2021 年 9 月

トピックモデルについて

KH Coder のメニューに「トピックモデル」という新しい機能が追加されました（図 1）．KH Coder では文書方向（横方向）からの第 1 段階の探索的な分析を行うためのツールの一つとして位置づけられています．KH Coder のマニュアルと文献を参考に，**本書**（「やってみようテキストマイニング［増訂版］）で扱った高齢者向けサービスのデータを使いながら，この新しい機能であるトピックモデルについて解説します．

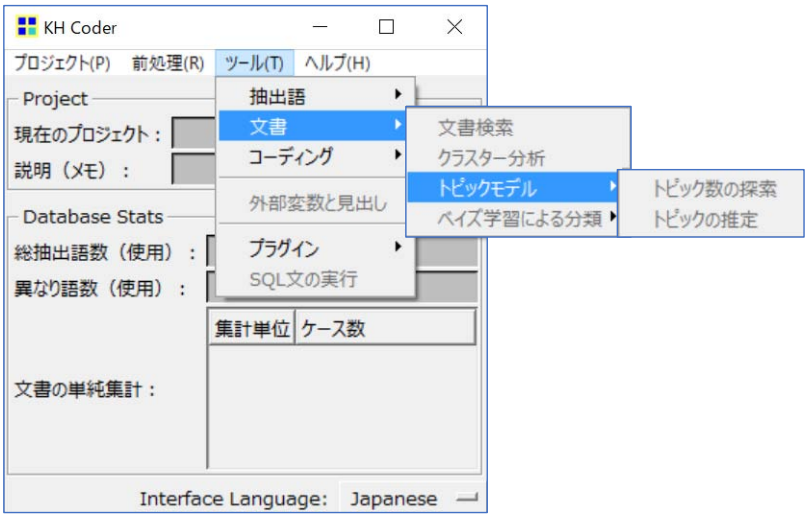


図 1．追加された新機能「トピックモデル」

1．トピックモデルと第 2 段階の分析

トピックモデルは理論的にはとても難しい手法です．文献によると，トピックモデルは次元の縮小と文書の分類をまとめて行うための確率モデルであるといえます．この意味では，本書の第 6 章で説明している分析者定義コードを利用する第 2 段階の分析と共通しています．つまりトピックモデルを「アンケート調査の自由回答」に適用することは，文書全体を要約し，分類するための理論的なアプローチである，といえます．KH Coder における 2 つの方法を比較して表 1 に示します．

2 つの方法とも次元を縮小して文書全体をトピックあるいはコードとして要約する方法ですが，トピックの方は語の出現頻度と共起性をベースに理論的に求めるのに対して，KH Coder における第 2 段階の分析の方は分析者が語を連結してコードとして定義します．トピックモデルの場合，トピックが語によってどのように構成されるのか，各文書サンプルがトピックとどのような関連性があるのかは 2 つのファイルに出力されます．これらの詳細は 2.2 節と 2.3 節で解説します．一方，第 2 段階の分析の方は，分析者がコードをコーディ

ングルール・ファイルに定義し，文書サンプルの分類結果は「文書×コード」表に出力されます（本書第 6 章参照）．2 つの方法はこのような関係にありますが，トピックモデルは，樋口耕一氏が KH Coder のマニュアルで述べている通り『トピックの推定はあくまで自動的に行われるので，分析者の観点を反映したトピックが見つかるとは限らない』という方法であり，したがって KH Coder の中では探索的な第 1 段階の手法という位置づけになっています．詳細はマニュアルを参照してください．

表 1．トピックモデルと第 2 段階の仮説検証型の分析

	トピックモデル	**第2段階の分析：仮説検証的な分析（本書第6章で解説）**
次元の縮小	**トピック**	**（分析者定義）コード**
語との関係	**「トピック×語」表**	**コーディングルール・ファイル**
文書の分類	**「文書×トピック」表**	**「文書×コード」表**

2．事例データをトピックモデルで分析

本書の事例（「高齢者向けサービスのアイディア」）に KH Coder のトピックモデルを適用してみます．なお，以降では，図 2 のように「強制抽出する語」と「使用しない語」を設定して分析を進めます（本書 3.4.2 参照）．

強制抽出する語　　使用しない語

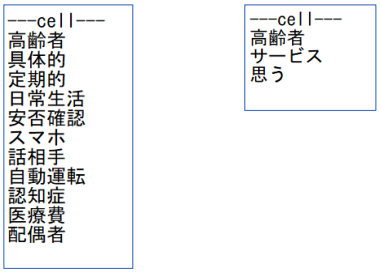


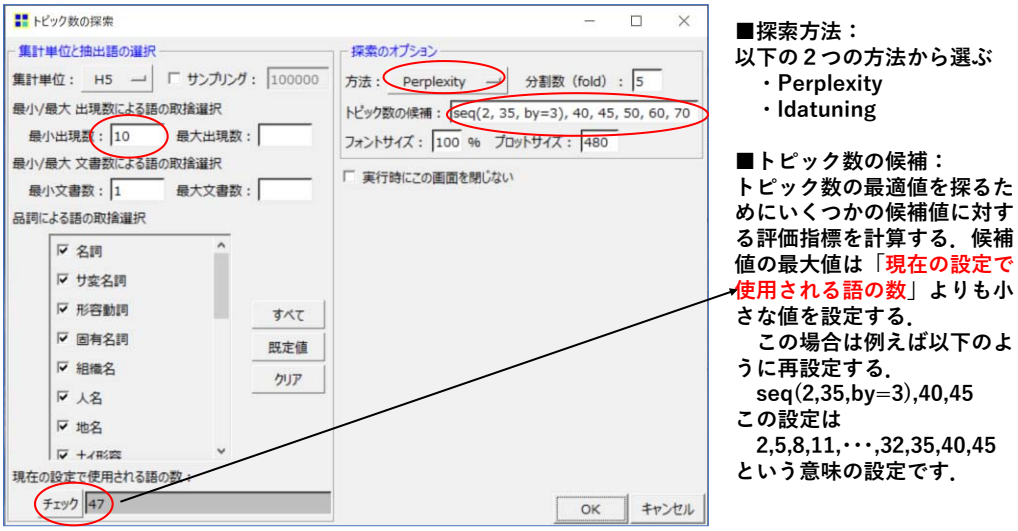
図 2．分析に使用する語の設定

2.1　トピック数の探索

トピックモデルでは分析者はあらかじめトピック数を決める必要があります．トピックモデルの 2 つのメニュー（図 1）の中の「トピック数の探索」によって，トピック数を決めるためのヒントを得ることができます．図 3 はパラメータ設定のための画面です．

トピック数を探索するための[方法]は 2 通り準備されていますが，両方とも実行してみてください．図 3 のコメントのように，[トピック数の候補]の最大値は「現在の設定で使用さ

れる語の数」（この例では 47 個）よりも小さな値を設定します．



■探索方法：
以下の 2 つの方法から選ぶ
・Perplexity
・ldatuning

■トピック数の候補：
トピック数の最適値を探るためにいくつかの候補値に対する評価指標を計算する．候補値の最大値は「現在の設定で使用される語の数」よりも小さな値を設定する．
この場合は例えば以下のように再設定する．
seq(2,35,by=3),40,45
この設定は
2,5,8,11,･･･,32,35,40,45
という意味の設定です．

図 3．トピック数の探索の設定画面

図 4 は 2 つの方法を実行した結果です．Perplexity 基準の場合は最小となる候補を，ldatuning 基準の 4 つの指標のなかの 2 つは最大の候補を，2 つは最小の候補をトピック数として選ぶように提案されています．図 4 を見るといずれの指標も横軸のトピック数の候補が 10 ぐらいまでは急激に大きくなったり小さくなったりしていますが，それ以降は変化が緩慢であり，最大値や最小値をとる候補は明確ではありません．したがって，この例の場合はトピックを 1 つに絞り込むのは難しそうです．

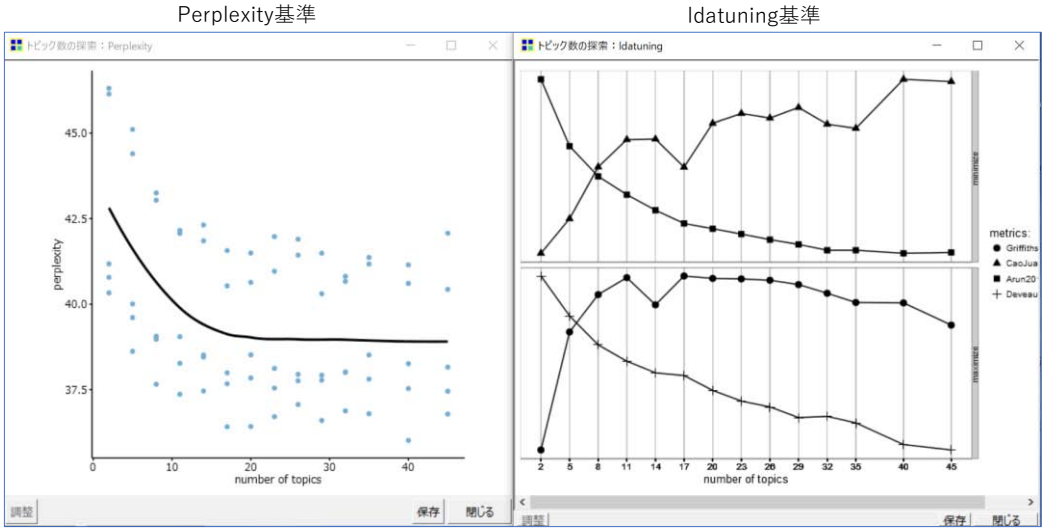
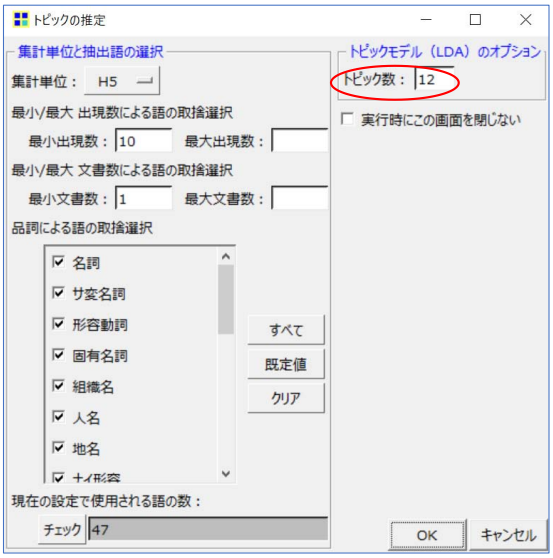


図 4．トピック数の探索結果

このような傾向は本書の事例データの場合だけに限らず，たとえば，金（文献[5]）では『トピック数を適切に推定できたという報告はあまり見られない．ツールによる推奨値はあくまでも参考程度である』と述べています．分析者はあくまでもこれを 1 つのヒントにして，試行錯誤してみましょう．

2.2　トピックの推定

図 5 は「トピックの推定」のためのパラメータ設定画面です．前述の通り分析者はトピック数を設定しなければなりませんが，何通りか変えてみて試行錯誤するのがいいでしょう．ここでは 2.1 節も参考にしながら，デフォルト値である 12 個のままで実行してみます．本書の第 5 章の文書のクラスター分析の場合には，クラスター数を 10 として分析していますが，これも 1 つの参考値にしてもいいでしょう．



■トピック数：
デフォルトは12であるが、「トピック数の探索」の結果などを参考にいろいろ変えて試行錯誤してみる

図 5．トピック推定の設定画面

図 6 は，分析に使用する語の数を 47（図中の②），トピック数を 12（図中の①）としてトピック推定を実行した結果の一部分（トピック#6 以降は隠れています）です．

トピック別の語と対応する値（確率値）は，「トピック×語」表（⑦をクリックして出力できる表の 1 つ）の一部を⑤にチェックして棒グラフ表示したものです．語は全部で 47 個ありますが，ここではトピック別に降順に 10 個まで表示した特徴語一覧表です（表示する語の数は⑥で設定可）．トピック別に語に対する値を合計すると 1 になります(2.3 節参照)．これがトピックモデルが確率モデルである所以です．これらの値を見ることによって各々のトピックがどのような特徴を持つトピックであるかが分かります．分析者はこの特徴をみてトピックを解釈することになります．たとえばトピック#1 は「施設」「支援」「生活」

「受ける」などの語によって特徴付けられるトピックということになります．トピック#2以降についても値の大きい順に出力されているので，それぞれの特徴を解釈していきます．

図６．トピックの推定結果画面

次に各文書（サンプル）がトピックによってどのように特徴付けられるのかを見てみましょう．この部分はもう１つのファイルである「文書×トピック」表と元の文書や外部変数を組み合わせて集計・分析しています．その一部の結果は図６の画面上の③，④，⑧をクリックすることで確認することできます．

図７は，図６の③をクリックした結果です．トピック#1 の特徴的な文書を KH Coder のマニュアルでトピック比率と呼んでいる値（図中の 0.128, 0.116, 0.108 などの値）の降順に示したものです．トピック#1 を特徴付けていた「施設」「支援」「生活」などの語を含む文書が並んでいます．文書を選択して[文書表示]をクリックすると詳細を確認できます．

図８は，図６の④をクリックした結果です．トピック#1 の文書別トピック比率の年代別の平均値を折れ線グラフで表した図です．微妙な違いですが「男性 55-59 歳」の年代がトピック#1 の特徴が強いようです．性年代は外部変数の１つであり，性別や就業形態などのほかの外部変数を選択してグラフ化することもできます．またグラフと一緒に 12 個のすべてのトピック比率の集計表が表示されます．この図の場合，トピック＃１のグラフのみが表示されていますが，[調整]ボタンをクリックすると，そのほかの複数のトピックを指定して折れ線表示することができます．

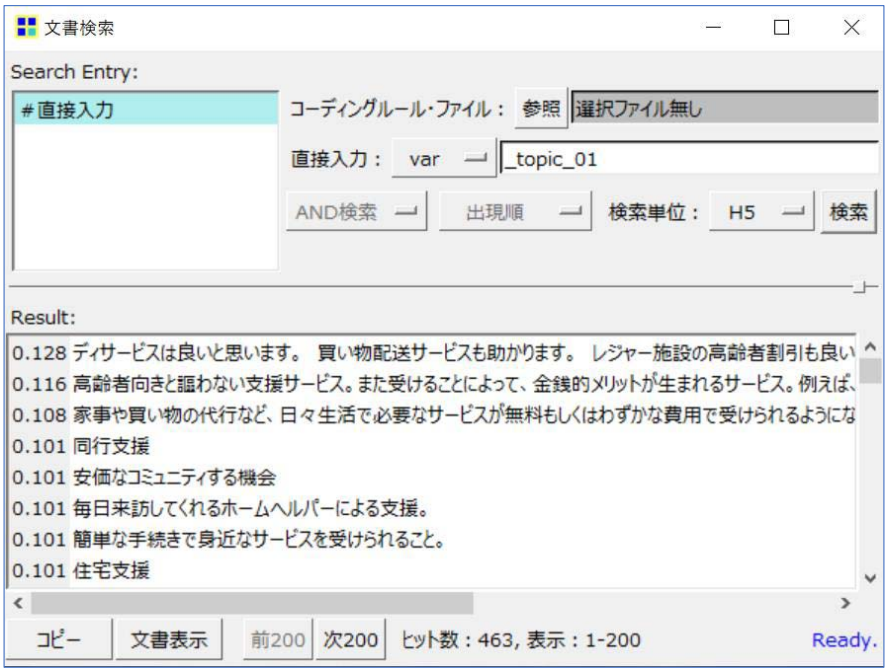


図７．トピック#1 の特徴的な文書

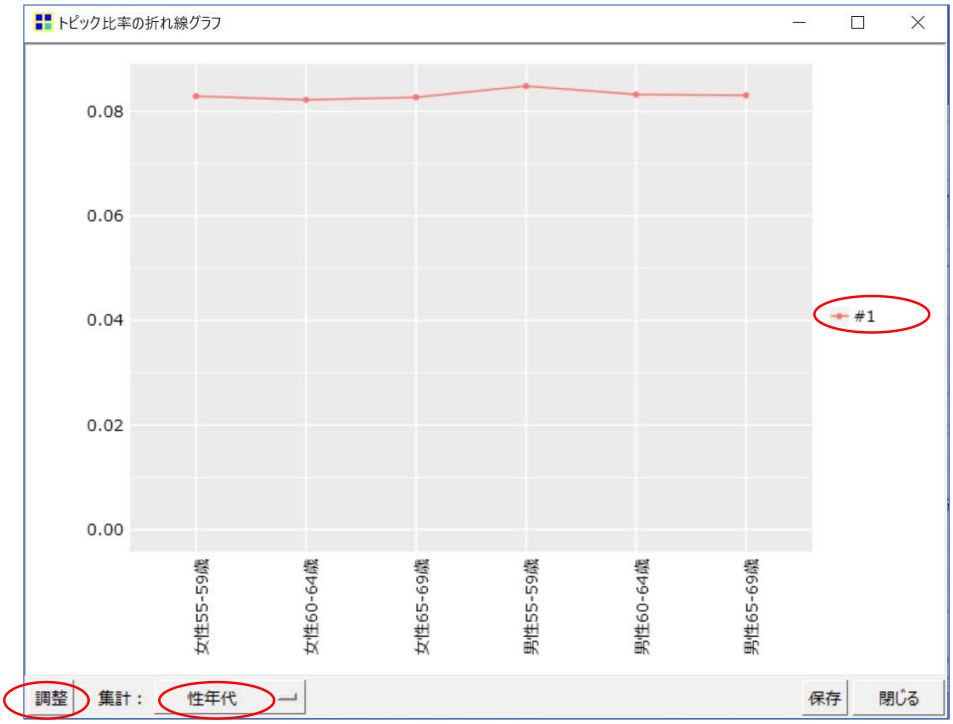


図８．トピック比率の「性年代別」折れ線グラフ

図９は，図６の⑧[マップ]をクリックしたときに表示される表とグラフです．すべてのトピックに関する性年代別のヒートマップと集計表が表示されます．性年代別にどのトピックに特徴があるのかを判断することができます．ほかの外部変数についても同様の分析を行うことができます．

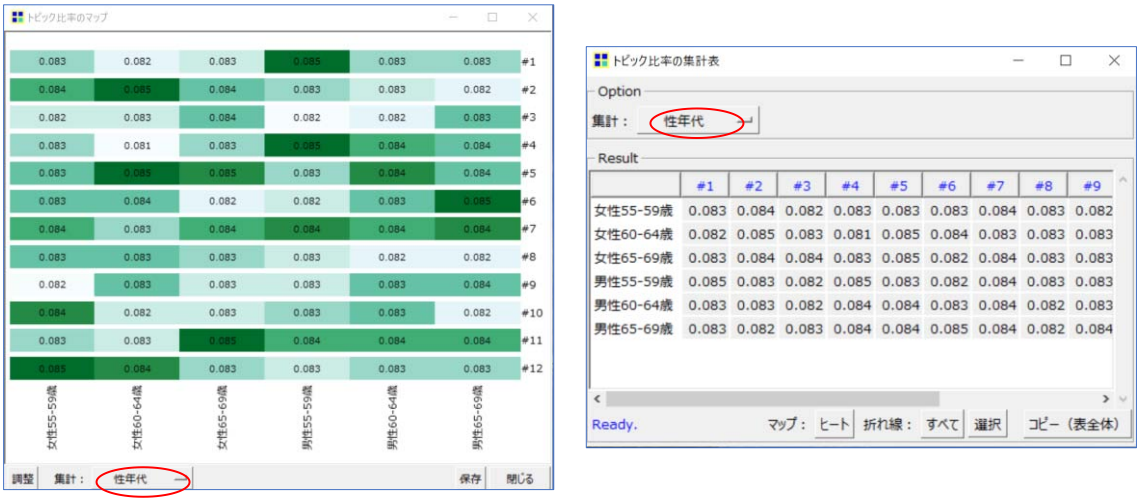


図９．トピック比率のヒートマップと集計表

2.3　トピック推定の２つの出力ファイル

トピック推定の詳細な結果が保存されている 2 つのファイルを改めて見ていくことにしましょう．2.2 節の図表はこれらをベースにして得られたものでした．

図 10 は，分析対象の 47 個のすべての語に関する「トピック×語」表に縦計，横計と語の出現頻度を追加し，**語の出現頻度の降順**に整列した図です．2.2 節の図６に合わせて，もとの出力ファイルの行と列を入れ替えて表示しています．図６はトピックごとに値を降順に並べ替えて先頭の 10 語だけを示しています（語数は任意で設定可）．トピックの特徴が判断しやすいようにするために図６と同様に，Excel の条件付き書式機能を利用して棒グラフ表示し，さらに 0.1 以上のセルを網がけしてみました．この図 10 から，すべての語とトピックの全体的な関係がよく分かります．どのトピックについても，0.1 以上の関連のある語は数個程度までしかなく，小数点以下 2 桁まで 0 の語がほとんどであることが分かります．つまり，どのトピックも数個程度までの語で特徴付けられています．また，出現頻度の高い語は複数のトピックと関連があり，一方出現頻度の低い語はどのトピックとの関連も強くないことも分かります．ひとつの語が複数のトピックと関連があるということは，トピックの解釈を難しくするという側面もあります．また，この場合は語の出現頻度と横計との相関係数を計算すると 0.99 と高い値になります．これらの特徴は，トピックは語の出現頻度と共起性をベースにした確率モデルであり，いろいろな回答パターンを反映しているこ

とを物語っています．確率モデルであることは各トピックの縦計が 1 であることから確認できます．

語	トピック#1	トピック#2	トピック#3	トピック#4	トピック#5	トピック#6	トピック#7	トピック#8	トピック#9	トピック#10	トピック#11	トピック#12	計	語の出現頻度
買い物	0.0012	0.0864	0.0014	0.0012	0.0009	0.0123	0.0010	0.0013	0.1792	0.0012	0.4309	0.0011	0.7179	65
介護	0.0012	0.0224	0.0014	0.1050	0.0102	0.0011	0.1470	0.0013	0.1665	0.0133	0.0010	0.2359	0.7062	63
施設	0.2162	0.0011	0.0014	0.1165	0.0009	0.0011	0.1276	0.0013	0.0013	0.0012	0.0010	0.0011	0.4706	41
生活	0.1685	0.1505	0.0014	0.0127	0.0009	0.0123	0.0010	0.0013	0.0521	0.0012	0.0010	0.0651	0.4678	40
代行	0.0012	0.0011	0.0014	0.0819	0.0009	0.3579	0.0010	0.0013	0.0013	0.0012	0.0010	0.0011	0.4511	39
家事	0.0012	0.0117	0.0014	0.0012	0.1216	0.0903	0.0010	0.0013	0.0013	0.0012	0.1648	0.0011	0.3980	38
人	0.0012	0.2572	0.0014	0.0012	0.0009	0.0011	0.0010	0.0140	0.0775	0.0012	0.0010	0.0438	0.4014	35
支援	0.2162	0.0011	0.0014	0.1742	0.0009	0.0011	0.0010	0.0013	0.0013	0.0012	0.0010	0.0117	0.4124	34
必要	0.0012	0.0117	0.0014	0.1280	0.0009	0.0011	0.0010	0.2046	0.0267	0.0012	0.0113	0.0011	0.3901	31
自分	0.0012	0.0011	0.0014	0.0012	0.1866	0.0011	0.0789	0.0013	0.0013	0.0012	0.0010	0.0011	0.2772	28
食事	0.0012	0.0011	0.2320	0.0012	0.0195	0.0680	0.0010	0.0013	0.0013	0.0375	0.0010	0.0011	0.3660	28
利用	0.0012	0.0117	0.1913	0.0012	0.0009	0.0011	0.0010	0.0013	0.0013	0.0012	0.1136	0.0011	0.3269	26
見守る	0.0012	0.0011	0.0014	0.0012	0.0009	0.0011	0.0107	0.0013	0.0140	0.2430	0.0420	0.0011	0.3189	26
出来る	0.0012	0.0651	0.0014	0.0012	0.0102	0.0011	0.0983	0.0140	0.0013	0.0012	0.0010	0.0758	0.2717	25
掃除	0.0012	0.0011	0.0014	0.0012	0.0009	0.0011	0.0010	0.0013	0.0013	0.0012	0.0931	0.1505	0.2552	23
システム	0.0012	0.0011	0.0014	0.0588	0.0009	0.0457	0.0107	0.0267	0.0013	0.0738	0.0215	0.0011	0.2441	20
向け	0.0012	0.0224	0.0014	0.0012	0.0474	0.0011	0.0010	0.0013	0.0013	0.0012	0.0522	0.0864	0.2180	20
今	0.0012	0.0011	0.0014	0.0012	0.1866	0.0011	0.0010	0.0013	0.0013	0.0012	0.0010	0.0011	0.1993	20
ロボット	0.0012	0.0011	0.0014	0.0012	0.0009	0.0011	0.0010	0.2427	0.0013	0.0012	0.0010	0.0011	0.2551	19
健康	0.0012	0.0011	0.0014	0.0012	0.0009	0.0234	0.1081	0.0013	0.0775	0.0012	0.0010	0.0011	0.2193	19
受ける	0.1207	0.0011	0.0014	0.0012	0.0009	0.0011	0.0497	0.0013	0.0521	0.0012	0.0010	0.0011	0.2326	19
病院	0.0131	0.0011	0.0014	0.0358	0.0009	0.0011	0.0010	0.0013	0.0013	0.1705	0.0010	0.0011	0.2295	18
交流	0.0012	0.0011	0.1642	0.0012	0.0009	0.0011	0.0010	0.0013	0.0013	0.0738	0.0010	0.0011	0.2490	18
仕事	0.0012	0.0011	0.0014	0.0012	0.1681	0.0011	0.0010	0.0013	0.0013	0.0012	0.0010	0.0011	0.1808	18
自宅	0.0012	0.1185	0.0149	0.0588	0.0009	0.0011	0.0010	0.0013	0.0013	0.0012	0.0010	0.0011	0.2023	17
無料	0.0012	0.0438	0.0014	0.0588	0.0009	0.0011	0.0010	0.0775	0.0013	0.0012	0.0010	0.0011	0.1902	15
充実	0.0012	0.0544	0.0285	0.0012	0.0009	0.0457	0.0010	0.0140	0.0394	0.0012	0.0010	0.0011	0.1895	15
相談	0.0012	0.0011	0.0014	0.0012	0.0009	0.0011	0.0302	0.0013	0.0013	0.1342	0.0010	0.0117	0.1865	15
補助	0.0012	0.0011	0.0014	0.0012	0.0009	0.1349	0.0010	0.0013	0.0013	0.0375	0.0010	0.0011	0.1837	15
気軽	0.0012	0.0011	0.0285	0.0012	0.0102	0.0011	0.1081	0.0013	0.0013	0.0012	0.0010	0.0117	0.1678	15
場所	0.0012	0.0011	0.0149	0.0012	0.0009	0.0011	0.0010	0.0013	0.0521	0.0859	0.0010	0.0224	0.1840	14
地域	0.0012	0.0331	0.0014	0.0012	0.0566	0.0123	0.0010	0.0521	0.0013	0.0012	0.0010	0.0011	0.1633	14
一緒	0.0012	0.0011	0.0014	0.0012	0.0288	0.0011	0.0010	0.0013	0.1283	0.0012	0.0113	0.0011	0.1788	14
社会	0.0012	0.0011	0.0014	0.0012	0.0102	0.1126	0.0010	0.0013	0.0013	0.0012	0.0010	0.0224	0.1557	13
外出	0.0729	0.0011	0.0014	0.0012	0.0009	0.0011	0.0010	0.0902	0.0013	0.0012	0.0010	0.0011	0.1743	13
宅配	0.0012	0.0011	0.0014	0.0012	0.0845	0.0011	0.0010	0.0013	0.0013	0.0012	0.0010	0.0438	0.1399	13
訪問	0.0012	0.0011	0.1642	0.0012	0.0009	0.0011	0.0010	0.0140	0.0013	0.0012	0.0010	0.0011	0.1892	13
使う	0.0012	0.0011	0.0014	0.0012	0.0009	0.0011	0.0010	0.1665	0.0013	0.0012	0.0010	0.0011	0.1788	13
保険	0.0490	0.0011	0.0014	0.0012	0.0009	0.0011	0.0010	0.0013	0.0013	0.0012	0.0010	0.0864	0.1468	12
旅行	0.0012	0.0224	0.0149	0.0012	0.0009	0.0123	0.0594	0.0013	0.0013	0.0012	0.0215	0.0011	0.1386	12
高い	0.0012	0.0117	0.0963	0.0012	0.0009	0.0123	0.0010	0.0394	0.0013	0.0012	0.0010	0.0011	0.1686	12
バス	0.0012	0.0011	0.0014	0.0242	0.0009	0.0011	0.0204	0.0013	0.0013	0.0859	0.0010	0.0011	0.1408	11
話し相手	0.0012	0.0011	0.0014	0.0012	0.0009	0.0011	0.1081	0.0013	0.0013	0.0012	0.0010	0.0011	0.1207	11
移動	0.0012	0.0011	0.0014	0.1050	0.0102	0.0011	0.0010	0.0013	0.0140	0.0012	0.0010	0.0011	0.1394	11
良い	0.0490	0.0011	0.0014	0.0012	0.0009	0.0123	0.0010	0.0013	0.0775	0.0012	0.0010	0.0011	0.1488	11
家	0.0012	0.0011	0.0014	0.0012	0.0009	0.0123	0.0107	0.0013	0.0013	0.0012	0.0010	0.0971	0.1306	11
コミュニテ	0.0490	0.0438	0.0014	0.0012	0.0195	0.0011	0.0010	0.0013	0.0013	0.0012	0.0010	0.0011	0.1227	10
計	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000		

（注）Excel の条件付き書式で棒グラフを，さらに 0.1 以上のセルを網がけしている．

図 10.「トピック×語」表（元の表の行と列を入れ替え）＋合計など

図 11 はもう 1 つの出力ファイルである「文書×トピック」表に外部変数と分析対象のテキスト部（「高齢者向けサービス」の自由回答）を加えて，最初の 20 サンプル分を示したものです．各サンプルのトピック比率について基準値（後述）よりも大きな値のセルを網がけしています．2.2 節の図 6 はトピック＃1 の比率の降順に整列した表です．また，図 8 と図 9 は外部変数「性年代」とクロスして平均値を求めグラフ化した結果です．図 11 の表に対して Excel のピボットテーブルや各種のグラフ機能などを適用すればさらにいろいろな集計や分析が可能になります．

さて，Excel の T 列にサンプルごとのトピック比率の合計を計算しています．分析対象の 47 語を含まない文書サンプルの場合は値が 0 であり，トピック比率の部分はすべて空欄になっています．トピック推定の対象にならなかったサンプルであることを示しています．そのほかのサンプルの合計値はすべて 1 になっています．図 10 のトピック別の語に対応する値の合計が 1 になったのと同様に，この点もトピックモデルが確率モデルであることの特徴を示しています．

図 11 と本書の第 6 章の図 6.21 と比較してみてください．図 6.21 の場合は分析者の定義したコード別に 0 か 1 の値をとっています．つまり文書と分析者定義コードが関連するか否かが 0 と 1 で区別されています．サンプル別に多重の分類（クラスタリングあるいはカテゴライズと言ってもいいでしょう．本書参照）が行われています．それに対して図 11 の場合は，サンプルとトピックの関連性の強さが 0 と 1 の間の実数で示されます．強引にひとつのクラスターに分類するのではなく，**クラスターへ属する度合い**を求めるファジィクラスタリングという統計手法はちょうどこのような方法です（文献[1]など）．トピック比率をみて，どのトピックとの関連性が強いのか否かを判断するのは難しいので，ここでは

トピック比率の計(1) ÷ トピック数(12) = 0.0833

を目安にして網がけをしてみました．分析対象の語を多く含む文書の場合（多くは文書が長い）には複数のトピックとの関連性が高くなっています．それだけ多くのこと，いろいろなテーマについて語っている文書ということになります．

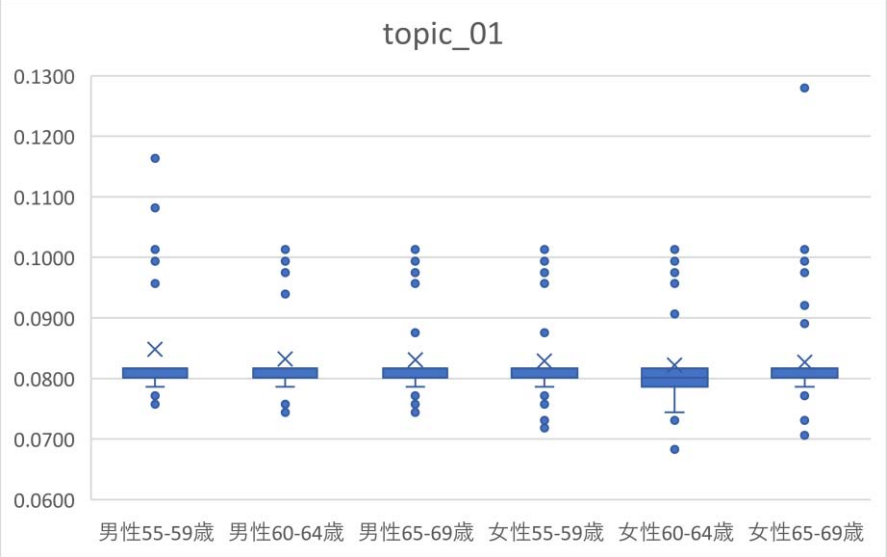
図 8 と図 9 は外部変数とクロスしたトピック比率の平均値を比較していますが，実際にはどのように分布しているのかという点も興味のあるところです．図 12 はトピック#1 の性年代別のトピック比率の箱ひげ図です（Excel の機能を利用しました）．図 11 で網がけの基準値とした 0.0833 よりも大きな値をとる特徴的なサンプルはどの性年代でも数件ずつしかないようです．性年代別の平均値に微妙な差しかないことが箱ひげ図の分布状況からも分かります．

以上のようにトピック推定の詳細なデータをいろいろ加工したり編集したり図表化することでトピックの特徴をさらに明らかにできるでしょう．

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
_topic_docid	性年代	性別	就業形態	子供の有無	家族構成	高齢者向けサービス	_topic_01	_topic_02	_topic_03	_topic_04	_topic_05	_topic_06	_topic_07	_topic_08	_topic_09	_topic_10	_topic_11	_topic_12	合計
1	男性55-59歳	男	3	2	1	家事のお手伝いさん。	0.0817	0.0817	0.0817	0.0817	0.0817	0.0817	0.0817	0.0817	0.0817	0.0817	0.1013	0.0817	1
2	男性55-59歳	男	3	1	4	ネットお見合い													0
3	男性55-59歳	男	3	1	4	対話重視のサービス。													0
4	男性55-59歳	男	3	2	3	話し相手	0.0817	0.0817	0.0817	0.0817	0.0817	0.0817	0.1013	0.0817	0.0817	0.0817	0.0817	0.0817	1
5	男性55-59歳	男	3	1	4	健康増進プログラム	0.0817	0.0817	0.0817	0.0817	0.0817	0.0817	0.1013	0.0817	0.0817	0.0817	0.0817	0.0817	1
6	男性55-59歳	男	1	2	2	社会的な老々介護の仕組み。 年金受給年齢になったら、一定の介護労働を義務化する。 老々介護に参加しない高齢者には年金を減額する。	0.0772	0.0772	0.0772	0.0772	0.0772	0.0772	0.0957	0.0772	0.1142	0.0772	0.0772	0.0957	1
7	男性55-59歳	男	3	2	1	時間はあると思うので社会貢献できるサークルがあり、地域貢献できる環境があれば良いと思います。	0.0801	0.0801	0.0801	0.0801	0.0801	0.1186	0.0801	0.0801	0.0801	0.0801	0.0801	0.0801	1
8	男性55-59歳	男	1	2	3	具体的なアイディアは浮かばないが、基本的には、機械やテクノロジーに頼ることなく、心の触れ合いを重視したサービスがあれば良いと思う。													0
9	男性55-59歳	男	1	1	4	混雑時の高齢者専用電車													0
10	男性55-59歳	男	1	2	2	宗教を介さずに死ぬことに対する恐怖心を軽減してくれるサービス													0
11	男性55-59歳	男	2	1	5	高齢者向けの医療サービスやレクレーション等が充実した施設	0.0786	0.0786	0.0786	0.0786	0.0786	0.0786	0.0975	0.0975	0.0786	0.0786	0.0975	0.0786	1
12	男性55-59歳	男	3	2	2	日常生活の支援	0.0817	0.0817	0.0817	0.1013	0.0817	0.0817	0.0817	0.0817	0.0817	0.0817	0.0817	0.0817	1
13	男性55-59歳	男	3	2	1	年金生活者への公的な機関の生活資金の貸しだし	0.0801	0.1186	0.0801	0.0801	0.0801	0.0801	0.0801	0.0801	0.0801	0.0801	0.0801	0.0801	1
14	男性55-59歳	男	5	1	3	終活													0
15	男性55-59歳	男	3	1	3	若い人との交流	0.0801	0.0994	0.0801	0.0801	0.0801	0.0801	0.0801	0.0801	0.0801	0.0994	0.0801	0.0801	1
16	男性55-59歳	男	3	1	4	同行支援	0.1013	0.0817	0.0817	0.0817	0.0817	0.0817	0.0817	0.0817	0.0817	0.0817	0.0817	0.0817	1
17	男性55-59歳	男	3	1	4	家事や買い物の代行など、日々生活で必要なサービスが無料もしくはわずかな費用で受けられるようになること。	0.1082	0.0731	0.0731	0.1082	0.0906	0.0731	0.0731	0.0731	0.1082	0.0731	0.0731	0.0731	1
18	男性55-59歳	男	3	2	1	買い物の足	0.0817	0.0817	0.0817	0.0817	0.0817	0.0817	0.0817	0.0817	0.0817	0.0817	0.1013	0.0817	1
19	男性55-59歳	男	4	1	5	定期的に様子を見に来てくれるようなこと													0
20	男性55-59歳	男	3	1	5	高齢者をただ見守って介護するだけでなく、リハビリ・ストレッチ等、手先だけでなく、脳の活性化や神経系の使い方を 補強できるサービスがほしい。	0.0801	0.0801	0.0801	0.0801	0.0801	0.0801	0.0994	0.0801	0.0801	0.0994	0.0801	0.0801	1

(注) トピック比率が合計(１)/トピック数(12)=0.0833 より大きなところに網がけしている

図 11.「文書×トピック」表＋外部変数とテキスト部



（注）×印は平均値を表す

図 12. トピック#1 のトピック比率の「性年代」別分布（箱ひげ図）

本書の第 6 章では，文書サンプル方向のクラスター分析（第 5 章）を第 2 段階の分析に生かす，というテキストマイニングの進め方を解説していますが，今回新たに追加されたトピックモデルも同様の利用の仕方をすればよいのではないかと考えます．その理由は，樋口耕一氏のマニュアルの一部を再掲すると『トピック推定はあくまで自動的に行われるので，分析者の観点を反映したトピックが見つかるとは限らない』という点に尽きます．

なお，図 10 と図 11 の Excel の表を同じフォルダ内に用意してありますので、ピボットテーブルやグラフ機能などを利用して読者自らいろいろな分析を試みてください．

参考文献

樋口耕一氏のマニュアルのほかに以下の文献を参照しました．

[1] 新納 浩幸「R で学ぶクラスタ解析」（2007 年 オーム社）

[2] 佐藤 一誠「トピックモデルによる統計的潜在意味解析」（2015 年 コロナ社）

[3] 岩田 具治「トピックモデル」（2015 年 講談社）

[4] 石田 基広「R によるテキストマイニング入門（第 2 版）」（2017 年 森北出版）

[5] 金 明哲 「テキストアナリティクスの基礎と実践」（2021 年 岩波書店）