Japan Physical Therapy Society for Policies and Administration　p.36~42

◎シリーズ

テキストマイニング技術を理学療法分野で活用するための基礎知識

Fundamentals to introduce text-mining methods into the field of physiotherapy

松下光範*，山西良典

Mitsunori Matsushita*，Ryosuke Yamanishi

要　旨

テキストマイニングは、収集されたテキストデータに統計的アプローチや機械学習アプローチを適用することでそのテキストに潜むパターンや関係性を見つけ出す手法の総称である。近年ではテキストマイニングを行うためのツールやライブラリが広く公開されるようになっており、分野外の人であっても簡単に試せるようになっている。しかしその一方で、自分が調べたいと考えていた内容を知る上で適切な手法が選べていなかったり、そのツールの特性を理解せずに使ってしまい誤った解釈をしてしまったりする懸念がある。そうならないためには、テキストマイニングにより何ができるかということだけでなく、その制約や限界についても理解しておくことが肝要である。本稿ではこうした立場から、理学療法分野で研究を行う人を想定し、基本的なテキストマイニングの手法や理学療法分野の研究での利用例を紹介する。

【キーワード】テキストマイニング、TF-IDF、共起頻度、情報抽出、クラスタリング

1.　はじめに

コンピュータが我々の日常的なツールになり、これまで手書きしていたテキストが電子的に産出・蓄積されるようになってきている。コンピュータでテキストを扱えるようになったことで、悪筆・くせ字に悩まされずに文章を読めるようになっただけでなく、蓄積された大量のテキストの中からある単語(キーワード)を含む文書を探し出すことができるようになった。情報学分野ではこうした大量のテキストの中から特徴や傾向を見出す方法論としてテキストマイニング[1]が注目されている。テキストマイニングは、収集されたテキストデータに統計的アプローチや機械学習アプローチを適用することでそのテキストに潜むパターンや関係性を見つけ出す手法であり、トピックの意味的分類やテキストの感情分析など様々な目的の下での活用が進められている。マイニング(mining)は採掘を意味する単語であり、1990 年代に大量のデータの中から特徴的な傾向やデータ間の相関を見つけ出す手法としてデータマイニングが提案されてからよく聞く用語となった。テキストマイニングはその名の通りテキストデータを対象としたデータマイニングであ

[1] 近年ではより広い概念を含むものとしてテキストアナリティクスと呼ばれることも多い。

関西大学総合情報学部　Faculty of Informatics, Kansai University

公開日：2024 年 3 月 29 日

【*責任著者】松下光範　関西大学総合情報学部　〒 569-1095　高槻市霊仙寺町 2-1-1

E-mail：m_mat@kansai-u.ac.jp

り、一つ一つのテキストの中には明示されてはいないけれど、大量のテキストを統計的に処理することでテキスト集合の中に潜在的に示唆されている価値ある事象を見つけ出すことがその主眼になる。

近年、テキストマイニングを行うためのプログラムやライブラリが広く公開されるようになっており、情報学分野以外の人でも学べるような入門書や解説書も数多く刊行されている(e.g., 1)、2))。テキストマイニングに関する詳細なアルゴリズムの説明や具体的な実装方法はそちらに詳しいので参照されたい。この解説ではテキストマイニングの考え方やいくつかの代表的な手法について基本的な考え方を紹介し、理学療法分野でどのような活用が可能か、その際にはどのような準備や注意が必要かを整理することで、テキストマイニングの理学療法分野への応用可能性について述べる。テキストマイニングは便利なツールが誰でも使えるようになったことで、分野外の人であっても簡単に試せるようになった反面、自分が調べたいと考えていた内容を知る上で適切な手法が選べていなかったり、そのツールの特性を理解せずに使ってしまい誤った解釈をしてしまったりする懸念がある。そうならないためには、テキストマイニングにより何ができるかということだけでなく、その制約や限界についても理解しておくことが肝要である。本稿ではこうした立場から、初学者を想定したテキストマイニングの概要を紹介する。なお、テキストマイニングの技術は多岐にわたり、近年では画像処理技術やAI技術の進化が著しく、先端の研究では画像とテキストを組み合わせた分析なども試みられているが、ここでは理学療法分野の人が利用することを想定し、情報学分野の専門家と連携せずにできる範囲として、テキストのみを対象とした基本的なものを紹介する。

2. テキストマイニングを始める前に

テキストマイニングはテキストをコンピュータによって解釈する手法であるため、テキストが電子的に利用可能になっているデータを処理する。以下では、分析対象となるテキストについて述べる。

2-1 テキストデータを準備する

理学療法分野で電子的に利用可能なテキストとしては、カルテや実施計画書などの診療記録、業務改善やアンケート、インシデントレポート、教科書、実習レポートなどが挙げられる。テキストマイニングの手法は多岐にわたり、「文章から何を見出したいのか」によって適切な手法は異なる。例えば、大量の改善要望アンケートを、書かれた内容の種類(例：スタッフへの要望、食堂のメニューに関する要望、待ち時間に関する苦情、など)に応じて仕分けるには、トピックごとにテキストを分類する手法が適しているであろうし、典型的・特徴的な事例を見つけるためには、各テキストを計量して特徴づける手法が適しているだろう。また、診療記録から、症状と治療方針の関係を明らかにしたいのであれば、文章内の単語あるいは文間を構造化し可視化する手法が適しているかもしれない。いずれにしても深い洞察を得られるような分析を行うためには、「どのようなデータを集め、何を明らかにしたいのか」という分析の目的を定めることが重要で、その文章(データ)から明らかにすべきコト(目的)を定め、その目的に適した手法を採用する必要がある。このとき、分析対象とするデータが目的を達成するために適しているかを予め考察しておくべきであることは言うまでもない。

数値を対象としたデータマイニング同様、テキストマイニングにおいても、一般に分析対象となるデータ、すなわちテキストの量は多ければ多いほどよく、データ数が少なければ、誤った結論や不確かな結論を導き出す懸念がある。ただし、「どの程度のテキスト量があればよいか」は悩ましい問題であり、対象とするタスクや適用する手法に依存する。目的や手法によっては少ないデータであってもできることはある。例えば、分類などを行うための機械学習用のデータとして使う場合には、文内の単語を類義語や別表記に置き換えたり、言い回しを変えたりするような水増し法 (Data Augmentation Method)[3] により、見かけのデータ数を増やすことができる。しかし、どのような不満が多いか、を知る目的でテキストマイニングを行うような場合では、こうした水増し法は利用できない。

2-2 テキストデータの前処理を行う

テキストを処理する際には前処理が必要である。テキストマイニングで扱うのは自然言語で記述された「テキ

スト」であり、その多くは人によって記述されているため、数値データを扱う以上に気をつけるべき点は多い。特に日本語の口語では主語が欠落しがちであるなどの特徴がある。正確な情報が欠損したテキストを対象として分析したとしても、本来目的としている情報が得られないおそれもある。

テキストマイニングの技術的背景は自然言語処理技術である。自然言語とは、人間が日常的に扱う言語情報のことを指し、コンピュータが扱う言語に対して人間にとって「自然な言語」ということである。この自然言語をコンピュータによって処理することを自然言語処理と呼ぶ。自然言語処理といっても様々で、テキストを扱う場合には、

- 形態素解析：テキストを意味の最小単位（＝形態素）に分割し、各形態素に品詞情報を振り分ける
- 構文解析：文法規則に基づいて形態素間の関係性（係り受けや句構造）を判断し、木構造と呼ばれる階層構造に変換する
- 意味解析：代名詞の指示対象の特定や語義の曖昧性解消など、テキストの意味を判断しアノテーションを付与するなどしてコンピュータで処理可能にする
- 文脈処理：前後のテキストやそのテキストが生成された状況など、対象となるテキストの周辺情報を考慮し、その文章の意図の推定や欠落語の補完などを行う。

といった処理が行われる。基本的なテキストマイニングでは、このうち、事前処理として形態素解析まで行っている場合がほとんどである。形態素解析を自動で行うテキストマイニングツールもあるが、多くの場合は形態素解析の処理がブラックボックス化しているツールも多い。例えば動詞の場合、その原形に変化させて用いるか（例：歩いて→歩く）や、どの品詞を利用するか（例：名詞だけを用いる、動詞だけを用いる、自立語を用いる）は、目的によって異なる。どのような意味の単語が頻出するかを見たい場合には自立語に変換したうえで「歩く」「歩きながら」「歩こう」といった様々な「歩く」を一元化して統計処理する方が適切である。一方で、例えば「話しながら歩こうとした」のように、特定の動作が他の動作と関連しているのかを確認したいといった場合には、注目する単語が連用形として記述されていること自体に意味が見いだせる場合もある。こうした処理は結果に大きく影響を及ぼすため、分析者自身がその特性を理解し、制御できるようにしておくことが望ましい。

形態素解析を行うツール（形態素解析器）は、語彙辞書を参照して文を形態素に分割する。近年の形態素解析器の分割精度は非常に高いものになってきてはいるものの、専門用語や新規語の語分割は一般的に苦手な場合が多い。このような場合には、予め分野に特化した語彙辞書を用意して反映させたり、形態素解析結果を加工して適切な分割単位に修正したりする必要がある。また、データが話し言葉（口語文）である場合には、誤った形態素解析を行う可能性があるため、事前に書き言葉にする、という前処理を行うことで、意図しない誤分割を低減させる事ができる[2]。

電子化されたテキストを計算機で処理する場合、半角カタカナや記号、特殊文字、絵文字の含まれたテキストはしばしばエラーを発生させる原因となる。このような場合には、半角カタカナを全角に変更したり特殊文字や絵文字を削除・置換したりするなどして処理することが求められる。

3. テキストマイニングではどのようなことができるのか

テキストマイニングはその目的によって様々な分析手法が用いられる。これらの手法はいずれも一長一短であり、それぞれの手法の特徴を十分に理解した上で適切な分析手法を選定することが求められる。

テキストを対象とした初歩的な分析としては以下のようなことが行われている。

3−1 語の出現頻度を見る

文章中に出現する語彙の傾向は、その文章の内容を理解する手がかりになる。語彙の傾向として最も素朴なものは、単語の出現頻度（Term Frequency; TF）である。これは、ある文章を特徴づけるのに「どのような単語が多く含まれているか」を指標とする考え方である。

例えば、**表 1** は、Wikipedia（英語版）の「physical

[2] もちろん，どのような口調の発話であるかに関心がある場合には，この処理を行ってはいけない。

therapy」の本文テキストを対象に TF を算出した場合の頻出語（上位 20 語）である。理学療法分野に関係する語彙が多く含まれており、TF がこの文章の特徴を表現していることが理解してもらえるだろう。 TF は、測る単位としてひとつのテキストを対象にすればそのテキストの特徴が、複数のテキストのまとまりを対象にすれば、そのテキスト集合の特徴が推察できる。

また、出現頻度の多寡をわかりやすく理解するため

表 1　英語版 Wikipedia の「physical therapy」項に出現する名詞の出現頻度上位 20 語

単語	出現頻度
physical	160
therapy	77
therapist	66
patient	35
care	26
program	26
health	25
physiotherapist	23
education	22
rehabilitation	21
United	20
degree	20
practice	18
treatment	18
States	17
Therapy	16
include	15
condition	15
professional	12
Physiotherapy	12

に ,TF を出現頻度に応じて語彙の大きさを変化させて表示するワードクラウドがしばしば用いられる（**図 1**）。

ただし、複数のテキスト集合を比較する場合、単語の出現頻度を用いることには注意する必要がある。長いテキストは単語の絶対数が多くなるため、出現率の低い単語の出現量が、短いテキストにおける最頻出の語より多くなることさえある。このようなケースでは、単語の出現頻度をそのまま用いて、テキスト同士を比較することはできない。長さの異なるテキスト同士を比較する場合は、テキストの長さで出現頻度を正規化する、出現頻度ではなく出現割合を用いて測る、といった手法がより適切である。

語彙数を計量する場合には、表記ゆれや略記が問題になることがある。例えば、BRS と Brunnstrom Recovery Stage は同一概念であるが、表記が異なるためコンピュータ上では別の語として扱われる。異なる単語として計量したあとで、後処理（post process）にてこれらの語を纏める処理（名寄せ処理）を行えば語の表記ゆれの問題は解決するが、大量の単語を扱う場合、名寄せ処理に多くの手間がかかるため現実的ではない。事前に類語や別表記の辞書を作成しておき、それを参照して名寄せするプログラムを用意すればその処理は軽減される。

3−2　語彙の偏りを見る

類似した分野のテキスト同士を比較する場合、共通する語彙が多く、それらのテキストを特徴づける単語が必ずしも上位の頻出語とならない場合がある。例えば医療分野に関連するテキストの多くは「患者」や「治療」といっ

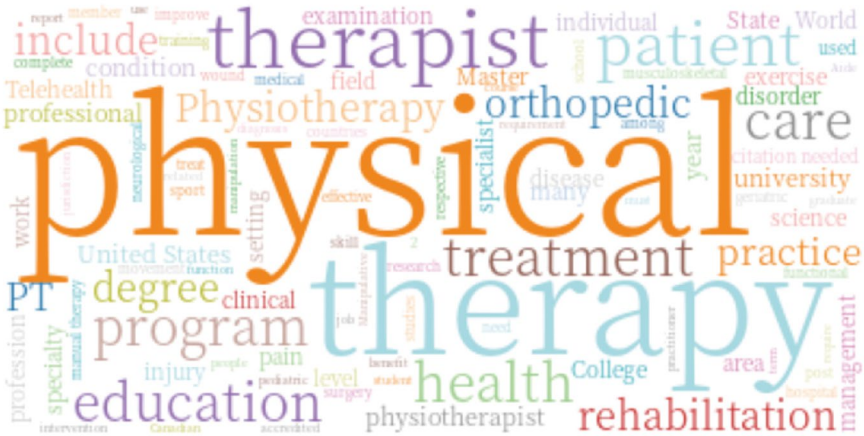

図 1　英語版 Wikipedia の「physical therapy」項に出現する名詞のワードクラウド

た語彙を含んでいるため、これらの単語を文章の特徴として利用しても、他のテキストとの差を知ることが難しい。

類似した分野における複数のテキスト（あるいはテキスト集合）を比較する場合、出現回数が比較的少なくても、他のテキストにはあまり表れず、ある特定のテキストで多く出現するのであれば、そのテキストを特徴づける手がかりとして扱うことができる。

このようなテキストごとの単語の偏りを知る方法として、逆文書頻度（Inverse Document Frequency; IDF）という考え方がある。

IDF は、ある語彙が、全体の文書のなかでいくつの文書で出現したかを逆数として数値化する方法である。多くの文書に出現すれば分母が大きくなるためその値は 0 に近づき、その語が現れた文書数が少ないほど大きくなる、という特徴を持つ。

一般的には TF と IDF をかけあわせた TF−IDF が用いられる。例えば、文献 4）では理学療法教育に 1 枚ポートフォリオ評価を導入することの検証として、学習前後や得点群などの基準で群分けし、TF−IDF を用いてポートフォリオに記述された内容の特徴比較を行っている。TF−IDF は単純な方法ではあるがテキスト集合の特徴を端的に掴むことが可能であるため、広く利用されている手法である。

3−3　単語の共起を見る

テキストに含まれる単語は互いに独立ではなく、特定の単語同士が同じテキストで一緒に用いられている（共起する）ことがしばしばある。この共起が多いほど、それらの単語に意味的な関係性があると考えられる。単語の共起はそのテキストのドメインによって異なるため、共起頻度を分析することでそのドメインにおける単語間の意味的関係を理解することができる。例えば、医療分野のテキストであれば「治療」という単語と「診断」という単語が共起する頻度は高いが、建築分野のテキストであれば、「診断」と「治療」の共起頻度よりも、「診断」と「故障」の共起頻度が高くなる。共起頻度の高さを意味的な関係性の強さと捉えることで、テキストから知識を抽出することが試みられている。

共起頻度を測る場合、ネットワーク図を用いて共起関係を可視化することがしばしば行われている。例えば、文献 5）では ,1982 年から 2017 年にかけて医学中央雑誌に掲載された緩和ケアに関連する論文（3,342 件）を対象に、文献 6）では、ISI Web of Science データベースに含まれる 2000 年から 2018 年の理学療法に関する論文（29,280 件）を対象に、それぞれ共起頻度を計量してネットワーク図で表現し、その対象分野の中心となるキーワードを調査している。ネットワークで可視化することは、特定のパターンや関連性を視覚的に理解しやすくするだけでなく、次数中心性（ある語がどれだけ多くの語と繋がっているかの指標）や媒介中心性（ある語がノード同士の最短経路に存在するかの指標）などの、ネットワーク分析の結果を読み取りやすくする効果もある。このように、ネットワーク図を用いた可視化は直観的な理解容易性からよく用いられる手法ではあるが、解釈を行う場合には注意が必要である。例えば上記の文献 6）では、「Physiotherapy（理学療法）」や「Rehabilitation（リハビリテーション）」、「Treatment（処置）」といった理学療法分野では自明の語が上位に現れており、分析の有用性に疑問が残る。また、共起頻度はあくまでも「二つの単語が文章中で一緒に使われることが多い」ことを示しているにすぎず、単語間の正確な関係をそのまま表すものではない。単純な共起頻度の計量では否定表現が反映されていなかったり、語の分節化処理が不十分なために意図しない関係が可視化されたりすることもある。共起する単語同士がポジティブな関係であるか、それともネガティブな関係かは、文脈を考慮して判断する必要があるため、解釈する場合には注意しなければならない。

3−4　情報を構造化して抽出する

情報抽出とはテキストの中から、意味的な構造を持つ情報を取り出すことである。人が記述するテキストでは、しばしばある一定のルールに則って記述されている。このようなテキストの構造を手がかりとして利用することで、テキストに含まれる知識を機械的に獲得することができる（**図 2** 参照）。例えば医療分野のテキストから、症状や診断の結果、治療方針、リスク因子などをひとまとまりの情報として抽出することで、臨床推論の過程や結論を検証したり、知識の構造的な蓄積を図ったりすることができるようになる。例えば文献 7）では、情報抽出手

法を活用して理学療法士が持つ知識を外在化し、知識共有の基盤にすることを試みている。この研究では、動作分析のテキストを対象とし、そのなかから、

- タイミング：歩行周期の表現や空間的な状況
- 身体部位：特定の身体部位を表す表現
- 状態(項目)：状態表現のうち、属性を表すもの
- 状態(value)：状態表現のうち、とり得る値を表すもの

の4項組で情報を抽出する。この項目に照らすと「左立脚初期において、下肢は股関節外転位で接地する」というセンテンスからは**表2**のような情報が構造的に抽出される。この論文では、こうした動作分析を構成する知識の最小単位をPBPU(Problem-based Physiotherapy Unit)と定義し、その関係性をネットワーク表現で可視化することでどのような論理の下で臨床推論が行われているかを示し、その推論過程を他者と共有することを狙っている。

情報を構造化して抽出するうえでは、どのような組で情報を抽出するかが解く課題によって異なるため、事前にどのような構造で情報を抽出すべきか明らかにしておく必要がある。このとき、人が書いたテキストを対象に情報の抽出を試みる場合､主語や目的語が省略されたり、暗黙の含意によって省略されたりすることがある。このような点を考慮して、事前準備の段階で欠落語を補完するなどして対象とするテキストの質を担保することが重要になる。

表2　動作分析のテキストから抽出・構造化された情報の例

タイミング	身体部位	状態	状態−value
左立脚初期	下肢	接地	股関節外転位

3−5　テキストを分類する

テキストの類似度に基づいてグループ(クラスター)に分類する手法を総じてクラスタリングと呼ぶ。例えば、アンケートなど人が記述したテキスト集合を内容の類似性に基づいてクラスター化したり、特定のテキストに類似したテキストを判別したりする場合に用いるものであり、様々な手法が提案されている。**図3**は、理学療法士のカルテの記述を対象としたクラスタリングの一例である。クラスタリングの技術を利用することで、テキストの特徴を手がかりとした機械的な情報分類が可能になる。テキスト集合を分類するだけでなく、抽出した単語やキーワード、語の共起関係をクラスタリングするような用い方もしばしば見られる。

クラスタリング手法を大別すると、事前にカテゴリーを定めずに自動で分類する教師なしクラスタリングと、あらかじめカテゴリーを定めてそこに該当するかどうかを判断させる教師ありクラスタリングの二種類に分けられる。教師なしクラスタリングでは、幾つのクラスターに分割するかを事前にコンピュータに与え、コンピュータがテキストの特徴を考慮してそのクラス数に分割する手法が多く使われてきた[3]。この場合、得られたクラスターがどのような意味を持つかは、分割された結果から人間が判断する必要がある。反対に、教師ありクラスタ

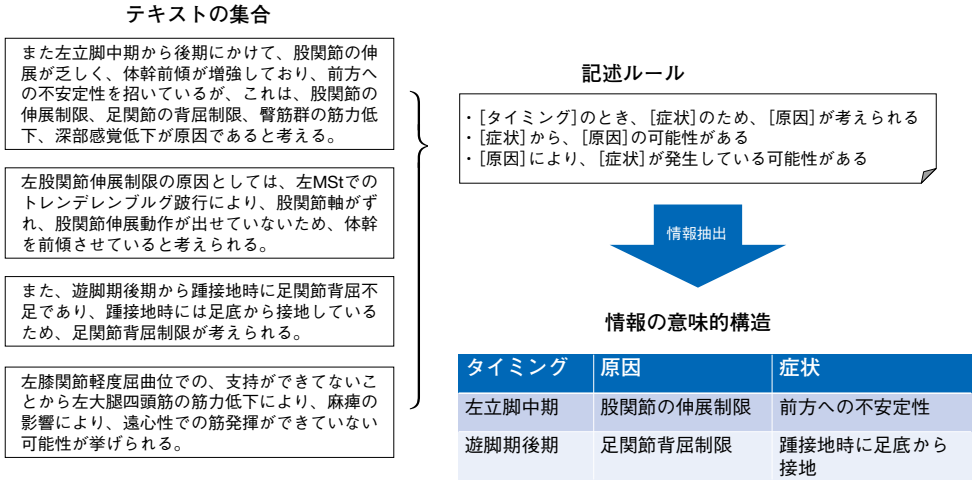

図2　情報抽出の一例

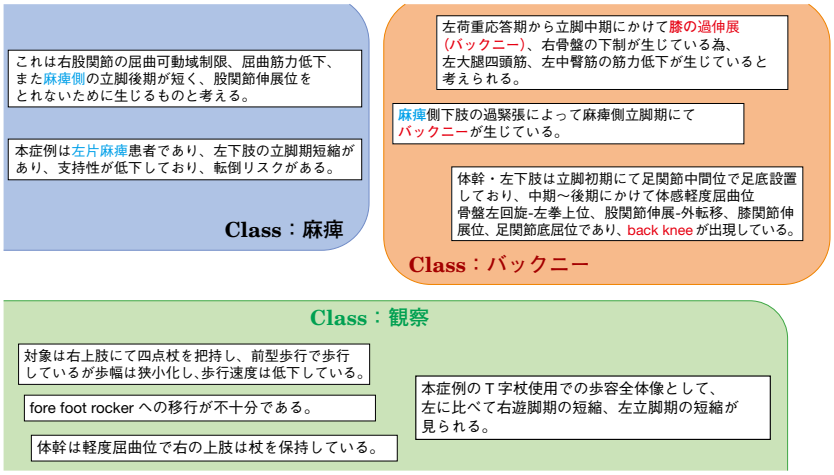

図 3　クラスタリングの一例

リングでは、事前にクラスラベル(≒カテゴリー)が付与されたテキストを学習データとして用意しておき、それらのデータを分割できるようにテキストから得られる特徴量とクラスラベルの関係性を学習した分類器を用いて、未知のテキストがどのカテゴリーに属するかを判断する。なお、教師ありクラスタリングの手法によってはクラスターの分類を階層的に行うこともできる。例えば、文献 8)では、臨床実習の指導者が指導に際して困ったことについて質問紙調査で収集し、得られたテキストを階層的クラスター分析によりクラスター化している。

4.　おわりに

本稿では、情報学分野で広く利用されているテキストマイニングについて、理学療法分野での活用を意識して紹介した。近年では、情報学の研究成果として作成されたツールやシステムがネット上に公開され、簡単に利用できるようになってきている。その特徴を理解しうまく活用することで、分野横断的に研究を加速させることが期待できる。冒頭でも述べたように、テキストマイニングを学ぶための環境は近年よく整備されてきている。文献 1)は手を動かすことを含めた初学者向けの入門書、文献 2)はテキストマイニング技術の全体を概観した解説書であり、いずれも技術の利用方法だけでなくその背景をも学べる良書である。分野横断的な研究推進が加速している昨今、こうした情報学分野の技術にも目を向けていただき、理学療法分野の発展につなげていただくことを願う。

文献

1)　榊剛史, 石野亜耶, 小早川健, 坂地泰紀, 嶋田和孝, 吉田光男：Pythonではじめるテキストアナリティクス入門, 講談社(2022).
2)　那須川哲哉, 吉田一星, 宅間大介, 鈴木祥子, 村岡雅康, 小比田涼介：テキトマイニングの基礎技術と応用, 岩波書店(2020).
3)　Feng SY, Gangal V, Wei J, Chandar S, Vosoughi S, Mitamura T and Hovy E : A Survey of Data Augmentation Approaches for NLP. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 968–988(2021).
4)　加藤研太郎：理学療法士の養成教育にOPPAを導入する意義について, 理学療法教育, Vol. 1, No. 1, pp.1_38–1_46(2022).
5)　李慧瑛, 下高原理恵, 緒方重光：テキストマイニングによる緩和ケア論文表題の可視化, 医療と社会, 2018, Vol.28, No.2, pp.259–275(2018).
6)　Emami Z, Joulahi L, Okhovatian F and Shahrokhifarid R : Mapping the Scientific Outputs in the Field of Physiotherapy: A Co-Word Analysis, Journal of Clinical Physiotherapy Research, Vol.5, No.3, e18(2020). DOI: 10.22037/jcpr.v5i3.32393
7)　宮本誠人, 松下光範, 高岡良行, 堀寛史：理学療法初学者の支援を目的とした動作分析テキストの構造の可視化, 2022年度人工知能学会全国大会(第36回)論文集, 1I1-OS-6-04(2022). DOI: 10.11517/pjsai.JSAI2022.0_1I1OS604
8)　二宮省悟, 吉村修, 楠元正順, 吉田勇一, 田崎秀一郎：臨床実習指導者のアンケート調査におけるテキストマイニングを用いた客観的分析, 理学療法科学, Vol.34, No.2, pp.205–209(2019).

[3] 教師なしクラスタリングで，コンピュータが適切なクラスター数を自動で判別する手法もある。