

## 研究資源としての Twitter

植 田 麦

### はじめに

掲題のとおり、本稿は Twitter を研究資源として利用することについて論じるものである。本稿執筆時現在、J-STAGE で「Twitter」と検索すると、5000 件近くヒットする（なお、論題に「Twitter」を含むものは 300 件程度である）。このように、すでに Twitter を用いて研究する方法は相当の広がりをもっている。

この状況に鑑れば、いまさら稿者がこのような文章を公にする必要はないように思われるかもしれない。しかしながら、以下に示す理由から、稿者は Twitter を研究資源とすることについて論じる必要があると考えるのである。

第一に、Twitter を利用した研究では「なぜ Twitter を利用するのか」についての言及がない場合がある。それらの研究では、Twitter を利用することが自明のものと考えられているために、その前提について説明されないのではないかと。また、Twitter には研究資源とするメリットが大きい、にもかかわらずそのメリットについての言及も少ない。そのため、Twitter が研究資源たりうることに、簡略であっても説明が必要と考えるのである。

第二に、これから Twitter を利用した研究を進めようとするときに、初歩的な知識を提示する情報が少ない。もちろん、本稿でも示すようにさまざ

まな手引き書はすでに公開されているし、ウェブ上には多くの情報が提示されている。しかし、それらはある程度の知識や技術をもつ利用者を前提にしたもので、全くの初心者进行を誘う情報は多くない。

稿者が所属機関で担当しているゼミナールにおいては、毎年のように卒業論文執筆に Twitter を利用する学生がいる。本稿は、彼らにむけた説明を再編成したものであることをあらかじめ示しておく。つまり、以下の言及は卒業論文で Twitter を利用したいと考える学生を读者の水準として設定するものである。

第三に、Twitter を利用した研究は「きれい」すぎる。公にされている論考で Twitter を研究に用いているものは、当然ながらデータのクレンジングが行われていると推測される。そこでは、クレンジングがいかに行われているのかについては触れられないことすらある。Twitter を研究資料に用いるものでは多く、テキストマイニングの技術が使用される。このテキストマイニングにおいては、データのクレンジングがきわめて重要である。また、Twitter を用いた研究では、ノイズになる要素が研究計画によって異なる。そのため、クレンジングの実例をもとにした説明が必要であるとする。

第四に、公刊論文では Twitter のデータ（ツイート）を取得する際に得られる変数についての言及が少ない。たとえば発信日時や発信ツールについての言及があってもよいのだが、そういった点に着目したものはあまりみられない。これらの変数は説明変数として設定することが可能であり、またデータのクレンジングにも利用できる。

以下、本稿では Twitter からのデータ抽出方法および利用方法について述べるとともに、データ利用時に考えるべき問題点を示す。

### 補論：テキストマイニング

テキストマイニングとは、大量のテキストを統計学の知見を用いて分析することで、そのテキストの価値を新たに掘り起こすものである。

テキストマイニングの考え方や利用法については、すでに多くの解説書が公刊されている。たとえば、金明哲（2021）はその著書の題（『テキストアナリティクスの基礎と実践』）のとおり、テキストマイニング（当該書に基づけば「テキストアナリティクス」）の基本的な考え方を示すとともに、実践的な分析方法を提示する。また、石田基広（2017）もテキストマイニングについての概説を行ったのち、様々なケースを想定した実践的研究方法を示す。

無料のテキストマイニング用アプリケーションソフトである KH Coder の解説については、樋口耕一（2020）に詳しい。また、樋口耕一・中村康則・周景龍（2022）は KH Coder の実践的使用法についての解説を行っている。

詳細なテキストマイニングについての紹介を本稿で行うことは、紙幅の都合からも稿者の能力からも難しい。ただし、論考の都合から、基本的な考え方と分析手法についてはここに示す。

テキストマイニングで必要となるのは、分析対象となるテキストと分析のための形態素解析器である。形態素解析器とは、ことばについてその意味を保持しうる最小単位に分解するためのプログラムである。形態素解析器は、辞書（解析の基礎となる、語の情報を集約した電子データ）を必要とする。この辞書が現代語に基づくものであれば、新聞記事や首相の所信表明といったテキストを分析することが可能である。古典語、たとえば上代語に基づくものであれば、『万葉集』や『古事記』といったテキストの分析をすることができる。

形態素解析器としては、MeCab や ChaSen 等が知られる。通常、これらの形態素解析器はそれ単独で使用されることが少なく、R や Python で運用される。また、形態素解析器を組み込んだアプリケーションソフトとして用いられる。たとえば上述の KH Coder は、R および複数の形態素解析器・現代語の辞書をひとつのパッケージにしたものである。金明哲の作成した MTmineR は R および MeCab 等を総合的に操作するアプリケーションソフト

トである。

テキストマイニングを行う際には、いかにして分析するか（＝いかなる解析器をどのように用いるか）を決定した上で、その分析手法に適したかたちで分析対象のテキストを整備しなければならない。

## 1. Twitter からのデータ抽出

### 1-1. SNS の種類と研究資源としての適性

SNS には、テキストコミュニケーションを主としたものと、映像や画像、音声・音楽による交流を目的とした非テキストコミュニケーションのものがある。前者の例としては Twitter や LINE、Facebook があり、後者の例としては YouTube・Instagram・TikTokなどを挙げることができる。

これらのサービスの中では、当然ながら前者がテキスト分析研究に適している。しかしながら、それらがすべて、テキスト分析による研究対象とはなりえない。たとえば、LINE は非公開のテキストが大半であり秘匿性が高いため、データの取得ができない。Facebook はテキストが公開されているが、データの抽出が難しく、テキスト分析の適性に欠ける。これらに対し、Twitter はテキストが公開されている上、Twitter API (Application Programming Interface、後述) が利用可能であり、テキスト分析による研究に向けた媒体であるといえる。

Twitter はすでに社会的インフラのひとつとして機能しているといっよういだろう。総務省の公開している「平成 24 年 情報通信メディアの利用時間と情報行動に関する調査」では、Twitter の利用者は調査対象全体の 15.7% であるが、同「令和 2 年度 情報通信メディアの利用時間と情報行動に関する調査」では 42.3% にまで拡大している。このように、利用者数・データ利用のしやすさからみて、Twitter は研究資源としての適性を備えている。

なお、非テキストコミュニケーションのサービスである YouTube にも一般利用が可能な API があり、各動画に付されたコメントの抽出が可能である。ただし、Twitter に比して利用条件が厳しく、現状では研究に不向きである。

## 1-2. Twitter を利用した研究

Twitter を対象とした研究は、日本では 2010 年代より多くみられる。たとえば、村井源（2012）は東日本大震災直後に発信されたツイートを対象として、そのツイートに付されたハッシュタグを分析している。この村井（2012）のように、社会的に大きなインパクトをもたらした事件や事態についての研究は多く、たとえば本稿執筆時点では COVID-19 に関連した研究として四方田健二（2021）や渡邊憲二・箕輪弘嗣（2021）がみられる。Twitter は社会状況を反映する「鑑」のような機能をもつため、ある時点での社会状況の分析や検討に適した媒体である。

このような社会状況の「鑑」としての Twitter は、利用者の指向性をみるための研究でも利用されている。高史明（2015）は、在日外国人に対するヘイトツイートを考察している。曹慶鎬（2018）も同様に、2016 年に発生した熊本地震の直後に発信されたツイートのうち中国（人）・韓国（人）に関わるものを分析している。また、上ノ原秀晃（2019）は 2017 年に行われた第 48 回衆議院選挙において、立候補者が Twitter をいかに利用したかについて考察している。

また、こういった何かしらの事態を反映したツイートを分析するものだけではなく、平常のツイートを対象とする研究もある。松本義之・井上仙子（2020）は下関市の地理データを有するツイートを収集し、下関市の観光資源を探求している。吉見憲二・上田祥二・針尾大嗣（2019）は売買春に関わるツイートについて、利用されるハッシュタグの分析を行い、問題投稿についての事前検知が有効であることを主張する。

このように Twitter では、大きな衝撃をもたらした事件・事故や社会に大きな意味をもつイベントといった、突発的あるいは一時的な事柄についてのツイートが多く発信される。また、利用者の日常的な感慨や意見を発露するツイートもあり、いわば非日常と日常を包摂する媒体といえる。この特徴は Twitter を有用な研究資源たらしめるとともに、研究目的によっては難点を生み出すこともある。この「難点」については 2-1-2. で詳説する。

### 1-3. Twitter API の取得と利用

先述のとおり、Twitter は API が利用可能である。本稿の目的に沿って API を簡略に説明すれば、Twitter で各アカウントが発信しているツイートの取得を可能とする仕組みである。ウェブ上のデータを機械的に収集する行為をスクレイピングというが、Twitter API は Twitter 社公式でツイートをスクレイピングすることができるものといってよいだろう。

Twitter API では、任意のワードを含むツイートの取得はもちろん、そのツイートを発信したアカウントの表示名、発信日時、発信ツール等のデータも抽出することができる。また、特定のアカウントのツイートを取得することもできる。

Twitter API を利用するためには、Twitter のアカウントが必要である。また、電話番号・メールアドレスをアカウントに紐付けなければならない。基本的に利用料金はかからない。ただし、高度な分析を行うためには料金が発生する場合もある。通常の研究においては、無料の範疇で十分である。

しかしながら、Twitter API を取得するためには、煩雑な手続きが必要である。また、不定期に取得方法が変更されるため、本稿では取得方法を説明しない。「Twitter API 取得」などでウェブ検索すれば、その時点で最新の取得方法を解説したウェブサイトにとどり着くはずである。

ただし、現時点で Twitter API を取得する際の留意点については、2 点、本稿でも述べておく。まず、Twitter API の取得申請にあたっては英語での

記述が中心である。いくつかの質問に対する回答も、すべて英語で記述する必要がある。次に、申請を完了してすぐに Twitter API が使用できる場合と、数時間後から数日後に追加の質問メールが送られてくる場合とがある。質問メールは日本語であるため、回答も日本語で行う。特段の問題がなければ、Twitter API の使用は許可される。なお、稿者が指導しているゼミナール所属の学生に Twitter API の取得を行わせたところ、ほぼ同内容の申請であるにもかかわらず、申請してただちに Twitter API が利用できたものと、追加の手続きが必要なものとがいた。その区分は不明である。

ツイートは 15 分ごとに 18000 件まで取得可能である。18000 件を越えるツイートを一括して取得したい場合は、制限が解除されるのを待つ必要がある。また、取得できるツイートは取得時点からおおむね 1 週間前のものまでである。

#### 1-4. データ取得に用いるプログラミング言語

Twitter API は、単独では利用が難しい。なにかしらのプログラミング言語と組み合わせて利用する。研究論文では Python を利用するものも多いが、稿者は R を利用している。これは、石田 (2020) のように比較的入手しやすい手引書が R を基本としているためである。なお、石田 (2017) でも Twitter を利用したテキストマイニング研究の解説が行われているものの、利用されている R のパッケージ (twitterR) が、現在は使用できない。以下、本稿では R と R のパッケージである「rtweet」の利用を前提として論を進めるが、Twitter API の仕様が利用するプログラミング言語によって制限される可能性は低いため、Python を利用した場合であっても、以下の言及に差はないものと考ええる。よって、Python を使用して分析を行う場合は、適宜環境に合わせて本稿を参照願いたい。

R で Twitter API を利用する場合、R そのものにコードを打ち込むよりは、R のための開発環境である RStudio を使用するとよい。導入について

は石田 (2017) に詳しい。ただし、Windows で R および RStudio を導入する場合、ホームディレクトリを半角英数以外で登録していると、rtweet を含めた各種パッケージのインストールに失敗することがある。また、OneDrive が障害となる場合がある。

ホームディレクトリを半角英数以外で設定している場合は、ID を半角英数にしたアカウントを新たに設定するなどの対処が必要である。新たにアカウントを作成した場合は、そのアカウントでログインした状態で R および RStudio のインストールを行う。パッケージをインストールする際、インストール先に OneDrive が選択されてしまう場合は、OneDrive のバックアップ機能をオフにする等、対応を行わなければならない。

### 1-5. 取得すべき変数

Twitter API では、多くのデータが取得できる。しかしながら、大半の項目はブランクであり、利用可能なデータの種類は多くない。取得可能なものの中で利用価値のあるデータは、第一にツイートそのもの (text) である。その他、発信日時 (created\_at)、発信者のアカウント名 (screen\_name)、発信ツール (source)、テキスト量 (display\_text\_width)、リプライの対象 (reply\_to\_screen\_name)、「いいね」の数 (favorite\_count)、リツイート数 (retweet\_count)、ハッシュタグ (hashtags)、発信者のフォロワー数 (followers\_count)、発信者のフレンド (フォロー中のアカウント) 数 (friends\_count)、発信者のアカウント作成日時 (account\_created\_at) などが利用できる。

これらの項目のうち、稿者は発信日時・発信者のアカウント名・発信ツールを抽出している。発信日時は分析の際、変数として利用することができる。また、アカウント名・発信ツールはデータのクレンジングに利用する (詳細は後述)。

研究計画によっては、これ以外の項目が有効となる場合もあるだろう。た



たとえば、任意のワードを含むツイートについて「いいね」の数 (favorite\_count) を取得し、発信者のフォロワー数 (followers\_count) と合わせて考えることで、「いいね」を獲得するメカニズムについて研究が行えるかもしれない。

また、後述するスパム対策 (2-2-8.) として、発信者のフォロワー数 (followers\_count) および発信者のフレンド数 (friends\_count) を取得し、両者ともに 0 もしくはフォロワー数が 0 のものをノイズとみて削除する考え方もあるだろう。なんとなれば、宣伝を目的とするアカウントの大半は、フォロワーが存在しないためである。さらに、そのようなアカウントは宣伝目的で他アカウントをフォローする場合もあるが、フォローをせず宣伝目的のツイートを発信するのみの場合もある。

## 2. 問題点

Twitter からツイートを取得して研究を行うにあたり、様々な面において考えておくべき問題点がある。以下、データの精度・Twitter のシステム・データの利用から、問題点をみていきたい。

### 2-1. データの精度に関わる問題

#### 2-1-1. 表記揺れ、同義語

ことばを対象とする研究において、常に問題となるのは表記揺れであろう。Twitter を研究資源とした場合も、同様の問題がある。Twitter では書き手が一律でないため様々な表記がなされる。そのため、表記揺れについても、そのまま受け入れるのかあるいは何らかの処理を行うのかを決定しておく必要がある。

たとえば、「いう」と「言う」は、形態素解析では別語として処理される。

これをどちらかに統一するのか、そのまま別語として扱うのか。

もしも表記の統一を図りたいのであれば、まず表記揺れを起こしている語の確定をしなければならない。そのためには、一度、すべての語の頻度表を作成し、表記揺れの確認をする必要がある。そののち、統一すべき語を決定し、処理を行う。

表記揺れに類する問題として、同義語の扱いがある。たとえば、「独身者」と「未婚者」はほぼ同義で使用される。文脈によっては、「単身者」も同じ意味をもつことがあるだろう。一方、「卒論」は「卒業論文」の略語であるから、両者の意味するところは同じであるように考えられる。しかし、日常の会話では「卒業論文」よりも「卒論」の出現頻度が高いであろう。とすれば、略さない表現である「卒業論文」が用いられる文脈は、「卒論」とは異なる可能性がある。この場合、「卒論」と「卒業論文」とを同一の語として扱うことには注意が必要である。

分析の目的によっては、同義語を統一することによって傾向がみえやすくなることもある。その場合、表記揺れを統一するのと同じ作業を行わなければならない。

KH Coder で分析を行う場合、表記揺れや同義語を統一するときは「表記揺れの吸収」を行うか、有料プラグインである「表記ゆれ&同義語エディター」を用いる。前者は、KH Coder に組み込まれているプラグイン「表記揺れの吸収」を利用するものである。使用の詳細については、KH Coder の公式サイトに詳しい。利用のための手順はやや煩雑であるが、分析者の意図に沿った設定が可能である。後者は KH Coder のバックアップを行っている企業である株式会社 SCREEN アドバンスドシステムソリューションズが公開しているもので、きわめて平易に表記揺れや同義語の統一を行うことができる。

### 2-1-2. 外れ値になりうるデータ

社会では時として、それまで話題にならなかったことが急速に語られ、そして時間をおかず忘れ去られることがある。社会状況を反映する Twitter でも、同様の現象がおきる。たとえば、ある企業で不祥事が発覚すると、普段はその企業名がツイートされることがほとんどないにもかかわらず、数日から一週間ほど、その企業に関わるツイートが多く発信される。仮に、「クッキー」についてのツイートを抽出しているとして、その抽出時期にクッキーを生産している会社でセクハラの実例が公表されると、ツイートを分析したときに「クッキー」と「セクハラ」が共起語として検出される。通常、「クッキー」は「セクハラ」と同時に語られることが少ないにもかかわらず。

このように、ツイートを取得する期間が長いと、一定時期に集中して他の時期にはみられないツイートが抽出されることもある。これらのツイートに対して、たとえば対応分析を行うと、一定時期に集中して発信されたツイートの内容が「特徴語」として表示される。しかしながら、それらの「特徴語」はデータ全体からみれば特異であって、不適切なものともいえる。

このような、いわば外れ値となるようなデータについての対応を考えておく必要がある。たとえば、そのようなツイートにみられる語を排除して分析を行う方法もあるが、しかしながら、そのような分析は恣意的な操作を加えているとみられうるため、相当な慎重さが求められる。

このような状況が出来た場合は、特異と考えられるデータが取得された背景を説明し、「外れ値」を含む分析と除外した分析を示すなどの対応が現実的であるかもしれない。

### 2-1-3. 用言、否定表現、研究の意図とは異なるツイート

ツイートを取得する際、用言の抽出は困難である。Twitter 自体での検索では、たとえば「あきらめる」と検索しても、「あきらめる」「諦める」「あきらめた」など、異なる表記・活用形でもヒットする。しかしながら Twitter API では、表出形でしか検索ができない。つまり、「あきらめる」を検索する場合、すべての活用形で検索をしなければならない。また、「諦める」のように異なる表記についても検索をする必要がある。この場合、さらに、「諦める」の例も含まれてしまう。

表現に関わる問題としては、否定表現も対応を考慮しておくべきである。たとえば、評価の賛否がわかるものを検索対象（仮に A とする）としていると、ツイートの中で「A はいいよね」と「A はよくない」が混在していても、語の頻度や共起としては後者も「A」と「よい」として計測されてしまう。よって、分析では「A」と「よい」とが共起しているものと検出され、結果、「A については肯定的評価が多かった」と結論を出してしまいかねない。

KH Coder で分析を行う場合、否定表現を別語として検出する有料プラグイン「否定表現チェッカー」がある。これは「歩か（ない）」を「歩く（否定）」のように抽出するものである。ただし、「歩く」と「歩く（否定）」は別語の扱いになるため、「歩く」の使用頻度数は下がる。そのため、否定語として検出しない場合は十分な頻度数を有していたことによって分析結果に表出していたのが、別語としたために使用頻度が低下し、分析結果に出てこなくなる場合がある。

なお、考えるべきであるのは検索対象そのものについての否定表現のみではない。たとえば「A はともかく、B はよくないよね」といったツイートのように、検索対象の A と関連して現れる B について多く否定表現があらわれる場合もある。しかし、A にのみ着目して分析を行うと、B についての評

価ではなく A についての評価として分析結果をみてしまうことがある。

また、非日本語の文字列を検索語としてツイートを取得すると、意図せざるものを抽出してしまうことがある。たとえば「ユニクロ」や「しまむら」を含むツイートを取得するとき、ユニクロやしまむらに関わるものの以外のツイートはほぼ抽出されない。しかし、「ZARA」や「GU」を含むツイートを取得すると、アカウントにそれらの文字列が含まれていたり、ツイートの中にそれらの文字列が含まれているが ZARA や GU とは関係のないものが抽出されてしまう。

このような、意図せざるデータの抽出は、非日本語のみならず日本語であっても起こりうる。たとえば「さんま」を検索したときに「○○さんまね」のように、「さんま」を含むが「秋刀魚」ではないツイートが抽出される場合が想定される。

Twitter を研究資源として研究を行う場合、検索語が計画の意図に沿ったものを取得するものであるのか、予備調査をするなど事前に検討しておかなければならない。

## 2-2. Twitter のシステムに関わる問題

### 2-2-1. 文字コード

ウェブ上のデータを取得する際、思わぬところでひっかかるのが文字コードの問題である。稿者は Windows ユーザであるため、たとえばウェブサイトにあるテキストをコピーアンドペーストする場合、その文字コードは Shift-JIS（以下、SJIS）である。RStudio でスクリプトを実行してツイートを取得するときも、基本的には SJIS のデータを取得する。しかしながら、タスクスケジューラを利用するなどして RStudio 以外のツールを作業に組み込むと、スクリプトの文字コードが SJIS ではなく UTF-8 となってしまうことがある。

ツイートを取得した際に文字化けを起こしている場合、どこかで文字コードのエラーが発生している。この場合、エラーの発生源を段階的に確認しなければならない。そののち、エラーを回避するため、文字コードを変換するように設定する必要がある。たとえば、稿者はツイートを取得する場合、自動化を行うため、

- 1) タスクスケジューラでバッチファイルを起動する
- 2) バッチファイルで R の Rscript.exe を起動する
- 3) R の source 関数でスクレイピング用スクリプトを起動する
- 4) スクレイピング用スクリプトでツイートを取得する

と 4 段階の手順を設定している。このうち、3) の段階で UTF-8 でエンコードを行うコードをスクリプトに加えている。ただし、Windows マシンであっても、この設定が不要である場合もある。

取得したデータは UTF-8 であるため、そのまま形態素解析器にかけると適切な分析がなされない。そのため、取得したデータを一度、SJIS に変換する作業が必要となる。

## 2-2-2. タイムゾーン

ツイートを取得するとき、変数として発信日時を取得する必要があるとする。その場合、日本でツイートを取得しても、取得される発信日時は日本標準時ではなく、グリニッジ標準時である。そのため、取得したデータを日本標準時に変更する必要がある。

R の使用に習熟している場合はスクリプトの中で変更すればいいが、(稿者のように) プログラミングに疎い場合は、ファイルを直接操作する必要がある (稿者は、取得した発信日時に対してエクセルの time 関数を使用し、日本標準時を新たな変数として設定している)。

### 2-2-3. ツイートの削除、鍵付きアカウントおよび再現性の問題

ツイートは一度発信したのち、任意のタイミングで削除することができる。そのため、ある時点では Twitter API で取得できたツイートが、そのツイートが削除されたあとは取得できなくなる。

また、ツイートは誰にでも見られる状態が基本ではあるが、自分のフォロワーの範囲のみに閲覧を制限することもできる。そのような状態のアカウントのことを鍵付きアカウントと呼ぶことがある。Twitter API では当然ながら、鍵付きアカウントのツイートは取得できない。

たとえば、非鍵付きアカウントがある時点から鍵付きアカウントに設定変更されたとする。その場合、任意の時点ではその非鍵付きアカウントのツイートが取得できていたにもかかわらず、設定変更以降は一切のツイートが取得できなくなる。

ツイートの削除および鍵付きアカウントのツイートは、Twitter API に関わる問題、再現性の問題として認識しておく必要がある。

さらに、ツイートの取得期間も再現性の問題として挙げられる。先述のとおり、Twitter API で取得できるツイートは、取得時点にさかのぼって1週間前までのものである。そのため、それ以前にさかのぼってツイートを取得したい場合は、専門の業者に依頼するなど、他者の力を借りることになる。また、業者に依頼しても、必ずしも全てのツイートは取得できない。

さらに、研究を行うものが自ら Twitter API を利用してツイートを取得する場合であっても、全てのツイートが取得できるわけではない。任意のワードを含むツイートが多いほど、取りこぼしは多くなる。そのため、たとえば衆議院選挙の時期に「衆院選」を含むツイートを取得するとしても、取得できなかったツイートがあるものとして研究を進める必要がある。

取得されるツイートの数は常に取得条件の100%に満たないものである、と理解した上で分析を行わなければならない。しかも、欠けたツイートが全

体のいかほどであるかを知ることは難しい。

#### 2-2-4. ハッシュタグ

ツイートを発信する際、任意の話題であることを示すための指標としてハッシュタグの付される場合がある。たとえば、「明治大学」に関わる話題のツイートをする場合、「# 明治大学」とする。ブラウザや公式アプリ等で Twitter を利用している場合、そのハッシュタグをクリックすることで同様のハッシュタグを含むツイートをまとめて見ることが可能となる。Twitter API でツイートを取得する際は、テキストの一部としてそのまま抽出される。

API では「c (" 明治大学 ")」のように、ハッシュタグをひとつの変数として取得することができる。そのため、ある検索語を含むツイートに対してどのようなハッシュタグが付されるのかを研究の要素として設定することも可能である。

このように、ハッシュタグを研究上の利点として利用する考え方のある一方、ノイズと判断する立場もありうる。研究計画を設定するなかで、ハッシュタグの扱いについての対応を決定しておく必要がある。

#### 2-2-5. 絵文字、URL、リプライの表記

ツイートを抽出してそのままテキストマイニングを行うと、絵文字や URL がノイズとなる。絵文字は「<U+0001F3C9>」のように「<U.」で始まり「>」で終わる文字列として表示される。URL は「https://t.co/.」のあとに 10 字の半角英数文字が続くものに変換される。そのため、いずれも正規表現で削除することが可能である（稿者は R のスクリプトの中に絵文字と URL を削除するコードを設定している）。

また、リプライについても、@ のうしろに続く文字列（アカウント名）がノイズとなることがある。リプライとは、あるアカウントの発信に対して別のアカウントが返信するものをいう（Twitter API では reply\_to\_user\_id



として、リプライの相手を取得することも可能である)。データを分析する際、リプライとなるツイートそのものを削除するのであればよいが、リプライであることは明確にしつつ、@ 以下の ID を削除したい場合、同じく正規表現を使用することでノイズとなる文字列を削除することができる。

なお、リプライの扱いについては、2-2-7. で再考する。

### 2-2-6. 重複するツイート

ツイートを分析する際、重複するツイートをどのように扱うかを事前に考えておく必要がある。そのひとつに、リツイートがある。

リツイートとは、あるツイートについて、発信者を明らかにした上でツイートをを行ったアカウントあるいは他のアカウントがそのツイートを繰り返し発信するものである。たとえば、アカウント X が「〇〇は A である」とツイートしたとする。そして、他のアカウント Y がそのツイートをリツイートする。

このとき、A を検索ワードとしてツイートを取得すると、「〇〇は A である」とするツイートが 2 件抽出される。リツイートの行われる回数が少なければ問題とならないが、数十を超えると、少なからず影響がでる。つまり、分析の際に「〇〇」と「A」との共起頻度が高いものとして検出される。延べの共起は多いものの異なりの共起が少ないのであれば、後者を正確な情報として分析するべきであると稿者は考える。

なお、重複するツイートはリツイートのみではない。引用ツイートもリツイートに同じであるし、正式なリツイートではないが、同一内容で複数回発信されるツイートなども同様である。

もしも、重複したものであってもひとつひとつのツイートとしてみなすのであれば、取得したデータをそのまま使用すればよい。しかし、同一内容の重複を排除したいのであれば、なにがしかの対策をとる必要がある。なお、稿者は、R でデータを取得する際、リツイートと引用ツイートは除外するよ

うにフィルタリングを行っている。また、取得期間中にリツイート・引用ツイート以外で重複するツイートがあった場合、1件のみを残して他のツイートを削除している。

これら、重複するツイートの影響を考える目的から、2022年1月1日0時台から同年1月7日23時台までに発信された「明治大学」を含むツイート3725件をモデルとして考えてみる。このモデルとするツイート群を、以下、「明治大学データ」と呼ぶ。ただしリツイートと引用ツイートは除外している。

図1は、明治大学データに何も手を入れずに KH Coder で作成した共起ネットワークである。なお、下処理時の強制抽出語は「明治大学」のみ、また以下の共起ネットワークはすべて Jaccard 距離で作成している。

あまりに雑多すぎて、傾向を読み取ることは不可能である。

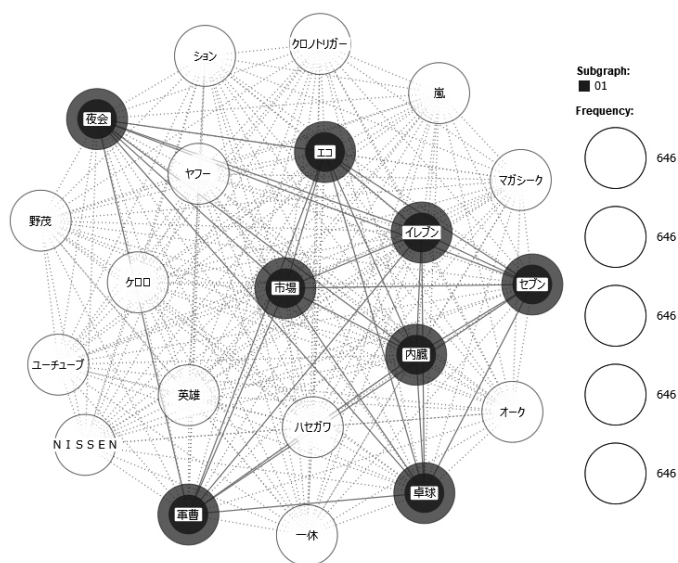


図1 「明治大学」を含むツイートの共起ネットワーク



## 2-2-7. リプライの扱い

リプライは会話である。たとえば、アカウント X が「〇〇は A だ」とツイートしたのに対し、アカウント Y が「@ アカウント X 〇〇は A ではなく B だ」と発信したとする。そして、検索ワードを A としてツイートを取得した場合、アカウント X とアカウント Y のツイートの両方、つまり会話全体を取得することができる。しかし、検索ワードを B とした場合、取得できるのはアカウント Y のもののみである。このとき、会話の一部を切り取ったことになる。

このように、リプライを抽出した場合、断片的な会話を取得する可能性が高い。そのため、もしも会話の文脈から切り離されたツイートを分析の対象外とするのであれば、リプライを削除する必要がある。図 3 は、明治大学

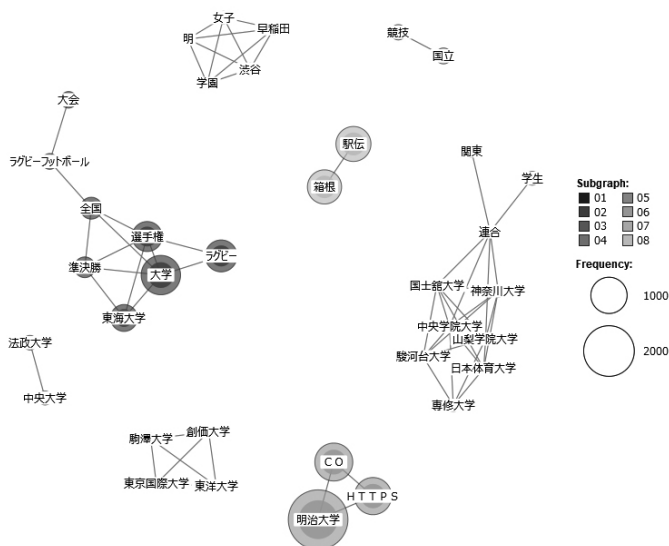


図3 リプライを削除した共起ネットワーク

データから重複ツイートを削除したのち、リプライを削除したツイートについての共起ネットワークである。

内容としては、リプライを含むもの（図2）とほぼ同じであるが、中学受験に関わるものについてわずかに変化を観測することができる。

なお、@ を含まないツイートであっても、そのあとにリプライが発信された場合、「つぶやき」から「会話」に変化する。そのため、@ を含まないツイートもまた、会話の一部である可能性は否定できない。また、リプライであるか否かを二項の変数として設定すれば、分析の幅が広がる。そのため、リプライの扱いについては、研究計画のなかで十分な検討が必要となる。

## 2-2-8. 同一アカウントによる複数ツイート、宣伝ツイート

リツイートなど重複するツイートとは別の問題として、宣伝を目的としたツイートの扱いを考える必要がある。たとえば「明治大学」を含むツイートの場合、大学スポーツなどのニュースサイトへのリンクを張ったアカウントのものが検出される。このようなツイートをそのまま利用して分析を行うのか、もしくは宣伝ツイートを可能な限り排除するのか、研究計画をたてる時点で考えておく必要がある。

稿者の場合、宣伝ツイートを排除する目的から、ツイートを発信するツール（Twitter API では source）について、iPhone もしくは Android 端末のもの（Twitter for iPhone および Twitter for Android）のみを残し、他をすべて削除することにしている。これにより、bot などが自動的に発信するツイートを除外することが可能である。

また、発信ツールが iPhone や Android であっても、宣伝を目的とするツイートを発信しているアカウントもある。そのようなアカウントの特徴として、同一日に同一あるいは類似したツイートを複数回発信していることを指摘できる。稿者は、同一アカウントによる同一日のツイートを1件のみ残し、

他を削除している。

図4は、明治大学データから重複ツイートを削除し、iPhone および Android 端末から発信されたもののみを残したものである。なお、同一日において2回以上ツイートしているアカウントはなかった。

先にみた、重複ツイートを削除したもの、あるいはリプライを削除したものと比較すると、中学受験に関わる内容がみられなくなった。また、先の分析でも傾向はあったが、おおむね箱根駅伝の上位(10位以上)と下位(11位以下)とがグループになっていることがわかる。

このように、ツイートを削除するか否かで、みえてくるものは変わることがある。Twitter を分析する際のデータクレンジングは、研究計画と密接な関係をもつため、事前に十分な検討を行った上、分析に合わせてその都度対応する必要がある。

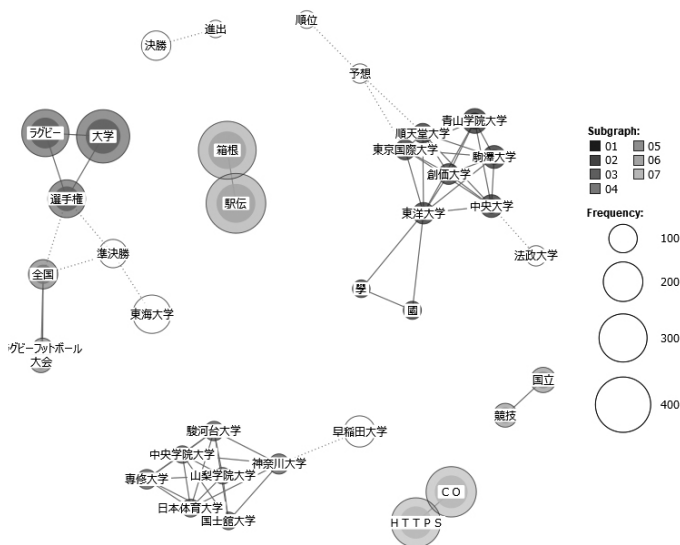


図4 共起ネットワーク

### 2-3. データの利用についての問題

先にみたとおり、本稿執筆時点で Twitter を利用した研究論文はすでに数千件にのぼる。利用の方法は様々で、本稿で想定しているようなテキストマイニング的手法に基づくものはもちろん、Twitter 利用の分析研究などもある。前者は「Twitter で研究するもの」、後者は「Twitter を研究するもの」といえる。

これらの研究において、ツイート自体の引用を行っているものもみられるが、倫理面からの言及がなされているものは少ない。わずかにみられる言及としては、たとえば四方田（2021）が挙げられる。四方田（2021）では、「公開データのみ収集を行い、投稿内容と投稿日時のみ収集にとどめ、投稿内容の引用掲載は行わず要約した内容を掲載する」ことを指針として示す。

ツイートを分析して研究を公にする場合、ツイート自体の引用は極力避けるべきであると稿者は考える。テキストマイニングツールを利用すれば、個々のツイートはむしろ、引用することがない。しかしながら、用例としてツイートを示す必要のある研究もあるだろう。その場合、倫理面に配慮して研究を行わなければならない。

### おわりに

以上、研究資源として Twitter を利用することについて、その利用方法の概略と問題点について述べた。

ここまでみたとおり、Twitter のデータ利用は、初期設定の難度が高い。しかし、それさえ済んでしまえばきわめて自由に利用することができる。また、その利用はいまだ切り拓かれ始めたばかりであり、さまざまな応用が考えられる。

現在、Twitter を資源として用いる研究は、多くが社会学あるいは自然科

学の分野によるものである。稿者のような人文学の研究者による研究はきわめて少ない。

人文学、特に日本語学ではコーパスを利用する研究が多い。すでに「中納言」のように、すぐれた研究環境が整っており、これを利用するものもある。しかしながら、既存のコーパスでは進めることのできない研究もあるはずで、そのような場合に大規模なテキストを容易に収集できる Twitter は大きな可能性を有す。

2-1-3. にみたとおり、Twitter API では用言の取得に難がある。その点は研究にとって大きな足かせであるかもしれない。しかし、活用しない語であればそのような制約は緩む。また、用言であっても表現が固定している場合、表出形のみでの検索であっても問題とならないことがある。たとえば、芦木亜彩湖 (2019) のように、「大丈夫です」の使用をみたいのであれば、「大丈夫でし (た)」などを検索する必要がある。このような場合、Twitter を用いたコーパスの作成を行うこともできる。

Twitter はことばの集積である。であれば、ことばを主たる研究対象とする人文学において、その利用はより積極的に行われるべきではないか、と考える次第である。

### 参考文献

- 芦木亜彩湖 (2019) 「ウェブコーパスを用いた「大丈夫です」の使用に関する実態調査 — 勧誘に対する拒否としての「大丈夫です」—」『実践國文學』96
- 石田基広 (2017) 『R によるテキストマイニング入門 (第2版)』森北出版
- 石田基広 (2020) 『実践 R によるテキストマイニング: センチメント分析・単語分散表現・機械学習・Python ラッパー』森北出版
- 上ノ原秀晃 (2019) 「2017 年衆院選とソーシャルメディア —— 候補者によるツイッター投稿の内容分析——」『人間科学研究』40
- 金明哲 (2021) 『テキストアナリティクスの基礎と実践』岩波書店



- 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
- 曹慶鎬 (2018)「インターネット上の災害時「外国人犯罪」の流言に関する研究：熊本地震発生直後の Twitter の計量テキスト分析」『応用社会学研究』60 巻
- 樋口耕一 (2020)『社会調査のための計量テキスト分析』ナカニシヤ出版
- 樋口耕一・中村康則・周景龍 (2022)『動かして学ぶ! はじめてのテキストマイニング フリー・ソフトウェアを用いた自由記述の計量テキスト分析 KH Coder オフィシャルブック』ナカニシヤ出版
- 松本義之・井上仙子 (2020)「下関地域における Twitter を利用した観光情報分析」『バイオメディカル・ファジィ・システム学会誌』22 巻 2 号
- 村井源 (2012)「東日本大震災後の Twitter 利用傾向 ——震災関連ハッシュタグの計量的分析——」『情報知識学会誌』22 巻 2 号
- 吉見憲二・上田祥二・針尾大嗣 (2019)「問題投稿に付帯されるハッシュタグの特徴 ——売買春を事例として——」『第 81 回全国大会講演論文集 (情報処理学会)』1 号
- 四方田健二 (2021)「新型コロナウイルス感染拡大に伴う休校に対する社会的関心：Twitter 投稿内容の計量テキスト分析と感情分析」『名古屋学院大学教職センター年報』5 号
- 渡邊憲二・箕輪弘嗣 (2021)「COVID-19 における Twitter の利用傾向に関する探索的研究」『情報知識学会誌』31 巻 2 号
- 総務省「平成 24 年 情報通信メディアの利用時間と情報行動に関する調査」  
([https://www.soumu.go.jp/iicp/chousakenkyu/data/research/survey/telecom/2013/00h24mediariyou\\_gaiyou.pdf](https://www.soumu.go.jp/iicp/chousakenkyu/data/research/survey/telecom/2013/00h24mediariyou_gaiyou.pdf)) (2022 年 5 月 31 日アクセス)
- 総務省「令和 2 年度 情報通信メディアの利用時間と情報行動に関する調査」  
([https://www.soumu.go.jp/main\\_content/000765135.pdf](https://www.soumu.go.jp/main_content/000765135.pdf)) (2022 年 5 月 31 日アクセス)

(うえだ・ばく 政治経済学部准教授)