Lista de exercícios 2

1) Considere a seguinte coleção composta por cinco documentos. Suponha que o vocabulário dessa coleção seja formado por cada uma das palavras que aparecem nos documentos abaixo:

   D1 = {homem estar tempo coisa dizer ir ter}

   D2 = {senhora estar dia moço moço senhora}

   D3 = {senhora vez senhora senhora tempo dizer filho}

   D4 = {casa ir ir dizer ter olho}

   D5 = {olho dia vez dia homem moço tempo}

   Calcule o grau de similaridade (modelo probabilístico) entre os documentos e as seguintes consultas: a) "homem moço" b) "dizer ir tempo" c) "dia senhora casa". **Dica**: considere R = r_i = 0 e use a equação que evita o comportamento anômalo n_i> N/2.

2) Compare os resultados dos exercícios 1) e do exercício 11) da Lista 1. O ranking vetorial é igual ao ranking probabilístico? Explique as diferenças.

3) Faça uma tabela ilustrando as principais vantagens e desvantagens dos modelos clássicos de RI (booleano, vetorial e probabilístico).

4) Qual é a ideia fundamental do modelo probabilístico? Explique as diferenças entre o cálculo de similaridade do modelo vetorial e do modelo probabilístico.

5) Considere o seguinte documento:
   Peer-to-peer (P2P) computing is the sharing of computer resources and services by direct collaboration between client systems. These resources and services often include the exchange of information (Napster, Freenet, etc.), processing cycles (distributed.net, SETI@home, etc.), and disk storage for files (OceanStore, Farsite, etc.). Peer-to-peer computing takes advantage of existing desktop computing power and networking connectivity, allowing of-the-shelf clients to leverage their collective power beyond the sum of their parts. Current research on P2P has evolved from very different research areas. Among others, P2P has attracted the attention of researchers working on classical distributed computing, mobile agents, parallel computing, or communications. Very interestingly, the P2P paradigm is different from those studied in all these areas. For instance, while in some sense peer-to-peer computing is very similar to classical distributed computing (as opposed to the client-server paradigm), some new characteristics emerge. These include the clear and present danger of malicious peers, high churn rate (peers joining and leaving the system), among others.

   Este texto é parte de uma coleção de um milhão de documentos indexados. Assuma que todos os documentos e consultas passam por um pré-processamento, e que somente os termos presentes na tabela abaixo são incluídos no índice. Adicionalmente, o índice armazena o número de documentos no qual cada termo aparece.

Term	documents
comput	300901
network	200019
system	110990
client	80921
agent	42003
traffic	40105
p2p	20909
peer	10979

a) Qual equação do modelo probabilístico deve ser usada nesse caso? A equação clássica conhecida como Robertson-Spark Jones ou a versão usada para evitar a inserção de termos negativos no cálculo? Justifique a resposta.

b) Calcule o grau de similaridade, usando o modelo probabilístico, do documento acima com a consulta q = “p2p computer systems”. Assuma nenhum conhecimento sobre os documentos relevantes.

c) Considere a consulta q = “p2p computer systems”. Suponha que em uma iteração inicial do algoritmo foi estimado que o número de relevantes para a consulta q é R = 1500. De posse dessas informações, calcule o grau de similaridade entre a consulta q e o documento acima usando o modelo probabilístico usando R=1500 e os valores de r_i de acordo com a tabela abaixo. Compare o resultado encontrado aqui com o grau de similaridade obtido na letra b).

Term	Documents
comput	1250
network	145
system	955
client	55
agent	67
traffic	18
p2p	1000
peer	542

6) Qual é a importância de um índice bem construído para os sistemas de recuperação de informação? Dê exemplos de estruturas de dados que podem ser usadas para a implementação de índices.