

テキストマイニングの実習

— 1日目 —

2019/7/4

ビジネス科学研究科
経営システム科学専攻

講義で使用するサイト

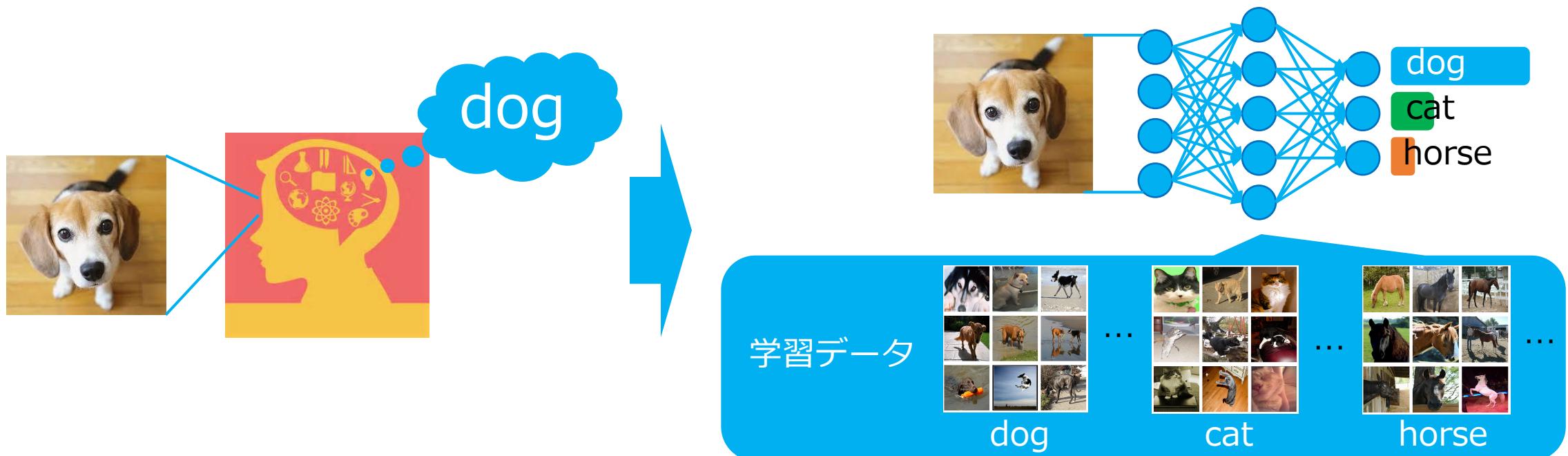
<https://github.com/haradatm/lecture/tree/master/gssm-201907>

自然言語処理のトレンド

—ディープラーニングによる自然言語処理—

ディープラーニング(深層学習)の成功

- ニューラルネット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



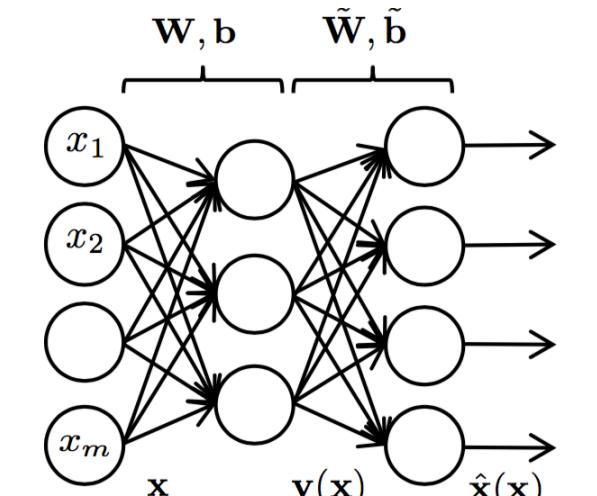
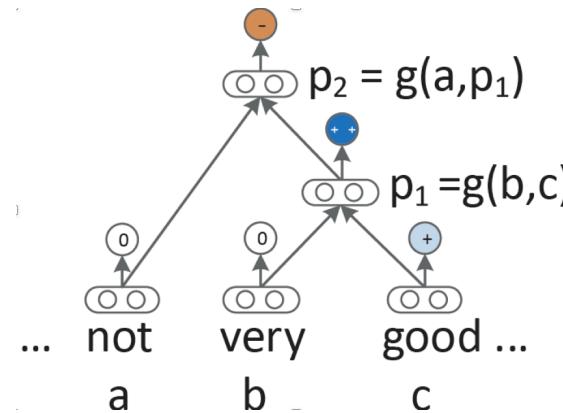
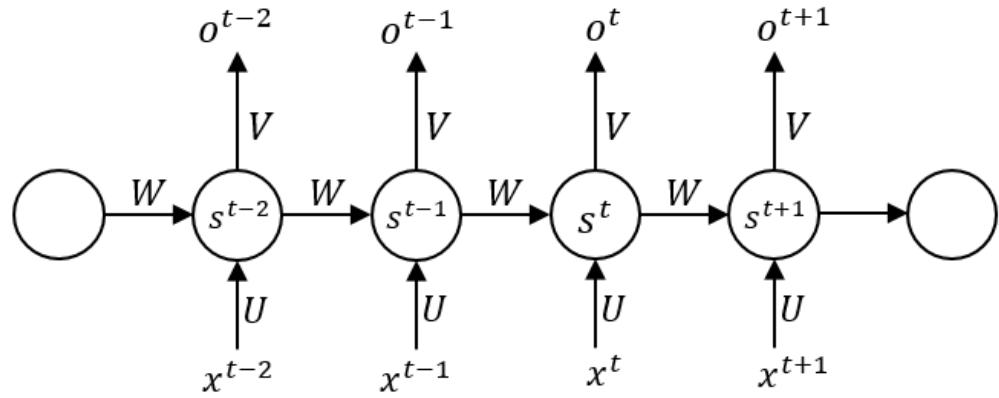
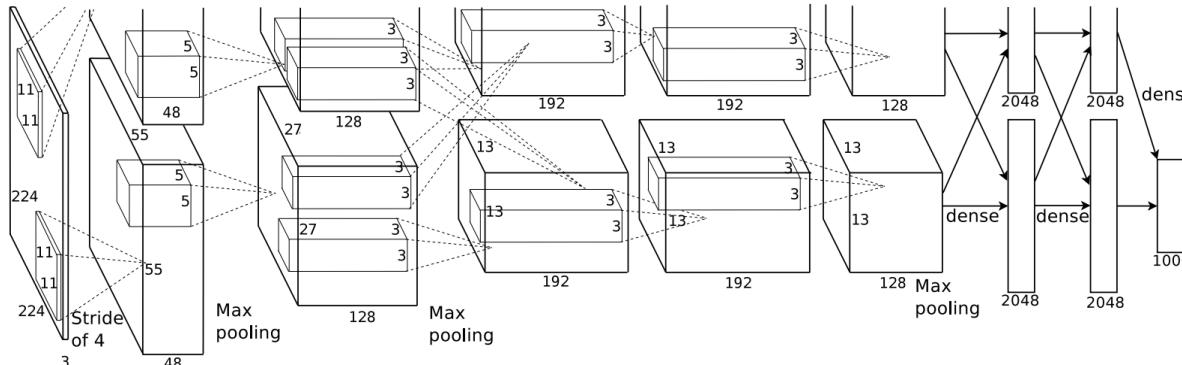
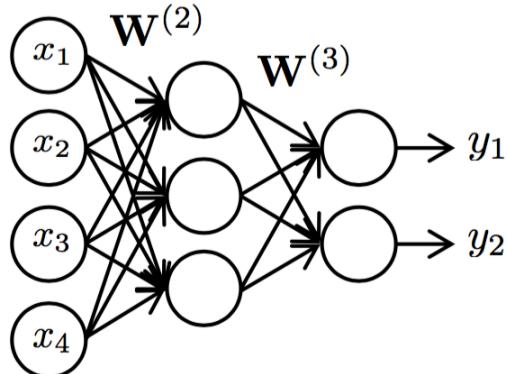
ニューラルネットの歴史

- 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]
冬	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]



様々なニューラルネット



音声認識での成功 [Seide+, 2011]

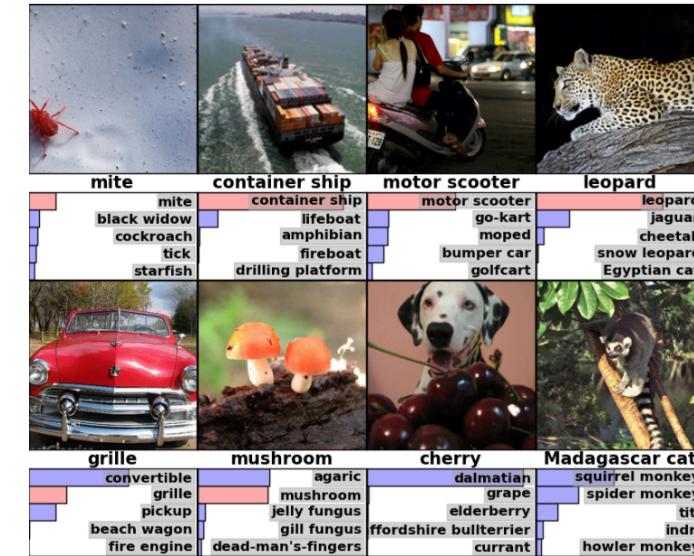
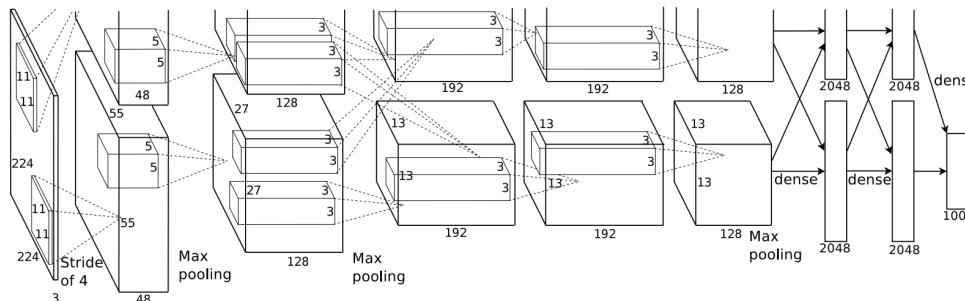
- Microsoft Research のグループ
 - 電話での会話音声の標準データセット
 - 入力(MFCC)-出力(HMM状態変数)の関係をDNNで学習
 - 従来 GMM-HMM → DNN-HMM (全結合7層, 事前学習あり)
 - 単語誤認識率で 10%前後の大幅な精度改善

acoustic model & training	recognition mode	RT03S		Hub5'00 SWB	voicemails		tele- conf
		FSH	SW		MS	LDC	
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]

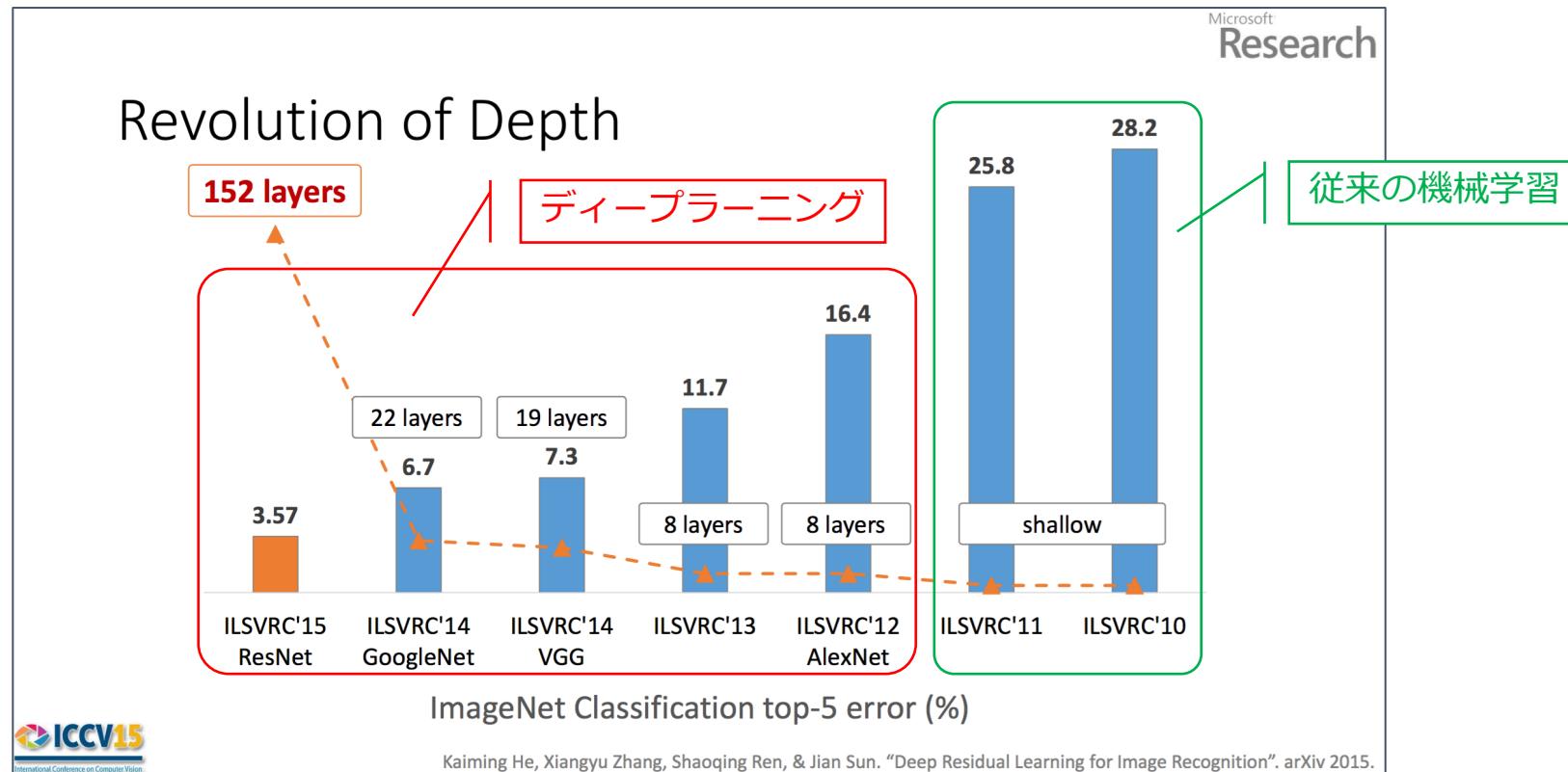
- ・一般物体認識 (Hintonのグループ)
 - ・ImageNet Large-scale Visual Recognition Challenge 2012
 - ・1000カテゴリ×約1000枚 = 100万枚 の訓練画像
 - ・畳込み層5, 全結合層3, 2つのGPUで2週間 (AlexNet)
 - ・誤識別率が10%以上減少 (過去数年間での向上は1~2%)



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

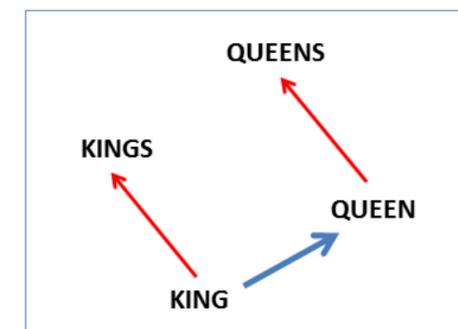
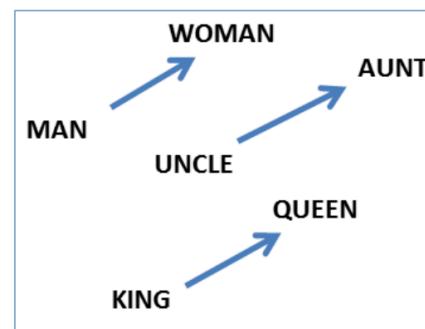
ディープラーニングによる自然言語処理

- 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
- $\text{king} - \text{man} + \text{woman} = \text{queen}$ で有名 →
※ 図上に $\text{king} + (\text{woman}-\text{man}) = \text{queen}$ を
描くとわかりやすい



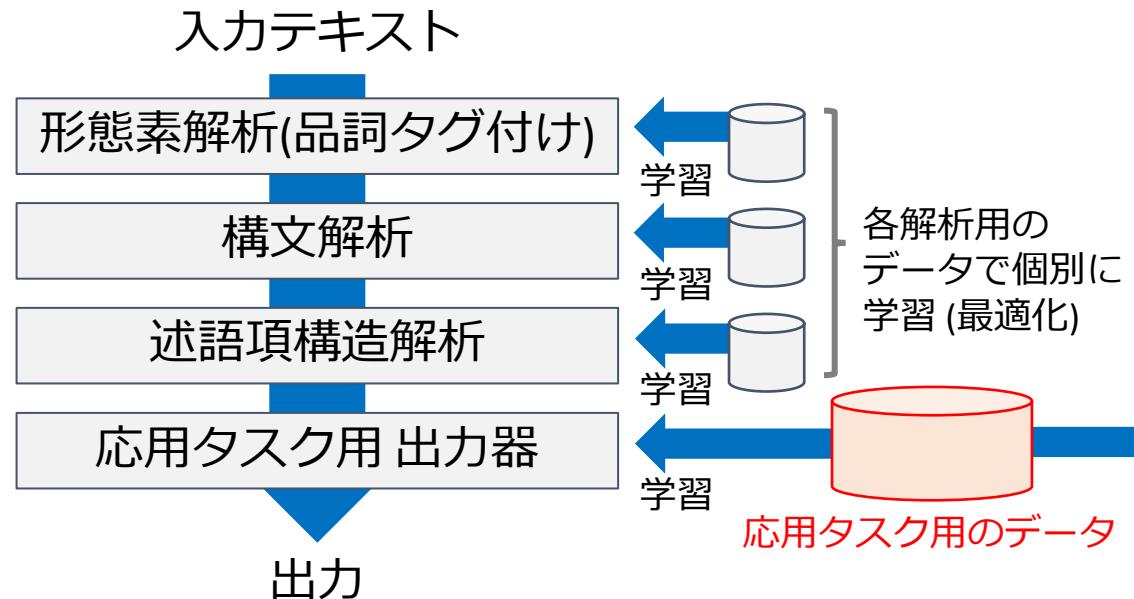
Tomas Mikolov, Wen-tau Yih, Geoffrey Zweig, 2013, NAACL

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發揮

- 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

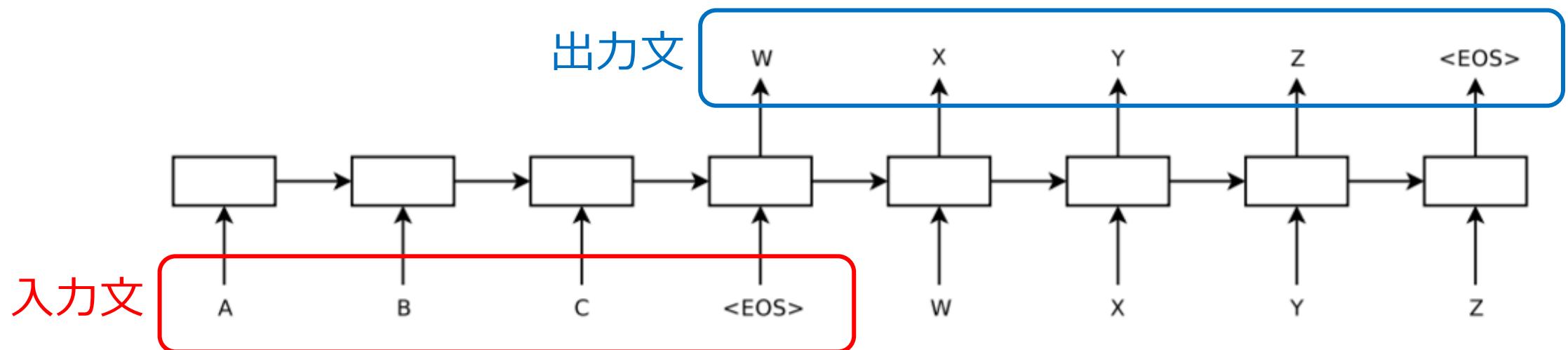
ディープラーニングによる自然言語処理

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 (RNNLM)	<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

Seq2Seq [Sutskever+, NIPS2014]

- ニューラル機械翻訳の基本となったモデル



“ABC”という単語列から“WXYZ”という単語列への翻訳

さらに,興味がある人へ

- 機械翻訳
 - 須藤 克仁. “ニューラル機械翻訳の進展 –系列 変換モデルの進化とその応用–.” 人工知能 34.4 (2019): 437-445.
- 文書要約
 - 西川 仁. “深層学習による自動要約.” 人工知能 34.4 (2019): 446-450.

などが詳しいです.

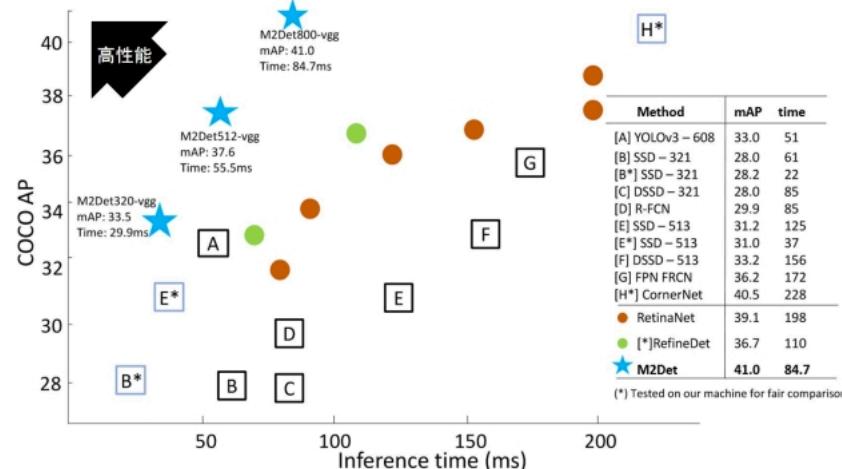


人工知能学会誌, 最新号

2018 → 2019 (5月時点)

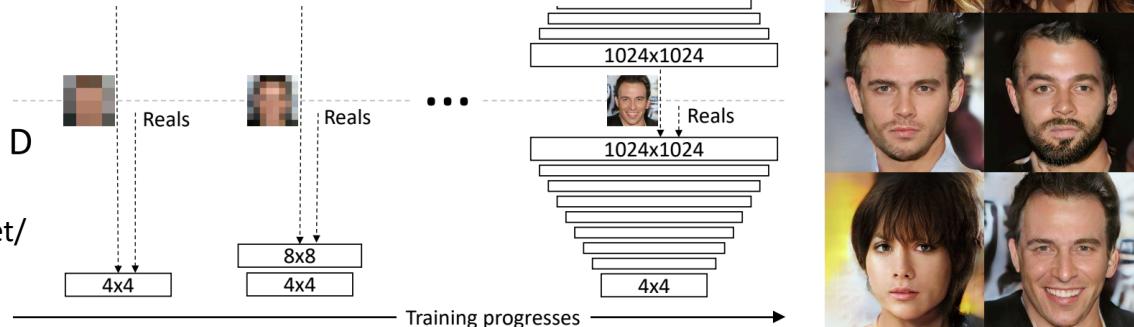
M2Det (物体検出) 精度と速度でSOTA

<https://qiita.com/kzykmyzw/items/1831f70dcade04db2210>



PG-GANs (顔画像生成) 1024×1024 のサイズの高画像

<https://openreview.net/pdf?id=Hk99zCeAb>



BERT (自然言語処理)

様々なタスクでSOTAを更新

タスク	概要	前SOTA	BERT
GLUE	8種の言語理解タスク	75.2	81.9
1. MNLI	2入力文の含意/矛盾/中立を判定	82.1	86.7
2. QQP	2質問文が意味的に等価か判定	70.3	72.1
3. QNLI	SQuADの改変。陳述文が質問文の解答を含むか判定	88.1	91.1
4. SST-2	映画レビューの入力文のネガポジを判定	91.3	94.9
5. CoLA	入力文が言語的に正しいか判定	45.4	60.5
6. STS-B	ニュース見出しの2入力文の意味的類似性をスコア付け	80.0	86.5
7. MRPC	ニュース記事の2入力文の意味的等価性を判定	82.3	89.3
8. RTE	2入力文の含意を判定	56.0	70.1
SQuAD	質疑応答タスク。陳述文から質問文の解答を抽出	91.7	93.2
CoNLL	固有表現抽出タスク。単語に人物/組織/位置のタグ付け	92.6	92.8
SWAG	入力文に後続する文を4つの候補文から選択	59.2	86.3

2018年10月: BERT の衝撃

Jacob Devlin, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." (2018)

- タスク特化のNN構造を持たずに入間のスコアを大きく超えた

SQuAD1.1 Leaderboard

Since the release of SQuAD1.0, the community has made rapid progress, with the best models now rivaling human performance on the task. Here are the ExactMatch (EM) and F1 scores evaluated on the test set of SQuAD v1.1.

Rank	Model	EM	F1
	Human Performance Stanford University (Rajpurkar et al. '16)	82.304	91.221
1	BERT (ensemble) Google AI Language https://arxiv.org/abs/1810.04805	87.433	93.160
2	BERT (single model) Google AI Language https://arxiv.org/abs/1810.04805	85.083	91.835
2	nlnet (ensemble) Microsoft Research Asia	85.356	91.202

<https://rajpurkar.github.io/SQuAD-explorer/>

- 特徴

- 双向Transformer言語モデルを大規模コーパスで事前学習し、出力層をタスク毎に1層のみ追加してfine-tuning
- マスク単語予測と次文章判定で事前学習

- 評価

- 11タスクでSOTA (含意、言い換え、文の分類など)
- 機械読解タスク(左)でも、完全一致と部分一致の両指標で最高精度
(2018/10/5)

2019年5月: XLNet に注目

Zhilin Yang, et al. "XLNet: Generalized Autoregressive Pretraining for Language Understanding". (2019)

- BERT の弱点を修正し, 20以上のタスクで BERT を超えた

SQuAD1.1 Leaderboard

Here are the ExactMatch (EM) and F1 scores evaluated on the test set of SQuAD v1.1.

Rank	Model	EM	F1
	Human Performance Stanford University (Rajpurkar et al. '16)	82.304	91.221
1 May 21, 2019	XLNet (single model) XLNet Team	89.898 XLNet	95.080
2 Oct 05, 2018	BERT (ensemble) Google AI Language https://arxiv.org/abs/1810.04805	87.433 BERT	93.160

<https://rajpurkar.github.io/SQuAD-explorer/>

- 特徴

- BERTではマスク単語を予測するが, マスクは通常発生しないためノイズにもなる問題を克服

- 評価

- 20タスクでBERTを超えた(2018/5/21)
- 機械読解タスク(左)では, single モデルで BERT の ensemble モデルを超えている

スケジュール

- 1日目: 7/4
 - 説明 – データ分析の手順
 - 演習 – データの理解 (Excel)
- 2日目: 7/11
 - 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: 7/18
 - 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, コールセンタの応対ログ
 - ・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: 1,042万件 ※年間約60~70万件増加

The screenshot shows the Rakuten Travel website at <https://travel.rakuten.co.jp/review/>. The main heading is 'お客様の声' (Customer Reviews) with the subtitle 'ホテルのクチコミ数No.1 10,402,126' (Number of hotel reviews No.1). Below this, there's a search bar for 'クチコミ (お客様の声) を検索' (Search reviews) and filters for '国内宿泊' (Domestic accommodation) and '海外ホテル' (Overseas hotels). A green banner at the bottom highlights '新着! 最新的クチコミ' (Newest reviews) from June 14, 2019, at 23:58:13, mentioning '吳ステーションホテル' (1062 reviews) with a 3.3 rating and 'ホテルエアウェイ' (1840 reviews) with a 4.17 rating. The page also features three testimonial bubbles with photos of happy guests.

経年変化:

780万件 (2015)
→ 836万件 (2016)
→ 900万件 (2017)
→ 973万件 (2018)
→ 1,042万件 (今回)

This image shows a screenshot of a travel review page from Rakuten Travel. The page is for Ryukyu Seaworld Hotel in Naha, Okinawa. It includes a navigation bar with links like 'Home', 'Travel', 'Hotels', 'Flight', 'Car', 'Tour', and 'Accommodation'. The main content area features a large photo of the hotel, its address, and a summary of guest reviews. Below this are sections for 'Check-in/Check-out Dates', 'Guest Details', and 'Search Options'. A 'Map' section shows the location of the hotel. The right side of the page contains a sidebar with various promotional offers and links.

⑧ 鴨川シーワールドホテル クラ ×

HARADA Tomohiko

travel.rakuten.co.jp/HOTEL/2910/review.html

★★お部屋★
●鴨川シーワールドホテル
★レストラン★
●鴨川シーワールドホテル
★温泉大浴場★
●鴨川シーワールドホテル
★館内施設★
●鴨川シーワールドホテル
★よくあるご質問★
●鴨川シーワールドホテル
★アクセス★
●鴨川シーワールドホテル
設備・アメニティ・基本情報
●鴨川シーワールドホテル
写真・画像
●鴨川シーワールドホテル
地図・アクセス
●鴨川シーワールドホテル
クチコミ
●鴨川シーワールドホテル
温泉

★★★★★ 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（品数が多く、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。

レビューを評価して不適切なレビューを報告するこのレビューは参考になりましたか？
いいえ いいえ

旅行の目的 … レジャー
同伴者 … 家族
宿泊年月 … 2015年06月

ご利用の宿泊プラン 【洋室 禁煙・特別室】 お部屋からシャチャやイルカも見える シーワールドと海一望宿泊プラン

ご利用のお部屋 【wa5シーワールド】が見える特別室禁煙【洋室】

★★★★★ 4
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 07:33:49
夫、2歳半と5ヶ月の子どもの4人で宿泊しました。
【立地】当たり前ですが鴨川シーワールドにとても近く、ゆっくり館内を見学できました。
【部屋】至って普通です。（古いからか、歯の声は少し聞こえます。）トイレ掃除などはしっかりされていました。清浄機などもTEL一本ですぐに届けて下さいました。
【食事】夜朝共にバイキング。イスですが子ども用イス、エプロン、ベビーベッドを用意して下さっています。キッズスペースも食事時間中に専門のスタッフの方がおりゆっくり食事ができました。
【風呂】小さな子ども（赤ちゃん）用のグッズ（ペビーベッド、コーナー、バス、おもちゃ、泡ソープ、支えのあるイス）が揃っていました。お子さん連れも多く気兼ねなく楽しめました。しかしそ風呂がひどくつかないので、温泉を楽しむという雰囲気ではなく、銭湯のお湯が温泉という感じです。
また、23時頃にお風呂に行くと、アメニティやシャンプーが空だったのは少し残念でした。
【サービス】受付スタッフの方さんとともに親切、丁寧です。チェックアウト後に子どもの薬を冷蔵庫にいておいて欲しいとダメ元で頼むと快く入

★★★★★ 2
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、大変申し訳ございませんでした。
重要改善として、早急に対応いたします。
今後は、この様な事の無いように、清掃・点検を強化いたします。

フロントスタッフへのお言葉、誠にありがとうございました。
モチベーションアップに繋がりますので、お客様からの声として、
スタッフと共に共有させて頂きます。

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。

★★★★★ 4
鴨川シーワールドホテル 2015年06月11日 19:25:48
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。

詳細にご感想頂きまして、ありがとうございました。
今後の参考にさせて頂きます。
また、スタッフ対応に関しまして、お褒めのお言葉を頂戴しまして、
とても嬉しく思います。
モチベーションアップに繋がりますので、お客様からの声として、
スタッフと共に共有させて頂きます。

最後に、「アメニティ・シャンプー」の件、
大変申し訳ございませんでした。
早急に対応をして、改善を行います。
貴重なご意見を、ありがとうございます。

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。

いい値！バリュープラン
【最安料金（目安）】 10,186円～
(消費税込11,000円～)
【当日15:50からアシカと記念写真】笑うアシカと一緒にパリピップラン
【最安料金（目安）】 10,278円～
(消費税込11,100円～)
【当日13:40～エコ・アクロームコミュニケーションタイム】1日3組限定
【最安料金（目安）】 10,278円～
(消費税込11,100円～)
【夜の水族館探検付】3月～10月の火・木曜日限定プラン
【最安料金（目安）】 10,278円～
(消費税込11,100円～)
【当日14:50からイルカと一緒にナリティ2室限定】鴨川シーワールド体験付プラン
【最安料金（目安）】 10,463円～
(消費税込11,300円～)
今しかない！★アワビ料理付＆シーワールド入園バスポート付で大満足♪5月～6月の月～木曜日限定プラン
【最安料金（目安）】 10,926円～
(消費税込11,800円～)
【便利な赤ちゃんギッズ付】初！お母さんはお母さんも嬉しい赤ちゃんなつ得プラン
【最安料金（目安）】 10,926円～
(消費税込11,800円～)
お子様にも大好評！オーシャンビュープラン
【最安料金（目安）】 11,112円～
(消費税込12,000円～)
【80cmのジャンボボササイズ】海の上のシャチぬいぐるみ付プラン
【最安料金（目安）】 11,204円～
(消費税込12,100円～)
房総2大テーマパーク満喫「マザーアウトドア」付プラン
【最安料金（目安）】 11,389円～
(消費税込12,300円～)
【当日14:50～イルカ

鴨川シーワールドホテルのクチコミ・お客さまの声

[●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

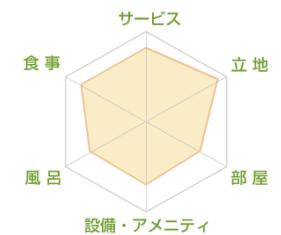
評価内訳

- 5点 ■■■■■
- 4点 ■■■■
- 3点 ■■■
- 2点 ■■
- 1点 ■

236件
302件
47件
15件
9件

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



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）



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（品数が多く、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、それ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ます。

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

情報



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す。

客室内清掃の件、大変申し訳

重要改善として、早急に対応いたします。

今後は、この様な事の無いように、清掃・点検を強化いたします。

テキストデータ

フロントスタッフへのお言葉

誠にありがとうございます。

セラベーションアップに繋がるお客様からの声として、スタッフと共有させて頂きます。

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2018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

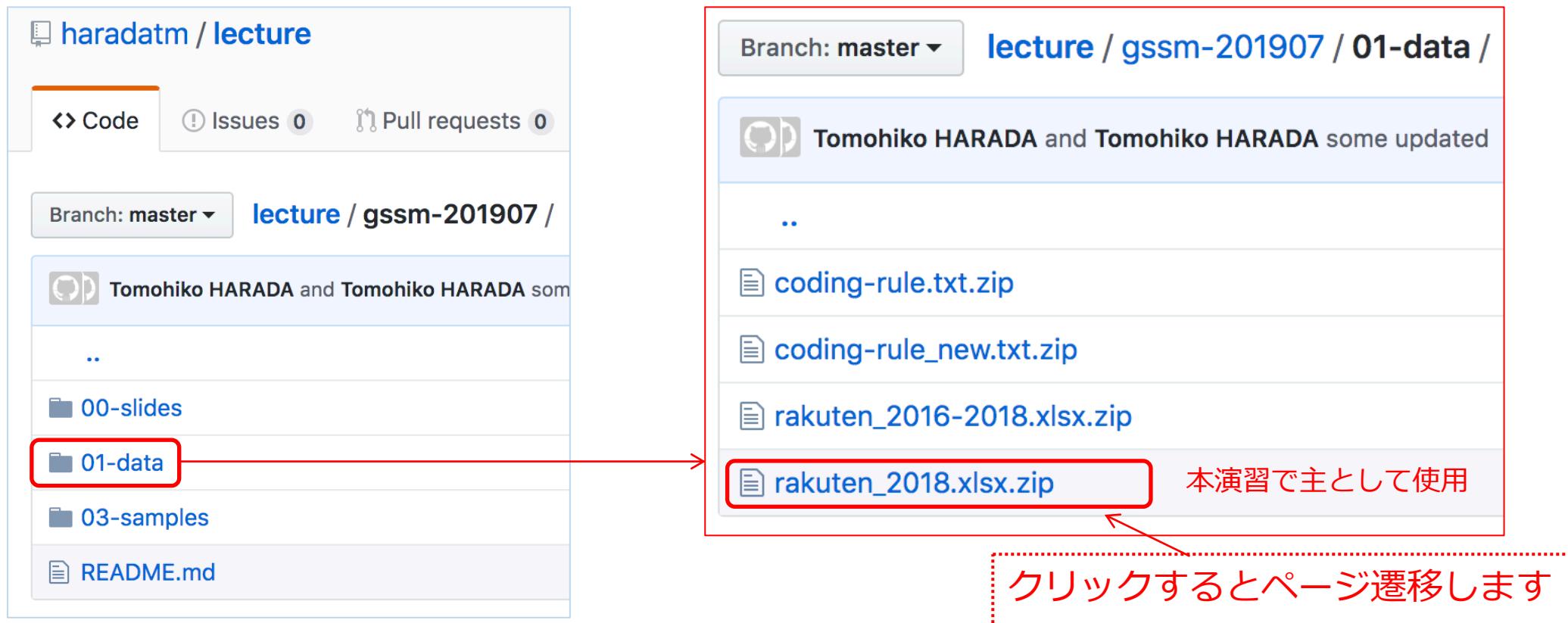
レジャー	5エリア	登別, 草津, 箱根, 道後, 湯布院	1,000件×10エリア = 計10,000件
ビジネス	5エリア	札幌, 名古屋, 東京, 大阪, 福岡	

- ・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

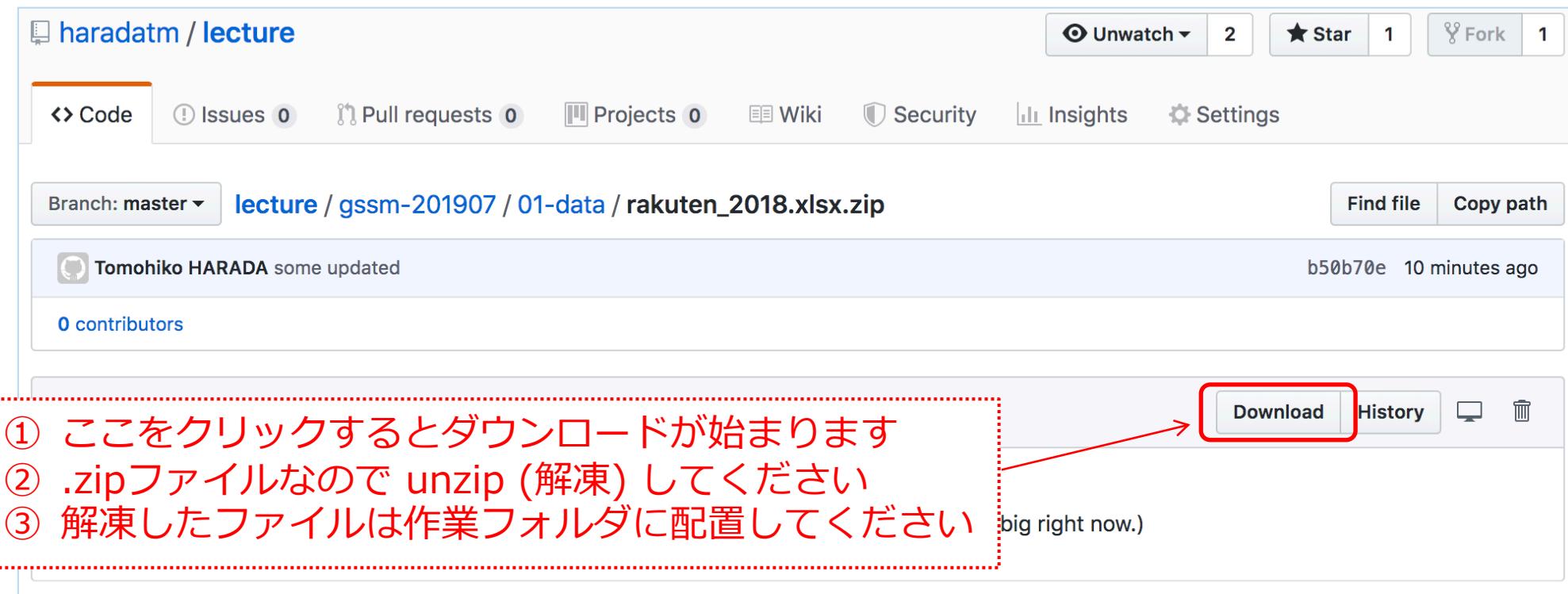
データの取得方法

- <https://github.com/haradatm/lecture/tree/master/gssm-201907>



ダウンロード方法

- Download ボタンをクリックするとダウンロードを開始

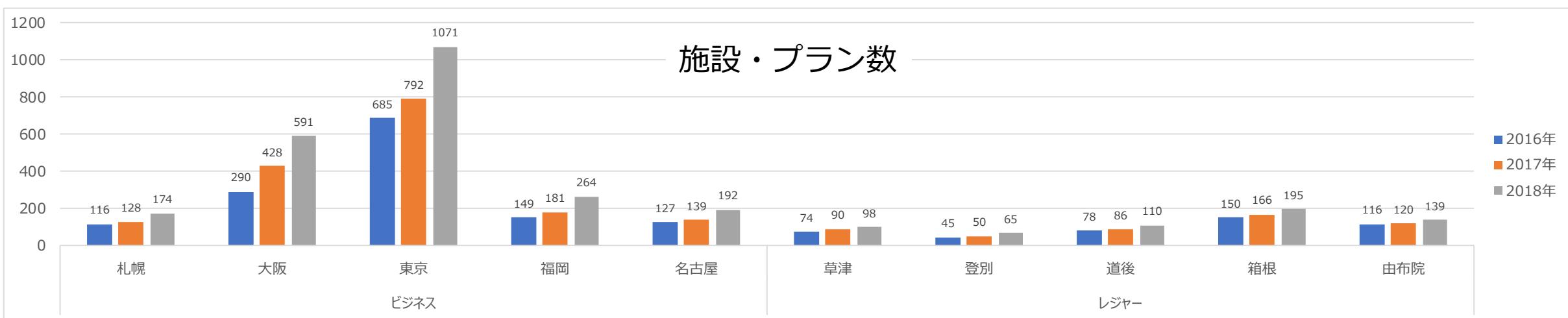
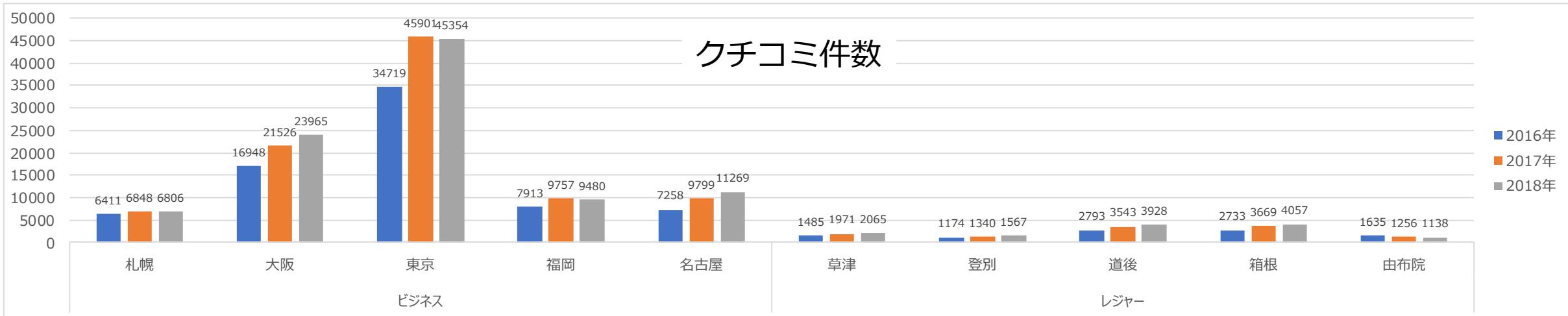


使用するデータ

データファイル名	件数	データセット
rakuten_2018.xlsx	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2018年のみ (1シート)
rakuten_2016-2018.xlsx	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2016, 2017, 2018年 (3シート)

参考－収集データについて

※演習ではこの一部(エリアごとに
1,000件をサンプリング)を使用



参考—サンプリングについて

NO.	カテゴリ	エリア	収集した全クチコミ件数と1000件サンプルでのカバー率						収集した全施設・プラン数と1000件サンプルでのカバー率					
			2016年 全件	カバー率	2017年 全件	カバー率	2018年 全件	カバー率	2016年 全件	カバー率	2017年 全件	カバー率	2018年 全件	カバー率
1	レジャー	登別	1,174	85.18%	1,340	74.63%	1,567	63.82%	45	100.00%	50	96.00%	65	96.92%
2		草津	1,485	67.34%	1,971	50.74%	2,065	48.43%	74	95.95%	90	94.44%	98	89.80%
3		箱根	2,733	36.59%	3,669	27.26%	4,057	24.65%	150	90.00%	166	87.35%	195	80.51%
4		道後	2,793	35.80%	3,543	28.22%	3,928	25.46%	78	91.03%	86	89.53%	110	86.36%
5		由布院	1,635	61.16%	1,256	79.62%	1,138	87.87%	116	93.97%	120	96.67%	139	97.84%
6	ビジネス	札幌	6,411	15.60%	6,848	14.60%	6,806	14.69%	116	93.10%	128	87.50%	174	86.78%
7		名古屋	7,258	13.78%	9,799	10.21%	11,269	8.87%	127	85.83%	139	88.49%	192	84.38%
8		東京	34,719	2.88%	45,901	2.18%	45,354	2.20%	685	56.93%	792	52.78%	1,071	46.03%
9		大阪	16,948	5.90%	21,526	4.65%	23,965	4.17%	290	71.72%	428	59.58%	591	53.64%
10		福岡	7,913	12.64%	9,757	10.25%	9,480	10.55%	149	87.25%	181	82.32%	264	76.14%
全体			83,069	12.04%	105,610	9.47%	109,629	9.12%	1,830	75.19%	2,180	70.09%	2,899	64.26%

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_2018.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 – 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)

行ラベル	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%	0.01%	0.00%	0.03%
20代	0.88%	1.02%	0.00%	1.90%
30代	2.89%	2.34%	0.00%	5.23%
40代	8.46%	3.78%	0.00%	12.24%
50代	10.82%	3.26%	0.00%	14.08%
60代	4.62%	1.11%	0.00%	5.73%
70代	0.67%	0.10%	0.00%	0.77%
80代	0.09%	0.02%	0.00%	0.11%
na	0.00%	0.00%	59.91%	59.91%
総計	28.45%	11.64%	59.91%	100.00%

投稿者の傾向 (エリア別)

行ラベル	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	25.74%	31.16%	28.45%
女性	14.20%	9.10%	11.65%
na	60.06%	59.74%	59.90%
総計	100.00%	100.00%	100.00%

- 無回答(na)の層がある年代や性別に偏っている場合もある
- 性別に偏りがある場合は、コメントにバイアスもある

参考 – データ集計の例

投稿者の傾向 (性別, 目的-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	06_札幌	07_名古屋	08_東京	09_大阪	10_福岡	
男性	30.40%	25.80%	18.60%	30.50%	23.40%	25.74%	34.50%	31.90%	28.90%	28.20%	32.30%	31.16%	28.45%			
女性	13.80%	15.70%	17.30%	8.80%	15.40%	14.20%	9.80%	7.30%	10.60%	9.90%	7.90%	9.10%	11.65%			
na	55.80%	58.50%	64.10%	60.70%	61.20%	60.06%	55.70%	60.80%	60.50%	61.90%	59.80%	59.74%	59.90%			
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%	100.00%	100.00%	100.00%

投稿者の傾向 (年代別, 目的-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	06_札幌	07_名古屋	08_東京	09_大阪	10_福岡	
10代	0.00%	0.10%	0.00%	0.10%	0.00%	0.04%	0.00%	0.10%	0.00%	0.00%	0.00%	0.02%	0.03%			
20代	1.10%	3.70%	3.70%	1.00%	2.60%	2.42%	1.10%	1.00%	1.90%	1.60%	1.30%	1.38%	1.90%			
30代	6.80%	6.80%	5.40%	4.20%	6.00%	5.84%	4.70%	3.80%	4.90%	5.30%	4.40%	4.62%	5.23%			
40代	13.30%	12.20%	8.70%	11.60%	10.10%	11.18%	15.70%	14.40%	12.20%	11.90%	12.30%	13.30%	12.24%			
50代	13.60%	12.70%	10.20%	15.90%	11.50%	12.78%	16.20%	14.60%	15.90%	14.60%	15.60%	15.38%	14.08%			
60代	7.50%	5.20%	6.70%	5.90%	7.60%	6.58%	5.90%	4.70%	4.30%	3.90%	5.60%	4.88%	5.73%			
70代	1.70%	0.70%	0.90%	0.60%	0.80%	0.94%	0.70%	0.60%	0.20%	0.70%	0.80%	0.60%	0.77%			
80代	0.20%	0.10%	0.30%	0.00%	0.10%	0.14%	0.00%	0.00%	0.10%	0.10%	0.20%	0.08%	0.11%			
na	55.80%	58.50%	64.10%	60.70%	61.20%	60.06%	55.70%	60.80%	60.50%	61.90%	59.80%	59.74%	59.90%			
110代	0.00%	0.00%	0.00%	0.00%	0.10%	0.02%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	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%	100.00%	100.00%	100.00%

参考 – データ集計の例

投稿者の傾向 (同行者別,目的-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						
一人	26.90%	14.60%	9.80%	56.30%	12.70%	24.05%	69.70%	76.10%	73.70%	66.00%	70.30%	71.16%			47.61%	
家族	58.50%	60.40%	64.80%	32.30%	65.50%	56.30%	20.40%	15.70%	16.50%	21.80%	18.20%	18.52%			37.41%	
恋人	7.00%	13.40%	13.80%	3.30%	9.50%	9.40%	2.30%	2.10%	3.30%	3.20%	3.30%	2.84%			6.12%	
友達	4.90%	9.00%	9.00%	4.30%	9.30%	7.30%	4.20%	2.90%	4.30%	5.10%	4.20%	4.14%			5.72%	
仕事仲間	2.20%	0.90%	1.80%	3.60%	1.30%	1.96%	2.70%	2.70%	1.80%	2.70%	3.10%	2.60%			2.28%	
その他	0.50%	1.70%	0.80%	0.20%	1.70%	0.98%	0.70%	0.50%	0.40%	1.20%	0.90%	0.74%			0.86%	
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 (評価値別, 目的-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						
5	35.20%	48.20%	47.00%	42.60%	66.60%	47.92%	39.00%	34.40%	31.90%	37.10%	32.10%	34.90%			41.41%	
4	42.20%	37.20%	35.90%	39.50%	23.70%	35.70%	40.90%	44.90%	46.60%	45.20%	44.90%	44.50%			40.10%	
3	13.60%	8.40%	8.90%	12.50%	5.60%	9.80%	13.40%	12.70%	14.00%	10.80%	16.20%	13.42%			11.61%	
2	5.20%	3.70%	4.90%	3.40%	2.50%	3.94%	4.20%	4.90%	5.40%	4.50%	4.60%	4.72%			4.33%	
1	3.80%	2.50%	3.30%	2.00%	1.60%	2.64%	2.50%	3.10%	2.10%	2.40%	2.20%	2.46%			2.55%	
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

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	4.15	4.21	4.06	3.96	4.23	4.22	4.22
01_登別	3.87	4.13	3.82	3.78	4.22	3.94	4.00
02_草津	4.18	4.27	4.04	3.91	4.30	4.16	4.25
03_箱根	4.18	4.10	4.05	3.97	4.16	4.27	4.18
04_道後	4.03	4.28	4.00	3.89	3.97	4.12	4.17
05_湯布院	4.50	4.27	4.38	4.28	4.46	4.60	4.51
■B_ビジネス	3.87	4.22	3.95	3.81	3.70	3.90	4.05
06_札幌	3.91	4.19	4.00	3.83	3.73	3.92	4.10
07_名古屋	3.85	4.11	3.95	3.81	3.71	3.84	4.03
08_東京	3.85	4.28	3.94	3.76	3.64	3.89	4.01
09_大阪	3.88	4.33	3.96	3.83	3.72	3.96	4.10
10_福岡	3.88	4.19	3.89	3.80	3.70	3.89	4.00

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	4.15	4.21	4.06	3.96	4.23	4.22	4.22
B_ビジネス	3.87	4.22	3.95	3.81	3.70	3.90	4.05

討論&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
 - ・グループごとで討論 (10分)
 - ・グループごとに発表 (3分 × 6グループ)

日経トレンド 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」

- ・温泉街の集客低下 → 浅間温泉の観光客は松本市内に宿泊
 - ・長野県 浅間温泉「界 松本」
- ・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

日経トレンド 2018年7月号 「コックroach脱皮缶」

- ・パッケージ描かれたイラストが嫌 → 前年比2倍の出荷



http://www.kincho.co.jp/seihin/insecticide/go_aerosol/gokiburi_u_spray/index.html

まとめ – 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	4.15	4.21	4.06	3.96	4.23	4.22	4.22
B_ビジネス	3.87	4.22	3.95	3.81	3.70	3.90	4.05

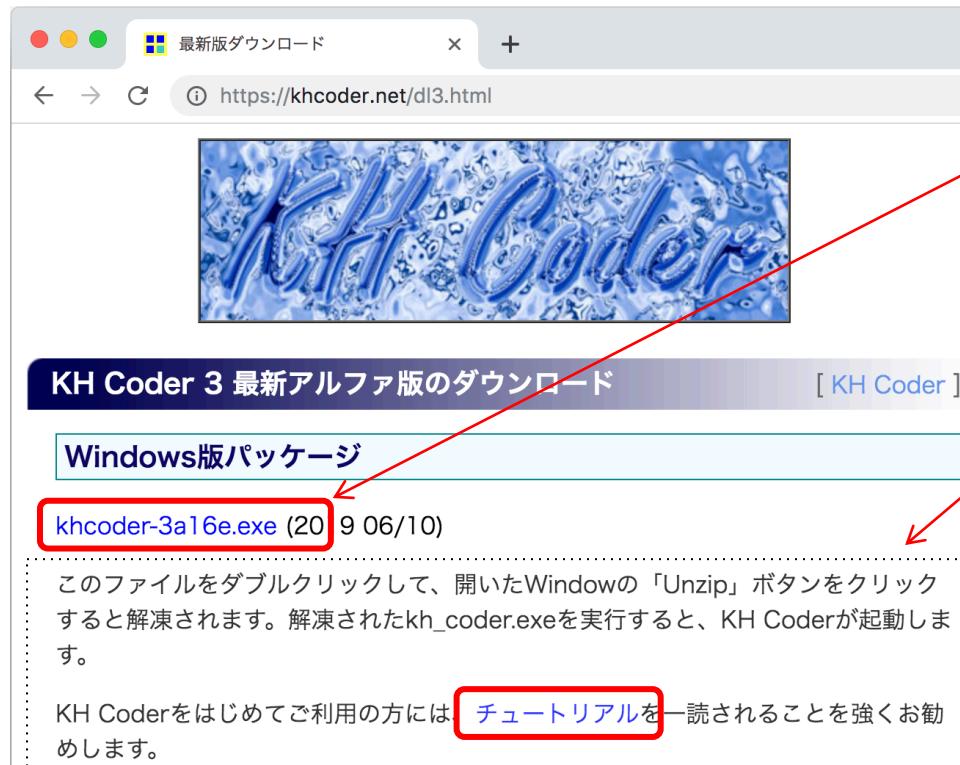
- 数値評価のみから違いを見つけるのは難しい!!

- 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

KH Coder のインストール (次回の演習で使用)

- ・ダウンロードとインストール <https://khcoder.net/dl3.html>

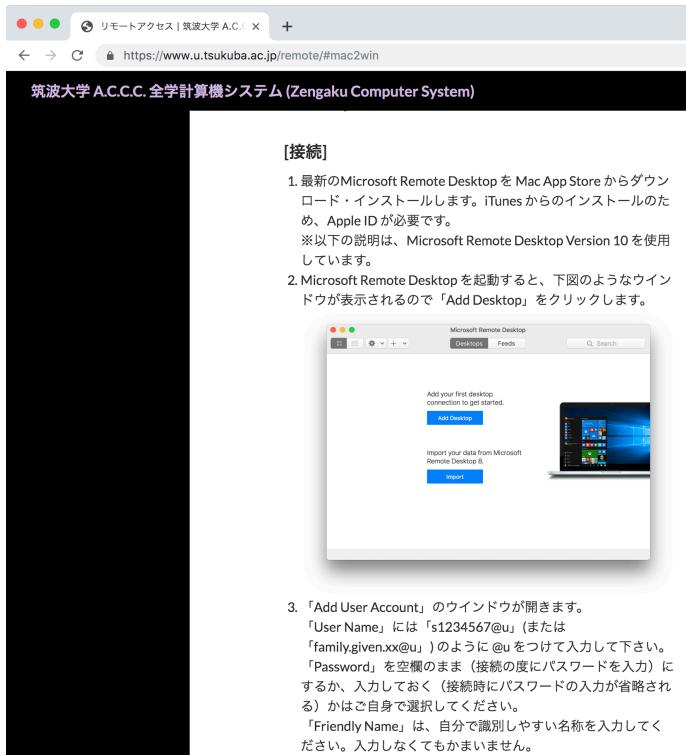


- ① ここをクリックすると遷移先のページからダウンロードが始まります
- ② 指示に従いインストール

自己解凍ファイルです。このファイルを実行（ダブルクリック）し、開いたWindowの「Unzip」ボタンをクリックすると、（特に変更しなければ）「C:\khcoder3」というフォルダにすべてのファイルが解凍されます。解凍された **kh_coder.exe** を実行すると、KH Coderが起動します。

参考 -Window OSがない場合 (同時実行未検証)

- ・全学計算機システムに Windows デスクトップがあります
 - ・<https://www.u.tsukuba.ac.jp/remote/#mac2win>



左記のページにある説明に従って、

- ① ツール (Microsoft Remote Desktop) のインストール
 - ② 全学計算機システム (Windows)へのログイン
- ができることを確認してください

KH Coder インストール時の注意:

全学の Windows の場合は、ログイン後の**デスクトップ上に「khcoder」というフォルダを作成**して、その中に解凍してください

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, 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.
- New [3] 牛澤賢二. やってみよう テキストマイニング－自由回答アンケートの分析に挑戦!. 朝倉書店, 2019**

(Windows環境によるCGM収集の参考に)

- [4] 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)

- [5] 金明哲. "テキストデータの統計科学入門." 岩波書店, 2009.
- [6] 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)

- [7] 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.
- [8] 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)

- [9] 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.