テキストデータの解析に活用可能な
自然言語処理手法の比較

2019 年 8 月

イノベーション推進部

志賀野　芳宏

PLANET, INC.

1．はじめに

昨今の技術の発達は、音声データのテキスト化、アンケートやお問い合わせの電子化など「大量のテキストデータ」の出現を引き起こし始めている。これらを背景として、「テキストデータ全体から、関連のある情報を探索する」タスクが、ビックデータ活用のシーンでこれまで以上に、重要な技術となってきた。

本検証では、主に 3 種類の自然言語処理手法の検証を行い、それぞれの特徴と活用イメージについて、整理を行う。

2．手法の検証

2．1．検証環境

本検証では、「From プラネット Vol.1～Vol.80」の自由回答欄に記述された、内容を各テーマと紐づけてデータベース化した。

データベース化に併せて、それぞれの文章を形態素解析で単語レベルに分割（分かち書き）した。

各手法では、分かち書きされたデータベースを使って、自然言語解析を行った。

2．2．各手法の概要

今回比較した 3 つの手法の主な特徴を表にまとめる

No.	名称	概要
1	Word2Vec	単語の意味をベクトル表現（数値化）に変換し、単語間の数値上の距離や演算、分類等を実行可能にする。
2	対応分析	二つの変数を二次元上にプロットし、関係を観察可能にする。
3	共起分析	単語間の相関関係を観察可能にする

表 1　自然言語処理手法の特徴

2．3．各手法を利用した自然言語解析

ここからは、各手法の例示を行い、手法の特徴について考察する。

（1）Word2Vec

この手法は、学習データ全体を単語（品詞）レベルで機械学習を用いて分類し、各単語間のつながりと出現箇所等を考慮して、単語をベクトル化（数値化）する。

下記に示す例では、From プラネットの柔軟剤のアンケート内の自由回答欄全体を学習した結果の上位 30 単語を平面上にプロットした。

チャート左側に「香る」「ほのか」「香水」「無臭」「香料」など「香り」に関連した単語が集待っていることがわかる。このように、「学習データ全体が持つ傾向を広く理解する」ために利用できる。

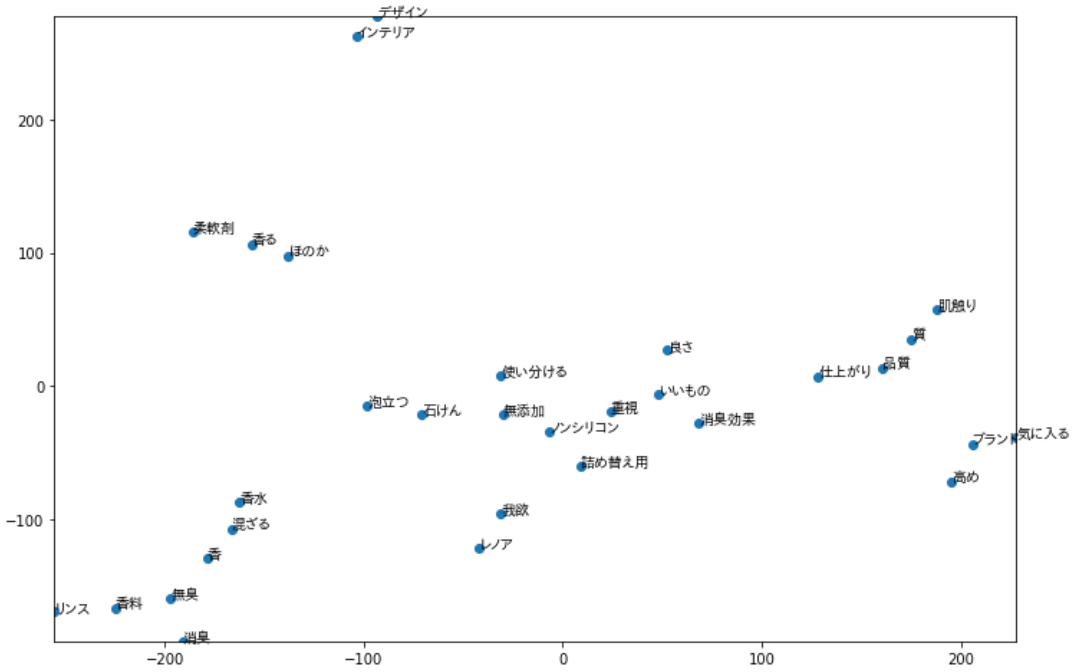

図１　Word2Vec を活用した単語の傾向分析

（２）対応分析

この手法は、カテゴリと呼ばれる変数を設定し、その変数と各単語との関係を２次元平面上にプロットする。

変数に「年代」と「ストレス解消の方法」を設定した分析結果を下記に示す。

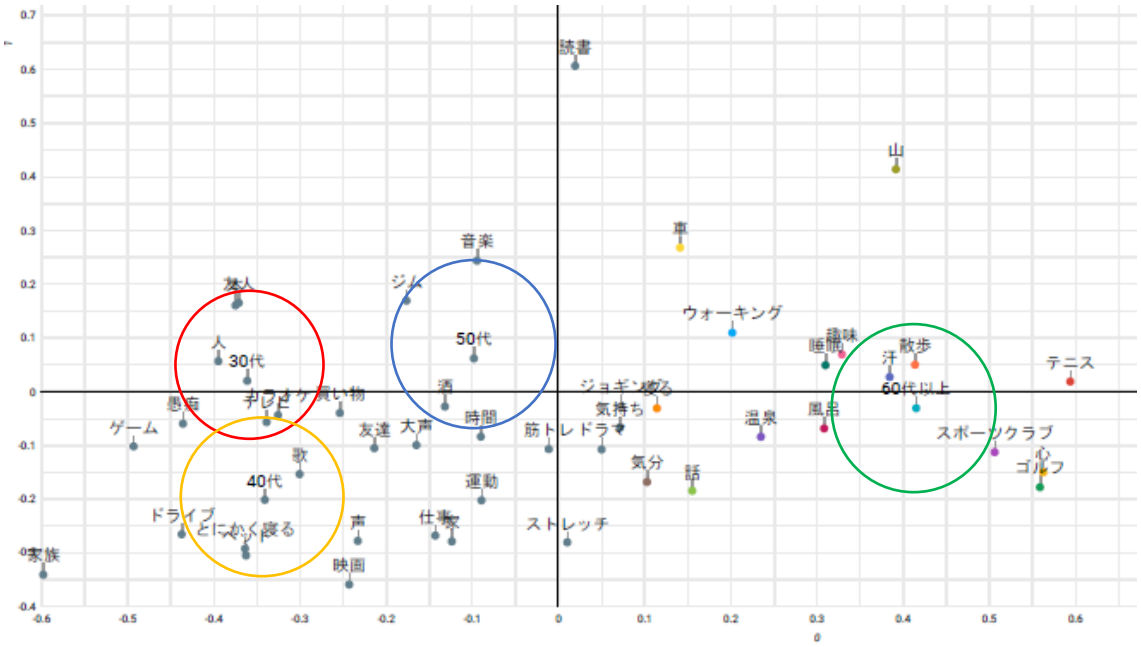

図２　対応分析を活用した年代ごとのストレス解消手段の傾向

年代ごとにストレス解消の手段が異なることが見える。下記に、表として整理する。

年代	近くにある手段に関する回答
30 代（赤）	友人、愚痴、カラオケ、テレビ、買い物
40 代（黄）	歌、とにかく寝る、ペット、ドライブ、映画
50 代（青）	音楽、ジム、酒、時間
60 代以上（緑）	散歩、汗、風呂、趣味、睡眠、スポーツクラブ

表１　分析結果から見える各年代のストレス解消手段の傾向

（３）共起分析

柔軟剤のアンケートに関する自由回答を共起分析した結果が下記である。

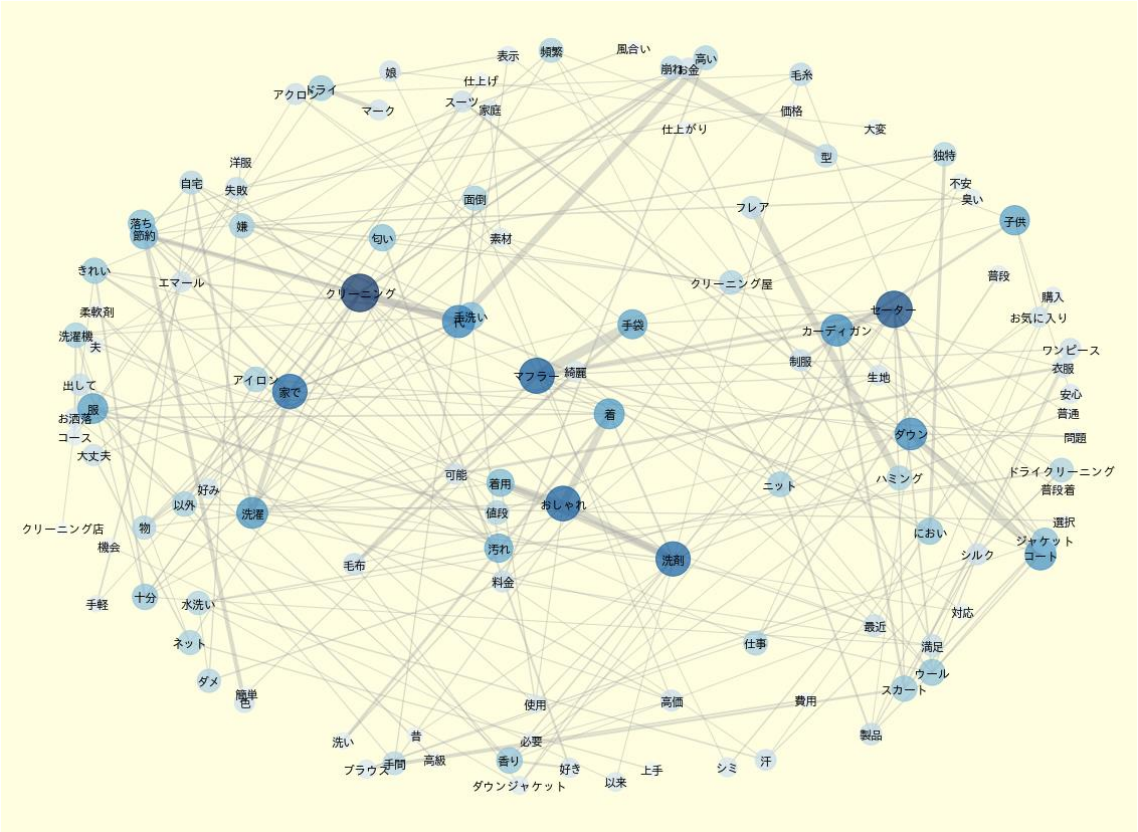

図３　柔軟剤の自由回答に関する共起分析結果

丸の大きさは出現頻度、単語間をつなぐ線は関連度を表現している。

比較的強い相関を持っているのは、「おしゃれ－洗剤」「マフラー－手袋」柔軟剤の利用に関連すると思われる単語であることがわかる。

「クリーニング」に連動して「代」「お金」といった反応が見えている点は、費用感に意識が向いていることがわかる。

出現回数は少ないが、「フレアーハミング」というメジャー商品が強い相関を持っているのも消費者の考える代表的な商品を表している点で興味深い。

3．まとめ

今回検証した手法はそれぞれ、データ抽出の手法、表現される内容ともに異なるものを調査、試行した。

自由回答形式のデータをより深く分析するためには、分析手法データが持つ特徴・傾向を理解したうえで分析と考察を行うことが重要である。

対応分析のように、定量データの解析で使われた手法でも、出現頻度を数えることで、テキスト情報の解析にも応用できる。また、今回はテキストのベクトル化を単純に表現したが、自社商品と近い単語の抽出や、特定の単語との組み合わせによって、生活者の認識を垣間見ることも可能であると考えられる。

データ解析の分野において、テキストデータはまだまだ発展途上であり、流通業界における活用も、これから拡がっていくと思われる。効率的な活用方法について、引き続き実験、考察を続けたい。

以上