【佐々木先生による解説シリーズ2 】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

2019年 4月 26日

お伝えしたいポイント

- テキストにおける単語の頻度を数える
- 形態素解析ツールMeCabなどによる単語の抽出
- 約3.5%の単語が80%の出現単語をカバー

前回はテキストを単語分割し、単語集合として表現することができれば、単語を特徴としてテキストを数値化することができると説明しました。

今回は数値化されたテキストを用いて、テキストの特徴を捉える簡単な方法について解説します。

テキストにおける単語の頻度を数える

特徴を捉える対象のテキストとして、株式会社QUICKが配信した2017年のニュースヘッドラインを使用します。この中には経済に関連したニュースだけではなく、日経新聞の一般的なニュースや財務省の公式発表を速報する記事が含まれています。今回は経済に関連したニューステキストの分析を目的とするため、ニュースヘッドラインの冒頭に「<QUICK>」もしくは「<NQN>」が含まれるニュースヘッドラインを対象としてテキスト分析を行います。ニュースヘッドライン1件あたりの文字数は数十～百数十文字で、以下の表のように長さの短いタイトルが記載されています。

QUICKニュースヘッドラインの記事例

配信日時	ニュースヘッドライン
2017/1/3 7:00:00	<NQN>◇ＮＹ投機筋　原油の買越幅が2年6カ月ぶり高水準　金の買越幅は縮小
2017/6/1 8:10:00	<NQN>☆ビール、値上げ過去最大級　安売り規制　投資家も悩ます
2017/12/1 12:13:00	<QUICK>2017年11月･投資信託(ﾌｧﾝﾄﾞ) 日本株投信 1年半ぶりの10兆円台 分類別純資産残高推移

（出所）QUICK社

対象となるニュースヘッドラインの総数は121,093件となり、これをテキスト集合と呼びます。このテキスト集合に含まれる冒頭の「<QUICK>」や「<NQN>」と記号の「☆」や「◇」はニュースタイトルと関係がないので、文字のマッチングなどを活用し、タイトルを求めます。

このテキスト集合のニュースヘッドラインすべてについて、形態素解析を行い、内容語を抽出します。前回のマーケットレターで解説したように、形態素解析ツールのMeCabと解析用辞書のmecab-ipadic-NEologdを使って、単語と品詞を取り出します。この品詞が名詞、動詞、形容詞である単語を内容語として抽出します。例えば、前ページのニュースヘッドラインから内容語を抽出すると、以下のような単語の集合が得られます。すべてのニュースヘッドラインについて、抽出した内容語を合わせて、ひとつの内容語の集合を作ります。

ニュースヘッドラインから抽出した内容語の集合

ニュースヘッドライン	内容語を抽出
<NQN>◇ＮＹ投機筋　原油の買越幅が2年6カ月ぶり高水準　金の買越幅は縮小	NY, 投機筋, 原油, 買, 越, 幅, 2年, 6か月, ぶり, 高水準, 金, 買, 越, 幅, 縮小
<NQN>☆ビール、値上げ過去最大級　安売り規制　投資家も悩ます	ビール, 値上げ, 過去, 最大級, 安売り, 規制, 投資家, 悩ます
<QUICK>2017年11月･投資信託(ﾌｧﾝﾄﾞ) 日本株投信 1年半ぶりの10兆円台 分類別純資産残高推移	2017年, 11月, 投資信託, ﾌｧﾝﾄﾞ, 日本株, 投信, 1年半, ぶり, 10兆円, 台, 分類, 別, 純資産, 残高, 推移

（出所）QUICK社のデータを元に茨城大学で加工・抽出

これから、この集合に含まれる内容語を数え上げて、内容語の頻度分布を求めましょう。それぞれの内容語の頻度を計算し、最も多く出現する上位25単語の頻度を記載した表と上位40単語をグラフ化したものを次のページに示します。

2017年のニュースヘッドラインにおける頻度の高い25単語

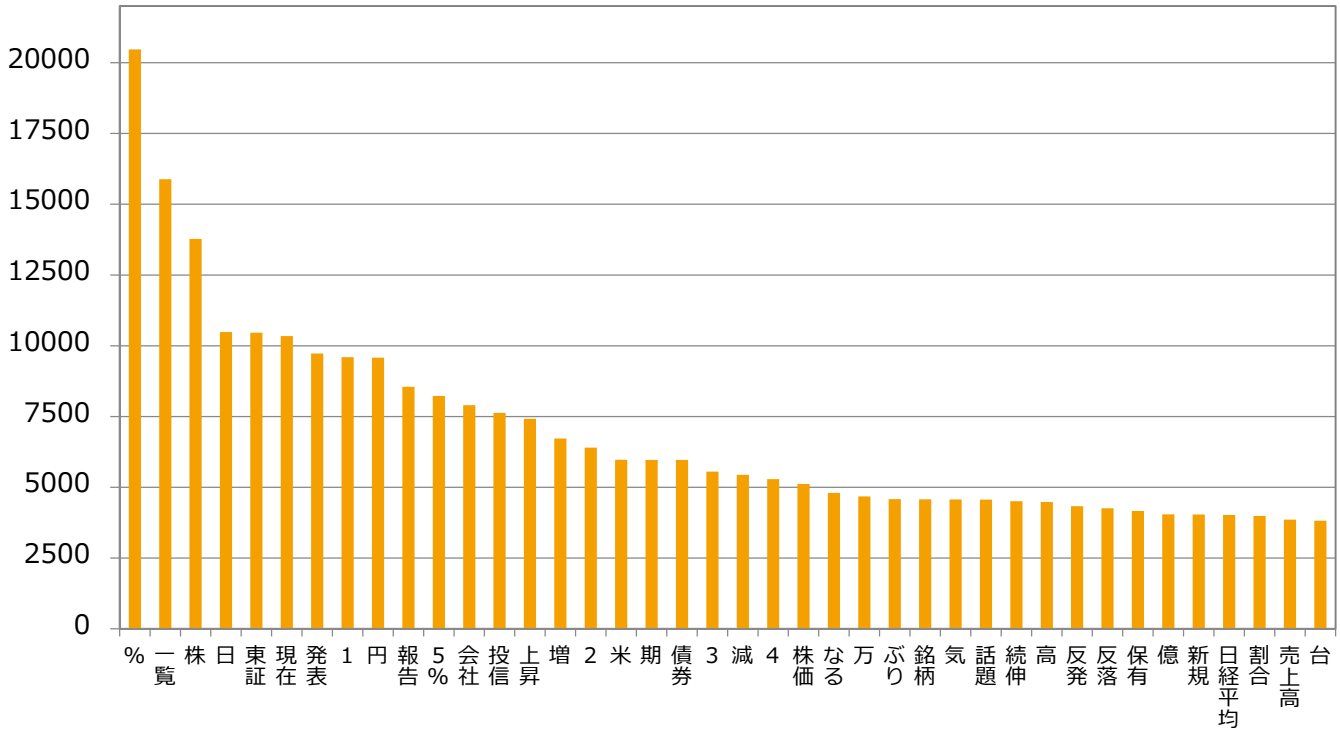
単語	頻度	単語	頻度	単語	頻度	単語	頻度	単語	頻度
%	20,472	現在	10,342	5%	8,230	2	6,401	減	5,443
一覧	15,882	発表	9,725	会社	7,902	米	5,975	4	5,288
株	13,778	1	9,598	投信	7,630	期	5,970	株価	5,118
日	10,487	円	9,581	上昇	7,423	債券	5,965	なる	4,802
東証	10,463	報告	8,550	増	6,726	3	5,557	万	4,677

（出所）QUICK社のデータを元に茨城大学で加工・抽出

最も頻度が多い内容語は記号の「%（品詞は名詞,接尾,助数詞）」でした。ニュースヘッドラインには割引短期国庫債券の利率や東京円金利スワップレートの速報値などのように、数多くの数値が存在することが分かります。二番目に頻度が多いのが「一覧」で、エクイティ・自社株買・決算短信の発表会社一覧や5%ルール報告の届出一覧などを示す記事のタイトルとして使われています。企業や個人が報告した内容をまとめた記事も多いことが分かります。三番目に頻度の多い「株」は、「ロンドン株」などの株式市場や「株二万円台回復」などの株価、「食品株」などの業種をまとめた株式銘柄というように、様々な意味で出現しています。このように、テキスト全体で出現頻度の高い単語を分析すると、全体的な傾向や特徴が見えてくることがあります。

次に、上位 N 単語の出現頻度の総数が総単語数に対して、どれくらいの割合を占めているのか調べます。このテキスト集合の延べ単語数は1,055,800単語存在します。1位の「%」は全体の約1.94%を占めています。上位10位までの単語は全体の約11.26%を占めます。これが上位948位までの単語までとなると、全体の80%に近い出現頻度数となります。テキスト集合の異なり単語数は27,132語なので、全体の約3.49%の単語が80%の出現単語をカバーしていることになります。

2017年のニュースヘッドラインに対する頻度の高い上位40単語のグラフ



（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

鈴木教授による解説シリーズ

・**001「AI運用に挑む」**
http://www.daiwa-am.co.jp/market/html_ml/ML20171207_1.html

・**002「集団化する人工知能」**
http://www.daiwa-am.co.jp/market/html_ml/ML20180125_1.html

・**003「２年目のジンクスを集合知AIで緩和」**
http://www.daiwa-am.co.jp/market/html_ml/ML20180301_1.html

・**004「 時系列データの見えない法則をつかむ」**
http://www.daiwa-am.co.jp/market/html_ml/ML20180409_1.html

・**005「愚かな人間心理・カモにするAI」**
http://www.daiwa-am.co.jp/market/html_ml/ML20180501_2.html

・**006「ナイトメア★アノマリーを狙え」**
http://www.daiwa-am.co.jp/market/html_ml/ML20180601_1.html

・**007「ブルーオーシャンAI戦略」**
http://www.daiwa-am.co.jp/market/html_ml/ML20180703_1.html

・**008「深層学習による株価予測 (前編)」**
http://www.daiwa-am.co.jp/market/html_ml/ML20180802_2.html

・**009「深層学習による株価予測 (後編)」**
http://www.daiwa-am.co.jp/market/html_ml/ML20181001_1.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