
Midterm (30 pts)

1 Vector Space Ranking (6 + 3 + 3 pts)

This exercise is based on the course assignment. Consider the following document collection $D = \{D1, D2, D3\}$ (given as one document per line):

```
D1 => Silly Sally Sleepy Sally
D2 => Seven Silly Sheep
D3 => Silly Sheep Should Sleep Silly
```

Assume that the stopword list contains the word **Should**, and words are stemmed (that is, converted to their root).

- Show the dictionary and the postings list including all the relevant statistics computed, such as *raw* tf-idf values shown explicitly as ‘(tf,idf)’ with each document in the postings list), for implementing (uncompressed) inverted index structure for Vector Space Ranked Retrieval in an easy-to-read format. Assume that raw term frequency factor is the *count* of the number of term occurrences in a document (rather than the normalized, log-dampened value) and the inverse document frequency factor is the *reciprocal of the fraction* of documents that contain the term (rather than its logarithm).
- What are the relevance scores and the ranking of the documents for the query: **Silly**?
- Does the ranking change if we define term frequency factor as the *normalized fraction* of the term occurrences in a document (rather than the *raw count*).

2 True/False with Justification (10)

Soundex, a phonetic algorithm patented in 1918, maps every word to a soundex code, which is a letter followed by three digits. The letter is always the first letter of the word, the digits are assigned according to some rules that groups similar sounding letters together. For example, M and N are both coded as 5 because they sound similar; B, F, P, and V code as 1; vowels are completely ignored (unless theyre the first letter).

Here are some sample soundex values:

- S-530 is the code for Smith and Smythe.
- A-450 encodes Allan, Allen, Alan, and even Allynn.

Assuming that we use soundex encoding to map tokens to terms, state whether each of the following claims is true or false, justifying your answer.

1. Soundex encoding reduces the size of the dictionary.
2. Soundex encoding reduces the maximum length of the positional postings list.
3. Soundex encoding improves precision.
4. Soundex encoding 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 Index Compression (2 + 2 + 4)

1. What is the largest gap that can be encoded in 3 bytes using the variable-byte encoding? (*Hint*: You may leave the answer in 2^i -form.)
2. What is the posting list that can be decoded from the variable byte-code 10001101 00000001 10000010 00000001 11111111?
3. What would be the encoding of the same posting list using a γ -code?