Universidade Federal de Uberlândia - UFU
Organização e Recuperação da Informação
Prof. Dr. Rodrigo Sanches Miani

Lista de exercícios 1

1) Defina Recuperação de Informação sob o contexto da Ciência da Computação. Qual era a principal aplicação dos primeiros sistemas de RI? Porque a área de RI está atualmente em destaque?
2) Qual é o objetivo principal de um sistema de RI?
3) Explique a diferença entre recuperação de dados e recuperação de informação.
4) Cite e explique cada um dos seis módulos presentes em um sistema de RI.
5) Um modelo de RI pode ser definido como uma quádrupla (D, Q, F e $R(q_i,d_j)$). Explique a importância de cada um desses elementos para a recuperação da informação. Qual desses quatro elementos pode ser dito como o mais importante para os modelos de RI?
6) No que consiste a representação conhecida como *bag of words*? Dê um exemplo do funcionamento dessa representação.
7) 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}

   a) Encontre o vocabulário dessa coleção. O vocabulário é formado por quantos termos de indexação?

   b) Construa a matriz de termos e documentos para essa coleção.

   c) Represente cada um dos documentos usando os componentes conjuntivos de termo.

   d) Usando o modelo booleano, verifique o número de documentos retornados pelas seguintes consultas: C1= {senhora AND moço}, C2 = {homem OR estar OR dizer}, C3 = {ir AND (dizer OR NOT olho)} e C4 = {dizer AND ir AND (homem OR moço)}.
8) Porque a ponderação de termos pode ser usada para melhorar a qualidade de recuperação do modelo booleano?
9) Explique as diferenças entre as ponderações TF, IDF e TF-IDF? Qual é o tipo de ponderação mais utilizada nos sistemas de RI?

10) Com base na coleção de documentos do exercício 7), faça a tabela de frequências $f_{i,j}$ e encontre os valores de TF, IDF e TF-IDF. Use as fórmulas logarítimicas apresentadas em sala de aula. Compare os resultados e esclareça as diferenças entre cada esquema de ponderação (TF, IDF e TF-IDF).

11) Use a coleção de documentos do exercício 7) para calcular o grau de similaridade (modelo vetorial) entre os documentos e as seguintes consultas: a) "homem moço" b) "dizer ir tempo" c) "dia senhora casa".

12) 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) Escreva a representação do documento acima no modelo booleano.
b) Escreva a representação do documento acima no modelo vetorial clássico.
c) Calcule o grau de similaridade, usando o modelo vetorial clássico, do documento com a consulta "p2p computer systems".

13) Seja d = (d1,d2) um vetor que representa um determinado documento e q = (q1,q2) um vetor que representa uma consulta. Use a lei dos cossenos para deduzir a fórmula do cálculo do grau de similaridade para vetores em um espaço com 2 dimensões.

14) A função seno poderia ser usada para expressar o grau de similaridade entre os vetores consulta e documento? Se sim, explique as diferenças com relação a função cosseno.