

ALMANAQUE PARA POPULARIZAÇÃO DE
CIÊNCIA DA COMPUTAÇÃO

SÉRIE 2
Inteligência Artificial

Volume 5

RECUPERAÇÃO DE
INFORMAÇÃO

Rafael Meneses Santos
Maria Augusta Silveira Netto Nunes
Sean Wolfgang Matsui Siqueira
Yargo Santana Vasconcelos



UNIVERSIDADE FEDERAL DE SERGIPE-UFS

REITOR

Prof. Dr. Angelo Roberto Antonioli

PRO-REITORA

Prof. Dra. Iara Campelo

RESPONSÁVEL PELA PRIMEIRA EDIÇÃO

Yargo Santana Vasconcelos

REVISOR TÉCNICO

Leonardo Nogueira Matos

REVISÃO GERAL

Maria Augusta Silveira Netto Nunes

RESPONSÁVEL PELA SEGUNDA EDIÇÃO

Viviane dos Santos Freire

Os personagens e as situações dessa obra são reais apenas no universo da ficção, não se referem a pessoas e fatos concretos, e não emitem opinião sobre eles.

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL
DA UNIVERSIDADE FEDERAL DE SERGIPE

R311r

Recuperação da informação [recurso eletrônico] / Rafael Menezes Santos ... [et al.]. - 2.ed. - Porto Alegre: SBC, 2017. 16p. : il. - (Almanaque para popularização de ciência da computação. Série 2, Inteligência artificial ; v. 5)

ISBN 978-85-7669-411-3

1. Sistemas de recuperação da informação. 2. Recuperação da Informação. 3. Ferramentas de busca na web. I. Santos, Rafael Menezes. II. Universidade Federal de Sergipe. III. Série.

CDU 004.775(059)



Cidade Universitária José Aloísio de Campos

CEP-490100-000- São Cristóvão- SE

ALMANAQUE PARA POPULARIZAÇÃO DE
CIÊNCIA DA COMPUTAÇÃO
SÉRIE 2:INTELIGÊNCIA ARTIFICIAL

VOLUME: 5
**RECUPERAÇÃO DE
INFORMAÇÃO**

Sociedade Brasileira de Computação-SBC
Porto Alegre-RS

AUTORES:

Rafael Meneses Santos
Maria Augusta Silveira Netto Nunes
Sean Wolfgang Matsui Siqueira
Yargo Santana Vasconcelos

Realização
Universidade Federal de Sergipe

São Cristóvão-2017

APRESENTAÇÃO

Essa cartilha foi desenvolvida pelo projeto de Bolsa de Produtividade CNPq–DTII nº306576/2016-3, coordenado pela prof^a. Maria Augusta S. N. Nunes em desenvolvimento no Departamento de Computação (DCOMP)/Programa de Pós-graduação em Ciência da Computação (PROCC) – UFS. É também vinculado à projetos de extensão, Iniciação Científica e Tecnológica para popularização de Ciência da Computação em Sergipe apoiado pela PROEX, COPES e CINTTEC/UFS. O público alvo das cartilhas são jovens pré-vestibulandos e graduandos em anos iniciais. O objetivo é fomentar ao público sergipano e nacional o interesse pela área de de Ciência da Computação.

As cartilhas da série de Inteligência Artificial descrevem sobre a área da Ciência da Computação que busca simular a inteligência humana através de mecanismos e software. Esta cartilha busca introduzir ao leitor os conceitos na área de Recuperação de Informação. A Recuperação de Informação trata dos problemas relacionados à representação, armazenamento, organização e acesso à informação, em geral considerando-se grandes coleções de documentos. É o principal conceito por trás das ferramentas de busca na Internet e está diretamente relacionado com as áreas: Inteligência Artificial e Processamento de Linguagem Natural.

(Maria Augusta Silveira Netto Nunes)

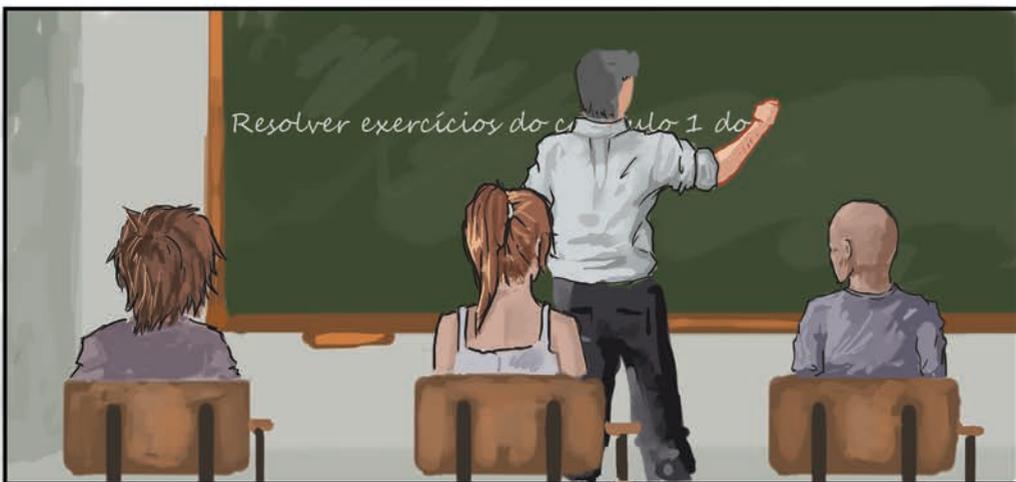
PRIMEIRA AULA DE PROBABILIDADE E ESTATÍSTICA DO CURSO DE CIÊNCIA DA COMPUTAÇÃO.



RICARDO E VANESSA, ALUNOS DO CURSO DE CIÊNCIA DA COMPUTAÇÃO.



Resolver exercícios do capítulo 1 do

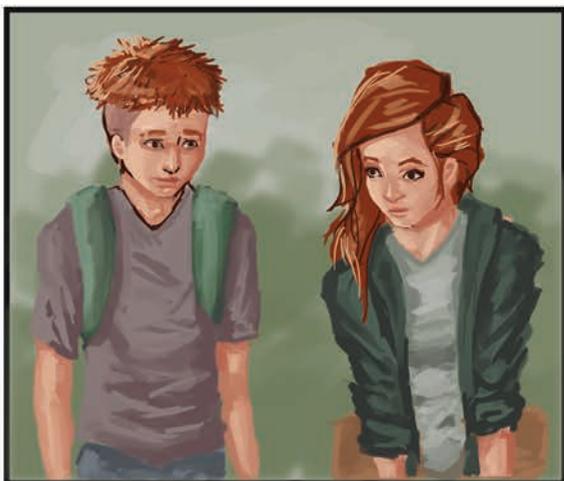


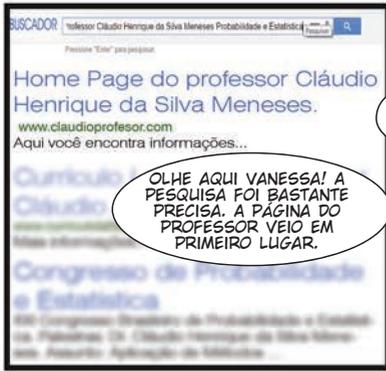
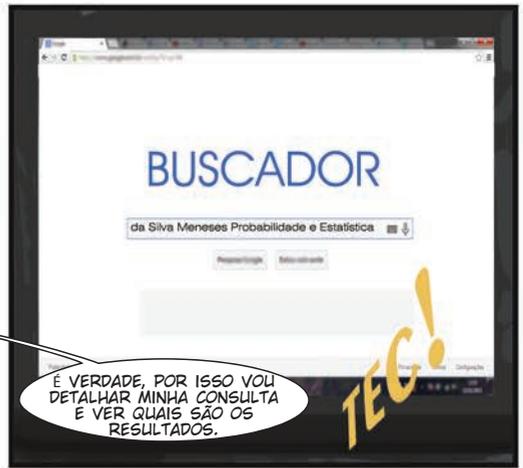
livro



BEM PESSOAL, PELO VISTO CHEGAMOS AO FINAL DA NOSSA PRIMEIRA AULA. GOSTARIA QUE TODOS RESOLVESSEM OS EXERCÍCIOS DO 1º CAPÍTULO DO LIVRO.









ESSA CONSULTA REPRESENTA UMA NECESSIDADE DE INFORMAÇÃO DO USUÁRIO E É TRADUZIDA PARA UM CONJUNTO DE PALAVRAS-CHAVE OU ATÉ MESMO UMA PERGUNTA MAIS ELABORADA. EXISTEM SISTEMAS QUE CONSEGUEM EXTRAIR SIGNIFICADO DESSE TIPO DE PERGUNTA ELABORADA E ENTREGAR UMA INFORMAÇÃO MAIS ESTRUTURADA. POR EXEMPLO, DIANTE DE UMA SENTENÇA, O SISTEMA CONSEGUE IDENTIFICAR PESSOAS, LUGARES, EMPRESAS ETC. ESSE TIPO DE SISTEMA É CONHECIDO COMO SISTEMA DE EXTRAÇÃO DE INFORMAÇÃO.



EM SEGUNDA, A CONSULTA VAI PARA O SISTEMA DE RECUPERAÇÃO DE INFORMAÇÃO. OS SISTEMAS DE RI DEFINEM DOIS MODELOS DE REPRESENTAÇÃO.

NO PRIMEIRO, TEM-SE UMA REPRESENTAÇÃO DA CONSULTA FEITA PELO USUÁRIO. AQUI ELE VAI PROCESSAR A CONSULTA QUE FOI FEITA EM LINGUAGEM NATURAL, QUE É A LINGUAGEM USADA PELO HOMEM, E FARA UMA REPRESENTAÇÃO QUE A MÁQUINA POSSA ENTENDER. JÁ O SEGUNDO

NA VERDADE, EXISTEM SISTEMAS QUE BUSCAM ARQUIVOS MULTIMÍDIA, COMO VÍDEOS, MÚSICAS E IMAGENS. NESSE CASO, ELAS DEVEM POSSUIR UMA DESCRIÇÃO QUE SERVIRÁ COMO REFERÊNCIA PARA A CONSULTA.

VOCÊ FALA EM DOCUMENTOS, ITENS DE INFORMAÇÃO E LINGUAGEM NATURAL. NESSE CASO, OS SISTEMAS DE RI SÓ RECUPERAM TEXTO.

SIM! SEMPRE VISITO SITES DESTE TIPO NA INTERNET.

EU ACHO QUE VOCÊ JÁ VIU SITES DE BUSCA DE VÍDEO, MÚSICA E FOTOS, NÃO É?



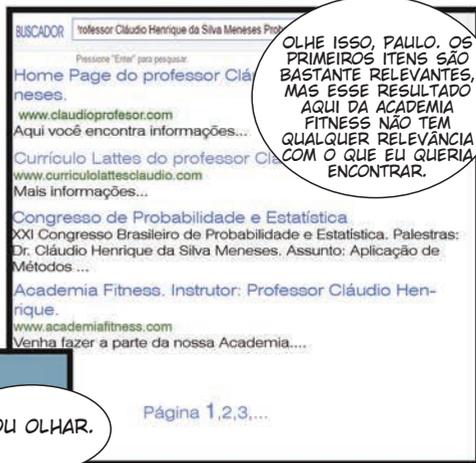
A PARTIR DO QUE FOI REPRESENTADO E ARMAZENADO NO COMPUTADOR É POSSÍVEL RECUPERAR OS ITENS DE INFORMAÇÃO (OU DOCUMENTOS) ATRAVÉS DOS SISTEMAS DE RI. FINALIZANDO O PROCESSO DE RECUPERAÇÃO, O SISTEMA DEVOLVE UM RESULTADO PARA O USUÁRIO COM OS ARQUIVOS ORDENADOS POR ORDEM DE RELEVÂNCIA.



EU JÁ HAVIA PERCEBIDO A PRECISÃO DESSE TIPO DE SISTEMA, PRINCIPALMENTE FERRAMENTAS DE BUSCA NA INTERNET. OS PRIMEIROS RESULTADOS SÃO, NA MAIORIA DAS VEZES, OS QUE EU ESTOU INTERESSADO.

SIM. HOJE EM DIA OS SISTEMAS DE RI EM GERAL CONSEGUEM CALCULAR ISSO MUITO





OLHE ISSO, PAULO. OS PRIMEIROS ITENS SÃO BASTANTE RELEVANTES, MAS ESSE RESULTADO AQUI DA ACADEMIA FITNESS NÃO TEM QUALQUER RELEVÂNCIA COM O QUE EU QUERIA ENCONTRAR.



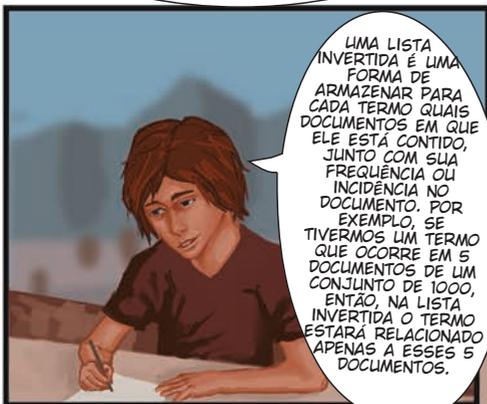
termo/Documento	Doc 01	Doc 02	Doc 03	Doc 04	Doc 05	Doc N
Termo 01	15	10	0	3	0	0
termo 02	0	0	2	0	0	0
termo 03	0	0	0	0	1	0
termo N	0	0	0	0	0	0

O GRANDE PROBLEMA DESSA ABORDAGEM ENVOLVE O PROCESSAMENTO DE UMA GRANDE COLEÇÃO DE DOCUMENTOS. EXISTE UM GRANDE DESPERDÍCIO DE ESPAÇO Nesses casos, pois a maioria dos termos não existe na maioria dos documentos. haveria um grande número de zeros nessa tabela. Essa abordagem é inviável para um grande conjunto de documentos.



NESSE CASO, A FORMA MAIS USADA NOS SISTEMAS DE RI, INCLUSIVE NAS FERRAMENTAS DE BUSCA NA INTERNET, É ATRAVÉS DE LISTAS INVERTIDAS.

LISTAS INVERTIDAS?



UMA LISTA INVERTIDA É UMA FORMA DE ARMAZENAR PARA CADA TERMO QUAIS DOCUMENTOS EM QUE ELE ESTÁ CONTIDO, JUNTO COM SUA FREQUÊNCIA OU INCIDÊNCIA NO DOCUMENTO. POR EXEMPLO, SE TIVERMOS UM TERMO QUE OCORRE EM 5 DOCUMENTOS DE UM CONJUNTO DE 1000, ENTÃO, NA LISTA INVERTIDA O TERMO ESTARÁ RELACIONADO APENAS A ESSES 5 DOCUMENTOS.



COM UMA LISTA DESSE TIPO, O GANHO EM DESEMPENHO É CONSIDERÁVEL.

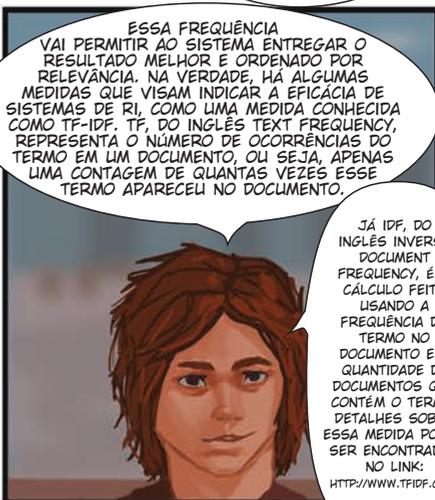


AQUI PODEMOS VER QUE NÃO OCORREM REPETIÇÕES EM CASOS NOS QUAIS OS TERMOS NÃO EXISTEM. CADA TERMO POSSUI UM FORMA DE IDENTIFICAR EM QUAL DOCUMENTO ELE ESTÁ CONTIDO E QUANTAS VEZES ELE APARECE.



REALMENTE FAZ SENTIDO. EU APOSTO QUE ESSAS FREQUÊNCIAS SÃO USADAS PARA ORDENAR OS RESULTADOS DE ACORDO COM SUA RELEVÂNCIA, FAZENDO AQUELE CÁLCULO LA QUE ESQUECI O NOME.

CÁLCULO DE SIMILARIDADE, VANESSA. AQUELA FORMA, APENAS INDICANDO SE EXISTE OU NÃO O TERMO NO DOCUMENTO, É CONHECIDA COMO MODELO BOOLEANO. QUANDO TEMOS A FREQUÊNCIA DA PALAVRA COM ALGUMA OUTRA MEDIDA QUE REPRESENTA O PESO DO TERMO NO DOCUMENTO, O MODELO É CONHECIDO COMO MODELO VETORIAL.



ESSA FREQUÊNCIA VAI PERMITIR AO SISTEMA ENTREGAR O RESULTADO MELHOR E ORDENADO POR RELEVÂNCIA. NA VERDADE, HÁ ALGUMAS MEDIDAS QUE VISAM INDICAR A EFICÁCIA DE SISTEMAS DE RI, COMO UMA MEDIDA CONHECIDA COMO TF-IDF. TF, DO INGLÊS TEXT FREQUENCY, REPRESENTA O NÚMERO DE OCORRÊNCIAS DO TERMO EM UM DOCUMENTO, OU SEJA, APENAS UMA CONTAGEM DE QUANTAS VEZES ESSE TERMO APARECEU NO DOCUMENTO.

JÁ IDF, DO INGLÊS INVERSE DOCUMENT FREQUENCY, É O CÁLCULO FEITO USANDO A FREQUÊNCIA DO TERMO NO DOCUMENTO E A QUANTIDADE DE DOCUMENTOS QUE CONTÉM O TERMO. DETALHES SOBRE ESSA MEDIDA PODEM SER ENCONTRADOS NO LINK: [HTTP://WWW.TFIDF.COM/](http://www.tfidf.com/)





BIBLIOGRAFIA

BAEZA-YATES, R.; RIBEIRO-NETO, B. Modern information retrieval. New York: ACM press, 1999.
MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. An Introduction to Information Retrieval. Cambridge University Press, 2008.

RUSSEL, S., NORVIG, P. (1995). Artificial Intelligence – A Modern Approach. Prentice Hall. (<http://aima.cs.berkeley.edu/>).

SALTON, G.; MCGILL, M.J. Introduction to Modern Information Retrieval. McGraw-Hill, New York, NY, 1983.

MAIS CARTILHAS EM:

<http://almanaguesdacomputacao.com.br/index.html>

<http://meninasnacomputacao.com.br/gutanunes/publication.html>

<http://meninasnacomputacao.com.br/>

SOBRE OS AUTORES

Maria Augusta Silveira Netto Nunes

Bolsista de Produtividade Desen. Tec. e Extensão Inovadora do CNPq - Nível 2 - CA 96 - Programa de Desenvolvimento Tecnológico e Industrial Professor Adjunto IV do Departamento de Computação da Universidade Federal de Sergipe. Membro do Programa de Pós-graduação em Ciência da Computação (PROCC) na UFS. Pós-doutora em Propriedade Intelectual no Instituto Nacional de Propriedade Industrial (INPI). Doutora em "Informatique pela Université de Montpellier II - LIRMM em Montpellier, França (2008). Realizou estágio doutoral (doc-sanduíche) no INESC-ID-IST Lisboa- Portugal (ago 2007-fev 2008). É mestre em Ciência da Computação pela Universidade Federal do Rio Grande do Sul (1998) e possui graduação em Ciência da Computação pela Universidade de Passo Fundo (1995). Possui experiência acadêmico-tecnológica na área de Ciência da Computação e Inovação Tecnológica/Propriedade Intelectual. Atualmente, suas pesquisas estão voltadas, principalmente na área de inovação Tecnológica usando Computação Afetiva na tomada de decisão Computacional. Atua também em Inovação Tecnológica, Propriedade Intelectual capacitando empresários na área de TI e fornecendo consultoria em Registro de Software e patente.
Lattes: <http://lattes.cnpq.br/9923270028346687>

Rafael Meneses Santos

Possui graduação em Sistemas de Informação pela UFS – Universidade Federal de Sergipe (2013.2) e cursa o mestrado em Ciência da Computação pela Universidade Federal de Sergipe (2014.1) na linha de pesquisa de Computação Inteligente. Tem experiência nas áreas de Mineração de Dados, Data Warehouse, Banco de Dados, Desenvolvimento Web e Processamento de Linguagem Natural.

Sean Wolfgang Matsui Siqueira

Jovem Cientista do Nosso Estado, da FAPERJ

Professor Associado da Universidade Federal do Estado do Rio de Janeiro (UNIRIO). Membro do Programa de Pós-graduação em Informática (PPGI) da UNIRIO. Doutor em Ciências - Informática, Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio, 2005). É mestre em Informática pela PUC-Rio (1999) e possui graduação em Ciências da Computação pela Universidade Federal de Goiás (1996). Tem experiência na área de Ciência da Computação, com ênfase em Sistemas de Informação e Informática na Educação, atuando principalmente nos seguintes temas: web semântica, web social, ontologias, redes sociais, aprendizagem apoiada por computador, objetos de aprendizagem, integração de dados, análise de dados, data warehousing, recuperação da informação, CRM, portais corporativos, gerência de conhecimento, modelagem de objetos complexos, sistemas de informação musical, mineração de dados, texto e web. Foi o coordenador do Programa de Pós-Graduação em Informática (PPGI) da UNIRIO de julho/2012 a setembro/2014 e atualmente está coordenando os comitês de programa do do Simpósio Brasileiro de Sistemas de Informação (SBSI 2015), além de ser o editor-chefe da iSYS: Revista Brasileira de Sistemas de Informação e um dos editores da edição especial "Computing for Human Learning, Behaviour and Collaboration in the Social and Mobile Networks Era" do periódico Computers in Human Behavior (CHB). Foi o coordenador do comitê de programa do Simpósio Brasileiro de Informática na Educação (SBIE) nos anos de 2012 e 2014 e é membro da Comissão Especial de Informática na Educação (CEIE) da Sociedade Brasileira de Computação (SBC).

Yargo Santana Vasconcelos

Bolsista COPES(IC)

Graduando em Design Gráfico Pela Universidade Federal de Sergipe e bolsista COPES(IC).

Experiência em ilustração com ênfase no digital.

AGRADECIMENTOS

Ao CNPq, CAPES, SBC, DCOMP, PROCC, PROEX, BICEN e CINTTEC/UFS.

APOIO:



Conselho Nacional de Desenvolvimento Científico e Tecnológico



UNIVERSIDADE FEDERAL DE SERGIPE
Aracaju (Sergipe), capital da qualidade de vida no Brasil!

<http://www.posgraduacao.ufs.br/procc>
e-mail: procc.ufs@gmail.com

<https://www.linkedin.com/company/proccuf>
@PROCC_UFS



ISBN 978-857669411-3



9 788576 694113