Revista de Letras
ISSN: 2179-5282

https://periodicos.utfpr.edu.br/rl

Contribuições da linguística computacional para a manipulação de dados

RESENHA

Valter de Carvalho Dias
valtinhodias@gmail.com
Instituto Federal de Educação, Ciência e Tecnologia, Simões Filho, Bahia, Brasil.

O livro Para conhecer a Linguística Computacional, de autoria dos professores Marcelo Ferreira e Marcos Lopes, busca refletir sobre o funcionamento dos sistemas computacionais que atuam diretamente na comunicação humana na atualidade, os quais são responsáveis por tarefas que já são consideradas naturais por qualquer usuário dos equipamentos tecnológicos, como a correção automática de palavras que são digitadas em um processador de texto ou, até mesmo, em um aparelho de celular; a possibilidade de reconhecimento da voz humana e transformá-la em texto escrito e vice-versa; entre outros.

Os autores são professores da Universidade de São Paulo, atuando no Departamento de Linguística da Faculdade de Filosofia, Letras e Ciências Humanas. Marcelo Ferreira é doutor em Linguística pelo MIT – Massachusetts Institute of Technology –, com estágios pós-doutorais na Universidade de São Paulo e Universidade de Maryland, nos Estados Unidos. É livre-docente na Universidade de São Paulo com os "Estudos formais sobre a semântica nominal e verbal do português", obtido em 2018. Atualmente, coordena o projeto de pesquisa "Modelos de Língua", o qual é dedicado à investigação e implementação de modelos probabilísticos de linguagem nos vários níveis de análise linguística do português. Além deste livro, é autor também do "Curso de Semântica Formal", publicado este ano em Berlim, pela Language Science Press.

Por sua vez, Marcos Lopes é doutor em Ciências da Linguagem pela Universidade Paris X, na França, e estágios pós-doutorais pela Universidade de São Paulo, Universidade de Quebec, em Montreal, Canadá, e pela Universidade de Bremen, na Alemanha. É coordenador de dois projetos de pesquisa: (i) "Fundamentação simbólica do léxico dicionarizado", através do qual busca, entre outras coisas, explicar como símbolos semanticamente fundamentados transmitem sua significação ao longo da cadeira computacional; e (ii) "Cálculo da Perspectiva Dêitica através do Raciocínio Espacial Qualitativo", que visa um estudo

semântico-computacional da espacialidade nos enunciados de língua natural, com o propósito de gerar modelos de descrição e inferência sobre as relações espaciais.

A proposta do livro é delineada claramente na apresentação, a partir de dois questionamentos: Como funcionam os sistemas computacionais? Que relações tem a sua organização linguística interna com o que se faz nas descrições e análises linguísticas tradicionais? Em resposta à essas perguntas, os autores esclarecem ao seu leitor que o livro traz alguns elementos que permitem a análise linguística e que será necessário experimentar não só as ferramentas que serão apresentadas, mas também a avaliação das outras possibilidades do seu conhecimento.

Na apresentação, ainda, esclarecem o caráter multidisciplinar da Linguística Computacional, conhecida também como "Processamento de Linguagem Natural – PLN", a qual é estudada não só pela Linguística e Ciências da Computação, mas também pelas Neurociências, Filosofia e Psicologia, compondo assim, conforme os próprios autores mencionaram, as "ciências cognitivas".

O leitor percebe logo no início que se trata de uma proposta bastante didática, uma vez que na apresentação são sugeridas leituras introdutórias à área proposta, a fim de que se possa conhecer melhor os conceitos básicos da Linguística Computacional, os quais serão mencionados ao longo do livro.

Isso é deixado claro, ao afirmarem que

> O material que você tem em mãos foi concebido como um curso, mais do que como um manual de linguística computacional. Nossa ideia central é oferecer o apoio necessário para o enfrentamento inicial nessa dupla formação necessária para se trabalhar nesse domínio. Não se espera, portanto, que você tenha qualquer experiência prévia com programação de computadores. (p. 10).

Uma outra obra significativa que parte dessa mesma premissa é o livro Linguística Computacional: teoria & prática, de autoria de Gabriel Othero e Sérgio Menuzzi, publicado em Parábola Editorial, em 2005. Diferentemente de Ferreira e Lopes que abordam a linguagem Python, Othero e Menuzzi fazem uso da linguagem Prolog de programação.

Voltando-se para o livro objeto deste texto, os quatro capítulos estão didaticamente estruturados com uma parte teórica e/ou teórico-prática; com leituras sugeridas para ampliação dos conteúdos tratados; e exercícios práticos a serem realizados no computador, de forma que o leitor poderá se colocar no papel de aluno e exercitar os conhecimentos vislumbrados a partir da leitura cuidadosa.

Além disso, cada capítulo inicia apresentando os objetivos gerais pretendidos para situar o leitor qual será o aprendizado esperado. E, ao longo deles, figuram-se alguns quadros explicativos, com as explicações específicas sobre os programas, com notas técnicas ou com notas bibliográficas, a fim de complementar as informações.

Cabe ainda ressaltar o caráter ilustrativo da obra. Todo capítulo é uma verdadeira aula. Não bastando apenas conceituar, explicar o funcionamento e como se processa cada código, os autores colocam toda a codificação em quadros com fundo cinza e borda escura, o que visivelmente chama a atenção do leitor para o que será necessário digitar em seu computador para processar os dados que estão sendo manuseados. Essas "ilustrações" mantém diálogo estrito com os

comentários, uma vez que neles são descritos cada símbolo empregado na codificação, para que serve e qualquer sua funcionalidade, principalmente se agrupados.

Um outro recurso empregado pelos autores, não de forma abundante, é um quadro contendo uma "dica". Fazendo o emprego da primeira pessoa do plural, os autores se aproximam ainda mais do leitor ao lhe assegurar melhor entendimento do que se pretende no capítulo ou numa seção específica, garantindo-lhe fluência em programação.

O primeiro capítulo é uma preparação prévia do leitor, introduzindo-o na linguagem Python, considerando-o um leigo em computação, destacando as noções básicas, inclusive a instalação passo-a-passo dessa ferramenta. Além disso, é possível perceber como essa linguagem figurará ao se programar no computador: de "modo interativo", ou seja, a informação que é digitada é imediatamente interpretada e o resultado disponibilizado; ou a partir de "scripts", os quais gravam a codificação a ser interpretada, podendo ser empregada a qualquer momento, de acordo com a necessidade do próprio programador.

Segue-se o capítulo com outros conceitos e códigos de processamento de dados que facilitarão o manuseio de todo e qualquer corpus estruturado para esse fim. Vale ressaltar que esta parte do livro é considerada pelos autores como bastante básica, sugerindo, inclusive, que se o leitor já tiver esses conhecimentos, pode seguir direto para o próximo capítulo.

Uma vez conhecidos os primeiros passos da linguagem Python, prossegue-se para o capítulo segundo, que tratará especificamente da Análise quantitativa de corpus. Sabendo que se trata de um breve manual, os autores iniciam esta parte definindo corpus, para que serve, os seus tipos e como obtê-lo para o desenvolvimento de uma pesquisa linguística. Eles citam alguns corpora já constituídos, mencionando-os nominalmente, inclusive com os endereços de suas respectivas páginas na Internet.

Os autores, a partir dessa parte, orientam como organizar um corpus não-estruturado de forma a proceder a análise quantitativa, a partir da estatística descritiva, a qual prevê a obtenção da média, mediana, moda, variância e desvio-padrão. Com essas noções esclarecidas para o leitor, inicia-se a orientação de como processar essas informações na linguagem Python, inclusive na criação de gráficos e a avaliação da relevância dos dados.

Percebe-se, principalmente nesse capítulo, que os iniciados na programação em R (linguagem empregada para o tratamento estatístico de dados, seja linguísticos e não-linguísticos) não terão muita dificuldade em acompanhar o desenvolvimento proposto por Ferreira e Lopes, uma vez que alguns conceitos e procedimentos são compartilhados por ambos.

O terceiro capítulo se volta para a análise probabilística a partir do modelo conhecido como n-grama, conceituado pelos autores como "[...] uma sequência de n elementos em um determinado nível de análise (letras, morfemas, palavras etc.)" (p. 129), isso estruturado diretamente em Python. Ainda de acordo com os autores, é possível utilizar um modelo previamente elaborado, como também criar um novo, de acordo com os objetivos pretendidos pelo pesquisador.

Antes de concluir esta parte, é possível saber alguns problemas que podem ocorrer ao programar e como resolvê-los: quando uma palavra não ocorre no

corpus de treinamento; a probabilidade de palavras com frequências muito baixas no treinamento se mostrarem pouco confiáveis; os inúmeros parâmetros que os modelos podem atingir quando apresentam vocabulários extensos; entre outros.

O quarto e último capítulo, Classificadores bayesianos ingênuos, faz um panorama sobre os classificadores, com breve apresentação, construção, treino, avaliação e implementação de um classificador. Segundo os autores, classificar é "agrupar dados, objetos ou observações em categorias predeterminadas" (p. 161), o que é muito importante para estudos linguísticos, pois possibilita verificar o quão confiáveis são os resultados obtidos a partir das probabilidades.

Por ser um livro bastante didático, um leigo em informática, bem como qualquer pesquisador que não seja da área da Linguística, não sentirão qualquer dificuldade em seguir a proposta de Ferreira e Lopes. Todos os capítulos apresentam a teoria que embasa cada análise linguística que pode ser desenvolvida, as noções da estatística aplicada ao modelo computacional empregado, sem contar com os inúmeros exemplos, o que permite com que o leitor faça associações com o seu próprio objeto de pesquisa. As leituras sugeridas e os exercícios ainda servem de auxílio para ampliar o conhecimento e sanar qualquer dúvida que apareça ao longo da leitura/treinamento.

Os autores, nas considerações finais, retomam os objetivos gerais pretendidos e o que se buscou realizar em cada capítulo, situando o leitor na tarefa que foi empreendida ao longo de sua leitura, como um verdadeiro curso de Linguística Computacional, organizado de forma a garantir a autonomia no aprendizado e no estímulo necessário para buscar em outras fontes as informações complementares para a sua formação na análise linguística mediada pela tecnologia.

Deixou-se claro também que não se tratava de um livro que buscaria explorar completamente todos os recursos disponíveis e que o leitor, ao realizar a leitura e os exercícios propostos não se tornaria um expert em linguística computacional. O objetivo não era esse, mas sim levá-lo a se familiarizar com a linguagem e permitir uma maior autonomia para a busca do aperfeiçoamento, caso esse se torne um dos seus objetivos futuros.

Nesse intuito, cabe trazer à baila mais uma vez a obra de Othero e Menuzzi (2005), cujo diálogo teórico continua como uma pauta importante para os estudos linguísticos após quase quinze anos do seu lançamento. Os aspectos tecnológicos que envolvem as duas obras são distintos, podendo haver até uma desatualização em relação à linguagem tratada nessa obra, mas a contribuição é deveras importante para o estabelecimento da subárea da Linguística que se preocupa com as ferramentas de análise linguística mediadas por computador.

As referências apontadas no final da obra de Ferreira e Lopes demonstram, como os próprios autores mencionaram, que a Linguística Computacional é "uma área recente, mas impulsionada por antigos anseios da ciência: criar máquinas capazes de compreender e produzira linguagem humana." (p. 188). Dessa forma, o leitor, além das leituras sugeridas ao longo de todo o livro, poderá buscar conhecer os textos que embasaram os ensinamentos de seus autores.

REFERÊNCIAS

FERREIRA, Marcelo; LOPES, Marcos. **Para conhecer linguística computacional**. São Paulo: Contexto, 2019.

OTHERO, Gabriel de Ávila; MENUZZI, Sérgio de Moura. **Linguística computacional**: teoria & prática. São Paulo: Parábola, 2005.

Recebido: 01 out. 2019
Aprovado: 21 nov. 2019
DOI: 10.3895/rl.v21n35.10910

Como citar: DIAS, Valter de Carvalho. Contribuições da linguística computacional para a manipulação de dados. *R. Letras*, Curitiba, v. 21, n. 35 p. 129-133, jul/dez. 2019. Disponível em: <https://periodicos.utfpr.edu.br/rl>. Acesso em: XXX.