

テキストマイニングの実習 ー 1日目 ー

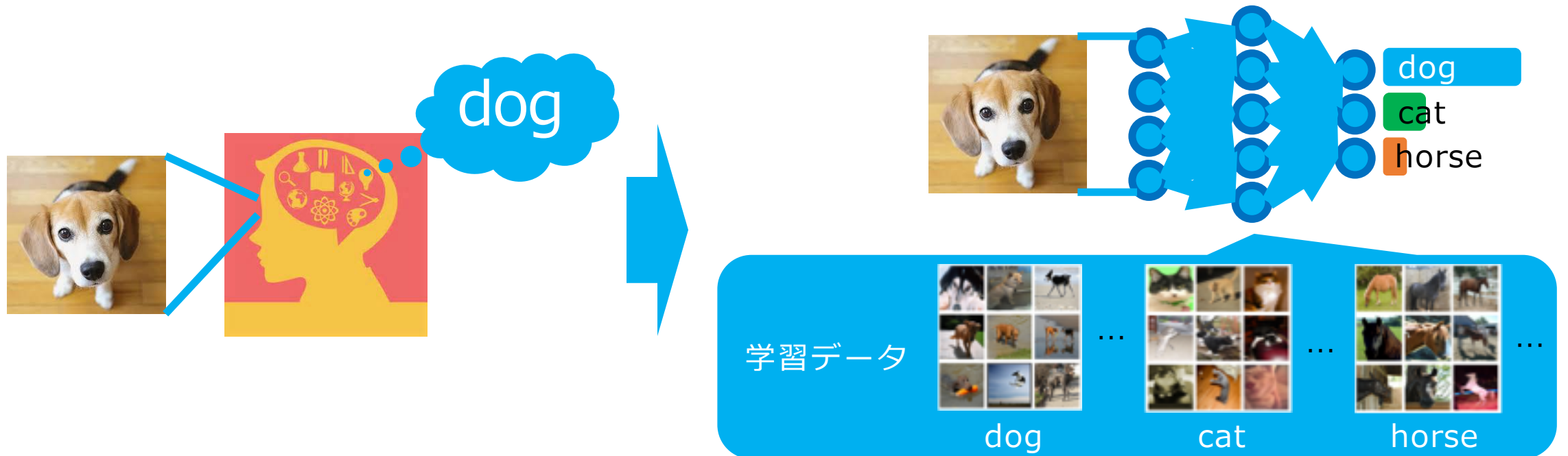
2018/7/5

ビジネス科学研究科
経営システム科学専攻

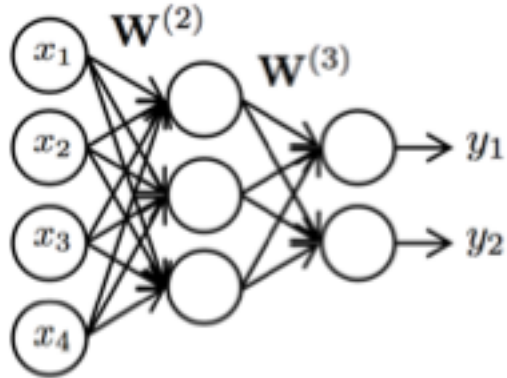
自然言語処理のトレンド

ディープラーニング (深層学習) の成功

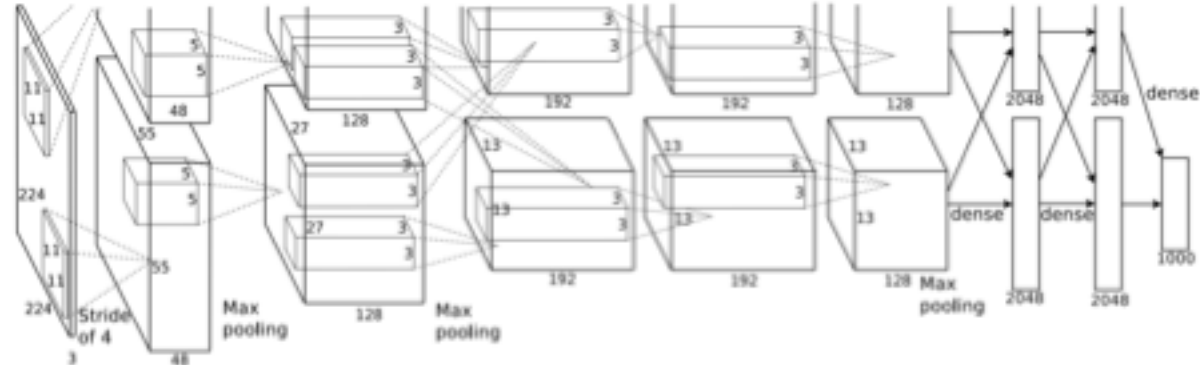
- ニューラルネットワークを用いた機械学習手法
 - 脳の神経細胞(ニューロン)の働きを模した
 - 機械学習とは,データを学習し,パラメータを獲得



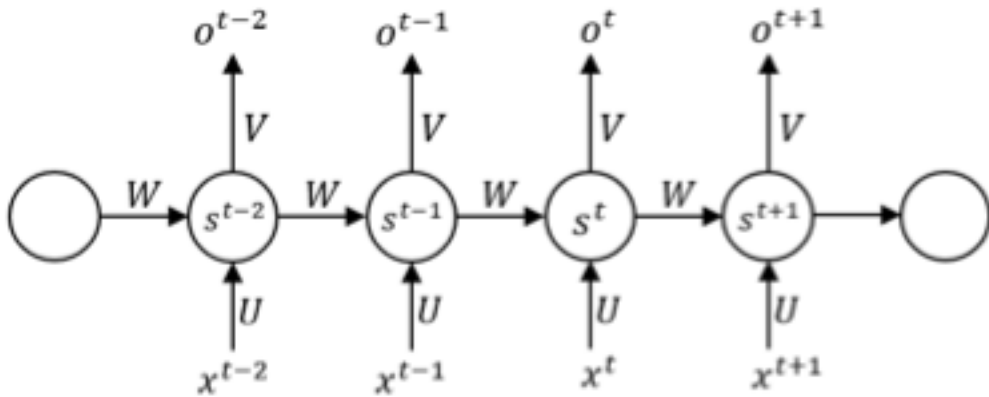
様々なニューラルネット



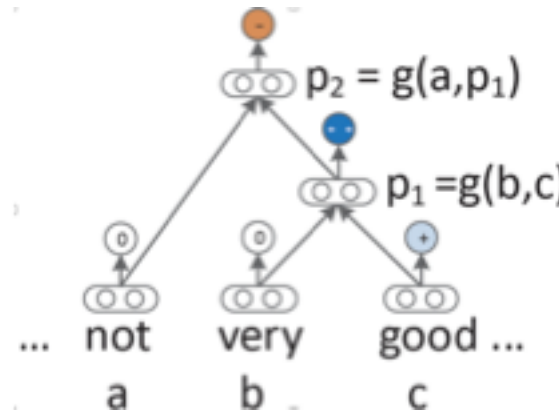
Feed forward NN



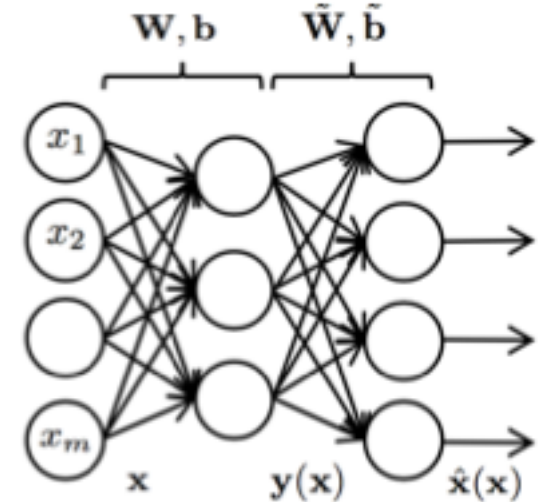
CNN (畳み込みNN)



RNN (Recurrent NN)



Recursive NN



AutoEncoder (自己符号化器)

ニューラルネットの歴史

- ・ 黎明～終焉を繰り返し,近年は 3度目のブーム

第1期	1940～	・ McCullochとPittsが形式ニューロンモデルを発表 [McCulloch-Pitts,43]
	1950～	・ Rosenblattがパーセプトロンを発表 [Rosenblatt,57]
	1960～	・ MinskyとPapertが単純パーセプトロンの(線形分離不可能問題への)限界を指摘 [Minsky-Papert,69]
冬	1970～	冬の時代 (階層的構造の学習方法が未解決)
第2期	1980～	・ Fukushimaらがネオコグニトロンを提案 [Fukushima,80] ・ Rumelhartらが誤差逆伝播法を提案 [Rumelhart+,86] ・ LeCunらが畳み込みニューラルネット Conv.net を提案 [LeCun,89]
	1990～	冬の時代 (学習時間や過学習に課題, 一方でSVMが流行)
第3期	2000～	・ Hintonらが事前学習とオートエンコーダを導入した多層NNを提案 [Hinton+,06]
	2010～	・ Seideらが音声認識のベンチマークで圧勝 [Seide+,11] ・ KrizhevskyらがReLUを提案し画像認識コンペで圧勝 [Krizhevsky,12]



音声認識での成功 [Seide+,2011]

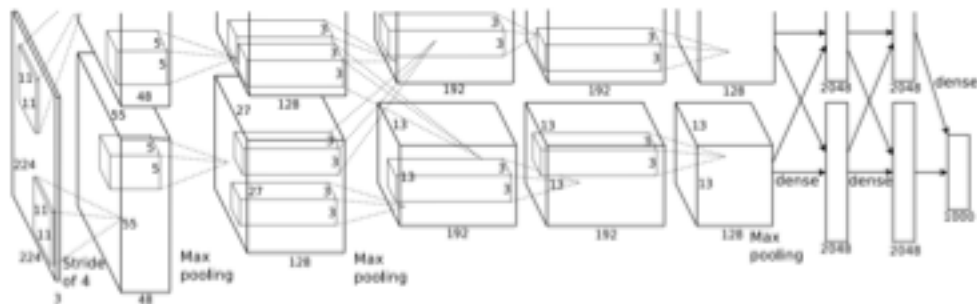
- Microsoft Research のグループ
 - 電話での会話音声の標準データセット
 - 入力(MFCC)-出力(HMM状態変数)の関係をDNNで学習
 - 従来 GMM-HMM → DNN-HMM (全結合7層, 事前学習あり)
 - 単語誤認識率で 10%前後の大幅な精度改善

acoustic model & training	recognition mode	RT03S		Hub5'00	voicemails		tele-
		FSH	SW	SWB	MS	LDC	conf
GMM 40-mix, ML, SWB 309h	single-pass SI	30.2	40.9	26.5	45.0	33.5	35.2
GMM 40-mix, BMMI, SWB 309h	single-pass SI	27.4	37.6	23.6	42.4	30.8	33.9
CD-DNN 7 layers x 2048, SWB 309h, this paper (rel. change GMM BMMI → CD-DNN)	single-pass SI	18.5 (-33%)	27.5 (-27%)	16.1 (-32%)	32.9 (-22%)	22.9 (-26%)	24.4 (-28%)

F. Seide, G. Li and D. Yu, "Conversational Speech Transcription Using Context-Dependent Deep Neural Networks." *Interspeech*. 2011.

画像認識での成功 [Krizhevsky+, 2012]

- 一般物体認識 (Hintonのグループ)
 - ImageNet Large-scale Visual Recognition Challenge 2012
 - 1000カテゴリ×約1000枚 = 100万枚 の訓練画像
 - 畳込み層5, 全結合層3, 2つのGPUで2週間 (AlexNet)
 - 誤識別率が10%以上減少 (過去数年間での向上は1~2%)

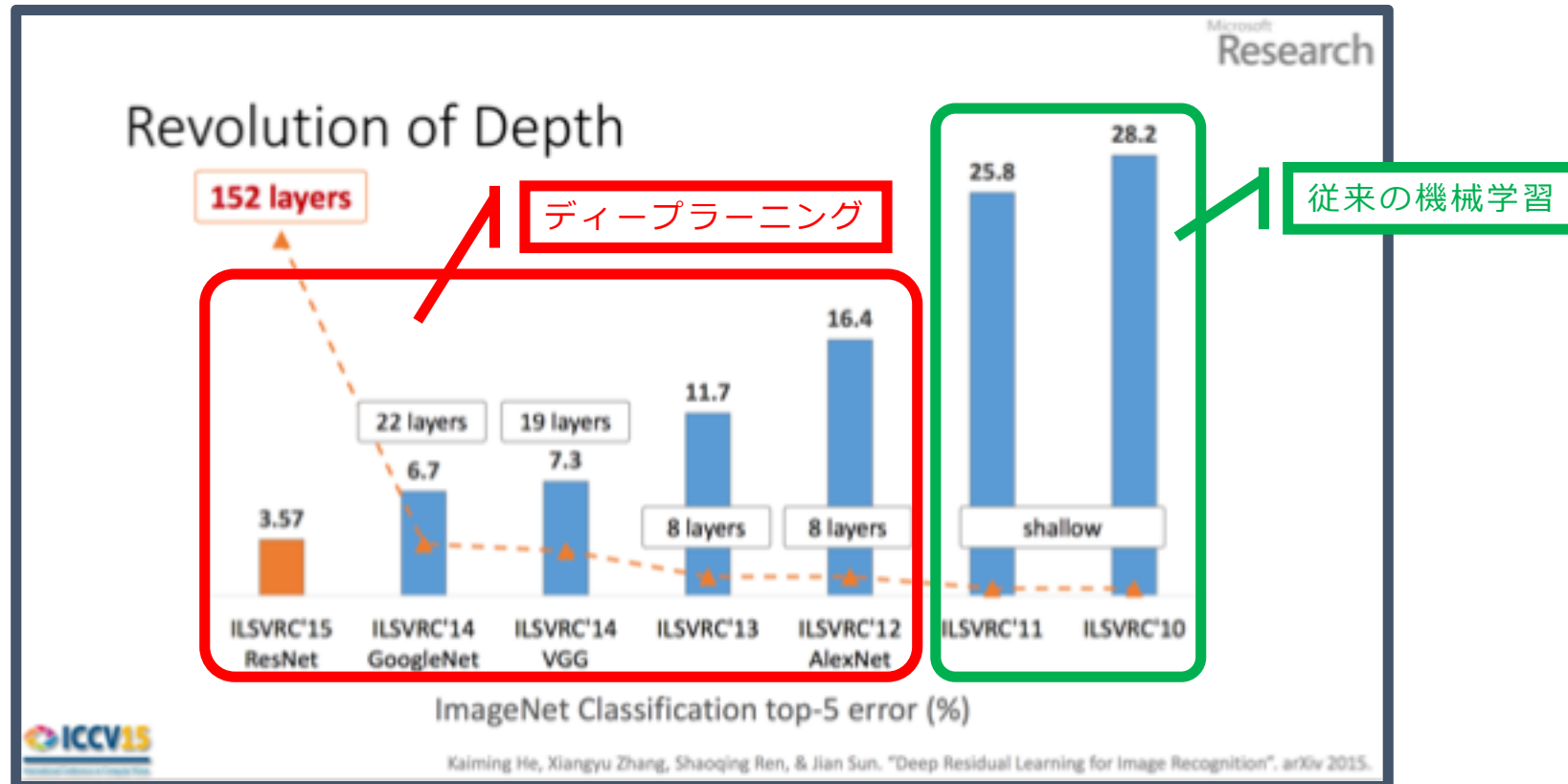


Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton.
"Imagenet classification with deep convolutional neural networks."
Advances in neural information processing systems. 2012.
<http://image-net.org/challenges/LSVRC/2012/supervision.pdf>



一般物体認識における認識精度の変遷

- 2015年,人の認識精度(5.1%)を超えたことが話題になった



ディープラーニング成功の背景

- 一定以上の規模のデータ → 改善
 - WebやIoT(センサ)などから十分な規模のデータを収集可能
- 学習の難しさ → 改善
 - 様々なテクニック (事前学習, dropout 等)
- 誤差逆伝搬法の計算量膨大 → 改善
 - 計算機能力の飛躍的向上
 - GPU, マルチコアCPU, PCクラスタの登場
- 性能を引き出すのに必要なノウハウ → 未解決
 - 「黒魔術」のまま

ディープラーニングによる自然言語処理

- 単語のベクトル表現

既存手法	最近の手法
TF-IDF , Okapi BM25 など (分布的, 高次元, スパース)	word2vec , Glove, fastText など (分散的, 低次元, 密)

- 代表格は「**word2vec**」

- 深層学習による分布仮説のモデル化
- king - man + woman = queen で有名 →

※ 図上に king + (woman-man) = queen を描くとわかりやすい



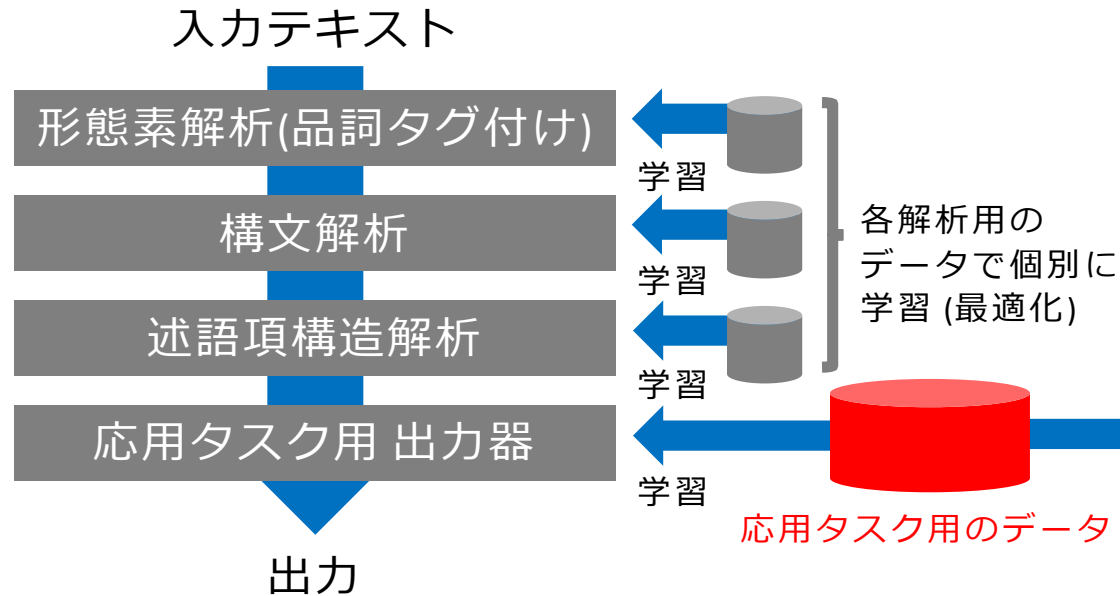
Tomas Mikolov, Wen-tau Yih, Geoffrey Zweig, 2013, NAACL

ディープラーニングによる自然言語処理

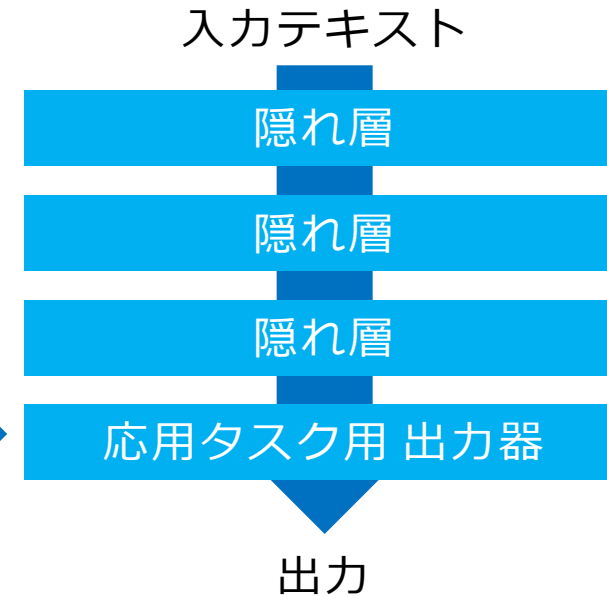
- 応用タスクでも効果を発揮

- End-to-end 学習: 応用タスク用の大規模な訓練データで全体を学習

従来の自然言語処理



ディープラーニングによる自然言語処理



坪井,海野,鈴木. 深層学習による自然言語処理. 講談社, 2017, p.4 の図を一部修正
テキストマイニング

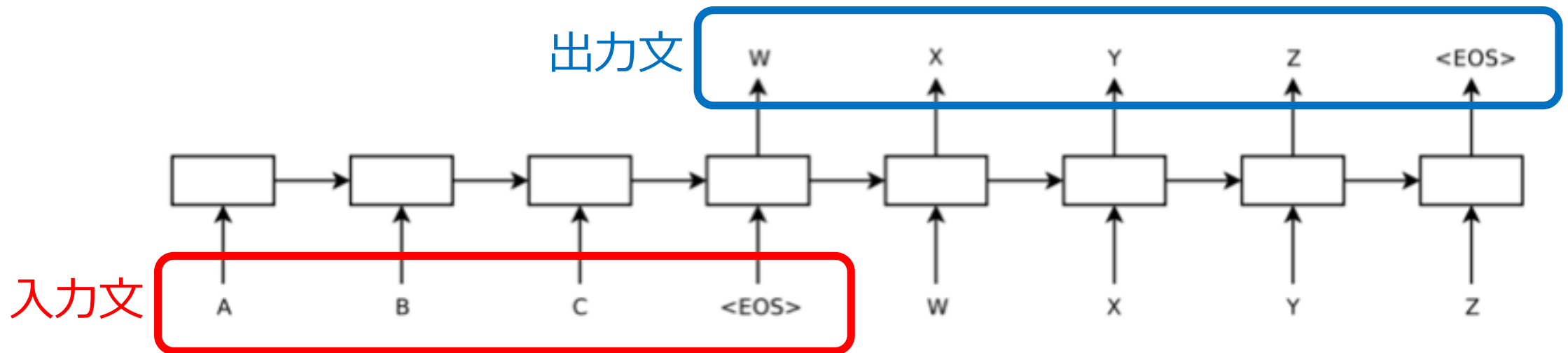
ディープラーニングによる自然言語処理

応用タスク	既存手法	最近の手法	応用先
文のベクトル化	<ul style="list-style-type: none">TF-IDFOkapi BM25	<ul style="list-style-type: none">Recurrent NN ※系列を考慮Recursive NN ※木構造を考慮Convolutional NN ※画像で成功	<ul style="list-style-type: none">文書分類極性(ポジ/ネガ)判定
文の生成 (言語モデル)	<ul style="list-style-type: none">N-gram	<ul style="list-style-type: none">Recurrent NN	<ul style="list-style-type: none">形態素解析音声認識, 文字認識
系列ラベリング	<ul style="list-style-type: none">CRFSVM	<ul style="list-style-type: none">Encoder-Decoder ※ Seq2Seq や Attention機構を含む	<ul style="list-style-type: none">品詞タグ付け固有表現抽出

さらに応用タスク	既存手法	最近の手法
機械翻訳	<ul style="list-style-type: none">統計的機械翻訳 ※ N-gram 言語モデル+ アラインメント(IBM)モデル+ フレーズテーブルなどの技術を複合的に使用	<ul style="list-style-type: none">Encoder-Decoder, Transformer ※ 対訳コーパスを end-to-end で学習する
文書要約	<ul style="list-style-type: none">SVM最大被覆問題	<ul style="list-style-type: none">Encoder-Decoder ※ 原文と要約文を end-to-end で学習する

ニューラル機械翻訳

- Recursive NN による Sequence to Sequence 機械翻訳モデル [Sutskever+,2014]



“ABC”という単語列から“WXYZ”という単語列への翻訳

スケジュール

- 1日目: 7/5
 - 説明 – データ分析の手順
 - 演習 – データの理解 (Excel)
- 2日目: 7/12
 - 説明 – テキストマイニングツールの使い方 (KHCoder)
 - 練習 – テキストマイニングツールの使い方 (KHCoder)
- 3日目: 7/26
 - 演習 – データ分析の実践 (KHCoder)

テキストマイニングとは

- 大量の文書データに記述されている多種多様な内容を対象として、その相関関係や出現傾向などから新たな知識を発見する
[那須川,1999]
- 市場調査や販売戦略の立案, 製品やサービス改善, 顧客対応の改善に役立てたい
 - 昔から
 - 営業日報, 自由記述のアンケート, コールセンタの対応ログ
 - 近年 → CGM (コンシューマー・ジェネレイテッド・メディア) など
 - レビューサイトの口コミ
 - ブログやマイクロブログ (Twitter, Facebook)

口コミサイトの例

- ホテルの口コミ数: 970万件 ※年間約60～70万件増加



経年変化:

780万件 (H27)
→ 836万件 (H28)
→ 900万件 (H29)
→ **973万件 (今回)**

[illegible][illegible]

鴨川シーワールドホテルのクチコミ・お客様の声

ホテル・旅行のクチコミTOPへ



総合評価

★★★★☆

4.12

アンケート件数：886件

評価内訳

5点

4点

3点

2点

1点

項目別の評価

サービス	★★★★☆	4.11
立地	★★★★★	4.61
部屋	★★★★☆	3.53
設備・アメニティ	★★★★☆	3.62
風呂	★★★★☆	3.53
食事	★★★★☆	4.10

サービス

食事

立地

部屋

設備・アメニティ

風呂

総合

★★★★☆

2

投稿者さんの 鴨川シーワールドホテル のクチコミ (感想 情報)

投稿者さん

2015年06月11日 17:03:57

良かったところ

- ・部屋からの景色（朝日最高でした）
- ・食事（品数も多く、朝夕とも良かったです）
- ・フロントの方の対応（お姉さんがとても頑張っていました）以上。

掃除が行き届いているとの口コミを多く見ましたが、そうは思いませんでした。

気にかかることは多々ありましたが、フロントのお姉さんが一生懸命で、その笑顔に救われた思いです。

評価

総合

★★★★☆

2

サービス	2
立地	4
部屋	4
設備・アメニティ	2
風呂	2
食事	4

旅行の目的

レジャー

同伴者

家族

宿泊年月

2015年06月

鴨川シーワールドホテル

2015年06月11日 19:32:50

この度は、ご利用頂きまして誠にありがとうございました。

客室内清掃の件、大変申し訳なく思っております。重要改善として、早急に対応いたします。今後は、このような事の無いように、清掃・点検を強化いたします。

フロントスタッフへお言葉をおかけし、誠にありがとうございます。モチベーションアップに繋がりますように、お客様からの声として、スタッフと共有させていただきます。

機会がございましたら、またご利用をお待ちしております。

テキストデータ

数値評価

テキストマイニングの手順

- データ理解
 - データ件数や構成比を集計 → データに詳しくなる
 - 旅行目的別の人気エリアは?
 - 同伴者別の人気エリアは?
 - 数値評価による人気エリアの差異は?
- テーマ設定
 - 解決すべき課題を決める → 分析目的を明確にする
 - 明らかにしたい事柄は?
 - 確認したい仮説は?
- テキスト分析
 - これら課題を解決するために,テキスト分析を実施

使用するデータ

- 楽天トラベルから収集した「お客様の声」のデータ
 - 宿泊日が2017年,下記の10エリアが対象
 - エリアごとに 1,000件ずつをランダムに選択

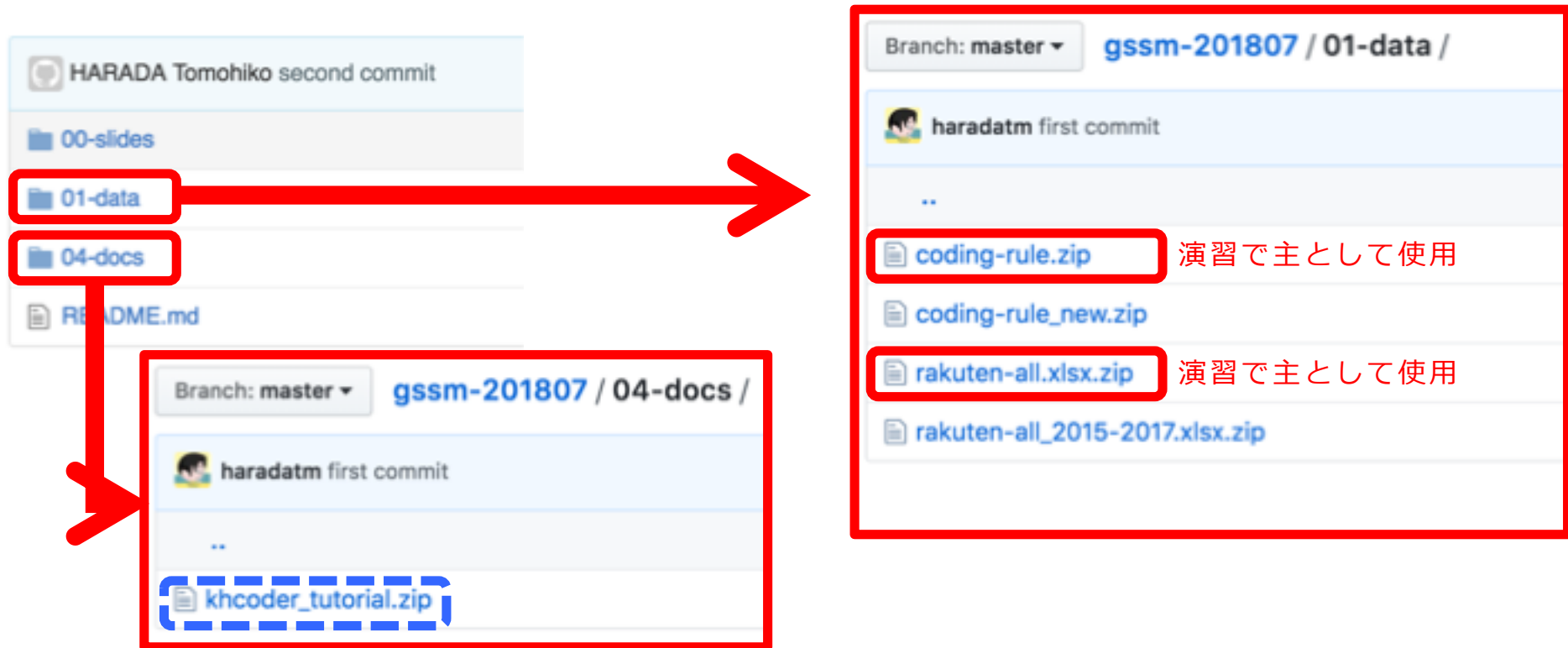
レジャー	5エリア	登別, 草津, 箱根, 道後, 湯布院	1,000件×10エリア = 計10,000件
ビジネス	5エリア	札幌, 名古屋, 東京, 大阪, 福岡	

- データ項目

施設情報	4項目	カテゴリ, エリア, 施設番号, 施設名
口コミ	1項目	コメント
ユーザー評価	7項目	総合, サービス, 立地, 部屋, 設備・アメニティ, 風呂, 食事
その他の分類	2項目	旅行の目的, 同伴者
宿泊日	1項目	宿泊年月
ユーザー情報	3項目	ユーザー, 年代, 性別

使用するデータ

- <https://github.com/haradatm/gssm-201807>



KH Coder のチュートリアル (Windows版に同梱)

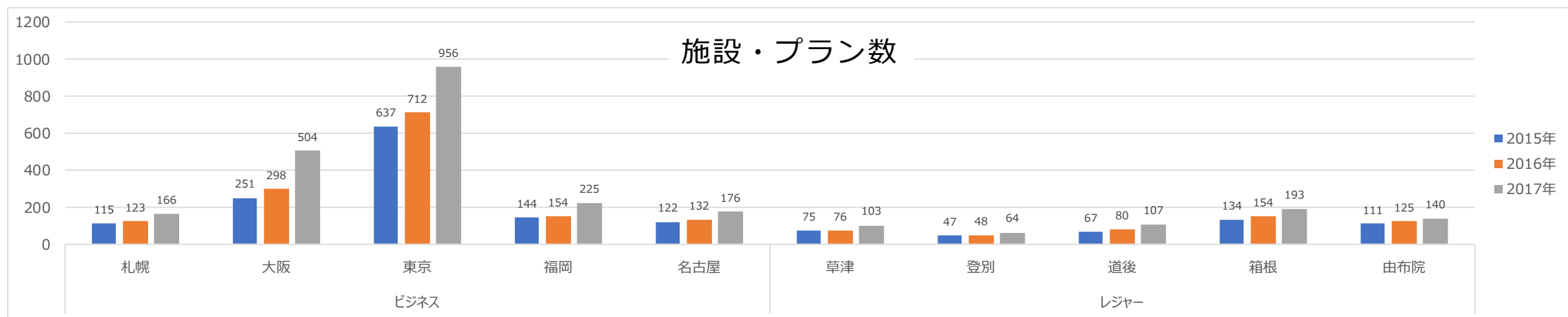
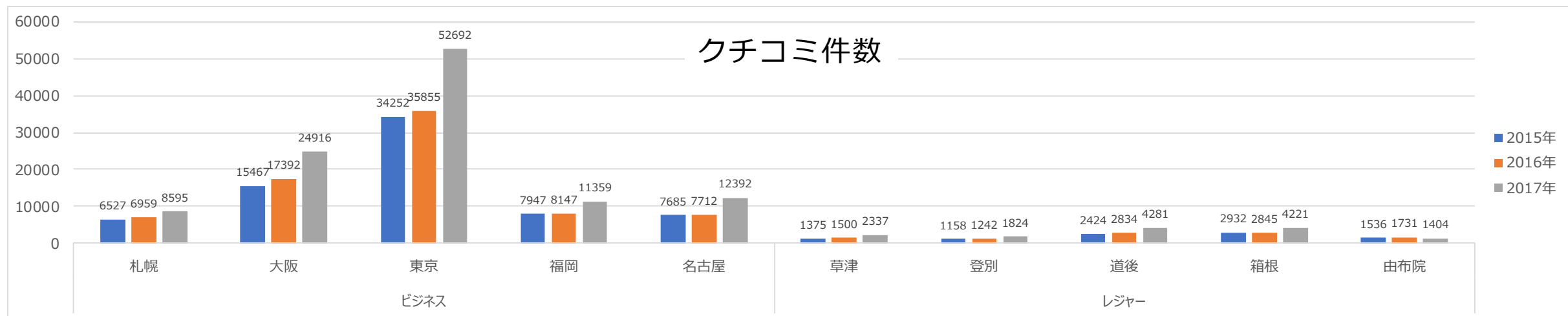
使用するデータ

データファイル名	件数	データセット
rakuten-all.xlsx.zip	10,000	<ul style="list-style-type: none">• レジャー+ビジネスの 10エリア• エリアごと 1,000件• ランダムサンプリング• EXCEL 形式• 2017年のみ (1シート)
rakuten-all_2015-2017.xlsx.zip	30,000	<ul style="list-style-type: none">• レジャー+ビジネスの 10エリア• エリアごと 1,000件• ランダムサンプリング• EXCEL 形式• 2015, 2016, 2017年 (3シート)

参考ー収集データについて

NO.	カテゴリ	エリア	収集したクチコミ件数とサンプリング率						収集した施設・プラン数とサンプリング率					
			2015年	サンプリング率	2016年	サンプリング率	2017年	サンプリング率	2015年	サンプリング率	2016年	サンプリング率	2017年	サンプリング率
1	レジャー	登別	1,158	86.36%	1,242	80.52%	1,824	54.82%	47	100.00%	48	100.00%	64	95.31%
2		草津	1,375	72.73%	1,500	66.67%	2,337	42.79%	75	98.67%	76	100.00%	103	92.23%
3		箱根	2,932	34.11%	2,845	35.15%	4,221	23.69%	134	82.09%	154	87.66%	193	81.35%
4		道後	2,424	41.25%	2,834	35.29%	4,281	23.36%	67	94.03%	80	90.00%	107	85.05%
5		由布院	1,536	65.10%	1,731	57.77%	1,404	71.23%	111	96.40%	125	94.40%	140	95.71%
6	ビジネス	札幌	6,527	15.32%	6,959	14.37%	8,595	11.63%	115	95.65%	123	93.50%	166	86.75%
7		名古屋	7,685	13.01%	7,712	12.97%	12,392	8.07%	122	90.16%	132	85.61%	176	85.23%
8		東京	34,252	2.92%	35,855	2.79%	52,692	1.90%	637	59.18%	712	57.58%	956	48.43%
9		大阪	15,467	6.47%	17,392	5.75%	24,916	4.01%	251	80.88%	298	69.80%	504	57.54%
10		福岡	7,947	12.58%	8,147	12.27%	11,359	8.80%	144	88.19%	154	90.26%	225	82.67%
合計			81,303	12.30%	86,217	11.60%	124,021	8.06%	1,703	77.98%	1,902	75.39%	2,634	67.24%

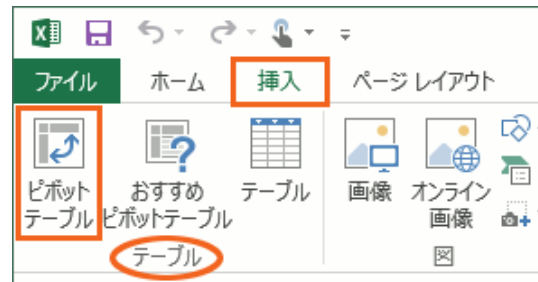
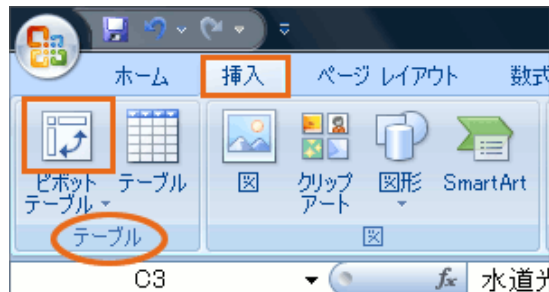
参考 ― 収集データについて



演習 ― データ理解

- ピボットテーブル(EXCEL)を使ってデータを集計する
 - ファイル **rakuten-all.xlsx** を開く
 - A～R 列を選択し,ピボットテーブルを作成する

【Windows】 Excel 2007・2010・2013



[挿入] タブ [テーブル] グループの [ピボットテーブル] ボタンをクリックします

課題ーデータ理解

- EXCELを使ってデータ集計を行い,発見した特徴や傾向をもとにデータセットを説明(要約)する

例) データセットを説明する観点

- 投稿者の属性(年代,性別)は?
- 旅行目的別の人気エリアは?
- 同伴者別の人気エリアは?

参考 — データ集計の例

件数 (エリア別)

行ラベル	個数 / コメント
A_レジャー	5000
01_登別	1000
02_草津	1000
03_箱根	1000
04_道後	1000
05_湯布院	1000
B_ビジネス	5000
06_札幌	1000
07_名古屋	1000
08_東京	1000
09_大阪	1000
10_福岡	1000
総計	10000

投稿者の傾向 (年代別・性別)

個数 / コメント	列ラベル			
行ラベル	男性	女性	na	総計
10代	0.02%	0.02%	0.00%	0.04%
20代	0.65%	0.89%	0.00%	1.54%
30代	2.73%	2.30%	0.00%	5.03%
40代	8.62%	4.21%	0.00%	12.83%
50代	10.20%	3.48%	0.00%	13.68%
60代	4.10%	1.07%	0.00%	5.17%
70代	0.71%	0.12%	0.00%	0.83%
80代	0.06%	0.00%	0.00%	0.06%
na	0.00%	0.00%	60.82%	60.82%
総計	27.09%	12.09%	60.82%	100.00%

投稿者の傾向 (エリア別)

個数 / コメント	列ラベル		
行ラベル	A_レジャー	B_ビジネス	総計
男性	25.44%	28.74%	27.09%
女性	14.04%	10.14%	12.09%
na	60.52%	61.12%	60.82%
総計	100.00%	100.00%	100.00%

参考 — データ集計の例

投稿者の傾向 (性別,エリア別)

個数 / コメント	列ラベル												総計
	A_レジャー					A_レジャー 集計	B_ビジネス					B_ビジネス 集計	
行ラベル	01_登別	02_草津	03_箱根	04_道後	05_湯布院		06_札幌	07_名古屋	08_東京	09_大阪	10_福岡		
男性	27.70%	27.60%	20.50%	30.90%	20.50%	25.44%	28.90%	29.20%	27.40%	28.20%	30.00%	28.74%	27.09%
女性	14.20%	14.60%	16.00%	9.80%	15.60%	14.04%	9.40%	9.20%	12.40%	10.70%	9.00%	10.14%	12.09%
na	58.10%	57.80%	63.50%	59.30%	63.90%	60.52%	61.70%	61.60%	60.20%	61.10%	61.00%	61.12%	60.82%
総計	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

投稿者の傾向 (性別,目的別,エリア別)

個数 / コメント	列ラベル													A_レジャー 集計	B_ビジネス 集計	総計
	A_レジャー					B_ビジネス										
行ラベル	01_登別	02_草津	03_箱根	04_道後	05_湯布院		06_札幌	07_名古屋	08_東京	09_大阪	10_福岡					
男性	27.70%	27.60%	20.50%	30.90%	20.50%		25.44%	28.90%	29.20%	27.40%	28.20%	30.00%	28.74%	27.09%		
レジャー	<div></div> 21.70%	<div></div> 26.70%	<div></div> 19.50%	<div></div> 17.80%	<div></div> 19.30%		<div></div> 21.00%	<div></div> 16.30%	<div></div> 13.20%	<div></div> 12.20%	<div></div> 14.50%	<div></div> 13.50%	<div></div> 13.94%	17.47%		
ビジネス	<div></div> 5.70%	<div></div> 0.50%	<div></div> 0.60%	<div></div> 12.40%	<div></div> 0.60%		<div></div> 3.96%	<div></div> 11.20%	<div></div> 14.70%	<div></div> 14.00%	<div></div> 13.10%	<div></div> 15.80%	<div></div> 13.76%	8.86%		
その他	<div></div> 0.30%	<div></div> 0.40%	<div></div> 0.40%	<div></div> 0.70%	<div></div> 0.60%		<div></div> 0.48%	<div></div> 1.40%	<div></div> 1.30%	<div></div> 1.20%	<div></div> 0.60%	<div></div> 0.70%	<div></div> 1.04%	0.76%		
女性	14.20%	14.60%	16.00%	9.80%	15.60%		14.04%	9.40%	9.20%	12.40%	10.70%	9.00%	10.14%	12.09%		
レジャー	<div></div> 13.90%	<div></div> 14.40%	<div></div> 15.50%	<div></div> 7.30%	<div></div> 15.10%		<div></div> 13.24%	<div></div> 7.70%	<div></div> 6.30%	<div></div> 8.10%	<div></div> 9.10%	<div></div> 6.60%	<div></div> 7.50%	10.40%		
ビジネス	<div></div> 0.20%	<div></div> 0.10%	<div></div> 0.00%	<div></div> 1.70%	<div></div> 0.10%		<div></div> 0.42%	<div></div> 0.90%	<div></div> 2.20%	<div></div> 3.20%	<div></div> 1.20%	<div></div> 1.50%	<div></div> 1.80%	1.11%		
その他	<div></div> 0.10%	<div></div> 0.10%	<div></div> 0.50%	<div></div> 0.80%	<div></div> 0.40%		<div></div> 0.38%	<div></div> 0.80%	<div></div> 0.70%	<div></div> 1.10%	<div></div> 0.40%	<div></div> 0.90%	<div></div> 0.70%	0.58%		
na	58.10%	57.80%	63.50%	59.30%	63.90%		60.52%	61.70%	61.60%	60.20%	61.10%	61.00%	61.12%	60.82%		
レジャー	<div></div> 48.60%	<div></div> 56.10%	<div></div> 60.90%	<div></div> 32.60%	<div></div> 61.80%		<div></div> 52.00%	<div></div> 33.00%	<div></div> 27.90%	<div></div> 29.00%	<div></div> 33.40%	<div></div> 30.60%	<div></div> 30.78%	41.39%		
ビジネス	<div></div> 7.30%	<div></div> 0.60%	<div></div> 0.50%	<div></div> 23.40%	<div></div> 0.80%		<div></div> 6.52%	<div></div> 23.80%	<div></div> 30.00%	<div></div> 25.80%	<div></div> 24.20%	<div></div> 26.80%	<div></div> 26.12%	16.32%		
その他	<div></div> 2.20%	<div></div> 1.10%	<div></div> 2.10%	<div></div> 3.30%	<div></div> 1.30%		<div></div> 2.00%	<div></div> 4.90%	<div></div> 3.60%	<div></div> 5.40%	<div></div> 3.50%	<div></div> 3.60%	<div></div> 4.20%	3.10%		
na	0.00%	0.00%	0.00%	0.00%	0.00%		0.00%	0.00%	0.10%	0.00%	0.00%	0.00%	0.02%	0.01%		
総計	100.00%	100.00%	100.00%	100.00%	100.00%		100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%		

参考 — データ集計の例

投稿者の傾向 (性別,エリア別)

個数 / コメント	列ラベル													
	A_レジャー					B_ビジネス					B_ビジネス 集計		総計	
行ラベル	01_登別	02_草津	03_箱根	04_道後	05_湯布院	A_レジャー 集計	06_札幌	07_名古屋	08_東京	09_大阪	10_福岡	B_ビジネス 集計		
一人	<div><div></div></div> 26.10%	<div><div></div></div> 11.80%	<div><div></div></div> 11.10%	<div><div></div></div> 55.50%	<div><div></div></div> 11.10%	<div><div></div></div> 23.12%	<div><div></div></div> 68.10%	<div><div></div></div> 74.50%	<div><div></div></div> 69.90%	<div><div></div></div> 66.20%	<div><div></div></div> 68.20%	<div><div></div></div> 69.38%		
家族	<div><div></div></div> 59.80%	<div><div></div></div> 66.70%	<div><div></div></div> 65.80%	<div><div></div></div> 30.30%	<div><div></div></div> 64.80%	<div><div></div></div> 57.48%	<div><div></div></div> 23.30%	<div><div></div></div> 14.70%	<div><div></div></div> 17.70%	<div><div></div></div> 20.70%	<div><div></div></div> 19.50%	<div><div></div></div> 19.18%		
恋人	<div><div></div></div> 4.30%	<div><div></div></div> 10.90%	<div><div></div></div> 11.20%	<div><div></div></div> 4.50%	<div><div></div></div> 12.00%	<div><div></div></div> 8.58%	<div><div></div></div> 2.60%	<div><div></div></div> 3.60%	<div><div></div></div> 3.90%	<div><div></div></div> 4.20%	<div><div></div></div> 3.80%	<div><div></div></div> 3.62%		
友達	<div><div></div></div> 6.20%	<div><div></div></div> 8.70%	<div><div></div></div> 9.40%	<div><div></div></div> 4.10%	<div><div></div></div> 9.10%	<div><div></div></div> 7.50%	<div><div></div></div> 2.50%	<div><div></div></div> 3.00%	<div><div></div></div> 4.50%	<div><div></div></div> 6.00%	<div><div></div></div> 3.40%	<div><div></div></div> 3.88%		
仕事仲間	<div><div></div></div> 2.80%	<div><div></div></div> 1.30%	<div><div></div></div> 1.50%	<div><div></div></div> 4.10%	<div><div></div></div> 1.20%	<div><div></div></div> 2.26%	<div><div></div></div> 3.00%	<div><div></div></div> 3.50%	<div><div></div></div> 2.10%	<div><div></div></div> 2.20%	<div><div></div></div> 4.10%	<div><div></div></div> 2.98%		
その他	<div><div></div></div> 0.70%	<div><div></div></div> 0.60%	<div><div></div></div> 1.00%	<div><div></div></div> 0.90%	<div><div></div></div> 1.80%	<div><div></div></div> 1.00%	<div><div></div></div> 0.50%	<div><div></div></div> 0.70%	<div><div></div></div> 1.90%	<div><div></div></div> 0.70%	<div><div></div></div> 1.00%	<div><div></div></div> 0.96%		
na	<div><div></div></div> 0.10%	<div><div></div></div> 0.00%	<div><div></div></div> 0.00%	<div><div></div></div> 0.20%	<div><div></div></div> 0.00%	<div><div></div></div> 0.06%	<div><div></div></div> 0.00%	<div><div></div></div> 0.00%	<div><div></div></div> 0.00%	<div><div></div></div> 0.00%	<div><div></div></div> 0.00%	<div><div></div></div> 0.00%		
総計	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	

数値評価の構成 (エリア別)

個数 / コメント	列ラベル													総計
	A_レジャー					B_ビジネス								
行ラベル	01_登別	02_草津	03_箱根	04_道後	05_湯布院	A_レジャー 集計	06_札幌	07_名古屋	08_東京	09_大阪	10_福岡	B_ビジネス 集計		
5	<div></div> 34.70%	<div></div> 42.80%	<div></div> 41.20%	<div></div> 34.50%	<div></div> 66.00%	<div></div> 43.84%	<div></div> 37.60%	<div></div> 29.50%	<div></div> 38.10%	<div></div> 34.70%	<div></div> 31.00%	<div></div> 34.18%	39.01%	
4	<div></div> 41.60%	<div></div> 40.60%	<div></div> 38.40%	<div></div> 46.40%	<div></div> 23.60%	<div></div> 38.13%	<div></div> 41.70%	<div></div> 48.00%	<div></div> 43.40%	<div></div> 45.70%	<div></div> 46.90%	<div></div> 45.30%	41.71%	
3	<div></div> 15.20%	<div></div> 9.60%	<div></div> 11.00%	<div></div> 13.00%	<div></div> 4.60%	<div></div> 10.68%	<div></div> 15.20%	<div></div> 16.10%	<div></div> 12.80%	<div></div> 14.30%	<div></div> 15.00%	<div></div> 14.68%	12.68%	
2	<div></div> 5.60%	<div></div> 4.40%	<div></div> 5.10%	<div></div> 4.30%	<div></div> 2.80%	<div></div> 4.44%	<div></div> 4.00%	<div></div> 3.40%	<div></div> 3.70%	<div></div> 2.70%	<div></div> 4.20%	<div></div> 3.60%	4.02%	
1	<div></div> 2.90%	<div></div> 2.60%	<div></div> 4.30%	<div></div> 1.80%	<div></div> 3.00%	<div></div> 2.92%	<div></div> 1.50%	<div></div> 2.10%	<div></div> 2.00%	<div></div> 2.60%	<div></div> 3.00%	<div></div> 2.24%	2.58%	
総計	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	

参考ーデータ集計の例

数値評価の平均 (エリア別)

行ラベル	平均 / サービス	平均 / 立地	平均 / 部屋	平均 / 設備・アメニ	平均 / 風呂	平均 / 食事	平均 / 総合
☐A_レジャー	4.08	4.16	3.97	3.89	4.16	4.16	4.16
01_登別	3.83	4.12	3.71	3.69	4.16	3.97	4.00
02_草津	4.11	4.23	3.90	3.79	4.22	4.12	4.17
03_箱根	4.09	4.01	4.00	3.88	4.10	4.16	4.07
04_道後	3.93	4.21	3.89	3.82	3.92	3.98	4.08
05_湯布院	4.44	4.21	4.37	4.26	4.40	4.51	4.47
☐B_ビジネス	3.91	4.25	3.92	3.79	3.66	3.88	4.06
06_札幌	3.98	4.22	3.96	3.86	3.78	3.91	4.10
07_名古屋	3.88	4.17	3.88	3.75	3.59	3.85	4.00
08_東京	3.94	4.38	3.99	3.87	3.68	3.91	4.12
09_大阪	3.89	4.25	3.94	3.77	3.71	3.93	4.07
10_福岡	3.86	4.21	3.86	3.71	3.57	3.83	3.99

数値評価の平均 (レジャー,ビジネス別)



行ラベル	平均 / サービス	平均 / 立地	平均 / 部屋	平均 / 設備・アメニ	平均 / 風呂	平均 / 食事	平均 / 総合
A_レジャー	4.08	4.16	3.97	3.89	4.16	4.16	4.16
B_ビジネス	3.91	4.25	3.92	3.79	3.66	3.88	4.06

議論&発表

- データ集計によって発見した, データセットに関する特徴や傾向について発表してください
- 時間配分
 - グループごとに議論 (10分)
 - グループごとに発表 (3分 x 8グループ)

日経トレンディ 2018年5月号

「都市観光ホテルを創造する、『星野流』の狙いと勝算」

- 長野県 浅間温泉「界 松本」
 - 温泉街の集客低下 → 浅間温泉の観光客は松本市内に宿泊
- 全国の都市部にあるビジネスホテルを調査
 - 宿泊客の6割はビジネス客でなく「観光客」
 - 一方で、料金に不満はないものの旅のテンションが下がる
- 都市型ホテルがどうか変われるか → 都市観光ホテル

関連研究

- 辻井康一 and 津田和彦「テキストマイニングを用いた宿泊レビューからの注目情報抽出方法」, デジタルプラクティス 3.4 (2012): 289-296.

数値評価の平均 (レジャー, ビジネス別)

行ラベル	平均 / サービス	平均 / 立地	平均 / 部屋	平均 / 設備・アメニティ	平均 / 風呂	平均 / 食事	平均 / 総合
A_レジャー	4.08	4.16	3.97	3.89	4.16	4.16	4.16
B_ビジネス	3.91	4.25	3.92	3.79	3.66	3.88	4.06

- 数値評価のみから違いを見つけるのは難しい!!
 - ユーザーの 8割が 4~5 の評価, 1~2をつけない
 - ユーザーは 注目の有無に関係なくすべての項目に回答
- レジャーとビジネスでは, 評価すべき項目も異なることを確認した
- テキストと対応付ければ, 同じ点数でも差異があることを確認した

参考書

(KH Coder)

- [1] 樋口耕一. 社会調査のための計量テキスト分析－内容分析の継承と発展を目指して－. ナカニシヤ出版, 京都, 2014.
- [2] 樋口耕一. テキスト型データの計量的分析－2つのアプローチの峻別と統合－. 理論と方法, 数理社会学会, 2004, 19(1): 101-115.

(Windows環境によるCGM収集の参考に)

- [3] テキストマイニングソフトを利用した新未来洞察手法の研究. 第10分科会, (財)市場創造研究会. http://www.shijo-sozo.org/news/%E7%AC%AC10%E5%88%86%E7%A7%91%E4%BC%9A_1.pdf

参考書

(R を使った参考書)

- [4] 金明哲. "テキストデータの統計科学入門." 岩波書店, 2009.
- [5] 石田基広. "RMeCab によるテキスト解析. R によるテキストマイニング入門." 森北出版, 2008, 51-82.

(他のツールを使った参考書)

- [6] 那須川哲哉. "テキストマイニングを使う技術/作る技術: 基礎技術と適用事例から導く本質と活用法." 東京電機大学出版局, 2006.
- [7] 上田隆穂, 黒岩祥太, 戸谷圭子. "テキストマイニングによるマーケティング調査." 講談社, 2005.

(統計解析を中心とした参考書)

- [8] 前田忠彦; 山崎誠. 言語研究のための統計入門. くろしお出版株式会社, 東京, 2013.