0.75 and k = 1.5).

   b. Using a language model with Jelinek-Mercer smoothing determine which document is more relevant to the query q= "Michael Jackson" (consider $\lambda = ½$)

   c. Using a language model with Dirichlet smoothing determine which document is more relevant to the query q= "Michael Jackson" (consider $\mu = 100$)

10. Consider the retrieval models that are derived from the Probability Ranking Principle.
    a. Explain the Probability Ranking Principle.
    b. Relate the BIM and the BM25 retrieval models with the Probability Ranking Principle.
    c. Relate the family of Language Models for retrieval with the Probability Ranking Principle.

11. Information Retrieval systems can implement several different weighting schemes and ranking functions.
    a. Relate the Binary Independence Model to the Inverted Document Frequency.
    b. What are the differences between standard vector space TF-IDF weighting and the BIM probabilistic retrieval model?
    c. Discuss the differences between the following term weighting functions: i) Binary; ii) frequency, and iii) tf-idf.
    d. Discuss the differences between the following ranking functions: i) Euclidean distance, ii) cosine distance and iii) BM25.

12. Show that models resulting from Dirichlet smoothing can be treated as probability distributions. That is, show that $\sum_t M_d^u(t) = 1$.

13. You have discovered that documents in a certain collection have a "half-life" of 30 days. After any 30-day period a document's prior probability of relevance p(r|D) is half of what it was at the start of the period. Incorporate this information into LMD. Simplify the equation into a rank-equivalent form, making any assumptions you believe reasonable.

14. Let X_t be a random variable indicating whether the term t appears in a document. Suppose we have $|R|$ relevant documents in the document collection and that $X_t = 1$ in s of the documents. Take the observed data to be just these observations of $X_t$for each document in R. Show that the MLE for the parameter $p_t = P(X_t = 1|R = 1, \sim q)$, that is, the value for $p_t$which maximizes the probability of the observed data, is $p_t = \frac{s}{|R|}$.

15. Consider making a language model from the following training text:

    the martian has landed on the latin pop sensation ricky martin

    a. Under a MLE-estimated unigram probability model, what are P(the) and P(martian)?
    b. Under a MLE-estimated bigram model, what are P(sensation|pop) and P(pop|the)?

16. Suppose we have a collection that consists of the 4 documents given in the below table.

docID	Document text
1	click go the shears boys click click click
2	click click
3	metal here
4	metal shears click here

Build a query likelihood language model for this document collection. Assume a mixture model between the documents and the collection, with both weighted at 0.5. Maximum likelihood estimation (mle) is used to estimate both as unigram models.

a. Work out the model probabilities of the queries "click", "shears", and hence "click shears" for each document, and use those probabilities to rank the documents returned by each query.

b. What is the final ranking of the documents for the query click shears?

17. Using the calculations in Exercise 17 as inspiration or as examples where appropriate, write one sentence each describing the treatment that the LM with Jelinek-Mercer smoothing gives to each of the following quantities. Include whether it is present in the model or not and whether the effect is raw or scaled.

a. Term frequency in a document

b. Collection frequency of a term

c. Document frequency of a term

d. Length normalization of a term

18. In the mixture model approach to the query likelihood model,

$$p(q|d,C) \approx \prod_{t\in\{q\cap d\}}\big(\lambda\cdot p(t|M_d)+(1-\lambda)\cdot p(t|M_c)\big),$$

the probability estimate of a term is based on the term frequency of a word in a document, and the collection frequency of the word. Doing this certainly guarantees that each term of a query (in the vocabulary) has a non-zero chance of being generated by each document. But it has a more subtle but important effect of implementing a form of term weighting, related to TF-IDF that was discuss in (Manning et al., Chapter 6). Explain how this works. In particular, include in your answer a concrete numeric example showing this term weighting at work.

Relevance Models

19. Suppose that a user's initial query is "ranking sites". The user examines two documents, A and B (the same from the previous question). She judges A, relevant and B nonrelevant. Assume that we are using direct term frequency (with no scaling and no document frequency). There is no need to length-normalize vectors.

    A: "This update is designed to reduce rankings for low-quality sites—sites which are low-value add for users, copy content from other websites or sites that are just not very useful."
    B: "We can't make a major improvement without affecting rankings for many sites. It has to be that some sites will go up and some will go down."
    Source: http://googleblog.blogspot.com/2011/02/finding-more-high-quality-sites-in.html

    a. Using Rocchio relevance feedback what would the revised query vector be after relevance feedback? Assume $\alpha = 1$, $\beta = 0.75$, $\gamma = 0.25$.

    b. Discuss the limitations of the Rocchio algorithm.

Learning to rank

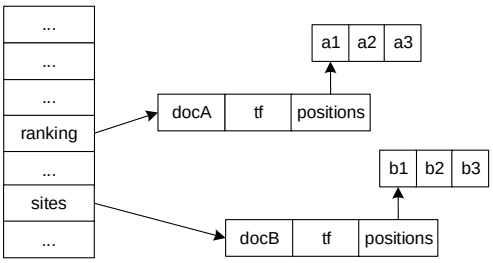
20. Suppose that your boss asks you to develop a test collection for a corporate search engine. The company wants the test collection to be useful for the next 2-3 years. Describe how you would build or acquire the different test collection components, and how much data is required for each component. Your answer needs to be practical, i.e., no magic, and your budget isn't infinite.

21. Rank fusion methods combine ranks in different manners. Compute the fused ranks for the following three lists with the CombSUM, CombMNZ, BordaFuse and RRF.

Rank 1 (id/score)	Rank 2 (id/score)	Rank 3 (id/score)
D3 / 0.5	D3 / 0.8	D9 / 0.9
D4 / 0.2	D8 / 0.8	D3 / 0.8
D2 / 0.19	D2 / 0.8	D1 / 0.7
D5 / 0.18	D1 / 0.5	D8 / 0.6
D6 / 0.07	D5 / 0.4	D2 / 0.5
D1 / 0.05	D6 / 0.32	D5 / 0.4
D7 / 0.01	D9 / 0.31	D6 / 0.3
D9 / 0.01	D7 / 0.30	D7 / 0.2

Indexing

22. Consider an indexing system implementing a specific postings structure including, the weight and the occurrence positions within the document.



a. Propose a term weighting scheme and a ranking algorithm considering the occurrence position information.

b. Modify the Block Sort-Based Indexing method to generate the above index structure with positional indexing.

23. Consider the indexing phase of a search engine.

a. Discuss the main differences between the BSBI and SPIMI indexing algorithms.

b. Describe an algorithm to implement the SPIMI algorithm on the Map-Reduce architecture.