論文発表　2018年5月

トピックモデルを用いた商品の評判要因分析に関する検討

B8EM1016　富田優(とみたゆう)

諸注意

* 本スライドは論文内容の紹介のために作られました
* なるべく分かりやすくするため、正確性を欠いた点があるかもしれませんがご了承ください
* 紹介する論文と本スライドの表記が異なる場合、原論文の方が正しく、このスライドのほうが間違っています
* 発表内容に関する誤りの責任は全て発表者にあります

アウトライン

* ①はじめに
* ②研究背景
* ③論文の概要
* ④数学の用意
* ⑤ユニグラムモデル
* ⑥トピックモデル
* ⑦論文の結果
* ⑧まとめ

はじめに

* 発表者：富田優
* 所属：経済学研究科1年
* 指導教員：石垣先生
* POSデータやレビューデータの分析をといったマーケティング・リサーチの分野に進む予定
* 興味ある分野：統計学、ベイズ統計、機械学習

○今日発表する論文

「月岡晋吾、吉川　大弘、古橋武：
トピックモデルを用いた商品の評判要因分析に関する検討」

研究背景

* ネットの普及でユーザーレビューが増加
  EX：amazon,楽天,価格.com 等々

* その解析には多くのニーズとメリットが存在
  EX：レコメンデーション、市場の発見(STP)、顧客満足度指数

* 自社製品に関する評判を商品開発へ反映させられる
  EX：コンジョイント分析、パスモデル

商品の評判分析

項目選択方式

○メリット

* モニターの負担が軽い

○デメリット

* 事前に項目を決める必要
* サンプル数が必要
* 人的労力と金銭的費用が大

自由記述方式

○メリット

* 事前に想定できなかった評判要因を知られる

○デメリット

* モニターの負担が重い
* 解析に労力が必要
* 多変量解析などの統計的解析手法が使いにくい

商品の評判分析②

ECサイト上のユーザーレビュー　　楽天トラベルのサイト

○メリット

* 容易に多くのデータを収集可能
* 統計処理しやすい評点情報
* 自由記述であるレビュー情報

論文の概要

○データ

楽天トラベルのサイト上の施設レビュー

○ユーザーレビューから評判要因を特徴量として抽出

→ここに**トピックモデル**を使用

○重回帰分析

全単語に占めるあるクラスに入る単語の割合を説明変数に、平均評点データを被説明変数に重回帰分析

○どの評判要因が平均評点を上げているのか分析

実験データ

1. 楽天トラベルの施設レビュー
2. ビジネス目的かつ一人で宿泊したものに限定
3. 東京・大阪・名古屋の中心地域にあり、レビュー数が100以上のホテルが対象
4. 30971件のレビューが対象
5. 東京46軒、名古屋37軒、大阪46軒
6. 総単語数154,477
7. 総語彙数34,302→品詞は名詞のみ

数学の用意

○全部を説明するわけにはいかないので以下は既知として話します(スライドに載せません)

① 確率の公理
② 同時確率
③ 周辺化
④ ベイズの定理
⑤ 事象の独立
⑥ 連続確率変数
⑦ 確率密度関数(二項分布、多項分布、ベータ分布)

数学の用意

以下は経済学部でやることは稀だと思うので説明します

* ①カテゴリ分布
* ②ディリクレ分布

カテゴリ分布　ディリクレ分布

○カテゴリ分布

複数の離散値から1つの値をとる確率分布

EX.コンビニの一番くじを1回引く、サイコロを1回振る

$$\{1,2,3,\cdots,V\}\ P(V) = \Phi_v$$

○ディリクレ分布

トピックモデルの事前分布に使われる

$$Dirichlet(\Phi|\beta) = \frac{\Gamma(\sum_{v=1}^{V}\beta_v)}{\prod_{v=1}^{V}\Gamma(\beta_v)}\prod_{v=1}^{V}{\Phi_v}^{\beta_v-1}$$

ディリクレ分布

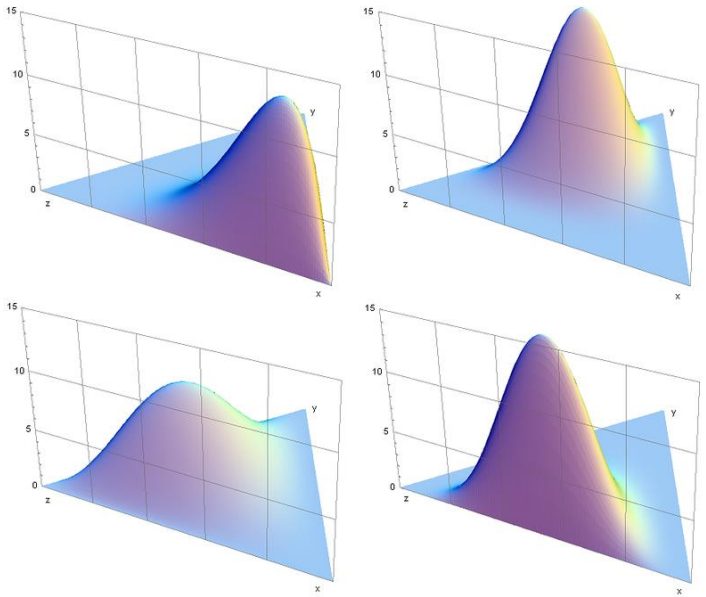
○ディリクレ分布

トピックモデルの事前分布に使われる

$$Dirichlet(\Phi|\beta) = \frac{\Gamma(\sum_{v=1}^{V} \beta_v)}{\prod_{v=1}^{V} \Gamma(\beta_v)} \prod_{v=1}^{V} {\Phi_v}^{\beta_v - 1}$$

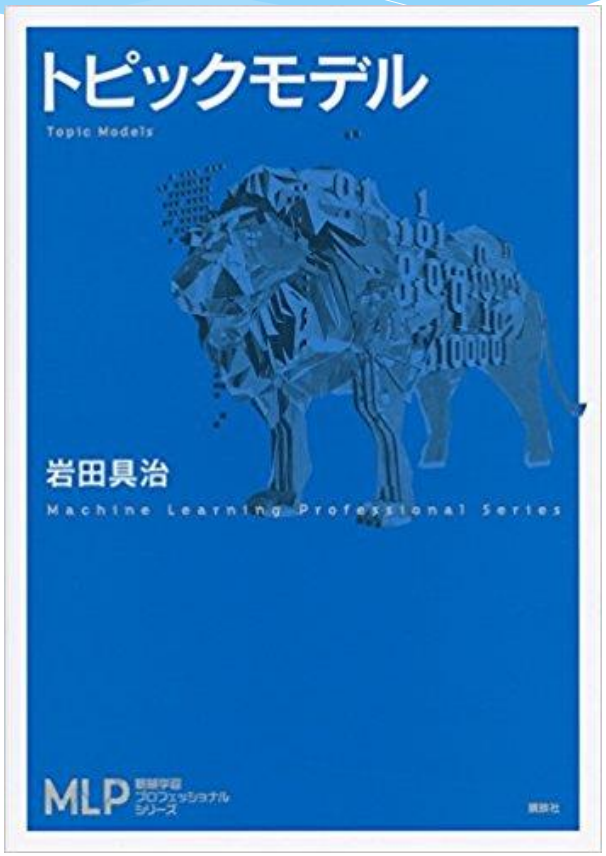
上を図にすると・・・

ディリクレ分布(V=3)の、様々なパラメータベクトルβにおける確率密度関数。左上から時計回りにβ=(6, 2, 2), (3, 7, 5), (6, 2, 6), (2, 3, 4)



トピックモデルに入る前に…

* ユニグラムモデルをやります
* ユニグラムモデルを拡張したものがトピックモデル
* 以降の説明は右の本を引用しています
* 「岩田具治:トピックモデル、講談社2015」



トピックモデルに入る前に…

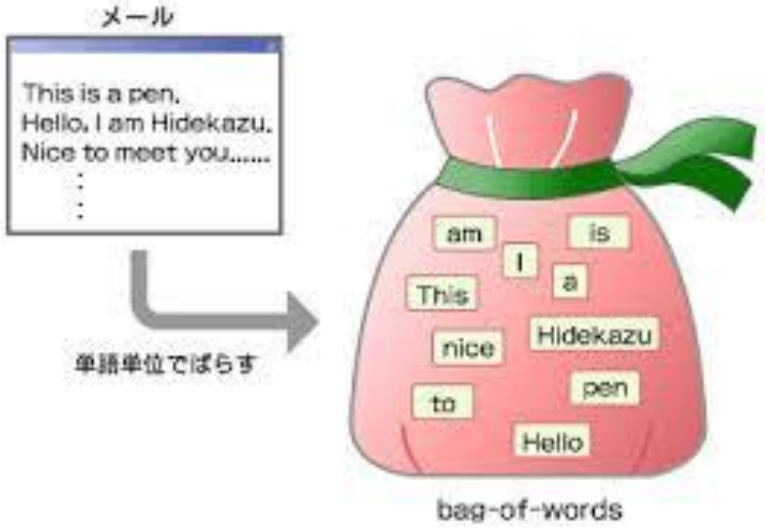
○文書を単語の多重集合(BAG)で表す

○多重集合

→重複OKな単語の集合

→BOW(Bag Of Words)表現と呼ぶ

実際には形態素解析を使って文章を単語に分割



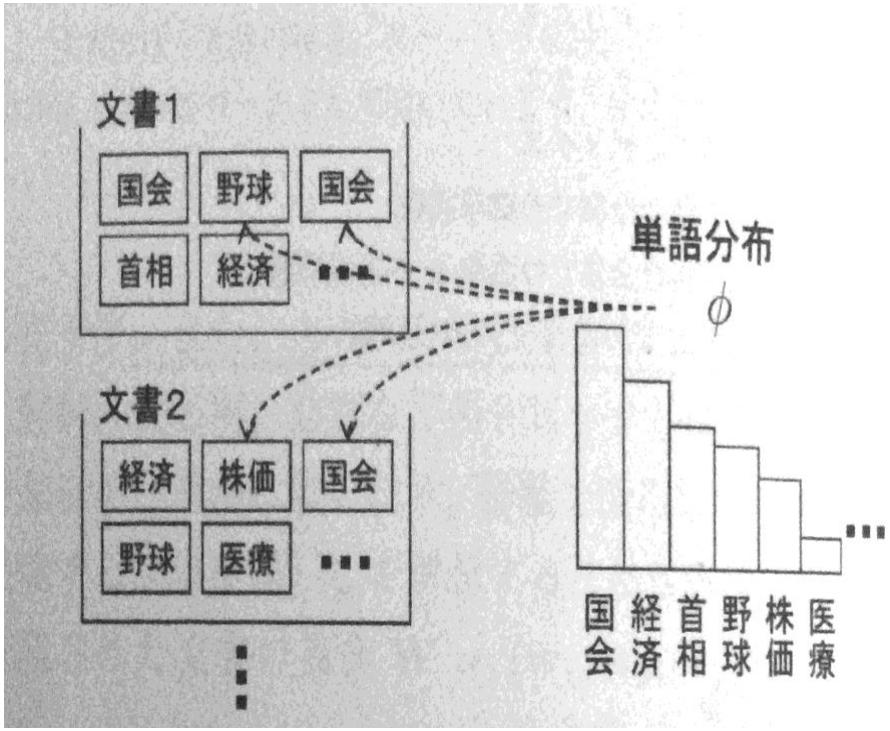
ユニグラムモデル

まずはイメージから

ある単語分布Φが存在し、そこからデータが生成されていると表現するモデル

生成モデルと呼ばれる

BOWで表現された文書集合を生成する
もっとも簡単なモデルがユニグラムモデル



ユニグラムモデル

パラメータΦが与えられたときの文書集合Wの確率は以下の通り

$$p(\boldsymbol{w}|\boldsymbol{\varphi}) = \prod_{d=1}^{D} p(w_d|\boldsymbol{\varphi})$$
$$= \prod_{d=1}^{D} \prod_{n=1}^{N_d} (w_{dn}|\boldsymbol{\varphi})$$
$$= \prod_{d=1}^{D} \prod_{n=1}^{N_d} \varphi_{w_{dn}}$$
$$= \prod_{v=1}^{V} \varphi_v^{\ N_v}$$

あとはこのΦ_vをデータから推定する

W:文書集合
Φ:φ_vのベクトル表示
w_d:文書dの単語集合
Φ_v:単語vが出現する確率
N_d;文書dに含まれる単語数
w_{dn}:文書dのn番目の単語
$\Phi_{w_{dn}}$:文書dのn番目の単語が出る確率

ユニグラムモデル

分かりづらいので言葉にすると、
(ある文書集合が生成される確率)
＝(n番目の単語の出現確率)^(出現回数)

〇モデルの前提
すべての文書の単語は同一の分布から生成される
→文書ごとに異なるトピックを持ってるんじゃないか
→一つの文書は複数のトピックを持ってるんじゃないか
→以上の2点を考慮してトピックモデルに拡張

ユニグラムモデル

* ちなみに推定方法は・・・

①最尤推定(最尤法でパラメータを推定)

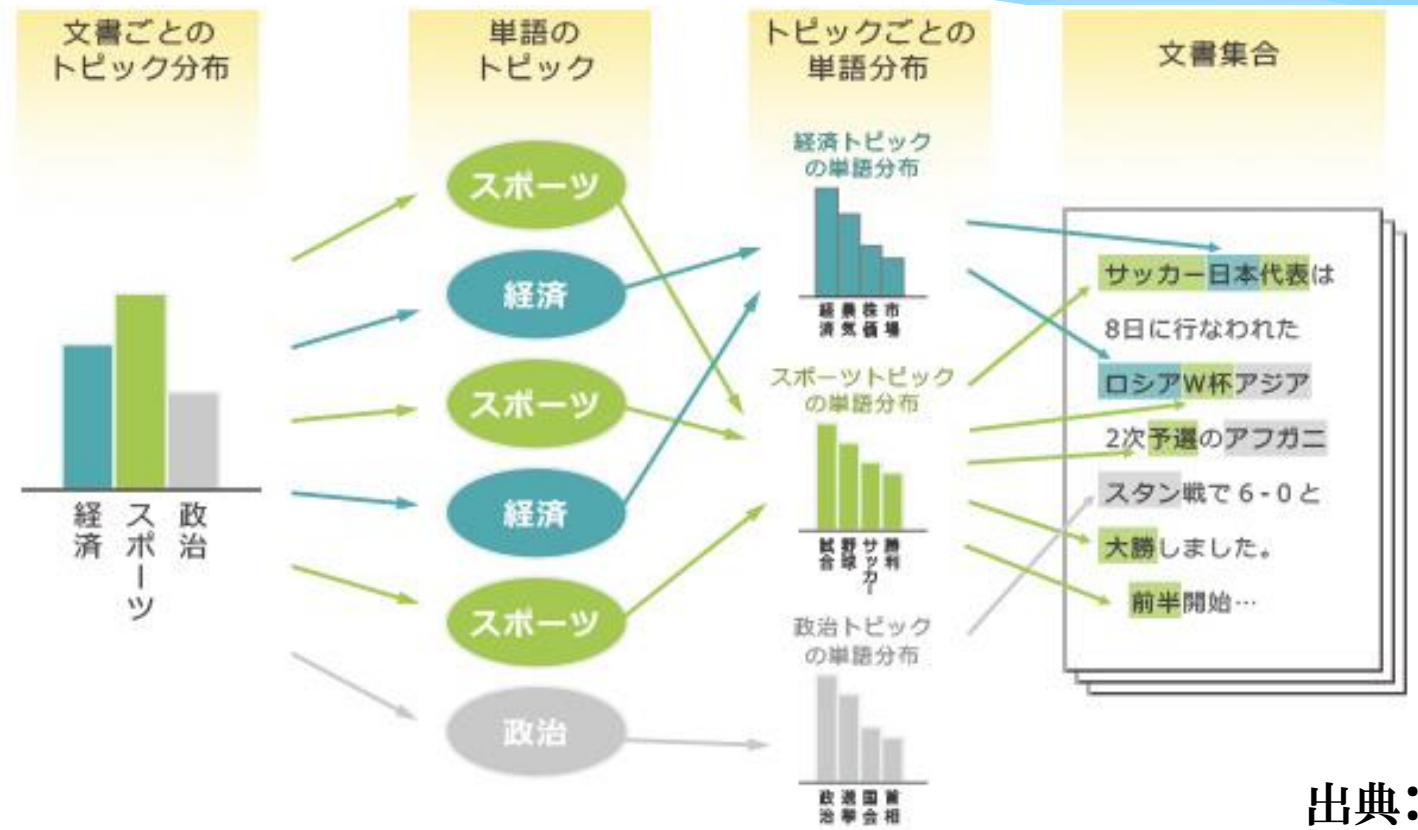
②MAP推定(事後確率が最大になるようにパラメータを推定)

③ベイズ推定(事前分布を仮定してデータを用いてパラメータの事後分布を推定)

があるが今回は割愛

トピックモデル

* またまたイメージから



出典：Albert

トピックモデル

パラメータΦが与えられたときの文書集合Wの確率は以下の通り

$$p(\boldsymbol{w}|\boldsymbol{\theta_d}, \boldsymbol{\Phi})$$

$$= \prod_{n=1}^{N_d} \prod_{k=1}^{K} p(z_{dn} = k|\theta_d)\, \boldsymbol{p(w_{dn}|\Phi_k)}$$

$$= \prod_{n=1}^{N_d} \sum_{k=1}^{K} \theta_{dk}\, \varphi_{kw_{dn}}$$

あとはこのθ_{dk}, $\varphi_{kw_{dn}}$,Kをデータから推定する

W:文書集合

Φ:φ_vのベクトル表示

w_d:文書dの単語集合

Φ_v:単語vが出現する確率

N_d;文書dに含まれる単語数

w_{dn}:文書dのn番目の単語

$\Phi_{w_{dn}}$:文書dのn番目の単語が出る確率

トピックモデル

* やっぱりわかりづらいので言葉にすると

(トピック分布と単語分布集合が与えられたとき、文書w_dが生成される確率)

＝(トピックが選択される確率)×(トピックごとの単語の出現確率)

以上を単語について掛け合わせ、トピックについてたしたもの

トピックモデル

○推定方法は
　①最尤推定
　②変分ベイズ推定
　③崩壊型ギブスサンプリング

がありますがこちらも省略（②、③に関しては私も完全に理解できていません）

○トピック数は階層ディリクレ過程を用いてデータドリブンで決められるがこちらも本日は省略

論文の結果

* レビュー毎にトピック分布を仮定
* 推定手法はギブスサンプリング
* ハイパーパラメータは

  α＝0.1（単語分布であるディリクレ分布のパラメータ）

  β＝0.1（トピック分布が従うディリクレ分布のパラメータ）

サンプリング回数1000回

論文の結果

ちなみにハイパーパラメータとは・・・

分析者が外生的に与えるモデルのパラメータのこと
これが多いことは基本的には好ましくない（できればデータから推定したい）

この論文にはハイパーパラメータの値の根拠は記載なし
トピック数に関してもなぜ10としたのか記載なし

論文の結果

各トピックの単語出現確率の上位20単語

Topic1 騒音	Topic2 再訪	Topic3 清掃	Topic4 部屋の設備	Topic5 対応	Topic6 満足性	Topic7 お風呂	Topic8 全体	Topic9 朝食	topic10 立地
部屋	いつ	部屋	部屋	対応	部屋	浴場	部屋	朝食	便利
音	今回	風呂	ベッド	フロント	宿泊	風呂	ホテル	パン	駅
隣	宿泊プラン	残念	快適	ホテル	お部屋	部屋	立地	満足	コンビニ
ホテル	予約	シャワー	ホテル	チェックイン	ホテル	満足	満足	部屋	近く
廊下	宿泊	掃除	アメニティ	スタッフ	今回	便利	対応	種類	ホテル
窓	お願い	水	テレビ	宿泊	満足	宿泊	価格	ホテル	立地
エレベーター	快適	ホテル	残念	丁寧	快適	駅	値段	無料	部屋
壁	プラン	トイレ	フロント	時間	予約	ホテル	宿泊	サービス	場所
問題	お世話	改善	仕事	お願い	ツイン	出張	朝食	コーヒー	大阪駅
外	部屋	清掃	お部屋	今回	風呂	朝食	サービス	食事	非常
立地	ホテル	今回	コンセント	チェックアウト	シングル	今回	フロント	朝	食事
残念	禁煙	お湯	加湿器	部屋	きれい	温泉	設備	メニュー	飲食店
声	出張	臭い	立地	親切	大変	いつ	非常	バイキング	快適
朝	今後	タバコ	非常	笑顔	次回	疲れ	駅	残念	出張
宿泊	ポイント	浴槽	きれい	女性	チェックイン	朝	スタッフ	サラダ	徒歩
エアコン	満足	問題	机	荷物	子供	仕事	きれい	豊富	宿泊
ドア	シングル	髪の毛	満足	残念	朝食	東京駅	出張	立地	地下鉄
空調	次回	立地	清潔	電話	対応	時間	ビジネスホテル	充実	きれい
夜	定宿	バスタブ	綺麗	サービス	アップグレード	最高	綺麗	スープ	雨
仕方	大阪	ユニットバス	風呂	仕事	綺麗	サウナ	リーズナブル	駅	大変

論文の結果

○重回帰分析

* $y_j = \beta_0 + \sum_{i=1}^{9} \beta_i x_{ij}$

y_j: ホテルの平均評点

x_{ij}:　商品ごとのトピックの割合(トピック8は除く)

β_i:各トピックの偏回帰係数

※トピックの割合すべてに回帰すると多重共線性が発生するので、今回は最も意味があいまいであるトピック8を削除

論文の結果

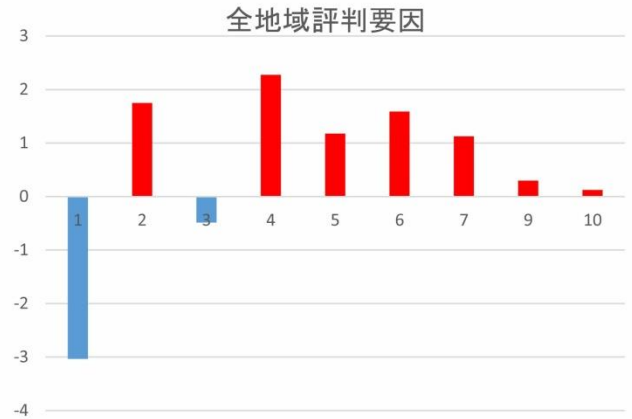
各地域の重回帰分析の結果

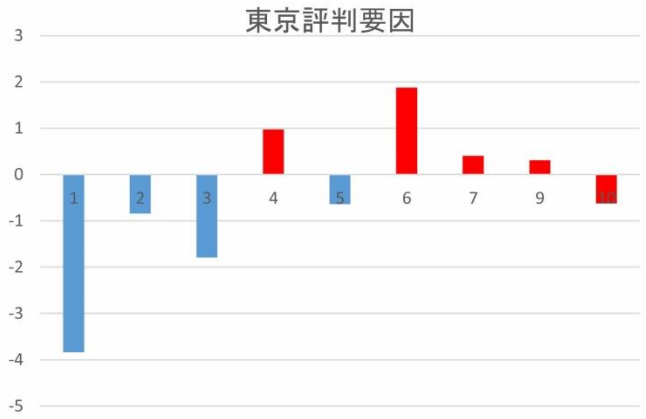
	全地域		東京		名古屋		大阪	
	偏回帰係数	p値	偏回帰係数	p値	偏回帰係数	p値	偏回帰係数	p値
切片	3.58	1.89E−09	4.50	3.44E−05	4.22	0.002	2.71	0.013
topic1	**−3.03**	**3.44E−04**	**−3.84**	**0.009**	**−4.17**	**0.024**	−1.85	0.151
topic2	1.75	0.081	−0.84	0.593	3.03	0.120	2.84	0.158
topic3	−0.49	0.528	−1.79	0.185	−1.32	0.464	1.00	0.483
topic4	**2.27**	**0.001**	0.97	0.387	0.95	0.570	**3.35**	**0.014**
topic5	1.18	0.138	−0.64	0.679	0.19	0.918	2.17	0.090
topic6	**1.59**	**0.006**	1.88	0.122	1.30	0.391	1.97	0.062
topic7	**1.12**	**0.047**	0.40	0.687	0.04	0.976	1.73	0.093
topic9	0.30	0.655	0.31	0.787	−1.02	0.472	1.88	0.292
topic10	0.12	0.861	−0.62	0.594	−0.68	0.690	0.75	0.578

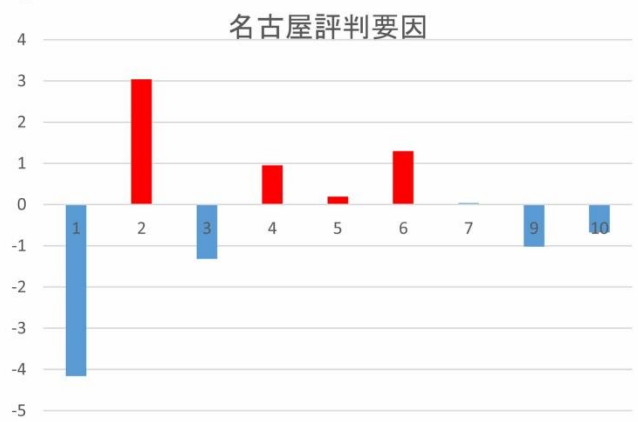
赤はP値が0.05未満

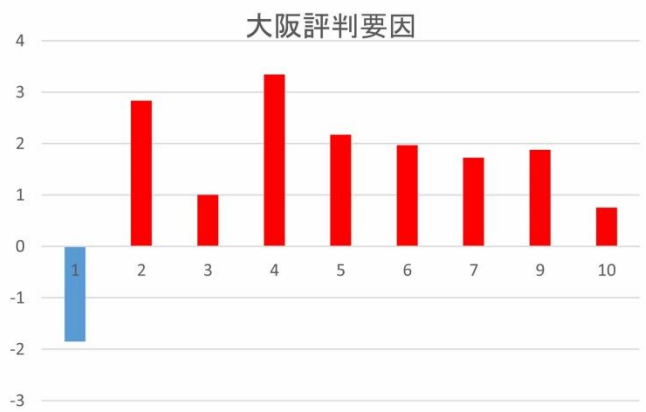
論文の結果

各地域における評判要因









論文の結果

名古屋における評判要因の可視化

topic1	topic2	topic3	topic4	topic5	topic6	topic7	topic9	topic10
部屋	いつ	部屋	部屋	対応	部屋	浴場	朝食	便利
音	今回	風呂	ベッド	フロント	宿泊	風呂	パン	駅
隣	宿泊プラン	残念	快適	ホテル	お部屋	部屋	満足	コンビニ
ホテル	予約	シャワー	ホテル	チェックイン	ホテル	満足	部屋	近く
廊下	宿泊	掃除	アメニティ	スタッフ	今回	便利	種類	ホテル
窓	お願い	水	テレビ	宿泊	満足	宿泊	ホテル	立地
エレベーター	快適	ホテル	残念	丁寧	快適	駅	無料	部屋
壁	プラン	トイレ	フロント	時間	予約	ホテル	サービス	場所
問題	お世話	改善	仕事	お願い	ツイン	出張	コーヒー	大阪駅
外	部屋	清掃	お部屋	今回	風呂	朝食	食事	非常
立地	ホテル	今回	コンセント	チェックアウト	シングル	今回	朝	食事
残念	禁煙	お湯	加湿器	部屋	きれい	温泉	メニュー	飲食店
声	出張	臭い	立地	親切	大変	いつ	バイキング	快適
朝	今後	タバコ	非常	笑顔	次回	疲れ	残念	出張
宿泊	ポイント	浴槽	きれい	女性	チェックイン	朝	サラダ	徒歩
エアコン	満足	問題	机	荷物	子供	仕事	豊富	宿泊
ドア	シングル	髪の毛	満足	残念	朝食	東京駅	立地	地下鉄
空調	次回	立地	清潔	電話	対応	時間	充実	きれい
夜	定宿	バスタブ	綺麗	サービス	アップグレード	最高	スープ	雨
仕方	大阪	ユニットバス	風呂	仕事	綺麗	サウナ	駅	大変

まとめ

〇まとめ

* トピックモデルを用いた商品の評判要因解析のための手法の提案
* 各地域ごとに正の評判要因と負の評判要因を確認
* 地域ごとの比較で特徴を検討

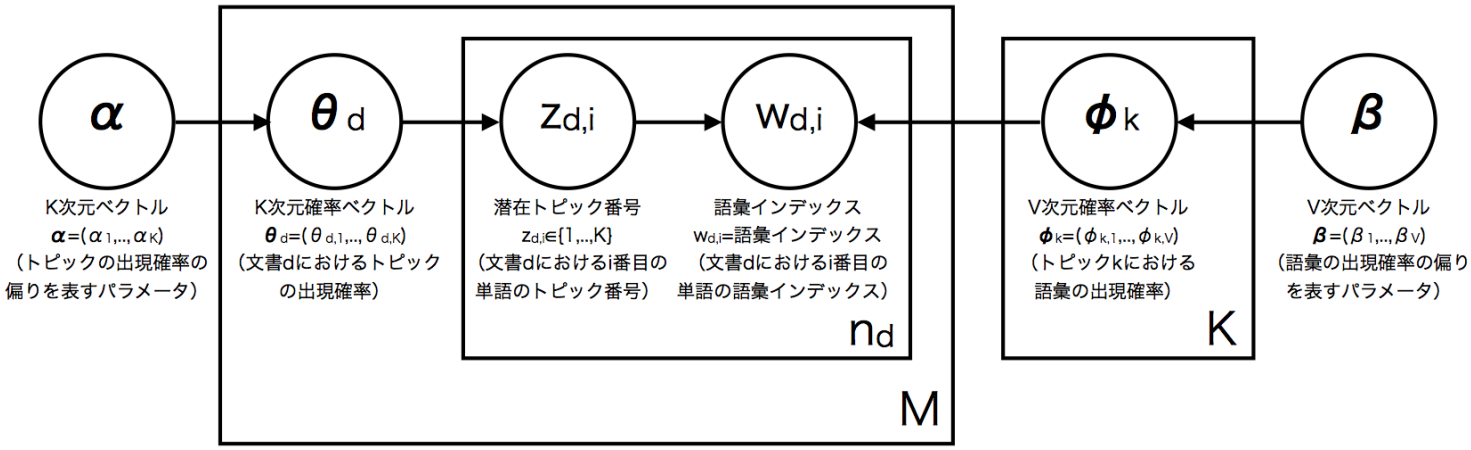
〇著者があげている今後の課題

* 異なるサービス・商品への適用
* 各レビュー評点に対しての重回帰分析

補足

* ○潜在的ディリクレ配分法（LDA）

トピック分布にディリクレ分布を仮定し、ベイズ推定する手法



M：文書数
K：トピック数
V：語彙数
n_d：文書dにおける単語数

参考文献

[1]月岡晋吾、吉川　大弘、古橋武：トピックモデルを用いた商品の評判要因分析に関する検討、講演論文集 31(0), 655-660, 2015

[2]岩田具治：トピックモデル、講談社,2015

画像取得元

* [1] https://www.amazon.co.jp
* [2] https://www.netmile.co.jp/contents/point/rakutensuperpoint/2238.html
* [3]http://www.atmarkit.co.jp/ait/articles/0803/05/news148_2.html
* [4] https://travel.rakuten.co.jp/HOTEL/67093/review.html
* [5]岩田具治：トピックモデル、講談社,2015
* [6]（http://ni66ling.hatenadiary.jp/entry/2015/05/04/163958）
* [7] https://www.albert2005.co.jp/knowledge/machine_learning/topic_model/about_topic_model