LDAによる主要ホテルチェーンの特徴抽出

B5EB1106 酒井洋輔

潜在トピックモデルの前提

潜在的意味

・・・複数の単語の共起性により創発される情報

潜在トピック

・・・潜在的意味のカテゴリ

潜在トピックモデルの前提

確率的潜在変数モデル

・・・データの背後に確率変数を仮定する。この確率変数は直接観測することができず、「潜在変数」と呼ばれる

例）データ $X = x_1, x_2, x_3 \ldots x_n$ は

潜在変数 $Z = z_1, z_2, \ldots z_k$ から生成される

潜在トピックモデルの前提

確率的生成モデル

・・・確率的潜在変数モデルの潜在変数をデータから推定するため、データの生成過程を数理モデルで表現したモデル

例）データx_iが確率分布$p\ \ (x_i|\Phi)$に従う時、

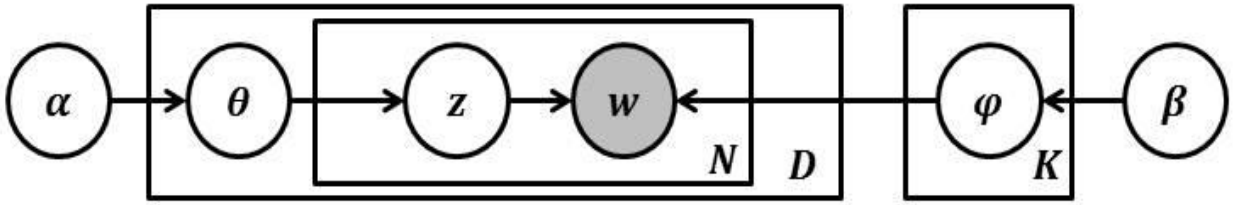
Φを確率分布のパラメータとすると

$$x_i \sim p\ \ (x_i|\Phi)$$

潜在トピックモデルの前提

グラフィカルモデル

・・・確率的生成モデルをわかりやすく図式にしたもの



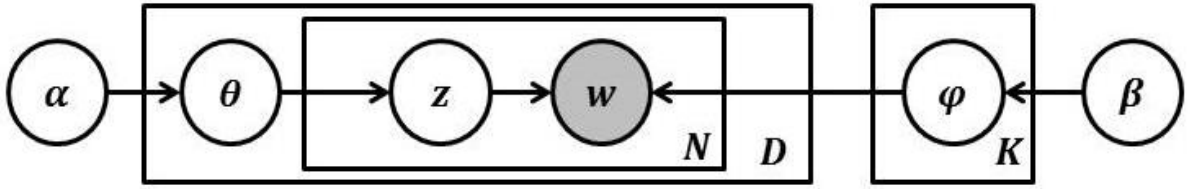
潜在トピックモデルとは

「単語同士の共起性を統計モデルとしてとらえたい！」

LDAは文書の確率的生成モデルとして提案された。

LDAを拡張した様々なモデルを総称して潜在トピックモデルという

Latent Dirichlet Allocation



トピックモデル

D:文書数
N:文書に含まれる単語数
W:単語
z:トピック
K:トピック数
α, β:ハイパーパラメーター
φ:単語分布
θ:トピック分布

$$w_{d,i} \sim Multi \quad (\Phi_{k,i})$$
$$z_{d,i} \sim Multi \quad (\theta_{k,i})$$

$$\theta_d \sim Dirichlet\ (\alpha)$$
$$\Phi_k \sim Dirichlet\ (\beta)$$

Latent Dirichlet Allocation

$\boldsymbol{\Phi_{k,i}}$

トピックKにおける単語w_iの出現確率

$$w_{d,i} \sim Multi \quad (\Phi_{k,i})$$
$$z_{d,i} \sim Multi \quad (\theta_{k,i})$$
$$\theta_d \sim Dirichlet\ (\alpha)$$
$$\Phi_k \sim Dirichlet\ (\beta)$$

$\boldsymbol{\theta_{d,k}}$

文書dにおけるトピックkの構成比率（出現確率）

LDAを使うメリット

①教師なし学習である

②単語のみの情報でその単語の極性を推定することができる

③潜在的な評価次元を抽出することができる

先行研究レビュー

"Mining marketing meaning from online chatter:

Strategic brand analysis of big data using latent dirichlet allocation,"

Tirunillai, S., & Tellis, G. J. (2014), __Journal of Marketing Research__, 51(4), 463-479.

先行研究レビュー

"Mining marketing meaning from online chatter:

Strategic brand analysis of big data using latent dirichlet allocation,"

Tirunillai, S., & Tellis, G. J. (2014), __Journal of Marketing Research__, 51(4), 463-479.

【data set】

5つのマーケット（ブランド）

パソコン（HP, Dell）、**携帯電話**(Motorola, Nokia, RIM, Palm)、**Foot Weat** (Skechers USA, TimberLand, NIKE)、**おもちゃ**(Mattel, Hasbro, LeapFrog)、**Data-Storage**(Seagate-Technology, Western-Digital-Corporation, SanDisk)

合計350,000のレビューデータ

先行研究レビュー

"Mining marketing meaning from online chatter:

Strategic brand analysis of big data using latent dirichlet allocation,"

Tirunillai, S., & Tellis, G. J. (2014), __Journal of Marketing Research__, 51(4), 463-479.

【Contribution】

①潜在的な評価次元（latent dimension）を抽出するとともにその極性を推定することができた

②あるブランドの消費者における、評価次元の異質性を数値化し、ブランド間、製品間で比較することができた

③抽出された評価次元を用いて、ブランドポジショニングを行い、可視化することができた

先行研究レビュー

①潜在的な評価次元の抽出及び極性の推定

トピックをラベリングするため、MI値を導入する。MI値を計算するためにトピックのエントロピーを計算する。この時$P(\eta = l)$のηは、あるレビューがトピックkについて言及しているeventを表しているそうですが・・・。全てのレビューがある一つのトピックでできている時、$E(k)$は最小値をとり、全てのトピックの出現確率が等しい時、最大値をとる

$$E(k) = -\sum_{\ell=0}^{1} P(\eta = \ell)\log_2 P(\eta = \ell)$$

$$E(k|w) = -\sum_{\ell=0}^{1}\sum_{w^*=0}^{1} P\left(\eta = \ell \middle| w = w^*\right)\log_2 P\left(\eta = \ell \middle| w = w^*\right)$$

先行研究レビュー

①潜在的な評価次元の抽出及び極性の推定

$$\mathrm{MI}(k|w) = \mathrm{E}(k) - \mathrm{E}(k|w) \geq 0 \ \forall (k, w)$$

MI値が高い語をあるトピックのlabeling wordとする

Table 1

DIMENSIONS OF QUALITY FOR MOTOROLA (MOBILE PHONES, QUARTER 4, 2008)

Instability (Negative)	*Portability (Positive)*	*Receptivity[a] (Positive)*	*Compatibility (Positive)*	*Discomfort[b] (Negative)*	*Secondary Features (Positive)*
Unstable	Smooth	Dependable	Universal	Cramp	Feature
Error	Handy	Reception	Expandable	Big	App
Crash	Portable	Sharp	Supported	Layout	Card
Freeze	Small	Quick	Compatible	Finger	Camera
Reboot	Compact	Crisp	Accessible	Heavy	Wi-Fi

先行研究レビュー

②あるブランドにおける、消費者の評価次元の異質性の数値化

ハーフィンダール指数を導入する

$$\alpha = \frac{\text{Total number of reviews citing the dimension}}{\text{Total number of reviews of the brand}}$$

$$H = \sum_{i=1}^{n} \alpha^2$$ →ハーフィンダール指数

値が高いほど、少数の次元に評価が集中している

先行研究レビュー

②あるブランドにおける、消費者の評価次元の異質性の数値化

HETEROGENEITY OF DIMENSIONS ACROSS BRANDS

Market, Brand	*Herfindahl Index of Concentration*	*Heterogeneity in Dimensions*	*Instability of Herfindahl Index over Time (%)*
Mobile Phones			
Nokia	45.78	Low	3.3
RIM	54.12	Low	3.5
Palm	43.58	Low	2.3
Motorola	48.18	Low	2.1
Computers			
Dell	24.80	Low	1.4
HP	31.68	Low	2.7
Toys			
Hasbro	12.82	Moderate	4.9
Mattel	11.64	High	5.4
LeapFrog	13.58	High	7.6
Footwear			
Timberland	25.74	Moderate	5.1
Skechers	21.52	Moderate	7.4
Nike	23.82	Moderate	8.9
Data Storage			
Seagate	52.44	Moderate	4.8
Western Digital	44.86	Low	3.6
Sandisk	61.02	Low	3.8

性能、品質といった垂直的属性で差別化がされる

「携帯電話、コンピュータ、Data-Strage」

⇒ 異質性が低い

デザインといった水平的属性で差別化がされる

「Toys, Footwear」

⇒異質性が高い

先行研究レビュー

③抽出された評価次元を用いたブランドポジショニング

あるトピックで最も高いMI値を持つ語の出現確率を比較する

そのための距離を導入する

Hellinger Distance

$$f\left(\theta_k^a, \theta_k^b\right) = \left[\frac{1}{2}\sum_k \left(\sqrt{\theta_k^a} - \sqrt{\theta_k^b}\right)^2\right]^{\frac{1}{2}}.$$

先行研究レビュー

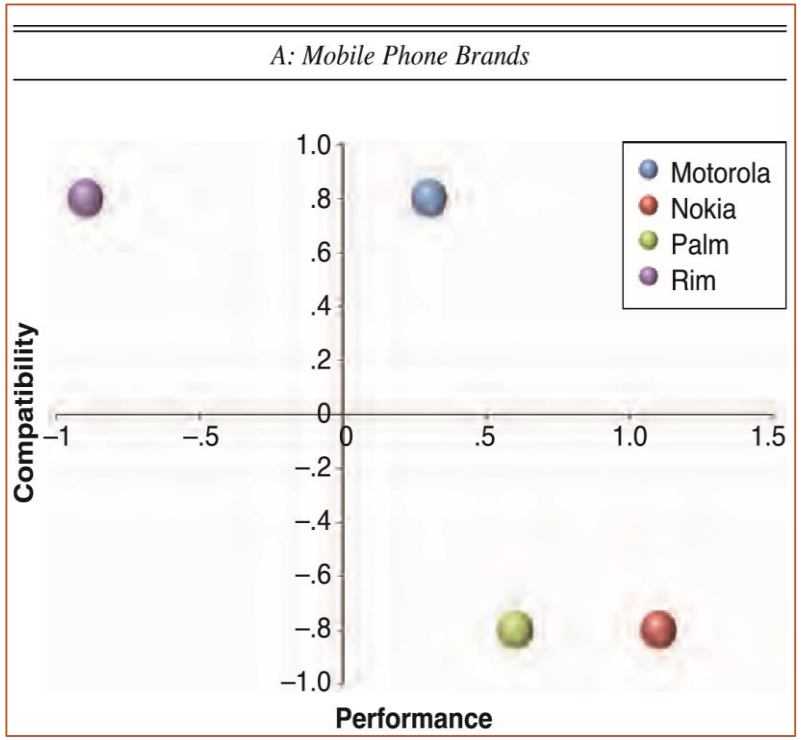
③抽出された評価次元を用いたブランドポジショニング

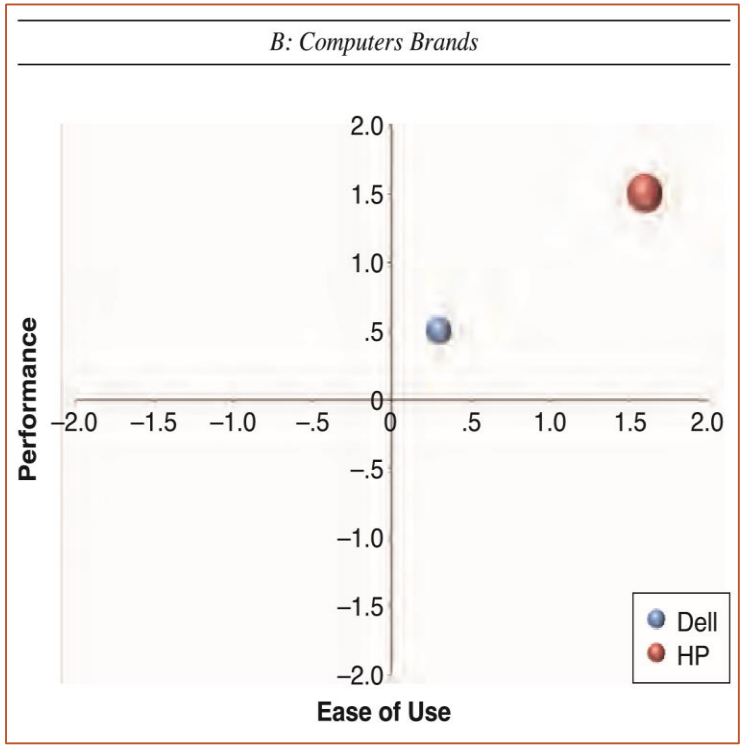
ポジショニングに用いる次元には、比較する二つのブランドに出現する頻度が最も高い2つの次元を採用する

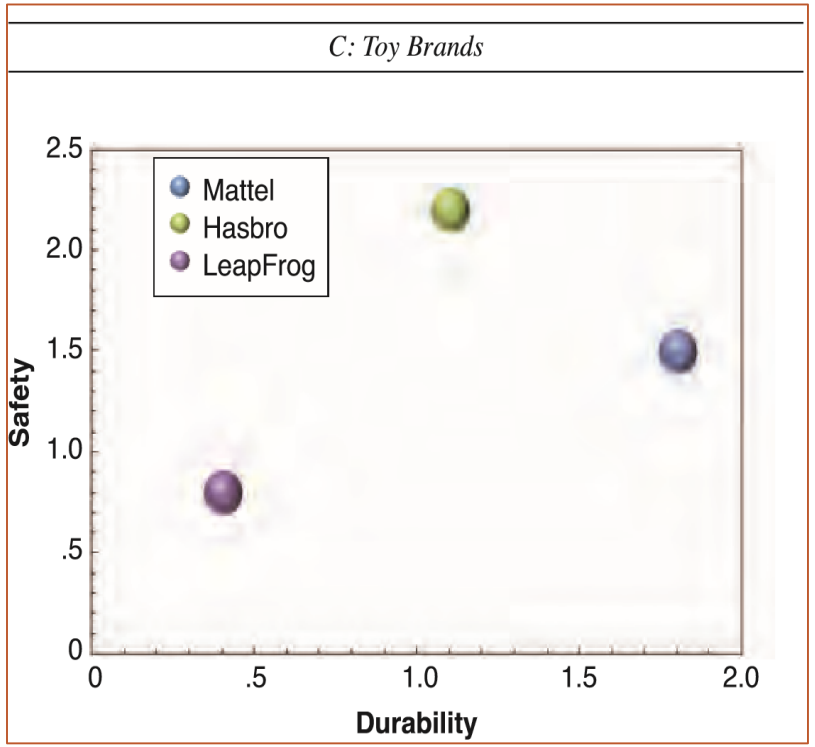
結果では、携帯電話、パソコン、おもちゃについて比較がなされている

先行研究レビュー

③抽出された評価次元を用いたブランドポジショニング







利用するデータ

楽天トラベル

施設データ（約13万施設）

レビューデータ（約620万レビュ）

筑波大学文単位評価極性タグ付きコーパス（1000件のレビューデータ）

Column	Sample
投稿番号	99887766
投稿日時	2012/01/01 12:34:56
ニックネーム	user12345
目的	レジャー
同伴者	家族
評価１（立地）	4
評価２（部屋）	4
評価３（食事）	0
評価４（風呂）	4
評価５（サービス）	3
評価６（設備）	3
評価７（総合）	4

↑ユーザデータ

レビューデータ→

Column	Sample
施設ID	121212
投稿日時	2012/01/01 12:34:56
ユーザ投稿本文	値段のわりには、きれいなホテルでした。以前泊まったときより、きれいに改装されたようで、入ったときにはびっくりしました。しかし、以前と同じくフロントでは丁寧な対応をして頂き、気分よく泊まることができました。接客には十分満足しましたが、駐車場が少し遠かったり、コンビニが近くになかったりしたところが少し残念です。朝食はバイキングでしたが、品数も多く、おいしかったですが、レストランスタッフの方を呼んでもなかなか出てこられなかったのは残念でした。
投稿番号	99887766
分類	感情・情報
プランID	242424
プランタイトル	新館、禁煙、朝食バイキング！平日お得プラン！
部屋種類	2
部屋名前	新館・禁煙・セミダブルルーム
施設回答本文	この度は、ご利用頂きまして誠にありがとうございます。また、貴重なご意見をお寄せ頂きありがとうございました。今年初めに改装しましたが、お褒め頂きましてありがとうございます。レスト

Column	Sample
施設ID	121212
facility name	品川楽天タワーホテル

個施設マスターデータ

Column	Sample
行ID	123
文書ID	10
文書内でのローカル文ID	1
作業者１の評価ラベル・アノテーション	p
作業者２の評価ラベル・アノテーション	p
文	女将さんをはじめ、スタッフの方みなさんのおもてなしに感動しました。

個評価極性付きデータ

モデルの作成

設定：

①α=0.1　β=0.1

②トピック数は5または10で固定

③データは5000件のレビュー

④名詞、形容詞を抽出

⑤ホテルチェーンは店舗数上位4社と、大江戸温泉物語を採用し比較

⑥宿泊した際のシーン別にも比較を行った

⑦該当するレビューのトピック割合の平均をグラフで表す

トピック数5　ホテル別

1	夕食	近い	部屋	ビジネス	旅館
2	露天風呂	コンビニ	ビジネス	コンビニ	夕食
3	子供	駅	価格	近い	露天風呂
4	美味しい	ビジネス	近い	駅	美味しい
5	おいしい	フロント	ルーム	サウナ	量
6	プール	ツイン	フロント	無い	海
7	バイキング	大阪	カード	徒歩	温泉
8	雰囲気	バス	鍵	車	湯
9	楽しい	立地	コンビニ	浴槽	風呂
10	刺身	シングル	LAN	タバコ	子ども

レビュー数:

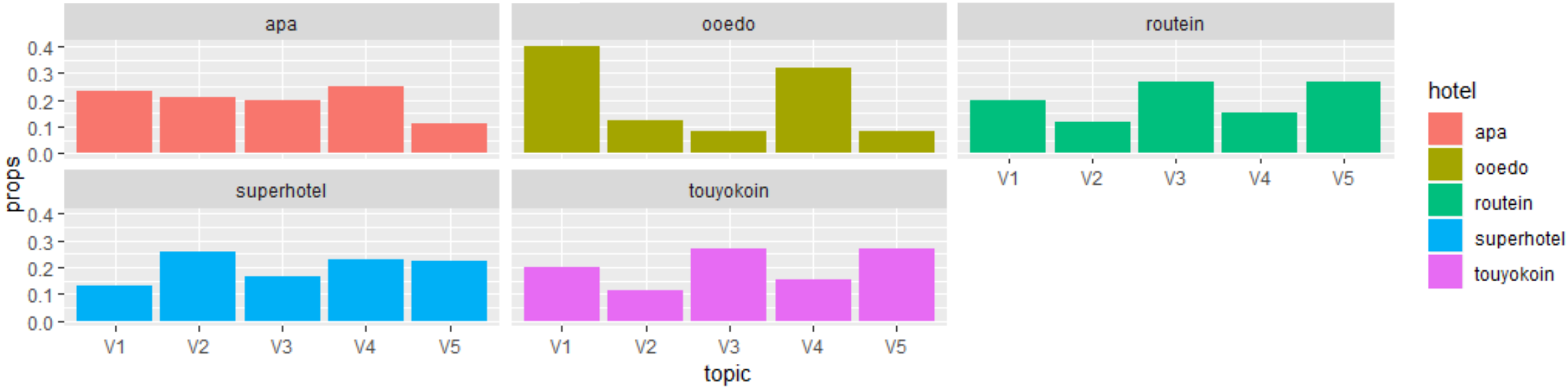
アパ　→106

ルートイン　→21

東横イン　→42

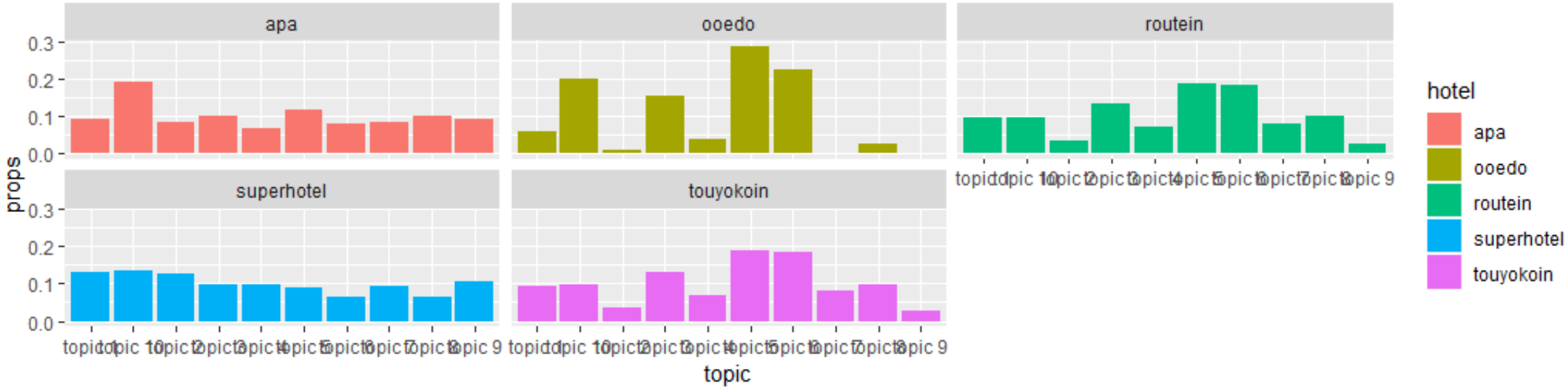
スーパーホテル　→64

大江戸温泉物語　→6



トピック数10　ホテル別

1	子供	美味しい	エアコン	温泉	夕食	駅	近い	近い	美味しい	駅
2	バス	夕食	部屋	露天風呂	美味しい	ない	フロント	良い	味	コンビニ
3	ベッド	温泉	女性	夕食	最高	近い	カプセル	繁華	露天風呂	バス
4	家族	海	暑い	雰囲気	温泉	コンビニ	バス	コンビニ	旅館	近い
5	大人	ご飯	無い	楽天	種類	料金	ベッド	値段	量	ルーム
6	トイレ	旅館	空調	湯	海	フロント	コンビニ	気持ち良い	夕食	ベッド
7	ユニット	家族	ほしい	スキー	風呂	他	テレビ	徒歩	皆さん	価格
8	プール	露天風呂	水	客	コス	少ない	ない	安い	温泉	アメニティ
9	ルーム	仲居	気	泉	パ	欲しい	駅	娘	パン	シングル
10	小さい	種類	ドリンク	最高	雰囲気	店	ルーム	おいしい	種類	立地



トピック数5　シーン別

1	風呂	温泉	近い	ダブル	近い
2	夕食	夕食	料金	部屋	駅
3	露天風呂	露天風呂	種類	LAN	コンビニ
4	湯	風呂	徒歩	ベット	立地
5	温泉	旅館	立地	カプセル	料金
6	子供	家族	コンビニ	シングル	冷蔵庫
7	旅館	量	パン	ネット	空港
8	味	子供	キー	ご飯	価格
9	ご飯	母	店	マイナス	徒歩
10	おいしい	湯	傘	声	無料

レビュー数：

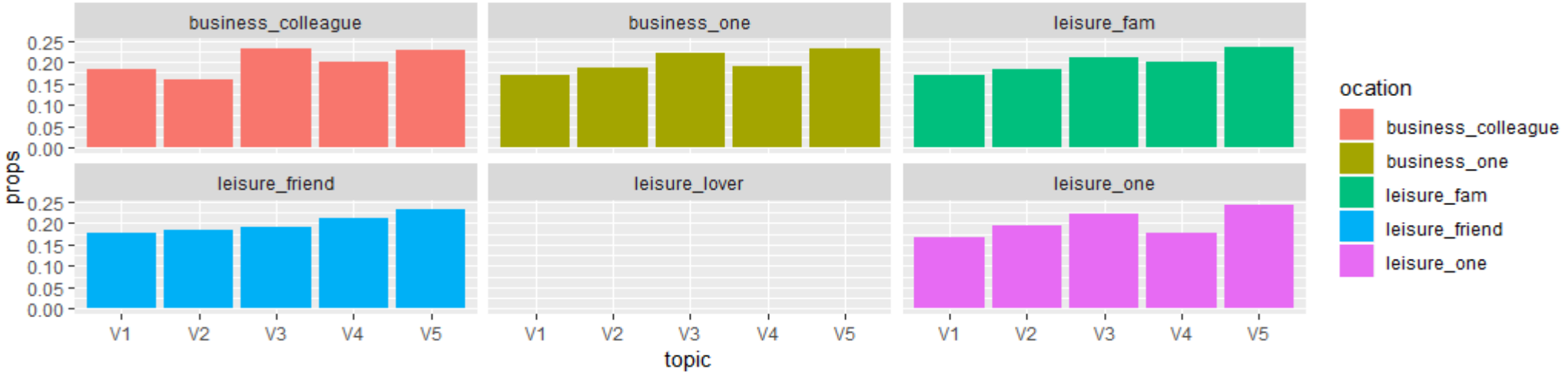
Business_colleague　→159

Business_one　→1351

Leisure_fam　→1638

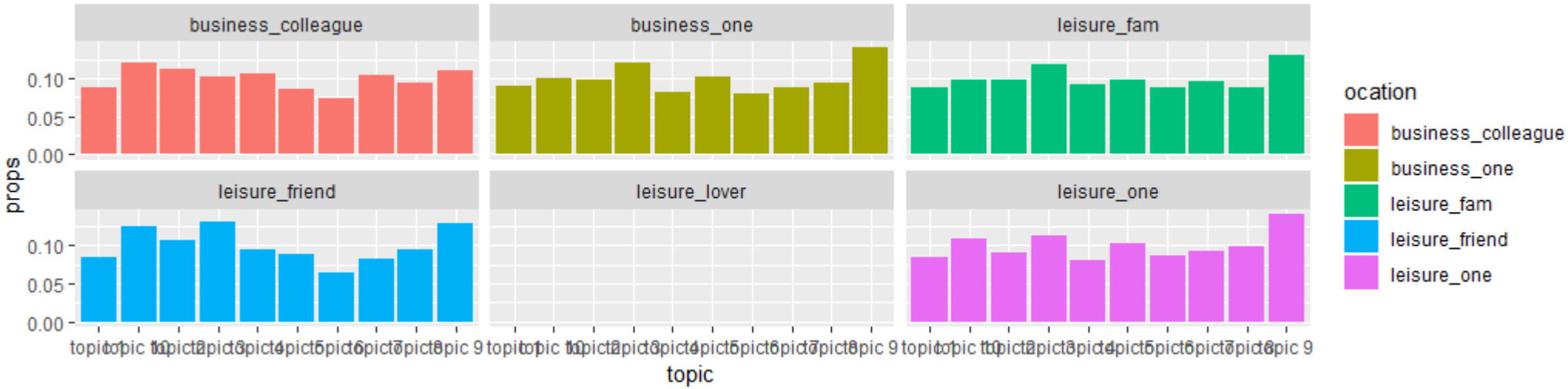
Leisure_friend　→227

Leisure_one　→999



トピック数10　シーン別

1	子供	美味しい	エアコン	温泉	夕食	駅	近い	近い	美味しい	駅
2	バス	夕食	部屋	露天風呂	美味しい	ない	フロント	良い	味	コンビニ
3	ベッド	温泉	女性	夕食	最高	近い	カプセル	繁華	露天風呂	バス
4	家族	海	暑い	雰囲気	温泉	コンビニ	バス	コンビニ	旅館	近い
5	大人	ご飯	無い	楽天	種類	料金	ベッド	値段	量	ルーム
6	トイレ	旅館	空調	湯	海	フロント	コンビニ	気持ち良い	夕食	ベッド
7	ユニット	家族	ほしい	スキー	風呂	他	テレビ	徒歩	皆さん	価格
8	プール	露天風呂	水	客	コス	少ない	ない	安い	温泉	アメニティ
9	ルーム	仲居	気	泉	パ	欲しい	駅	娘	パン	シングル
10	小さい	種類	ドリンク	最高	雰囲気	店	ルーム	おいしい	種類	立地



問題点

①LDAを求める上で、ハイパーパラメータα、β、トピック数が最適化されていない

②求められたトピックが明確な評価次元とはなっていない

③ホテル別比較は、該当するレビューが少なかった

④今回行った可視化ではこれが有意な差なのかどうかわからない

現在考えていること

極性付きのデータを教師とし、全ての単語に極性をつける

極性付きのトピックモデルならば、もう少しトピックの解釈性が向上する？