Datamining rep3 level4 205713B 朝比奈太郎 2022/07/06

Level4

文書を眺めてみよう２。

・仮説

対象となるDataFrameには”love”と”others”の２カテゴリあり、このカテゴリは、前回のlevel3でtextの内容から後付けたものである。”love”には恋愛や結婚について書かれた文書、”others”にはそれ以外の文書という基準で分類した。従って、(5)のカテゴリ別の単語の出現数をカウントするという問題で、”love”というクラスにおいて、”恋愛”、”結婚”、”男性”、“デート”が上位にくると考える。また、(6)のscattertextでは、２カテゴリとも独女について書かれた文書であるため、両文書で出現回数の多い単語(Frequency)は”女性”ではないかと考える。

**(5) カテゴリ別に単語の出現回数をカウントし、積み上げ棒グラフにより描画せよ。グラフから分かることを述べよ。**

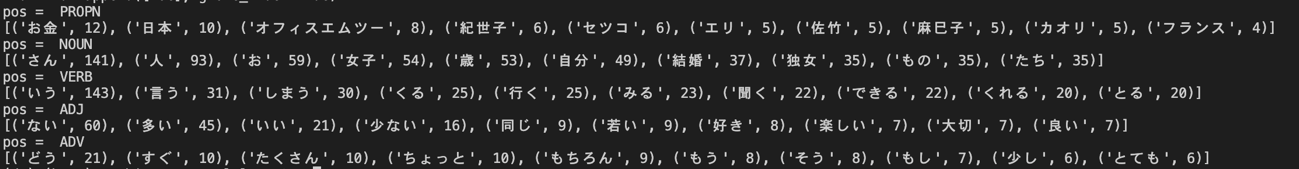
カテゴリの種類とカテゴリカル文書の数

グラフィカル ユーザー インターフェイス, テキスト, アプリケーション

自動的に生成された説明

カテゴリ数は2つで、”love”, “others”となっている。それぞれのカテゴリカルな文書の数は10文ずつである。

品詞別カウント出力結果



積み上げ棒グラフ

グラフ, 棒グラフ

自動的に生成された説明

グラフからわかること

“love”, “others”の2つの積み上げ棒グラフを見て、Y座標が0に近い単語(下に積み上がっている単語)ほど共通しているwordが多く、Y座標が400に近い単語(上に積み上がっている単語)ほど、それぞれ特有の単語が出てきていることがわかる。”love”において、仮説で高頻出単語に出てくるであろうと考えた、”恋愛”、”結婚”、”男性”、“デート”のうち”結婚”, “男性”の2単語が上位20単語の中に入っていた。また、独身女性の略語である、”独女”や”女子”が”others”と比べて多いことがわかった。

**(6)任意の2カテゴリを選び、scattertextにより単語出現分布を可視化せよ。(3)で述べたこと以外で、scattertextから分かることを述べよ。**

Scattertext出力結果



カテゴリは”love”, “others”の2カテゴリのみであるため、そのまま2カテゴリを採用した。20文書のうち、”love”にカテゴライスされた文書数、”others”にカテゴライズされた文書数、共に10文書である。グラフの真ん中を通る右肩上がりのライン状に位置する単語はcharacteristicの下位の単語が多いことから、”love”, “others”の２カテゴリにおいて同じくらいの数が出現している単語であるといえる。また、右肩下がりのライン状に位置する単語は、love軸またはother軸に近くなるにつれて、片方では出現頻度が高いが、片方では出現頻度が低い単語が分布されていることがわかる。仮説で唱えた、両文書において出現頻度が高いであろう”女性”は、グラフの真ん中を通る右肩上がりのライン状に位置しているかつ、両frequentの部分に位置していることから、両文書において”女性”は出現頻度が高いといえ、仮説が支持された。