Lista de exercícios 2

1) Defina *avaliação* de recuperação no contexto da recuperação da informação.
2) Explique a importância do paradigma Cranfield para a avaliação da recuperação de informação.
3) O que é uma coleção de referência? Como ela pode ser utilizada para avaliar a recuperação de um sistema de RI?
4) Suponha que os sistemas de RI - Zoodle e Ping foram propostos. Explique cada um dos passos exigidos para realizar uma avaliação da recuperação de informação usando as métricas de precisão e revocação.
5) Considere uma coleção de referência e um conjunto de consultas para teste. Suponha que os conjuntos R1, R2 e R3 de documentos relevantes para as consultas q1, q2 e q3, respectivamente, tenham sido determinados por um grupo de especialistas. Os conjuntos R1, R2 e R3 são dados da seguinte forma:

   R1 = {d3, d7, d12, d13, d26, d68}
   R2 = {d1, d2, d9, d24, d51, d52, d70, d82}
   R3 = {d2, d3, d6, d16, d20}

   Considere que um novo algoritmo de recuperação chamado XYZ foi recém projetado. Suponha que esse algoritmo retorne, para as consultas q1, q2 e q3, os seguintes rankings de documentos (primeiras quinze posições):

   Consulta q1 (algoritmo XYZ) = {d1, d9, d26, d15, d2, d10, d74, d68, d32, d3, d53, d39, d56, d11, d4}.
   Consulta q2 (algoritmo XYZ) = {d3, d7, d8, d9, d19, d16, d37, d24, d20, d80, d67, d50, d46, d51, d29}.
   Consulta q3 (algoritmo XYZ) = {d2, d30, d25, d3, d9, d7d6, d39, d75, d19, d26 d16, d20, d51, d1}.

   a)Calcule os níveis de precisão e revocação para cada uma das consultas.

   b) Construa o gráfico de precisão versus revocação para cada uma das consultas.

   c) Encontre a precisão média do algoritmo XYZ e faça o gráfico dos valores médio de precisão versus revocação.
6) Liste, no mínimo, duas vantagens e desvantagens das medidas de precisão e revocação.
7) Com base na coleção de documentos apresentada na exercício 4) encontre:
   a) Os valores de precisão P@n nas posições 5, 10 e 15 para cada uma das três consultas.
   b)Para qual consulta a impressão dos usuários será mais positiva? Justifique.
8) Com base na coleção de documentos apresentada na exercício 4) calcule a MAP (Média das Precisões Médias) para as consultas q1, q2 e q3.

9) Um algoritmo de RI chamado A1 produz o seguinte ranking R_{A1} = {*d1, d9, d26, d15, d2, d10, d74, d68, d32, d3*}. Outro algoritmo A2 produz o seguinte ranking R_{A2} = {*d1, d68, d26, d32, d2,d10 , d74, d9,d15 , d3*}. Compare a ordenação relativa das respostas produzidas pelos dois algoritmos usando o coeficiente de correlação de Spearman.

10) Pesquise e explique o funcionamento das métricas conhecidas como *taxa de cobertura* e *taxa de novidade*.

11) Explique as diferenças entre as informações explícitas e implícitas na realimentação de relevância. Cite exemplos.

12) Explique o funcionamento do método de Rocchio para realimentação de relevância do modelo vetorial.

13) Considere a coleção abaixo formada por sete documentos. Suponha que a coleção seja indexada com base no modelo vetorial clássico.

d1 = The apple is the pomaceous fruit of the apple tree, Malusdomestica.
d2 = Delicious and crunchy apple fruit is one of the popular fruits.
d3 = An apple a day keeps the doctor away.
d4 = Apple trees take four to five years to produce their first fruit.
d5 = Apple Inc. is an American multinational corporation headquartered in Cupertino, California.
d6 = Apple designs the Mac, along with OS X, iLife, and iWork.
d7 = iPhone is a line of smartphones designed and marketed by Apple.
d8 = Android smartphones are better than iPhone?

a) Seja q1 = "apple" a consulta inicial do usuário. Se cada termo de indexação possui quatro ou mais letras, escreva o vetor consulta q1.
b) Qual foi o ranking gerado pela consulta q1?
c) Suponha que o usuário não ficou satisfeito com o resultado e que o Método de *Rocchio* para realimentação de relevância tenha sido implementado. Nesse caso, o usuário selecionou os documentos d5, d6 e d7 como relevantes. Encontre o vetor de documentos relevantes (segundo termo da fórmula de *Rocchio*) e o vetor de documentos não relevantes (terceiro termo da fórmula de *Rocchio*).
d) Suponha que *alfa* = *beta* = *gama* = 1. Qual é o vetor modificado da consulta? Qual seria o novo ranking gerado por tal vetor consulta?
e) Em termos de realimentação de relevância, o que significa fazer *beta* = 0.5 e *gama* = 0.25 ou seja *beta*>*gama*?

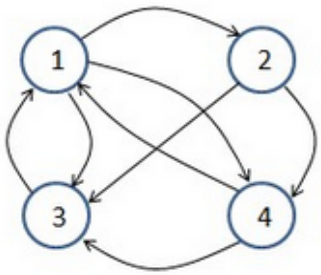
14) O que são tesauros? Pesquise e apresente exemplos de tesauros. Como funciona a realimentação ímplícita com tesauros?

15) Os cliques podem ser usados como indicador **direto** de relevância? Qual abordagem pode ser empregada para adotar dados de cliques no contexto da realimentação de relevância explícita? Dê exemplos.

16) O funcionamento básico de recuperação de informação na Web consiste em dois passos em dois passos bem definidos. Explique cada um deles.

17) O PageRank depende de dados da consulta para ser inicializado? Justifique.

18) Encontre o valor do PageRank para o grafo a seguir. Use 3 iterações completas para chegar no resultado.



site	sites apontando p/ ele
1	3, 4
2	1
3	1, 2, 4
4	1, 2

19) Explique a importância do fator de amortização no cálculo do PageRank.

20) Explique as diferenças entre o ranqueamento de dados na Web e em coleções de dados convencionais.