【論文】

Python, Embedding Projectorを用いたTwitterデータ分析
2016年東京都知事選挙を事例に

和　田　伸一郎

■　はじめに

本稿の目的は大きく分けて二つある。一つは、2016年7月に東京都知事選挙について言及されたTwitter全数データを用いて、Twitterユーザーが、何について、どのような投稿を行っているかを調べることである。この調査手法として、クラスタ分析（クラスタリング）を行う。クラスタ分析とは、「機械学習」の三つのカテゴリ、すなわち、「教師あり学習」、「教師なし学習」、「強化学習」のうちの、「教師なし学習」を指す[1)]。もう一つは、ビッグデータである全数データを、Pythonを用いて、極力、人間を介入させることなく機械学習を行うことによって、どこまでクラスタリングの精度が出るか（しかも日本語のテキストデータを使って）を確認することにある。

クラスタリングについては、奥村、高村（2010：77-78）が、非常に分かりやすい説明を行っているので、少々長いが、以下に引用する。

> ある製品に関する“お客様の声”が大量に集まったとしよう。これを分析したいのだが一つ一つすべて読むのは手間がかかるので、とりあえずどういった種類の不満や要望などがあるのかを概観したいとする。そのため、類似した不満や要望は自動的にグループ化したい。そうすれば、各グループにつき少数の“お客様の声”を読めば、全体を概観したことになる。
>
> このように、非常に多くの文書があり、とりあえずこれらをいくつかのグループに分けたいという状況に直面したことがある方は少なくないだろう。文書だけでなく例えば単語でも同様である。似た単語を一つのグループにまとめることにより、同じグループの単語には同じ処理を施すことが可能である。このようなグループ化の作業をクラスタリングと呼ぶ。また、できあがったグループをクラスタとよぶ。

本稿で行うのは、ここで述べられているような意味でのクラスタリングである。そして、ここで述べられている文書にあたるものがTwitterデータとなる。さしあたりここで注意が必要なのは、クラスタリングと「分類」（上の例で言えば「教師あり学習」）が異なるということである[2)]。

> どんなクラスタができあがるかはわれわれは知らない［･･･］。できあがった各クラスタをみて、「ああ、この単語クラスタはこういう意味の単語の集合なんだ」という類推をすることは可能であるが、前もってそれを知ることはできない。これは、はじめから目的のグループがあって、各事例（文書、単語など）がそのグループに属するか否かを推測する、分類［･･･］とは異なる。（奥村、高村（2010：78））（強調は引用者による）

しかし、クラスタリングには大きな問題がある。奥村、高村は、そのことについて「クラスタリン

グは計算に時間がかかることが多い。これは、アルゴリズム中で、あらゆる事例の組合せを試したり、繰返し計算を行ったりするからである。」という点を挙げている。

この欠点の大きな克服が2013年になされた。自然言語の大きなデータを非常に高速に処理し、クラスタリングするword2vecアルゴリズムが、当時、Googleの研究所にいたトマス・ミコロフ（2018年12月現在はFacebookの研究所に所属。Tomas Mikolovプロフィール参照）らによって考案された。本稿では、Pythonライブラリのgensim版word2vecを用いて、後で述べるTwitterデータからモデルファイルを作成することとした。word2vecについては後述する。

■　選挙における有権者の関心について。

それでは最初に、今回の分析対象とするTwitterデータの内容となる、東京都知事選挙についての背景についてざっと触れておきたい。

八代尚宏（2016）によれば、国政選挙に関しては、年代別に投票行動を見ると、若い世代の投票率が低く、60歳代が最も多い。この理由について八代は以下のように述べている。

> 高齢者の投票率が高い理由は、引退者が大部分を占める高齢者世代にとって、投票のために費やす時間コストが低いことや、長年住み続けている地域社会との結びつきが強いことが挙げられる。[･･･] このことが、選挙での集票に左右される政治家の行動を通じて、高齢者への社会保障給付をいっそう増やすことに結びつくという悪循環をもたらしている。
> 八代（2016：13）（強調は引用者による）

この傾向については、今回の分析対象とする都知事選挙で、一種の逆転現象が起きていた。逆転現象というのは、東京新聞が選挙期間中に行った調査によれば、有権者が重視している政策が、二位に「医療・福祉」で、トップが「教育・子育て支援」となっていたからである。つまり、八代が投票を集めやすいとしていた高齢者に対する「医療、福祉（介護）」を、「教育・子育て支援」が上回る結果となっていた。くらしに関して都民が最も重視する政策は「「教育・子育て支援」「医療・福祉」がともに三割弱で上位を占めた。[･･･] 中でも女性は、「教育・子育て支援」への関心が高く、18～29歳は51.6%、30代は60.4%、40代も41.5%がトップに挙げた。」（東京新聞　2016年07月24日　朝刊）また、朝日新聞による調査でも「新知事に一番力を入れてほしい政策（5択）は、「教育・子育て」27%が最多」（朝日新聞2016年7月25日　朝刊）となっていた。

ところで、子育て支援で、最も問題となっているのが、「待機児童」問題である。待機児童問題とは、「子育て中の保護者が、仕事や家庭の事情などで保育園への入所を希望し、申請しているにもかかわらず、入所できないで待機を余儀なくされている児童」（コトバンク「待機児童問題」）の問題のことを指す。とくに0から2歳児を預ける保育園が足りず、働く親たちにとって深刻な問題となっている。また、待機児童問題は、図1にあるように、特に東京都において深刻な問題となっている。

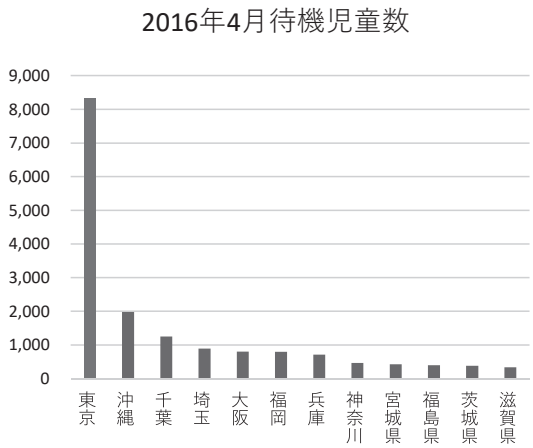

図1　2016年4月　待機児童数（都道府県別）

この厚生労働省が公開している 2016 年 4 月のデータによれば、東京には約 8300 人の待機児童がいる。この状況は、現在もほとんど変わっていない。ただし後述するように、計算方法のトリックによって、この数になっているにすぎず、実質的には、東京都の待機児童は、この三倍以上の数、存在するという指摘がなされている。

■　「人口統計学」の「都市エリア」という観点

東京都知事選挙の場合、投票できる有権者は東京都民であるが、しかし都政の影響を受けるのは、東京都という行政区画内に住む人びとだけではない。

人口統計学では、市町村などといった行政区画単位での地理的境界で人口を算出せず、行政区画よりも、実質的に都市部が形成されている域内を一つの都市エリアとして計測している。実質的に、というのは、労働者が自由に移動できること、建物が連続していることを指す。この定義から「東京-横浜」と標記されているエリアには、千葉、川崎、前橋、相模原、埼玉、宇都宮も含まれている。このエリアが示しているのは、都内への、あるいは、都内からの、通勤圏と解釈できよう。他国と比較して、エリア面積がニューヨークについで二位と大きいが、日本の場合、これだけ広い範囲で交通網、建物が連続している、また、通勤が行われている可能性が高いことが分かる。

この定義からすれば、東京・横浜エリアには、日本の人口、約 1 億 2700 万人（2018 年 11 月現在）の内、約三割の人々が住んでいる、あるいは、働いていることになる。このように、東京・横浜エリアが大きいことは、都知事選挙にも関係してくる。というのは、埼玉県、千葉県、神奈川県といった東京都に隣接する県から都内に通勤している人びとは、一日の多くを東京都内で過ごしているにもかかわらず、投票の権利がないことになるからである。

これについて本稿の分析対象としたTwitterデータ（データの詳細については後述する）を調べたところ、他道府県民のツイートのほとんどが、「～道府県民なので投票権がない」、あるいは「関係ないのだが」、「気になる」といったもの、また

表 1　「世界最大開発都市エリア 2018」
（Demographia “World Urban Areas” studyより）

	都市	国	推計人口（人）	面積（km^2）	人口密度（km^2 あたり）
1	東京-横浜	日本	38,050,000	8,547	4,500
2	ジャカルタ	インドネシア	32,275,000	3,302	9,800
3	デリー、DL-UP-HR	インド	27,280,000	2,202	12,400
4	マニラ	フィリピン	24,650,000	1,787	13,600
5	ソウル-仁川	韓国	24,210,000	2,745	8,800
6	上海、SHG-JS-ZJ	中国	24,115,000	4,015	6,000
7	ムンバイ、MH	インド	23,265,000	881	26,400
8	ニューヨーク、NY-NJ-CT	アメリカ	21,575,000	11,875	1,700
9	北京、BJ-HEB	中国	21,250,000	4,144	5,100
10	サンパウロ	ブラジル	21,100,000	3,043	6,900

「都民のみなさんの良識を信じたい」、「都民のみなさんの良識が問われている」といったツイートがなされ、また、リツイートされていた。また、調べていく中で「埼玉都民」といった言葉が少ないが見られた。この語について説明するツイートがあったので、以下、引用しておく。「いよいよ明日は都知事選か…埼玉都民　住まいは埼玉、勤め先は都内の人々をこう呼ぶらしい。なので投票権はないがとても気になっております。平日は日中起きてる間はほとんど都内にいるからね」。参考までに、東京都と隣接する埼玉県民（都民）、千葉県民（都民）、神奈川県民（都民）という語が入ったツイート数を挙げておく。埼玉県民（5,863）、埼玉都民（50）、神奈川県民（1,168）、神奈川都民（73）、千葉県民（627）、千葉都民（37）。

話しを元に戻そう。先に述べた、有権者の関心の逆転については、東京都世田谷区長（2016 年当時、日本の自治体でもっとも待機児童を抱えていた自治体）である保坂展人が 2016 年 7 月 25 日のブログで次のように書いている。「有権者が重要視する政策の中で、「教育・子ども支援」がトップになったのは、これまでになかったことです。とくに女性層の関心が高いことが特徴です」（強調は引用者による）。また彼は、「待機児童問題」が「都知事選挙で各候補も避けて通れない最優先課題となった」とも述べている。

待機児童問題の当事者である、第一子を出産する母親の全国平均年齢は、厚生労働省が行った 2016 年の人口動態調査によると、30.7 歳であり、父親の第一子誕生時の平均年齢は 32.8 歳である。年齢階級別でみると「30〜39 歳」が 59.2％で最も多く、これについで 20〜29 歳が 34.1％となっている（厚生労働省「人口動態調査」（2016 年）による）。

ここで確認しておきたいのは、この調査で示された、乳児をもつ両親の分布が、先に引用した都知事選についての東京新聞の調査で示された「教育・子育て支援」に最も高い関心を示した 30 代（60.4％）、18 歳〜29 歳（51.6％）の女性たちにほぼ重なることである。

本稿の目的の一つは、Twitter データから、都知事選について、Twitter ユーザーたちは何に関心をもったのか、また、同じ関心を抱いたユーザーたちの間でどのようなやりとりが行われたのか、について調べることである。もちろん、Twitter ユーザーは日本全国に散らばっており、その中から、東京都民であるユーザーだけを抽出することは、様々な観点から難しい。とはいえ、先述したように、人口統計学の定義に基づいた東京都を含む都市エリア内に日本の人口の約三分の一が居住しているということから考えると、その三分の一の人びとが都知事選の結果、何がしかの影響を受ける可能性があると推測できる。また、「待機児童問題」、「介護問題」が、東京だけの問題にとどまらない日本全国で起きている問題であることを考えると、残りの三分の二の人びとにとっても無関心ではいられない部分があると言える。

先に選挙結果を示しておく。主要候補者は、小池百合子（当時 64 歳）、増田寛也（当時 64 歳）、鳥越俊太郎（当時 76 歳）であったが、選挙の結果、小池が約 290 万票（増田：約 170 万票、鳥越：約 130 万票）を獲得し、東京都知事に当選した。

■　都知事選Twitterデータについて

Twitter データは、ユーザーローカル社（東京都港区）の特別な協力を得て、2016 年 7 月 13 日〜8 月 1 日（選挙告示日が 7 月 14 日で投開票日が 7 月 31 日）の間の以下の検索ワードを含む全数データを収集することができた（ただし、もちろん鍵つきアカウントの投稿は入っていない）。検索ワードは次のとおりである。「小池 OR 増田 OR 鳥越 OR 百合子 OR 寛也 OR 俊太郎 OR 都知事選 OR 都知事選挙 OR 知事選挙 OR 知事選」。その結果、以下の数のツイートを csv ファイルに

表２　ツイート数

	総ツイート数
All	4,825,560
RT	3,588,302
OT	1,237,258

て収集することができた。

RTは公式リツイートを（以後、RTデータと呼ぶ）、またOTはオリジナル・ツイートのことを（以後OTデータと呼ぶ）指す（以後、すべて足したデータをAllデータと呼ぶ）。なお、AllデータのcsvファイルのデータTwitterコーパスにおいては、拡散されたツイートの数が全体の74％と多いものとなっている。

Twitterデータは、その性質上、トピックによって、リツイート数のボリュームが異なる。これについては、高（2015）による調査が存在する。Twitterにおけるリツイートの分布について、高（2015：31）は、Twitterより当時サービスとして提供されていたRSS機能を使って、ワイルドカードを用いて日本語のTwitterデータを取得した結果について述べている（ただし高も述べているように取得できるデータ数は制限されている）。ワイルドカードを用いると、一定期間になされたツイートをトピックに関係なく時間順に収集することができる。高によれば、収集された114,932件のうち、リツイートは10,817件と、全体の9.4％を占めていたとのことである。これに対し高が研究対象としていたヘイトスピーチが多い在日コリアンについてのリツイートは44.7％と、後者が拡散されやすい内容であることを指摘している。これを踏まえると、本調査で収集したTwitterコーパスは、高の調査で収集されたコーパス以上に拡散された投稿の多いコーパスであることが分かる。

■　データの前処理（形態素分析、データクリーニング）について

word2vecによって自然言語を単語ベクトル化する前に必要なのが、データの前処理である。ここで、前処理の一つである形態素分析について説明しておきたい。

日本語は、英語などとは異なり、品詞毎に単語が分かれていない。そのため、一つの文章を品詞毎に分解する必要がある。これを形態素分析、あるいは分かち書きと呼ぶ。日本語の形態素分析エンジンとして、最も有名なのはMecabである（Kudo（2013））。またそれとセットにしばしば使われる辞書が、Mecab-ipadic辞書である。

ここで問題になるのは、とりわけSNSテキストデータには、それぞれのプラットフォームに固有のスラングや、多岐にわたるトピックごとに多種多様な語彙群が存在することである。なぜ問題なのかといえば、ipadicは、標準的な辞書レベルの語彙を十分網羅的に含んでいるが、特殊な語彙を欠いているからである。これを解決するために本研究では、Sato（2015）によって、いまもなお定期的に更新され続けている、こうしたネット上のスラングなどを多く含むMecab-ipadic-NEologd辞書を用いた。

とりわけ、ipadic辞書がSNSデータ分析にとって致命的なのは、氏名を一つの単語として認識しないことである。都知事選の場合でいえば、「小池」とカウントされた単語が「小池百合子」なのか共産党議員の「小池晃」なのか、が分からない。つまり、「小池」と「百合子」、「晃」が分解されてカウントされてしまう。さらには、「桜井」とカウントされた単語が、候補者の１人であった元在特会会長の「桜井誠」なのか、一時期候補すると噂された「桜井俊」（アイドルグループ嵐のメンバーの櫻井翔の父親）なのか、保守派論客である「櫻井よしこ」なのか、が分からないということが起きる（これらの氏名はすべて分解されてしまう）。こういったことがクラスタリングで大き

な欠陥となりうる理由は、それらの氏名が、相当異なる文脈に出現する可能性が高い以上、氏名が分解されてしまうと、それぞれの文脈の差異が学習不能になってしまうからである。

NEologdは、こうした問題の多くを解決してくれる（ただし、NEologdも完璧な辞書など存在しないのと同じ意味で、もちろん完璧ではない。例えば、マイナーなアニメの主人公の名前などは、登録されていないため、氏名が分解されてしまう）。

二つの辞書の精度の違いを示すために、先述した、Mecab-ipadic-NEologdで分かち書きしたファイルに加えて、同じデータを、Mecab-ipadic辞書を用いて分かち書きしたファイルを作成した。その上で、Google社がオープンソースで公開しているEmbedding Projector（このツールについては後で詳しく説明する）と呼ばれるウェブUIの検索機能（これによって表示される予測候補単語を見ると、Googleの検索エンジンと同様のアルゴリズムが用いられているように思われる）を用いて、二つのデータを読み込ませ、語彙を比較、確認した。ターゲット単語を「都」としたところ、その予測候補単語として、表3と表4のような結果が出た。ここから、NEologd辞書の語彙がいかに豊富かが分かるだろう。と同時に、クラスタリングを行なうにあたって、辞書の

表3　Mecab-ipadic　39語

也都	都市	都度	都会
都内	都道	都議	大都市
御都合主義	都合	都県	都心
不都合	都留文科大学	宇都	京都大
都々逸	都区	舫都	帝都
都立	都道府県	都民	都政
都	ご都合主義	都留	都議会
都知事	都営	都庁	遷都
州都	都下	宇都宮	新都
好都合	京都	首都	

表4　Mecab-ipadic-NEologd　364語の内の100語

都知事	宇都宮けんじ	都庁職員	東京都政	都道府県知事
都	自民党東京都連	宇都宮大	都市伝説	不都合
都市	都有地	都営住宅	万票宇都宮万票細川万票田母神万票	舛添都政
都政	自民党都連	三大都市	革新都政をつくる会	東京都庁
都知事選	不都合な真実	都留文科大学	東京都港区赤坂	東京都港区
宇都宮	都庁	都立高校	東京都議会議員選挙	多摩都市モノレール
都知事選挙	京都大学	東京都議会	石原都政	都下
首都	都市部	都議会民進党	都道府県議会	京都市長選挙
都議会	日本の首都	京都	首都直下地震	宇都宮市
都合	美濃部都政	東京都選挙区	都知	都政の刷新
都民	都内	舛添都知事	不都合な事実	大都会
東京都知事選	消滅可能性都市	都議会議員選挙	新都	い都
東京都	東京都内	消滅都市	都度	首都大
宇都宮健児	石原都知事	首都圏	都会	東京都心
東京都知事選挙	都立	首都高	都営	東京都議選
帝都	都庁前	都心	大阪都構想	京都府民
東京都知事	都市外交	青島都政	副首都構想	東京都議会議員補欠選挙
都議	大都市	東京都議会議員	好都合	遷都
東京都民	都議会議員	都道	仁都	都議会選挙
都道府県	都議選	宇都	小池都知事	ご都合主義

語彙の豊富さがいかに重要かも容易に推測されるだろう。表 4 に含まれる単語から例を挙げると、NEologd で「青島都政」、「石原都政」、「舛添都政」と一語で認識できているものが、ipadic だと、この表を見る限り「青島（石原、舛添）／都政」と分かれてしまう。これだと、ツイート文に「都政」という単語がある場合、どの知事の時のものを指しているのかをアルゴリズムが学習できなくなる。このことは、クラスタリングの精度にきわめて深く関係する。

また、前処理として重要なデータクリーニングとして、以下のことを行った。すなわち、Twitter データには、URL や記号など、自然言語処理を行うには、不要な情報が多く含まれているため、ここでは、アルファベット、ローマ数字、記号をすべて除去した。そのうえで、上の形態素分析を行って、分かち書きファイルをつくった。とはいえ、それでもいくつかの記号や不要な文字が残ってしまう。しかしこれについては後述するように、ビッグデータ分析においては、むしろある程度ノイズがあったほうが精度が上がるため、そのまま残した。

分かち書きファイルを作った上で、それぞれのデータの単語の分布を見るために、単語数をカウントした。カウントにあたっては、EKWords というフリーソフトを用いた。その理由は、このソフトが、単語のカウントを行う際、ひらがなだけの単語、また、一文字だけの語を除去してくれるからである。ひらがなだけの単語は、助詞や接続詞といった情報量の少ない単語が多い。ただし「ゆりこ」（「ひろや」という単語は 1 語もなかった）のような、情報量のある単語が除去されてしまうが、ここでは、単語の分布を見るためだけに、便宜上、このソフトを使った（つまり、最終的にデータ分析を行うために使用する三つのアルゴリズム、word2vec、PCA、t-SNE に学習させるデータは、元の、ひらがな、一文字の単語を含むデータを使用した。なお、その元データの単語出現数の最も多かったものを順に示しておく。「の（10,176,529 回）」、「に（5,967,698 回）、「は（5,606,293 回）」、以下、「を」、「が」、「て」、「た」、「で」、「と」、「し」、「も」、「ない」、「氏」、「な」と続いた後に「 鳥越（1,462,436 回）」がくる」。

EKwords を使ってカウントしたところ、表 5 のようになった。All が約 15 万語（元データは 199,287）、RT が約 9 万 5 千語（元データは 123,192）、OT が約 14 万語（元データは 187,394）となっており、同じツイートが大量に拡散されるリツイートの場合、その分、単語数が少なくなっていることが分かる。なお、ここで総単語数というのは、単語の種類の総数ということであって、それぞれの単語の出現回数を合わせた総数ではない。なお、後者の意味での総数（つまりノイズを除去していない単語数）をカウントしたところ、All データが 177,439,525 語、RT データが 143,181,473 語、OT データが 34,258,052 語となった。

表 5　単語数

	総ツイート数	総単語数
All	4,825,560	154,439
RT	3,588,302	95,856
OT	1,237,258	145,035

■　三つのトピックにおける、それぞれの特徴語の出現回数について

次に、先述した有権者の関心が高かった二つの問題、「教育・子育て支援」、「医療・福祉」に関連する特徴ある単語を、それぞれ三つずつ選んだ。すなわち、「待機児童」、「保育」、「子育て」と、「老人」、「介護」、「高齢者」である。これらに加えて、都民以外の日本全国の人びとの関心となったと思われる「オリンピック」という単語も含めた。「オリンピック（五輪、パラリンピック）」は、単語出現回数ランクは 78 位で、出現回数は 78,006 回だった。表 6 の単語と比べて、最も出

現回数が多かった。これについては、今回の都知事選が、2020年開催予定の東京オリンピック開催時の知事を決めるということでもあったため、東京都民以外の人々の関心の対象にもなったと考えられる。なお、表6の「単語出現回数ランク」とは、先述した単語の種類の総数、154,359の単語の中で出現回数がどれだけあったかについての、1位から154,439位の間のランクを示す数となる。

これらの出現回数を見ると、総じて、待機児童など子育て支援のほうが、介護問題より投稿数が多いことが分かる。このことから先述した新聞社による調査、すなわち、「子育て支援」のほうが「介護・福祉」よりも関心が高いという結果と、Twitterコーパスでの投稿内容が、概ね一致しているといえよう。

それぞれの単語において興味深い特徴としては、いずれも総じてリツイートでの出現回数が多いことである。これは、特定のツイートが、大規模に拡散されていることを意味する。つまり、拡散されないツイート＝OTよりも、拡散されたツイートが多いことを意味する。

冒頭で本稿の目的を二つ挙げておいた。再度確認しておくと、都知事選のツイートから、Twitterユーザーが何に関心を持ち、どのようなやりとりがあったのかを調べること、と、Pythonを用いた機械学習によるクラスタリングの精度を確認することである。ここで実はもう一つ目的があることについて述べておかねばならない。それは、従来の、新聞社や調査会社などによるアンケート調査などからは収集できないような人びとの声を、TwitterというSNSから収集できるのではないか、という試みを行なうことである。これについてマイヤー＝ショーンベルガーとクキエが、次のように述べている。

> ただの戯言や無駄口にすぎないと見る向きもある。あながち間違いではない。しかし、ツイッターが実現しているのは、人々の考え、心情、反応のデータ化であり、これまで集めようがなかった情報なのだ。（マイヤー＝ショーンベルガー、クキエ、2013=2013：143）

Twitterに特徴的なコメントといったものはどういうものなのか。これについては後でみることにするが、その前にデータの分析手法について説明しておく。

表6　各トピックの特徴語の単語出現回数ランクと単語出現回数

		単語出現回数ランキング	単語出現回数			単語出現回数ランキング	単語出現回数
待機児童	All	187	43,794	老人	All	604	19,347
	RT	170	38,836		RT	683	13,977
	OT	390	4,958		OT	361	5,370
保育	All	169	47,573	介護	All	383	26,955
	RT	146	44,417		RT	350	23,477
	OT	721	3,156		OT	641	3,478
子育て	All	398	26,190	高齢者	All	804	14,947
	RT	366	22,690		RT	749	12,594
	OT	629	3,500		OT	1,014	2,353

■　「自然言語処理」とword2vec

自然言語処理では、近年、単語ベクトル表現が用いられている。その中でも、「単語分散表現は、テキスト中の単語を数値ベクトルに変換する方法のひとつで」、「テキストを数値ベクトルに変換することで、数値ベクトルの入力を必要とする機械学習アルゴリズムでテキストを解析できるようにな」（グッリ、パル、2017=2018：137）る。この分散表現によって、「ある単語の意味を、その単語の文脈中に出現する別の単語との関係において捉える」ことができるようになる。言い換えれば、ある文脈に出てくる単語は、それと同じ（か、あるいはそれと似たような）文脈に出てくる単語と意味が近い（類似している）確率が高い。つまり、似た文脈にある単語は、意味が近いということになる。

単語のベクトル化の手法としては、TF-IDF、潜在意味解析（LSA）、トピックモデル[3]などが存在する。これらについては、グッリ、パル（2017=2018：137-138）が次のように述べている。「これらの手法は文章中の単語に着目しており、単語それ自体の意味を捉えようとする単語分散表現とは異なる手法」になる。

単語分散表現として、2018年12月時点で有名なのは、先述したように、2013年当時、Googleの研究所にいたTomas Mikolvらによって考案された、word2vecである。

これによって、例えば、「フランス」→「パリ」、「日本」→「？」の「？」にあたるものが「東京」だということを算出することができるようになった。意味そのものを理解しているわけではないが、与えられたデータが大きいほど、互いの単語の類似度を多次元、例えば300次元などで関係づけることができるため、近い単語のグループを見つけることが可能となった、すなわち、クラスタリングの精度が格段に上がったと言われる。

■　「べき乗」分布

word2vecが自然言語処理の手法の中で、SNSデータ分析に向いていると考える理由は、いくつかある。一つは、SNSデータの分布が、「べき乗」（「ロングテール」）分布なっていることを挙げることができる。後で述べるように、word2vecは、べき乗分布になっているデータを学習することに適している。そこでまず、べき乗分布について簡単に説明しておきたい。

標準的なアンケート調査などで収集されたサンプリングデータは、一般的に正規分布になることが前提とされている。その場合、平均値というものが全体の分布を把握する上で重要となる。これに対し、ロングテール分布では平均値はあまり意味がない。これについては、平均所得についてすでに言われていることである。すなわち、少数の富裕層が平均を、上へと釣り上げているので、平均値がその分上がってしまう。そのため、中央値のほうがより適正とされる。しかしロングテール分布の場合、中央値もあまり意味をなさない。なぜなら一方で、極端に出現回数が多い単語がいくつか存在するからであり、他方で出現回数がわずかしかない単語が極端に多いからである。

自然言語処理において重要な前処理で必要とされる作業として、情報をほとんどもたない単語、例えば、日本語であれば、助詞（「である」、「だ」など）や接続詞（「そして」、「つまり」など）を除去することを挙げることができる。これを避けるために、自然言語処理の先行研究でしばしば見られるのは、名詞、形容詞、動詞などのみ取り出して、頻出度が高い単語に重みをもたせることである。例えば、小説や歴代内閣総理大臣所信表明演説のテキストファイルを用いて、名詞、形容詞、動詞などの出現回数を見るなどである。こうしたデータの場合、出現回数の多い名詞や形容詞、動詞に重みをつけることには意味がある。例えば、複数の作家の小説をデータ化し、重みづけの作業をすると、小説家によって使用する語彙の違いを

見ることができ、そこから小説家の特徴を引き出すことができるかもしれない。

これに対して、本稿で用いたTwitterデータの場合、出現回数が多いからといって、その単語に情報量があるとは限らない。むしろ、あまりにも多く出現する頻出語は、情報量をもたないものがほとんどである。このことは図2に表れている。これはAllデータに含まれる154,439の単語の分布をツリーマップで示したものである。

この図から、Twitterコーパスの分布が、極端な「べき乗（ロングテール）」分布になっていることが分かる。上から五から七つ目の三単語（「小池」、「選挙」、「候補」）が全体の中央値にあたる。最も頻度の高い一つ目の「鳥越」の出現回数は1,462,436回であるが、上位にきている頻出語は情報量の少ない単語であり（都知事選に関するコーパスであるため、どんなトピックの文章にも出現するという意味で情報量が少ない）、また分布の点から言っても、中央値ですら全く意味がないことがわかる（上部右にある色別出現回数の左の値が0となっているのは、ファイルの行の最後に余計な改行が一つ入っており、その改行箇所に単語が入っていないためである）。ロングテール分布の語彙を使えば、左端のショートヘッドにあたる数語以外は、X軸の右へとひたすら1に近づく、ほとんど平行線に近い下降線（ロングテール）が続く。

■ word2vecについての説明

それでは、word2vecアルゴリズムが、どのような点で、べき乗分布のデータを処理するのに適したものなのか。word2vecは、データの学習にあたって、いくつかのオプション[4]が設けられ

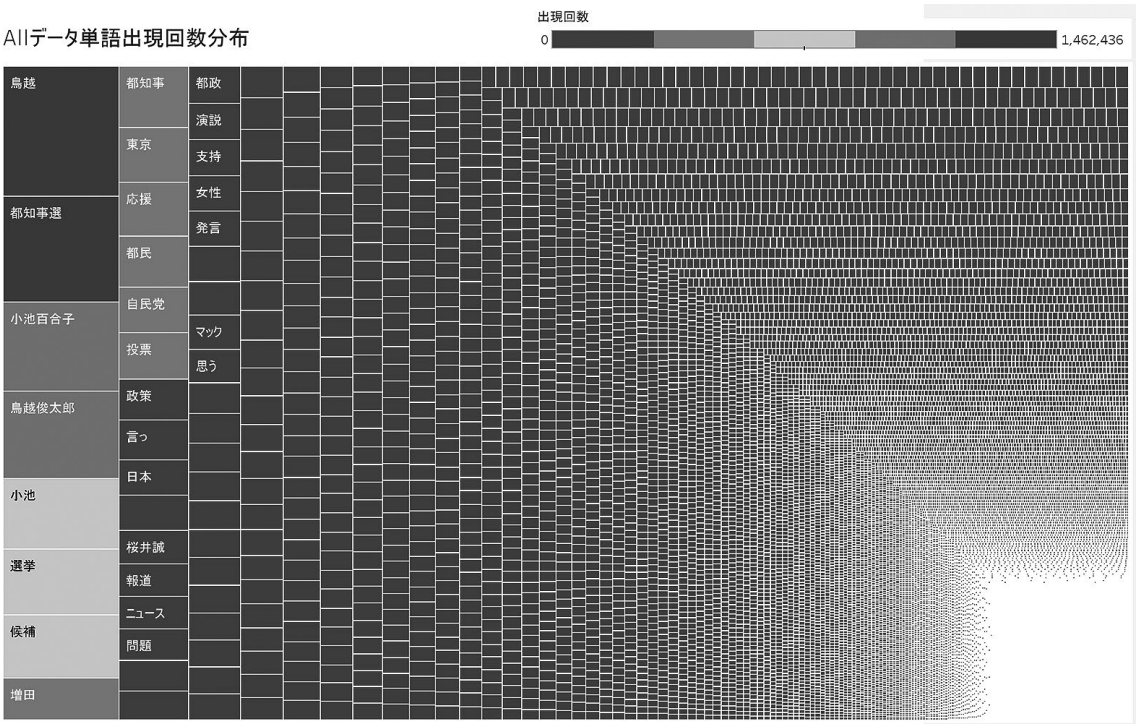

図2　Allデータ単語出現回数分布（tableauにて作成）

ており、パラメータの数値を変えることによって、学習精度を改善することができるようになっている。高頻出語の除去については、sampleオプションで調整可能である。このオプションは、単語の出現についての閾値を設定するものであり、非常に高い頻度で現れるものを、ランダムにダウンサンプリングしてくれる。これによって、ロングテール分布の左端のショートヘッドにあたる、あまり情報量のない単語を一定程度減らすことができる。有効範囲は0,1e−5となっており、本稿では、0.001とした。一方、ロングテール部分にあたる、極めてまれにしか出現しない単語については、min_countオプションを使って除去することができる。ここでは、5に設定した。これによって、5回未満しか出現していない単語を、学習する前に除去することができる。

また、冒頭に述べたように、大きな自然言語データを機械学習するにあたって学習を高速度で行うことは、非常に重要なことである。word2vecでは、高速化するためのオプションとして、「ネガティブ・サンプリング」という学習手法を利用できる。これは、ある単語がおかれている文脈、またその単語といっしょに出てくる単語については関連性の確率を高くしつつ、その単語とは無関係な文章からランダムに選ばれた単語については確率を低くするように学習していく手法である。元論文によれば、学習させるデータの量が大きければ、2〜5個の関連性のない単語を選ぶだけで十分な結果が得られると書かれている（Mikolov et al.（2013）“Distributed Representations of Words and Phrases and their Compositionality”）。推奨値は5から20となっているが、本稿では5とした。この手法は、後でも述べるが、ビッグデータにおいては“ゴミ”データが、分析精度を上げてくれるという傾向を利用したものである。

こうして、word2vecアルゴリズムによって、互いの類似度が300次元（一つの単語が300の要素から成っている）で数値化された単語ベクトル群のコーパスのモデルファイルを作成し、学習結果として、98,417の単語が残った。

■ Google社によって提供されているEmbedding Projectorの説明

先述したように、元データは、約1億7千語、480万行のツイートからなっており、単語の種類の数で言えば、約20万語の単語があった。これが上の分析処理の結果、約半分にあたる98,417単語残った。ただし後で述べるように、まだ、98,417語だと可視化するにあたって物理的に多すぎるという問題がある。

本稿では、冒頭で述べたクラスタリングを行うために、Embedding Projectorを用いた。Embedding ProjectorはGoogle社がオープンソースで提供しているディープラーニングフレームワークであるTensorFlowのパワフルな可視化ツールであるTensorBoardの、スタンドアローン版ウェブUIである。これだと、複雑なPythonスクリプトを必要とするTensorFlowを使うことなく、Embedding Projectorに必要なデータさえ用意できれば、TensorBoardと同じ可視化ツールを使うことができる（なお、TensorBoardもブラウザで動くものであり、TensorFlowで生成したモデルファイル（正確には学習されたログデータ）を受け取るだけで、それ以降の可視化するにあたっての処理の速度はEmbedding Projectorと変わらないことを確認している）。

これを用いる理由としては、これが、三次元でTwitterコーパス空間を可視化してくれるだけでなく、クラスタリングの学習のプロセスそのものを可視化してくれるため、どのようなクラスタが出てきているか、学習中に目視で確認でき、クラスタが出てきた時点で学習を止めることができるからである（クラスタを三次元で可視化することの大きなメリットについては、このツールを作成したチームメンバー、Smilkov, Danielらによるペーパー“Embedding Projector: Interactive Vi-

sualization and Interpretation of Embeddings.”を参照のこと）。またこのツールには、学習中に、インタラクティヴに三次元のコーパス空間を自由に回転させることができ、また、ズームイン、ズームアウトできるため、クラスタを発見しやすいという大きなメリットがある（このツールのインタラクティヴ性については、言葉では説明しにくいため、先と同じく、これを作ったチームメンバーによる説明動画（Google Developers“A.I. Experiments: Visualizing High-Dimensional Space”を参照のこと）。そしてまさにこのメリットこそ、冒頭に述べたような、どのようなクラスタができるか、われわれは前もって知らされないといった、機械学習や深層学習の欠点としてしばしば指摘される、学習の「ブラックボックス」性を、一定程度、取り除くことにつながると考えることができる。

ただし、この手法にも欠点が存在する。それは、ヴィジュアライゼーションツールを用いてクラスタリングを行うことに不可避的に伴う欠点である。その欠点とは、ブラウザの画面の中で可視化させながらクラスタを見つけるため、あまりにも多い単語ラベルが出てきてしまうと、三次元空間が単語ラベルで覆いつくされてしまい、クラスタを目視で確認できない、という物理的な制約である。

しかしながら、おそらくこうしたことを踏まえた設計がEmbedding Projectorにはなされている。後で述べるように、Embedding Projectorは、最初に、単語ベクトルファイルを読み込む際に、主成分分析（PCA（Principle Component Analysis））で処理を行い、三次元（あるいは二次元）でコーパスのグローバル空間を可視化してくれる。ただしその際、Embedding ProjectorのPCAは、50,000以上の単語を含むデータを、50,000語にダウンサンプリングし、300次元を200次元にダウンサイジングした結果を表示している。これは、ブラウザで表示させるための適正な数を考慮に入れているためだと思われる。

さて、可視化に伴うこの物理的限界については、図3を見れば端的に理解可能だと思われる。単語ラベルが多すぎて、クラスタが出ないだけでなく、語の判別そのものができない。なお、これは、word2vecを使ってつくったAllデータのモデルファイルを、Pythonでよく使われるライブラリであるscikit-learnとmatplotlibを使って非線形次元圧縮手法の一つt-SNEアルゴリズムで次元圧縮を行い（詳しくは後述する）、二次元で描画した結果である。

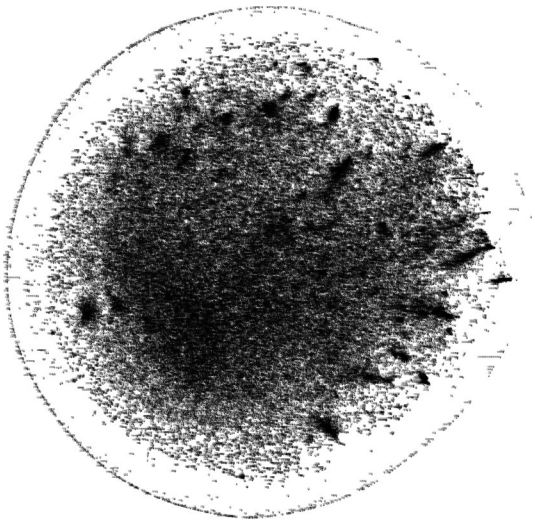

図3　t-SNEを使ってAllデータを二次元に圧縮しクラスタリングした結果

こうしたヴィジュアライゼーションの限界を避けるために、Embedding Projectorでは、ターゲット単語に近い単語（近傍値が小さいほど単語間の距離が近いことを意味する）を最大1,000語、表示するよう設計されている。そして、ターゲット単語に近い1,000語（ターゲット単語を含めて1,001）を、近さが1,002語以降の単語群、つまり、全体のコーパスから隔離した上で、その1,000語のコーパス内で、次元圧縮アルゴリズムであるt-SNEアルゴリズム（後述する）を使って、クラスタリング学習を行うことができる。

■　グローバル空間：PCAアルゴリズムを使った次元圧縮

それでは改めて、Embedding Projectorで行うヴィジュアライゼーションの手順について説明したい。Embedding Projectorにデータをロードさせると、主成分分析（PCA）をした結果を可視化してくれる。PCAとはデータの分散が最大になるような軸（主成分）を探してくれる次元圧縮アルゴリズムである。300次元ある単語ベクトルを、分散が最大となる上位10個の主成分を計算し、そこから三つ（三次元）あるいは二つ（二次元）選ぶことができる（詳しくはSmilkov, Danielら、“Embedding Projector: Interactive Visualization and Interpretation of Embeddings.”を参照のこと）。三次元の場合だと、単語ベクトルを、例えば、第一主成分軸：X軸、第二主成分：Y軸、第三主成分軸：Z軸を中心として、配置してくれる。Embedding Projectorで使用可能な主成分分析の描画の優れたメリットは、これによってデータの全体像（グローバル空間）がつかめることである。

まずは、PCAの学習結果の、都知事選Twitterコーパスの全体をみておきたい。先述した都知事選について言及したツイートのうち、直接的には都知事選に関係がないと思われる投稿が相当数、存在していることを確認できた。ただし、ビッグデータに不可避的に含まれる、こうした「ノイズ」データについては、マイヤー＝ショーンベルガーらが次のように述べている。ビッグデータをコンピュータにトレーニングさせるにあたって、むしろ「“ゴミ”も欠かせない」（マイヤー＝ショーンベルガー、クキエ、2013=2013：65）。なぜなら、データ量が多いほど、“ゴミ”がむしろ分析の精度を上げる[5]からである。

図4は、ブラウザで開いたEmbedding ProjectorのウェブUIである。右端にターゲット単語に近い近傍語リストが下に向かって並んでいる。ここに表示されているのは、98,417の単語からなる全体を、PCAアルゴリズムを使って、「ポケモン」をターゲット単語にして二次元で可視化させた結果である。グローバル空間の右中ほどから上

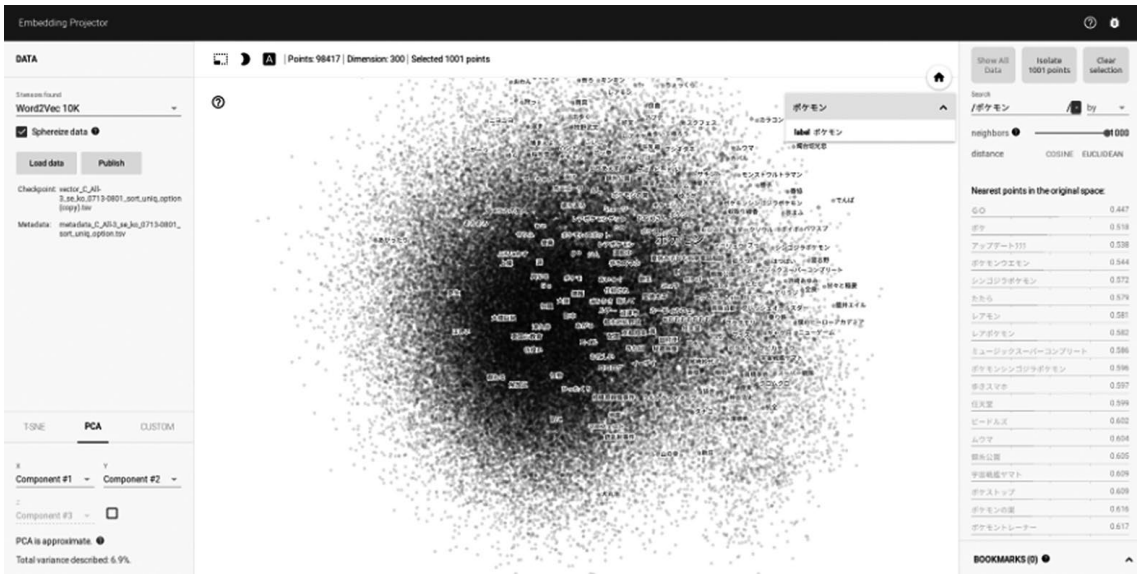

図4　Embedding Projector　「ポケモン」クラスタ

部にかけて、この単語に近い上位 1,000 の近傍語（Embedding Projector の仕様上、検索語に近い単語を表示できるのは最大 1,000 語である）がオレンジ色のバブルで表示されている。

ポケモン GO は日本では 2016 年 7 月 22 日にサービスが開始された。サービス開始時から、日本でも爆発的に人気が出て利用された。なお、この単語を含むツイートは、All データの 4,825,560 ツイート中、59,011 ツイートあった。7 月の選挙期間中、多くの人びとがこのゲームをプレイしていたが、ここでなぜポケモン関係の単語が入ってしまったか、All データで確認したところ、ツイートの内容としては「選挙に行くついでにポケモンゲットした」、あるいは、「投票所にモジュールをさしておいたから、みんなボールをとっていってね」といったものが多かった。他には、日本全国で放送されているある人気テレビバラエティ番組で、都知事選の話題より、ポケモンの話題のほうが大きく取り上げられていたことに対して、東京都民ではない日本人が、都知事選に関心がないのは当たり前だ、と指摘するものなどもあった。

なお本稿では、こうした直接、選挙や候補者の公約に関係ないツイートを排除しなかった。理由の一つは、こういった一部の直接ないと判断されたツイートを削除すると、コーパス全体の分布が歪んでしまうからである。何が直接関係があり、直接関係がないのかといった判断を行うことも、データの量が膨大になるほど、困難になる。一方、候補者についてツイートしているものであっても、選挙、あるいは政治的関心といった観点からすれば、無関係だと判断できるものも数多く存在する（例えば、候補者の容姿を揶揄したものなど）。しかし本稿では、検索単語で引っかかったものすべてを分析対象として残すこととした。

このことについて、マイヤー＝ショーンベルガー、クキエは、ビッグデータ分析にあたっては「乱雑なデータ」を受け入れる必要があることを何度も強調している。少し長いが以下に引用する（この観点から、数が非常に少ないことを確認した上で、本稿で分析する Twitter データのうち、あるツイートへのリプライ、また、bot ツイートを除外しなかった）。

> これまで標本による分析を担ってきた人々は、［･･･］乱雑さは受け入れがたいはずだ。統計学者は、標本収集の際に誤り率を抑えるための対策を総動員し、構造的な偏りが潜んでいないか標本を検証したうえで分析結果を発表している。標本の収集は、特別な訓練を受けた専門家が正しい手順に沿って実施する。たとえ測定値の数が限られていても、誤りを減らす対策には費用がかかる。
>
> ところが、すべてのデータを収集するとなれば、そのような対策はまず実現不可能だ。費用がかかりすぎるうえ、膨大なデータに対し、厳格な収集基準を一貫して維持することなどまず無理だ。人手に頼らない方法でも解決は難しい。
>
> ビッグデータの世界に足を踏み入れるためには、「正確＝メリット」という考え方を改める必要がある。［･･･］情報が少ない時代には、1 つひとつの測定値が分析結果を大きく左右したから、分析を歪めないように細心の注意を払う必要があった。
>
> 今、我々が暮らしている世界は、そんな“情報飢餓社会”ではない。目の前で起こっている現象のほんの一部だけでなく、大部分あるいは全体を取り込んだ包括的なデータ集合が手に入るなら、個々の測定値の良し悪しにいちいち悩む必要もない。（マイヤー＝ショーンベルガー、クキエ、2013 = 2013：66–67）

実際、図 5 にあるように、「ポケモン」のクラスタは、主成分の軸から、右にずれたところに単語が密集していることがわかる。軸は二次元であるため、XY 軸が縦と横に引かれており、中心が

0 になる。つまり、中心から遠いほど選挙というトピックから遠いことになる。これは「ポケモン」の類似単語群が、選挙の文脈や単語とは遠いものであることを、word2vec、PCA アルゴリズムが学習できていることを意味している。このことが、冒頭に述べた本稿の目的の二つ目にかかわることをここで改めて確認しておきたい。すなわち、ビッグデータに対し極力、人間の判断を入れない状態で機械学習を行うことによって、どこまでクラスタリングの精度が出るかを確認するという目的のことである。

■ ローカル空間：t-SNE アルゴリズムを使ったクラスタリング

非線形次元圧縮手法である t-SNE（t-distributed Stochastic Neighbor Embedding（t 分布型確率的近傍埋め込み））アルゴリズムは、PCA とは逆に、コーパスのグローバル空間内での位置を犠牲にして、類似した単語を互いに近づけ、同時に、類似しない単語を互いに遠ざける。これは、コーパス空間内で、複数の、似た単語のグループ（＝クラスタ）のローカル性を重視するよう学習を行なうことを意味する。本稿では、これを冒頭に述べた、クラスタリングを行うアルゴリズムと考え、これを使ってクラスタを特定することにした。

先述したように、Embedding Projector では、データをロードする際に PCA アルゴリズムによる次元圧縮が行われるが、t-SNE を考案した元論文著者の van der Maaten と Hinton 自身、PCA で前処理を行った上で、t-SNE を使用することを推奨している。つまり、Embedding Projector は、この推奨に従った仕様になっている。

さて、Embedding Projector 上では t-SNE の学習をリアルタイムで可視化してくれる。それと同時に、先に述べたように、学習中であってもインタラクティヴに三次元空間を左右上下自由に回転させることができ、また、ズームイン、ズームアウトすることによって、どの辺りにクラスタが出現しつつあるかを、直感的に目視で判断することができる。

しかしながら、非線形アルゴリズムであるため、クラスタがどれだけあるか、どこからどこまでが一つのクラスタなのかを厳密に確定することはできない（とはいえ、この非線形の柔軟な性質こそが、クラスタリングの高い精度を生んでいるわけだが）。ただし、t-SNE には、クラスタの大きさを変えるオプションが存在している。それが Perplexity である。t-SNE 元論文では「Perplexity は、近傍語の有効数を決める滑らかな尺度として解釈することができる。」と書かれてある。加えて、Embedding Projector では、Perplexity について、以下のようなヘルプメモがつけられている。「最も適正な Perplexity の値は、データの密度に左右される。大まかに言えば、より大きな密度をもったデータは、Perplexity の値を大きくする必要がある。」。より分かりやすく言えば、Perplexity の値を変えることによって、クラスタの大きさを変えることができる。値が大きいほど、クラスタが大きくなり、小さいほどクラスタは小さくなる。元論文同様、ここでも Perplexity は、5 から 50 が推奨値とされている。なお、Embedding Projector では、ロードさせたデータの密度によって、自動的に、Perplexity を算出してくれる。本論文では、この自動設定にしたがった。1,000 語（繰り返しになるが、Embedding Projector の仕様上、ターゲット単語に近い単語（近傍語）を表示できるのは最大 1,000 語である）であれば、8 が設定される。また、iteration（学習回数）については、筆者が、都知事選データを含む様々なデータを使って何度も学習してきた結果、2,000 回ほどでクラスタが収束してくることを経験的に確認してきたため、2,000 回とした。

Embedding Projector の画面の右端に縦に並んでいる単語と数値は、検索語の近傍語を近さ順にリストされたものである。単語間の近さを示す近傍値は、主成分分析がなされた主成分 XYZ 軸を基軸とするオリジナル空間での近傍値である。

t-SNEでは、このグローバル空間での位置、それぞれの単語間の近さを無視して、word2vecで生成されたモデルファイルに基づいて、類似する単語を学習しなおした上で、クラスタをつくる。

■　1.「待機児童」クラスタ

それではここからは、t-SNEを使ってクラスタリングを行った結果をみていく。

図5は、ターゲット単語を「待機児童」で登録し、待機児童に近い単語1,000語を、グローバル空間から切り離した上で、2,000回学習させた結果、収束した「待機児童」のローカル空間の分布全体である。

空間は三次元であり、奥行もある程度深くなっているため、奥のほうの単語まですべてを表示さ

図5　「待機児童」ローカル空間

せることができないことが分かる。これもEmbedding Projectorの仕様によるもので、自動的に手前にあるバブルに付されている単語ほど濃い字で表示され、奥にあるものは、奥まっていくほど薄く表示されるようになっている。

また、このローカル空間に表示されている単語を見ると分かるように、先述した、助詞や接続詞といったものは完全に除去され、また「ポケモン」のようなTwitterデータに含まれていた、待機児童問題とは直接的に無関係な単語が排除されていると同時に、「待機児童」に関係がある単語と思われるものが残っていることも確認できる。このローカル空間のうち、「待機児童」という単語ラベルの左下にあるクラスタを拡大したのが、図６である。

ここは、保育士の待遇が悪いこと（給料が低い）に関係するクラスタである。

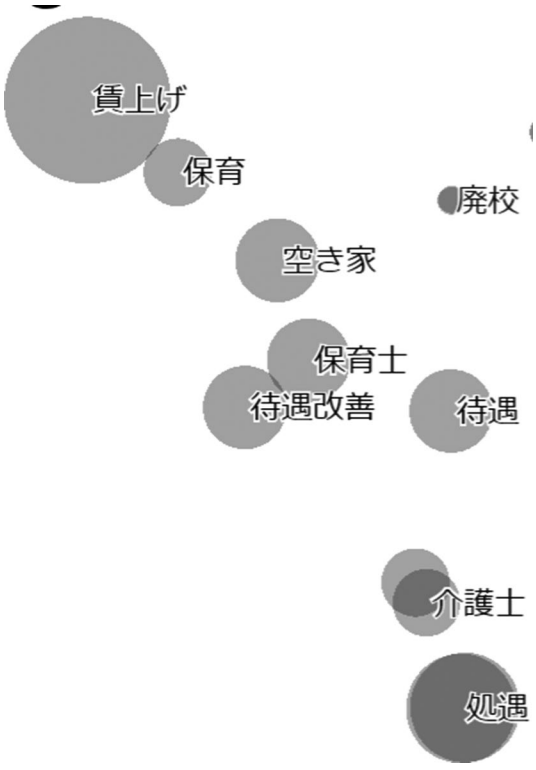

図６　待機児童クラスタ１「保育士・待遇改善」

これに対して、改善を求めるツイートが多くみられた。これは「事実」についての指摘であり、客観的な内容だといえる。また、小池百合子が公約として掲げていた、空き家や廃校を利用することについての単語も見える。

ここで、「介護士」の単語が近傍語として出ているが、後でも述べるように、待機児童問題と介護問題は、共通するところがいくつかある。すなわち、保育士と介護士いずれも不足しており、これらの職業いずれもが待遇が悪いこと、また、施設が足りていないこと、などである。このため、待機児童と介護のローカル空間は重なっている箇所がいくつかあり、このクラスタはその一つである。補足しておくと、特定の単語を登録し、t-SNEで学習させてできるローカル空間は、98,417単語から成る同じ一つのグローバル空間から、当該の単語の近傍語1000語を引き抜いて、生成されるものであるため、このクラスタのように、ローカル空間の間で重なり合う箇所が出てきうる。

次に、最上部にあるクラスタを拡大したものが図７である。

ここには、似たような表現が、「ギュウギュウ」、「詰め込める」、「つめこみ」、「詰め込ん」など、複数あることが分かるが、これは、似たようなツイートが、少なくともその数の分だけ存在し、さらにそれらがリツイートされていることを示している。その一つである単語「詰め込」が入ったツイートをカウントしたところ11,111存在しており、類似するツイートが相当数多いことが推測できる。加えて言えば、この推測は、類似する語を一か所に引き寄せようとするt-SNEアルゴリズムのパフォーマンスがあるからこそ、成り立ちうる。

ツイートの内容についていえば、ほとんどすべてが小池百合子の待機児童を減らすために提案した、保育園延べ床面積あたりの児童数を増やすという規制緩和論への批判だった。小池百合子選挙

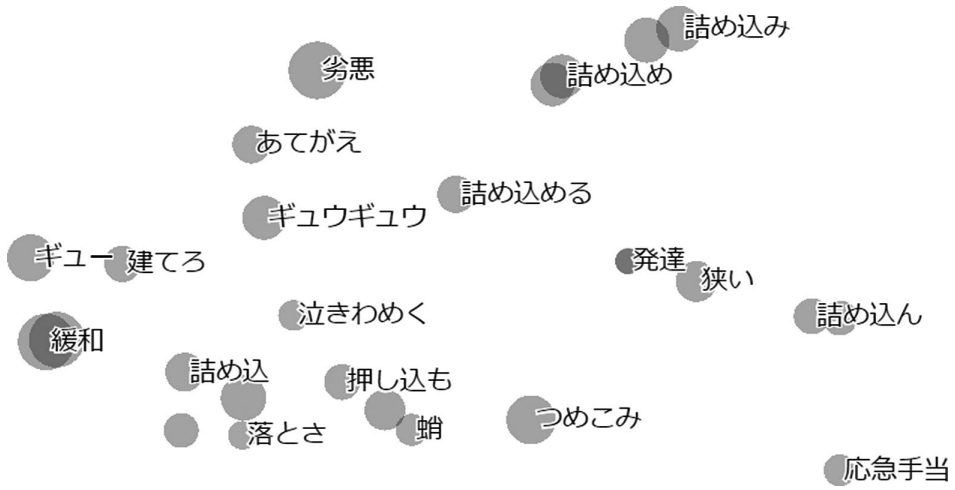

図７　「待機児童」クラスタ２「詰め込み・ギュウギュウ」

事務所のウェブサイトに示された公約には「『待機児童ゼロ』を目標に保育所の受け入れ年齢、広さ制限などの規制を見直す」という記述はあるが、広さ制限の規制見直しについて具体的な数値は示されていない。しかしTwitter上では、保育所の教室が「ギュウギュウ」「詰め」になるかのように書かれた投稿がほとんどだった。また、「ギュウギュウ」については、以下の日刊ゲンダイのウェブ記事「言動を知るほど危うい　小池百合子の『子育て・教育論』」2016年7月29日にその表現が使われており、これをリツイートしたものが1,461あった。

> 園児１人あたりの広さ制限を緩和し、保育所の面積を変えずギュウギュウに園児を詰め込む考えです。子を持つ親が望む認可保育所の増設とは逆行しています。（ジャーナリスト・横田一氏）

なお、選挙期間中の「待機児童」、「保育」、「子育て」に関係する朝日新聞と読売新聞の記事すべてをテキストファイルにして、「ギュ」、「ぎゅ」で検索をかけたところ、一件もヒットしなかった。

ここでのクラスタから推測しうることは、公約について、正しい事実に基づかない誇大表現的な解釈を行った投稿が、その公約を提案した候補者への批判として相当数、拡散されたということである。

■　2.「子育て」クラスタ

図8は、ターゲット単語を「子育て」と設定し、この単語に近い1,000語をそれ以外の単語から切り離し、2,000回学習させて収束してできたローカル空間である。

特徴的なクラスタとして、「子育て」の単語ラベルのあるあたりのクラスタを拡大したものが図9である。

「親学」で検索したところ、16,701ツイートあった。「親学推進協会」とは、2012年に発足した超党派の議員連盟のことを指す。ツイートはほとんどがリツイート（15,889）であり、その内容は小池百合子がこの議員連盟に入っていたことを批判するものだった。「発達障害」や、「日本の伝統」、「下村」という単語が出てきているのは、以下の出来事についてのツイートが多かったためで

図８　「子育て」ローカル空間

ある。

超党派の国会議員でつくる「親学推進議員連盟」が５月末「発達障害を予防する伝統的子育て」をテーマに勉強会を開いたことが分かった。配布資料には［･･･］「発達障害児は笑わない」「予防は可能」などの記述もあった。発達障害は子育ての問題だと受け取られかねない内容に、関係者の抗議が殺到、議連側は最終的に陳謝した。勉強会は５月 25 日、衆院第２議員会館で開かれ［･･･］た。翌 26 日に議連事務局長の下村博文衆院議員（自民）がブログで内容を報告すると、発達障害者や親から［･･･］「発達障害について学んでほしい」と抗議が殺到した。下村議員は今月２日のブログで「誤ったメッセージになった」と陳謝。（「親学議連：『発達障害、予防は可能』　勉強会内容をブログで報告、抗議殺到し陳謝」毎日新聞　2012 年６月 12 日、朝刊）

認定NPO法人フローレンス代表の駒崎弘樹、

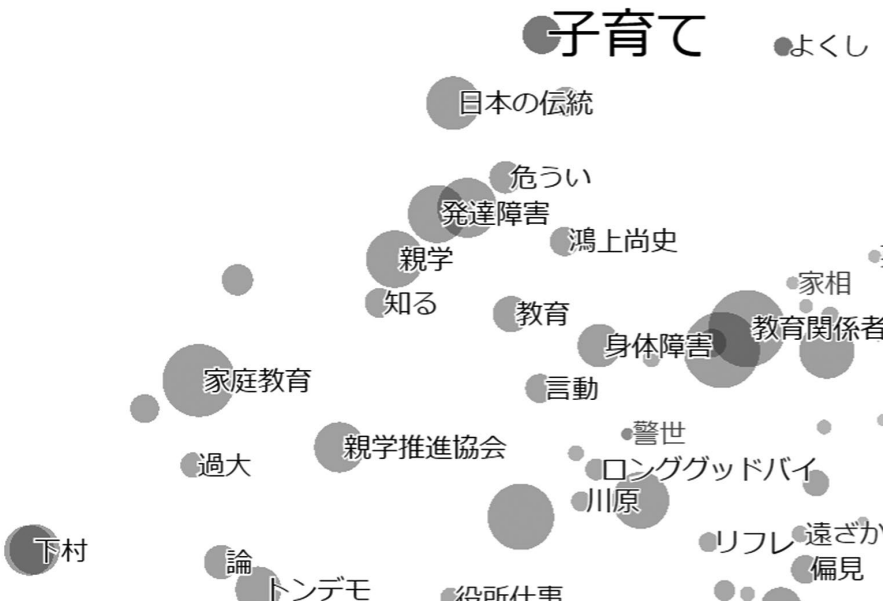

図９　「子育て」クラスタ１「日本の伝統・親学・発達障害」

日経DUAL編集長羽生祥子によって2016年11月6日に公開されたインタビューの中で、「親学」にかかわっていたかどうかについて小池百合子は「かかわっていない」と述べた後に次のように答えている。「一度だけ誘われて講演を聴きに行ったことはありますが、ちょっと私の考えとは方向性が違うと考えて、以降は何も関わっていません。」

なお、「親学」を単語登録しクラスタリングしたところ、「日本会議」という単語が出てきた。これは、小池百合子が、日本会議の「国会議員懇談会副会長」を務めていたことについてのものであり、これもまたほぼすべてが批判的ツイートであった（53,439ツイート）。ただし、都知事選に立候補した7月14日に、衆議院議員を退職した（自民党から離党した）ことによって、脱会したようである。

以上の二つのローカル空間に存在している「待機児童」、「子育て」クラスタから見えてきた側面として、二点指摘できる。一つは、これらの問題に関して、どのような実態があるか、このことについての情報が拡散され、共有されているということである。もう一つは、それらの社会問題についてコメントするよりも、候補者という人物に関してコメントすること、中でも候補者を何らかのネガティブなエピソードとからめて批判する投稿内容が多いということである。

■　3.「保育園落ちたの私だ／保育園落ちた日本死ね」クラスタ

この「保育園落ちた」を含むツイートは、2,883と数としては多くないが、都知事選で待機児童問題に関心が集まった原因の一つをつくった出来事だと考えたため、ここで取り上げることにした。

「保育園落ちた日本死ね」という言葉をめぐる一連の出来事については、遠藤（2016）に簡潔にまとめられているため、それを引用する。「2016

図 10　「保育園落ちたの私だ／保育園落ちた日本死ね」ローカル空間

年 2 月中旬にブログに『保育園落ちた日本死ね！！！』という記事が投稿された。書いたのは 30 代の女性であると後に判明した。内容は、自治体の認可保育園への入園が叶わなかった女性が、はたらきながら子育てをすることの困難さを訴えたものだった」。この記事は、すぐにウェブや Twitter で話題になり、テレビのバラエティー番組などでも取り上げられた。また、民主党の山尾志桜里議員が、2 月下旬に国会で、この件について安倍首相に質問したところ、「匿名による投稿であるため、それが本当なのかを確認しようがない」などと発言したことから、ネット上ですぐに拡散され、Twitter 上で、3 月に「＃保育園落ちたの私だ」というタグが流れ始めた。

一方、「これを合い言葉にした人びとが自然発生的に国会前でのスタンディングを行い」、他方で、署名ウェブサイトとして有名な Change.org にわずか 1 週間で保育制度の充実を求める約 2 万 7000 人の署名が集まり、山尾議員がこの署名をその母親らと一緒に政府に手渡したということがあった。

そして、2016 年 12 月、この言葉はその年の流

行語大賞に、政治トピックで唯一受賞することとなった。受賞に際し山尾議員は、「私がこの賞を受け取っていいのかとてもためらっている」としながらも、当時を「奇跡のようなことだと思っている」と振り返り、「そこから、みんなの力でこの待機児童問題を政治課題の隅っこからど真ん中に、場所移動することができた。ここからは、またみんなでシェアしながら解決するときだと思っている。」（民進党広報局（2016））とコメントし、署名した母親たちに謝意を表明した。なお、ローカル空間には、「山尾」議員の名前が下部に見える。

「保育園落ちた」を含むツイートは、2,883 あり、そのうちのほとんどである 2,568 がリツイートだった。これは、特定のツイートが、大規模に拡散されたことを意味する。つまり、それらの投稿は、一度もリツイートされない人のつぶやき＝ OT ではなく、拡散されたものであることがわかる。

また、図 11 にあるように、「保育園落ちたの私だ」という単語に最も近い単語が「スタンディング」、「いらっしゃら」だった。それぞれ調べたところ、「スタンディング」という単語が入ったツイートは 3,409 存在し、その内容は「保育園落ちたの私だスタンディング」の集会に参加した人びと、また、鳥越俊太郎を応援する有権者たちが、駅前など都内各地で集会を行っているものを指しているものがほとんどだった。これに対し、「いらっしゃら」は、小池百合子が「保育園落ちたの私だ」スタンディングに「いらっしゃら」なかったことについてのツイートのリツイートがほとんどだった。そのため「いらっしゃら」と「小池」いずれをも含んだツイートのみカウントしたところ、[6] 1,329 ツイート中、1,325 が以下のツイートのリツイートだった。「小池百合子さんは保育園問題を盛んにアピールしているようですが保育園落ちたの私だスタンディングにもいらっしゃらなかったしその後行われた超党派の院内集会にも本人も秘書さんもいらっしゃいませんでした」。つまり、このローカル空間では、鳥越俊太郎を支持する内容がある一方で、小池百合子を批判する内容があることが確認できた。

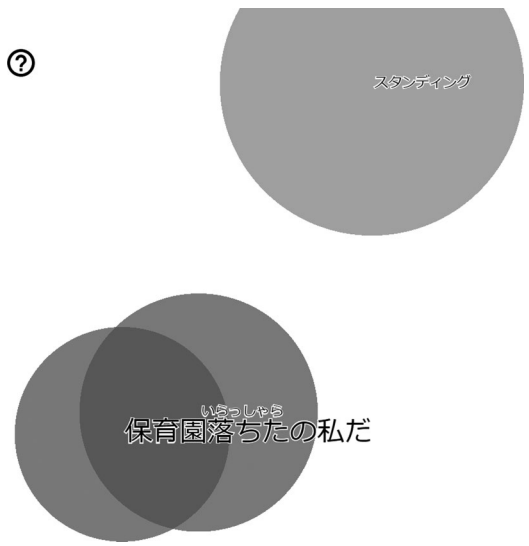

図 11　「保育園落ちたの私だ」、「スタンディング」、「いらっしゃらな（い）」クラスタ

遠藤（2016）は、「『保育園落ちた』というフレーズで語られる一連の動きは、〈世論〉あるいは社会運動を構成したといってもよいかもしれない」と述べている。本稿では、選挙期間中の Twitter データしか収集していないため、ブログが書かれ、流行語大賞を受賞するまで、これについてのツイートやリツイートがどのぐらいあったかは分からない。しかし Twitter 上で、「＃保育園落ちたの私だ」というハッシュタグが流れたことから、この都内各地で開かれた集会へと人びとを動かした「運動」を駆動する力の一端をなしていたのが、Twitter であったことは間違いないだろう。そしてここに、先述した三つ目の本稿の目的を確認することができる。すなわち、ここには、Twitter の大きな特徴の一つである、ある種の（ポジティヴな）共感が Twitter ユーザーたちの間ですぐさま広がるモメントが見られる。

■　4.「演説」クラスタ

このローカル空間では、図 12 にあるように、都内での候補者たちの演説場所のクラスタが出ていた。「保育園落ちたの私だ」ローカル空間に出てきたものであるということは、このトピックに関連する演説が、それぞれの場所で行われていた可能性がある。したがって、演説場所が特定の場所だけに偏っている可能性があるが、他のローカル空間には、これほどまとまって演説場所のクラスタが見られなかったため、演説クラスタについては、このローカル空間で出たものを扱うことにする。

それぞれの単語の入ったツイートを目視で確認し、候補者ごとの数をカウントしたところ、以下のようになった。なお場所については特徴のあるもの、数が多いものだけを調べた。

中ほどにみえる、「板橋」を検索したところ、ほとんどが街頭演説に関するものであり、3,588 ツイート存在した。そのうち、「板橋・小池」の All データを AND 検索したところ 1,672、RT データは 1,542、存在した。「板橋・鳥越」の All データは 1,649、RT データが 1,523、存在した（「板橋・増田」の All データは 485（ただし、鳥越、小池と重複したツイートが多かった）存在した）。3,588 のうちほとんどがこれら両候補に関するものであり、また、数がほぼ同じであり、いずれもリツイートが 9 割以上占めていることが分かった。

図 12　演説場所クラスタ

上部にみえる「小平」を検索したところ、2,623 ツイート存在した。そのうち、「鳥越・小平」を検索したところ、2,150 ツイートと最も多かった（ただし、演説とはかかわりのないものも少ないながら存在した）。そのうちリツイートは、1,957 だった。「小池・小平」が 351、そのうちリツイートが 273、存在した（「増田・小平」は 132、存在したが、鳥越と重複したものがいくつかあった）。鳥越俊太郎についてのツイートのリツイートのみ、9 割を超えていた。リツイートされたものとして多かったものの一つが以下の演説告知ツイートであった（190 回リツイート）。「28 日（木）18:30 開場、19:00 開会ルネこだいら大ホール 西武新宿線小平駅徒歩 3 分 応援弁士 民進党：山尾しおり 共産党：田村とも子 社民党：福島みずほ 生活と山本太郎と～：山本太郎 他 #鳥越俊太郎を東京都知事に」。他にも「参加します」や「集まりましょう」といった呼びかけるようなツイートがリツイートされていた。また、演説の結果を報告するツイートとして「昨夜、小平市で『鳥越俊太郎』演説会が開かれ、会場のルネ小平を 2 階席まで埋め尽くした。鳥越候補は、待機児童ゼロ、待機高齢者ゼロ、原発ゼロ、の 3 つのゼロを訴えた。多くの国会議員、19 区（国分寺、国立、小平、西東京）の自治体議員などが参加」が 73 回リツイートされていた。しかし、小平市の開票結果を確認したところ（「東京都知事選挙投開票結果」（2016））、鳥越は 21000 票と、小池：約 38000 票、増田：約 27000 票には及ばなかった。これらのことから、Twitter 上での盛り

上がりが、必ずしも選挙結果に結びつかないことが分かった。とはいえ、こうした呼びかけやその出来事の最中になされる投稿といったものも、Twitterに固有のデータ、現象である。こういった、同時多発的に、各所からなされる投稿データは、サンプリング抽出によってなされるアンケートでは収集することが難しいデータだといえる。

左にみえる「石神井」を検索したところ、894ツイートあった。すべて、石神井公園駅前で演説することを告知する内容だった。「石神井・鳥越」が420ツイート、リツイートが391あり、「石神井・小池」が397ツイート、リツイートが374あった。いずれも総数は少ないがリツイートが9割を超えていた。上部少し下にある「西東京」を検索したところ、893ツイートある中で、「西東京・小池」のAllデータが191、RTデータが172、「西東京・鳥越」Allデータが186、RTデータが123が存在した。

■ 5.「介護」クラスタ

介護ローカル空間上の、図13にみられるクラスタは、先の図7の「待機児童」クラスタ1（「保育士・待遇改善」）と同じ問題、すなわち、保育士と同様、介護士についても、人手が不足していること、待遇が悪いこと（「低賃金」、「薄給」）に加えて、「離職率（が高い）」ことについて投稿が行われていることが分かる。これらについては、図7で指摘したのと同じく、介護問題の実態について行われた客観的な投稿とそれの拡散といっていいだろう。

ここで注目したのは、図7の待機児童クラスタ1にはなかった「離職率」と「介護離職」、「介護難民」という単語である。「離職率」を含むツイートをカウントしたところ、1,015存在した。そのうち、「離職率」、「保育」でAND検索を行ったところ、67件ヒットし、そのうち一つを除き、「鳥越俊太郎さん『保育士や介護士は他の職業より十万円くらい給料が低い。離職率も高い。だからちゃんと行政が手当をして行くべき』これは問題分析で誰でも出来ること。都知事は『ちゃんと行政が』の当事者で、都民が聞きたいのはそこのアイデアや政策。大丈夫か鳥越さん。」というツイートのリツイートが65件であることが分かった。一方、「離職率」、「介護」でAND検索を行ったところ、1,005ヒットし、そのうち最も多かったリツイートは鳥越俊太郎自身のツイート、「高齢者の介護を若者に押しつけないためには、介護士をきちんと確保しなければならない。だが、介護士の離職率は20%。介護の職に就いた方の5人に1人は辞めてしまっている。それで本当にいいのか」が290回リツイートされたものであった。また、他にも鳥越俊太郎が離職率に触れたツイートがリツイートされたものがあったが、1,015のうち65あった先の「鳥越俊太郎さん［･･･］大丈夫か」という先のツイートのリツイート以外は、

図13　「介護離職・薄給・離職率・介護難民」クラスタ

すべて鳥越俊太郎が介護士の離職率を問題にしたものであることが分かった。

次に、「介護離職」という単語についてだが、これが入ったツイートは、違った意味合いで使われていることが分かった。この単語を含むツイートは、5,395 あり、ほとんどがリツイートで、鳥越俊太郎を批判するものだった。内容は、立川駅前での演説で「介護離職」の意味をはき違えて使っていたことへの批判だった。鳥越俊太郎は演説内で、介護離職を、低賃金ゆえ結婚できないから介護士をやめるという意味で使っていたようだが、本来は、親の介護をするために仕事を辞めることを介護離職という。

最後に「介護難民」についてだが、約 480 万ツイート中、17 しか存在しなかった。あまりに少ないため、この単語に近い「待機老人」も調べたところ、これも 127 しかなかった。Twitter コーパス内で、いかにこの問題への関心が低いかということが分かった。しかし関心が低いとはいえ、少ない人たちがこれについて触れていたことを発見できたのは、この単語を介護クラスタ内に引き寄せたt-SNEの精度ゆえである。

さらに選挙期間中の新聞記事も検索してみたが、朝日新聞では「待機老人」、「介護難民」いずれの単語もゼロだった。東京新聞では「待機老人」は一件、「介護難民」は三件、読売新聞は「待機老人」が一件、「介護難民」が一件だった。なお、これらの記事の中で、有権者の声が紹介されたものが一件、東京新聞の記事にあったため、引用しておく。「一人暮らしの高齢者が住民の三割を占める都営住宅。新宿区の本庄有由さん（78）は高齢者の孤独死を防ぐ活動をしてきたが、病気で続けられなくなった。『都内には老人ホームなどの施設に入れない待機老人が四万人以上いる。問題は待機児童だけではない』」（東京新聞　2016 年 7 月 23 日　朝刊）。

■　6.「高齢者」クラスタ

「老人」については、近傍語にノイズが多く、また、とくにクラスタが出なかったため、ここでは省略する。「高齢者」を単語登録し、2000 回学習させたところ、これについてもあまりクラスタがでなかったため、4000 回まで学習させた結果、特徴的なクラスタとしては、図 15 のようなものがみられた。

「情弱」（同じ意味をもつ「情報弱者」を含む）という単語を含むツイートをカウントしたところ、少ないが、608 あった。ところで、ここで「情報弱者」という語が意味するものは、一般的なものとは異なる。ここでは、情報をネットで収集するものが情報強者であり、ネットを見ずに、テレビ、新聞から情報を収集している人びと（ここではそれが「高齢者」になっている）が「情報弱者」と呼ばれている。ツイートの一例を挙げておく。「不都合な情報をカットされたテレビしか見ていない情弱層が選挙で鳥越氏に入れそうでこわいですね。ネットを活用している若い世代に期待したいです」。このクラスタも、高齢者を批判するよりは、候補者である鳥越俊太郎を批判するツイートが多かった。

■　「潜在的待機児童」

最後に、有権者の関心の高かった待機児童問題で、最も重要だと思われる点は、冒頭で触れたように、東京都がカウントしている待機児童の数の三倍の待機児童が実際には存在しているという、計算のトリックについて触れているツイートが、極端に少なかったという点である。これについて記事にしていたのは、All データで調べたところ、共産党のしんぶん赤旗（ウェブ版）（「隠れ待機児」）、毎日新聞（「隠れ待機児童」）、東京新聞（「潜在待機児童」）だった。まず、しんぶん赤旗の記事の内容を一部引用する。

図 14　「高齢者」ローカル空間

図 15　「情弱」クラスタ

東京都は 19 日、4 月 1 日時点の都内の待機児童数が 2 年ぶりに増え、8466 人になったと発表しました。前年比 652 人の増加。認可保育所に入れず都独自の認証保育所に入ったり、育児休暇を延長する人など“隠れ待機児”を加えると 2 万人を超えるとみられています。(「東京の待機児童増加　知事選の重要争点　2 年ぶり“隠れ”含め 2 万人超」(しんぶん赤旗　ウェブ版　2016 年 7 月 20 日)

次に、毎日新聞が独自の調査を行い、このことを記事にしたのは、7 月 23 日だった。

全国 152 自治体への毎日新聞の調査で、待機児童の約 3 倍に当たる 5 万人以上が確認された「隠れ待機児童」。自治体が待機児童数を見かけ上だけ少なくもできることを示した数字で、保護者からの不満は強い。国に基準の明確化を求める自治体もある。(「隠れ待機児童：遠い保育所、断れば待機外　集計法に自治体「裁量」」(毎日新聞、2016 年 7 月 23 日、朝刊))

続いて、東京新聞も独自調査を行い、それを 7 月 26 日に記事にしている。

東京二十三区で、今春の入所を希望して認可保育所などに新規で申し込んだ子ども約六万七千人のうち、約二万四千人が入れなかったことが、本紙の調査で分かった。都が十九日に発表した二十三区の「待機児童」は約五千六百人だが、国の定義に含まれず各区が計上しない「潜在的な待機児童」が、その三倍以上の約一万八千五百人いる実態が浮かんだ。(「潜在待機児童 1 万 8500 人　都発表 5600 人　集計外が 3 倍超　認可 2 万 4000 人入れず　東京 23 区本紙調査」東京新聞、2016 年 7 月 26 日、朝刊)

All データの中から東京新聞が用いていた表現、「潜在（的）待機児童」という単語を入ったツイートを調べたところ、480 万のツイートのうち、264 ツイート存在した。そのうち、252 がリツイート、OT データが 12 ツイートであった。つまり、新聞のこの記事についてツイート、あるいはそのツイートをリツイートしたアカウントは、約 480 万ツイートの中で 12 アカウント、すなわち 12 人しかおらず、それが 252 回リツイートされたにすぎなかったということである。なお、目視で確認したところ、そのうちのリツイートは、その 12 アカウントのうち、2 アカウントのツイートをリツイートしたものがほとんどだった。また、毎日新聞が用いていた「隠れ待機児童」という単語は、106 ツイート存在し、共産党のしんぶん赤旗の記事についてのツイート、「都知事選問われる保育政策 鳥越氏開発優先やめて保育施設増設へ しんぶん赤旗 小池氏詰め込み策を推進 増田氏隠れ待機児童無視」といったツイートは 22、存在した。

上に引用した、東京新聞の記事は、2018 年 11 月の時点で検索したところ、ウェブ版には存在しなかった。しかし、選挙期間中に、当該記事の紙面を撮影した画像を貼り付けていたブログをいくつかウェブで確認することができた。また、毎日新聞の記事は、全文読むには有料会員になる必要があるとはいえ、見出しと記事冒頭は読むことができ、また、そのサイトにはツイートボタンが存在するため、ツイートしやすい環境にはなっていた。しんぶん赤旗ウェブ版は、全文読むことができ、またツイートボタンが置かれていた。

さきほどの「待機老人」の場合と同じく、重要な新聞記事がこれほどまでに拡散されないということは驚くべきことである。このことを深刻だと考える理由は、待機児童についての実態の把握が、実質的な数の三分の一としてしか示されていなかったことは、その少ない数を想定して提案されていた公約の意味をその分、失わせることにつながるからである。つまり、主要候補らによって提

案されていた、「保育園を増やす」、「空き家の利用」、「保育士の待遇改善」、「延べ床面積での児童数の規制緩和」などの公約は、実際の数の三分の一を想定して提案されたものである以上、実態それ自体が想定外となり、「待機児童ゼロ」など実行できるはずがなくなるからである。これだけ数が異なると、どの候補者が何を提案していたか、その分意味を失うし、また公約の内容そのものが、その分実行性を失うだけでなく、それらの公約についてTwitter上で行われていた投稿、すなわち先述した、規制緩和すると「ギュウギュウ詰め」になる、「保育の質が低下する」といった反論も含めて、すべてが意味のない議論に成り下がってしまう。また、有権者が最も強い関心を示していたのは、待機児童問題という新聞の調査結果が報告されていたが、これに関心を持っていた有権者たちは、候補者によって、見させられていた解決案の方向が、意味のないものだったということになる。

■ 分析を終えて

まず、分析手法について述べておきたい。word2vecによる単語ベクトル化による類似語の学習について元論文で示されていたのは、pythonスクリプト上でターゲット単語に類似する近傍語を出力させることであり、その精度が高いことだった。しかしこれをいくら繰り返しても、コーパスの全体像をつかむことができず、クラスタがどこにあるか、word2vecアルゴリズムだけだと特定できなかった。これを一定程度解決してくれたのが、word2vecで生成した高次元の単語ベクトルデータを次元圧縮するPCAアルゴリズムとt-SNEアルゴリズムであり、また、その結果を三次元で可視化してくれるEmbedding Projectorであった。これによって、類似語の全体像、つまり本稿でローカル空間としたもの、あるいは、グローバル空間が見えるようになり、また、クラスタをデータ・ヴィジュアライゼーションを使って特定することができた。

語彙についても、NEologdの語彙が豊富であったため、その分、クラスタリングの精度が上がったのは間違いない（例えば「介護難民」）。また、データのクリーニングを日本語のみにしたことによって、クラスタをその分はっきり出すことができた。というのは、t-SNEは、似た単語を近づけると同時に、似ていない単語を遠ざけるという性質をもつが、クリーニングができていないデータを学習させると、たとえば、ノイズにあたるローマ字と、日本語の大きな二つのクラスタができてしまい、学習の主な作業が、そのノイズクラスタとシグナルクラスタ二つを遠ざけることだけになってしまうということがあった。ノイズをきれいに除去しておくと、シグナルにあたるグループの中で、さらに細かいグループを分けることできていることを確認することができた。

次に、Twitterコーパスから得られた知見についてまとめておきたい。一つ目として、各候補が掲げる公約にからめながら、東京都内の社会の諸問題について、客観的な事実の共有が見られた。その裏面として、公約を事実に基づかない、誇大な表現での解釈が拡散され、さらにそれと似たようなツイートを生みながらさらに拡散されるということが見られた。二つ目として、Twitterの拡散という特性が利用されることによって、市民社会での呼びかけのようなものが拡散され、市民運動が現実空間で起きることを確認することができた。「保育園落ちたの私だ」の場合、当事者などが、現実空間でのスタンディングや、集会へと現れ、また、街頭演説についての告知の拡散が、それへと参加する人びとたちを街頭に引き寄せた。同時に、小平市での鳥越候補の演説に関するツイートから分かったのは、Twitter上での盛り上がりが、必ずしも現実に反映されるものではないことであった。三つ目として、投票権を持たない人たち（とりわけ東京都内に勤めている隣接県民たち）からの、期待や、不安や、不満などが共有、拡散されたことが確認できた。四つ目として、こ

の選挙で最も関心を持たれた待機児童問題の根本に関わる実態の誤った把握について、また、待機児童問題と同じく深刻な問題である介護難民について、少なくともTwitter空間では、それを問題視した人が極めて少なかったことが確認できた。

■　最後に

今後、AI（人工知能）が、またそれを支えるアルゴリズムが、自然言語を学習し、分析し、意思決定する機会が様々な場面で増えていくのは間違いない。この論文の目的の最後のものとして挙げておきたいのは、コードを、アルゴリズムを「ブラックボックス」のまま利用するのではなく、中の仕組みを理解した上で利用する必要があることを示すことだった。なぜなら、アルゴリズムが人間によって書かれたコードの束である限り、必ずそこには人間の偏見が含まれるからであり、それを修正できるのもまた、人間しかいないからである。とはいえ、本稿では、word2vecなどのアルゴリズムのオプション設定値を変えるなどして、アルゴリズムのパフォーマンスそのものを計測する、あるいは、複数の学習結果の精度を比較する、などを行なう余裕がなかった。稿を改めて検討することにしたい。

本稿は、JSPS科研費 16K 13189「新デジタルメディア時代におけるソーシャル・デザインのためのデータ利活用法研究」（代表：筆者）の助成を受けたものある。

注

1）「強化学習」で、最も分かりやすい事例は、2017年に囲碁の世界ランカーを破ったGoogleDeepMindのAlphaGoである。AlphaGoは、膨大な棋士たちの手を学習し、かつ、それを学習したマシン同士を何度も対戦させることによって、勝率を強化できることを証明した。

2）ただし、クラスタリングアルゴリズムの最もシンプルなものと呼ばれる、K-Meansアルゴリズムは、クラスタリングを行うにあたって、データに含まれる単語の数を事前に指定できるほか、クラスタの数とクラスタ毎の単語数をも事前に指定することができる。そこで、本稿で分析対象としたファイルを用いて、word2vecで単語ベクトル化したモデルファイルをつくり、Pythonライブラリのscikit-learnに実装されているK-Meansを使ってクラスタリングを行った（3,000単語を対象としクラスタ数は1,000でクラスタ毎の単語は10とした）。しかし、精度が非常に低かったため、本稿では用いなかった。

3）LDA（Latent Dirichlet Allocation）を使用してトピック分析を行った先行研究として、山縣、梅原（2018）がある、また他にも同様の先行研究として、木田（2017）がある。

4）なお本稿で設定したword2vecのオプション設定値は以下のとおりである。sg=1、size=300、min_count=5、window=10、hs=0、negative=5、iter=10、sample=0.001。オプションについての詳しい説明については、GitHubで公開されているPythonライブラリgensimのソースコード（word2vec.py）に記載されているので、そちらを参照されたい。

5）このことについての最も有力な証拠となるのは、Google翻訳が、2016年11月からニューラルネットワークアルゴリズムを用いた機械学習を導入して、ネット上に存在する膨大な自然言語データを学習した結果、翻訳精度が劇的に上がったことを挙げることができる。例えば、以下の名古屋大学の中岩浩巳氏へのインタビュー記事を参照「Google翻訳の精度はなぜ上がった？翻訳者は不要になる？専門家に聞いてみた」。

6）本稿では、このような二つの単語が入ったツイートを検索する際には、サクラエディタのGrepコマンドを用いて、AND検索を行った。

■　参考文献，参考サイト

Demographia “World Urban Areas” study. http://www.demographia.com/db-worldua.pdf（最終アクセス日　2018年12月1日）

EKWords http://www.djsoft.co.jp/products/ekwords.html

Embedding Projector https://projector.tensorflow.org/（最終アクセス日　2018年12月1日）

遠藤薫（2016）「間メディア民主主義と〈世論〉―2016年都知事選をめぐるスキャンダル・ポリティクス」，『社会情報学』，第5巻1号，2016。

Google Developers（Smilkov, Daniel, Viégas, Fernanda, Wattenberg, Martin.）（2016）“A.I. Experiments: Visualizing High-Dimensional Space” https://www.youtube.com/watch?v=wvsE8jm1GzE&feature=youtu.be（最終アクセス日　2018年12月1日）

グッリ，アントニオ，パル，サジット（2017=2018）『直観Deep Learning』，大串正矢，久保隆宏，中山光樹訳，オライリー・ジャパン。

保坂展人（2016）「『都立認可保育園設置』は，『待機児童ゼロの切り札』になる」https://www.huffingtonpost.jp/nobuto-hosaka/tokyo-election_b_11175212.html（最終アクセス日　2018年12月1日）

木田勇輔（2017）「ソーシャルメディアとポピュリストの動員―2016年東京都知事選挙におけるTwitterデータの分析から―」，『文化情報学部紀要』，椙山女学園大学，第17巻，pp.83―92。

コトバンク「待機児童」https://kotobank.jp/word/%E5%BE%85%E6%A9%9F%E5%85%90%E7%AB%A5%E5%95%8F%E9%A1%8C-895398（最終アクセス日　2018年12月1日）

小池ゆりこ選挙事務所公約ウェブサイト（2016）https://www.yuriko.or.jp/senkyo/kouyaku.pdf（最終アクセス日　2018年12月1日）

駒崎弘樹，羽生祥子によるインタビュー「小池百合子都知事　保育士の確保に本気で取り組む」日経DUAL 2016.11.07　https://dual.nikkei.co.jp/article/093/94/?P=2（最終アクセス日　2018年12月1日）

厚生労働省「保育所等関連状況取りまとめ（平成28年4月1日）」https://www.mhlw.go.jp/stf/houdou/0000135392.html（最終アクセス日　2018年12月1日）

厚生労働省「人口動態調査」（2016年）https://www.e-stat.go.jp/stat-search/files?page=1&layout=datalist&tstat=000001028897&cycle=7&year=20160&month=0&tclass1=000001053058&tclass2=000001053061&tclass3=000001053064&result_back=1&second2=1&toukei=00450011&stat_infid=000031621289（最終アクセス日　2018年12月1日）

Kudo Taku.（2013）“MeCab:Yet Another Part-of-Speech and Morphological Analyzer”, http://taku910.github.io/mecab/（最終アクセス日　2018年12月1日）

工藤拓（2018）『形態解析の理論と実装』近代科学社。

中岩浩巳（2016）「Google翻訳の精度はなぜ上がった？翻訳者は不要になる？専門家に聞いてみた」https://gotcha.alc.co.jp/entry/20161220-it-translation（最終アクセス日　2018年12月1日）

Maaten, Laurens van der, and Hinton, Geoffrey. 2008. “Visualizing data using t-SNE.” Journal of Machine Learning Research, Vol 9(Nov), pp. 2579-2605.

Mikolov, Tomas, Chen, Kai, Corrado, Greg, Dean, Jeffrey. “Efficient estimation of word representations in vector space”. CoRR, abs/1301.3781, 2013.

Tomas Mikolovプロフィールhttps://scholar.google.com/citations?user=oBu8kMMAAAAJ&hl=en（最終アクセス日　2018年12月1日）

Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Corrado, Gregory S., Dean, Jeffrey. 2013. “Distributed representations of words and phrases and their compositionality”. In Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States, pages 3111-3119, 2013.

マイヤー＝ショーンベルガー，ビクター，クキエ，ケネス（2013=2013）『ビッグデータの正体』，斎藤栄一郎訳，講談社。

民進党広報局（2016）「『保育園落ちた日本死ね』新語・流行語大賞トップ10入りで山尾議員が受賞」https://www.minshin.or.jp/article/110505（最終アクセス日　2018年12月1日）

奥村学監修，高村大也著（2010）『言語処理のための機械学習入門』コロナ社。

Sato Toshinori.（2015）“Neologism dictionary based on the language resources on the Web for mecab-ipadic”, https://github.com/neologd/mecab-ipadic-neologd/（最終アクセス日　2018年12月1日）

Smilkov, Daniel, Thora, Nikhilt, Nicholson, Charles,

Reif, Emily, Viégas, Fernanda, Wattenberg, Martin. “Embedding Projector: Interactive Visualization and Interpretation of Embeddings.” 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

高史明（2015）『レイシズムを解剖する：在日コリアンへの偏見とインターネット』，勁草書房。

東京都選挙管理委員会事務局（2016）「東京都知事選挙投開票結果」http://www.senkyo.metro.tokyo.jp/election/tochiji-all/tochiji-sokuhou2016/（最終アクセス日　2018年12月1日）

山縣史哉，梅原英一（2018）「平成28年度東京都知事選挙のTwitter分析」，信学技報，電子情報通信学会。

UserLocal Social Insight，株式会社ユーザーローカル https://social.userlocal.jp/

八代尚弘（2016）『シルバー民主主義　高齢者優遇をどう克服するか』中公新書。

word2vec.py. https://github.com/RaRe-Technologies/gensim/blob/develop/gensim/models/word2vec.py#L436（最終アクセス日　2018年12月1日）