【佐々木先生による解説シリーズ３ 】テキストデータからの特徴抽出

単語の頻度から分かるニュース記事の特徴②

2019年 5月 31日

お伝えしたいポイント

- テキスト中の単語と頻度の関係を表すジップの法則
- 語彙の多様性を示す指標タイプ・トークン比
- ニュースヘッドラインテキストは限られた語彙で形成

前回はニュースヘッドラインから出現する単語を抽出し、その頻度を分析すると少数の単語が大部分の出現単語をカバーしていることを説明しました。

今回は単語の頻度と頻度順位の関係を表すジップの法則と語彙の多様性を表す指標であるタイプ・トークン比について解説します。

テキスト中の単語と頻度の関係を表すジップの法則

日本語や英語などの自然言語では、単語の頻度分布に関してZipf(ジップ、ジフ)の法則があてはまると言われています。Zipfの法則は単語を出現頻度の多い順に並び替えると、単語の出現頻度がその順位の k 乗に反比例する傾向があるというものです。経験則ですが、単語の出現頻度だけではなく、ウェブページのアクセス頻度や都市の人口などの様々な現象でもあてはまると言われています。このZipfの法則がニュースヘッドライン1年分のテキスト集合でもあてはまるかどうか検証します。

Zipfの法則は単語の出現頻度を f 、その順位を r と置くと、出現頻度 f と順位 r の関係は以下の式で表されます。この式にある c は対象とするテキストの長さや語彙の種類によって変化する定数を意味します。

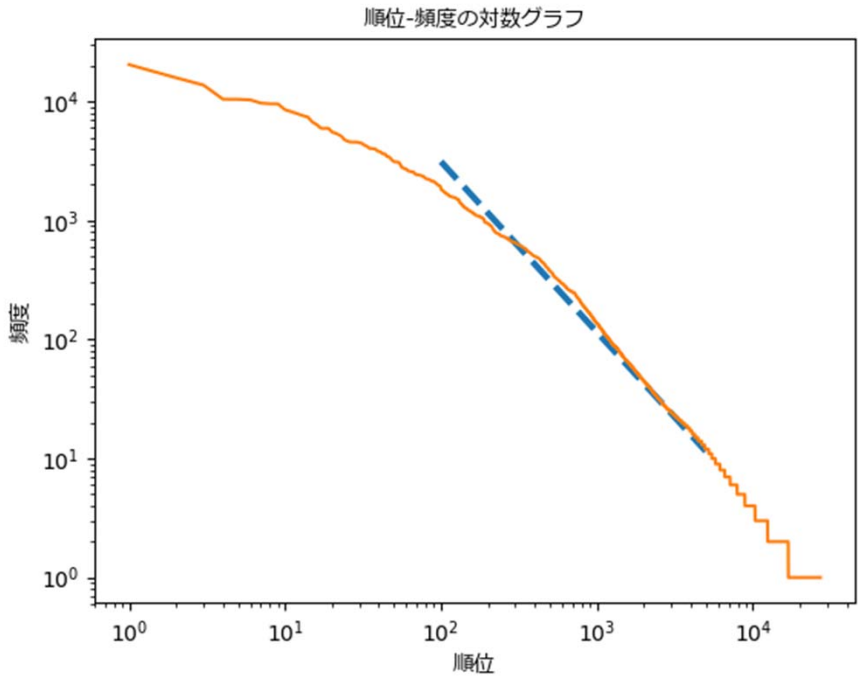
$$f = \frac{c}{r^k}$$

この式に対して両辺の常用対数を取ると、出現頻度の対数 $\log_{10} f$ と順位の対数 $\log_{10} r$ の関係が得られます。したがって、順位と出現頻度を両対数グラフを使って描画すると傾きが $-k$ の直線になります。

$$(\log_{10} f) = \log_{10} c - k(\log_{10} r)$$

この法則は本来 $k=1$ で、一般的に両対数グラフの傾きが -1 に近い直線になると言われています。しかし、様々なテキストを分析すると、傾きが -1 とは異なる値を取ることも確認されています。そのため、本稿では単語の出現頻度 f がその順位 r の k 乗に反比例するとし、両対数グラフにおける近似直線の傾きをマイナス k としています。

さて、前回使用したニュースヘッドラインのテキスト集合に出現する内容語27,132語に対して、出現頻度を高い順に並び変えます。その順位を横軸、出現頻度を縦軸とし、どちらの軸も対数で表した両対数グラフを作成します。この両対数グラフは下に示すグラフになります。



（出所）QUICK社のデータを元に茨城大学で算出

グラフは直線となるのが理想的なのですが、この例では曲線となっています。Zipfの法則に従っていたとしても、高い順位と低い順位においてグラフは一般的に近似直線からずれると言われています。使用したテキスト集合では、例えば順位が300～3,000において直線で近似可能だと考えられます。そこで、順位が300と5,000の点を結んだ直線（グラフ中の青い点線）をグラフにプロットしました。順位-頻度の曲線とこの直線を比較すると近似可能だということが分かります。

Zipfの法則は元々ジェームス・ジョイスの「ユリシーズ」という小説に出現する26万単語を対象として分析をしていました。この分析では対象単語の中には冠詞や前置詞などの高頻度で出現する機能語も使用しています。内容語だけではなく、機能語も含めて出現頻度を求めると、高い順位におけるグラフのずれは改善されるのではないかと考えられます。

テキストにおける語彙の多様性を分析する

語彙の多様性を示す指標として最もシンプルな方法は、テキスト集合における延べ単語数 N に対する異なり単語数 V の比を求めることとなります。この指標 R_{TT} は「タイプ・トークン比」と呼ばれ、この値が 1 に近いほど多くの種類の単語が使われていて、語彙に多様性があるということができます。

$$R_{TT} = \frac{V}{N}$$

ニュースヘッドラインのテキスト集合の延べ単語数 N は1,055,800単語で、異なり単語数 V は27,132単語ですので、このテキスト集合のタイプ・トークン比は

$$R_{TT} = \frac{V}{N} = \frac{27132}{1055800} \fallingdotseq 0.0257$$

と、約2.57%という低いタイプ・トークン比になります。すなわち、ニュースヘッドラインには語彙の多様性がなく、ほとんど同じ一連の単語が使われているということを示しています。このタイプ・トークン比がどれほど低いかを検証するために、夏目漱石の長編小説「こころ」に対して同様にタイプ・トークン比を計算してみます。「こころ」においても、名詞、動詞、形容詞の内容語を抽出して、単語の集合を作成します。

この集合に対する延べ単語数 N は 43,538 単語で、そのうち異なり単語数 V は 5,460 単語でした。これから、タイプ・トークン比を計算すると、

$$R_{TT} = \frac{V}{N} = \frac{5460}{43538} \fallingdotseq 0.1254$$

と、約12.54%のタイプ・トークン比となります。夏目漱石の小説は深みのある単語が多く使われ、語彙力を増やすのに有効だと言う方もいます。この値とニュースヘッドラインのタイプ・トークン比を比較することで、ニュースヘッドラインのタイプ・トークン比の約2.6%が小さい値であることが分かると思います。

ニュースヘッドラインでタイプ・トークン比が小さい値となるのは決まった表現の多さが要因のひとつとして挙げられます。同じパターンで出現する表現が多ければ多いほど、延べ単語数は増加しますが異なり単語数はあまり増加しません。このニュースヘッドラインで、同じパターンで出現する表現を調べてみたところ、数多く出現するパターンとして、下記の表のようなフレーズがありました。この結果を見ると、冒頭に「<QUICK>」や「<NQN>」があるニュースヘッドラインは「5%ﾙｰﾙ報告」の対象となる銘柄や決算短信を発表した企業を知ることに適していることが分かります。

頻繁に出現するフレーズ

フレーズ	頻　度
5%ﾙｰﾙ報告	8,231
決算短信　　発表会社一覧	1,880
株投信 ﾃﾞｲﾘｰﾓﾆﾀｰ	1,319
銘柄ﾗｳﾝﾄﾞUP	1,313
自社株買　　発表会社一覧	1,148
配当予想　　発表会社一覧	1,095

（出所）QUICK社のデータを元に茨城大学で算出

佐々木先生　プロフィール

- **佐々木稔（ささきみのる）**

  徳島県徳島市生まれ．平成13年徳島大学大学院博士後期課程修了．博士(工学)．平成13年茨城大学工学部情報工学科助手を経て，平成17年より同専任講師。研究分野は，機械学習や統計的手法による情報検索，自然言語処理等に従事．情報処理学会，言語処理学会，計量国語学会，電子情報通信学会各会員．

バックナンバー

佐々木先生による解説シリーズ

- **001「テキストデータからの特徴抽出　ニュースからの単語による特徴表現」**
  https://www.daiwa-am.co.jp/specialreport/quants/20190204.html
- **002「テキストデータからの特徴抽出　単語の頻度から分かるニュース記事の特徴①」**
  https://www.daiwa-am.co.jp/specialreport/quants/20190426.html

鈴木教授による解説シリーズ

- **001「AI運用に挑む」**
  https://www.daiwa-am.co.jp/specialreport/quants/2017/quantst_20171207.html
- **002「集団化する人工知能」**
  https://www.daiwa-am.co.jp/specialreport/quants/2018/quantst_20180125.html
- **003「２年目のジンクスを集合知AIで緩和」**
  https://www.daiwa-am.co.jp/specialreport/quants/2018/quantst_20180301.html
- **004「 時系列データの見えない法則をつかむ」**
  https://www.daiwa-am.co.jp/specialreport/quants/2018/quantst_20180409.html
- **005「愚かな人間心理・カモにするAI」**
  https://www.daiwa-am.co.jp/specialreport/quants/2018/quantst_20180501.html
- **006「ナイトメア★アノマリーを狙え」**
  https://www.daiwa-am.co.jp/specialreport/quants/2018/quantst_20180601.html
- **007「ブルーオーシャンAI戦略」**
  https://www.daiwa-am.co.jp/specialreport/quants/2018/quantst_20180703.html
- **008「深層学習による株価予測 (前編)」**
  https://www.daiwa-am.co.jp/specialreport/quants/2018/quantst_20180802.html
- **009「深層学習による株価予測 (後編)」**
  https://www.daiwa-am.co.jp/specialreport/quants/2018/quantst_20181001.html
- **010「ニュースを読んで投資判断する集合知AI」**
  https://www.daiwa-am.co.jp/specialreport/quants/2018/1204.html