CM News -

Tweetar

Classificação requintada

24/08/2015 - Software desenvolvido na Universidade de São Paulo filtra grande quantidade de dados digitais e os separa a partir de associações próprias

» ROBERTA MACHADO

Para muitos, organizar as pastas e os arquivos virtuais do computador é uma tarefa continuamente adiada. Assim também acontece com a obrigação de ler e separar as mensagens acumuladas na caixa de entrada do e-mail, uma ideia que se torna mais insuportável a cada novo recado que chega. Imagine, então, como difícil seria examinar todo o conteúdo publicado em sites da internet, como portais de notícias, blogues e redes sociais. Um desafio impossível para humanos. Para máquinas, porém, um trabalho que pode ser cumprido sem dificuldades. Um software em desenvolvimento no Instituto de Ciências Matemáticas e de Computação (ICMC) da Universidade de São Paulo (USP) em São Carlos consegue classificar automaticamente grande quantidade de textos digitais.

Trata-se de um algoritmo que identifica os termos usados em cada tipo de texto e analisa a relação entre as palavras para classificar um novo documento. Tudo é feito de acordo com os exemplos dados por humanos. Se uma biblioteca virtual tiver vários tipos de arquivos científicos, por exemplo, bastaria cadastrar no programa alguns trabalhos relacionados a cada assunto. A partir de alguns exemplos de cada categoria, o programa conclui a organização por conta própria.

A maioria dos programas de classificação automática de textos considera a frequência com que certas palavras-chave aparecem nos documentos. No entanto, os algoritmos desenvolvidos pelo aluno de doutorado do ICMC Rafael Rossi também são capazes de interpretar as redes formadas por associações entre termos, o que permite ao computador identificar padrões não assimilados em outros tipos de representações, tornando o software mais eficiente. Por meio do aprendizado de máquina, o sistema pode se aperfeiçoar na sua tarefa, imitando o discernimento de um humano sem que ele tenha de ser especialmente programado.

“O que propomos é considerar a similaridade entre termos em uma coleção de documentos. Se tenho a palavra ‘banco’ e ‘dados’ no mesmo documento, elas serão similares. Se, por outro lado, eu tiver os termos ‘banco’ e ‘redes’, que são áreas distintas, elas não serão similares. Assim, definimos o que chamamos de valor de relevância. Seria o peso ou a força que um termo tem para determinado documento. O objetivo é usar a relação de similaridade para definir essa relevância”, explica Rafael Rossi.

Testes
Graças a esse modelo inteligente, o sistema pode aprender um tipo de classificação com menos de 10 exemplos — em um software de classificação comum, é necessário cadastrar ao menos 100 documentos para garantir a mesma taxa de acerto. O programa foi usado inicialmente como ferramenta de monitoramento de aprovação dos candidatos à Presidência do país nas eleições do ano passado, mas também já foi testado em coleções de artigos acadêmicos, páginas da web e como filtro de e-mails.

O programa pode ser usado para classificar qualquer tipo de texto, desde que uma pessoa determine os parâmetros que precisam ser seguidos. Uma empresa preocupada em saber o grau de satisfação dos clientes nas redes sociais poderia, por exemplo, selecionar alguns exemplos de posts considerados negativos e positivos. Não é necessário usar hashtags ou determinar que palavras são importantes: o programa determina por conta própria os termos que definem cada tipo de texto. Assim, o software pode filtrar milhares de publicações e indicar quantas mensagens elogiosas e quantas reclamações sobre a companhia foram publicadas na rede.

Outras possíveis aplicações para o software seria a classificação de obras literárias, de notícias e até mesmo de e-mails, conforme o conteúdo do texto. “Esse método pode ser aplicado com facilidade para o atendimento pós-venda ou para selecionar os e-mails que determinada empresa recebe. Imagine que os e-mails podem ser classificados e respondidos pelos funcionários de acordo com o assunto. Então, o método já pode encaminhar a mensagem para o empregado correto”, exemplifica Solange Rezende, professora do Departamento de Ciências da Computação da USP São Carlos e orientadora do projeto.

Sensor de doenças
O websensor ainda poderia ser usado como um agente automatizado, feito para detectar sentimentos, tendências sociais, indicadores políticos e econômicos e até mesmo o surgimento de epidemias, conforme as palavras publicadas na internet. “Temos conseguido resultados muito impressionantes. Acreditamos que esse trabalho vai revolucionar a área”, diz Rezende.

O trabalho brasileiro foi premiado na International Conference on Intelligent Text Processing and Computational Linguistics, uma das principais conferências de linguística de mineração de textos do mundo. O evento, realizado em abril, no Egito, recebeu mais de 300 artigos, e a proposta de Rafael foi uma das únicas duas que receberam o maior reconhecimento do quadro internacional de especialistas.

Fonte: Correio Braziliense / DF

Onde Estamos

Marília - SP (Sede)
Rua Coronel José Braz,1443 - CEP 17.502-010 - Fone: (14) 3402-3333 - Fax: (14) 3402-3331 - diretoria@cmconsultoria.com.br

Brasília - DF
SC/Norte-Quadra 05, Bloco A, sl. 1022 à 1025 - Ed. Brasília Shopping -CEP 70.715-900- Fone/Fax: (61) 3328-7305 - brasilia@cmconsultoria.com.br

Serviços
Consultoria
CM On-line
Programas de Capacitação
Vídeo Conferência
Seminários e Eventos
Plantão de Dúvidas

Consultores
Blog do Presidente
Consultores CM

Central de Conhecimento
Vídeos
PodCast
Links

Clientes

Sobre a CM

Portal @prender

Receba

Contato