吉野貴晶 のクオンツ　トピックス：*NO8*

AIによるテキスト情報の解析（テキストデータの特徴を掴む）

AIによる大量テキストデータの自動分類とネットワーク構造の可視化

- 連載形式でAI（人工知能）と投資手法の関係性を紹介。
- AIを活用してテキストデータをグループ化、さらに可視化に挑戦。

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。

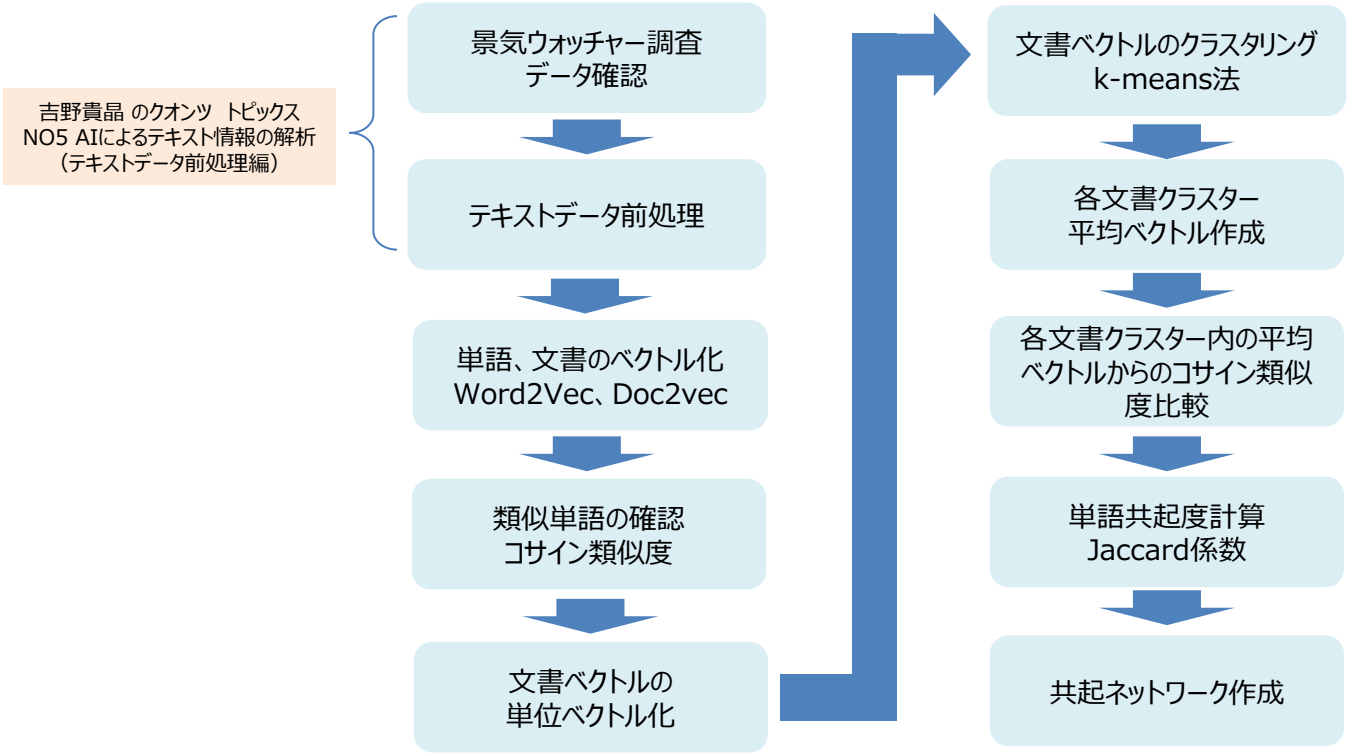
今回のテーマは経済テキストの特徴を掴むための様々な手法になります。

1. 大量のテキストデータから特徴を掴む

AI技術の発展により、様々な場面でテキスト情報を活用しようとする試みが増えています。テキストデータの例では、アンケート結果やwebニュース記事、web掲示板情報、twitter等のSNSなどが考えられますが、これらのテキスト情報を活用しようとする場合、まずデータの特徴を掴む必要があります。このテキストデータはどこから取得され、何について記述されているか。加えて、どのようなグループ分け、ラベル付けが出来るかが重要な特徴になります。このようなデータの特徴は、従来は人がデータ全体を俯瞰した上で考えていました。しかし、昨今では扱うデータ量も膨大になったため、人手では全てを俯瞰するのは困難です。

このような状況の中、AIを利用して、自動でテキストデータの特徴を把握できないか？というニーズが提起されます。今回はこのニーズを意識した上で、AIによるテキストの分類を実施します。また、実際に分類されたデータの特徴を人間が掴むための可視化にも挑戦します。

図1. 今回のレポートテーマ



単語と文書のベクトル化

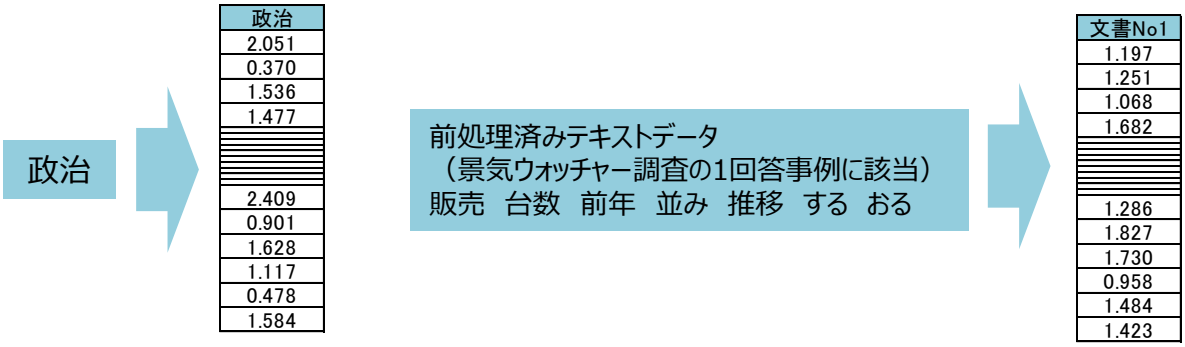
2. ベクトル化とは？

昨今のAI領域においては、ベクトル化という技術が頻繁に使われています。簡単に言うと、単語や文章を数字情報に置き換える処理になります。そもそも機械であるAIは日本語の文字をそのまま扱うのは難しいので、扱いやすいようにベクトル（数字情報）に変換する必要があります。

2_1. Word2Vec、Doc2Vec

単語や文書をベクトル化をする手法として、Word2VecとDoc2Vecがあります。これは、機械学習を利用して、単語や文書をn次元の数字情報に変換する手法です。このn次元の数値情報に変換する際に、AIが単語または文書の特徴が表せるように数字を決定します。

図2. 単語や文書のベクトル化



2_2. 類似単語の確認

今回は文書のベクトル化としてDoc2Vecを利用します。この過程で単語のベクトル表現もWord2Vecと同様に取得されています。今回のベクトル化において、単語の特徴が上手く表現されているか確認したいと思います。具体的には、単語を指定した際に、ベクトルが類似※している単語を上位いくつか表示し、人間の感覚とずれていないかを確認します。実際の結果が図3です。感じ方は人によるところがありますが、概ね対象の単語に関連している単語が上位に位置しているかと思います。あえて経済寄りの単語を対象に選びましたが、政治と為替は似た単語が上位に来ています。

※コサイン類似度による類似度確認をしています。この説明は後述になります。

図3. 単語のベクトル表現から分かる類似単語

対象	政治

類似度順位	該当単語
1	北朝鮮
2	離脱
3	EU
4	英国
5	大統領

対象	求人

類似度順位	該当単語
1	労働
2	正規
3	転職
4	正社員
5	就業

対象	経済

類似度順位	該当単語
1	政治
2	政策
3	米国
4	情勢
5	世界

対象	為替

類似度順位	該当単語
1	円高
2	米国
3	大統領
4	乱効果
5	離脱

対象	石油

類似度順位	該当単語
1	軽油
2	鋼材
3	原料
4	原材料
5	値下がり

文書ベクトルのクラスタリング

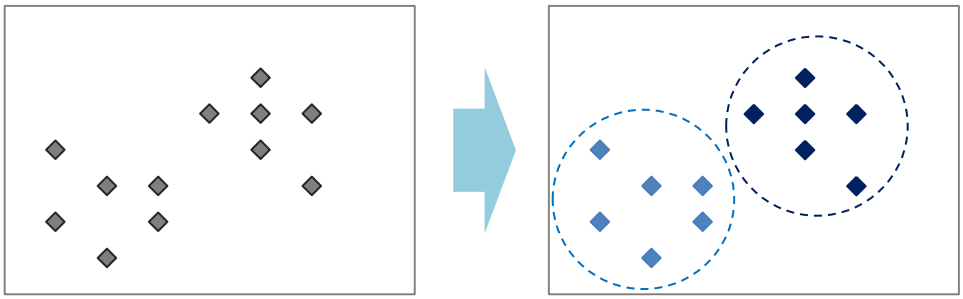
3. クラスタリングとは？

Doc2Vecを適用し、景気ウォッチャー調査を1回答ずつ100次元の文書ベクトルに変換します。総回答数が数万もの大量データになります。この文書ベクトルが、上手く文書の特徴を捉えていると仮定し、文書を特徴毎にグループ分けしたいと思います。

3_1. k-means法によるクラスタリング

グループ分けについて、今回はk-means法によるクラスタリングを利用します。クラスタリングとは、任意の数のグループ（クラスター）に分類することであり、その分類手法がk-means法になります。K-means法とは、AI（教師無し機械学習）の一種として考えられます。

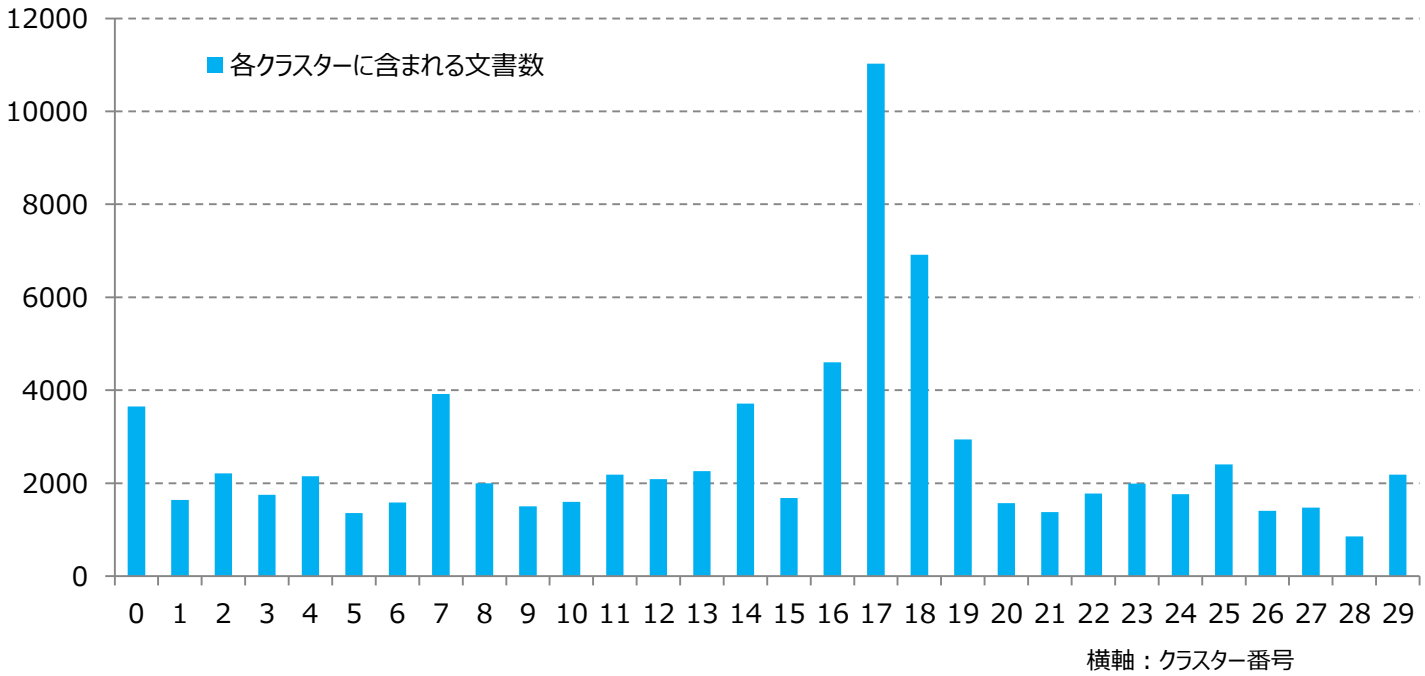
図4. クラスター分けの二次元イメージ



3_2. 全ての文書ベクトルを30個のクラスターに分類

今回はクラスター数30で分析を行いましたが、その結果が図5です。AIが文書ベクトルから判断した基準により、各文書が各クラスターに特徴毎に分類されたことになります。結果を見ると、各クラスターにおいて含まれる文書（回答）の数に偏りが見られます。1クラスター当たり、含まれる文書数は大体2000前後が多いですが、クラスター17番は10000以上の文書を含み、やや突出しています。

図5. クラスター分析の結果：各クラスターに含まれる文書数



平均ベクトルとコサイン類似度

4. 各クラスター内でのベクトル分布を確認する

クラスター内の文書がどの程度確からしく分類されているか確認したいと思います。

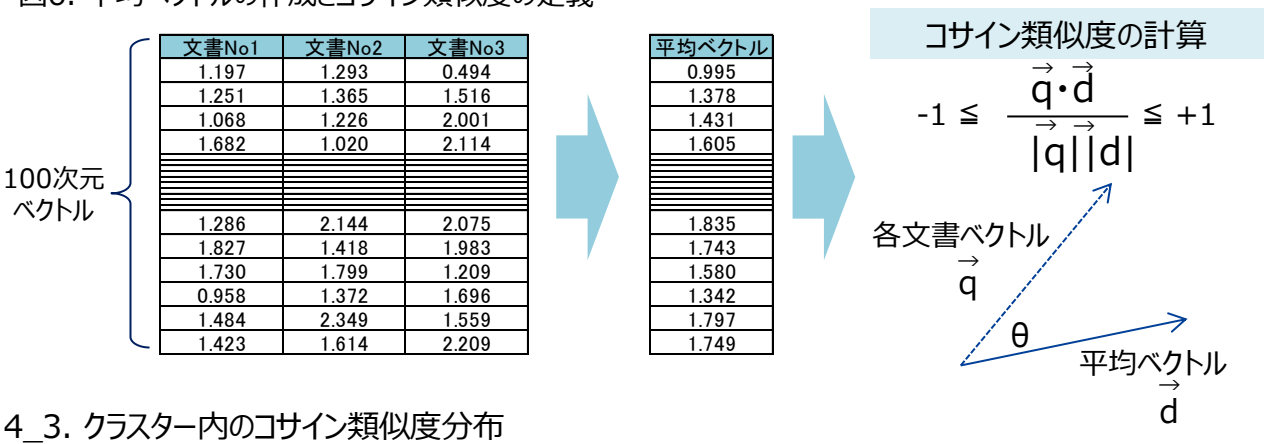
4_1. 平均ベクトル

各クラスターは先述のようにk-means法で分類されています。この分類の過程で重心（平均）ベクトルが使われることから、各クラスターの特徴を表すベクトルは、各クラスター内文書ベクトルの平均ベクトルといえます。各クラスター内の文書ベクトルとこの平均ベクトルとの類似度を考えます。

4_2. コサイン類似度

ベクトル間の類似度を測る方法は複数考えられますが、今回はコサイン類似度を利用します。コサイン類似度とは、二つのベクトルの為す角度のコサイン値であり、結果の数値は-1から+1までの範囲となります。「数字が大きいほど類似している」、と解釈されます。ここで、K-means法では類似度に重心（平均）ベクトルからのユークリッド距離を活用しており、コサイン類似度と尺度が違うことに違和感を覚える方もいるかもしれません。しかし、今回はクラスタリングする前に各文書ベクトルを単位ベクトル化しており、ユークリッド距離とコサイン類似度の順序関係は同じになります。

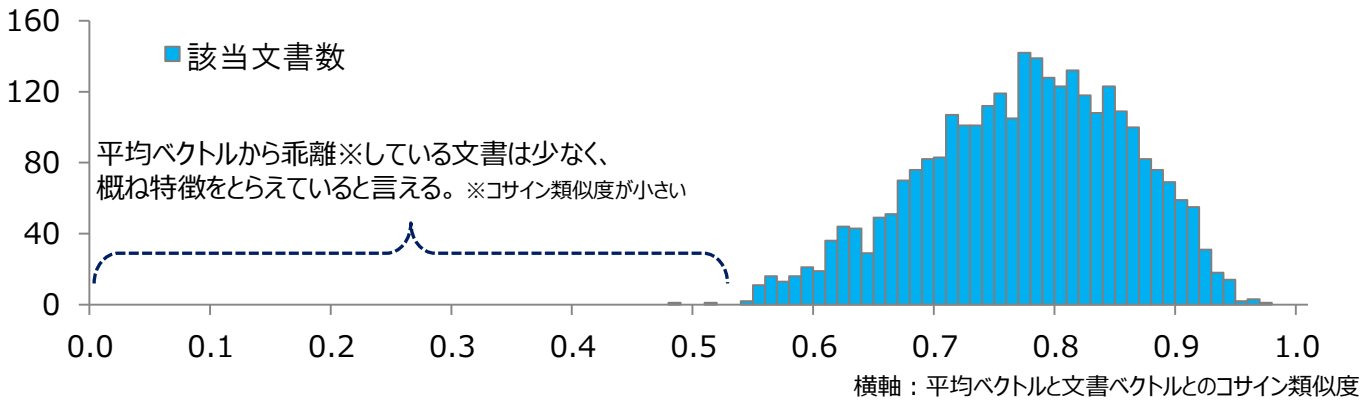
図6. 平均ベクトルの作成とコサイン類似度の定義



4_3. クラスター内のコサイン類似度分布

あるクラスター内における、平均ベクトルと文書ベクトルとのコサイン類似度を実際に計算し、分布を図示したのが図7になります。分布が+1側に偏っており、概ね平均ベクトルの特徴に沿ってクラスタリングが行われていることが分かります。一方、コサイン類似度が小さい（1から離れている）文書ベクトルもある程度存在します。これは、今回計算された30個のクラスターでは分類が難しく、半ば無理やりに分類された文書と推測されます。

図7. クラスター内平均ベクトルとクラスター内各文書ベクトルとのコサイン類似度



単語間の共起とは？

5. 共起分析

さて、クラスタリングされた文書について、その文書の特徴を人間が把握する方法を考えてみたいと思います。簡単な方法としては、各クラスター内で平均ベクトルとのコサイン類似度が高い文書を眺める方法や、各クラスター内で頻出する単語が何かを比較する方法、等が考えられます。計算コストが低く、差が出れば納得しやすい手法だと思います。一方、文書数が非常に多い場合や、単語単体に着目するだけでは特徴が分からない場合も考えられます。そこで、単語と単語の関係性に着目する手法として、今回は共起分析を試してみます。なお、共起分析自体は以前からある手法であり、AI（機械学習）の領域ではありませんが、AIの結果を人が確認する手法の例として今回は提示します。

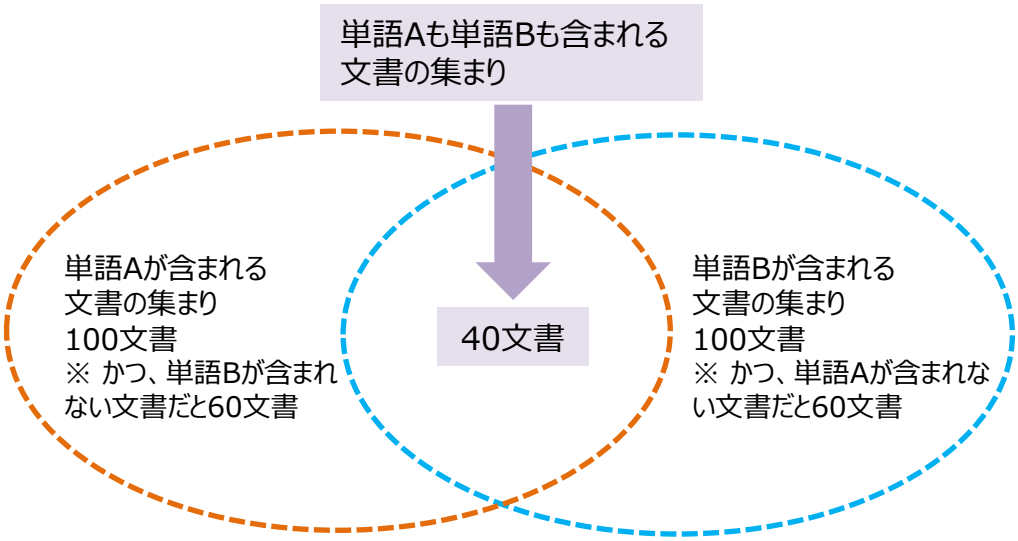
5_1. 共起の定義

まず始めに、共起とは何か？を定義しないといけません。今回の分析における共起とは、「ある2つの単語のペアが、1つの文書（回答）の中に含まれていれば共起である」とします。また、共起関係の強さを表す指標には、今回はJaccard係数を利用します。整理したものが図8です。

図8. 共起関係とJaccard係数の定義

「販売」と「売上」の共起度を計算する例

「販売」を含む文	販売 台数 前年 並み 推移 する おる
「販売」と「売上」を含む文 →共起関係	販売 量 売上 前年 並み 景気 上向く いる 思える
「売上」を含む文	売上 前年 上回る いる もの 問い合わせ 受付 件数 前年 大きい 下回る いる



Jaccard係数
= 単語Aも単語Bも含まれる文書の数
÷（単語Aが含まれるが単語Bが含まれない文書の数
＋単語Bが含まれるが単語Aが含まれない文書の数
＋単語Aも単語Bも含まれる文書の数）
= 40÷（60+60+40）
= 0. 25

共起ネットワークの作成

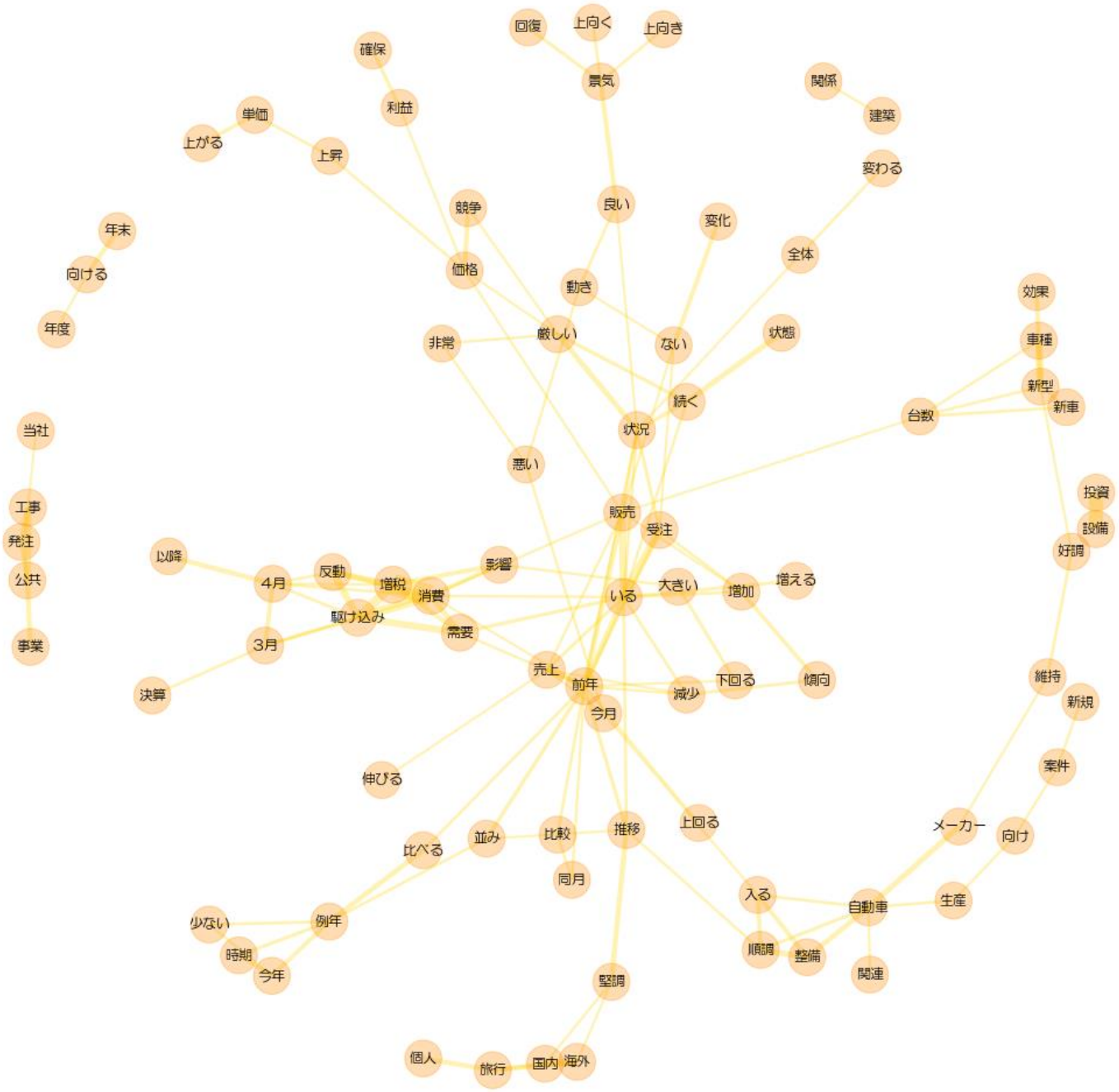
6. 共起関係を図示する

先ほど計算した共起度について、人が見て考えられるように図示したいと思います。今回はネットワーク分析を利用します。

6_1. 共起ネットワークの図示

先ほど計算した共起関係を図示したものが図9です。丸い部分（ノード）は単語を表し、共起関係にある単語間が線（エッジ）で結ばれています。共起度が大きいほど、線が太くなっています。円状に表現したため、一部の単語は円の外周上に表示されていますが、関係性は確認できます。なお、この共起ネットワークを作成する過程では、余りに共起度が低いものや、余りに出現頻度が高く、その単語が意味を成しづらいものは除外しています。

図9. クラスター0における共起ネットワーク



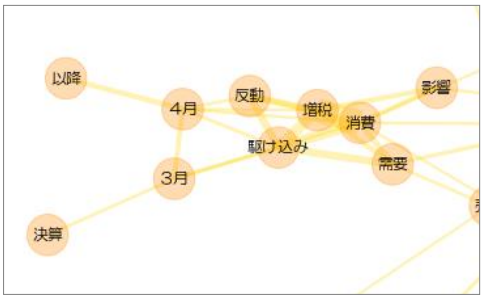
共起ネットワークから何が読み取れるか？

6_2.共起ネットワークを部分的に眺める

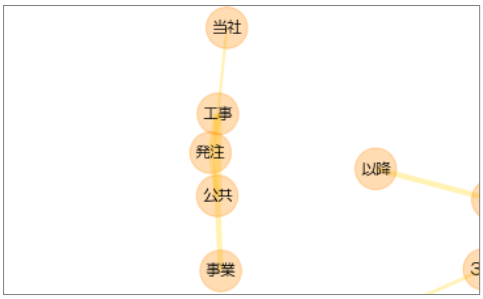
共起ネットワークの全体を俯瞰するのも重要ですが、一部に焦点を絞って眺めるのも有効です。以下の図10は筆者が切り取ったネットワークの一部です。このクラスターには、図10に記載したテーマの単語、文章が含まれていることが推測されます。また、単語群を繋ぐ単語にも注目することで、各単語群の関係性を発見する手助けになることも期待されます。このように、共起ネットワークを作成すると、大量の文書から特徴を見つけ出せる可能性があります。

図10. 共起ネットワークの部分的な関係性

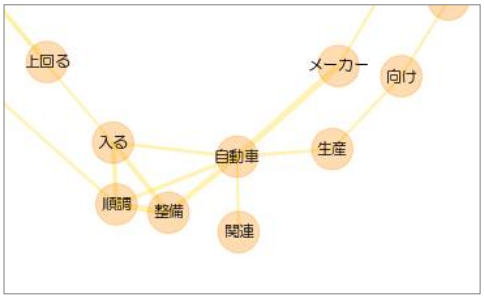
消費税増税に関連する単語群



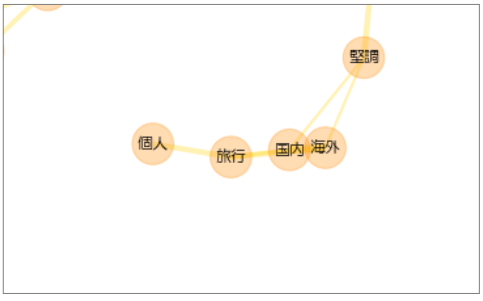
工事に関連する単語群



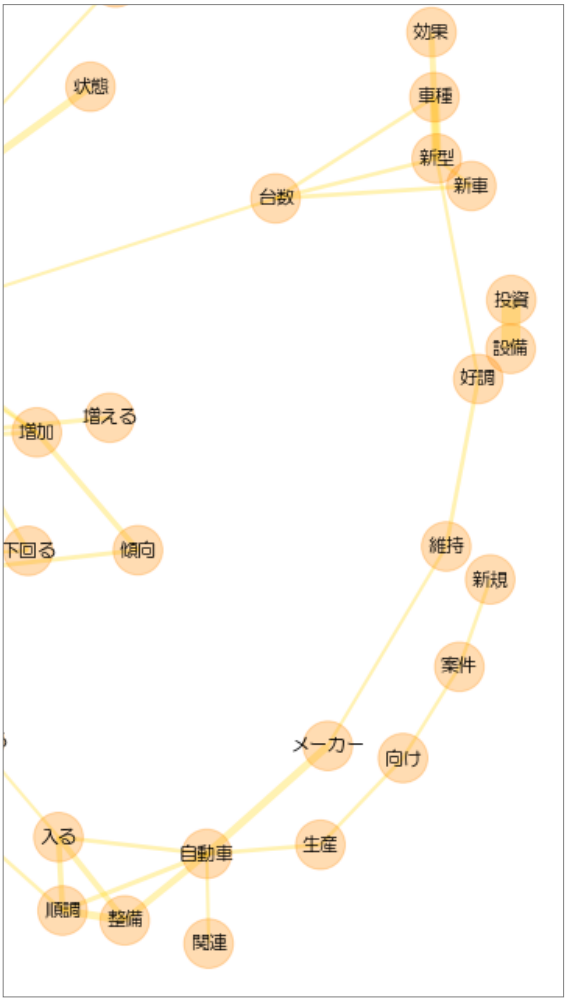
自動車メーカーなどに関する単語群



旅行に関連する単語群



自動車メーカーや設備投資に関する単語群



様々な手法を比較検討することが重要

7. 他の手法との兼ね合い

今回のレポートでは、大量のテキストデータに対してAIによる自動分類、並びに人が理解するために共起ネットワークの可視化を実施しました。この手法は大量のデータセットがある場合に、人手をかけずに知見を得る上で効果を発揮します。または、このクラスタリングをした後に、さらに後続として別のAI処理を噛ませる場合も有効だと思います。一方で、そもそもAIによって文章をベクトル化してグループ分けすることに対しては、どうしても解釈性の問題が付きまといます。結果の解釈が困難な場合が考えられるためです。

テキストデータを分類する方法は他にも考えられます。ある程度人手で付けたラベルデータが既にあるならば、それを教師データとして、ニューラルネットなどでラベル付け用モデルを構築することも考えられます。また、トピックモデルという手法もあります。これはAIが文書をトピック（単語など）でグループ分けする手法です。

上記のように、様々な手法がありますが、どの手法が適しているかはケースバイケースです。ユーザーのニーズにも依存しますし、複数の手法を組み合わせた方がベストな場合もあります。投資工学開発室が取り組んでいる投資手法の開発においても、同じ問題に直面します。技術が進化する一方、複数の手法を理解し、実装して比較できることが、現在は最も重要なのかもしれません。

8. 終わりに

次回レポートでも引き続きAIをテーマに取り扱う予定です。投資手法に活用するための挑戦をベースとし、テキストデータの解析手法を継続するか、画像処理系を予定しています。
（都合により変更になることがあります）

Appendix

A-1. 参考文献

1. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean
Distributed Representations of Words and Phrases and their Compositionality.

2. PySpark, NetworkXを利用した単語共起ネットワークの並列分散処理と可視化

～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。