トピックモデル可視化ツールの開発に向けて

黒田　絢香

大阪大学大学院言語文化研究科

〒 560-0043 豊中市待兼山町 1-8

E-mail:　kuroda22a@gmail.com

あらまし　本研究の目的は，トピックモデルの一つである LDA を用いた文学作品の分析をサポートするため，効果的なビジュアライゼーションツールを開発することである．数多くの文書とトピックから成る高次元で複雑なトピック構造から有益な情報を見つけ出すことは非常に労力を要するものであり，また最適なトピック数の選択やモデルの機械的な評価が難しい現状において，試行錯誤のプロセスは必要不可欠と言える．そこで本研究では，モデル出力結果から複雑な設定をすることなく直接グラフを描画することのできるトピック可視化ツールを目指す．複数のモデルを比較しながら探索的な分析を行うことのできるインタラクティブなツールを開発することは，LDA を用いる分析者に大いに寄与すると考えられる．

キーワード　LDA，トピックモデル，ビジュアライゼーション，ツール開発

Development of an Interactive LDA Visualization Tool

Ayaka Kuroda

Graduate School of Language and Culture, University of Osaka

1-8 Machikaneyama-cho, Toyonaka, Osaka, 560-0043 Japan

Abstract　Textual analysis using LDA, a generative probabilistic topic model, is recently attracting attention among the field of literary analysis. However, we have had difficulty in choosing the number of topics when it comes to long literary texts, because each of them usually contains various topics. Thus trial and error is required in the analysis process.

We, therefore, attempt to develop an interactive LDA visualization tool which can compare multiple models we created by changing the number of topics. The aim of this tool is to make the trial-and-error process easier, and to help our understanding of complex but meaningful topical structures.

Keywords　LDA, topic model, visualization, tool development

1.　はじめに

1.1.　文学作品とトピックモデル

近年，機械学習アルゴリズムの一つであるトピックモデルをテクストマイニングに応用する研究が数多く見られる．トピックモデルには author-topic model (Rosen-Zvi et al. 2004) や dynamic topic model (Blei and Lafferty 2006) など様々な種類が存在するが，それらの基礎と

なった LDA (Latent Dirichlet Allocation) (Blei et al., 2003) は，最も有名かつ最も使用されるトピックモデルの一つであると言える．

LDA は文学作品の分析にも用いられており，従来の質的な研究にはなかった新たな観点から作品を見直すことができるとして注目されている．例えば，Jockers and Mimno (2013) では，19 世紀のフィクションを集めたコーパスに対してトピックモデルを適用することで，男性著者，あるいは女性著者が特徴的に使用する話題を発見した．また田畑 (2017) は，小説だけでなく様々なレジスターの文書が含まれた大規模コーパスである FLOB コーパスに対しても，そのコーパスの背後にある潜在的な意味構造を読み解く上でトピックモデルが有用であることを示した．

文学作品のように人文学の領分とされてきた研究対象に対して情報工学の手法を持ち込む，いわゆるデジタル・ヒューマニティーズと呼ばれるアプローチは，近年のテクノロジーの進化に伴いより活発となっており，質的研究が中心であった分野に新たな知見を数多くもたらしている．

1.2. 最適な設定を選ぶには

LDA を実行する際に，分析者は抽出するトピックの数を任意に設定することができる．分析対象の文書が少ない場合には 10 前後とすることが多いが，規模の大きなデータを扱う場合では 100 から 500 ものトピックを設定することもある．一般的には，この値が小さすぎれば抽出されるトピックは汎用的な語ばかりが含まれる解釈の難しいものとなり，一方で大きすぎれば過剰に細分化されるあまり，比較的出現頻度が低い各文書に特有の語が多く出現してしまうと言われている．そのため，実験対象コーパスの規模や分析目的に応じて適切な数値を都度判断する必要がある．

最適なトピック数の決定方法については数多くの議論がなされているが，どのような場合でも使用できる普遍的な最適解は未だ存在しない．現状ではトピックモデルの評価指標である Perplexity と Coherence の値を参考にして決定されることが多いが，この手法には問題点も存在する．

まず，Chang et al. (2009) が指摘した通り，機械学習モデルの精度を確率論に基づいて示す値である Perplexity は予測モデルの評価において有用であるものの，探索的な分析を行なうトピックモデルには適合しないことが多い．

一方 Coherence ではトピックの質，つまり生成されるトピックに含まれる単語の一貫性，解釈可能性を評価する．「人間にとってわかりやすいかどうか」という定量化しにくい評価を機械的に測定する必要があるため，その算出方法は物議を醸している．例えば Newman et al. (2010) のように WordNet や Google などの外部コーパスをベースとして単語間の類似度を計算する手法もあれば，Mimno et al. (2011) のように各トピックの頻出単語同士の共起頻度をもとに算出し，外部データを利用しないものもある．Lau et al. (2014) ではこれらを含めた複数の Coherence 算出方法を比較評価しているが，いずれも，特にトピックベースでの評価の場合，人間による評価とは必ずしも一致しないと結論づけられている．

また，Sbalchiero and Eder (2020) もトピック数の決定方法について検討を行っているが，学術論文のアブストラクトやニュース記事など単一の話題について書かれた文書を用いた実験が数多く行われている一方で，長編小説のように多様な話題で構成されている比較的長い文書については十分に検討されていないと指摘している．

以上のことから，トピック数を選択する際には，複数のモデルを構築した上で比較評価を行なう試行錯誤プロセスが必要不可欠であると言える．Perplexity や Coherence の値は参考になるものの，決定打とすることは難しい．実際に生成されたモデルを確認し，どのようなトピックが生成されているか，どれほどのトピックの粒度が分析に最適かを判断し，ときには複数のモデルによる結果を並行して表示しながら分析を進めていくことが重要であると考えられる．

1.3. LDA ビジュアライゼーションツール

前述のような試行錯誤プロセスにおいて，モデルビジュアライゼーションの効率化は重要な課題である．何百，何千もの文書と数多のトピックによって構成される高次元で複雑なトピック構造から分析に有用な情報を見つけ出すには的確な可視化が不可欠であり，バブルチャートやヒートマップ，ワードクラウド，ネットワークグラフなど様々なグラフが使用されてきた．しかし，そのグラフ描画プロセスが複雑で時間のかかるものであったり，大量，もしくは複雑なグラフによって表現されるものであるならば，特に複数のモデルを作成して比較する場合，分析には多大な労力を要する．

これまで幾多のトピックモデルのビジュアライゼーションシステムが開発されてきたが，中でも Sievert and Shirley (2014) によって提案された LDAvis[1]，Python 版である pyLDAvis[2]は代表的な LDA 可視化ツールであると言える．その開発の主眼は"both compact and thorough is interactivity" (Sievert and Shirley, 2014) とされており，インタラクティブなグラフを描画することによって，画面遷移を挟むことなく大量のトピック情報をコンパクトにまとめ上げることに成功している．

しかし一方で，多くとも 50 前後のトピック数を想定しているため抽出トピック数が大きすぎる場合にグラフが煩雑になりすぎる点や，各トピックがどの文書にどれほどの確率で出現しているかを示すことができない点など，限界も存在する．

1.4. 研究の目的

本研究の目的は，文学作品の分析を視野に入れた LDA のビジュアライゼーションツールを開発することである．前節までの内容を踏まえ，このツールの目標を以下の 4 つに設定する．

- 静的なグラフではなく，インタラクティブな出力であること
- トピック数を変化させて実験した複数のモデルの出力結果を比較できること
- 抽出するトピック数が 100 を超える場合でも対応できること

[1] https://github.com/cpsievert/LDAvis
[2] https://github.com/bmabey/pyLDAvis

- 各トピックの出現単語とその比率だけでなく，各トピックがどの文書に出現しているかを示すこと

このようなツールを開発することで，多大な労力と試行錯誤を要するトピックモデル分析をより効率的なものにし，文学作品の量的研究に寄与することができると考えている．

2. 手法

2.1. データと前処理

本稿では，サンプルコーパスとして Arthur Conan Doyle の長編小説コーパスを用いる．収録されているのは 29 作品，推理小説や歴史小説，ノンフィクションなど 5 種類のジャンルが含まれており，総語数は約 2,180,000 語である．

このコーパスに対し，Jockers and Mimno (2013) で用いられていた手法を参考に，以下の前処理を行った．

- チャプタータイトルや注など本文以外の要素を削除
- TreeTagger[3]によるタグ付けに基づき普通名詞のみを取り出す
- 1,000 語ごとのブロックに分割する

以上を行った結果生成された 409 ファイルが本稿での分析対象である．

2.2. MALLET の出力構成

モデリングには，Java ベースの自然言語処理ツールキットである MALLET[4]を用いた．このツールではモデルを様々な形式のファイルで出力できるが，今回必要となるのは主に (1) output-doc-topics コマンドによって生成される，各ドキュメントにおけるトピックの出現確率を示すデータ， (2) topic-word-weights-file コマンドによって生成される，各トピックにおける頻出単語リストとそれらの重みを示すデータ， (3) diagnostics-file コマンドによって生成される，複数の指標に基づいて各トピックの品質を示すデータの 3 種類である．

特に (3) の diagnostics データは，トピックに割り当てられた単語の数を示す tokens，トピックを指定された際の文書の予測確率からそのトピックの一般性を計る document_entropy，トピックの単語分布とコーパス全体の単語分布との距離を示す corpus_dist，Mimno et al. (2011) の手法に基づく coherence など，12 種類の指標を用いて各トピックの特徴やモデル全体における位置づけを表す有用なファイルである．

[3] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
[4] http://mallet.cs.umass.edu/

2.3. ツール開発

開発では，Python のグラフ描画ライブラリの一つである Plotly[5]，及びそれを組み込むことのできる Web アプリケーションフレームワークの Dash[6]を主に用いた．これらの特徴はインタラクティブなグラフを HTML 形式で出力できることにある．グラフの拡大や縮小，ドロップダウンメニューやスライダーなどによるグラフコントロール，クリックやホバーによるイベントの発生など，様々なカスタマイズが可能である．

3. 開発中ツールの概観と特徴

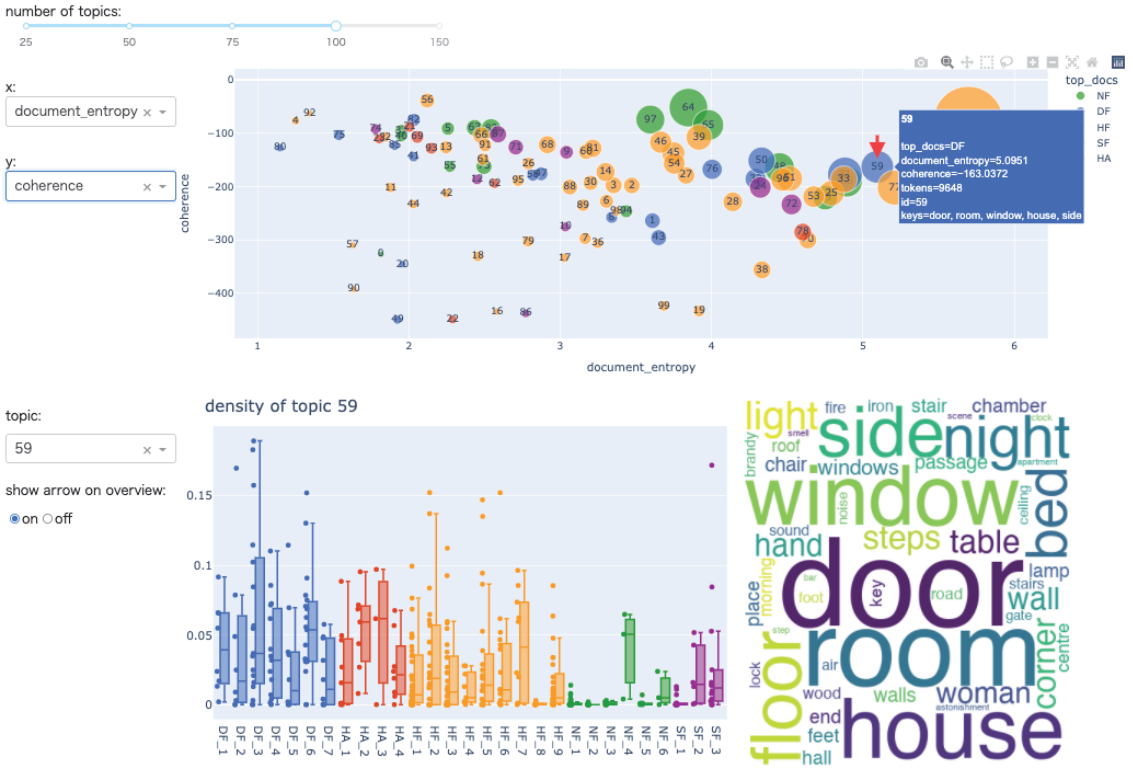

図 1: 開発中ツールの概観

開発中ツールの概観を図 1 に示す．大きく分けて 2 つの表示部があり，上部は前述の Diagnostics を参照してモデルの全体像を表す model view，下部は選択されたトピックの詳細をボックスプロット・散布図とワードクラウドで示す topic view である．model view でのマウスホバーによって各トピックのデータがツールチップの形で表示され，また下部の topic view に即座に反映されるため，どのようなトピックが存在するか，モデルの全体像を見ながら探索的に分析することが可能となっている．

5 https://plotly.com/python/
6 https://dash.plotly.com/

3.1. model view

前述の通り，上段の model view では Diagnostics に表示される 12 種類の指標に基づいてバブルチャートがプロットされる．バブルサイズは tokens に対応しており，大きなものほどトピックに含まれる単語数も多くなる．

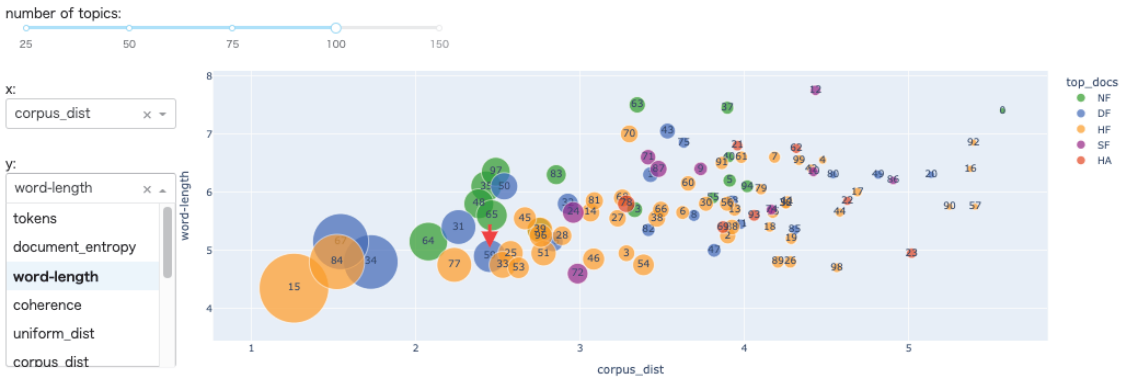

図 2: model view の x 軸と y 軸の変更

x 軸と y 軸に設定する指標はツール使用中に切り替えることが可能である．例えば図 1 では x 軸を document_entropy，y 軸を coherence に設定しているため，数多くの文書に出現する一般的なトピックは右に，比較的マイナなトピックは左に配置され，上部のものほど一貫性 (ここでの一貫性は単語の意味に基づく評価ではなく，構成要素の単語同士が同じ文書に出現する確率に基づく評価である) が高い．図 2 のように指定すれば，コーパス全体の単語分布に近い，つまり一般的なトピックほど左に配置されるようになる．y 軸で word-length を指定しているため，上部には平均単後長が長い高尚でアカデミックなトピックが配置されやすい．

ここに挙げた以外にも様々な指標が組み込まれているため，x 軸と y 軸の設定を切り替えることで，自身の分析目的に適うトピックを探し出すことが容易となる．

また，Plotly の特徴として，グラフの一部分をマウスドラッグで拡大することができる．そのため tokens の値が低い故に小さなバブルで表示されているトピックを選択する際，密集地帯であっても問題なく目的のトピックを探し出せる (図 3 を参照)．この機能は，トピック数の設定が 100 以上の大きな数値であった場合，特に真価を発揮すると考えられる．

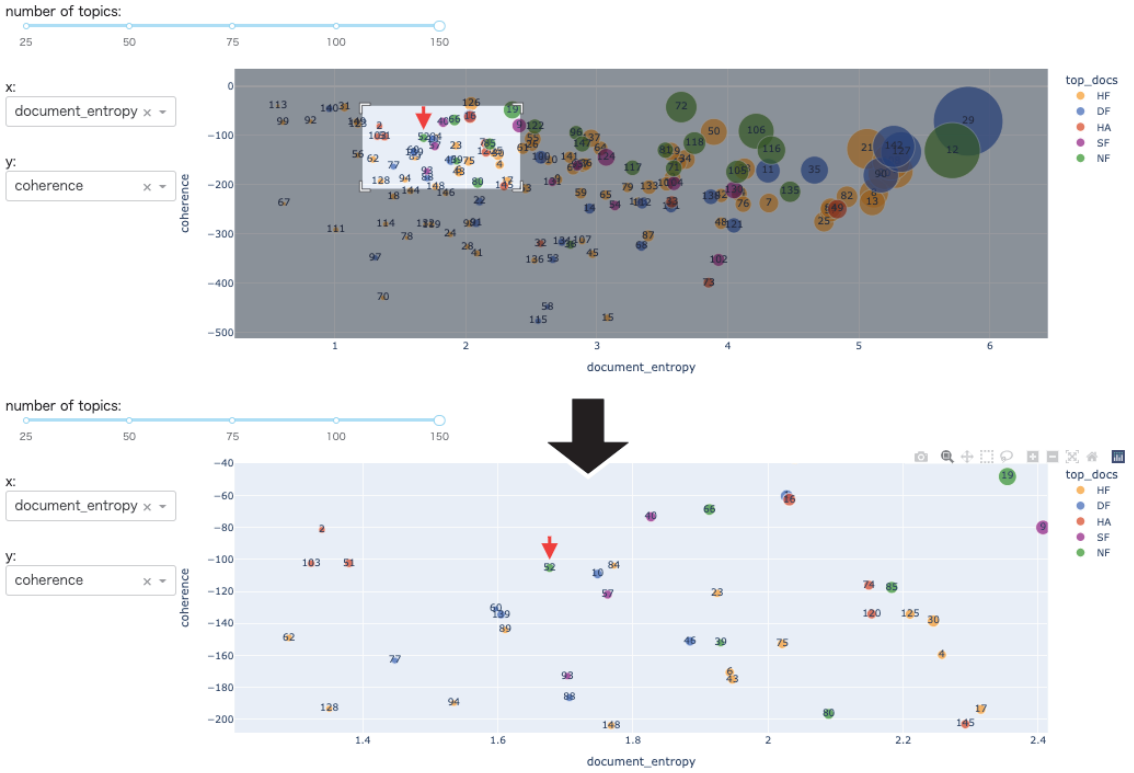

図 3: 拡大機能プレビュー

3.2. topic view

topic view では，当該トピックがどの文書にどの程度の確率で出現するかを示す散布図とボックスプロット，そして右側にトピックの構成単語とその重みが反映されたワードクラウドが描画される．この 2 つの図を読み解くことで，トピックや文書の特徴を解釈することができる．

例えば，図 1 に表示されている topic 59 は，ほとんどの文書において出現する一般的なトピックであることがわかるが，NF から始まるファイル，つまりノンフィクションでは NF_4 を除いてほとんど出現していない．そこでワードクラウドに注目すると，この topic 59 は，door, room, house, window, bed, floor など室内のものに関する語を中心に構成されていることがわかる．

NF_4，*The Case of Oscar Slater* は実際に発生した殺人事件について検証を重ねる形式の小説である．一方で，それ以外のノンフィクション小説は戦争を題材としたものがほとんどであった．この topic 59 の出現確率が高い作品は，推理小説を始め室内を中心に物語が展開するものが多く，逆に出現確率の低い作品は屋外で物語が展開する，あるいはより大局的な視点から出来事が描かれている作品が多いのではないかと考えられる．

3.3. トピック数の変更

図 4: トピック数の変更イメージ

最上段のスライダーを操作することで，図 4 のように全トピック数を切り替えることができる．

これは，スライダーの入力に応じて設定を変更したモデルを生成するのではなく，事前にトピック数値を変更しながら複数のモデルを用意しておき，スライダーにその数値を登録することでアウトプットを切り替える形式となっている．モデルの生成は負荷が大きいため，都度生成する方式では細かな数値変更に対応できる一方，切り替えに時間がかかりすぎてツール全体の処理が遅くなることが考えられる．事前に様々なパターンのアウトプットを用意しておくことで，スムーズにトピック数を切り替えながら比較し，どのモデルが自身の分析目的に適っているか考察することができる．

図 4 では 25, 50, 75, 100, 150 の 5 種類のモデルを登録しているが，25 以下，150 以上の数値でも問題なく設定できるほか，より多くのパターンを一度に登録することも可能である．

4. 結び

本稿では，LDA に基づくトピックモデルを効果的に可視化するツールの必要性について論じ，開発中のツールの概観と主要な機能を示した．静的なグラフではなくインタラクティブなグラフを出力することで，トピック数が多い場合でも快適に探索的分析を行うことが可

能となる．またスライダーを使ってトピック数を変化させることができる機能は，トピック数の決定に試行錯誤が必要不可欠である現状において，分析者に大いに寄与するのではないかと考えられる．

今後の課題として，インプットされたコーパスがサンプルコーパスよりも極めて多い文書数であった場合など，どのような種類のデータであっても topic view の表示が崩れないよう UI を改善していくことが挙げられる．

また，トピックリストから表示したいトピックのみを選択したり，特定の文書群に出現するトピックのみを選択したり，tokens など特定指標の値が一定の範囲を超えるものを除外するフィルタ機能を設けたい．分析目的に沿わないトピックを model view から取り除くオプションを追加することで，よりシンプルなプロットを描画できるようになると考えられる．

MALLET の diagnostics ファイルには，トピック間の類似性を評価する指標は組み込まれていない。一方で pyLDAvis では，多次元尺度法（MDS）に基づいてトピックを配置することで，トピック同士の類似性を距離に変換し可視化している。今後は類似性を評価する指標を組み込むことで，model view 部分のグラフ表示を切り替えたり，選択したトピックに類似したトピックをレコメンドする機能を追加することも検討している．

文　献

[1] Blei, M., Ng, A. and Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3: 993–1022.

[2] Blei, M, Lafferty, D. (2006). Dynamic topic models. *Proceedings of the 23rd International Conference on Machine Learning*, 113–120.

[3] Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D. Blei. (2009). Reading tea leaves: How humans interpret topic models. *Neural Information Processing Systems*.

[4] Jockers, M. and Mimno, D. (2013). Significant themes in 19th-century literature. *Poetics* 41: 750–769.

[5] Kuroda, A. (2018). Topic Representation across Texts and Genres: Finding Key-words through Topic Models. (Unpublished master's thesis). Osaka University.

[6] Lau, J., Newman, D. and Baldwin, T. (2014). Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 530–539.

[7] Mimno, D., Wallach, H., Talley, E., Leenders, M. and McCallum, A. (2011). Optimizing semantic coherence in topic models. *EMNLP '11: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 262–272.

[8] Newman, D., Lau, J., Grieser, K. and Baldwin, T. (2010). Automatic Evaluation of Topic Coherence. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 100–108.

[9] Rosen-Zvi, M., Griffiths, T., Steyvers, M. and Smyth, P. (2004). The author-topic model for authors and documents. *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, 487–494.

[10] Sbalchiero, S. and Eder, M. (2020). Topic modeling, long texts and the best number of topics. Some Problems and solutions. *Quality & Quantity*, 54, 1095–1108.

[11] Sievert, C. and Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, 63–70.

[12] 田畑 智司 (2017)「FLOB コーパスの意味構造: 確率論的トピックモデルによる言語使用域の特徴付け」『統計数理研究所共同研究リポート』386: 1–17.