単語の出現頻度に着目した病院評判情報の分析

和多，太樹
九州大学大学院システム情報科学府

関，隆宏
九州大学大学評価情報室

田中，省作
立命館大学文学部

廣川，佐千男
九州大学情報基盤センター

出版情報：情報処理学会研究報告 ： 自然言語処理. 2005 (50), pp.15-20, 2005-05-26. 情報処理学会
バージョン：

単語の出現頻度に着目した病院評判情報の分析

和多太樹 †1　関隆宏 †2
田中省作 †3　廣川佐千男 †4

多様で膨大な量の文書が混在する WWW から、目的に合致する文書だけを抽出する技術が求められている。従来の検索では、主に単語、特に名詞の集合による文書の特徴付が使われている。しかし近年、「もの」ではなく、評判、評価、信頼性のように、感性に関わる検索の要求が高まっている。本研究では、対象文書群に現れる特徴的単語を名詞、動詞、形容詞、副詞の品詞ごとに抽出する方式を提案し、WWW 上の病院評判情報サイトの文書群について実験を行なった。単語の特徴判定の方法としては、病院の評判情報文書群における出現頻度と、一般の文書群における出現頻度の差を用いた。

An Analysis of Reputation Information of Hospitals using Frequency of Words

Taiki Wada †1 ,Takahiro Seki †2 ,Shosaku Tanaka †3
and Sachio Hirokawa †4

Information Retrieval is a key technology to utilize the vast amount of texts available on WWW. The conventional search technology mostly uses nouns to focus on target documents and characterize documents. A new trend of search concerns not objects or facts but emotional information, such as reputation, evaluation and reliability. For such purpose, this paper proposes a method to collect characteristic words using the difference of frequency of each word in the target documents and in other general documents. Characteristic words obtained with this method are reported for reputation information of hospitals.

1. はじめに

WWW が急速に普及し、世界中の誰もが容易に情報を発信できるようになった。WWW 上にはレビュー記事や個人の日記、新聞記事のように様々な文書が混在している。混在するそれらの文書群から、目的に合致するものだけを抽出する技術が求められている。しかし、現在の自然言語処理技術の水準では、文書の意味まで抽出することは困難で、文書の内容という点で有用かどうかを機械的に判定することは不可能である。また、従来の検索は、主に知識や事実を対象とするもので、利用者が使う検索のキーワードとしても、文書を特徴付ける単語としても名詞が使われることが多かった。しかし、このような検索だけでなく、評価、評判、信頼性のように人間の感性に関わる検索の必要性が高まっている。

例えば、目良等 [4] は、自然言語の文章が格フレーム表現に変換されたものとしてそこから情緒を数値として求める方式を述べている。Turner[6] は、レビューの文章の評価として、肯定、否定を表す単語と形容詞句、副詞句の共起から文を肯定的か否定的かを判断し、それらの平均として文章の評価を求めている。緒方等 [5] は、Web 上の書き込み記事からその記事を書いた人の共感度などの評価値を求めるためテキストマイニングを使っているが、評価値のベースとなる辞書は人手で作られている。

これに対し、本研究では、文書の表層的情報、ここでは特に単語の頻度という観点から文書の有用性を判定することを提案する。また、感性に関わる特徴を捉えるため、文書あるいは文書群の特徴を名詞、動詞、形容詞、副詞などの品詞ごとの単語の集合として捉え

†1 九州大学大学院システム情報科学府
Kyushu University, Graduate School of Information Science and Electrical Engineering
†2 九州大学大学評価情報室
Kyushu University, Office for Information of University Evaluation
†3 立命館大学文学部
Ritsumeikan University, College of Letters
†4 九州大学情報基盤センター
Kyushu Univeristy, Computing and Communications Center

る方式を提案する。与えられた文書群の特徴的単語を抽出するのに、頻度を使う研究としては有村等 [1] の研究がある。そこでは、複数の文字列の列である近接相関パターンについて、対象文書群（正例）で頻度が高く、同時に一般的文書群（負例）で頻度ができるだけ低いものを求めることにより、対象文書群に特徴的なパターンを発見する高速なアルゴリズムを示している。本稿では、複雑なパターンではなく、一つの文字列をパターンとする。しかし、感性情報が目的なので、単純な文字列ではなく、形態素解析を行なって品詞を確定した単語を文書群の特徴とする。

本稿では、WWW で公開されている病院の評判情報を具体的な分析対象とする。WWW の普及により、医療についても多くの情報が簡単な検索により得られるようになってきた。利用者（患者）にとって、どの病院がよい病院か、あるいは、現在利用している病院がどれだけ信頼できるか知りたいという関心も高まっている。単純な検索では、病院の住所や電話番号などは分かっても、そのような評価や評判についての情報に効率よくたどり着くことは容易ではない。これに対し、病院や医療関係の公的機関でも積極的な情報公開が始まっている。また、病院に対する利用者（患者）の評価・評判を書き込むことができる評判情報サイトやクチコミサイトが現れている。

例えば、病院の評判情報を集めたサイト ARA!患者推薦病院 NAVI ☆ には病院について次のようなコメントが書かれている。

> （病院 A）先生、看護婦さんがみんなやさしい。説明が丁寧で、納得して治療を受けられます。女性の先生もいて、話しやすい。

> （病院 B）医師とカウンセラーが別々に居て、カウンセラーの方がとてもよく話しを聞いて下さり、心の病気専門だけあって医師はもちろん、看護婦の方も親身になって下さいます。、、、

「先生」、「看護婦」、「治療」、「病気」など病院に関連する名詞が現れていることが分かる。また、「やさしい」、「優しい」、「丁寧」、「よく」（形容詞）、「られます」、「下さい」（動詞）、「とても」（副詞）など、病院に対する患者の評価を表す言葉が多く見られる。これらの単語は病院評判情報について特徴的なものであり、ニュース記事のサイトや下のような小説などの一般的な文書群では、稀にしか出現しないと考えられる。

> （夏目漱石：吾輩は猫である）吾輩は猫である。名前はまだ無い。どこで生れたかとんと見当がつかぬ。何でも薄暗いじめじめした所でニャーニャー泣いていた事だけは記憶している。、、、

本稿では、特定のテーマの文書群と一般的文書群におけるそれぞれの単語の出現頻度を比較することにより、そのテーマの特徴的な単語を抽出する方法を提案する。WWW で公開された病院評判情報について、夏目漱石の小説の文書群を比較対象として特徴的な名詞、動詞、形容詞、副詞を求める。

2. 出現頻度を用いた特徴語抽出アルゴリズム

本稿で提案する特徴語抽出アルゴリズムは、異なる文書群におけるそれぞれの単語の出現頻度の差異に着目する。すなわち、一般的な単語は、同種類の文書群の間で比較しても、異なるジャンルの文書と比較しても、出現頻度は同様な分布になるが、そのジャンルに特徴的な単語は、その特定の文書群で出現頻度は高く、一般の文書群では出現頻度は低い。

そこでまず、対象とする文書群に現れるすべての単語について、文書群 P における出現頻度と文書群 N における出現頻度をそれぞれ y 軸、x 軸として回帰直線を求める。２つの文書群が同種の文書群であれば相関が高く、ほとんどの単語がこの回帰直線の周囲に分布するものと予想される。しかし、２つが異なるジャンルの文書群であれば、一般的な単語は回帰直線の周囲に分布するが、それぞれの文書群に特徴的な単語は、回帰直線から乖離したところに現れる。そこで、回帰直線から乖離したところに現れる単語を、それぞれの文書群の特徴的単語とする。

２つの文書群 P,N が与えられたとき、それぞれの文書群に現れる特徴的な単語を次のようにして求める。

ステップ１：文書群 P,N に対し形態素解析を行ない現れる単語 $w_1, w_2, \cdots, w_n$ を求める。その際、各単語は品詞ごとに分ける。

ステップ２：i 番目の単語 w_i に対し、 それぞれの文書群 P,N における出現頻度 x_i, y_i を求める。

$$x_i = tf(w_i, P), y_i = tf(w_i, N)$$

ステップ３：n 個の２次元データ $(x_1, y_1), (x_2, y_2),$

☆ http://www.mscn.net/ara/ara_index.htm

$\cdots, (x_n, y_n)$ に対し次式により回帰直線 $y = ax + b$ を求める。

$$a = \frac{(n(\sum_{i=1}^{n} x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i))}{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} y_i)^2}$$

$$b = \frac{(\sum_{i=1}^{n} x_i^2)(\sum_{i=1}^{n} y_i) - (\sum_{i=1}^{n} x_i y_i)(\sum_{i=1}^{n} x_i)}{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} y_i)^2}$$

ステップ 4：$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ の各点 (x, y) について、回帰直線より上に現れるものを文書群 P の特徴的キーワード、下に現れるものを文書群 N の特徴的キーワードとして出力する。ただし、次式で与えられる回帰直線からの距離 $d(x, y)$ を乖離度とし、乖離度の大きなものから上位 10 位までを出力する。

$$d(x, y) = \frac{(ax + b - y)}{\sqrt{a^2 + 1}}$$

出現頻度の代わりに、キーワードの重要度としてよく用いられる tf*idf 値も考えられる。しかし、対象が同じような文書群の場合、一般的単語だけでなく、その文書群に特有の単語も df 値が高くなりその結果、tf*idf 値が小さくなる。そこで本稿では、単純な出現頻度を用いることとした。

3. 病院評判情報サイト

1 節の例で述べたサイト「ARA！患者推薦病院 NAVI」には、図 1 のように病院名、住所、利用者のコメントが書かれた 97 個の HTML ページがあった。それぞれのページからコメントの部分だけを切り出し合計で 2547 件、1 件あたり平均 140 字のコメントを集めた。他に 9 件のサイトについても同様に特徴語抽出を行なった。比較対象としては、青空文庫☆ で公開されている夏目漱石の著作 95 作品、全体で約 680 万字の文書を使った。それぞれのサイトについて、品詞ごとの異なり数は表 5 の通りである。

本稿では WWW で公開されている 9 個の病院評判情報サイト 1 について、次のような予想を検証すべく、単語の出現頻度の分析と、一般的文書との比較を行なった。

(病院評判情報の特徴予想)

- 形容詞が多く出現する。
- 医療関係の用語が多く出現する。
- 専門的な用語は使われていない。
- 肯定的か否定的かという評価や意見を含む。

☆ http://www.aozora.gr.jp

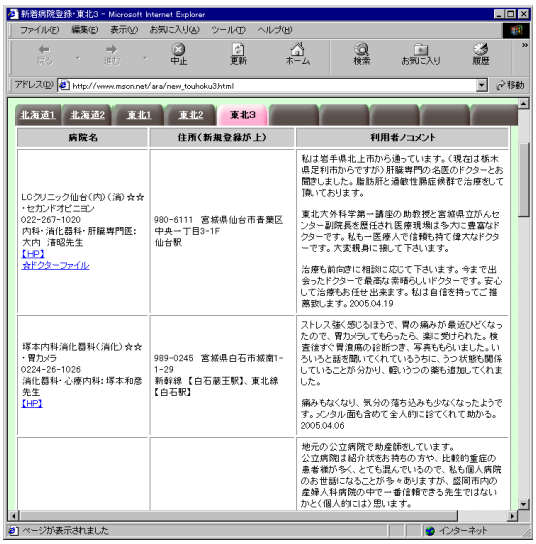

図 1 ARA！患者推薦病院 NAVI

表 1 病院評判情報サイト分析データ

サイト番号	件数	名詞	動詞	形容詞	副詞
1	2547	25177	11081	2107	1806
2	3613	18809	9667	2131	1777
3	280	4606	1810	407	344
4	2633	28706	12772	2622	2384
5	2633	12026	6157	1137	1089
6	1735	38769	19663	3303	3462
7	997	18791	9524	1987	1792
8	1786	35265	15470	1970	2399
9	470	4424	1918	469	296
10	947	24358	10523	3197	1953
夏目漱石小説群	95	587853	311565	36483	63156

4. 病院評判情報文書に対する仮説の検証

本稿では比較対象の一般的文書の例として、青空文庫で公開されている夏目漱石の小説を使い、この仮説を検証した。またその結果得られた、病院評判情報に頻出する単語について品詞ごとに報告する。

1 節で述べた仮説を次のようにして検証した。

まず、収集した『ARA!』の 541 個の txt ファイルを 270 個のファイル群 P と 271 個のファイル群 N とに分割した。これらのファイル群は同じ評判サイトから収集したものなので、2 つの間に何ら大きな違いはないファイル群である。この 2 つのファイル群に対して、P,N に出現する単語について、y 座標にファイル群 P での出現数、x 座標にファイル群 N での出現数をとった座標をグラフ上にプロットし、同じグラフにそれらの点に最もフィットする回帰直線も挿入した。名詞についてのグラフを図 2 に示す。また、各座標の直線までの距離の上位部分を表 4 に示す。図より、点は

表 2　病院評判情報サイト URL

1. ARA!患者推薦病院 NAVI
   http://www.mscn.net/ara/ara_index.htm
2. デンターネット街の歯医者さん
   http://www.ix3.jp/
3. Do Select
   http://www.bekkoame.ne.jp/ yokomi/aaa/hos.html
4. オススメの歯医者さん
   http://www2.odn.ne.jp/good-dentist/
5. 茨城クチコミ健康ネット
   http://www.iarc.jp/health/index/
6. クリニックジネコ
   http://www.jineko.net/clinic.html
7. クチコミ病院情報
   http://www.kuchikomi.tv/
8. SBSCorporation
   http://www.sbspet.com/animal.html
9. わんさか Net 福岡
   http://wansakanet.com/
10. SaferSex
    http://www.wbs.ne.jp/bt/d-net/safersex.htm

回帰直線上の周囲に集まっていることが確認できる。これは他の品詞についても同様であった。したがって、1 節で述べた通り、似たような文書群どうしの場合、各単語の座標は回帰直線付近に集まることが確認できた。

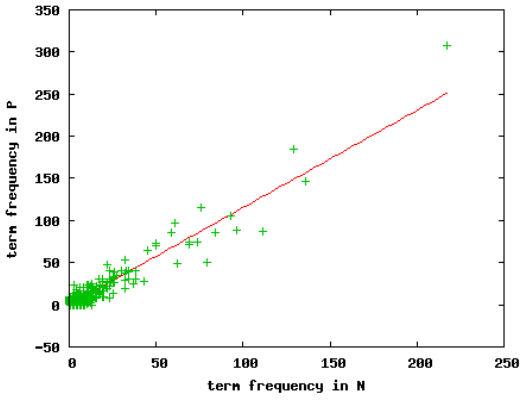

図 2　病院評判情報を 2 分割したときの出現頻度の比較 (名詞)

次に、病院評判情報における単語の出現頻度を縦軸に、小説における単語の出現頻度を横軸に取って同様に相関を調べたのが図 3 である。

図 3 を見ると、病院評判情報同士で比較した場合よりも点が分散している。これは、夏目漱石の著書に出現する単語と評判情報に出現する単語とでは種類が異なり、同じ単語でもそれぞれのファイル群では使われる頻度がかなり異なるという事実を表した結果となっている。乖離度が病院評判情報側に大きく乖離したものが病院評判情報の特徴語と考えることができる。

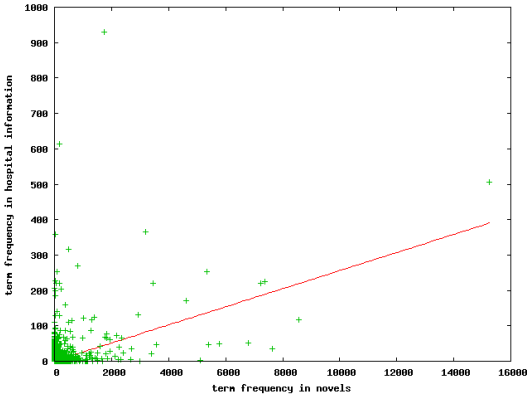

図 3　病院評判情報と小説における出現頻度の比較 (名詞)

5. 病院評判情報の特徴語

2 節で述べたアルゴリズムをサイト 1、サイト 2 について適用して得られた病院評判情報の特徴語（上位 10 個）を表 3,4 に品詞ごとに示す。サイト 1 の形容詞については、病院情報、小説それぞれの上位 50 個を回帰直線からの乖離度、出現頻度などの詳細データと合わせて表 5 に示す。

表 3　病院評判情報特徴語 (サイト # 2)

順位	名詞	動詞	形容詞	副詞
1	先生	くれる	良い	とても
2	病院	できる	やすい	とっても
3	治療	受ける	優しい	きちんと
4	説明	もらう	やさしい	初めて
5	丁寧	診る	すごい	しっかり
6	手術	くださる	痛い	かなり
7	患者	もらえる	詳しい	じっくり
8	出産	わかる	多い	一見
9	親切	いただく	ひどい	とりあえず
10	診察	分かる	おいしい	ゆったり

表 4　病院評判情報特徴語 (サイト # 2)

順位	名詞	動詞	形容詞	副詞
1	治療	くれる	良い	とても
2	先生	もらう	痛い	きちんと
3	歯医者	いただく	優しい	本当に
4	歯科	みる	やすい	初めて
5	丁寧	言う	やさしい	しっかり
6	説明	できる	すごい	とっても
7	歯	頂く	いい	かなり
8	院長	通る	多い	たまたま
9	患者	くださる	怖い	結構
10	歯科医	てる	素晴らしい	全然

前述の特徴語抽出で得られた病院評判情報の特徴語を見てみると、病院評判情報特有の表現が多く抽出されている。品詞ごとに見てみると、名詞については、

「先生」「病院」「療」といった病院関係（医療分野）の単語が上位を占めていることがわかる。特に 1 位の「先生」については、夏目漱石の著書で 1745 回も出現するにも関わらず、回帰直線からの乖離度が最も病院評判情報側に離れており、病院評判情報において特に頻繁に見られる単語であると言える。次に動詞について見てみると、「くれる」「くださる」「いただく」といった、敬語表現もしくは受身的な表現が目立つ。また、「診る」「治る」といった治療に関する語も目に付く。形容詞に関しては、やはり評判情報でよく目にするような語が多い。意味的にポジティブ・ネガティブといった正負を含む語、言い換えるとその形容詞で文章がある対象に対して肯定的であるか否定的であるかが決まってしまう語が多いということである。副詞については比較対象が文学作品ということもあり、口語的なものが多い。

具体的な病院評判情報で見られる文章は、「看護婦さんが優しい」などのように名詞と形容詞の組み合わせ、もしくは、「先生が詳しく説明してくれる」のとおうに名詞、形容詞、動詞の受身的な表現の組み合わせで構成されたものが多い。表 3,3 の単語はまさしくそのような語である。2 節で述べた特徴語判定アルゴリズムは２つのファイル群を比較した場合、一方特徴語を見出す有効な方法だと言える。

6. まとめと今後の課題

「もの」や事実についての検索だけでなく、評判、評価、意見などの感性に対する検索の要求が高まっている。本稿では、その例として、病院評判情報における特徴的単語を抽出する方法を提案した。病院評判情報と一般的文章における単語の出現頻度を比較し、病院評判情報における出現が一般的文章における出現より極端に多いものを抽出した。WWW 上の病院評判情報 10 サイトについて実験を行ない、本手法により抽出される単語が病院評判に関する特徴的な単語であることを確認できた。名詞、動詞、形容詞、副詞の品詞ごとに区分することにより、感性情報の切り分けができた。形態素解析は使うが、他に自然言語処理の技術も病院や医療に関する知識を一切使っていないので、病院評判情報に限らず、他の分野や対象でも適用可能と考える。

本稿では特徴語抽出を上位 10 件としたが、乖離度の分布を分析し適当な閾値を設定することが考えられる。表 5 における乖離度の分布のより詳細な解析によりこの閾値を求めることが可能と考える。比較対象を小説としたため、動詞や副詞として口語的な表現が多く得られた。対象を別の文書群とする実験を行ない、抽出される特徴語がどのようなものになるか分析する必要がある。表 3,4 で分かるように、サイトによって得られる単語の集合と順位が微妙に違う。複数のサイトの結果を統合することで、病院評判情報についてのより一般的な特徴語のリストを作ることが考えられる。また、これとは逆に、サイトごとの出現頻度の微妙な差で、サイト間の違い、例えば、信頼度の評価なども分析できると考える。実際、サイト１からは、ほとんどが肯定的な形容詞しか得られなかった。これは、サイト１がそもそも患者が推薦する情報を集めていたからだった。

今回は、単語の集合として文書群の特徴を捉えたが、出現する単語の位置も考慮して＜対象、属性、評価＞のような組とすれば、飯田等 [2] や小林等 [3] のような評価表現の収集が考えられる。単純な単語の集合のままでも、対象を評判情報に限定したクローラー [7] への応用が考えられる。

参 考 文 献

1) Hiroki Arimura, Jun-ichiro Abe, Ryoichi Fujino, Hiroshi Sakamoto, Shinichi Shimozono, Setsuo Arikawa, Text Data Mining: Discovery of Important Keywords in the Cyberspace, Proc. Kyoto International Conference on Digital Libraries 2000, Kyoto University, British Library and National Science Foundation (U.S.A.), 121–126, 2000.

2) 飯田龍, 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一, 意見抽出を目的とした機械学習による属性-評価値対同定, 情報処理学会研究報告, NL165-4, 21–28, 2005.

3) 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, テキストマイニングによる評価表現の収集, 情報処理学会研究報告 NL154-12,77–84, 2003.

4) 目良和也, 市村匠, 相沢輝昭, 山下利之, 語の好感度に基づく自然言語発話からの情緒生起手法, 人工知能学会論文誌 17(3),186–195,2002.

5) 緒方進, 池田真司, 牟田高信, 木本勝敏, Web 上のテキスト情報を用いた人物評価手法, 情報処理学会研究報告, NL165-2, 9–14, 2005.

6) Peter D. Turney, Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews, Proc. 40th Annual Meeting of the Association for Computational Linguisitics(ACL), 417–424, 2002.

7) 山田信太郎, 松永吉広, 伊東栄典, 廣川佐千男, Web シラバス情報収集エージェントの試作, 電子情報通信学会論文誌 D1, J86-D1(8),566–574, 2003.

表 5 抽出された特徴的形容詞 (上位 50 個)

	病院評判情報 (サイト# 1)				夏目漱石小説			
順位	単語	乖離度	頻度 (病院)	頻度 (小説)	単語	乖離度	頻度 (病院)	頻度 (小説)
1	良い	193.26	194	21	ない	-748.50	123	7702
2	やすい	163.29	174	110	好い	-135.71	0	1222
3	優しい	110.57	115	58	悪い	-60.71	28	806
4	やさしい	80.67	80	15	面白い	-51.29	3	504
5	すごい	44.28	42	3	黒い	-49.28	0	460
6	痛い	38.08	49	119	白い	-48.15	0	450
7	詳しい	37.08	44	84	暗い	-46.96	2	457
8	多い	32.37	67	327	寒い	-33.29	0	319
9	ひどい	25.98	31	68	深い	-32.95	0	316
10	おいしい	25.72	24	9	赤い	-32.75	1	323
11	明るい	20.88	35	148	高い	-31.32	36	617
12	楽しい	18.96	18	16	くい	-29.43	0	285
13	素晴らしい	17.45	15	3	強い	-27.12	15	396
14	美味しい	14.81	12	0	遠い	-26.76	10	349
15	短い	14.68	17	45	広い	-25.48	9	329
16	新しい	13.62	31	177	固い	-24.22	0	239
17	少ない	13.36	28	153	細い	-23.31	0	231
18	すばらしい	12.49	10	3	早い	-23.27	40	581
19	心強い	11.83	9	0	重い	-22.86	0	227
20	わかり易い	9.85	7	0	うい	-22.54	5	268
21	安い	9.43	18	100	苦しい	-21.06	5	255
22	気持ちよい	8.85	6	0	くらい	-20.89	7	271
23	上手い	8.85	6	0	長い	-19.05	56	684
24	酷い	8.51	6	3	淋しい	-18.89	0	192
25	うれしい	8.49	11	47	恐ろしい	-18.32	0	187
26	厳しい	7.61	6	11	狭い	-17.44	1	188
27	よい	7.57	153	1299	薄い	-17.07	0	176
28	辛い	7.52	9	38	嬉しい	-16.38	8	240
29	忙しい	7.44	12	65	わるい	-16.30	1	178
30	ほしい	7.35	7	22	むずかしい	-16.05	0	167
31	易い	7.04	6	16	おかしい	-15.99	2	184
32	こわい	6.50	5	12	美しい	-15.37	0	161
33	速い	6.41	4	4	若い	-15.11	28	404
34	ものすごい	5.87	3	0	えらい	-13.21	0	142
35	すっごい	5.87	3	0	古い	-12.81	6	191
36	難しい	5.65	3	2	旨い	-12.76	0	138
37	珍しい	5.22	6	32	善い	-12.76	0	138
38	有り難い	4.88	2	0	低い	-12.67	1	146
39	数少ない	4.88	2	0	つまらない	-12.19	0	133
40	イイ	4.88	2	0	暑い	-11.88	1	139
41	きつい	4.77	2	1	丸い	-11.85	0	130
42	気安い	4.43	2	4	濃い	-11.85	0	130
43	あたたかい	4.28	3	14	青い	-11.74	0	129
44	力強い	4.09	2	7	ええ	-11.29	0	125
45	温かい	3.94	3	17	よろしい	-9.73	1	120
46	ままならない	3.89	1	0	偉い	-9.70	0	111
47	聞きづらい	3.89	1	0	惜しい	-9.47	0	109
48	気持ち良い	3.89	1	0	おとなしい	-8.93	1	113
49	ややこしい	3.89	1	0	厚い	-8.79	0	103
50	きめ細かい	3.89	1	0	乏しい	-8.68	0	102