階層的トピックモデルに基づく宿泊レビューの時間変化分析

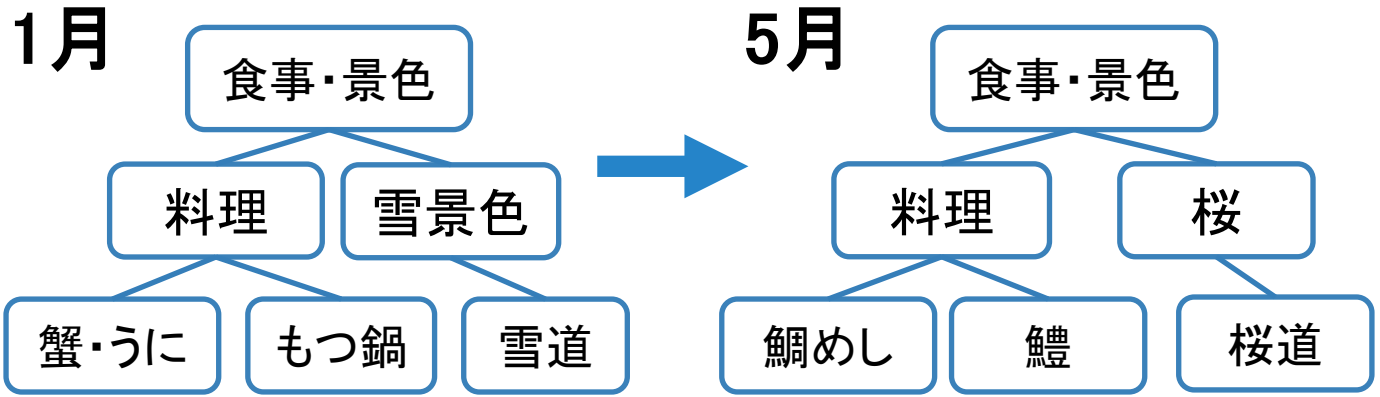
第23回インタラクティブ情報アクセスと可視化マイニング研究会（SIG-AM）
2019/11/23(土)

佐藤 裕次郎*（立命館大学）
山西 良典（立命館大学）
西原 陽子（立命館大学）

1　研究の概要

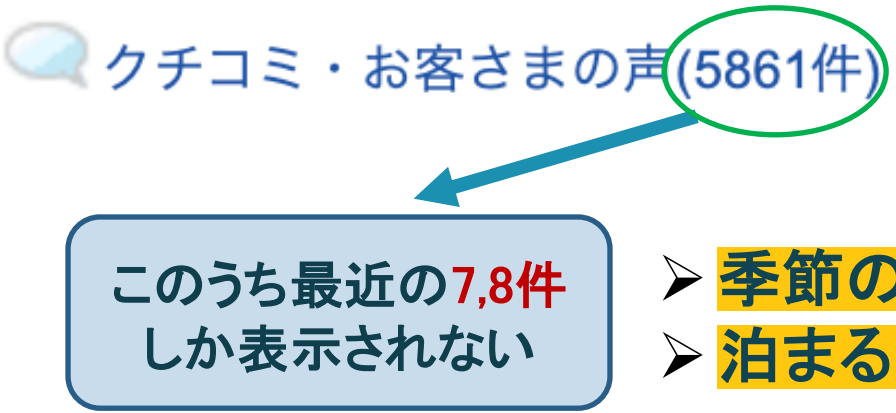
宿泊レビューから

季節の特徴を階層構造で抽出



1 研究背景

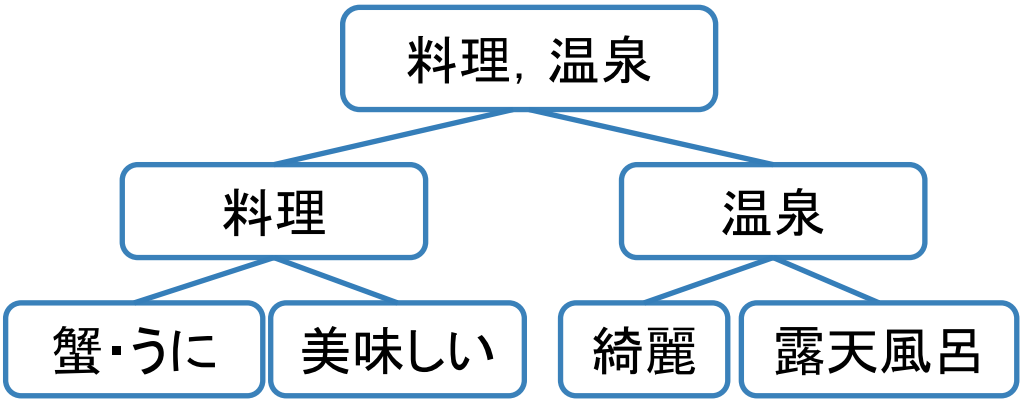
- 宿泊レビューはユーザの感想が得られる情報源
- 人気な宿泊施設はレビュー数が多く，欲しい情報が見つからない



1　研究目標

- 宿泊施設の特徴
  - ➢ カテゴリーの階層構造

この構造で季節の特徴を抽出



1. 研究背景・目標

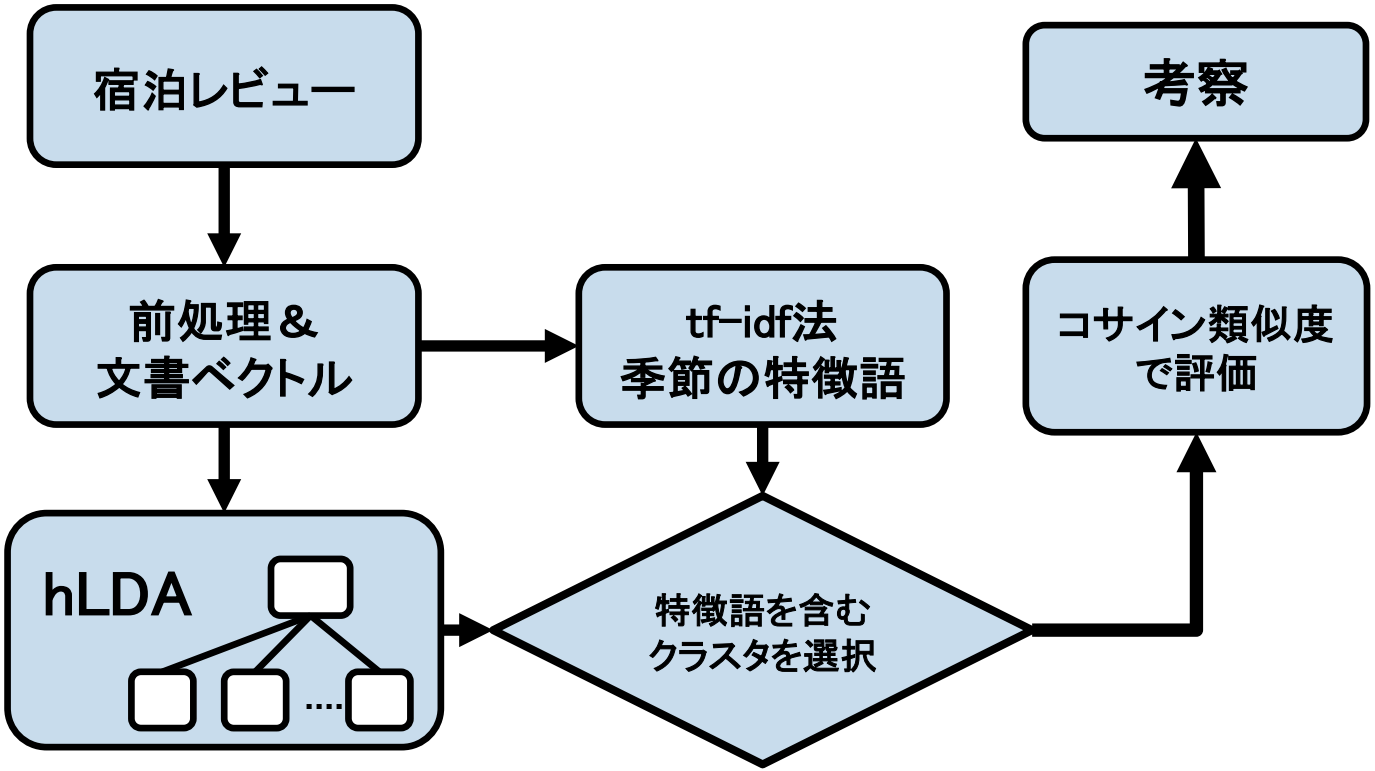
2. 分析手法・評価方法

3. 抽出結果

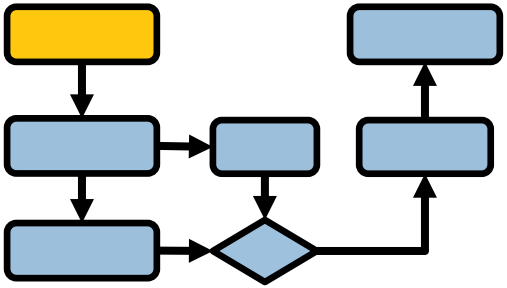
4. 考察

5. 結論・展望

2 分析フロー


宿泊レビュー
前処理＆
文書ベクトル
hLDA
tf-idf法
季節の特徴語
特徴語を含む
クラスタを選択
コサイン類似度
で評価
考察

2 入力データ

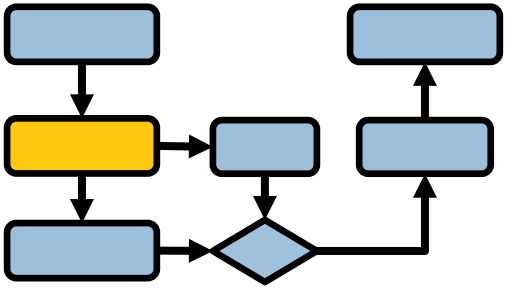


- 楽天トラベルのレビューデータ
  - 1996年～2016年に収録
  - レビュー数上位10%から選出

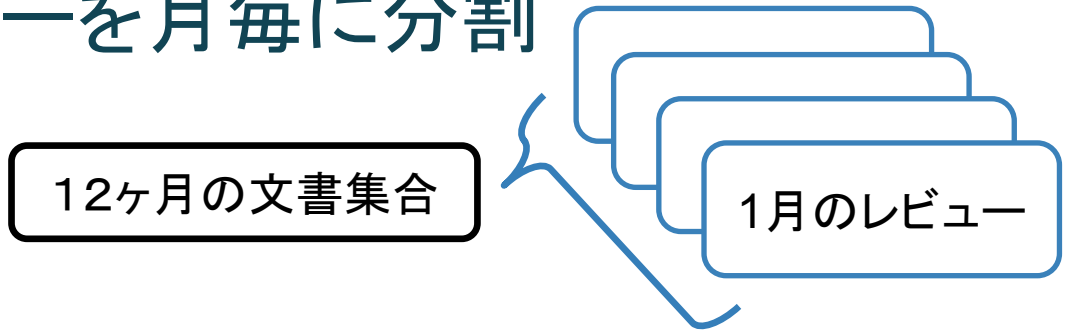
◆粟津温泉 旅亭懐石 のとや（3,118件）

◆シーサイトホテル舞子ヒラ神戸（2,770件）

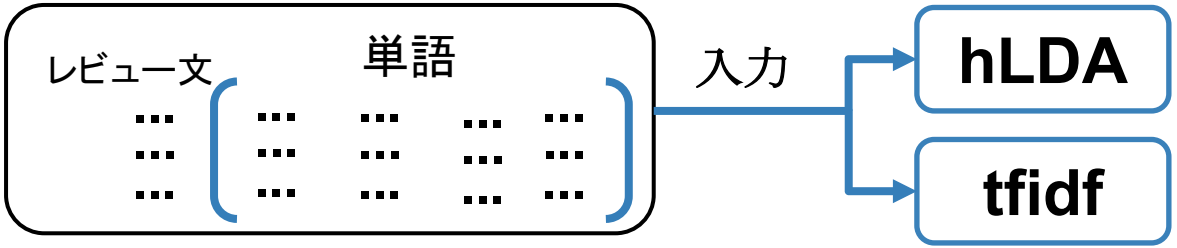
2 文書ベクトル化



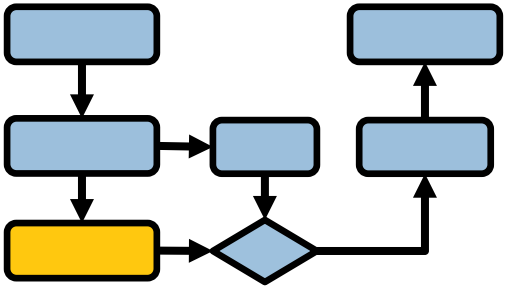
- レビューを月毎に分割



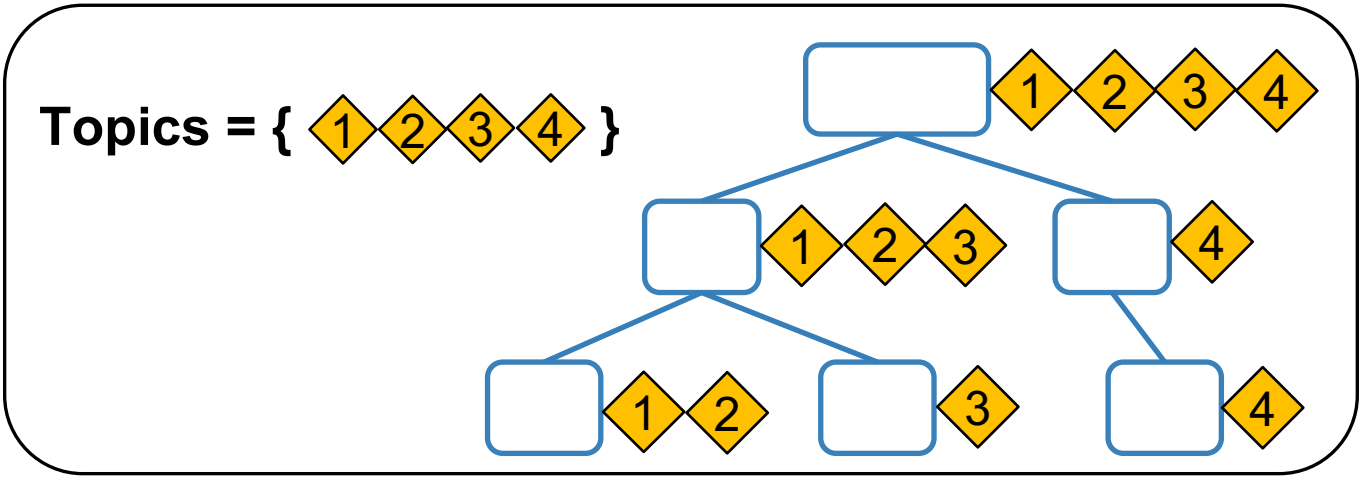
- 文書ベクトルに変換



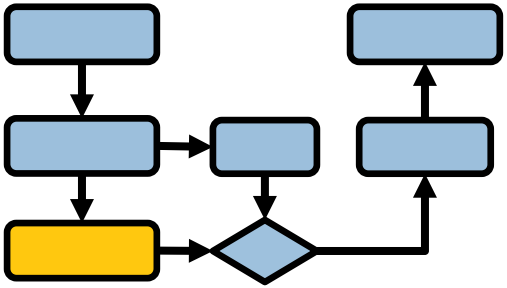
2 hLDA(1/2)



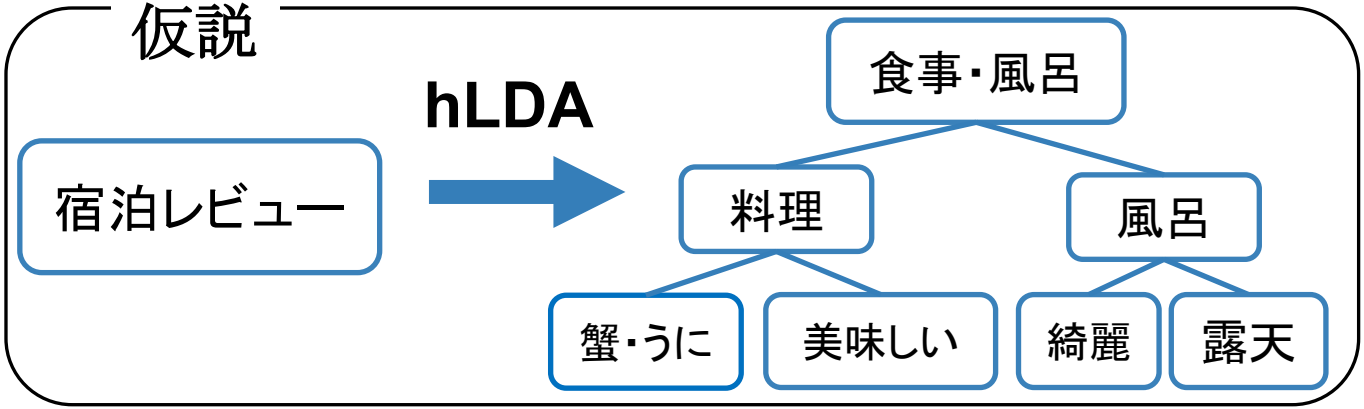
- hLDAによる階層型トピック分析
  - ✓ トピックモデル(LDA)の拡張



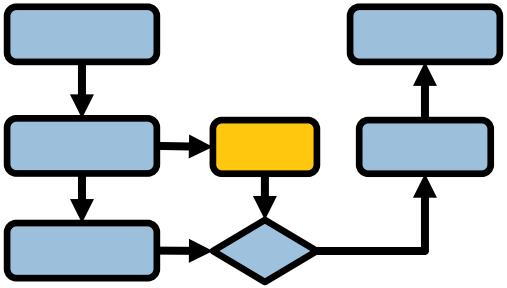
2 hLDA(2/2)



- hLDAで宿泊施設のカテゴリを階層構造で抽出できる
  - ✓ 本研究では名詞が分類される



2　tf-idf法



✓ tfidf値：文書に含まれる単語の重要度

→つまり本研究では季節（月）の特徴量と言える

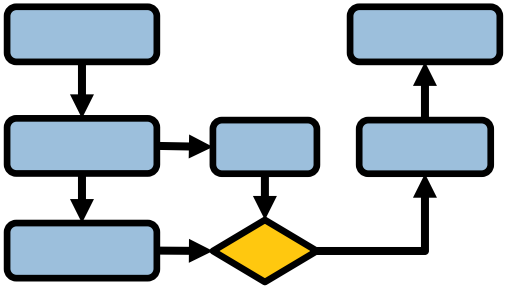
12個のベクトル

tf-idf法

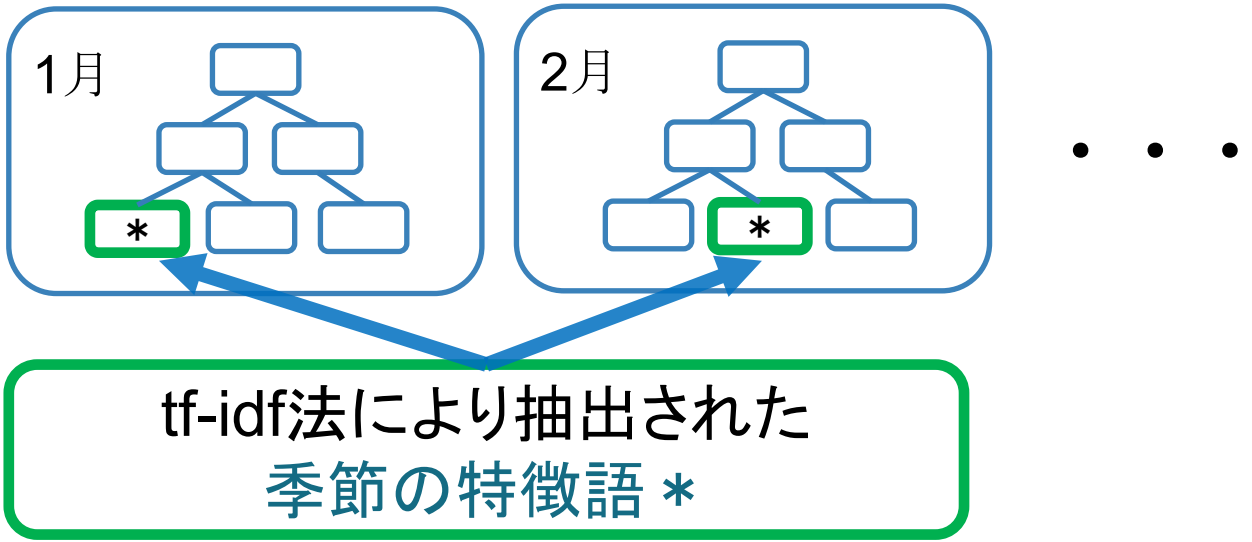
季節の特徴語

各月のtfidf値の大きい
上位5件の名詞集合

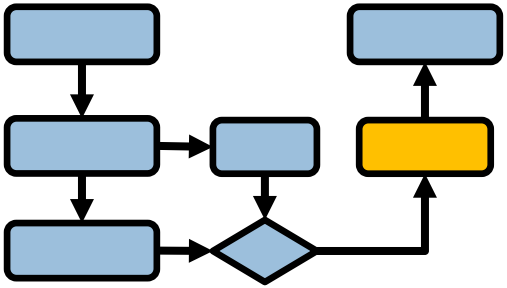
2　クラスタの選択



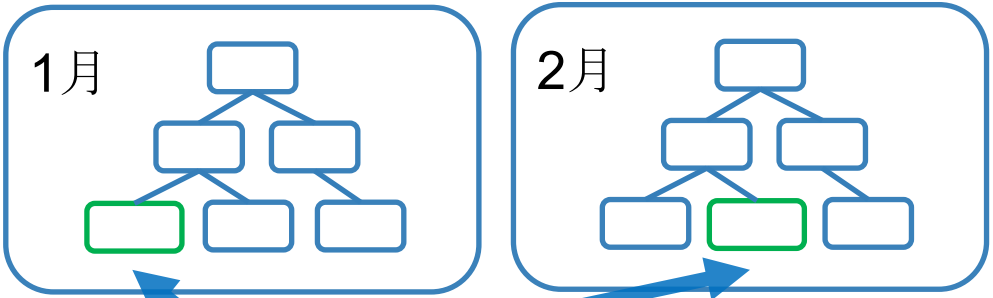
- 季節の特徴語を含むクラスタの抽出



2 コサイン類似度(1/2)

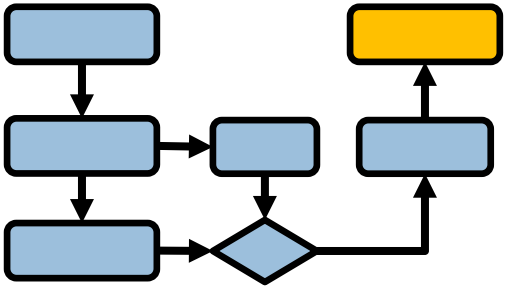


- コサイン類似度による評価
  - ✓ コサイン類似度：ベクトル間の類似度

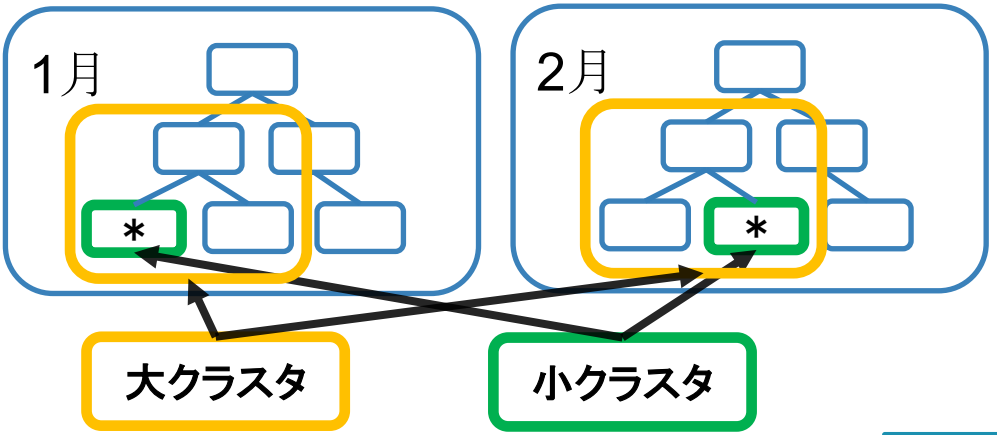


クラスタ間の類似度 → 高い：似た特徴を持つ
低い：季節的な特徴が変化した

2　コサイン類似度(1/2)



- 類似度の変化量を評価



・・・

✓ 2種類のクラスタ間類似度を分析 → 季節の大きな変化
小さな変化

1. 研究背景・目標

2. 分析手法・評価方法

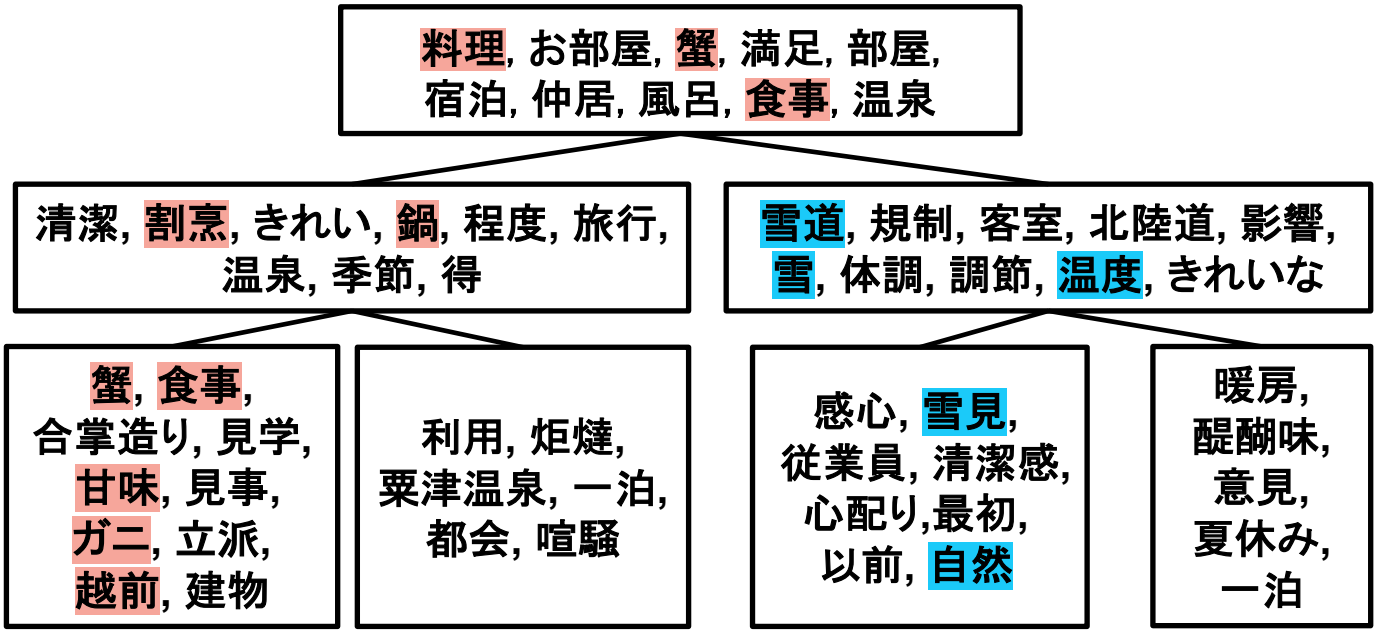
3. 抽出結果

4. 考察

5. 結論・展望

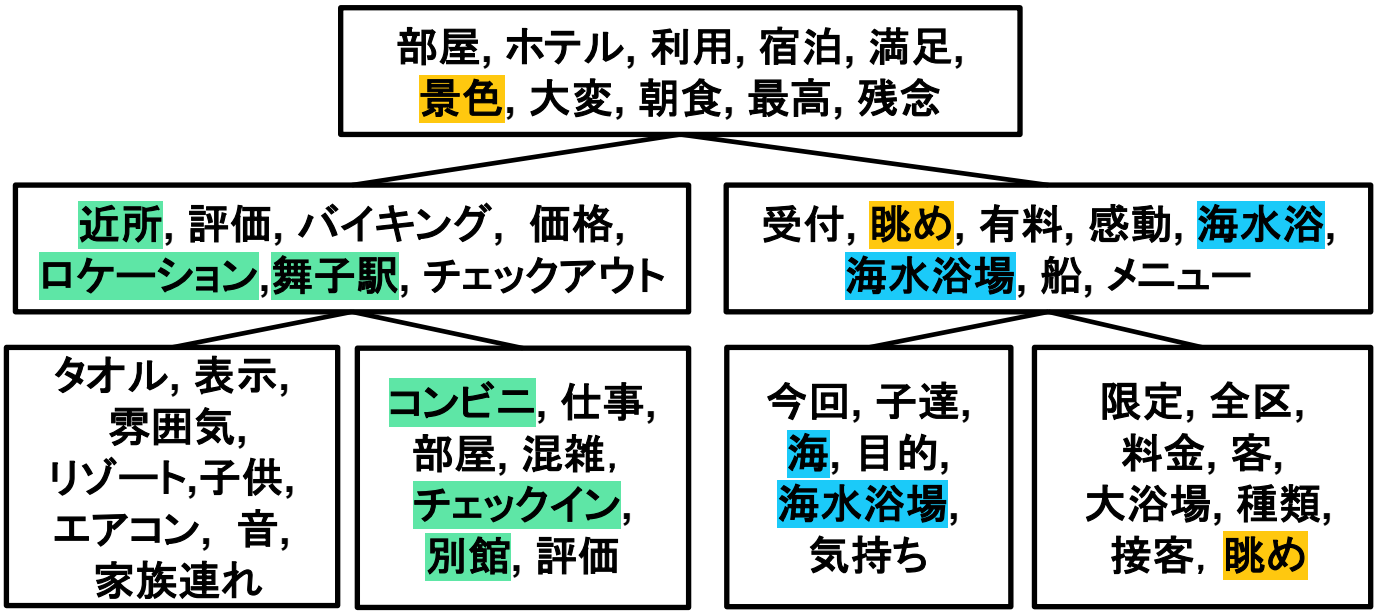
3　hLDAの結果(1/2)

- hLDA, 1月, 粟津温泉 旅亭懐石 のとや



3　hLDAの結果（2/2）

- hLDA，8月，シーサイドホテル舞子ビラ神戸



3 tfidf法の結果

- 宿泊レビューから抽出したtfidf値が高かった上位5件の名詞

✓ 宿泊施設の繁忙期の月を表示

	粟津温泉 旅亭懐石 のとや	シーサイトホテル舞子ヒヲ神戸
1月	お正月，年越し，年始年末，餅搗き，最上級	お正月，お年玉，お菓子，周年，スへ゜シャル
5月	コ゛ールテ゛ンウィーク，鯛，春の，愛，連休	室温，コ゛ールテ゛ンウィー ク，週末，ク゛レート゛，エステ
8月	お盆，合掌，道場六三郎，鮎，納涼	プール，夏休み，海水浴場，海水浴，バーベキュー
12月	クリスマス，めし，突破，名物，最上級	ルミナリエ，クリスマス，冬，おもてなし，スへ゜シャ ル

1. 研究背景・目標

2. 分析手法・評価方法

3. 抽出結果

4. 考察

5. 結論・展望

4　考察(1/3)

- 「粟津温泉 旅亭懐石 のとや」
- “鯛”(5月)を含むクラスタの類似度

	1月～3月	4月	5月	6月	7月	8月	9月	10月	11月，12月
大クラスタ	0.0	0.029	**1.0**	0.046	0.018	0.018	0.021	0.028	0.0
小クラスタ	0.0	0.099	**1.0**	0.199	0.099	0.099	0.099	0.105	0.0

✓ **4月, 6月, 10月**の類似度が高い → 季節による変化が緩やか

✓ 小クラスタより大クラスタの方が，変化の詳細がわかる

4月,6月に共起した名詞
鯛，鯛茶漬け，香り，懐石

→

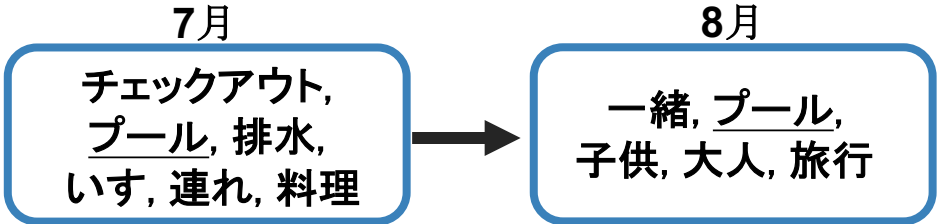
4月～6月にかけて
“鯛”が注目

4　考察(2/3)

- 「シーサイドホテル舞子ビラ神戸」
- “プール”(8月)を含むクラスタの類似度

	1月～5月	6月	7月	8月	9月	10月~12月
大クラスタ	0.0	0.110	0.042	**1.0**	0.018	0.0
小クラスタ	0.0	0.099	0.099	**1.0**	0.099	0.0

✓ 7月, 9月の類似度が低い　→　季節の変化が激しい
✓ 7月, 8月小クラスタの比較　→　8月は家族旅行が人気



4　考察(3/3)

- 「粟津温泉 旅亭懐石 のとや」
  - ✓“鯛”が4月から6月にかけて人気
  - ✓食事に関しての注目度が高い
- 「シーサイドホテル舞子ビラ神戸」
  - ✓プールは6月～9月に楽しめる
  - ✓8月は家族旅行が人気

1. 研究背景・目的

2. 分析手法・評価方法

3. 抽出結果

4. 考察

5. 結論・展望

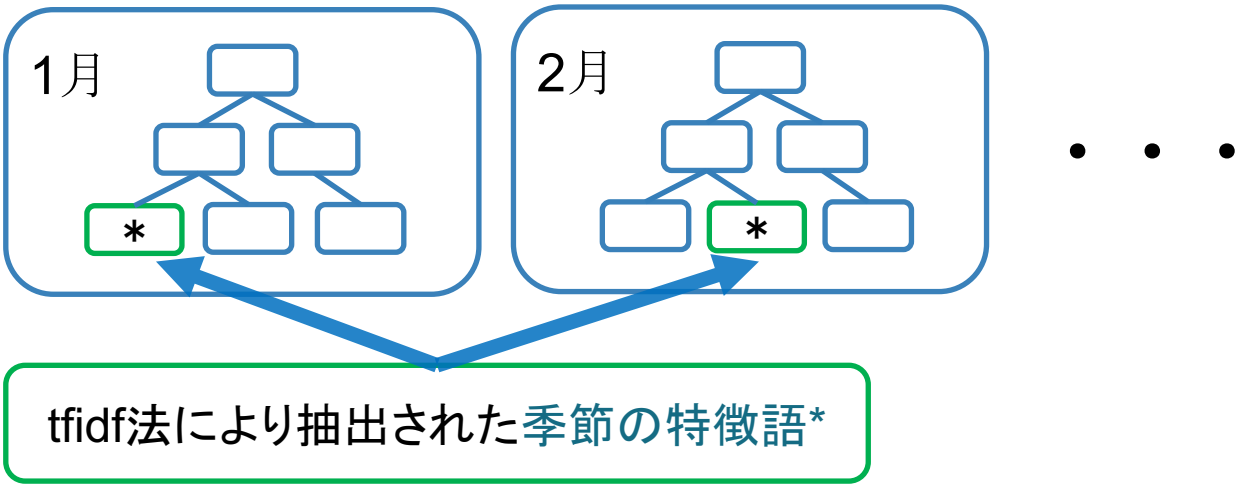
5 課題・展望

- hLDAの精度
  - ハイパーパラメータの調整
- 地域における特徴量の差
- 宿泊施設決定支援へのモデル化

5　まとめ

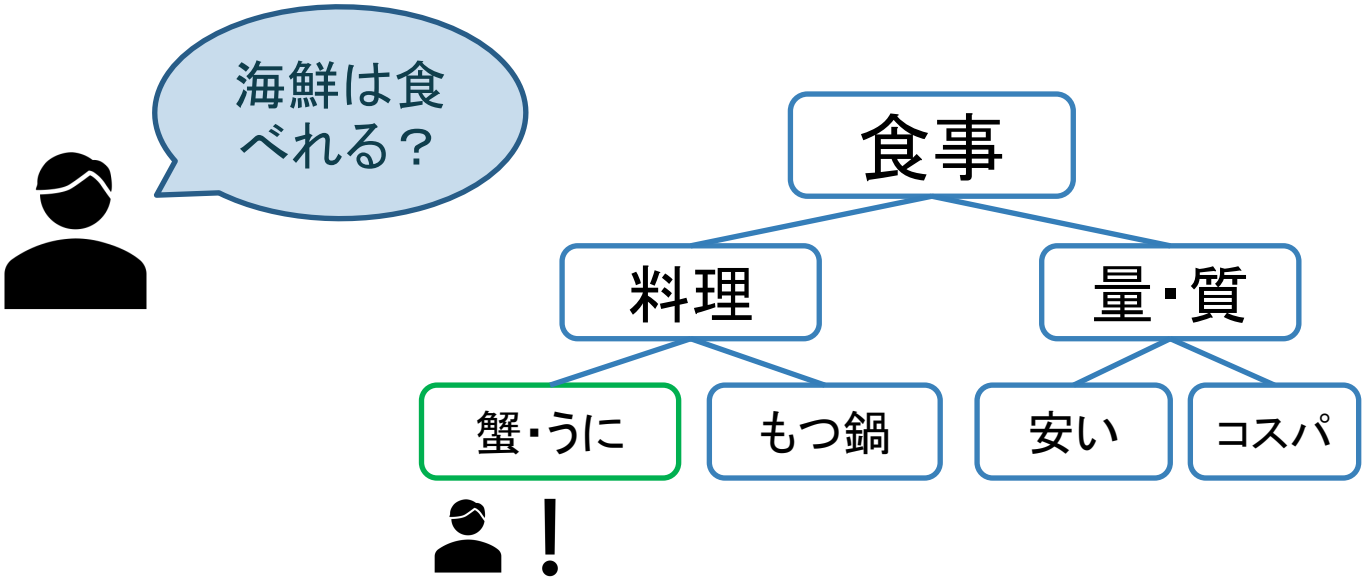
1. 宿泊レビューの季節の特徴を階層構造で抽出することを目的とした
2. hLDAとtfidf法の組み合せによる分析
3. その結果を類似度により評価
4. それぞれのクラスタには季節差が存在
5. 今後は地域における特徴量の差を抽出

- 季節の特徴語を含むクラスタを抽出



1　研究目標

- カテゴリの細分化により，欲しい情報が手に入る
  - ➢ **宿泊施設決定要因になる**



1　研究目標

1. 宿泊レビューにおけるカテゴリを階層構造で提示
2. 宿泊レビュー内の季節的な特徴を抽出

ラスタを抽

tfidf法により抽出された季節の特徴語*

・・・

5. 結論・展望

1. 研究背景・目的

2. 分析手法・評価方法

3. 抽出結果

4. 考察

5. 結論・展望