# Information Retrieval (CS60092), Autumn 2013 – 2014

## Tutorial - I

**Q. 1>** Create the term-document incidence matrix and the postings lists from the following (meaningless) text:

Doc-1: *The Boolean retrieval model is a model for information retrieval in which we can pose any information retrieval query which is in the form of a retrieval Boolean expression of terms, that is, in which terms are combined with documents as the operators AND, OR, and NOT.*
Doc-2: *The collection model views each document as just a set of words to be a model collection.*
Doc-3: *We will refer to the group of documents over which we perform retrieval as the (document) collection.*
Doc-4: *In it, a collection system aims to provide documents from within the collection that are relevant to an arbitrary user information collection need, communicated to the system by means of a one-off, user-initiated query.*

Use the stopword list provided at http://jmlr.org/papers/volume5/lewis04a/a11-smart-stop-list/english.stop . Assign term indexes in the order in which new words are encountered. Use case-folding.

**Q. 2>** What is query optimization? For which Boolean search operator is query optimization provide the most significant gains?

**Q. 3>** What are some common cases encounted during token normalization, or the equivalence classing of terms?

**Q. 4>** What is the difference between stemming and lemmatization? What is the aim of performing these processes?

**Q. 5>** What are skip pointers? For which class of queries (with respect to Boolean operators) is a skip pointer useful? What is a commonly used heuristic for the number and locations of skip pointers in postings lists?

**Q. 6>** What are positional and non-positional indexes?

**Q. 7>** How are wild card queries handled in tolerant retrieval?

**Q. 8>** What is the edit distance between *click* and *trips*? Show all steps of the computation.

**Q. 9>** Why is the knowledge of Heaps' law and Zipf's law useful in system design?

**Q. 10>** Provide the VB gap encoding for the following postings list of a term: 345, 346, 745, 987, 10112.

**Q. 11>** Consider the four documents in Q. 1. Apply the vector-space model on this collection for the top-6 most frequent words (barring stopwords as usual). Choose arbitrarily in case of ties. Index these terms in the order in which they are encountered, i.e. the first new word gets the index 1, the second new word 2, and so on. Use the TF-IDF term-weighting scheme, with raw frequencies and $\text{IDF}_t = \log_{10}(N/\text{DF}_t)$. Now rank the documents in response to the following queries: (a) *document model*, (b) *information retrieval collection*. Use the simple overlap score measure which adds the weights of matching terms of the query in the document. How does the scoring function change when the cosine similarity is used along with the query vector?

**Q. 12>** What are some standard test collections used by the global IR community?

**Q. 13>** Consider the following top-ten result lists for three queries (Q) by three systems (S) (1 = relevant, 0 = irrelevant):

S1Q1: 0 0 0 0 0 1 1 1 1 1
S1Q2: 0 0 0 1 0 1 0 1 1 0
S1Q3: 0 0 0 0 0 0 1 1 0 1

S2Q1: 0 1 0 1 0 1 0 1 0 0
S2Q2: 1 0 1 0 1 0 1 0 0 0
S2Q3:  0 1 0 1 0 1 0 1 0 1

S3Q1: 1 1 1 1 0 0 0 0 0 0
S3Q2: 1 1 1 0 0 0 0 0 0 0
S3Q3: 1 1 1 1 0 0 0 0 0 0

The numbers of relevant documents in the collection are 6, 4 and 5 respectively for the three queries. Compute the precision, recall and F1 of each result set. Rank the systems using MAP and average R-precision. What is the nature of the P-R curve? What is usually done to smoothen it? Why is *F*-measure chosen as the harmonic mean of precision and recall? State the generalized formula for *F* in terms of *alpha* or *beta*.

**Q. 14>** What is an ROC curve? Which measure is usually used for graded judgments in place of MAP?

**Q. 15>** Suppose that a user's initial query is *cheap CDs cheap DVDs extremely cheap CDs*. The user examines two documents, *d*1 and *d*2. She judges *d*1, with the content *CDs cheap software cheap CDs* relevant and *d*2 with content *cheap thrills DVDs* nonrelevant. Assume that we are using direct term frequency (with no scaling and no document frequency). There is no need to length-normalize vectors. Using Rocchio relevance feedback, what would the revised query vector be after relevance feedback? Assume *alpha* = 1, *beta* = 0.75, *gamma* = 0.25.

**Q. 16>** Draw the DOM tree for the following XML document:

```
<db>
 <customer>
   <name>
     <firstname>John</firstname> <lastname>Doe</lastname>
   </name>
   <phone>
     <areacode>512</areacode> <number>471-9558</number>
   </phone>
   <purchases>
     <item>
       <camera>
         <type>Canon digital</type> <price>200</price>
       </camera>
     </item>
   </purchases>
 </customer>
</db>
```

**Q. 17>** What are the differences between standard vector space tf-idf weighting and the BIM probabilistic retrieval model (in the case where no document relevance information is available)? State the RSV expressions for the BIM and the BM25 models.

**Q. 18>** Suppose the document collection contains two documents only:
*d*1: Xyzzy reports a profit but revenue is down
*d*2: Quorus narrows quarter loss but revenue decreases further
Use the MLE unigram model from the documents and the collection, mixed with $\lambda = 1/2$. Suppose the query is *revenue down*. Rank the documents in response to this query.