CSE 586 Final Exam

Olcay Taner YILDIZ

I. Question (16 points)

Let say there are N documents in a collection. Given that the terms x and y occur in n_x and n_y distinct documents respectively, how many results will be returned from these queries (minimum and maximum)?

a.) x and y
b.) x or y
c.) x and not y
d.) x or not y

II. Question (14 points)

Let say you have the following terms in the dictionary: adana, istanbul, ankara, izmir, erzurum, tekirdag, hakkari. In order to get

a.) a balanced
b.) a maximumly unbalanced

binary tree, in what order would you insert those terms to the dictionary?

III. Question (20 points)

Consider the following documents:

Doc1: breakthrough drug for schizophrenia
Doc2: new schizophrenia drug
Doc3: new approach for treatment of schizophrenia
Doc4: new hopes for schizophrenia patients

a.) Construct the vector space model (with tf-idf weighting) for the documents. ($\log_2(3)$ = 1.7).
b.) Calculate the similarity between documents 1 and 4.
c.) Construct a query for which all documents score the same.

IV. Question (15 points)

Consider the following scenarios for unranked evaluation of document retrieval:

a.) There are 10 relevant documents for the given query in the collection, system A returns 5 relevant and 5 irrelevant documents.
b.) There are 10 relevant documents for the given query in the collection, system B returns 10 relevant and 90 irrelevant documents.
c.) There are 10 relevant documents for the given query in the collection, system C returns 4 relevant documents.

Order the systems A, B, and C in terms of precision, recall and F-measure.

V. Question (20 points)

docID	words in document	in Turkey?
1	Turkish Ankara Turkish	yes
2	Turkish Turkish Adana	yes
3	Athens Greece	no
4	Athens Greece Turkish	no

Based on the data given above,

a.) Estimate a multinomial Naive Bayes classifier
b.) Apply classifier to the document Turkish Turkish Turkish Athens Greece.
c.) Estimate a Bernoulli Naive Bayes classifier
d.) Apply classifier to the document Turkish Turkish Turkish Athens Greece.

VI. Question (15 points)

Let say we have a collection of M documents with N terms. What will be the time complexity of

a.) Learning the mean vectors in the Rocchio algorithm with K classes?
b.) Testing a new document in the 1-NN algorithm?
c.) Finding which cluster the document belongs to if we have used K-means clustering?