テキスト分析のプロセス

PANGAEA

企業組織活性化コンサルティングカンパニー

パンゲア合同会社
加藤 昌生
2023.7.7

CONTENTS

- データ収集から知見を得るまでのプロセスChatGPTとは？
- テキストアナリシスとは企業活動とChatGPT
- テキストアナリシスのプロセス事例１：Jalan AI Chat
- 新たな言葉を創る

PANGAEA

データ収集から知見を得るまでのプロセス

Twitterデータ

サイエンス論文データ

その他のデータ

データ収集

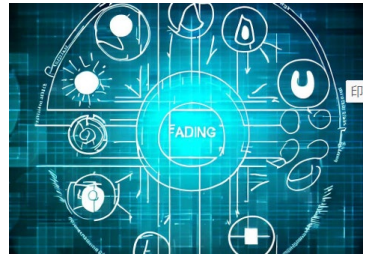
Tweetデータとサイエンス論文データ収集の**アプリは開発済み**

データ基盤

開発予定
現状は簡易DB

テキストアナリシス

分析作業中



分析結果から知見

テキストアナリシスの結果から**ヒント、インスパイア、知見を得る**

図はOpenAI社のDALLEを使用して作成しました

PANGAEA

テキストアナリシスとは

- テキストアナリシスは、テキストデータを解析して情報を抽出するプロセスです。主な目的は、大量のテキスト情報から意味やパターンを把握し、有用な洞察を得ることです。

- テキストアナリシスでは、自然言語処理（NLP）や機械学習の技術が使われます。具体的な手法には、単語の出現頻度や共起関係を調べる「単語のカウント」といった基本的なものから、感情分析やトピックモデリングなどの高度な解析手法まで幅広く存在します。

テキストアナリシスのプロセス

テキストデータ

前処理
テキストクリーニング
形態素解析

分析
分析アルゴリズム
分析手法
の適用

オリジナル辞書
の作成

オリジナル辞書

複合語を抽出
それをベースに
作成

テキストアナリシスに使える分析手法リスト

- 頻出語の抽出
- KWICコンコーダンス
- コロケーション集計
- 共起ネットワーク
- 対応（コレスポンデンス）分析
- 多次元尺度構成法
- トピックモデリング
- クラスター分析
- 主成分分析

いくつかの
手法を用い、
分析を行う

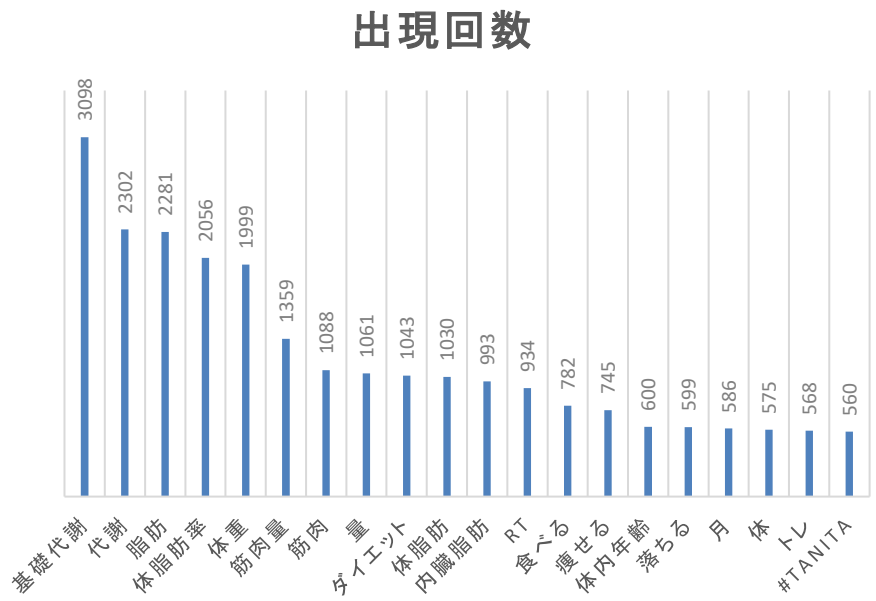
分析結果

図はOpenAI社のDALLEを使用して作成しました

PANGAEA

頻出語の抽出

- 頻出語の抽出とは、Tweetなどの文字データから出現回数の多い単語やフレーズを見つけることです。頻出語の抽出は、テキスト分析や自然言語処理の一部として使われることがあります。

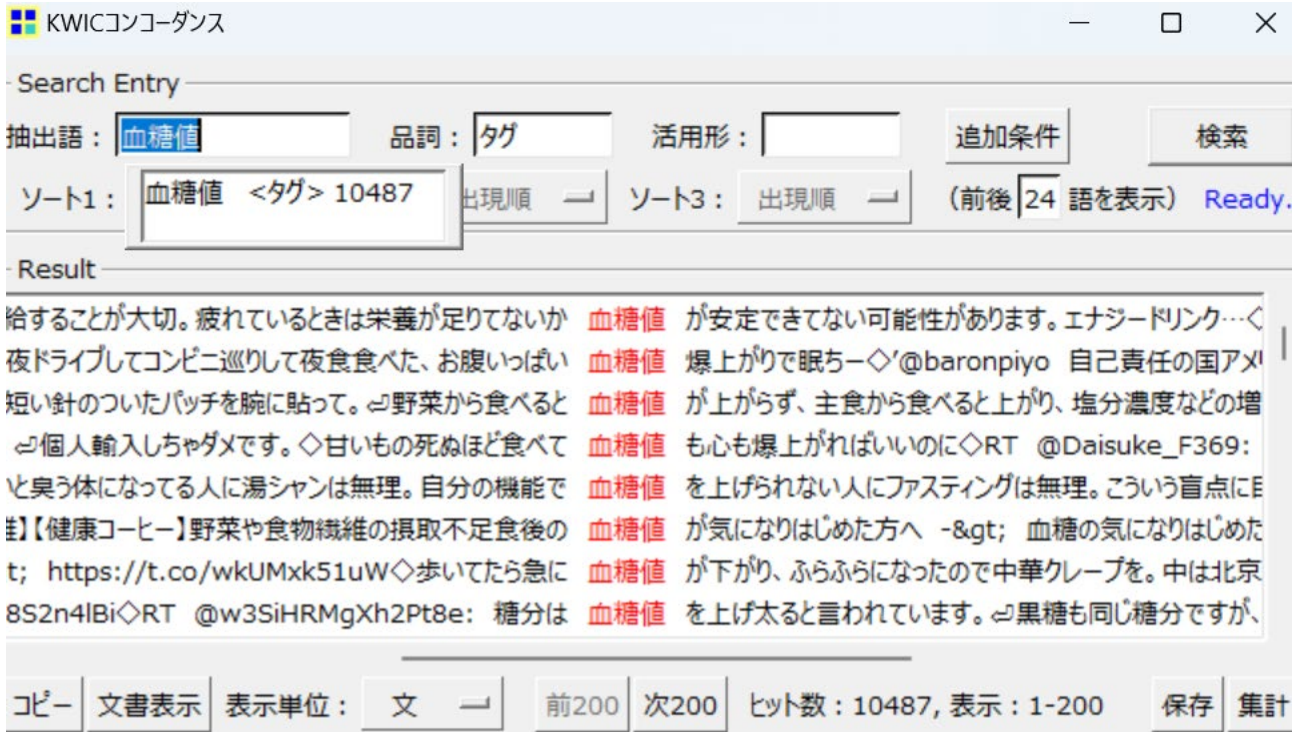


抽出語	出現回数
基礎代謝	3098
代謝	2302
脂肪	2281
体脂肪率	2056
体重	1999
筋肉量	1359
筋肉	1088
量	1061
ダイエット	1043
体脂肪	1030
内臓脂肪	993
RT	934
食べる	782
痩せる	745
体内年齢	600
落ちる	599
月	586
体	575
トレ	568
#TANITA	560

PANGAEA

KWICコンコーダンス

- KWICコンコーダンスとは、キーワードとその文脈を表示する検索ツールです。
- KWICコンコーダンスを使うと、テキストからキーワードとその周辺の単語やフレーズを見つけることができます。
- KWICコンコーダンスは、テキスト分析や自然言語処理の一部として使われることがあります[2]。



PANGAEA

コロケーション集計

- コロケーション集計とは、キーワードとその関連の強い単語やフレーズを統計的に調べることです。
- コロケーションとは、「二つ以上の単語の慣用的なつながり」のことです。
- コロケーション分析を使うと、テキストからキーワードとその周辺に多く出現する単語やフレーズを見つけることができます。

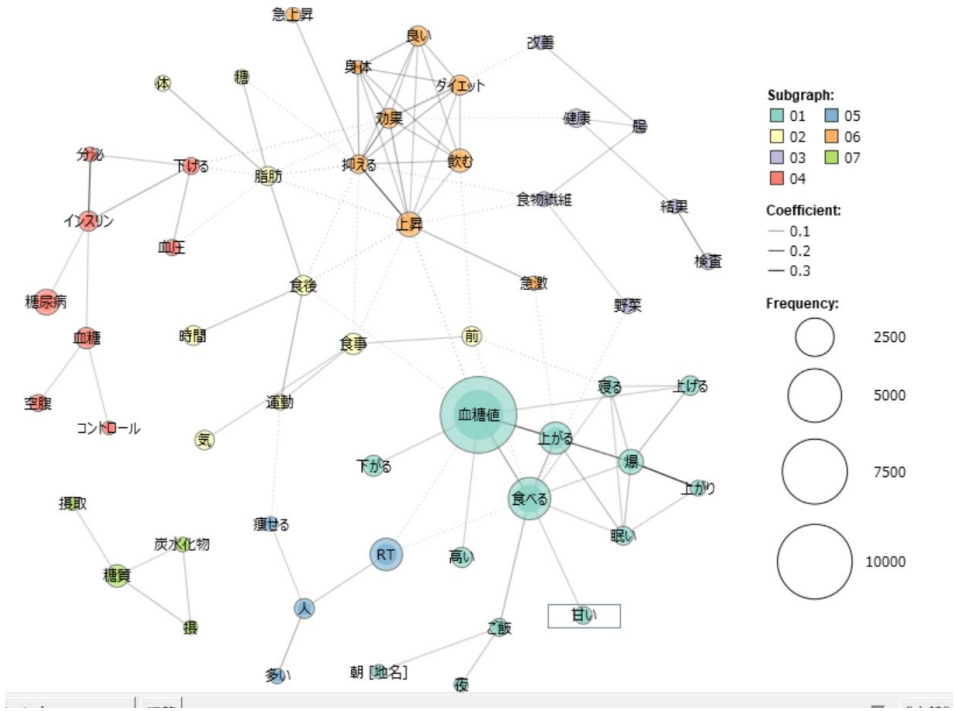
N	抽出語	品詞	合計	左合計	右合計	左5	左4	左3	左2	左1	右1	右2	右3	右4	右5	スコア
1	上がる	動詞	1561	50	1511	13	5	11	19	2	342	875	142	125	27	882.500
2	爆	未知語	925	37	888	8	11	10	8	0	734	120	10	9	15	814.267
3	上昇	サ変名詞	925	29	896	9	5	4	10	1	202	582	12	66	34	530.683
4	下がる	動詞	537	28	509	4	4	8	9	3	119	291	48	38	13	304.567
5	食べる	動詞	839	756	83	112	153	195	290	6	0	4	12	27	40	297.400
6	ない	否定助動詞	853	362	491	86	85	86	94	11	0	82	156	117	136	274.567
7	上げる	動詞	529	14	515	6	2	3	2	1	55	369	50	35	6	270.817

PANGAEA

共起ネットワーク

- 共起ネットワークとは、文章中に出現する単語と単語が同時に出現する関係性を図にしたものです。
- 共起ネットワークを使うと、文章の構造的な特徴を直感的に理解するのに役立ちます

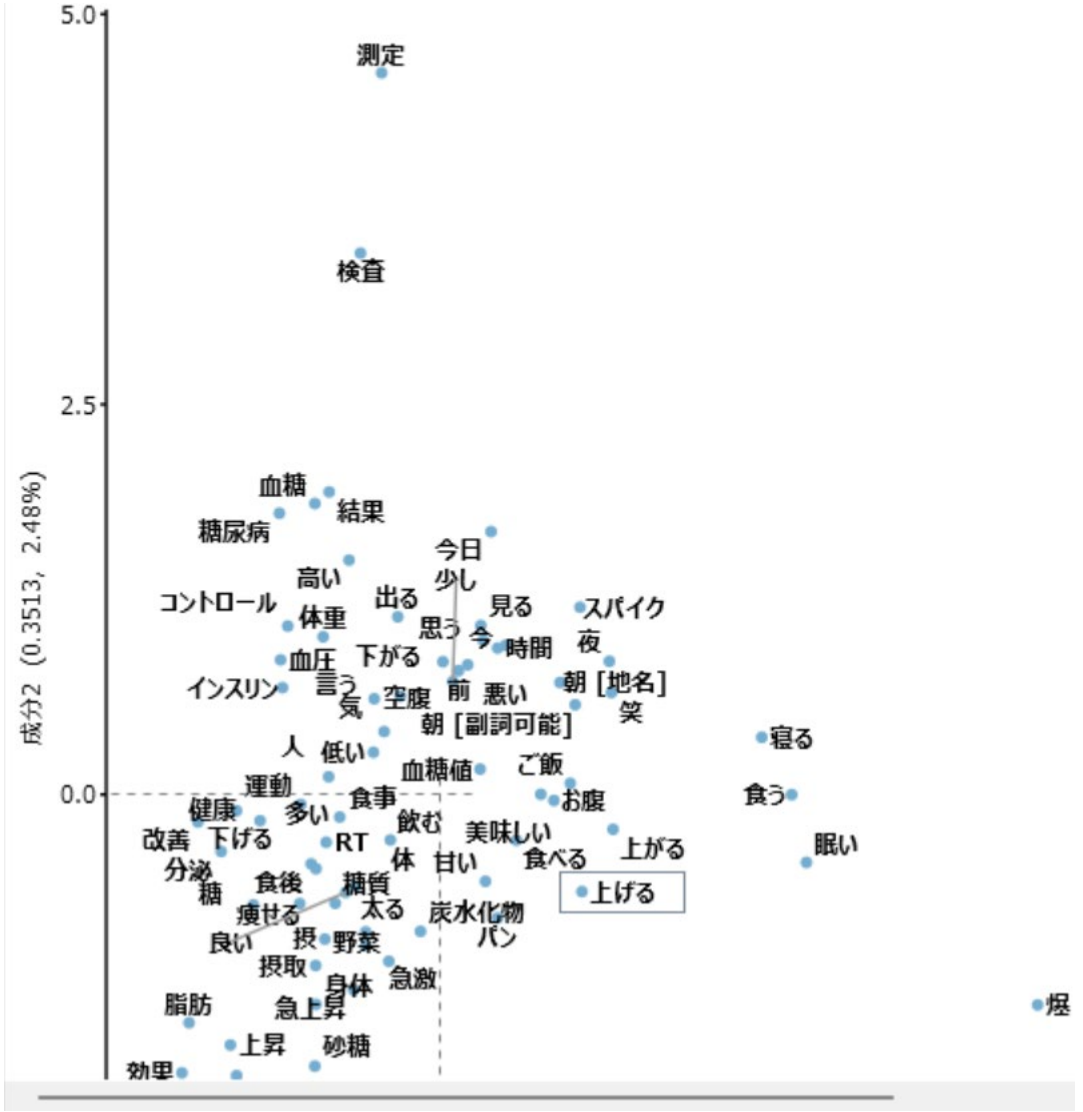
- 活用事例
  - テキストデータの要約: 大量のテキストデータをまとめる際に、共起ネットワークを作成して、出現回数が多い単語や関連性が強い単語を抽出することで、テキストデータの概要や傾向を把握できる[1]。
  - 文化や思想の分析: 文学作品や歴史資料などに共起ネットワークを適用することで、作品や時代の背景にある文化や思想を探求できる。例えば、夏目漱石の小説『こころ』に共起ネットワークを作成すると、物語の登場人物やテーマが浮かび上がる[2]。
  - 言語の構造的特徴の定量化: 共起ネットワークには、密度や中心性などの数値指標がある。これらを用いることで、言語の構造的特徴を定量化できる。例えば、日本語と英語の共起ネットワークを比較すると、日本語は英語よりも密度が高く中心性が低いことが分かる[3]。



PANGAEA

対応（コレスポンデンス）分析

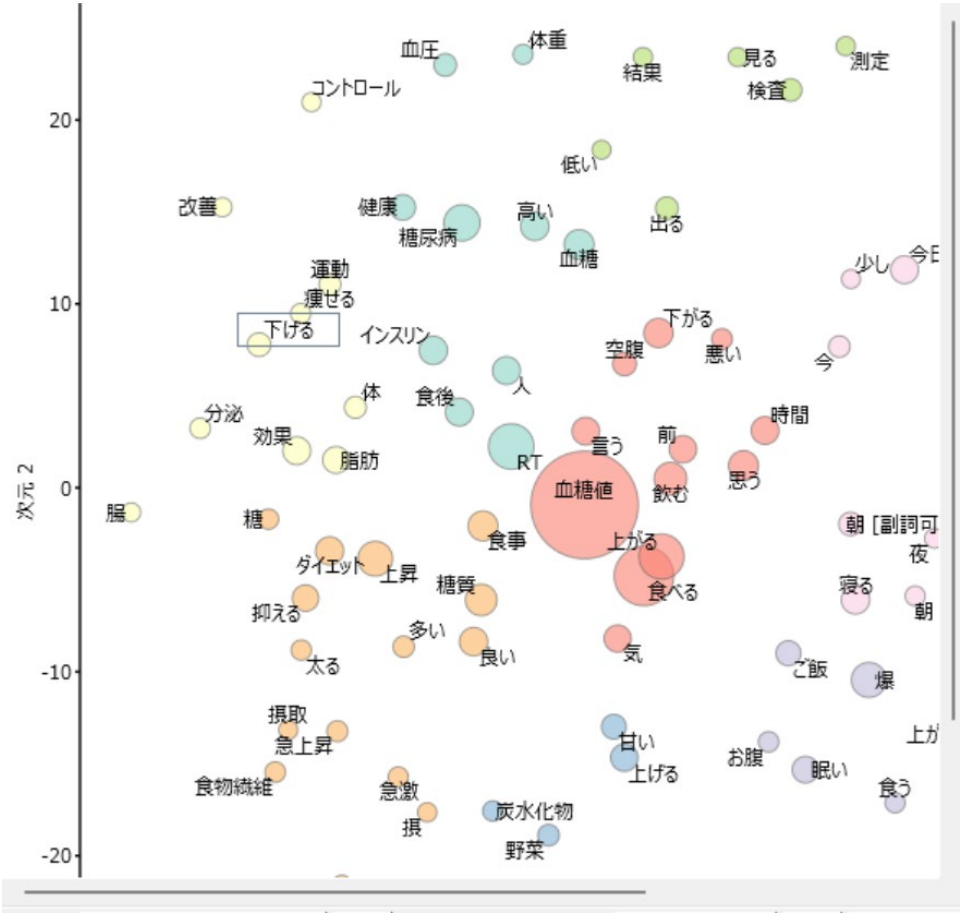
- 対応分析とは、クロス集計表など、行と列からなるデータの特徴を図示し、項目間の関係を視覚的に把握する方法
- 対応分析は、偏りの小さい項目は原点付近に、偏りの大きい項目は原点から遠くに配置
- 互いに関連の強い項目どうしは、原点からみて同一方向に布置される性質があります[1]。別名でコレスポンデンス分析とも呼ばれます[2]。
- 活用事例
  - ブランドイメージや商品イメージのポジショニングマップを作成
  - アンケート調査の回答者属性や購買行動と回答内容の関係を分析
  - 文書や単語の類似性や特徴を可視化



PANGAEA

多次元尺度構成法

- 多次元尺度構成法は、類似度を元に対象の関係を視覚的に分かりやすい形に変換する分析手法
- 多次元尺度構成法を用いることで、グループや変数の関係性を距離で示すことができ、データの解釈が容易に
- 活用事例
  - 競合他社製品のポジションや関係性をポジショニングマップで把握し、自社製品の開発方針を検討
  - 顧客行動の各地域別の特性をポジショニングマップで把握し、類似性の高い地域でグルーピングして施策を検討
  - 自社製品をポジショニングマップで比較し、売れる商品と売れない商品の差を検討



PANGAEA

トピックモデリング

- トピックモデリングとは、Tweetのような文字データから抽象的な「トピック」を発見する統計的なモデル
- トピックモデリングは、文字列の隠れた意味構造を探索
- トピックモデリングの手法が
  - LSI（Latent Semantic Indexing）：単語の意味的な関係を考慮して、文章ベクトルの次元を圧縮する手法
  - LDA（Latent Dirichlet Allocation）：文章中の潜在的なトピックを推定し、文章分類や文章ベクトルの次元削減に用いる手法
  - BERTopic：BERTという深層学習モデルをベースにして、単語の文脈をより正確に理解し、データから最適なトピック数を自動的に決定する手法
- 活用事例
  - 消費者セグメンテーション: 購買ログデータやアンケートデータなどをトピックモデルに適用することで、消費者の購入意向やニーズを反映したトピックを抽出し、消費者をトピックに基づいて分類することができる[1]。
  - 文書分類や検索: 文書に含まれるトピックを特徴量として用いることで、文書の分類や検索を効率的に行うことができる。例えば、ニュース記事や口コミなどにトピックモデルを適用して、関連性の高い文書を探したり、文書のカテゴリーを予測したりできる[2][3]。
  - テキスト生成: トピックモデルを逆に用いて、あるトピックに関連する単語や文書を生成することもできる。例えば、ある商品のレビューを生成する際に、トピックモデルからその商品に関連する単語やフレーズを選択して組み合わせることで、自然なテキストを生成できる[4]。

クラスター分析

- クラスター分析とは、データ全体の中から似たもの同士をグループ分けする方法
- クラスター分析は、データの特性や傾向を単純化して理解しやすくするために、市場調査やマーケティングなどの分野でよく使われます
- クラスター分析は、データから似たもの同士をグループ分けすることで、様々な分野で有効に活用できる分析手法です
- 活用事例
- ラスター分析の活用事例は、以下のようなものがあります。
  - アンケートや市場調査：データから顧客やターゲットの傾向や特徴を把握し、セグメンテーションやポジショニングなどのマーケティング戦略に活用できます[1234]。
  - One to Oneマーケティング：データから顧客のニーズや志向を分析し、個別に最適な商品やサービスを提供できます[1]。
  - メルマガやDM配信：データから顧客のプロファイリングを行い、顧客の志向に合わせてメルマガやDM配信などを行うことで、効果的なコミュニケーションができます[1]。
  - 広告配信：データから顧客の興味や関心を分析し、最適な広告を配信することで、コンバージョン率を高めることができます[4]。

主成分分析

- 主成分分析とは、多くの変数を持つデータを集約して主成分という新しい変数を作成する統計的分析手法
- 主成分分析は、データの総合力や変数間の関係性を把握したり、データをグラフ化したりするのに役立つ
- 主成分分析のメリットは、データ数を少なくして調査を効率化したり、総合力に影響している項目を把握したりできることです[2]。主成分分析のデメリットは、取りこぼされる情報が出てしまったり、分析内容が分析者の判断に依存したりすることです
- 活用事例
  - マーケティング (顧客の分類): 顧客アンケートによって集めた「接客の質」「店舗の雰囲気」「過ごしやすさ」「メニューの豊富さ」「メニューの味」「価格」といった変数を主成分分析することで、総合的な顧客満足度が算出できる。また、顧客満足度に最も影響している要素も把握できるため、戦略立案に役立つ[1]
  - 作品・製品評価: 新聞や口コミサイトで見かける総合ランキングは、主成分分析が活用されていると考えられる。また、自社が販売している商品それぞれの総合評価を知ることで、顧客が求めている商品を把握できる
  - 人事評価や人員配置: 新規立ち上げ部署のマネージャーを選出したい場合や、人事評価を行う場合に、「これまでの実績」「立ち上げ分野に関する知識」「上司からの評価」「部下からの評価」「部下の実績」といった変数を主成分分析することで、総合力が高く最もふさわしい人物を選出可能だ。また、総合力を上げるために影響が大きい要素も分かるため、教育や研修にも活かせる