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

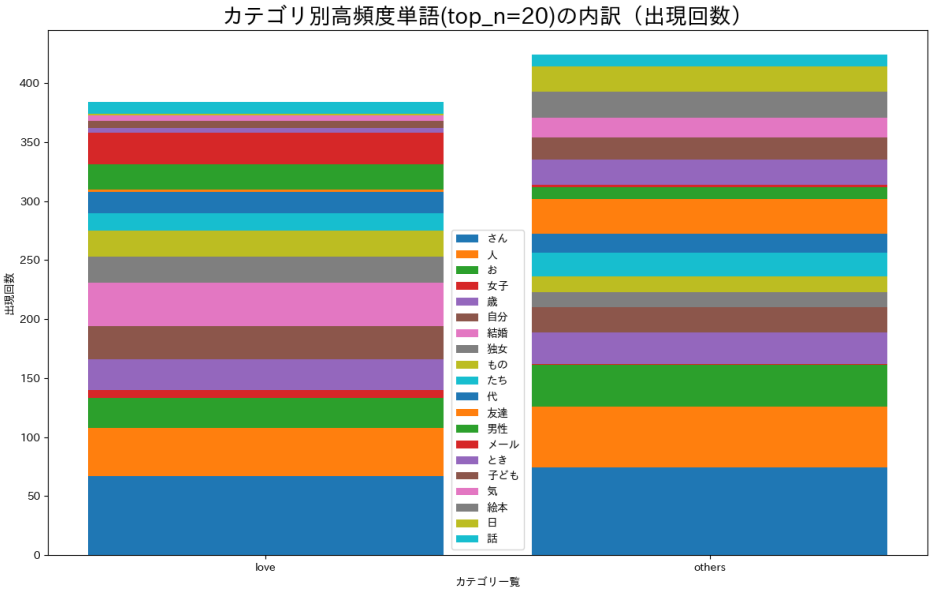
```
love      10
others    10
Name: class, dtype: int64
```

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

```
pos =  PROPN
[('お金', 12), ('日本', 10), ('オフィスエムツー', 8), ('紀世子', 6), ('セツコ', 6), ('エリ', 5), ('佐竹', 5), ('麻巳子', 5), ('カオリ', 5), ('フランス', 4)]
pos =  NOUN
[('さん', 141), ('人', 93), ('お', 59), ('女子', 54), ('歳', 53), ('自分', 49), ('結婚', 37), ('独女', 35), ('もの', 35), ('たち', 35)]
pos =  VERB
[('いう', 143), ('言う', 31), ('しまう', 30), ('くる', 25), ('行く', 25), ('みる', 23), ('聞く', 22), ('できる', 22), ('くれる', 20), ('とる', 20)]
pos =  ADJ
[('ない', 60), ('多い', 45), ('いい', 21), ('少ない', 16), ('同じ', 9), ('若い', 9), ('好き', 8), ('楽しい', 7), ('大切', 7), ('良い', 7)]
pos =  ADV
[('どう', 21), ('すぐ', 10), ('たくさん', 10), ('ちょっと', 10), ('もちろん', 9), ('もう', 8), ('そう', 8), ('もし', 7), ('少し', 6), ('とても', 6)]
```

積み上げ棒グラフ

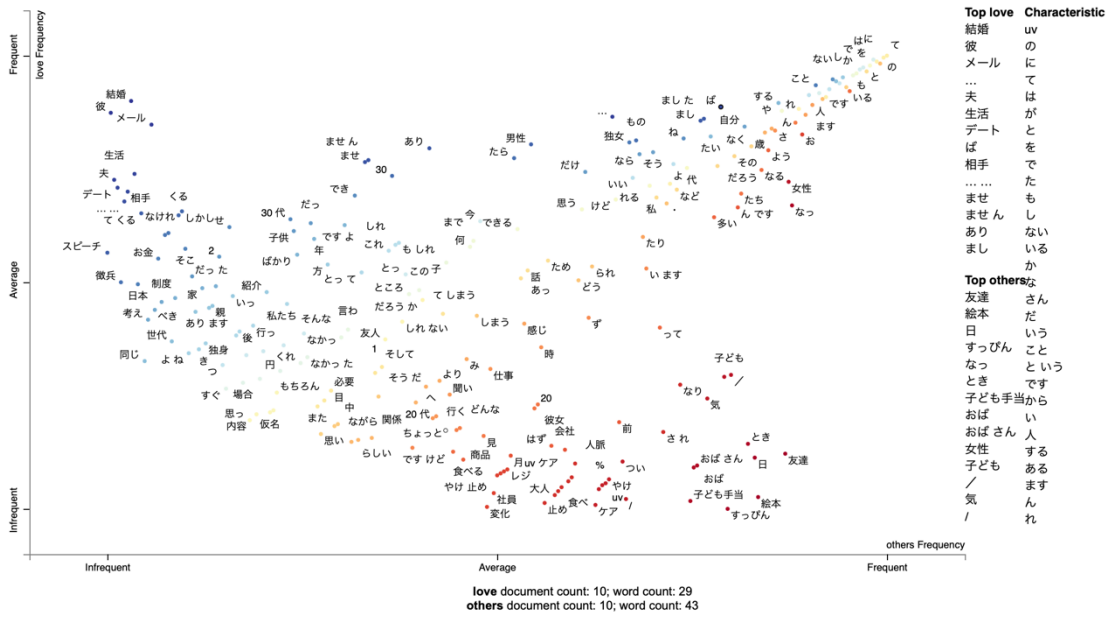


グラフからわかること

“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。