

CS 276B Problem Set #2

Assigned: Thursday, February 24, 2005

Due: Tuesday, February 8 by 5:30 p.m.

Delivery: Submit your solutions in class or under Professor Manning's door. *Please do not submit your assignment via email.*

Late policy: Refer to the course webpage.

Honor code: Please review the collaboration and honor code policy on the course webpage. Also note: because some questions may be drawn from previous years' problem sets or exams, students are forbidden to consult solution sets from previous years unless we explicitly provide them.

#1. XML retrieval

```
<document>
  <chap>
    <sec>
      <p>
        <s>
          <w t="PRP">I</w> <w t="VB">know</w> ...
        </s>
        <s>
          <w t="NNP">Bill</w> <w t="VB">saw</w> ...
        </s>
        ...
      </p>
      <p>
        ...
      </p>
    </sec>
    <sec>
      ...
    </sec>
    ...
  </chap>
  ...
</document>
```

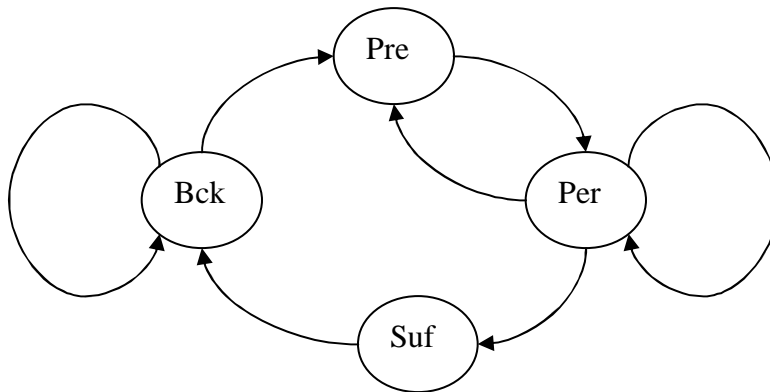
Suppose we have a document collection where the documents are richly annotated XML documents (as might be found in a humanities or legal setting), such as the one shown partially in outline above. Consider doing IR over such a collection where we could do (passage) retrieval over larger or smaller units of text (e.g., documents, sections, paragraphs, etc.). That is, we would choose some element as the indexing level, and return as matches units of that level. Because each term occurrence in the document is marked up with its part of speech, the “smallest” granularity is at the term level.

Consider the following propositions, and give a brief (1 – 2 sentence) response. For each statement, we want to know if it is **true or false**, but also want an explanation as to why. Assume for concreteness that we are issuing a two-word query to the IR engine.

- i. As the size of the indexing unit gets smaller the difference between using a Boolean term presence/absence model and a model using term frequency gets smaller and approaches zero.
- ii. As the size of the indexing unit gets smaller the difference between using and not using document frequency scaling gets smaller and approaches zero.
- iii. As the size of the indexing unit gets smaller, use of a binary “term-incidence vector” (a.k.a. a bitmap) becomes an increasingly attractive way to index the text, in terms of space and time efficiency.
- iv. As the size of the indexing unit gets larger, the *relative* overhead in storage cost for doing a positional index vs. a simple frequency index increases.

#2. Hidden Markov Model-based Information Extraction

Suppose that we are training a four state HMM for information extraction. The transition structure of the model is as shown below, with a background state, a prefix and suffix state (which do not allow self-loops) and a “Per” state for extracting Person Names.



Suppose that we train this model on the following training text, where Person Name words are tagged with /PER. Assume that each line is a separate document (i.e., that you start with a start probability each time).

President George/PER Bush/PER and Vladimir/PER Putin/PER will meet in Slovakia today
Mr Bush/PER meets Wednesday in Germany with Chancellor Gerhard/PER Schroeder/PER

Suppose that we train the HMM on this data. The Π and A matrices should just be estimated by maximum likelihood estimation. But the B matrix definitely has to be

smoothed to be able to use the HMM on new data. Assume a vocabulary of 20 words, consisting of the 19 word types found in the two documents above, and a reserved symbol UNK to which we will map all word types in test data that were unseen in training data. Estimate the B matrix by doing add one smoothing (as in CS 276A Lecture 11, slide 20).

a) Show what the Π , A, and B matrices are if you estimate the HMM on the above data as indicated.

b) Consider tagging the following document:

French president Jacques Chirac meets with Tony Blair

Run the Viterbi algorithm with the HMM trained in part a) over this document. Show (i) the δ matrix of state by time maximum probability paths and (ii) the ψ matrix of state by time backpointers, and finally (iii) the Viterbi state sequence for this document (either separately or by drawing it on the ψ matrix).

c) Consider using the trained HMM of part a) to generate text. i) According to the model what is the probability distribution of generating Person Names of different lengths (that is, what is the probability $P(\text{Len} = N | \text{generating person name})$, where Len is how many words long the person name is)? ii) How does this model distribution compare with the empirical distribution over name lengths in the training data? iii) How could one more closely model the empirical distribution of name lengths within an HMM-based model?

#3. XML

Consider the XML document on Slide 37 of Lecture 15.

(a) In the JuruXML scheme, what would the structural terms be?

(b) Consider a second document that is identical to this document, except that *Gates* is replaced by *Clinton*. What is the cosine similarity between these two documents? Does this depend on what the other documents/axes of the vector space are?

#4. INEX

In the INEX assessments, each result is assigned a pair of ratings from $\{0,1,2,3\} \times \{N,S,L,E\}$. Of the 16 possible pairs, only 9 are considered in the generalized f -measure computed on Slide 47 of Lecture 15. Explain why the remaining 7 pairs are ignored, if need be by breaking the 7 pairs into subsets for which there is a common reason to ignore them.