605.744 Information Retrieval
Paul McNamee

Problem Set (Module 1)

1. [20%] List five English stopwords.

2. [20%] What is the difference between a disjunctive clause and a conjunctive clause in a Boolean query? Give an example of each.

3. [20%] Section 1.3 in IIR describes how two postings lists of lengths *x* and *y* can be merged in a conjunctive Boolean query in O(*x*+*y*) operations (see Figure 1.6). Describe how the *Intersect* algorithm could be modified to handle queries where one term is negated. For example: "PIE AND NOT PEACH" which should return all docids that have PIE but do not also contain PEACH. Your modified algorithm should have a similar fast computational complexity even if "NOT PEACH" matches 10x more documents than "PEACH". Note: no program needs to be written or submitted to answer this question.

4. [20%] Give one specific and clear example of a word where case-folding (*i.e.*, lower-casing text) can cause an IR system to make an error that would not normally occur if case distinctions were retained. Briefly explain the error.

5. [20%] Suppose we are using a biword index as described in IIR Section 2.4. Give an example of a short plausible English document that is 1 or 2 sentences in length and that would be retrieved for the query "GREEN PARTY FAVORS", but which is actually a false positive and does not contain all three words in consecutive order.

An example true positive document might be:

"I asked June to pick up some green party favors before Todd's birthday party."