Recuperação de Informação
Lista de Exercícios

Instituto Tecnológico Educacional da Amazônia (ITEAM) 09.03.2020

Matrícula: _____

Nome: _____

Questão 1

Explique as diferenças entre os seguintes conceitos:

(a) Ranking, similaridade e relevância

(b) Precisão e revocação

(c) Recuperação de informação e recuperação de dados

Questão 2

Por que não se usa o comando *grep* (que faz busca por substrings em documentos) para fazer recuperação de informação?

Questão 3

Considerando o modelo vetorial, o que os vetores no espaço vetorial representam? Como os pesos dos termos (por exemplo, calculados através do tf-idf) são usados para posicionar objetos nesse espaço? Por que as técnicas de pesagem de termo (*term weighting*) são importantes para uma recuperação de documentos efetiva?

Questão 4

O que é suposição da ortogonalidade (*orthogonality assumption*) empregado pelos modelos clássicos? Por que essa suposição pode ser falsa, e porque ela pode levar a resultados ruins de precisão? Descreva de forma sucinta um dos modelo de recuperação de informação que procurou melhorar a qualidade do ranking partindo da ideia de que essa suposição é falsa.

Questão 5

Desenhe o vocabulário e as listas invertidas resultantes da indexação dos 3 documentos da Tabela 1. Informe o IDF (Inverse Document Frequency) de cada termo do vocabulário, e o TF (Term Frequency) de cada elemento da lista invertida.

	Doc1	Doc2	Doc3
amazônico	2	12	0
boi	8	4	12
festival	0	8	16
maior	6	0	10

Tabela 1: TF dos termos da coleção formada pelos documentos Doc1, Doc2 e Doc3

Questão 6

Explique com suas palavras o que é um modelo de linguagem (*language model*), dando um exemplo de um modelo criado para ordenação de respostas em sistemas de busca textual. Explique como os scores dos documentos são computados no modelo fornecido.

Questão 7

A seguinte lista representa os documentos relevantes (R) e não-relevantes (N) retornados por um algoritmo de ranking para uma dada consulta. O ranking foi ordenado da esquerda para a direita. A consulta possui 6 relevantes, todos presentes no ranking retornado.

$$RNRNN \quad NNNRN \quad RNNNR \quad NNNNR \tag{1}$$

Pergunta-se:

(a) Qual é a precisão do ranking para o ponto de revocação igual a 50%?

(b) Qual é o valor do MAP para este ranking?

(c) Qual é o valor do P@5 e P@10 para este ranking?

Questão 8

A figura 1 mostra o ranking de dois sistemas de recuperação de informação sobre duas consultas. Apenas os 15 primeiros documentos do ranking são mostrados, e os sistemas não retornaram documentos relevantes da posição 16 em diante. Os documentos assinalados com X foram julgados relevantes por um especialista, enquanto que os demais documentos foram julgados não relevantes.

(a) Explique as seguintes métricas de avalição, e mostre os resultados dessas métricas para os resultados obtidos para a consulta Q1 em ambos os sistemas: (i) Precisão no ponto 10 do ranking; e (ii) Revocação no ponto 10 do ranking.

(b) Escreva a fómula da métrica Mean Average Precision (MAP), e mostre o uso desta fórmula calculando o MAP do primeiro sistema.

System 1		
Rank	Q1	Q2
1	—	X
2	X	—
3	X	—
4	X	—
5	—	—
6	—	—
7	—	—
8	X	—
9	X	—
10	X	—
11	X	—
12	—	—
13	—	X
14	—	X
15	X	—

System 2		
Rank	Q1	Q2
1	X	X
2	X	—
3	X	—
4	—	X
5	X	X
6	X	—
7	—	—
8	—	—
9	—	—
10	—	—
11	X	—
12	X	—
13	—	—
14	—	—
15	X	—

Figure 1: Ranking de dois sistemas de RI

(c) Para a consulta Q2, desenhe a curva de precisão e revocação para ambos os sistemas. Em sua opinião, qual é o melhor modelo de acordo com essa métrica?

(d) O MRR (*Mean Reciprocal Rank*) seria uma boa métrica para avaliar os sistemas de recuperação 1 e 2?

Questão 9

O que é e para que serve o método de pooling? Em que situações ele deve ser usado?

Questão 10

Em um experimento da área de RI, um especialista foi convidado a avaliar o nível de relevância da resposta de uma máquina de busca para uma dada consulta. Para cada documento foi atribuída uma nota de 0 a 3, onde 0 significa irrelevante, 1 pouco relevante, 2 mediamente irrelevante, e 3 muito relevante. A máquina de busca retornou a seguinte ranking de documentos: D_{57} , D_{12} , D_{33} , D_{78} , D_{9} , D_{25} , que foram avaliados com as notas 3, 2, 3, 0, 1, 2, respectivamente. Desenhe as curvas de NDCG para esse experimento.