

インタラクティブなデータ・ヴィジュアルライゼーション・ツールを用いた Twitter データのクラスタ分析

Cluster analysis of Twitter Data, using Interactive Data visualization Tool.

和田 伸一郎^{*1}
WADA, Shinichiro

立教大学大学院社会学研究科
Graduate School of Sociology, Rikkyo University.

Abstract: This study attempts cluster analysis of Twitter data posted on Tokyo Governor's Election held in 2016, using Python (July 13 - August 1, 2016, 4.8 million tweets, 170 million words). For cluster analysis, words were vectorized using gensim version word2vec algorithm which is a library of Python, and attempt to visualize clusters in three dimensions using t-SNE (t-distributed Stochastic Neighbor Embedding) which is dimensionality reduction algorithm. In particular, in this research, we used the data visualization tool Embedding Projector for clustering. By using this tool, we attempted to visually identify clusters by moving the three-dimensional space interactively while visualizing the dynamic learning process in the three-dimensional space. As a result, we could identify multiple clusters with high accuracy. This made it possible to clarify what in this election Twitter users were interested in.

1. はじめに

本研究では、2016 年 7 月に行われた東京都知事選挙に関する Twitter データのうち、選挙期間中(とその前後一日ずつを含む)に投稿されたものを全数収集し、クラスタ分析を行うことによって、選挙に関する、どのような投稿がなされたのかを分析した。

2. データと前処理について

2.1 収集した 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。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、表1のようなデータを csv ファイルにて収集することができた。RT は公式リツイートのみ数、OT はオリジナルツイートを指す。

表1

	総ツイート数	語彙数
ALL	4,825,560	199,287
RT	3,588,302	123,192
OT	1,237,258	187,394

なお ALL データ(RT データと OT データを足したすべてのデータ)の単語数は、177,439,525 語となった。

この表からも分かるように、このデータのうち、約 74%がリツイートからなっていた。語彙数(単語の種類)でみても、RT の語彙数が OT よりも少ないのは、同じツイートが数多くリツイートされているためである。したがって、Twitter データの場合、しばしばテキストマイニングで行われる、単語の出現回数を出しても、あまり意味がない。多くリツイートされた文章に含まれる単語の出現回数が多くなってしまったためである。

そこで、単語の類似度(意味の近さ)を計算する、Python のライブラリ gensim に実装されている word2vec を使って、単語のベ

クトル化を行い、どのような類似する単語群、つまりクラスタがあるかを調べることにした。なお、単語ベクトル化をする際に設定した word2vec のオプションの値は以下の通りである。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)。

2.2 データの前処理(形態素解析)

その前に、これらのデータをそれぞれ形態素解析する必要がある。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 SNS 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、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」なのか分からない。つまり、「小池」と「百合子」、「晃」が分解されて学習されてしまう。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桜井翔の父親)なのか、保守派論客である「桜井よしこ」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。これだと、いくら word2vec, t-SNE などのアルゴリズムの精度が高くても、学習結果が混乱したものになる可能性が高くなってしまふ。

連絡先: 和田伸一郎 立教大学大学院社会学研究科, 東京都豊島区西池袋3-34-1

3. word2vec と Embedding Projector による可視化

約 20 万もある語彙の分布のどこにクラスタがあるのかを調べることは、単語ごとに類似する単語をリスト化してくれる word2vec の Python スクリプトを実行するだけでは難しい。例えば、約 20 万もの単語すべてに類似単語リストをつくることは、現実的に難しい。

そこで、コーパス全体を可視化した上で、その全体の中に、どのようなクラスタが現れているかを視覚的に把握するツールとして、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、TensorBoard のスタンドアローン版 WebUI である Embedding Projector を用いることにした。

word2vecを使って 300 次元で単語ベクトル化したコーパスを、Embedding Projector 上で、次元圧縮アルゴリズムである、PCA、t-SNE を用いて、選定した特徴語とそれに類似する 1000 語からなる、三次元のローカル空間を可視化し、クラスタを出すことを試みた。特徴語として、選挙期間中に大きな争点となっていた「待機児童」、「介護」、また、全国的に注目されていた「オリンピック／パラリンピック」などを選定した。なお、「待機児童」、「介護」、「オリンピック／パラリンピック」を含んだツイートは、それぞれ、43,794、26,955、78,006 存在した。

図1は、「待機児童」をターゲット単語として設定し、2000 回学習した結果、収束した、その単語と意味の近い 1000 語からなる三次元のローカル空間である。

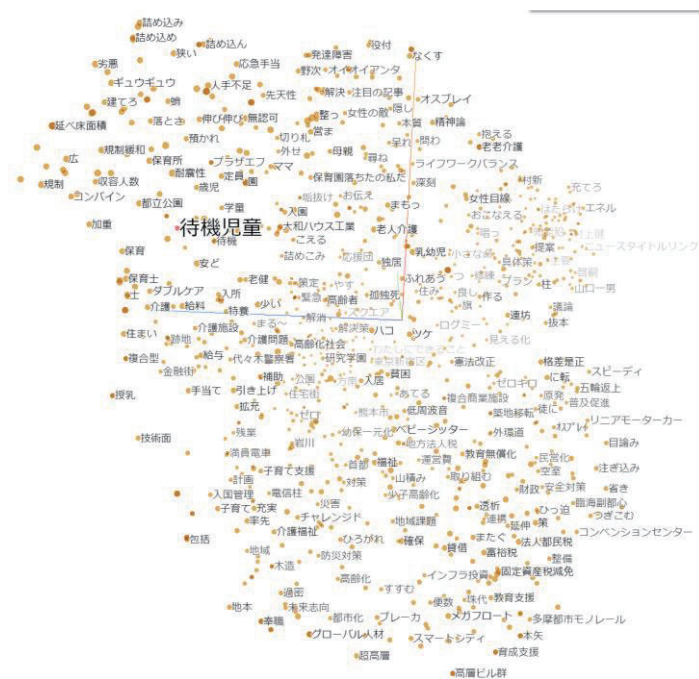


図1.「待機児童」ローカル空間

Embedding Projector では、学習中であってもインタラクティブ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。このツールの大きなメリットとして挙げることができるのは、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、このツールが一定程度、取り除いてくれることである。

4. 分析結果と課題

いくつか特徴語を選定し、学習を行った結果、日本語でも高い精度でクラスタを確認することができた。この精度の高さは、まずは word2vec, PCA, t-SNE のアルゴリズムの精度の高さによりもたらされたものであるが、加えて、NEologd 辞書が、例えば、候補者や有識者などの名前や政党名、政治団体名、イベント名、地名、市民会館などの建物の名前といった固有名詞を一語で拾ってくれたことからもたらされたものでもあることが分かった。

他にも、Instagram のテキストデータを学習させ、高い精度でク
ラスタが出ることも確認できた。

課題としては、今回、word2vec に用意されているオプションの値を変えて、それぞれのデータで、最も精度が高い学習結果を比較する余裕がなかったことを挙げるができる。これについては、今後検討したい。

参考文献

- [Smilkov 16] Google Developers (Smilkov, Daniel, Viégas, Fernanda, Wattenberg, Martin). : A.I. Experiments: Visualizing High-Dimensional Space. <https://www.youtube.com/watch?v=wvsE8jmlGzE&feature=youtu.be> (2016)
- [木田 17] 木田勇輔. : 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, 榎山女学園大学, 第 17 巻, pp.83—92 (2017)
- [Kudo 13] Kudo Taku. : MeCab:Yet Another Part-of-Speech and Morphological Analyzer, <http://taku910.github.io/mecab/> (2013)
- [Maaten 08] Maaten, Laurens van der, and Hinton, Geoffrey. : Visualizing data using t-SNE. Journal of Machine Learning Research, Vol 9(Nov), pp. 2579—2605. (2008)
- [Mikolov 13a] Mikolov, Tomas, Chen, Kai, Corrado, Greg, Dean, Jeffrey. : Efficient estimation of word representations in vector space”. CoRR, abs/1301.3781. (2013)
- [Mikolov 13b] Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Corrado, Gregory S., Dean, Jeffrey. : 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)
- [Sato 15] Sato Toshinori. : Neologism dictionary based on the language resources on the Web for mecab-ipadic”, <https://github.com/neologd/mecab-ipadic-neologd/> (2015)
- [Smilkov 16] 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 (2016)
- [山縣 18] 山縣史哉, 梅原英一 : 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 (2018)