テキストマイニングでできること

―理学療法分野で活用するためのコツと注意点―

関西大学 総合情報学部
松下 光範

はじめに：このセミナーの目的

- 理学療法分野で記録や報告に使っているテキストをどのように活用したらいいだろう？
- 最近，理学療法分野でも「データサイエンス」とか「質的研究」という言葉を聞くようになってきたが，どんな利点があるんだろう？
- 情報系のような他の分野って，理学療法とどんな関わりがあるんだろう？

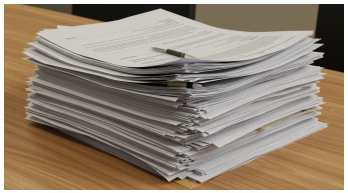
・・・とお思いの理学療法士の皆様に**思考の補助線**を提供します

テキストデータを統計的に扱う？

テキストデータ

理学療法に関わるテキストデータはいろいろ存在する

- 経過記録・診療記録
- カンファレンス議事録
- インシデントレポート
- 退院時アンケート　　etc



これらのテキストデータ，**ちゃんと活用**できてますか？

テキストはどんなふうに活用できるだろう？

テキストデータの可能性

患者のQoL向上のために

・自由記述アンケートから満足度や生活上の困難を定量化
・リハビリ日誌の記述からモチベーション要因を抽出

理学療法士の臨床支援のために

・診療記録の分析を通じて，症状の類型や治療のベストプラクティスを把握
・回復の過程を時系列的に理解

教育・熟達支援のために

・症例検討会のレポートを分析して熟練者と初学者の知識構造を比較
・学習記録を分析して指導カリキュラムの改善点を発見

テキストデータはよりよい理学療法の実践に有益

テキストデータの特徴

数値データでは得られない特徴がある

・大量の情報がテキスト（自然言語）で記録されている
・数値データや画像では理解しづらい「意味」が記述されている

こうしたテキストをコンピュータで処理するのが一般的になりつつある
テキストマイニングはその一つのアプローチ

テキストマイニングとは

・テキストを統計的に分析し、意味や傾向を取り出す技術
  ・ 文章から要点を要約
  ・ 重要な語句の抽出も可能
・テキストデータを対象としたデータマイニング

近年では**テキストアナリティクス**と呼ぶことが多い

なぜテキストマイニングをするべきなのか？

テキストは記述者の「頭の中」を理解する手がかりの宝庫

・数値の分析は「意味」が扱えない
・テキストを解析することで，「意味」や「理由」を把握したり，「知識」を洗い出すことができる

  ・診療記録：なぜそのような判断をしたか，注目はどこか
  ・アンケート：満足度を「2 (やや低い) 」と回答したのはなぜか

定性分析で十分じゃない？

定性分析でできること

- 少量のテキストから洞察を得る
- 文脈やニュアンス，背景を読み取る
- 仮説や理論を発想するきっかけになる

定性分析の限界

- 大規模な記録を全て読み解くことは時間・労力的に不可能
- 定量的な議論が難しいため，分析者の主観が入りやすい

テキストマイニングでできること（1）

定性分析では十分でなかった点を解決できる

1. 大規模なデータを処理できる

- 人手では処理しきれない量（数千〜数百万件）のテキストから傾向を把握できる.
- 「量は質を凌駕する」

2. 客観的かつ定量的に傾向を把握できる

- 頻出語や共起関係、感情極性など，分析に役立つ指標を数値化できる.
- 印象ではなく，データに基づくエビデンスを提示できる.

3. 時間的な変化や関係性の分析ができる

- 語の使われ方や感情などの時間的な推移が理解可能になる.
- 関係性をネットワーク表現を用いて視覚的に把握できる.

テキストマイニングでできること（2）

テキストマイニングのさらなる利点

4. 再現性の高い分析が可能になる

- 人手の主観に依存せず，同じ手法で再度分析すれば同じ結果が得られる
- 再現性は学術的知見を得るために必要な観点

5. 隠れたパターンを発見する手がかりになる

- トピックモデルやクラスタリングにより，人間の直感では気づきにくいテーマや特徴を抽出できる

テキストマイニングは難しい？

難しくない，とは言わない

- **なぜ処理が難しい？**
  - テキストは非構造データなので，構造化処理が必須
  - 文脈や前提知識（常識）を理解していないと，意味を正しく理解することができない
- **「簡単に処理できるツール」の功と罪**
  - わからなくてもそれなりの「答え」が得られてしまう

テキストマイニングを体験してみよう

テキストマイニングをやってみよう

Webツールを使ってチャレンジ

- 簡単なツールはWebツールとして公開されている
  - ワードクラウド
    - ファンブライド社
    - UserLocal社
  - 形態素解析器（茶まめ）
  - TF-IDF（松下研究室 作成）

※ 一般的に，情報系ではPythonなどのプログラミング言語を用いて
自前で作成するので，現状のWebツールは上記の初級ツールが中心です

体験してみよう

- 松下研究室のHP
  - http://mtstlab.org/ のリンク集にある「理学療法支援プロジェクト」に入る
  - https://www.japanpt.or.jp/about_pt/therapy/

WordCloudを作ってみよう

ファンブライト社のツール

https://lab.fanbright.jp/wordcloud/

WordCloudは何ができる

特徴

- テキストで頻出する語に気付けるようになる
  - 仮説生成に向く

使い道

- OK: ポスター発表で扱っているテキストの概要を伝える
- OK: 仮説を生成する
- NG: 論文でエビデンスとして使う（この図から〜であることが窺える）

懸念点1：Bag of Wordsの副作用

文法を無視した分析法（よく利用される）

- 構文情報を「捨てる」方法
  - 同じ文単位に含まれている単語群を一つのグループとして分析
    - 通常は自立語（名詞，形容詞，形容動詞，など）

例文

抗体反応がないのが問題だ

形態素解析

言葉を「形態素」単位に区切る

「抗体反応」を1語として処理したい．．．．

抗体反応がないのが問題だ

形態素解析結

表層形	語彙素読み	品詞
抗体	コウタイ	名詞-普通名詞-一般
反応	ハンノウ	名詞-普通名詞-サ変可
が	ガ	助詞-格助詞
ない	ナイ	形容詞-非自立可能
の	ノ	助詞-準体助詞
が	ガ	助詞-格助詞
問題	モンダイ	名詞-普通名詞-一般
だ	ダ	助動詞

代表的な語の扱い方：Bag of Words

文法を無視した分析法（よく利用される）

- 構文情報を「捨てる」方法
  - 同じ文単位に含まれている単語群を一つのグループとして分析
    - 通常は自立語（名詞，形容詞，形容動詞，など）

名詞 名詞 形容詞 名詞

抗体/反応/が/ない/の/が/問題/だ

どちらも

Bag

ない 抗体 問題 反応

Bag of Wordsの問題

文法を無視したことによる弊害

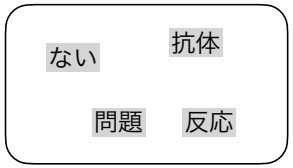
- かかり受けや否定の情報は無視

A: 抗体反応がないのが問題だ

B: 抗体反応は問題がない

どちらも

Bag



懸念点2：複数のテキストの比較が難しい

頻出することはどのくらいの意味を持つのか？

- 文章の長さがテキストによって異なる
  - 長い文章の場合，出現する語彙の回数は高まる傾向にある
  - 異なる文長のテキストを比較する場合，語数を比較することに意味はない
- 同じジャンルの文章だと，特定の語彙が両方に頻出し，本当の違いが分かりにくくなる

どうすればいい？

Word Cloudで二つのページを比較する

理学療法士協会：理学療法とは

https://www.japanpt.or.jp/about_pt/therapy/

理学療法士協会：理学療法士とは

https://www.japanpt.or.jp/about_pt/therapist/

二つのページの共通性/相違性を見るには？

共通する語彙

二つのページの共通性/相違性を見るには？

「理学療法とは」ページの固有語彙

「理学療法士とは」ページの固有語彙

二つのページの共通性/相違性を見るには？

意図した情報を見るには，適切な前処理が不可欠

「理学療法士とは」ページの固有語彙

適切な前処理のために

- 解決策1
  - 形態素解析を自分で行い，必要な語彙に絞る
  - 単語境界の間違いも修正できる（例：理学/療法　→　理学療法）
- 解決策2
  - TF-IDFを使ってみる
    - 単語の出現頻度だけじゃなく，その語の「偏り」を反映する方法
  -

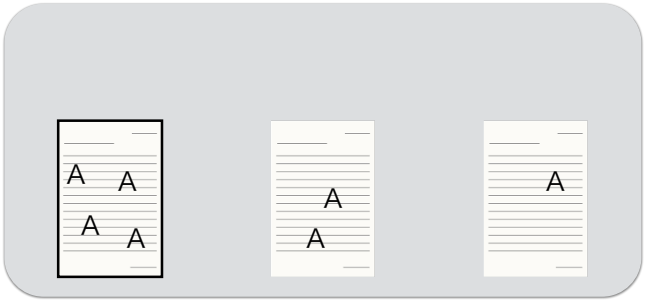
形態素解析

- 茶まめ：Webツールの無料形態素解析用ページ
  - https://chamame.ninjal.ac.jp/index.html
  - 国立国語研究所が提供
  - 自然言語処理分野で標準的な形態素解析器であるMeCab（めかぶ）が裏で動いている
  - 5MBまでのテキストを解析可能

茶まめの使い方

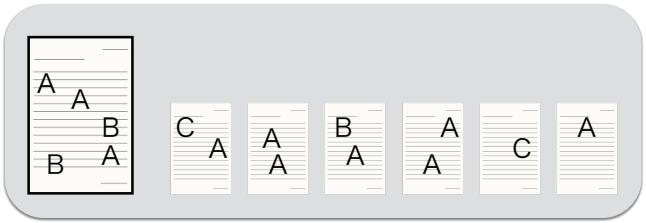
出現頻度（Term Frequency)

- 検索キーワードが多く含まれているページは関連度が高いページ



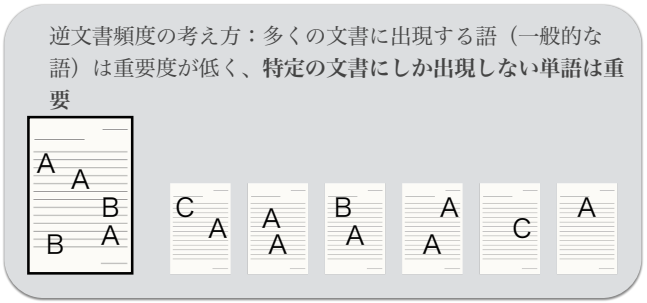
逆文書頻度（inverse document frequency)

- 逆文書頻度の考え方：多くの文書に出現する語（一般的な語）は重要度が低く、**特定の文書にしか出現しない単語は重要**



TF-IDF

- 出現頻度（TF）と逆文書頻度（IDF） を掛けあわせた指標



TF-IDFツールを使ってみよう

Excel形式で複数のテキストデータを比較する

テキストマイニング活用のさらなる活用可能性について

テキストマイニング技術は日々発展している

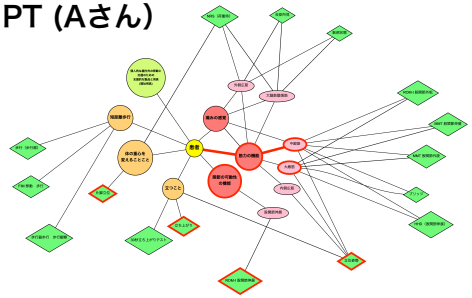
- ワードクラウドやTF-IDFは20年以上前の研究成果
- 機械学習（特に深層学習）の発達により，文脈を考慮したテキスト処理（クラスタリングや推定）が可能に
  - 最近のトレンド：単語を文字列ではなく，ベクトルとして表現する

`King - Man + Woman → Queen`

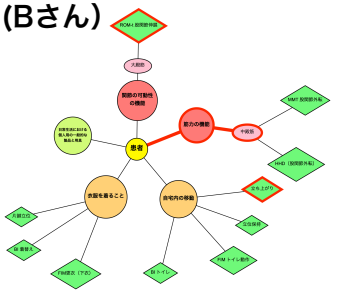
臨床推論知識の可視化

模擬患者に対する問題点の列挙

PT (Aさん)



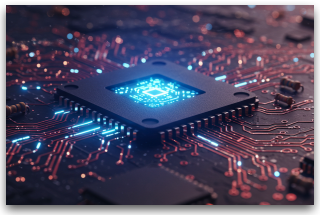
PT (Bさん)



同じ患者でも，人によって検査項目や判断が異なる
（診療記録に基づく知識獲得は要注意）

テキストマイニングしなくても良くない？

「AIに任せたらいいんじゃない」論



生成AIに分析を丸投げするのはありか？

- （私見ですが）現状では限りなく否定的

理由

- 生成AIは「極めて物知りな素人」であり，実体験を伴っていない
- 「わかりません」が言えないので，必ず答えを見つけようとする

→ハルシネーション（幻覚）の問題

臨床現場の理学療法士が生成AI使ってみた

- 統合と解釈を書く，実施計画書を書くなど，根本的にベースデータが外に落ちていないものはかなり間違う。
- 感覚的データの扱いが言語化されていないため腑に落ちない表現が多い
- 生理学的な順序を教えてもくれますし，引用文献も教えてくれるのですが，どこか飛躍した考え（知らない単語が多い）に感じる
- 「論文から」というプロンプトにも関わらず引用を覗くとHPだったりするので、鵜呑みにすると危険だなぁと思う場面が少なからずある印象

テキストマイニングのために「すべきこと」

（1）テキストデータを蓄積する

- テキスト処理が有効であるための前提
  - 良好な**質**のテキストが**十分な量**存在していること

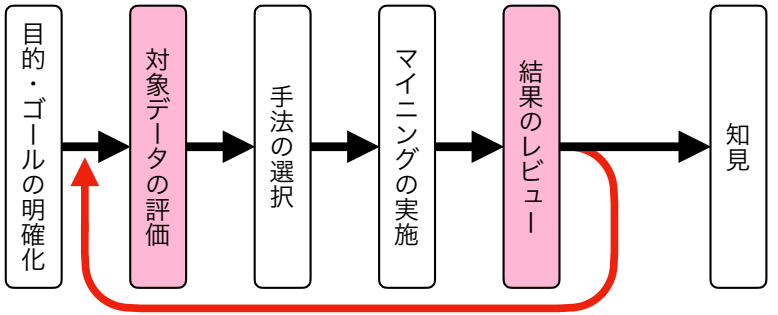
    → 頭の中にあるものを適切に外在化（Garbage in，Garbage out）
  - あくまでも「形式的な処理」であることを理解して使うこと

    → 意思決定を委ねるのは論外（あくまで**エビデンスの一つ**として）

    → 自分の能力にレバレッジを効かせるという意識が大切

テキストマイニングのために「すべきこと」

- テキストマイニングの利用モデル



テキストマイニングのために「すべきこと」

（2）分析の「単位」を考える

与える文章の量や単位によって結果は大きく変わる

- 文章（特に長文）はどう扱うべきか
  - 知りたいことによって，適切な処理単位は異なる
    - 発言の意味役割を知りたい → 一文
    - 文章に含まれるトピックを知りたい→パラグラフ

おわりに

- テキストマイニングは，未来の理学療法分野の発展に貢献するはず.
- 今回紹介したツールは初歩の初歩
- 理学療法分野でのテキストマイニングの地位向上を目指して，今後も色々と簡単に使えるツールを作っていこうと思っています.
- 分析の仕方で迷ったらぜひ相談してください.
  - 理学療法 x 情報学 は，伸び代の大きな原野