Wright State University
Department of Computer Science and Engineering

CS707 Winter 2012 Prasad

Final Exam (40 pts)

1 Evaluating Search Engine (3 + 3 + 1 + 1 pts)

Given a search engine benchmark (such as Medline dataset containing binary relevance judgements), explain clearly why you would prefer to use *Mean Average Precision* metric over *Precision-At-Fixed-Recall* metric to evaluate a ranked retrieval system based on Vector Space Model?

The table below shows the output of an IR system on two queries. Only top 5 ranks are shown. Crosses correspond to documents which have been judged relevant by a human judge; circles correspond to irrelevant documents. There are no relevant documents in lower ranks (> 5). Compute the MAP.

Rank	Q1	Q2
1	O	X
2	X	O
3	X	O
4	X	O
5	O	X

Is MAP appropriate to evaluate a search engine that uses LSI? Justify.

Is MAP appropriate to evaluate a search engine that uses graded/multi-level relevance measure in place of boolean/binary relevance measure? Justify.

2 Vector Space Ranking (6 + 2 + 2 + 2 pts)

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

```
Asterix: Asterix the Gaul
Asterix and the Golden Sickle
Asterix and Cleopatra
```

Assume that the stopword list contains the word `the`, and words are not stemmed. For the given example, show the dictionary and the postings list including all the relevant statistics computed (such as tf-idf values shown explcitly 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 term frequency factor is the *count* of the number of term occurrences in a document (rather than the normalized 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 "relative" ranking of the documents for the query `Asterix`?

Does the ranking change if we were to define term frequency factor as the normalized *fraction* of the term occurrences in a document (rather than the raw *count*).

3 Clustering and Classification (4 + 4 + 6 pts)

1. Consider a dendrogram created by an agglomerative hierarchical clustering algorithm. Will that yield the same results as a k means clustering algorithm if the algorithm is terminated when only k clusters remain (assuming the same comparison function is used)? That is, will the resulting k clusters be the same? Why or why not?

2. Explain clearly why Naive Bayes classifiers are so robust in spite of making several simplifying independence assumptions to ensure tractability.

3. 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.

4 Map Reduce Paradigm (6)

Given a directed graph as an adjacency list (list of nodes which are targets of (out-)links): `Ids -> list(Ids)`, determine the list of pairs `(sn, cnt)`, such that `cnt` is the number of (in-)links incident on `sn`, using MapReduce paradigm. E.g., for the following graph, the expected pairs are `(s1,1), (s2, 1), (s3,2), (s4,1).`

```
s1 :  s2, s3.
s2 :  s1, s3, s4.
```

Explain the *map* task and the *reduce* task, and define these functions using *set* or *list* notation.