
Final Exam (40 pts)

1 Bridging Similarity and Distance (6 pts)

If the query q and documents d_i s are all normalized to unit vectors, then what is the relationship between the result set rank ordering produced by Euclidean distance to that produced by cosine similarity? Justify your answer.

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

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

```
new york times
new new york post
los angeles times
```

Assume that the stopword list contains words **the** and **is**, and words are not stemmed. For the given example, show all the relevant data structures constructed and all the relevant statistics computed (such as 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. 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 the fraction* of documents that contain the term.

What are the "relative" relevance scores and the ranking of the documents for the query : **new**?

3 Short Answer Questions (5 + 3 + 3 + 8 pts)

1. What are the variable byte codes encoding for the following postings list:
2, 30, 286?
2. Is it true that clustering can be used to improve result set navigation in a search engine? Justify your answer.
3. Explain clearly why Naive Bayes classifiers are so robust in spite of making several simplifying independence assumptions to ensure tractability.
4. Discuss bias-variance tradeoffs in (i) Rocchio classifier, (ii) kNN classifier, and (iii) SVM classifier in terms of the nature of class separation boundaries and the data points in the training set that can effect them.