書評

佐藤進也（著）

「Pythonではじめる情報検索プログラミング」

- ・ 出版社：森北出版
- ・ 発行日：2020 年 12 月 18 日
- ・ 菊判，160 頁，2,640 円
- ・ ISBN：978-4-627-81861-3

柴田 淳司（評）

東京都立産業技術大学院大学

この本は情報検索に特化した自然言語処理とそのプログラミングの入門書である。読者の中には、「入門書としてなら特定分野に特化した本より全体を網羅するような本のほうがいいのでは」と思いかもしない。しかし、昨今の人工知能分野に関していえば、本書のような参考書を私はお勧めしたい。

近年、情報技術、特に人工知能関連の技術は大いに発展し、実社会の様々な場面で利用されるようになった。そして同じ人工知能といっても、その適用先の分野によって必要とされる知識と技術も多様化する傾向にある。例えばテキストを対象とした自然言語処理でも、翻訳、要約、意味理解、対話システムなど、その対象は多岐にわたり、使われる手法や用語も完全に共通とはいえない。

こうした背景から、現在の人工知能分野は非常にとっつきづらいものとなっている。そのため、全体を網羅する本よりも、本書のように一つの内容に特化した本を使って、できるところから学んでいくことは初学者にとって重要なことだろう。

この本の特徴として、初学者向けに簡潔に工夫して書かれている点が挙げられる。まず、数式などどっつきづらくなる話は極力省かれている。プログラミングに関しても、既存のパッケージやフレームワークを用いるものがほとんどで、複雑な演算処理を手書きさせるようなことはない。また、コードも配布されており、実行しながら学ぶことができるようになっている。頁数も 160 頁と手軽に読めるボリュームである。最先端の難解な理論についても省かれているため、大学 1～2 年生の独習に向く。

まずはこの本で基礎を学んだあと、理論、プログラミング、自然言語処理全体や最新の手法など、コアなところを学びたい方は別の専門書に進むのが良いだろう。

本書は 10 章と付録から構成されている。それぞれの章は 10 ページ程度にまとめられている。各章の初めにはその章で学べるキーワードも付いているため、必要なところをピンポイントで学べるようになっている。

1 章では文字について、文字コードや正規表現などの文字をどう表すかについて書かれている。

2 章では語について、形態素解析によりテキストデータを語に分けることについて書かれている。また、その結果を視覚的

にわかりやすくするようワードクラウドを利用している。

3 章では特徴語について、不要語や TF・IDF を用いて、2 章で得られた語の特徴を数値化している。

4 章では文書のランキングについて、文書を検索単語に合わせて順序付けするために類似度の概念について解説している。

5 章では検索システムの評価について、4 章で上げた文書のランキングの妥当性を評価する方法について解説している。このシステムの評価やテスト方法は他の機械学習分野でも使われる重要なものである。

6 章では文書のクラスタリングについて、k 平均法や階層的クラスタリングなど基本のクラスタリング手法を紹介、解説している。デンドログラムを用いて視覚的に提示しているため、初めてクラスタリングを行う人にも感覚がつかみやすいだろう。

7 章では文書のクラス分類について、ナイーブベイズやサポートベクターマシンなどの基本的な機械学習の手法を紹介している。絵を用いてできるだけ視覚的にわかりやすいよう配慮していることが伺える。ただし、7、8 章にある手法の理論を本格的に学ぶには本書の頁数では足りていないため、補助の参考書が必要だろう。

8 章ではクエリの修正について、クエリを修正するという考え方と適合性フィードバックを解説している。頁数が少なく、登場する数式や意味を深く理解するのはやや難しいかもしれない。

9 章ではトピックについて、潜在的意味を導入した後の検索について解説している。本章で登場する内容は意味要約や推薦システムなどにも使うことができる。

10 章では表現学習について、ニューラルネットワークを用いた Word2vec、Doc2vec の解説をしている。この手法は文脈自由モデルとして自然言語処理でよく利用されているもので、解説にも力が入っていることが窺える。

また、付録として Python の基本的な利用方法といくつかの Web データの紹介を行っている。

総論として、本書は情報検索とそのプログラミングに焦点を当てて書かれた入門書である。頁数が少なく、初学者でも必要最低限の内容を押さえて独学するために向いている。一方、理論やプログラミングの詳解、Transformer や BERT などの最新手法までは押さえていない。総じて、大学学部 1～2 年生の演習などに向けた自然言語処理の入門書であると言える。本書でまずは自然言語処理で行われていることに触れて感覚を掴み、他の参考書を合わせてより深く学んでいくことが良いだろう。