CSE 586 Final Exam

Olcay Taner YILDIZ

I. Question (15 points)

Write the algorithm that handles the boolean query x Δ y (symmetric difference) using inverted index.

II. 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

III. 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?

IV. Question (12 points)

Suppose that the task is to cluster the following eight points (with (x, y) representing location) into three clusters. A_1(2, 10), A_2(2, 5), A_3(8,4), A_4(5, 8), A_5(7,5), A_6(6,4), A_7(1,2), A_8(4,9). The distance function is Euclidian distance. Suppose initially we assign A_1, A_4, and A_7 as the center of each cluster, respectively. Run the k-means clustering one iteration on these data.

V. Question (14 points)

docID	words in document	in Turkey?
1	Turkish Ankara Turkish	yes
2	Turkish Turkish Adana	yes
3	Turkish Turkish Adana Istanbul	yes
4	Athens Greece	no
5	Athens Greece Turkish	no
6	Athens Greece Selanik	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`

VI. Question (14 points)

Based on the data given above

a.) Implement Rocchio classifier, that is find the mean vector of each class.
b.) Apply classifier to the document
`Turkish Turkish Turkish Athens Greece`

VII. Question (14 points)

Based on the data given above, estimate the class of the document

`Turkish Turkish Turkish Athens Greece`

for

a.) k = 1
b.) k = 3