投資工学開発室

（審査確認番号：他2018-135）

吉野貴晶 のクオンツ　トピックス：*NO5*

AIによるテキスト情報の解析（テキストデータ前処理編）

テキスト情報をどのように活用するか？

- 連載形式でAI（人工知能）と投資手法の関係性を紹介。
- テキストデータの前処理はAIの最終精度を左右する重要な工程。

最近、AI（人工知能、以下AI ）に関連するニュースが増えています。投資の分野でも研究開発が盛んに行われており、実際に投資手法として利用可能な段階まで進展しています。本レポートでは、AIと投資手法の関係性をご紹介したいと思います。

今回はAIに学習させるデータの準備、一般的に前処理と言われる手順をご紹介します。

1. テキスト情報を用いた投資手法の開発

前回のレポートに続き、テキスト情報の利用についてご紹介したいと思います。まず最初に、テキストデータをどのように投資手法の開発まで繋げるのかを考える必要があります。様々なアプローチがありますが、一例としては図1の手順が考えられます。本レポートでは手順①と②をご紹介します。

図1.テキストデータの活用アプローチ

①：テキストデータを取得
②：テキストデータを綺麗な状態に整形
③：AIが読み取れるようにデータを加工（数値情報に変換）
④：単語 または 文章単位でスコア化（AI）
⑤：算出したスコアと投資対象との関連性を確認（スコアとTOPIXとの関係 等）
⑥：実際に投資してリターン獲得

2. データの準備

まずはテキストデータを準備します。データの条件としては、景気や経済、市場に関連したデータであり、かつ豊富なサンプルが得られることです。

2-1. 景気ウォッチャー調査

今回は景気ウォッチャー調査を利用します。景気ウォッチャー調査とは、内閣府が毎月集計している経済統計データで、街角景気の調査を目的に、各地域で働く人々を調査対象としています。身の回りの景気について、良い～悪いまでの5段階評価で回答する形式ですが、対象者がその選択肢を選んだ理由を、景気判断理由集として内閣府が公表しています。今回はこの景気判断理由集を活用します。

図2.景気判断理由集（現状）の例

分野	景気の現状判断	業種・職種	判断の理由	追加説明及び具体的状況の説明
家計動向関連 （北海道）	◎	一般小売店［土産］（経営者）	来客数の動き	・前月に引き続き国内客、外国人観光客共に増加しており、売上も順調である。６月の売上は前年比114.4%、一昨年比112.4%となっている。また、外国人観光客による売上は全体の２～３割を占めている。
	▲	スーパー（企画担当）	来客数の動き	・このところ来客数の前年割れが続いている。絶対数が減少していることもあるが、むしろ来店頻度が落ちていることに起因しており、買物を控えたいという消費者の節約行動の表れとみられる。

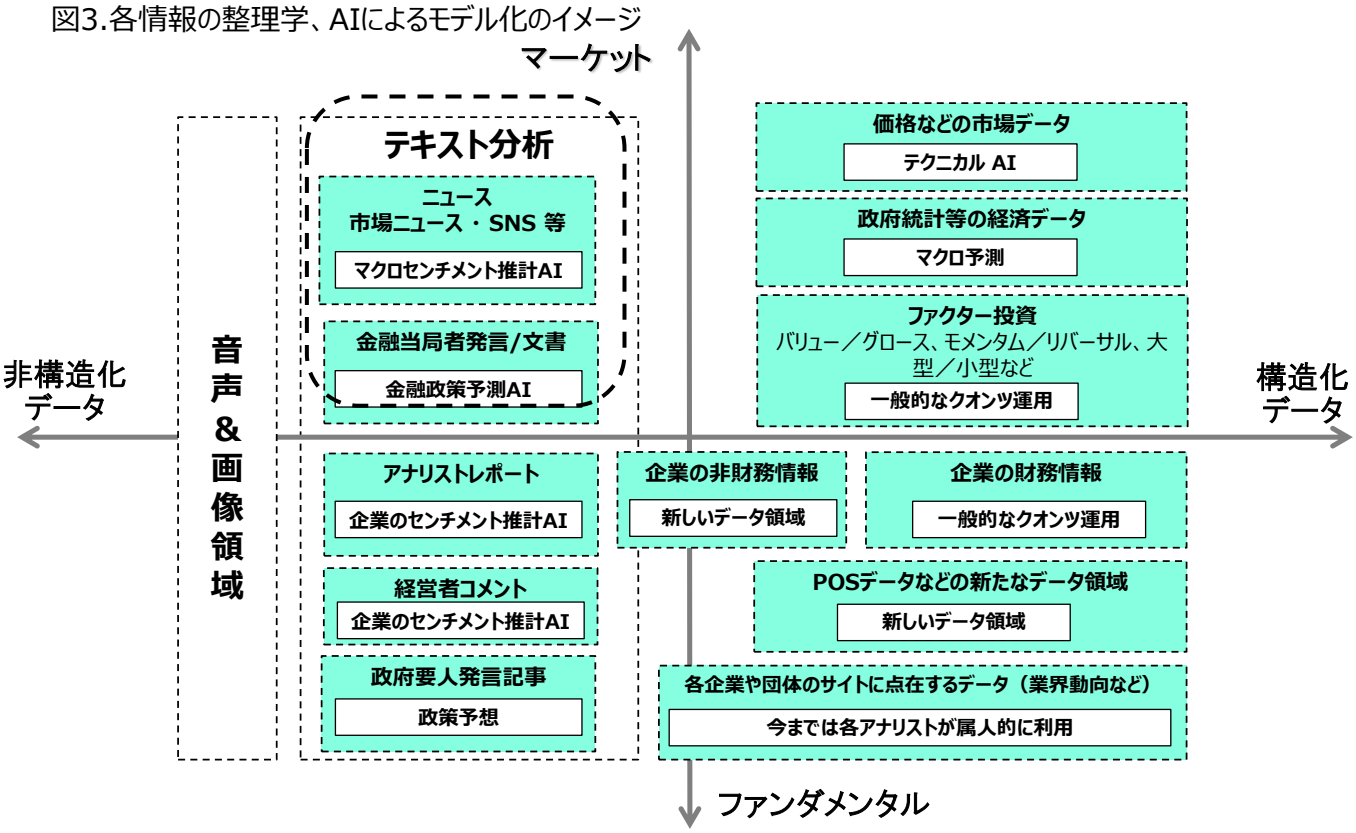
データの特性と構造

2-2. データ特性の確認

今回利用する景気ウォッチャーの景気判断理由集について、データの特性を図3を用いて確認したいと思います。これは、データの特性を切り口に、データ領域（どのデータを使うか）とAIモデル（どのような結果を目指すか）をマッピングしたものになります。

今回のデータはテキスト情報なので非構造化データであり、左側に該当します。上下の軸であるマーケットデータとファンダメンタルデータの観点で考えると、マーケットデータ寄りといえるかと思います。結果、図上では左上部分に分類されます。

図3.各情報の整理学、AIによるモデル化のイメージ



※マップ上の各項目とも複数の領域に跨る可能性があるが、分類案の一例として図示

補足：データ構造の説明（2018年8月15日発行　吉野貴晶 のクオンツ トピックス：NO4、より一部抜粋）

1. マーケットデータとファンダメンタルズデータ

縦軸は、データの特性がマーケット寄りか、それともファンダメンタル寄りか、を表しています。マーケット寄りのデータとは、株価情報や出来高など、一般に市場から取得されるデータになります。対してファンダメンタル寄りのデータとは、企業の財務状況など、その企業の状態を表すデータになります。

2. 構造化データと非構造化データ

横軸はデータの構造を表しています。そのデータが構造化データか、非構造化データか、を切り口にしています。構造化、非構造化データの定義は様々ですが、本レポートでは、株価のような最初から扱いやすい数字データになっているものは構造化データとし、テキストや画像、音声データは非構造化データ、と括っています。

テキストデータの前処理にはどのような種類があるか？

3. データ前処理（テキストデータ編）

さて、この景気ウォッチャー調査のデータですが、分析を実施する前にデータに手を加える必要があります。これが「前処理」と呼ばれる工程です。大まかに言うと、テキストデータを使いやすい状態に加工する処理になります。図4はいくつかの前処理手法をまとめたものです。

図4.テキストデータにおける前処理手法

処理名	処理内容
PDFからのテキスト抽出	PDFに掲載されているテキスト情報を抽出する処理。様々なソフト等を使ってのアプローチがあるが、この工程だけではテキストのデータとしての精度が低い場合があり、なんらかのデータ処理工程が別途必要になることが多い。
半角全角変換	同じ意味の単語だが、半角での記載と全角の記載が混ざっている場合にどちらかに寄せる処理。 （例）全角：アセット、半角：ｱｾｯﾄ
表現の揺らぎ調整	同一の物に対して複数の名称が有るような場合に、名前を統一する処理。 （例）日本株式、内株、国内株
分かち書き	文章を文節や単語単位で区切る処理。文章の構成要素に分解することで、後工程でAIが単語を認識することが可能になる。
ストップワードの取り扱い	助詞など、それ単体では意味を持たず、かつテキストデータを解析した際に大量に出現する単語に対する処理。場合によっては解析時点で該当単語を無効になるように処理する。
単語の原型への修正	日本語における単語の活用形を全て原型に戻す処理。
句読点の取り扱い	句読点（。や、）の処理。これ単体では意味を成さないため削除する場合が多い

3_1 半角全角変換

半角と全角の違いが混在している場合には、加工処理が必要です。ある時には「アセット」と書かれているのに、別の時には「ｱｾｯﾄ」記載されているケース等です。この場合、加工処理を行わない場合、両単語が別々の単語としてAIに取り扱われる可能性があり注意が必要です。対応策としては、全角のカタカナや数字は全て半角に変換または認識させるようにプログラムを作成することが考えられます。

3_2 表現の揺らぎ

また、表現の揺らぎも問題になります。これは言い方のバリュエーション、と言うこともできます。例えば、日本株、という単語があったとします。日本株式、内株、国内株、と言ったように、同じ意味を表したいのに表現にバリュエーションがある場合、これらを揃えてやる必要がでます。この表現のゆらぎを解消するために機械学習による専門のプログラムを作ることがあります。

テキストデータの前処理にはどのような種類があるか？

3_3 分かち書き

テキスト情報を活用する上で、文節や単語単位で分解する場合があります。この作業を「分かち書き」と呼びます。この工程にて、文章を構成要素に分解することで、後工程で単語別に文章を認識することが可能になります。

3_4 ストップワードの取り扱い

ストップワードとは、それ単体では意味を成さない単語を指します。例としては、「て、に、を、は」にあたる助詞や助動詞などです。これら単語は一般的なテキスト解析では意味を成さないものとして除外されるケースが一般的です。しかし、画一的に除外、という選択肢を取るだけではなく、助詞や助動詞まで含めた構文解析や係り受け解析を実施したい場合は、あえて削除せずに残す、という選択肢も考えられます。
今回の前処理では、助詞や助動詞を削除して、名詞、動詞、形容詞のみを抽出しました。

3_5 単語の活用形と原形

日本語において、単語が原形から活用形になる場合が多く見られます。「良く」と「良い」、「下がって」と「下がる」等です。これらを同じ単語として認識させたい場合、原型に揃える作業をすることになります。この処理をすることで、認識される単語の種類数を減らすことができます。

3_6 句読点処理

日本語文に頻出する句読点は、それ単体では意味を成さないので、削除することが一般的です。ただし、文章の切れ目の判定や文脈の活用のために、句読点を利用する場合もあります。

3_7 前処理実例

今回の対象テキストデータに対して、前処理として「分かち書き」、「句読点処理」、「一部のストップワードの除去」、「単語の原型への修正」を実際に実施した結果が図5になります。この前処理済みデータを起点とし、データ解析をしていくことになります。

図5.分かち書きから原型への修正まで

1、元の文章
高額商品の動きがやや良くなってきた。

↓

2、分かち書き ※eos = end of statement、調査回答の終了を示す。
高額商品 の 動き が やや 良く なっ て き た <eos>

↓

3、品詞の確認（上記の分かち書き文章の品詞）
名詞 助詞 名詞 助詞 副詞 形容詞 動詞 助詞 動詞 助動詞

↓

4、名詞、動詞、形容詞のみ抽出
高額商品 動き 良く なっ き <eos>

↓

5、単語を原型に修正
高額商品 動き 良い なる くる<eos>

単語リストの作成

4. 単語リスト

前処理後のテキストデータを使って、どのような単語の出現回数が上位に来るか、下位に来るかを確認して見たいと思います。図6が今回の対象データにおける、出現頻度が多い単語と少ない単語の一部抜粋リストになります。

「前年」という言葉が頻度上位に来ていますが、これは回答例で、「前年と比べて」というような表現が多かったのではないかと推測されます。また、「客」や「消費」、「販売」といった、景気の動向に関連しそうな単語の出現頻度が高くなっています。一方、出現頻度が少ない単語の一部抜粋リストに目を移してみると、特定の業種や業態、特殊な条件でしか文章に出てこないような単語は、出現回数が極端に低くなる傾向にあるようです。

図6.前処理後の単語例

出現回数：多	出現回数
前年	23470
客	21067
数	21001
売上	16222
来客	12391
良い	11224
消費	10564
販売	10123
動き	9851
状況	9697
減少	9672
単価	9210
増える	9192
増加	8994

出現回数：少	出現回数
宿所	1
誤り	1
開戦	1
つぶる	1
売れ高	1
厳	1
硝子	1
緻密	1
併行	1
地理	1
甚だしい	1
借地	1
重厚	1
長大	1

5. データ前処理の重要性と次回レポートについて

本レポートでは、AIに学習させるデータの準備、一般的に前処理と言われる手順をご紹介しました。
なぜ AIの手法ではなくデータの前処理の話をするのか、と思われた方もいるかもしれません。しかし、この前処理がAIの最終結果を大きく左右する重要な分野であり、避けては通れない工程となります。

次回レポートでは、この前処理済みのテキストデータを使って実際に解析していく手法、手順をご紹介する予定です。

～執筆者の紹介～

吉野貴晶（写真：右）
「日経ヴェリタス」アナリストランキングのクオンツ部門で16年連続で1位を獲得。ビックデータやAIを使った運用モデルの開発から、身の回りの意外なデータを使った経済や株価予測まで、幅広く計量手法を駆使した分析や予測を行う。

髙野幸太（写真：左）
ニッセイアセット入社後、ファンドのリスク管理、マクロリサーチ及びアセットアロケーション業務に従事。17年4月に投資工学開発室に異動後は、主に計量的手法やAIを応用した新たな投資戦略の開発を担当する。