Information Retrieval

Assignment 3

Due:Tuesday, 6th May.

---

1. **(**5 Points**) Are the following statements true or false?**
    a. In a Boolean retrieval system, stemming never lowers precision.
    b. In a Boolean retrieval system, stemming never lowers recall.
    c. Stemming increases the size of the vocabulary.
    d. Stemming should be invoked at indexing time but not while processing a query.

2. **(**5 Points**) Suggest what normalized form should be used for these words (including the word itself as a possibility):**
    a. ’Cos
    b. Shi’ite
    c. Los-Angeles
    d. Hawai’i
    e. O’Rourke

3. **(**5 Points**) The following pairs of words are stemmed to the same form by the Porter stemmer. Which pairs, would you argue, should not be conflated? Give your reasoning.**
    a. abandon/abandonment
    b. absorbency/absorbent
    c. marketing/markets
    d. university/universe
    e. volume/volumes

4. **(**3 Points**) List 3 problems for tokenizing documents of different languages.**

5. **(**2 Points**) What is the main different between Lemmatization and Stemming techniques?**

6. **(**2 Points**) Why are skip pointers not useful for queries of the form x OR y?.**

7. (5 Points) **We have a two-word query. For one term the postings list consists of the following 16 entries:<[4,6,10,12,14,16,18,20,22,32,47,81,120,122,157,180] and for the other it is the one entry postings list: [47].**
   Work out how many comparisons would be done to intersect the two postings lists with the following two strategies. Briefly justify your answers:
   a. Using standard postings lists
   b. Using postings lists stored with skip pointers, with a skip length of $\sqrt{p}$.

8. (3 Points) **Assume a biword index. Give an example of a document which will be returned for a query of Suez Canal University but is actually a false positive which should not be returned.**

9. (5 Points) **Shown below is a portion of a positional index in the format:**
   term: doc1: <position1, position2, . . . >; doc2: <position1, position2, . . . >; etc.

   angels: 2: <36,174,252,651>; 4: <12,22,102,432>; 7: <17>;
   fools: 2: <1,17,74,222>; 4: <8,78,108,458>; 7: <3,13,23,193>;
   fear: 2: <87,704,722,901>; 4: <13,43,113,433>; 7: <18,328,528>;
   in: 2: <3,37,76,444,851>; 4: <10,20,110,470,500>; 7: <5,15,25,195>;
   rush: 2: <2,66,194,321,702>; 4: <9,69,149,429,569>; 7: <4,14,404>;
   to: 2: <47,86,234,999>; 4: <14,24,774,944>; 7: <199,319,599,709>;
   tread: 2: <57,94,333>; 4: <15,35,155>; 7: <20,320>;
   where: 2: <67,124,393,1001>; 4: <11,41,101,421,431>; 7: <16,36,736>;

   Which document(s) if any meet each of the following queries, where each expression within quotes is a phrase query?

   a. “fools rush in”
   b. “fools rush in” AND “angels fear to tread”