National University of Computer & Emerging Sciences
FAST-Karachi Campus
Information Retrieval (CS317)
Class Activity # 1

Dated: April 20, 2020 Marks: 30
Time: 45 min. + Chapter Reading 8 and 9
Std-ID: _____

Problem No. 1

Consider a very specific collection of 10 documents D={d_1,d_2,..d_{10}}, the human judges ranked these documents for two queries Q= {q_1, q_2} as below q_1 -> { 1,0,0,2,0,2,2,1,1,0} and q_2-> { 0,2,2,1,1,0,0,2,0,1}on a scale of (0-2, 2 most relevant, 0 non-relevant, and 1 means neutral). The information systems retrieved the following results for the two quires: q_1 -> {d_4, d_5, d_6, d_7, d_1, d_8, d_9} and q_2 -> {d_2, d_3, d_8, d_4, d_5, d_9, d_1}. Assume system does some kind of filtering and does not return all documents. Answer the following questions.

a. Compute the Cumulative Gain (CG) for the two queries.
b. Compute the Discount Cumulative Gain (DCG) for the two queries.
c. Compute the Normalized Discount Cumulative Gain (nDCG) for the two queries.

Problem No. 2

Consider the set of 8 documents on which 2 human judges performed relevance judgement for a given query q, where a 0 means non-relevant and 1 means relevant to query, given as below:

Doc-ID	1	2	3	4	5	6	7	8
Judge1	0	0	1	1	1	1	0	0
Judge2	0	1	1	0	1	0	1	0

An information system returns the following set of documents against the same query q -> {3,4,5,2,1}. Answer the following questions:

a. Calculate the kappa measure between the two judges.
b. Calculate precision, recall, and F1 of your system if a document is considered relevant only if the two judges agree.
c. Calculate precision, recall, and F1 of your system if a document is considered relevant if either judge thinks it is relevant.

Problem No. 3

Consider a set of document vectors given below:

$d_1 = <0.2, 0.4, 0.1, 0.1, 0.3, 0.4, 0.1>$

$d_2 = <0.01, 0.3, 0.2, 0.1, 0.1, 0.2, 0.2>$

$d_3 = <0.11, 0.21, 0.11, 0.11, 0.12, 0.12, 0.01>$

$d_4 = <0.4, 0.01, 0.01, 0.01, 0.15, 0.21, 0.11>$

For a given query vector $q = <0.01,0.22,0.11,0.01,0.01,0.22,0.1>$ if a user marks d_1, d_2 and d_4 as relevant.

a. Compute the modified query vector using Rocchio's algorithm. Assume α, β, and γ are 0.1,0.2 and 0.4 respectively.
b. What will be the modified query if we only use weighting for relevant documents? What value of both α and β you should use?
c. Under what conditions would the modified query q_m be the same as the original query q_0? In all other cases, is q_m closer than q_0 to the centroid of the relevant documents?