
Final Exam (45 pts)

1 Inverted Index (5 pts)

Consider the following document collection $C = \{C1, C2, C3\}$ (given as one document per line):

```
clear blue sky
the blue car
sky is nice
```

Assume that the stopword list contains words `the` and `is`, and words are not stemmed.

Explain briefly all the relevant data structures constructed (including their characteristics such as sortedness) for implementing (uncompressed) inverted index structure for Boolean Retrieval, and their instantiation for the above example.

2 Ranking using Vector Space Model (3 + 8 + 4 pts)

Consider the following (filtered) document collection $D = \{D1, D2, D3\}$ (given as one document per line):

```
new york times
new new york post
the times
```

Show all the relevant data structures constructed and all the relevant statistics (such as tf-idf values shown as '(tf,idf)' explicitly) computed for implementing (uncompressed) inverted index structure for Vector Space Ranked Retrieval for the given example. Assume that term frequency factor is just the count of the number of term occurrences in a document (rather than the normalized value) and the (inverse) document frequency is the (reciprocal of) fraction of documents that contains a term.

What are the “relative” relevance scores and the ranking of the documents for the query : `new`?

3 Short Answer Questions (3 + 6 + 4 + 6 + 6 pts)

1. Discuss the pros and the cons of a linear classifier.
2. What are the variable byte codes encoding AND the γ -codes encoding of the following postings list: 1, 20, 276?
3. Compute the eigenvector associated with the largest eigenvalue for the following matrix.

$$A = \begin{bmatrix} 1 & 1 \\ 0 & -2 \end{bmatrix}$$

4. State and explain clearly all the independence assumptions made to ensure tractability of Naive Bayes Multinomial model.
5. What is the fundamental difference between (i) Rocchio classifier, (ii) kNN classifier, and (iii) SVM classifier?