INLS 509 Information Retrieval
Assignment 1

Due: Monday, August 31 (by 5pm) via Sakai

1: Survey of a Controlled Vocabulary (20%)

So far, we've discussed two ways of representing documents within a collection to facilitate information access: (1) *controlled vocabularies* and (2) *free-text indexing*. This question is about controlled vocabularies.

Find an information access provider or a digital library online that uses a controlled vocabulary, possibly in conjunction with free-text indexing. An example would be Wikipedia (you can use Wikipedia, but are encouraged to find another). Then, answer the following questions.

(a) Provide a description of the controlled vocabulary. Approximately how many concepts does it contain? How are the concepts related to each other, if at all? For example, are the concepts organized in a hierarchy? If so, can concepts appear in multiple places in the hierarchy?

(b) How is the controlled vocabulary maintained? For example, who maintains it? How often is it modified? How are changes to the controlled vocabulary initiated? Are there ways in which the organization detects missing, underutilized, or confusing concepts? How are changes to the controlled vocabulary communicated to users?

(c) How is the controlled vocabulary used to facilitate information access? Is it used for navigation? Query-based search? Both? How do users find controlled vocabulary concepts they may not know of?

(d) In your opinion, what are some strengths and limitations of this particular controlled vocabulary?

You may not find exact answers to all these questions. For example, it may be difficult to determine exactly how often a controlled vocabulary is re-visited and modified. If you can't find the exact information, make an informed guess based on the current state of the controlled vocabulary, the nature of the collection, and the user community.

2: Precision and Recall (15%)

Precision and recall are two metrics that can be used to evaluate a set of unranked results. These metrics consider differences between the set of documents *retrieved* for a given query and the set of documents that are *relevant* to the user's need. Answer the following questions about precision and recall.

(a) Compute precision and recall for the following retrieval:

- relevant documents: 4, 25, 39, 63, 769, 1563
- retrieved documents: 4, 26, 38, 63, 569, 769, 790, 1565, 1589

The numbers correspond to the document id's.

(b) Precision and recall are often discussed together because they focus on complementary information. If precision is important, the user does not want to see any non-relevant documents. That is, whatever is retrieved, should be relevant. If recall is important, the user wants to see all the relevant documents, even if it requires sifting through some non-relevant ones.

Give two information-seeking tasks where you think precision is more important than recall. Give two information-seeking tasks where you think recall is more important than precision. Make sure you justify your choices. Hint: this is the part that will be graded, not the particular choices.

(c) The trade-off between precision and recall may also be user-specific. That is, some users may care more about precision and others may care more about recall. Without explicitly asking a user, how might a search engine try to guess whether a particular user cares more about precision than recall, or vice versa? Hint: think of different ways in which users interact with a search engine and be creative!

3: Ranked Boolean Retrieval Model (30%)

We've discussed two boolean retrieval models: unranked and ranked boolean. Both boolean retrieval models return *only* the set of documents that match the query. Their difference is that the *ranked* boolean model orders the results by the number of ways the document satisfies the query. As we saw in class, for an AND boolean constraint, this is equivalent to taking the *minimum* term-frequency. For an OR boolean constraint, this is equivalent to taking the *sum* of the term-frequencies. This question is about the *ranked* boolean retrieval model.

Consider the following two boolean queries:

- (houses `OR` for `OR` sale `OR` in `OR` durham `OR` nc)
- (houses `AND` for `AND` sale `AND` in `AND` durham `AND` nc)

Suppose these are issued to a search engine that uses the ranked boolean retrieval model. Assume, for simplicity, only four documents in the collection (with document ids 1-4). Answer the following questions. The following table gives the number of times each query-term occurs in each document.

docid	houses	for	sale	in	durham	nc
1	40	0	35	0	39	43
2	10	10	9	10	8	0
3	10	12	9	11	6	1
4	1	37	1	166	1	1

(a) Compute the document scores and the ranking associated with the query (houses `OR` for `OR` sale `OR` in `OR` durham `OR` nc).

(b) How is the ranking produced probably sub-optimal and why does this happen?

(c) Compute the document scores and the ranking associated with the query (houses `AND` for `AND` sale `AND` in `AND` durham `AND` nc).

(d) How is the ranking produced probably sub-optimal and why does this happen?

4: Extending Ranked Boolean Retrieval Model (15%)

As mentioned in Question 3, the ranked boolean retrieval model scores documents using the *minimum* term frequency for `AND` and the sum of the term frequencies for `OR`.

(a) How would you extend the boolean retrieval model to handle `AND NOT` constraints (e.g., houses `AND NOT` durham)? Your proposed solution should give a higher score to documents that contain *fewer* occurrences of the term to the *right* of the `AND NOT` (e.g., durham). Please be as mathematical as possible. In other words, saying: "I would reduce the score for documents that contain the word to the right of `AND NOT`." is too vague.

(b) Using the index from Question 3, what would be the scores given to documents 1-4 by your proposed scoring method for the query "houses `AND NOT` durham"?

5: Zipf's Law (10%)

Zipf's Law expresses the proportion of term-occurrences associated with a term as a function of the term's frequency-based rank. Zipf's law can be stated as follows:

$$P_t = \frac{c}{r_t},$$

where P_t is the proportion of term occurrences associated with term t, r_t is the frequency-based rank associated with term t, and c is a constant (for English, $c = 0.1$).

What proportion of term-occurrences would be removed from the collection if we ignored all occurrences of the five most frequent terms in the collection?

6: Heap's Law (10%)

Suppose we have three collections A, B, and C. Collection A has 100 documents, collection B has 10,000 documents, and collection C has 100,000 documents. Now, suppose we add the documents in collection A into both collection B and into collection C. According to Heaps' Law, which collection is likely to see more new terms added to its vocabulary: B or C? Why?