令和 6(2024)年 3 月 14 日
国 立 国 会 図 書 館

ウェビナー「日本研究のための情報源活用法」(令和 5 年度)に係る
講師への質問及び講師からの回答

■講義:全文テキストデータを利用した近現代日本文学の研究
■講師:日比　嘉高氏(名古屋大学大学院人文学研究科教授)

(質問 1)
　講義で紹介されたような解析データは指針とするもので、検索データに制限があることから研究結果にはならないでしょうか。時代を特定すれば問題ないのでしょうか。

(回答)
　今回使ったデータでも、「～という点で制限(限界)がある」ということを明示すれば、一定の研究成果にはなると考えます。とはいえ、今回のやり方で構築したデータセットはかなり簡易なもので、文学研究の観点からすると不十分であることも確かです。再刊本・再々刊本が含まれていたり、短編集が各作品に切り分けできていなかったり、パラテクスト(奥付や柱、ノンブル、広告など)が含まれていたり、文字認識ミスが数パーセント残っていたり、ということなどが問題点として挙げられます。文学研究としての精度を上げていくならば、これらをできるだけ整えていく必要があります。

(質問 2)
　著作権などの関係からダウンロードできる範囲に制約を受けることもあろうかと思いましたが、AI を通して分析を行う場合、活用が可能な資料の範囲で工夫を行う機会は多くなっておられるでしょうか。

(回答)
　はい、工夫次第で扱えるデータは増えていると思いますし、今後もどんどん増えるだろうと思います。工夫やデータ上の課題については質問 1 についての回答もご参考になるかもしれません。

（質問 3）
　例 1「トピック分析」の手順③にある、本文テキストを単語ごとに分割する工程についてですが、これも LDA でできるのでしょうか。それとも別に何らかのツールがあるのでしょうか。

（回答）
　これは「形態素解析」という技術を用いて行います。自然文を、コンピュータを用いて単語に切り分けます。形態素解析を行うツールとしては、 MeCab、Juman++、Janome などがよく使われます。各ツール（辞書）によって、得手不得手があります。ツールを実行するにはさまざまな方法があるようですが、私は Python を使って実行しました。自分で Python のコードを書けない場合でも、Chat-GPT を使うとコードを生成してくれます。（もっとも、生成されたコードがそのままエラーなく使えることはあまりありませんので、修正が必要です。表示されたエラーを Chat-GPT に示すことで、修正も自動で行ってくれます。）

（質問 4）
　Voyant-tools を使った事例があれば、教えていただけるとありがたいです。

（回答）
　日本文学研究に関しては、今のところ見つけられていませんが、以下の文献では Voyant-tools を使っているようです。

須永 恵美子. 南アジア研究トピックの変遷 : 学術論文データベースを活用したテキストマイニングによる分析. 高崎商科大学紀要 = The journal of Takasaki University of Commerce / 高崎商科大学メディアセンター 編. (37):2022,p.211-216.
https://ndlsearch.ndl.go.jp/books/R000000004-I032787831

（質問 5）
　キーワード検索のご説明のなかで、関連するワードが連鎖的に増えるとありましたが、これは、キーワード検索の結果、ご自身で関連ワードを発見されたのか、それとも、AI が自動的に関連ワードを増やしているのでしょうか？

（回答）
　自分自身で、という意味で話しました。このあたりが AI の利用が進んだとしても、人の出番がなくなるわけではない部分と思います。つまり、示された検索結果や分析結果を読んで、次のアイデアに結びつけていく際に、専門知の蓄積が物を言うということです。

（質問 6）
　全文テキストに対するトピック分析の前処理で、「本文テキストを単語ごとに分割し、かつ形態素解析する」とありましたが、ここにコードをご利用されたのでしょうか？文章を単語ごとに分割できるということ自体が驚きでした。ここの辺りをもう少し詳しく伺いたいです。

（回答）
　質問 3 への回答に、あわせて記述しましたのでご覧下さい。

（質問 7）
　例 1「トピック分析」では、手順③についても Python で行ったという理解でよろしいでしょうか。次世代デジタルライブラリーからダウンロードした全文テキストデータは、ページごとにファイルが分かれており、テキストデータを統合するなどの作業も必要だったのではないかと推察します。そういった作業も含めて Python で行うことができるということでしょうか。

（回答）
　はい、Python を使いました。Chat-GPT による支援を受けています。これについては質問 3 への回答もあわせてご覧下さい。
　次世代デジタルライブラリーからダウンロードした全文テキストデータは、zip フォルダで得られます。含まれるファイルには txt、json の 2 種類があります。txt と json それぞれについて、ページごとのデータと、全体を統合したデータがふくまれています。ページごとのデータは「1083062_0000001.txt」「1083062_0000001.json」などのように _000…という連番が付いています。全体を統合したデータは、txt ファイルの場合、「11083062.txt」のように連番がありません。json についても連番はなく、また拡張子が「11083062.jsonl」というものに変わっています。利用の目的に合致したファイル形式を選びます。

（質問 8）
　講義の中で、Python の入門書を 1～2 冊読むと Chat-GPT4 に Python のコード作成を頼めるようになるというお話でした。今回のようなデータ解析を行うためのおすすめの Python の入門書、参考書がありましたらご教示いただけますと幸いです。

（回答）
　Python の入門書については、適切なアドバイスができるか自信はありません。私が読んだ本は超初心者向けの森巧尚 著『Python 1 年生 : 体験してわかる!会話でまなべる!プログラミングのしくみ』(翔泳社)、それから定評がありますがそれなりに手強い(分厚い…)Bill Lubanovic 著／鈴木駿 監訳『入門 Python3』(オライリー・ジャパン)でした。ネット上に記事や動画も多く、参考にしました。自然言語処理に関わる入門書としては、youwht 著『キテレツおもしろ自然言語処理』(翔泳社)、中山光樹 著『機械学習・深層学習による自然言語処理入門』(マイナビ出版)を参照しました。いずれにしても最新の版や記事を読むことが大事と思います。変化の早い領域ですので。

（質問 9）
　研究に使用されている新しいツールやその情報などは、どこから仕入れられているのでしょうか。

（回答）
　「やりたいこと」をもとに、ウェブ検索をすることがほとんどです。その場合、文学研究者としてのやりたいことを、情報学的な語彙に変換して検索するのが、コツと言えばコツかもしれません。Chat-GPT などの AI チャットへの質問もそれなりに有効だと感じています。既存の論文からツールや方法論のヒントを得ることももちろんあります。