WRIGHT STATE UNIVERSITY
Department of Computer Science and Engineering

CS707 Winter 2012 Prasad

Midterm (30 pts)

1 Positional Index and Querying (3 + 4 + 3)

Consider the following fragment of a positional index with the format:
`word: doc#: <posn, posn,...>; doc#: <posn,...>.`

```
Gates: 1: <1>; 2: <6>; 3: <2,15>; 4: <1>.
IBM: 4: <3>; 7: <14>.
Microsoft: 1: <2>; 2: <12, 16,21>; 3: <13>; 5: <21,25>.
```

The `/k` operator, `word1 /k word2` finds occurrences of `word1` within k words of `word2` (on either side), where `k` is a positive integer argument. Thus `k = 1` demands that `word1` be adjacent to `word2`.

- Describe the set of documents that satisfy the query: `Gates /2 Microsoft`.
- Describe the set of values for `k` for which the query: `Gates /k Microsoft` returns the set of documents {1,3} as the answer.
- Describe the set of values for `k` for which the query `Microsoft /k Microsoft` returns a non-empty set of documents as the answer.

2 True/False with Justification (2*5)

Feel free to assume Reuters dataset for specificity.

1. Stemming is more effective than stopwords elimination at reducing the size of the dictionary.
2. Stemming is more effective than stopwords elimination at reducing the size of the index.
3. Stopwords elimination reduces the maximum length of the positional postings list.
4. Indexing documents on synonyms improves precision.
5. Indexing documents on synonyms improves recall.

Precision is defined as the number of *relevant* documents retrieved by a search divided by the *total* number of documents retrieved by that search. *Recall* is defined as the number of *relevant* documents retrieved by a search divided by the total number of *existing relevant* documents (which should have been retrieved).

3 Estimating Time Units for Sorting (4)

If we need $Tlog_2T$ comparisons (where T is the number of termID-docID pairs) and two disk seeks for each comparison, how much time would it take to sort 1 million terms if we used disk for storage and an unoptimized sorting algorithm (i.e., not an external sorting algorithm)? What is the closest time unit that captures the order of magnitude of the time required: a microsecond, a second, an hour, a day, a month or a year?

Repeat this calculation for the case when the complete data is in the main memory, using in-memory sort?

Use the system parameters given below. ($log_2 10 = 3.3$)

Symbol	Statistic	Value
m	memory transfer time per byte	5 ns
s	average seek time	5 ms
p	typical ALU operation	10 ns

4 Index Compression (1 + 2 + 3)

1. What is the largest gap that can be encoded in 2 bytes using the variable-byte encoding?

2. What is the posting list that can be decoded from the variable byte-code `10001001 00000001 10000010 11111111`?

3. What would be the encoding of the same posting list using a γ-code?