研究ノート

グルメレビューサイトにおける口コミ内容の地域比較

Regional Characteristics Analysis of Review Content on Gourmet Review Sites

久保田　貴　文*
Takafumi KUBOTA

キーワード：テキストマイニング、トピックモデル、口コミ、Python
Keywords：Text mining, Topic model, Review, Python,

1. はじめに

本研究では、グルレビューサイトである食べログ [1] において、「うどん」のレビューのうち東京都と香川県の口コミ（レビュー）を比較することを目的とする。研究動機としては、コロナ禍において、行動制限がかかっている中で、レストラン利用者がコロナ禍以前に比べて、慎重にレストラン選定を行っていることに加え、その内容が都心と田舎でどのように異なるのかを調べることが重要だと考えたためである。また、著者が指導しているゼミの学生が本件に興味をもち、Web スクレイピングの学修にもつながると考えたことも研究のきっかけである。

研究方法としては、プログラミング言語 Python[2] を用いて該当するデータを収集し、そのデータについて統計アプリケーション Exploratory[3] を用いて基礎集計を実施し、さらに、ワードクラウド、単語ペアネットワーク、トピックモデル等により分析を実施した。

2. 研究方法

データの収集方法としては、プログラム言語 Python とそのライブラリーである Beautiful Soup[4] を用いて、Web スクレイピングと HTML 解析により、食べログのデータを収集した。そのうち、東京都内の「うどん」を提供しているレストランにおいて、レビューの星数が 3.6 以上のレビューを抽出し、そのうち、口コミにあたるテキストデータを用いて分析を進めることとした。同様に、香川県内においてもデータを収集した。ただし、店舗の全レビューを取得すると、店舗ごとにデータ件数の多少が存在するため、1 店舗当たりのレビュー数は 20 件以内とすることとした。なお、実際の食べログのページに表示されるレビューの件数は 1 ページあたり 20 件であり、それを超えると次のページに遷移する。ここでは、データは 2022 年 8 月 16 日に収集したため、それ以前のデータを含んでいる。

データ分析手法としては、分析用のアプリケーションとして Exploratory（version 6.10.6.1）

* 多摩大学経営情報学部　School of Management and Information Sciences, Tama University

を用いて、ワードクラウドおよび単語ペアネットワークを作成し、2つの地域について比較を行った。さらに、トピックモデルを用いてトピック数3で分析を行い、地域ごとの特徴を抽出した。

3. 分析結果

分析としては、Exploratory のアナリティクスにおいて、テキスト分析の単語のカウントでワードクラウドと単語ペアネットワークを実行した。両者で共通の分析における設定において、追加のストップワードとしてを「うどん」、「ですが」、「という」を適用した。その理由としては、「うどん」を対象としているので、どちらの地域においても「うどん」はかならずすべてのテキストで出現することが想定されており、比較が困難なためである。また、「ですが」や「という」などの文言については、レビューを書く際に頻繁に使われる用語であるため、レビューの内容を比較したいという理由からこの分析では排除することとした。

ワードクラウドとは、最も頻繁に出現する語が中央部分に大きなフォントの文字で描画され、その周りに次に頻繁に出現する語が少し小さなフォントの文字で描画され、と言ったように頻出語を中央部分に目立つように配置する視覚化の方法である。

図1にワードクラウドのうち、東京都内（左）と香川県内（右）を示す。この結果から、共通点として、「店」、「食」、「麺」、「円」、「天」（天ぷら）、「注文」がともに頻出している。一方で、東京都内では特定のメニュー：「カレー」や、「味」についての説明が比較的多かった。香川県内では、提供される食材であるうどん自身の「出汁」、「コシ」、場所の名前「香川」、「讃岐」、「高松」が多かった。

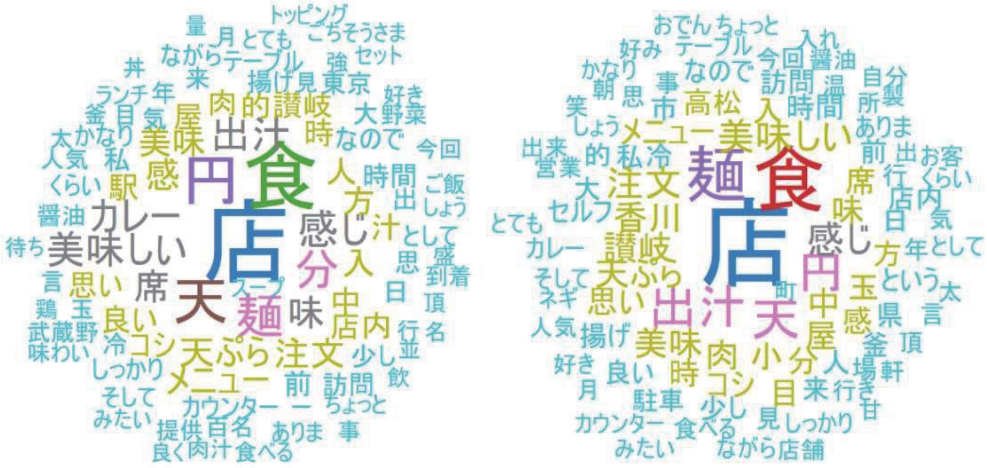

図1　ワードクラウドの結果（左：東京都内の描画結果、右：香川県内の描画結果）

次に、1レビュー内での単語のペアの共起を算出し、単語ペアネットワーク図に描画した。単語ペアネットワーク図とは、単語自体の出現頻度だけでなく、同じテキスト内（今回の場合であれば、同一レビュー内）において共起する確率も合わせて算出し、その値に応じてグラフにおいて結合する視覚化の方法である。

図２に単語ペアネットワーク図のうち､東京都内（左）と香川県内（右）を示す。これより、共通していることとしては､「店」や「食」はほかの言葉と頻繁に共起している。一方で東京都内では、比較的「注文」の方法や、うどんの金額である「円」がほかの言葉とよく共起しており、香川県内では､「麺」がほかの言葉と共起していた。

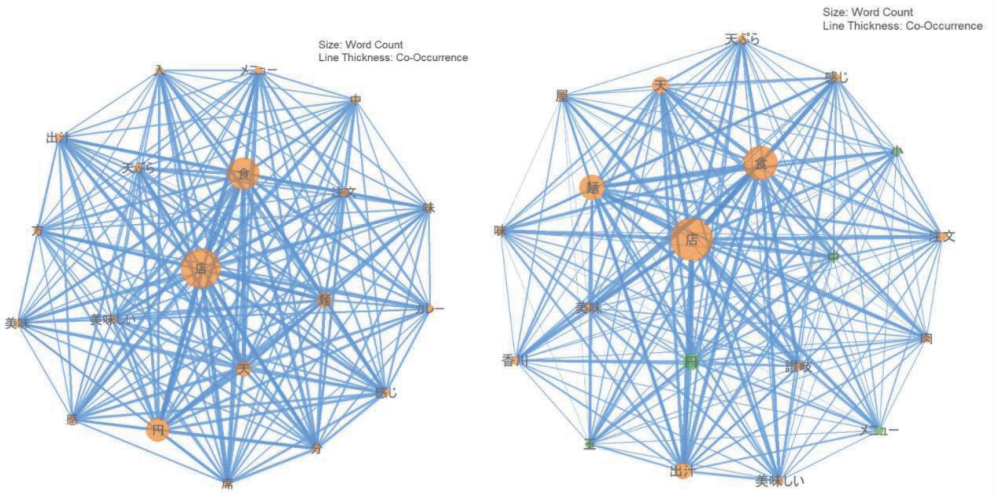

図２　単語ペアネットワーク図の結果（左：東京都内の描画結果、右：香川県内の描画結果）

最後に、トピックモデルを用いてレビューをクラスターに分類し、その頻出後の比較を行った。トピックモデルとは、潜在的ディリクレ配分法を用いた分析の一種で、文書（ここではレビュー）が複数の潜在的なトピックから確率的に生成されると仮定したモデルである。この分析では、トピック数を３として分析を進めた。図３は、トピックモデルの分析結果のうち、トピックごとの頻出語を東京都内と香川県内で比較した図である。

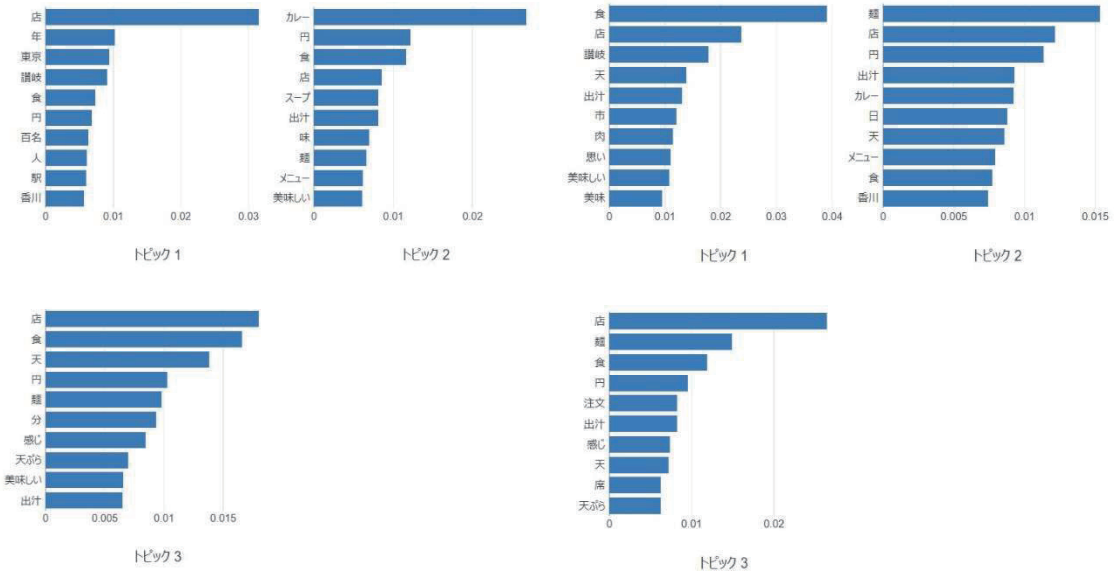

図３　トピックモデルの結果：トピックごとの頻出後（左：東京都内、右：香川県内）

図 3 からは、東京都内の 3 つのトピックについては、店舗の環境を説明しているトピック 1、カレーうどんに関連するトピック 2、そして提供されるうどん自体のことを説明しているトピック 3 と分類されたが、一方で香川県内については、きれいなトピックに分かれているとは言えないことが判明した。

4. まとめと今後の課題

本報告ではプログラミング言語 Python を用いて、東京都内と香川県内の「うどん」を提供するレストランの口コミ（レビュー）データを収集し、そのデータについて統計アプリケーション Exploratory を用いて基礎集計を実施し、さらに、ワードクラウド、単語ペアネットワーク、トピックモデル等により分析を実施した。

ワードクラウドの結果から、東京都内では、味や店の環境についての言及および特定のメニューについて多く説明されており、香川県内では、提供されるうどん自身の内容や、うどん屋の場所を表す説明が多くなされていることが判明した。また、同一口コミ内の同時出現の言葉の違いを比較したところ、東京都内ではうどんの注文の方法や金額が頻出していた一方で、香川県内では麺に関する情報が説明されていた。

トピックモデルでは、東京都内の口コミについては、環境のこと、うどんのこと、そして特定のメニューのことと 3 つのトピックに分類できたが、香川県内についてはうまく分類することはできなかった。

今後の展望としては、同様のデータ収集を全国で行い、グルメレビューの口コミにおいて、特定の麺類でどのように情報提供されるのか、その分析と分類を実施することが可能である。また、今回の分析では、収集する口コミの数を限定したが、収集と分析の時間が許せれば、さらなるデータ収集数を拡大して分析する予定である。

参考文献

[1] Kakaku.com, Inc. (2022)、食べログ・グルメ・レストラン予約サイト、
URL：https://tabelog.com/ （参照日：2022 年 9 月 21 日）

[2] python.jp (2022)、プログラミング言語 Python 総合情報サイト、
URL：https://www.python.jp/（参照日：2022 年 9 月 21 日）

[3] Exploratory, Inc. (2022)、Exploratory、
URL: https://ja.exploratory.io/（参照日：2022 年 9 月 21 日）

[4] Crummy (2022)、Beautiful Soup、
URL：https://www.crummy.com/software/BeautifulSoup/ （参照日：2022 年 9 月 21 日）