

テキスト分析の応用の動向と分析の実施例

v1.1

機械学習を活用した内容による分類などを題材に、最近のテキスト分析手法を、実際にツールを使いながら説明します。

舘野

Tateno.masakazu@gmail.com

自己紹介

館野 昌一(たての まさかず)

1980年、慶應義塾大学大学院工学研究科管理工学専攻修士課程修了。富士ゼロックス株式会社入社後、おもにシステム製品の計画に従事。

Smalltalk-80、Interlisp-Dなどのソフトウェアと、そのためのワークステーション(1100SIP)の日本市場導入を行った。1987年1月～1990年6月、米国ゼロックス(パロアルトリサーチセンター)で計算言語学(日本語処理)の研究を行った後に帰国。日本語処理の研究を続ける。

2003年～2008年、慶應義塾大学(SFC)准教授。この間、慶應義塾大学・深谷昌弘教授との共同研究によりテキスト意味空間分析法の確立へ向けて研究とソフト開発を進める。

2015年4月、富士ゼロックス株式会社退職。

興味のある分野

自然言語処理、統計処理、機械学習、オープンソース、データサイエンティスト育成、など

使用プログラミング言語

Lisp、Prolog、Python、R、(Jupyter、Knime)、など



著作物

基礎からのSmalltalk-80：オブジェクト指向のプログラミング

<http://ci.nii.ac.jp/ncid/BN01248914>

参考URL

J-GLOBAL

<http://jglobal.jst.go.jp/public/200901021955393860>

テキスト解析×数学

http://news.mynavi.jp/column/sugaku_recipe/016/

PARCについて

<http://textmagic.dip.jp/PARC/>

フェイスブック

<https://www.facebook.com/masakazu.tateno.1>

インスタグラム

<https://www.instagram.com/masakazutateno/>

木星ペンギンの会(作成中)

(HRデータサイエンティスト育成研究会)

<https://hrds.jimdo.com/>

目次

1. テキスト分析方法

- 内容の集約

1. 観測変数(語の頻度)による統計的方法
2. 潜在変数(語の共起)による統計的方法
3. 機械学習による方法

- 仮想的な対話

4. Chatbot

2. ツール類(無償のもの)

1. Jupyter
2. Knime

3. 関連情報

1. HRデータサイエンティスト育成研究会

4. 分析事例

1. Jupyterを使った分析事例
 1. 潜在クラス分析

その前に

観測データの尺度(種類)

種類	説明	例
比例尺度	0が原点であり、間隔と比率に意味があるもの	
間隔尺度	目盛が等間隔になっているもので、その間隔に意味があるもの	
順序尺度	順序や大小には意味があるが間隔には意味がないもの	
名義尺度	他と区別し分類するための名称のようなもの	

データ分析手法

数量化理論（すうりょうかりろん、Hayashi's quantification methods）は、統計数理研究所元所長の林知己夫によって1940年代後半から50年代にかけて開発された日本独自の多次元データ分析法である。数量化理論にはI類、II類、III類、IV類、V類、VI類までの6つの方法があるが、現在、I類からIV類までがよく知られている。この何類という名称は、1964年に社会心理学者の飽戸弘によって命名されたもので、以後その名称が定着した。

(<https://ja.wikipedia.org/wiki/数量化理論>)

数量化1類は、目的変数が**数量データ**、説明変数が**名義尺度**のデータです。**重回帰分析**で適用できるデータは、目的変数、説明変数どちらも数量データです。**ロジスティック回帰分析**で適用できるデータは、目的変数は2群のカテゴリーデータ、説明変数は**数量データ**です。

数量化2類は、目的変数が**名義尺度**のデータ、説明変数が**名義尺度**のデータです。**判別分析**は、目的変数が**名義尺度**のデータ(群データ)、説明変数が**数量データ**です。

数量化3類は、目的変数がなく、説明変数が**名義尺度**のデータです。**コレスポンデンス分析**はクロス集計結果を散布図で表現する解析手法です。

クラスター分析は類似している質問項目や回答者をグルーピングする解析手法です。

目的変数がなく、説明変数が**数量データ**(比例尺度、間隔尺度、順序尺度)の場合の手法は、主成分分析と因子分析があります。どちらの手法も、数多くの変数から新しい概念の変数を作ります。新しく作られた概念の変数を**潜在変数**といいます。これに対し、元の変数を**観測変数**といいます。

主成分分析は、分析を通し新しく見出す潜在変数に、個々の変数では表現されない総合点があります。

因子分析は、総合点が存在しません。潜在変数一つ一つが一つの概念を表現します。

共分散構造分析は、アンケート調査の回答データ、テスト得点、実験データなどの観測データにおいて、分析者が項目間（変数間）の因果関係について仮説を立て、これが正しいかどうかを検証する解析手法です。

潜在クラス分析とは、**名義尺度**のデータ(観測変数)の背後に**名義尺度**のデータ(潜在変数)があることを仮定して潜在構造を読み解くことを言います。

https://istat.co.jp/ta_commentary/

https://www.jstage.jst.go.jp/article/ojjams/24/2/24_2_345/pdf

着目点：観測変数から潜在変数を導く：因子負荷量

因子分析において、得られた共通因子が分析に用いた変数（観測変数）に与える影響の強さを表す値で、観測変数と因子得点との相関係数に相当する。-1以上1以下の値をとり、因子負荷量の絶対値が大きいほど、その共通因子と観測変数の間に（正または負の）強い相関があることを示し、観測変数をよく説明する因子であると言える。

因子負荷量 | 統計用語集 | 統計WEB

<https://bellcurve.jp/statistics/glossary/660.html>

オリジナル項目分析（因子分析：項目の削除）

因子数を5として、因子負荷量の小さい（基準値 .35）V17とV26を外し、複数の因子に高い寄与を示しているV23も外した17項目を対象として主因子法、バリマックス回転をしてみます。

回転後の因子行列^a

	因子				
	1	2	3	4	5
V29体調管理に気を付けている	.864	-.013	.074	.028	-.046
V30バランスのとれた食事を心掛けている	.757	.060	.116	.014	.014
V283食きちんと食べるようにしている	.667	.067	.092	.025	-.105
V27充分睡眠をとるようにしている	.349	-.045	-.063	-.158	.114
V11初対面の人と話すのが得意だ	-.028	.834	.045	-.097	.080
V12人の輪の中にすぐ溶け込める	.065	.792	.057	-.036	.243
V18大切な人がいる	-.012	.075	.736	.009	-.012
V19居場所がある	.102	.007	.734	-.182	.001
V20直接会って悩みを相談できる人がいる	.053	-.026	.438	-.010	.298
V16毎日が充実している	.131	.290	.410	-.060	.239
V13周りの目が気になる	-.100	.068	-.115	.765	-.159
V15人に嫌われるのが怖い	-.008	-.058	-.081	.731	-.031
V22心配性だ	.020	-.126	.074	.395	-.321
V14親しい相手の前でも緊張してしまう	.008	-.286	-.085	.389	.131
V24のう天気だ	-.181	.003	.125	-.068	.606
V21自分を過大評価する	.059	.099	.007	.004	.504
V25何事もいい結果に結びつく気がする	.082	.242	.206	-.182	.465

因子抽出法: 主因子法

回転法: Kaiserの正規化を伴うバリマックス法

a. 5回の反復で回転が収束しました。

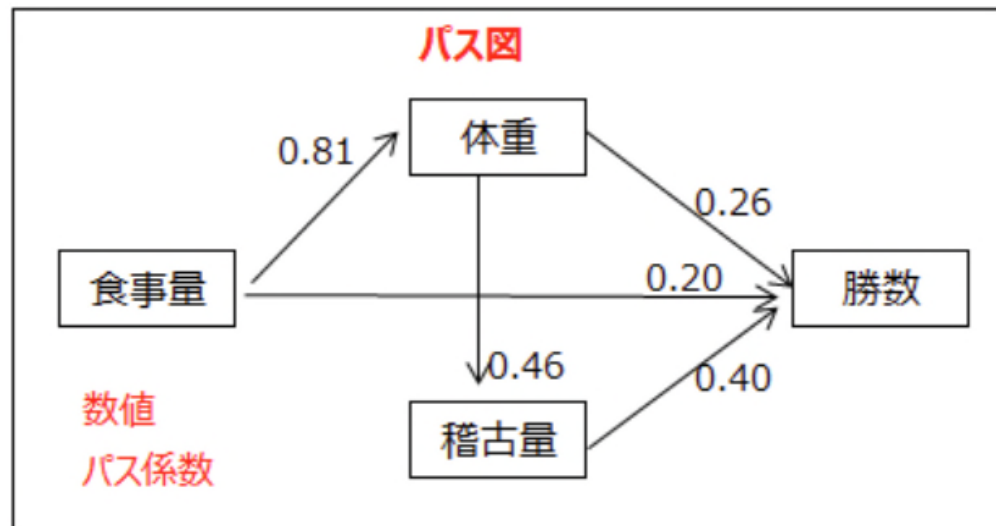
着目点：観測変数から潜在変数を導く：パス係数

共分散構造分析は、アンケート調査の回答データ、テスト得点、実験データなどの観測データにおいて、分析者が項目間（変数間）の因果関係について仮説を立て、これが正しいかどうかを検証する解析手法です。

共分散構造分析から次のことが把握できます。

- ・ 項目間の相関関係、因果関係を解明します。
- ・ 潜在変数を導入することによって、潜在変数と項目との間の因果関係を解明します。
- ・ 潜在変数から、類似した傾向を示す項目をまとめることができます。
- ・ 潜在変数の間で因果関係を検討すれば、多くの項目の間の関係を直接扱うより効率よく扱えます。

因果関係の仮説は項目間を矢印で結んだ**パス図**と呼ばれる図で表します。共分散構造分析を行うことにより、項目間の関係の強さを表す**パス係数**と呼ばれる値が求められ、パス図の矢印線上に記載されます。パス係数の大小によって因果関係を解明します。



1-1 観測変数(語の頻度)による統計的方法

1. 語の頻度
2. TF-IDF
3. 相互情報量
4. ...

<https://www.slideshare.net/aliabasi/an-introduction-to-data-mining>

<http://brandonrose.org/clustering>

	Document 1	Document 2	Document 3	Document 4	Document 5	Document 6	Document 7	Document 8
Term(s) 1	10	0	1	0	0	0	0	2
Term(s) 2	0	2	0	0	0	18	0	2
Term(s) 3	0	0	0	0	0	0	0	2
Term(s) 4	6	0	0	4	6	0	0	0
Term(s) 5	0	0	0	0	0	0	0	2
Term(s) 6	0	0	1	0	0	1	0	0
Term(s) 7	0	1	8	0	0	0	0	0
Term(s) 8	0	0	0	0	0	3	0	0

Document Vector

Word Vector
(Passage Vector)

TF-IDF: AN EXAMPLE

Consider words “apple” and “the” that appear 10 and 20 times in document 1 (d1), which contains 100 words.

Consider $|D| = 20$ and word “apple” only appearing in d1 and word “the” appearing in all 20 documents

$$tf-idf(t, d, D) = tf(t, d) * idf(t, D)$$

$$tf(t, d) = \frac{f(t, d)}{\max\{f(w, d) : w \in d\}}$$

$$idf(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$

The total number of documents in the corpus

The number of documents where the term t appears

$$tf-idf("apple", 1) = \frac{10}{100} \times \log \frac{20}{1} = 0.13$$

$$tf-idf("the", 1) = \frac{20}{100} \times \log \frac{20}{20} = 0.$$

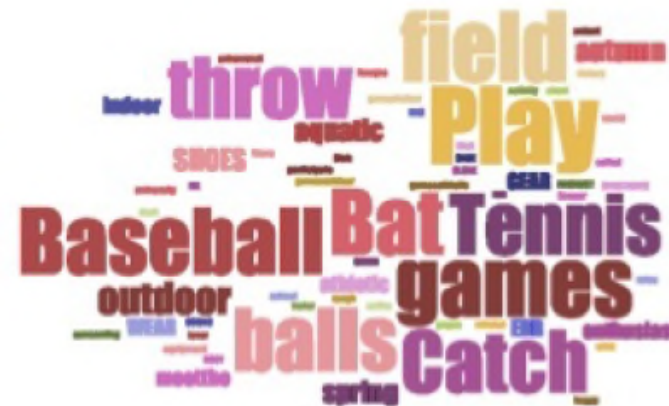
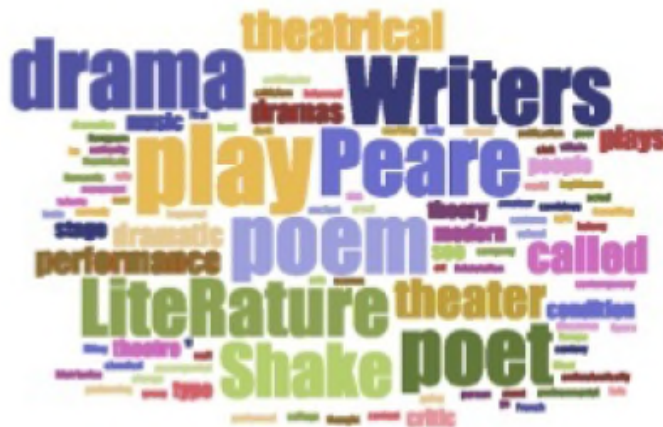
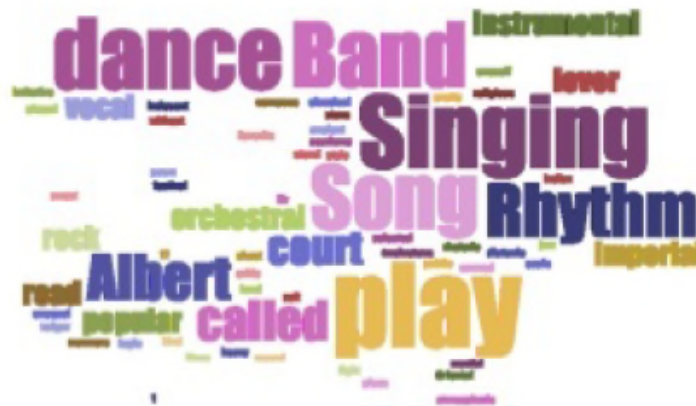
参照:

Data Mining: an Introduction

P78~

1-2 潜在変数(語の共起)による統計的方法

- 1-2-1 潜在意味分析(LSA)
 - <https://www.slideshare.net/ksmzn/topicmodel>



1-2-1潜在意味分析(LSA)

潜在セマンティック分析 (LSA) は、[自然言語処](#)

[理](#)、特に[分布セマンティクスの](#)技術であり、文書

と用語に関連する一連の概念を生成すること

によって、文書セットとその用語との関係进行分析

します。LSAは、意味の近い単語が類似したテキス

ト ([分散仮説](#)) で発生すると仮定します。段落ご

との単語数を含む行列 (行は一語の単語と列は各

段落を表します) は、大量のテキストと[特異値分](#)

[解](#)と呼ばれる数学的手法から構成されます

(SVD) は、列間の類似構造を維持しながら行数

を減らすために使用されます。その後、任意の2つ

の行によって形成される2つのベクトル (または2

つのベクトルの[正規化](#)間の[ドット積](#)) の間の角度

のコサインを取ることによって、単語が比較され

ます。1に近い値は非常に類似した単語を表し、0

に近い値は非常に異なる単語を表します。^[1]

[Scott Deerwester](#)、[Susan Dumais](#)、[George](#)

[Furnas](#)、[Richard Harshman](#)、[Thomas Landauer](#)、

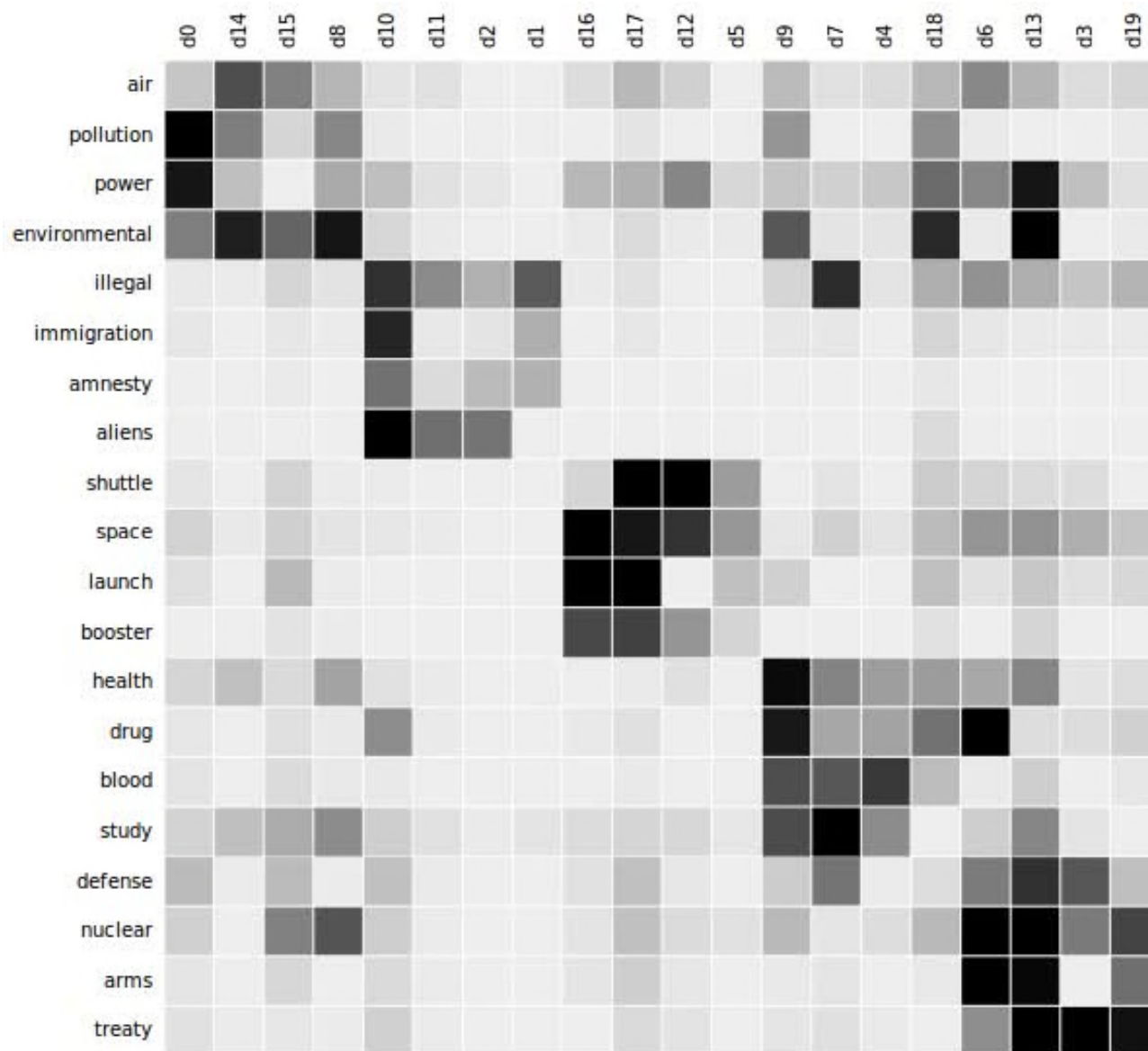
[Karen Lochbaum](#)、[Lynn Streeter](#)が1988年に潜在

的意味構造を用いた情報検索技術の特許取得した

([米国特許第4,839,853号](#)、期限切れ)。[情報検索](#)

への応用の文脈では、**潜在意味索引 (LSI)** と呼ば

れることがあります。^[2]



1-2-1潜在意味分析(LSA)

潜在的意味

- ▶ 「音楽」や「スポーツ」という単語が無かったとしても、単語群を見て想起できる
- ▶ 複数の単語の共起性によって創発される情報

トピック

- ▶ 潜在的意味のカテゴリを**トピック**と呼ぶ
- 「単語の共起性をいかに数学的にモデル化するか？」

特異値分解

- ▶ 単語文書行列 X を 3 つの行列に分解

$$X = USV^T$$

- ▶ U, S, V の各列ベクトルを特異値が大きい順に K 個用いて、 $\tilde{U}, \tilde{S}, \tilde{V}$ を作り、ランク K の**低ランク近似行列** \tilde{X} を得る

$$\tilde{X} = \tilde{U} \tilde{S} \tilde{V}^T$$

特異値分解による潜在意味解析

文書に含まれている単語を抽出し、それらの頻度から単語文書行列 X を作成する

	drive	automobile	car	play	music
文書1	2	3	0	0	0
文書2	2	0	2	0	0
文書3	0	0	0	2	2
文書4	0	0	0	3	1

- ▶ 「car」で検索しても、文書1は発見できない
- ▶ 「automobile」でも、文書2は発見できない

→単語の持つ潜在的な意味を考える

→特異値分解

特異値分解の結果

	drive	automobile	car	play	music
文書1	2.38	2.29	0.85	0	0
文書2	1.32	1.27	0.47	0	0
文書3	0	0	0	2.36	1.37
文書4	0	0	0	2.67	1.55

文書1・2ともに、「car」「automobile」の頻度が0でない！

→「drive」との共起性から、潜在的な意味が抽出されている

[参照](#)

1-2-2 潜在クラス分析(LCA)

- ESRA2015 course: Latent Class Analysis for Survey Research (<https://www.slideshare.net/DanielOberski/esra2015-course-latent-class-analysis-for-survey-research>)

Small example: data from GSS 1987

Y1: "allow anti-religionists to speak"

Y2: "allow anti-religionists to teach"

Y3: "remove anti-religious books from the library"

(1 = allowed, 2 = not allowed),

(1 = allowed, 2 = not allowed),

(1 = do not remove, 2 = remove).

	Y1	Y2	Y3	Observed frequency (n)	Observed proportion (n/N)
	1	1	1	696	0.406
	1	1	2	68	0.040
	1	2	1	275	0.161
	1	2	2	130	0.076
	2	1	1	34	0.020
	2	1	2	19	0.011
	2	2	1	125	0.073
	2	2	2	366	0.214

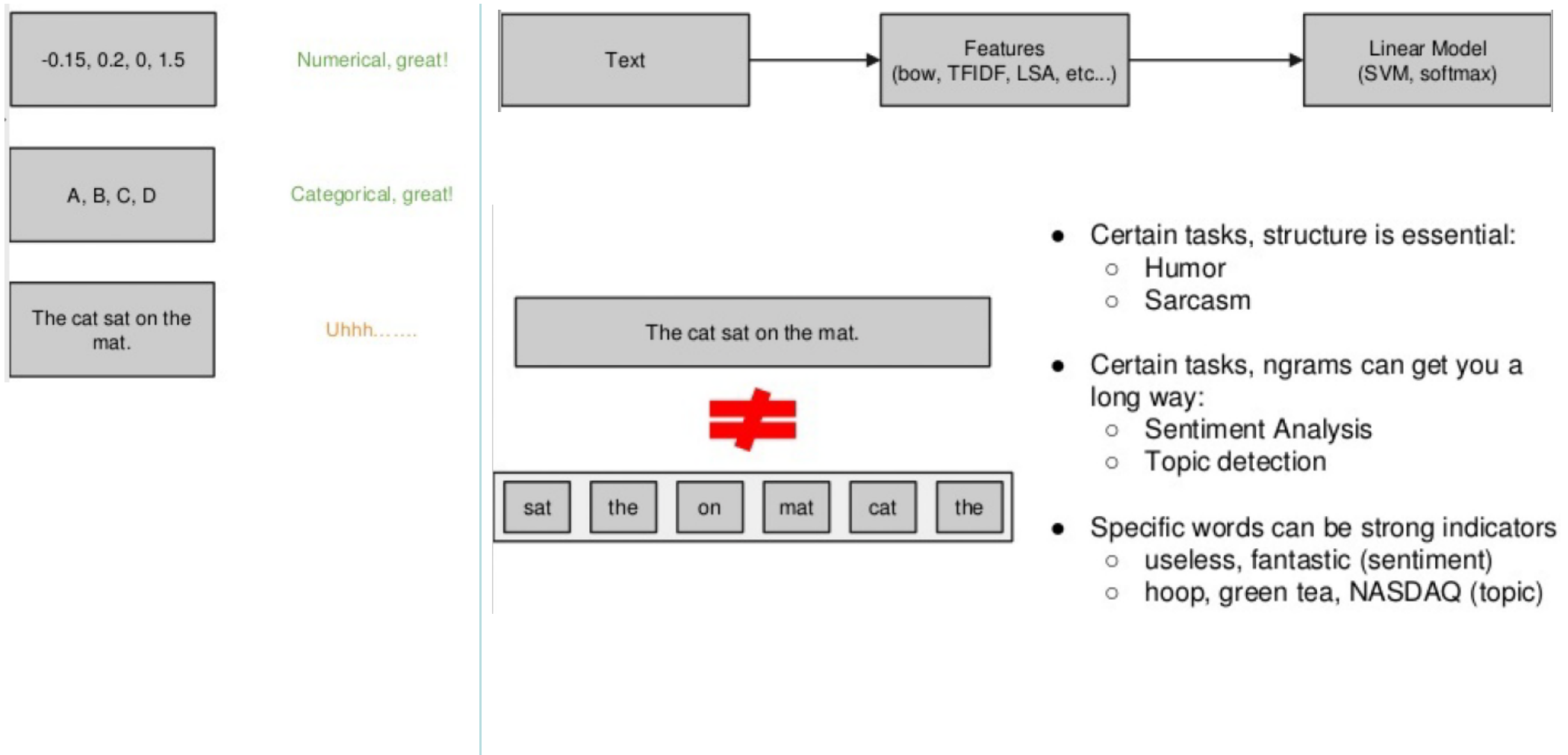


	Y1	Y2	Y3	P(X=1 Y)	P(X=2 Y)	Most likely (but not sure!)
	1	1	1	0.002	0.998	2
	1	1	2	0.071	0.929	2
	1	2	1	0.124	0.876	2
	1	2	2	0.832	0.169	1
	2	1	1	0.152	0.848	2
	2	1	2	0.862	0.138	1
	2	2	1	0.920	0.080	1
	2	2	2	0.998	0.003	1

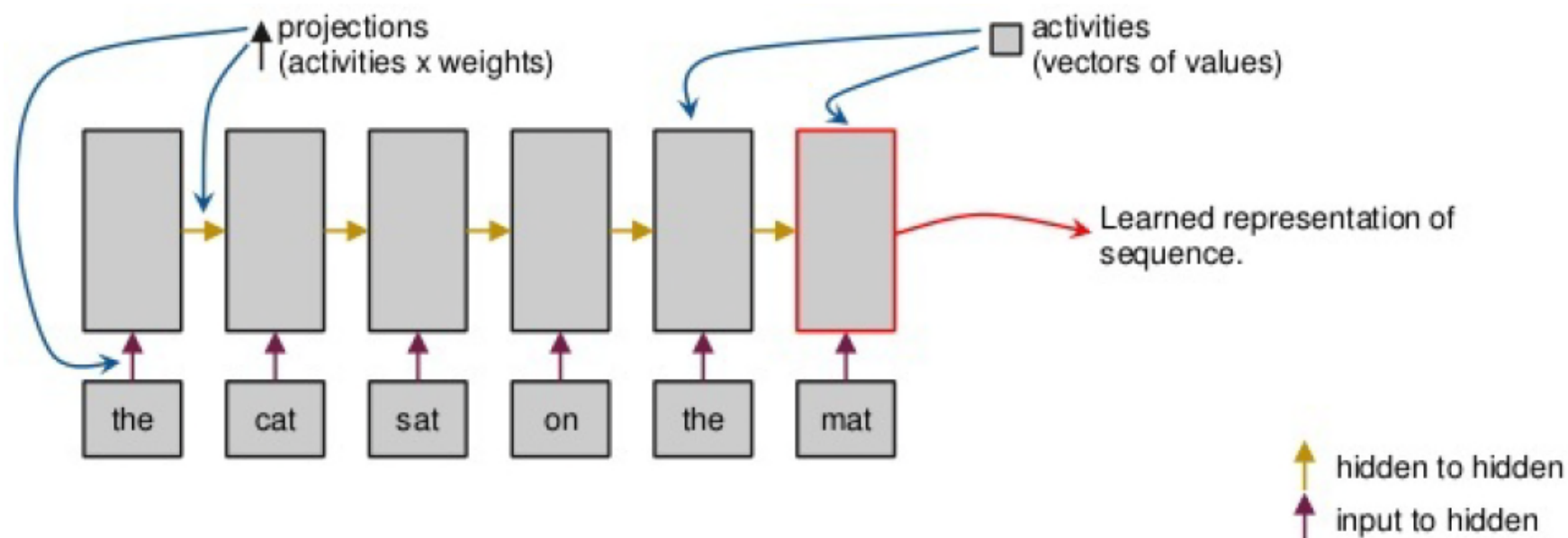
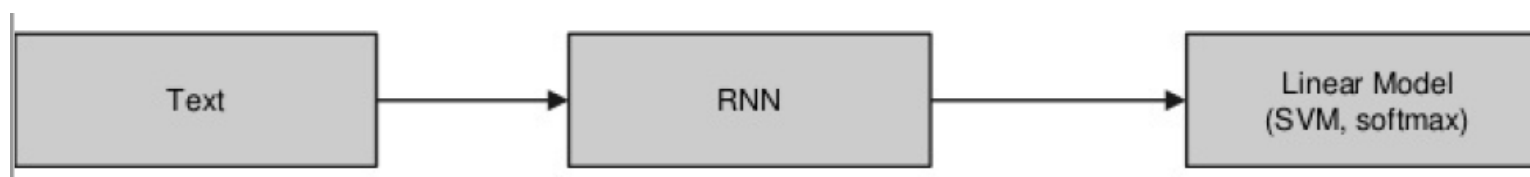
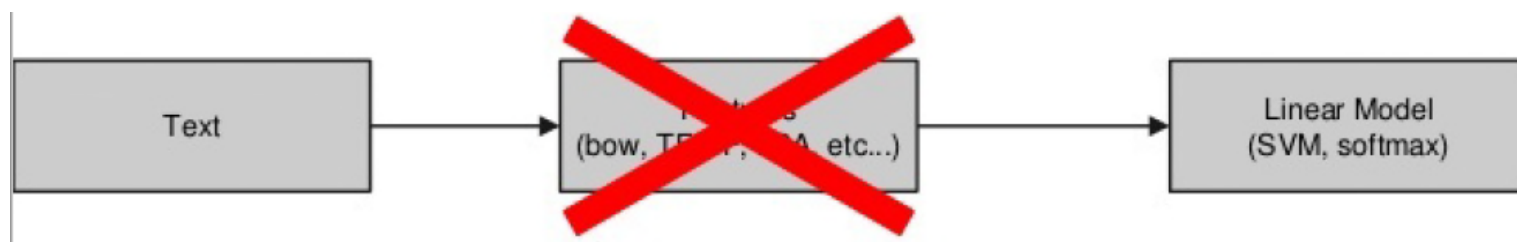
1-3 機械学習による方法

1-3-1 リカレントニューラルネットワーク(RNN)

– Recurrent Neural Networks for text analysis




1-3-1 リカレントニューラルネットワーク(RNN)



1-4 Chatbot

```
=====
EEEEEEEE L          IIIIII ZZZZZZZ AAA
E         L          I      Z      A    A
E         L          I      Z      A    A
EEEEEE   L          I      Z      A    A
E         L          I      Z      AAAAAA
E         L          I      Z      A    A
EEEEEEEE LLLLLLLL IIIIII ZZZZZZ  A    A
=====
ELIZA > (PLEASE TYPE IN ALL CAPS) WHAT'S YOUR NAME DEAR ?
AMIT
ELIZA > HI, I'M ELIZA. WHAT DO YOU WANT TO TALK ABOUT ?
AMIT > CAN I TALK ABOUT MY PROBLEM ?
ELIZA > SURE... !
AMIT > DUE TO ON GOING EXAMS IN COLLEGE, I HAVE A LOT OF STRESS.
ELIZA > PLEASE RELAX A LITTLE. SLEEP WELL.
AMIT > THANKS FOR YOUR ADVICE.
ELIZA > NO MENTION.
AMIT > BYE
ELIZA > BYE AND KEEP IN TOUCH...
=====
```

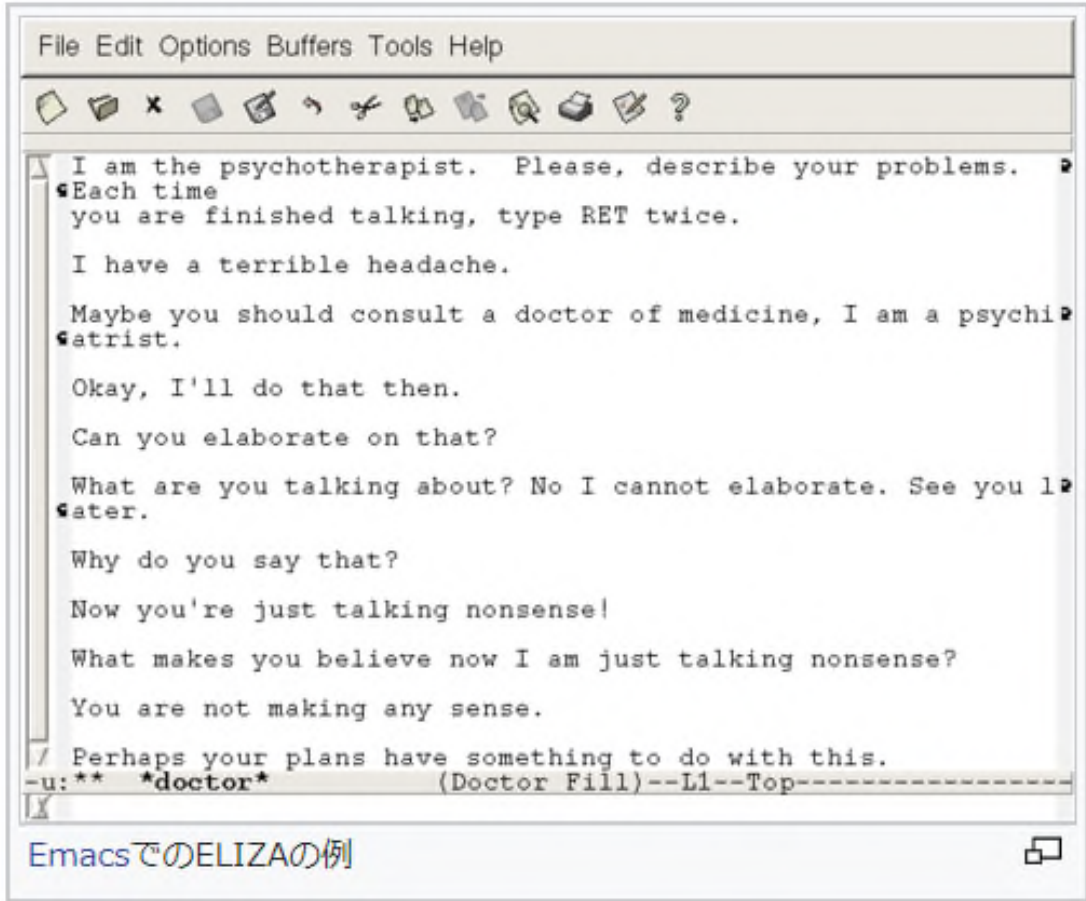


1-4 Chatbot

ELIZA

ELIZA（イライザ）は初期の素朴な自然言語処理プログラムの1つである。対話型（インタラクティブ）であるが、音声による会話をするシステムではない。スクリプト (script) へのユーザーの応答を処理する形で動作し、スクリプトとしては**DOCTOR**という来談者中心療法のセラピストのシミュレーションが最もよく知られている。人間の思考や感情についてほとんど何の情報も持っていないが、DOCTORは驚くほど人間っぽい対話をすることがあった。MITのジョセフ・ワイゼンバウムが1964年から1966年にかけてELIZAを書き上げた。いわゆる人工無脳の起源となったソフトウェアである。

ユーザー（患者役）の入力する文がDOCTOR内の非常に小さな知識ベースの範囲外のものだった場合、DOCTORは一般的な応答を返す。例えば、「頭が痛い」と言えば「なぜ、頭が痛いとおっしゃるのですか？」などと返し、「母は私を嫌っている」と言えば「あなたの家族で他にあなたを嫌っている人は？」（この場合「母」が「家族」の下位概念である、という知識ベースは必要である）などと返す。単純なパターンマッチ技法を使っているが、一部のユーザーはワイゼンバウムがその仕組みを説明しても納得せず、ELIZAの応答を真剣に受け止めた。



1-4 Chatbot: 従来からの方法



チャットボットとはコミュニケーション自動化するプログラム - AIベースのビジネスチャット InCircle

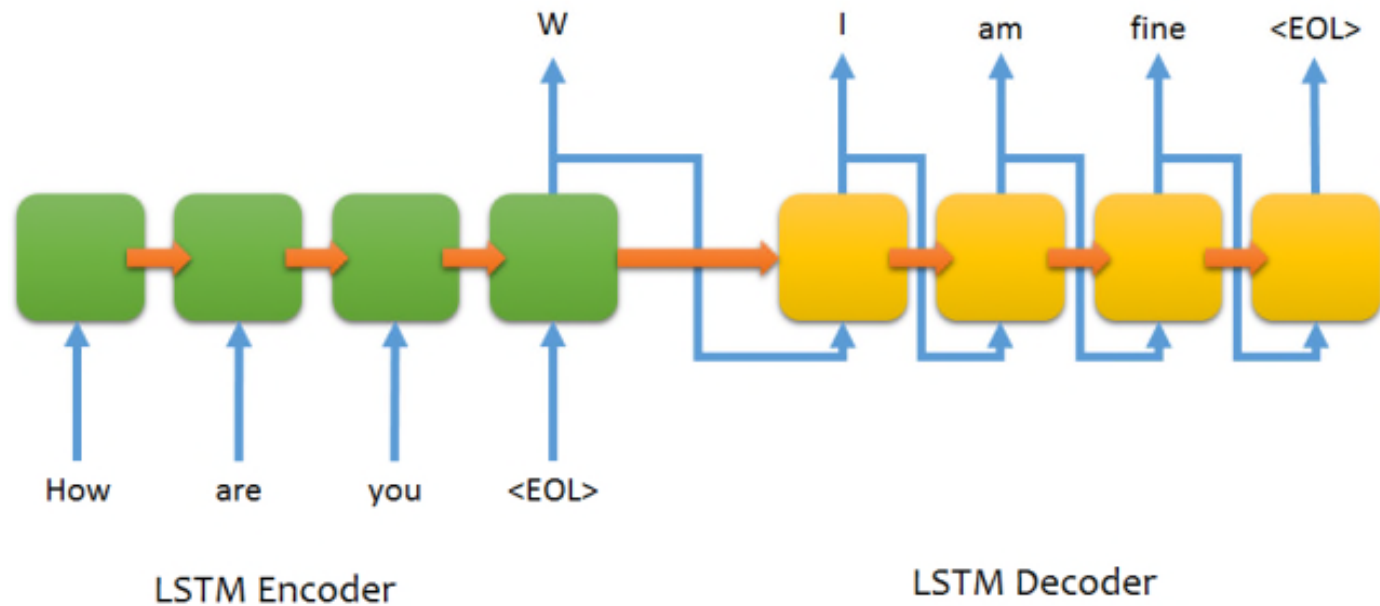
チャットボットの普及が人工知能を強化し、よりよい未来を引き寄せる（書籍出版のお知らせ） | okinawa.io

【LINE】chatbotの開発・普及に向けて新たな展開を発表、新たなMessaging APIを公開し、開発者への正式提供を開始 | LINE Corporation | ニュース

Line Messaging API：目指せ 1000万円！！AI と雑談ができる Line Bot を作ってみた

1-4 Chatbot: 機械学習(RNN)による方法

Chatbots with Seq2Seq



1. https://youtu.be/5_SAroSvC0E
2. <https://youtu.be/SJDEOWLHYVo>
3. <https://youtu.be/t5qgjJIBy9g>
4. https://qiita.com/K_Yagi/items/7148c533ad7ea4226361
5. <https://youtu.be/bUwiKFTvmDQ>

Perform sentiment analysis with LSTMs, using TensorFlow

“rnn text classification tensorflow”で動画検索

- **Deep Learning for NLP**

- Question Answering - The main job of technologies like Siri, Alexa, and Cortana
- Sentiment Analysis - Determining the emotional tone behind a piece of text
- Image to Text Mappings - Generating a caption for an input image
- Machine Translation - Translating a paragraph of text to another language
- Speech Recognition - Having computers recognize spoken words

Sentiment Analysis with LSTMs

You can download and modify the code from this tutorial on GitHub here.

In this notebook, we'll be looking at how to apply deep learning techniques to the task of sentiment analysis. Sentiment analysis can be thought of as the exercise of taking a sentence, paragraph, document, or any piece of natural language, and determining whether that text's emotional tone is positive, negative or neutral.

This notebook will go through numerous topics like word vectors, recurrent neural networks, and long short-term memory units (LSTMs). After getting a good understanding of these terms, we'll walk through concrete code examples and a full Tensorflow sentiment classifier at the end.

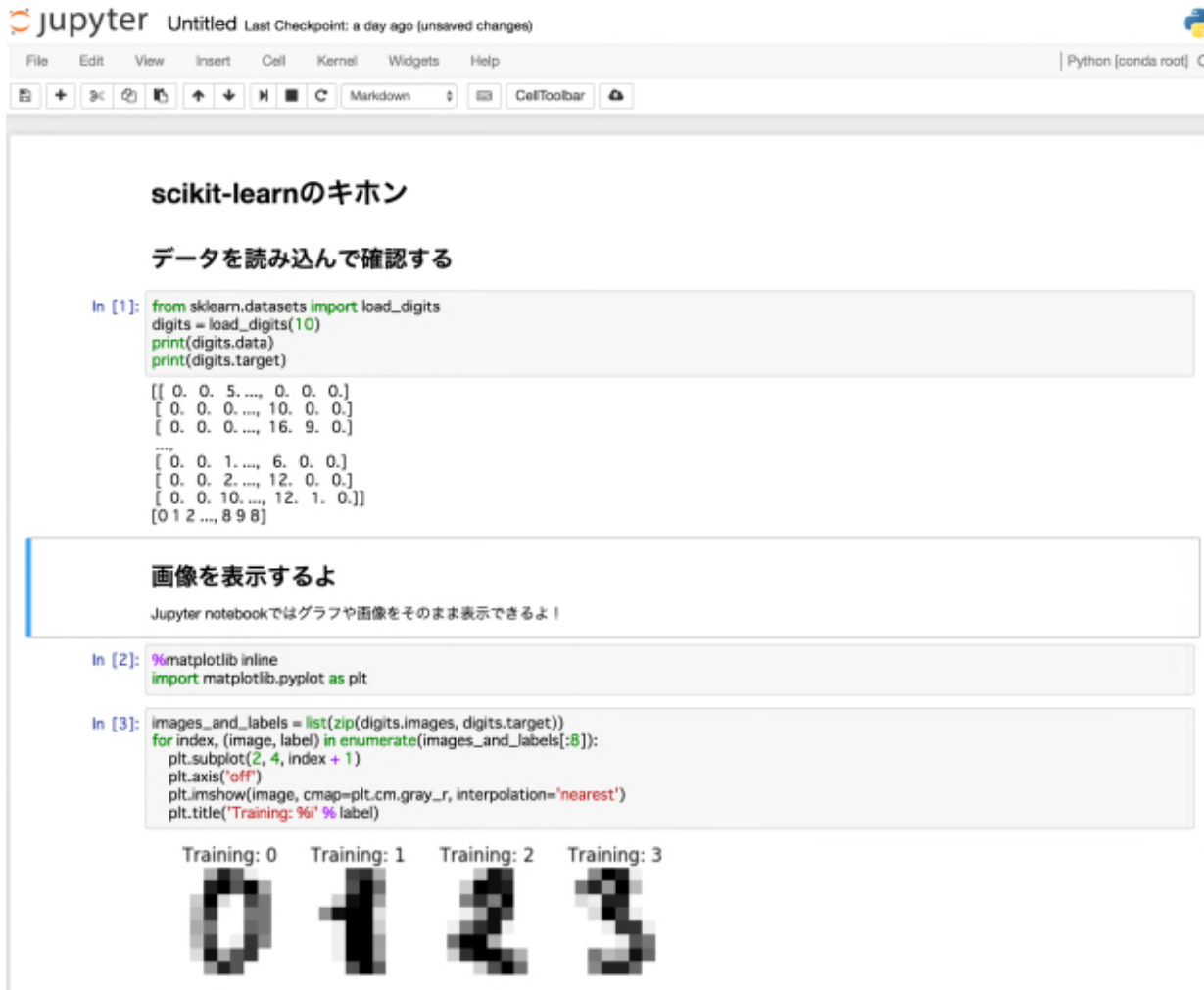


Let us know what you think!

☒ auto scroll/pause

2. ツール類(無償のもの)

- Jupyter notebook



The screenshot shows a Jupyter Notebook window titled "Untitled" with a menu bar (File, Edit, View, Insert, Cell, Kernel, Widgets, Help) and a toolbar. The notebook content is in Japanese and includes the following sections and code:

scikit-learnのキホン

データを読み込んで確認する

```
In [1]: from sklearn.datasets import load_digits
digits = load_digits(10)
print(digits.data)
print(digits.target)
```

```
[[ 0. 0. 5. ..., 0. 0. 0.]
 [ 0. 0. 0. ..., 10. 0. 0.]
 [ 0. 0. 0. ..., 16. 9. 0.]
 ...,
 [ 0. 0. 1. ..., 6. 0. 0.]
 [ 0. 0. 2. ..., 12. 0. 0.]
 [ 0. 0. 10. ..., 12. 1. 0.]]
[0 1 2 ..., 8 9 6]
```

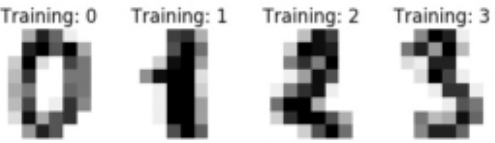
画像を表示するよ

Jupyter notebookではグラフや画像をそのまま表示できるよ！

```
In [2]: %matplotlib inline
import matplotlib.pyplot as plt
```

```
In [3]: images_and_labels = list(zip(digits.images, digits.target))
for index, (image, label) in enumerate(images_and_labels[:8]):
    plt.subplot(2, 4, index + 1)
    plt.axis('off')
    plt.imshow(image, cmap=plt.cm.gray_r, interpolation='nearest')
    plt.title('Training: %i' % label)
```

Training: 0 Training: 1 Training: 2 Training: 3



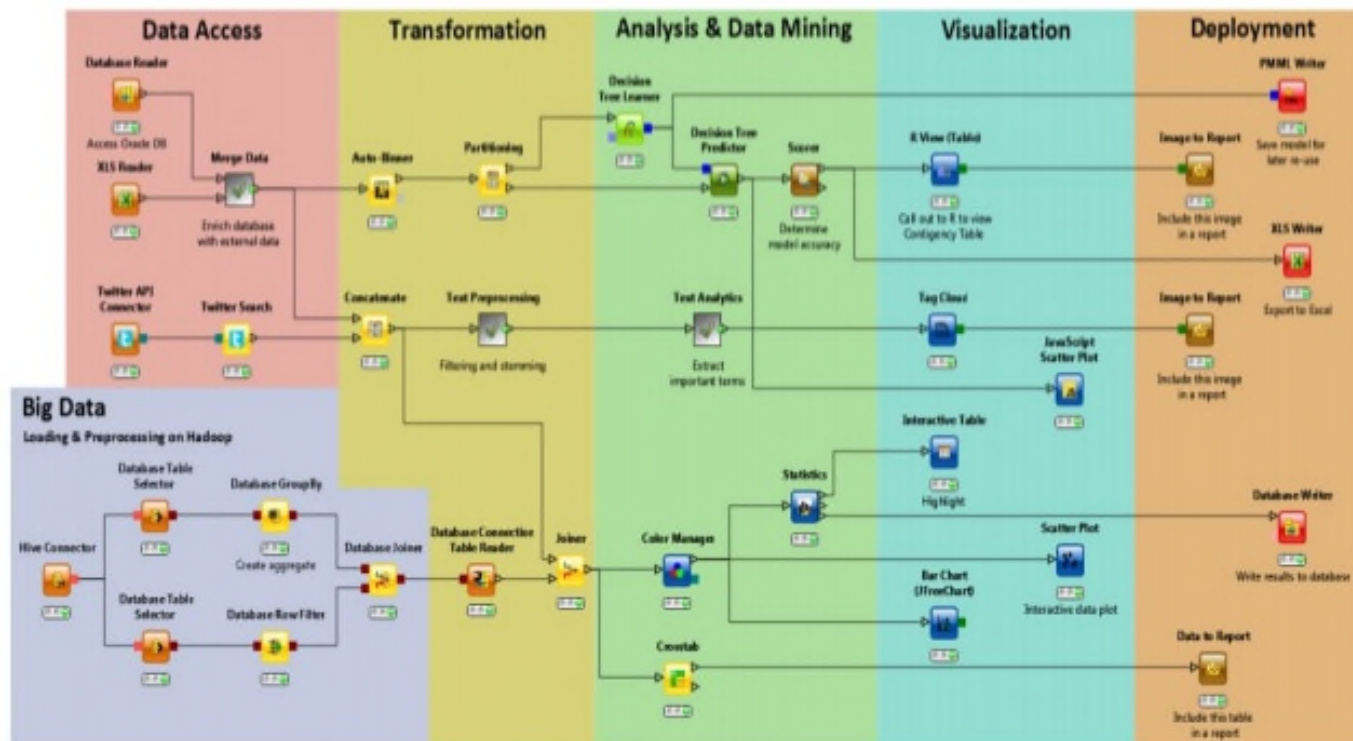
1. [エクセルの次はとりあえず、jupyter notebook – Qiita](#)
2. [IPython Notebook\(Jupyter\)って何ができるの？ - Fire Engine](#)
3. [データ分析の必需品「Jupyter Notebook」の魅力とは – DeepAge](#)
4. [jupyterの環境構築をして、簡単にPythonを書いてみよう | Pythonで機械学習vol.1 | TechClips](#)

2. ツール類(無償のもの)

- Knime



KNIME Analytics Platform



KNIME provides over 1000 nodes to cover every aspect of the analytic process

1. [アジャイル・データ分析
— 先進的なワークフロー
と R の統合](#)
2. [KNIMEトレーニングコー
スのお知らせ | クラソル |
CrowdSolving](#)

見える化との関係の確認

1. テキスト分析の基本(v1.01).pptx

4 関連情報

- HRデータサイエンティスト育成研究会
 - <https://github.com/t-magic/HRDS/wiki>
 - <https://github.com/t-magic/HRDS/wiki/STEP>
 - 問い合わせ
 - tateno.masakazu@gmail.com

付録

3 分析事例

poLCA (<https://github.com/t-magic/poLCA/blob/master/poLCA.md>)

原文表

- ここではサンプルデータとして、フリーアンサー10,859件中に重要度の高い100個の名詞が含まれているかどうかを示した原文表を用います。「現在使用しているファンデーションの気に入っているところは何ですか？」がフリーアンサーの問いです。

	A	E	O	D	E	F	G	H	OR	OS	OT	OU	OV	OW	OX
1	番号	コメント(FA)	肌(2259)	自然(721)	価格(868)	色(650)	自分(468)	感じ(46)	安価(81)	効果(38)	肌質(37)	香料(51)	お値段(6)	厚化粧(22)	さそう(14)
2	1	じぶんにあってる	0	0	0	0	0	0	0	0	0	0	0	0	0
3	2	しっとり肌になじむ感じ、密っぽく厚くない。	1	0	0	0	0	1	0	0	0	0	0	0	0
4	3	安い	0	0	0	0	0	0	0	0	0	0	0	0	0
5	4	毛穴が消える。厚くならない。表情が明るくなる。	0	0	0	0	0	0	0	0	0	0	0	0	0
6	5	あんまりあらくなくてもかばーしてくれるところ。	0	0	0	0	0	0	0	0	0	0	0	0	0
7	6	カバーメイクができる 紫外線カット率が高い	0	0	0	0	0	0	0	0	0	0	0	0	0
8	7	微粒子	0	0	0	0	0	0	0	0	0	0	0	0	0
9	8	使い勝手が良い 海綿でつきやすくべとつきも早く汗で取れる事も悪い	0	0	0	0	0	0	0	0	0	0	0	0	0
10	9	保湿効果が高く、パウダーなのにしっとりした仕上がる。	0	0	0	0	0	0	0	1	0	0	0	0	0
11	10	肌の付き具合、伸びが良い。肌がきれいに見える。	1	0	0	0	0	0	0	0	0	0	0	0	0
12	11	値段はそんなに高くないのにカバー力がある	0	0	1	0	0	0	0	0	0	0	0	0	0
10848	10847	長年コーセーを使用しているので、カウンターですめられて、	0	0	0	0	0	0	0	0	0	0	0	0	0
10849	10848	値段が安く、UVにも対応しているところ。	0	0	1	0	0	0	0	0	0	0	0	0	0
10850	10849	自分に合った色があり、カバー力もある。	0	0	0	1	1	0	0	0	0	0	0	0	0
10851	10850	伸びがよくて軽い感じがする	0	0	0	0	0	1	0	0	0	0	0	0	0
10852	10851	のりがいい。	0	0	0	0	0	0	0	0	0	0	0	0	0
10853	10852	暑さが高くても自然な仕上がりになるところ。	0	1	0	0	0	0	0	0	0	0	0	0	0
10854	10853	肌にあっていい。使いやすい。くずれにくい。	1	0	0	0	0	0	0	0	0	0	0	0	0
10855	10854	特に無い	0	0	0	0	0	0	0	0	0	0	0	0	0
10856	10855	香料の匂いがないところ。刺激が少ないところ。顔に塗ったときノリがいいところ。	0	0	0	0	0	0	0	0	0	1	0	0	0
10857	10856	伸びが良い。自分の肌に合っている。	1	0	0	0	1	0	0	0	0	0	0	0	0
10858	10857	カバー力がよい上にナチュラルに仕上がる。	0	0	0	0	0	0	0	0	0	0	0	0	0
10859	10858	安いところ。	0	0	0	0	0	0	0	0	0	0	0	0	0
10860	10859	肌にやさしい	1	0	0	0	0	0	0	0	0	0	0	0	0

- この原文表は、PPMPで作成しました。

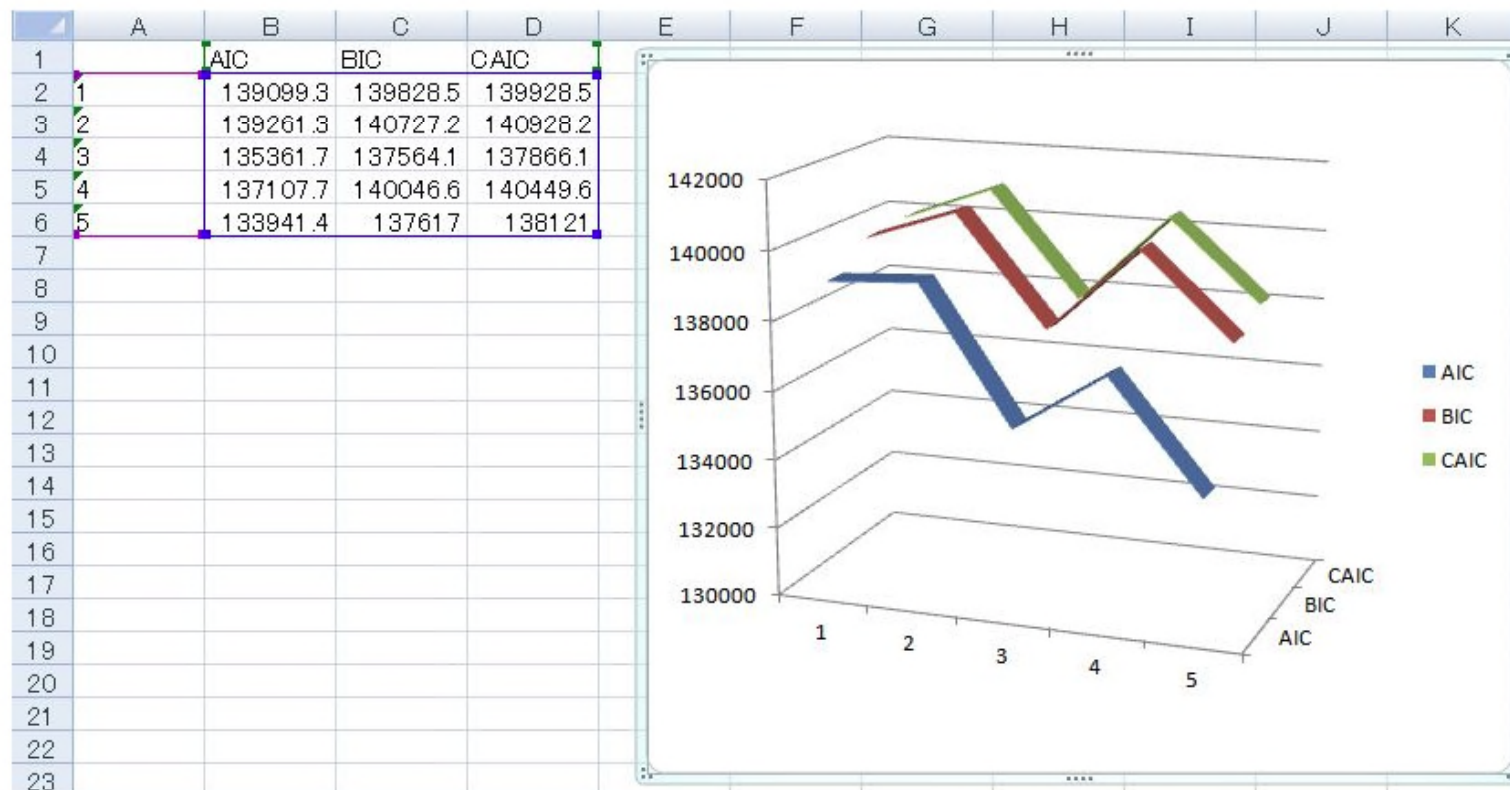
3 分析事例

poLCA (<https://github.com/t-magic/poLCA/blob/master/poLCA.md>)

処理結果

最適な潜在クラス数の確認

- M3.xlsx
- AICで見ると、潜在クラス数が5個のときに最小値なので、5が最適な潜在クラス数である。
- BICで見ると、潜在クラス数が3個のときに最小値なので、3が最適な潜在クラス数である。



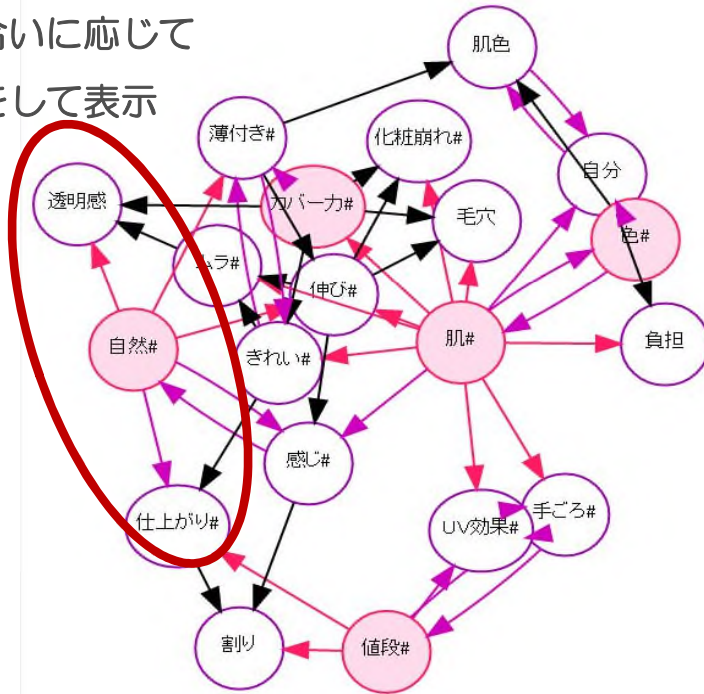
poLCAによるクラスターの利用例

多語グラフやウニグラフのノードをまとめ、コメントを付与する

「使用中ファンデーションの良い所は？」に対し、女性3,000名に自由記述アンケート実施した内の、40代(870名)の回答結果

多語グラフ

つながりの多い語とそのつながり方を、
その度合いに応じて
色付けをして表示



【凡例】

キーワードの重み

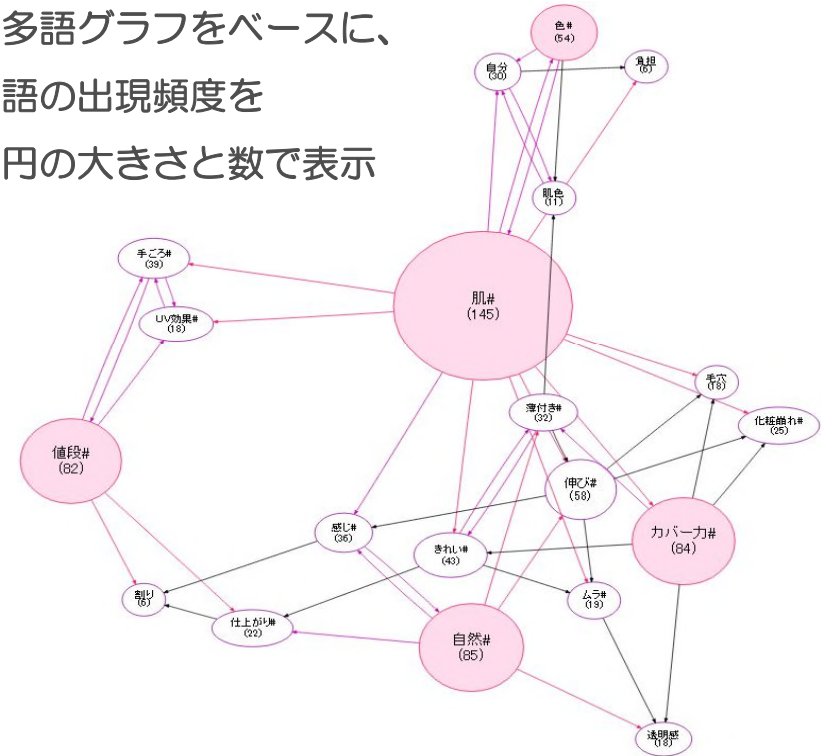
- 多くの関係を持つ語のTOP5
- の次に関係を多く持つ語
- 上記以外

キーワードの関連性

- ● のノードのTOP3からの矢印
- 一緒に語られることが多い関係を示す
- 上記以外

ウニグラフ

多語グラフをベースに、
語の出現頻度を
円の大きさと数で表示



➡ 多く語られている内容の概観が可能

TextMagic(PPMP)表示例

分析対象設定と対象数傾向（モザイクロット）

語の出現頻度による傾向比較（比率差グラフ）

語のつながりによる意味概要（多語グラフ・ユニグラフ）

解釈支援表示（意味チャンク・原文表）

分析対象設定と対象数傾向

表示例 1

テキストの分類分けが設定されていれば
(例えば発言者の年代など)・・・

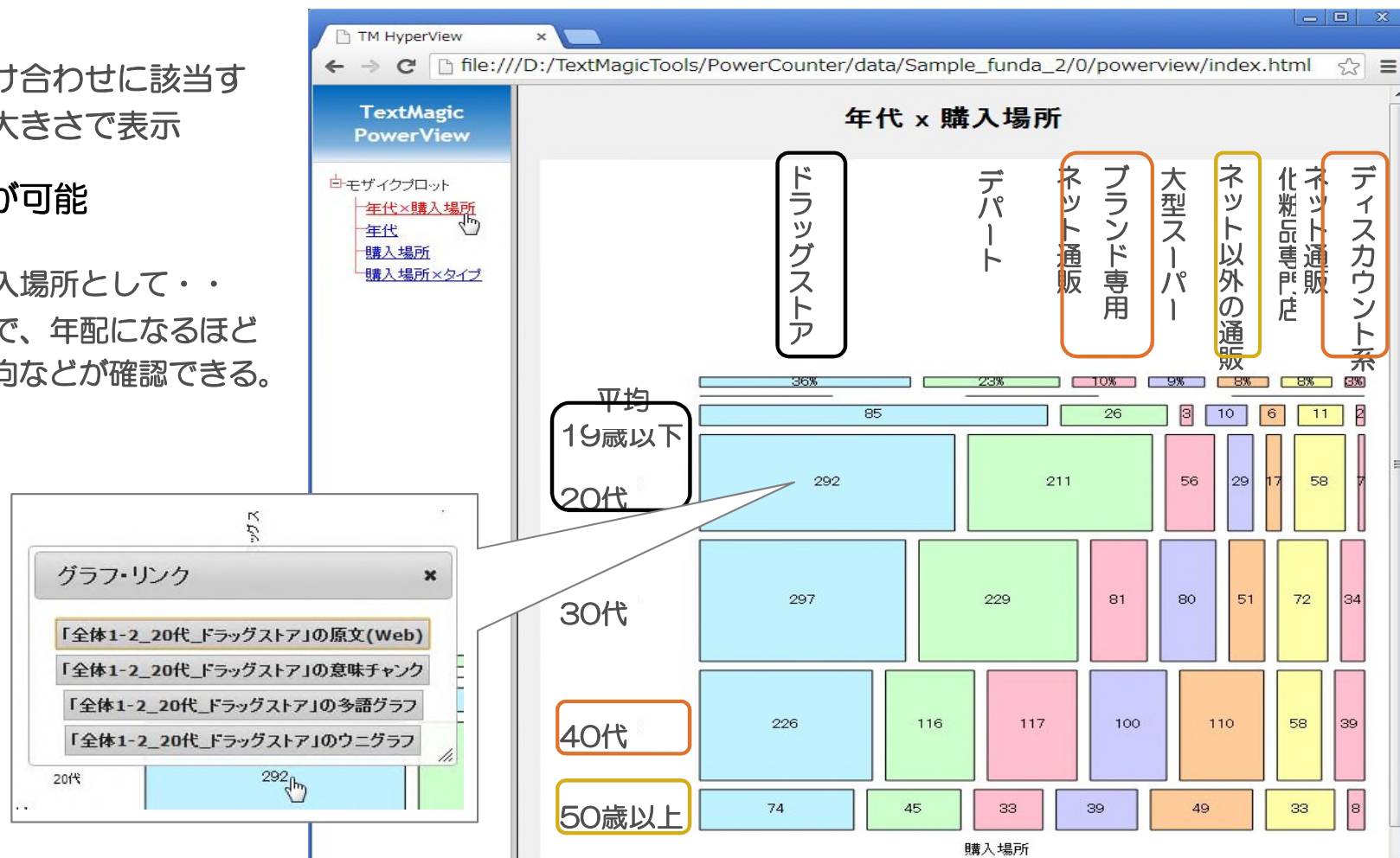
2つの分類(属性)の掛け合わせに該当する
テキスト数を、矩形の大きさで表示

➡ 分類数の傾向把握が可能

【右の例では】化粧品の購入場所として・・・
若者はドラッグストアなどで、年配になるほど
通信販売の利用が増える傾向などが確認できる。

それぞれの矩形もしくは
縦軸ラベル集計単位で
所属するテキストの
分析等の表示が可能

モザイクプロット



語の出現頻度による傾向比較

表示例 2

比率差グラフ

分類別のキーワードの出現比率と全体における比率との差により、多い場合(+)は青色で、少ない場合(-)は赤色でその度合いを表示

語	全体集合	20代	30代	40代
1 肌#	9.75%	0.17%	-0.22%	-1.03%
2 カバー力#	5.08%	0.73%	0.35%	-0.13%
3 自然#	4.88%	-0.85%	-0.32%	0.13%
4 色#	4.10%	0.68%	0.20%	-0.80%
5 値段#	3.89%	-1.29%	0.21%	0.94%
6 きれい#	3.24%	1.20%	-0.32%	-0.59%
7 伸び#	3.23%	0.19%	-0.20%	0.19%
8 薄付き#	2.48%	0.66%	0.44%	-0.60%
9 感じ#	1.99%	-0.69%	-0.09%	0.14%
10 自分	1.95%	0.03%	0.40%	-0.19%
11 手ごろ#	1.90%	-0.88%	0.20%	0.39%
12 心地#	1.82%	0.44%	0.43%	-0.64%
13 毛穴	1.64%	0.55%	0.10%	-0.58%
14 仕上がり#	1.51%	0.27%	0.03%	-0.21%
15 透明感	1.09%	0.48%	-0.32%	-0.03%
16 化粧崩れ#	0.98%	-0.09%	-0.52%	0.50%
17 リキッドタイプ#	0.94%	0.36%	-0.23%	-0.00%
18 ムラ#	0.79%	-0.38%	-0.03%	0.33%
19 UV効果#	0.79%	-0.11%	-0.13%	0.27%
20 ファンデーション#	0.73%	0.02%	-0.06%	-0.08%

■ 全体に比べて多い

■ 全体に比べて少ない

➡ 分類別の多く語られている語の傾向確認が可能

【左の例では】

使用中ファンデーションの好きなところとして・・・

20代では・・・

色・きれい・毛穴・透明感といった仕上りの見た目に関する語が、

40代では・・・

値段が手頃なこと、及び化粧崩れ、ムラ、UV効果等の機能的なことに関する語が、

解釈支援表示

意味チャンク一覧
(語のつながり)



原文表
(語による絞込)

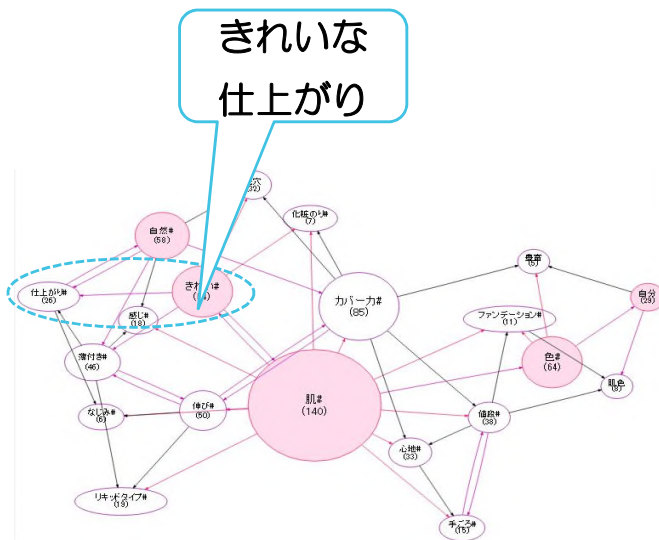


特徴的な言葉の
意味概要等を付加

20代女性

白-きれい# (37)
 白-見える (16)
 白-に (15)
 ...自然_で;肌_が;きれい_に;; 見える;; ; 791
 ...する;感じ_で;見た目_が;きれい_に;; 見える;; ; 954

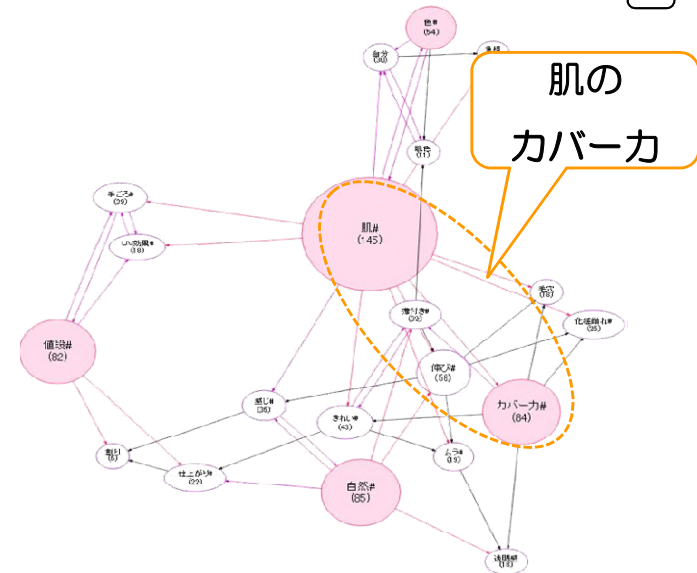
好きなところ (FA)	肌#	色#	きれい#	自然#	自分#
薄付きなのに、 肌がきれい に見えるところ、 自然な感じ に仕上がるところが好きです。	1	0	1	1	0
仕上がり が 自然 になるところ。素肌より きれい に見えるところ。	0	0	1	1	0
外装がしゃれている。 きれいな肌 に見える。	1	0	1	0	0
手軽に肌を 綺麗 に仕上げられる。使う量が少量ですむので、経済的。粒子が細かくて、軽いので、塗っても違和感がさほどない。	1	0	1	0	0
付け心地がよい。発色がよく、 肌がきれい に見える。	1	0	1	0	0
ぶきっちょな私でも、 綺麗 な 肌 に見えるようになるところ。 薄付き で 自然 な 肌 になる。	1	0	1	1	0



40代女性

白-カバー力# (73)
 白-ある (58)
 白-が (49)
 ...いる;適度;カバー力_が;; ある;; ; 22
 ...細かい;カバー力_が;; ある;; ; 26

好きなところ (FA)	肌#	自然#	値段#	カバー力#	色#
薄付きなのに カバー力 が有り 化粧 れしにくく私には合っていると思います。	0	0	0	1	0
伸び が良くて 薄つき けど意外 カバー力 がある。 化粧 れしにくい。	0	0	0	1	0
粒子が細かくて カバー力 があり、 自然な感じ に仕上がる。	0	1	0	1	0
オイルフリー、無香な点。 伸び がよく、 カバー力 もあるところ。 値段 が 手ごろ なところ。インターネットでいつでも購入手配ができるところ。	0	0	1	1	0
カバー力 があるのに 自然な仕上がり が好きで気に入っている	0	1	0	1	0
値段 が安い 割 りに適当な カバー力 もあって、手間がかからず使いやすい。	0	0	1	1	0



参考資料

機械学習

機械学習とは、データから反復的に学習し、そこに潜むパターンを見つけ出すことです。そして学習した結果を新たなデータにあてはめることで、パターンにしたがって将来を予測することができます。 人手によるプログラミングで実装していたアルゴリズムを、大量のデータから自動的に構築可能になるため、さまざまな分野で応用されています。

(https://www.sas.com/ja_jp/insights/analytics/machine-learning.html)

[センサ](#)や[データベース](#)などから、ある程度の数のサンプルデータ集合を入力して解析を行い、そのデータから有用な規則、ルール、知識表現、判断基準などを抽出し、[アルゴリズム](#)を発展させる。なお、データ集合を解析するので、[統計学](#)との関連が深い。(<https://ja.wikipedia.org/wiki/機械学習>)

起源	1959年、 アーサー・サミュエル は、機械学習を「明示的にプログラムしなくても学習する能力をコンピュータに与える研究分野」だとした ^[4] 。
アルゴリズムの分類	教師あり学習 : 入力とそれに対応すべき出力(人間の専門家が訓練例にラベル付けすることで提供されることが多いのでラベルとも呼ばれる)を写像する関数を生成する。例えば、 分類 問題では入力ベクトルと出力に対応する分類で示される例を与えられ、それらを写像する関数を近似的に求める。 教師なし学習 : 入力のみ(ラベルなしの例)からモデルを構築する。 データマイニング も参照。
技法	決定木学習 、 相関ルール学習 、 ニューラルネットワーク 、 遺伝的プログラミング 、 帰納論理プログラミング 、 サポートベクターマシン 、 クラスタリング 、 ベイジアンネットワーク 、 強化学習 、 表現学習
ソフトウェア	SAS ・ RapidMiner ・ LIONsolver ・ KNIME ・ Weka ・ ODM ・ Shogun toolbox ・ Orange ・ Apache Mahout ・ scikit-learn ・ mlpy ・ MCMLL ・ OpenCV ・ XGBoost ・ Jubatus などがある。

潜在意味解析

潜在意味解析（英: Latent Semantic Analysis, LSA）は、ベクトル空間モデルを利用した自然言語処理の技法の1つで、文書群とそこに含まれる用語群について、それらに関連した概念の集合を生成することで、その関係を分析する技術である。**潜在的意味解析**とも。

1988年、アメリカ合衆国でLSAの特許が取得されている^[1]。情報検索の分野では、**潜在的意味索引**または**潜在意味インデックス**（英: Latent Semantic Indexing, LSI）とも呼ばれている。

潜在意味解析

<https://www2.deloitte.com/jp/ja/pages/deloitte-analytics/articles/analytics-plsa.html>

<https://www.slideshare.net/ksmzn/topicmodel>

<https://www.slideshare.net/kojiono507/topic-modelchapter21to3>

<https://www.targetingnext.com/plsa/>

text rnn

<https://www.slideshare.net/odsc/alec-radfordodsc-presentation>

Which is better for text classification: CNN or RNN? Which areas of NLP do they better suit to?

<https://www.quora.com/Which-is-better-for-text-classification-CNN-or-RNN-Which-areas-of-NLP-do-they-better-suit-to>

https://en.wikipedia.org/wiki/Latent_semantic_analysis

確率論的潜在意味解析

確率的潜在意味解析（PLSAとしても知られる）、**確率的潜在的意味インデキシング**（PLSI特に情報検索界では、）である[統計的手法](#) 2モードと共起データの分析のため。事実上、PLSAが進化した[潜在意味解析](#)と同様に、隠れ変数に対する親和性の観点から観測変数の低次元表現を導くことができる。

[線形代数](#)に由来し、発生表を（通常は[特異値分解](#)を介して）縮小する標準的な[潜在意味解析](#)と比較して、確率的潜在意味解析は、[潜在クラスモデル](#)から導出された混合分解に基づく。

潜在クラスモデル

統計、**潜在クラスモデル（LCM）**は、観察された（通常は離散）のセットに関する多変量のセットに変数潜在変数。潜在変数モデルの一種です。潜在変数は離散的であるため、潜在クラスモデルと呼ばれます。クラスは、変数が特定の値を取る可能性を示す条件付き確率のパターンによって特徴付けられます。

潜在クラス解析（LCA）は、構造方程式モデリングのサブセットであり、多変量のカテゴリデータの症例のグループまたはサブタイプを見つけるために使用されます。これらのサブタイプは「潜在クラス」と呼ばれます。

https://en.wikipedia.org/wiki/Latent_class_model