605.744: Information Retrieval
Problem Set (Module 1)

Sabbir Ahmed

September 6, 2022

1. (20%) List five English stopwords.

   Answer:

   - the
   - a
   - an
   - is
   - I

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

   Answer:

   A disjunctive clause returns the postings found in the union of all the postings while a conjunctive clause returns the intersection of all of the postings lists in the query. For example, consider the following documents:

$$\begin{aligned} Doc1 &= \{apple, oranges, banana\} \\ Doc2 &= \{denim, nylon, silk\} \\ Doc3 &= \{apple, bottom, denim\} \end{aligned}$$

   The disjunctive clause: ‘"apple OR denim"‘ will yield $result = \{Doc1, Doc2, Doc3\}$. "apple" is found in both Doc1 and Doc3, but not Doc2. However, the other term "denim" is found in Doc2 (as well as Doc3, which has already been added to the results set).

   The conjunctive clause: ‘"apple AND denim"‘ will yield $result = \{Doc3\}$. "apple" is found in both Doc1 and Doc3, but not Doc2. The other term "denim" is found in Doc2 and Doc3. Making an intersection of the 2 results sets: $\{Doc1, Doc3\} \cap \{Doc2, Doc3\}$ yields $\{Doc3\}$ as the final result set.

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.

Answer:

The following snippet is the original Intersect algorithm:

Listing 1: Intersect Algorithm from Figure 1.6

```
INTERSECT(p1, p2)
    answer ← <>
    while p1   NIL and p2   NIL do
        if docID(p1) = docID(p2) then
            ADD(answer, docID(p1))
            p1 ← next(p1)
            p2 ← next(p2)
        else
            if docID(p1) < docID(p2) then
                p1 ← next(p1)
            else
                p2 ← next(p2)
    return answer
```

The Boolean expression $A \cap \overline{B}$ is equivalent to the set difference $A \setminus B$. The intersect algorithm can be modified to perform a set difference operation of the 2 sets.

Note that this operation is not commutative, i.e. $A \cap \overline{B} \neq B \cap \overline{A} \implies A \setminus B \neq B \setminus A$.

Listing 2: Intersect Algorithm Modified to Negate the Second Term

```
DIFFERENCE(p1, p2)
    answer ← <>
    while p1   NIL do
        if docID(p1) = docID(p2) then
            p1 ← next(p1)
            p2 ← next(p2)
        else
            if docID(p1) < docID(p2) then
                ADD(answer, docID(p1))
                p1 ← next(p1)
            else
                p2 ← next(p2)
        if p2 = NIL then
            while p1   NIL do
                ADD(answer, docID(p1))
                p1 ← next(p1)
    return answer
```

The conditionals in the main loop are essentially reversed, where the docIDs get added to the result set when they are not equal. To account for cases where the negated term has fewer postings than the first term, an additional loop is added to transfer all of the postings of the first term to the result set. Since the additional loop runs over the remaining postings of the first term, the total number of iterations is preserved to $O(x + y)$.

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.

   Answer:

   One potential issue raised by unconditionally case-folding words is found in proper nouns. For example, suppose a document Doc where $\{fielding, Fielding\} \in Doc.$ ”fielding” (all lowercase) indicates a verb used in sports where a player is sent out to be active in the field, among other usages. ”Fielding” (capitalized) indicates a name of an entity, such as a person, company, location, etc. Both of those terms would be considered identical without context, and an IR system may run into additional issues when those words are stemmed to ”field”.

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.”

   Answer:

   ”Have you heard that Todd is also an avid supporter of the Green Party? No wonder he was passing out little succulents as party favors!”

   The above document would contain the biwords: ”green party” and ”party favors” and would be retrieved by the query ”GREEN PARTY FAVORS”.