顧客レビューにおける テキストデータと6段階評価 の関係性の分析

東京理科大学経営学部経営学科3年

白井陽樹

目次

1.研究背景

2.目的

3.利用データ

4.データの特徴

5.研究の流れ

6.分析
6.1.データ抽出
6.2.主成分分析
6.3.テキストマイニング
6.4.内容一致度のチェック

7.分析結果

8.考察、まとめ

参考文献
Appendix

研究背景

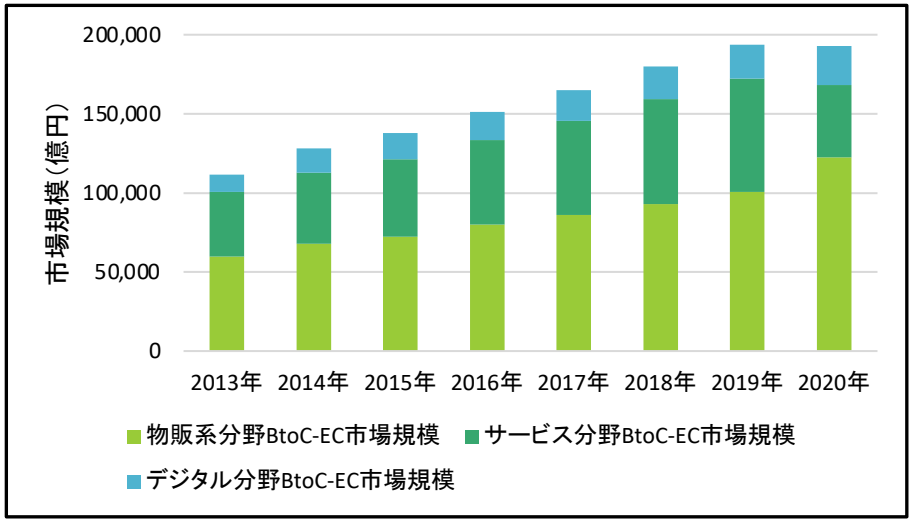


図1　BtoC-EC市場規模の計年推移[1]

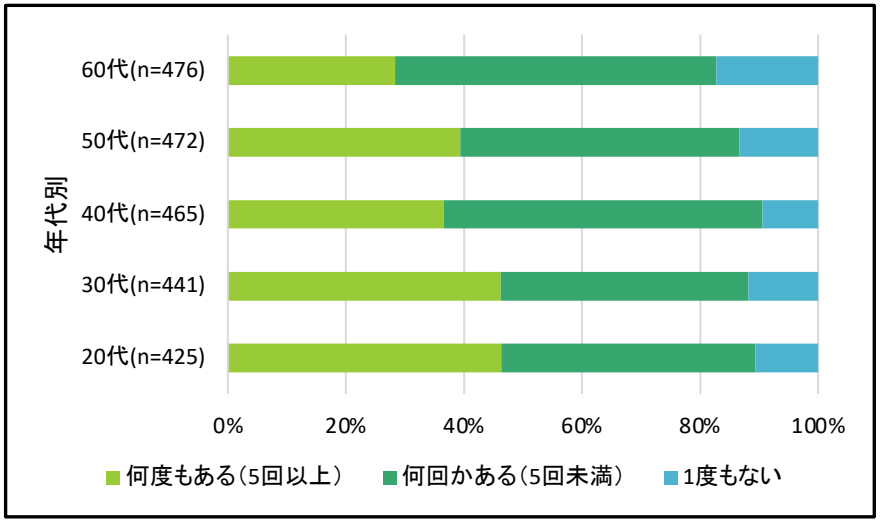


図2　レビューを読んだことで、購入する商品を決定した経験[2]

EC市場の規模拡大に伴い、レビューの重要性も大きくなっている。

研究背景

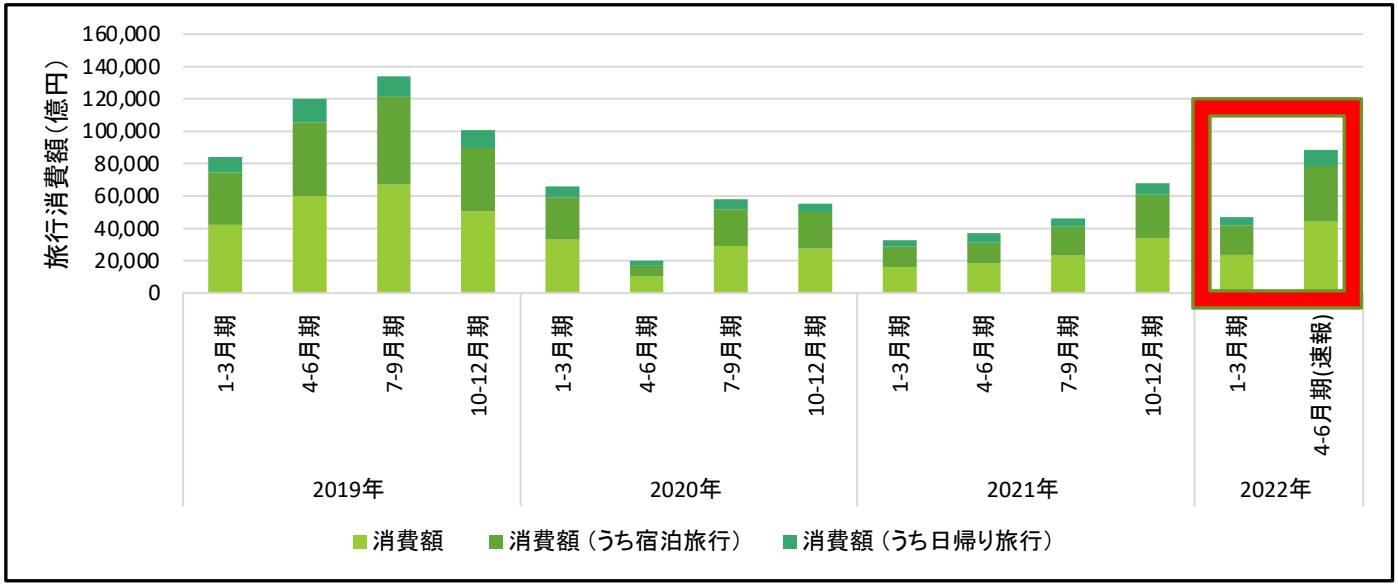


図3　日本人国内旅行消費額の推移

全国旅行支援や外国人の入国規制の撤廃に伴い、日本の観光市場は再び大きな規模となり始めている。
そこで、成長している観光市場に着目して、分析してみたいと考えた。

目的

レビューの件数を元に抽出した宿泊施設のデータに主成分分析とテキストマイニングをし、どのような傾向が見られるか分析をする。

利用データ

・楽天グループ株式会社 (2020): 楽天トラベルデータセット

国立情報学研究所情報学研究データリポジトリ　より

URL： https://doi.org/10.32130/idr.2.2

・楽天公開データのページ

URL：https://rit.rakuten.com/data_release/

データの特徴

表1　利用データのデータ項目

投稿者ID	**評価1（立地）**
投稿日時	**評価2（部屋）**
施設ID	**評価3（食事）**
プランID	**評価4（風呂）**
プランタイトル	**評価5（サービス）**
部屋種類	**評価6（設備）**
部屋名前	評価7（総合）
目的	**ユーザ投稿本文**
同伴	施設回答本文

（評価1～7は0～5の6段階評価、ユーザ投稿本文と施設回答本文はテキスト）

今回は赤字の評価1～6とユーザ投稿本文を利用する。

研究の流れ

元データから基準数以上のレビューがある宿泊施設のデータを抽出する。

↓

抽出した宿泊施設のデータに主成分分析とテキストマイニング行う。

↓

主成分分析で得られた傾向とテキストマイニングで得られた傾向の違いを確認する。

↓

得られた結果から考察をし、結論を出す。

データの抽出

・2018年の楽天トラベルのデータを施設IDごとに並べ、データ数が多い上位50の施設を抽出する。

（最小のデータは274件）

データの選定理由

・2018年にした理由

手元に楽天トラベルのデータが2019年分まであり、今後の研究で年毎の比較ができるようにするため。

・施設数を50件にした理由

今回自力で分析できるデータ数が50件が上限だったかたら。

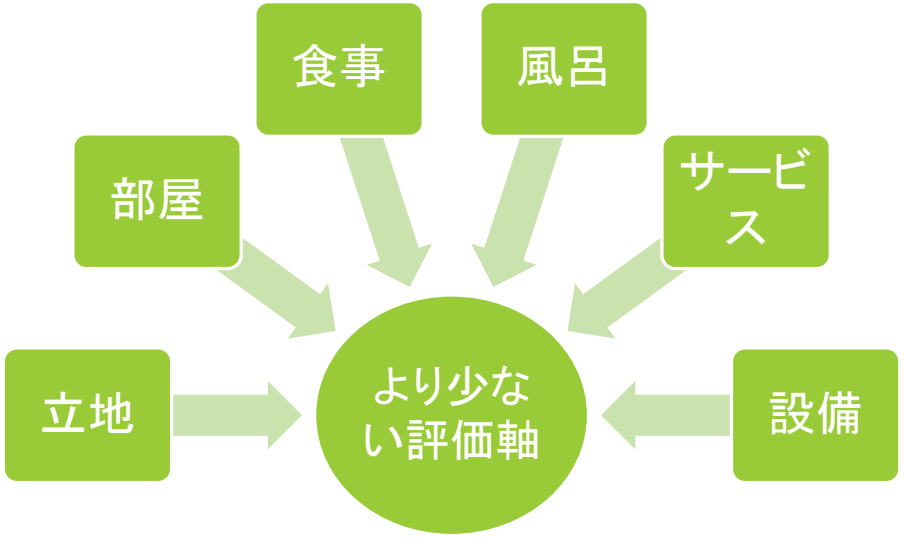
主成分分析

主成分分析とは

「多次元のデータのもつ情報をできるだけ損なわずに低次元空間に情報を集約する方法」[4]

今回で言うと、

Alkanoを用いて、分析するデータである「立地」「部屋」「食事」「風呂」「サービス」「設備」の6項目を可能な限り少ない数の軸で評価できるようにする。



主成分分析

主成分分析で得られた分析結果を以下のルールで利用することとする。

第2主成分時点での累積寄与率が80%以上

- ただし、第1主成分で80%を超えていたら第2主成分は切り捨てる。

各主成分内での主成分負荷量が絶対値0.2未満の要素は切り捨て

- 対象施設の90%の主成分に絶対値0.85以上の要素が存在するため。

第1主成分と第2主成分の平面にプロットされたデータのグループのうち、最大のものを顧客からの評価のメインとする

- マジョリティに絞らなかった場合、分析が困難なため。

主成分分析後（例）

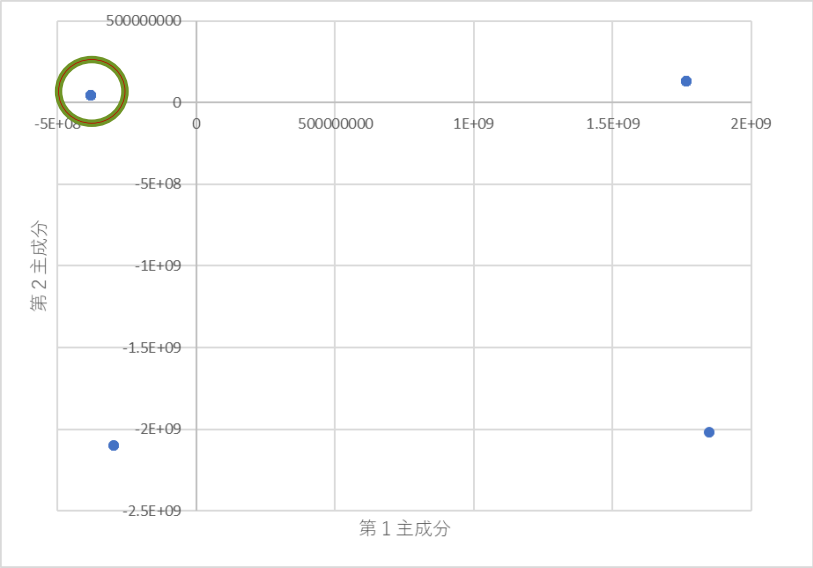


図4　第1主成分と第2主成分の散布図
（図上の赤丸の部分が最多グループ）

表2,3　主成分分析後の出力データ

components	標準偏差	寄与率	累積寄与率
第1主成分	672609901744810240.000000	0.839285	0.839285
第2主成分	128798396196018432.000000	0.160715	1.000000
第3主成分	84.008007	0.000000	1.000000
第4主成分	42.589385	0.000000	1.000000
第5主成分	1.032863	0.000000	1.000000
第6主成分	0.311591	0.000000	1.000000

主成分負荷量	第１主成分	第２主成分	第３主成分	第４主成分	第５主成分	第６主成分
立地	0.000000	0.000000	-0.005881	-0.999954	-0.007594	0.000509
部屋	0.000000	0.000000	-0.999964	0.005926	-0.006082	-0.000864
食事	-0.999256	-0.038571	0.000000	0.000000	0.000000	0.000000
風呂	-0.038571	0.999256	0.000000	0.000000	0.000000	0.000000
サービス	0.000000	0.000000	-0.003616	-0.005598	0.691620	-0.722231
設備	0.000000	0.000000	-0.005020	-0.005103	0.722196	0.691651

赤で囲まれた部分が着目しているポイントで、ここでの例は第１主成分が80%を超えているので、第１主成分のみを考える。
第１主成分内でも風呂の主成分負荷量は絶対値0.2未満なので考慮しない。

主成分分析後（例）

図4　第1主成分と第2主成分の散布図
（図上の赤丸の部分が最多グループ）

表2,3　主成分分析後の出力データ

components	標準偏差	寄与率	累積寄与率
第1主成分	672609901744810240.000000	0.839285	0.839285
第2主成分	128798396196018432.000000	0.160715	1.000000
第3主成分	84.008007	0.000000	1.000000
第4主成分	42.589385	0.000000	1.000000
第5主成分	1.032863	0.000000	1.000000
第6主成分	0.311591	0.000000	1.000000

主成分負荷量	第１主成分	第２主成分	第３主成分	第４主成分	第５主成分	第６主成分
立地	0.000000	0.000000	-0.005881	-0.999954	-0.007594	0.000509
部屋	0.000000	0.000000	-0.999964	0.005926	-0.006082	-0.000864
食事	-0.999256	-0.038571	0.000000	0.000000	0.000000	0.000000
風呂	-0.038571	0.999256	0.000000	0.000000	0.000000	0.000000
サービス	0.000000	0.000000	-0.003616	-0.005598	0.691620	-0.722231
設備	0.000000	0.000000	-0.005020	-0.005103	0.722196	0.691651

前スライドでの条件を踏まえて点の最多グループに着目すると、
ここの結果では**「食事を評価している人が多い」**ことが読み取れる。

テキストマイニング

今回はTMStudio(Text Mining Studio)を用いてテキストマイニングをする。

ユーザ投稿本文のうち、類義語をまとめる処理を行なった後に名詞に絞って単語の出現頻度を解析した。

解析結果のうち上位5位の名詞の中に6段階評価の要素の中に関連する単語（例：朝食→食事など）があるかを確認する。

その理由：名詞だけに絞ってみてもユーザ投稿本文中には1000種類を超える名詞が存在し、その中で上位5位だった場合、その投稿文の中でも重要な要素を占めていると言えるから。

テキストマイニング

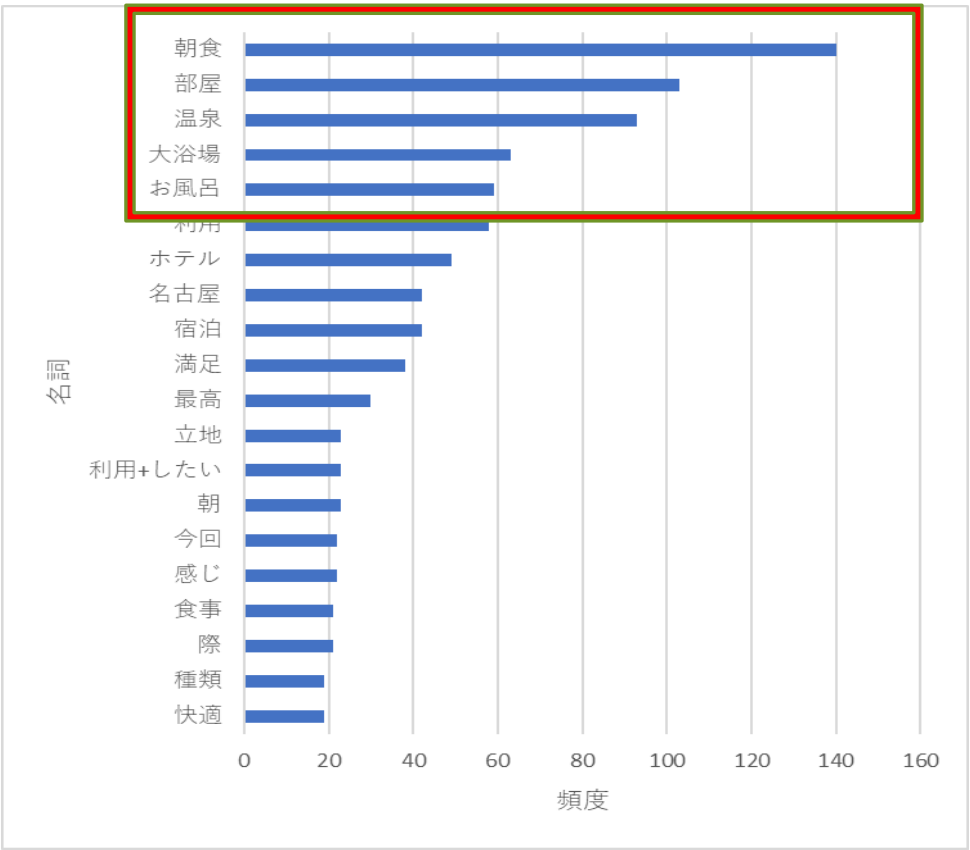
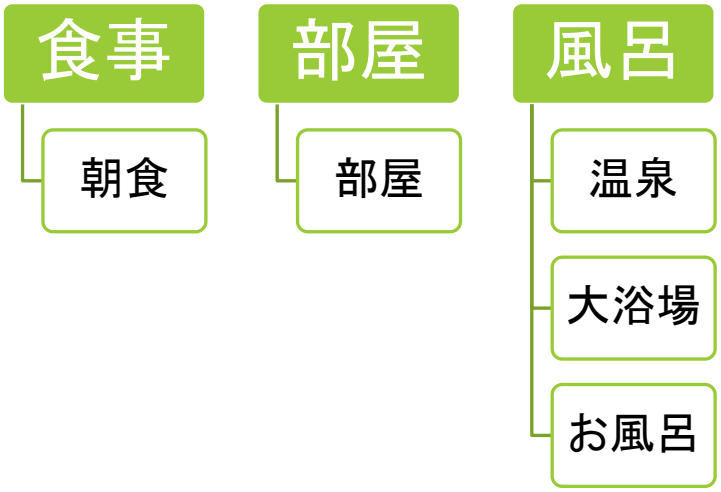


図5　名詞頻度上位20位（例）

左の図の赤で囲まれた部分が抽出された名詞の頻度上位5種類になる。



今回の例では、上のように分類される。

内容一致度のチェック

主成分分析とテキストマイニングでそれぞれ得られた重要視されている要素を下記の基準で比較し、その一致具合を確認する。

完全に一致

- それぞれの分析で出現した要素が全く同じとき

部分的に一致

- どちらかの要素において不足している部分があるとき

不一致

- 出現した要素が全く異なるとき。

例）主成分分析にて「食事」「風呂」、テキストマイニングで「食事」「部屋」

このような場合は「食事」のみが一致しているので“完全に一致”に分類される。

分析結果

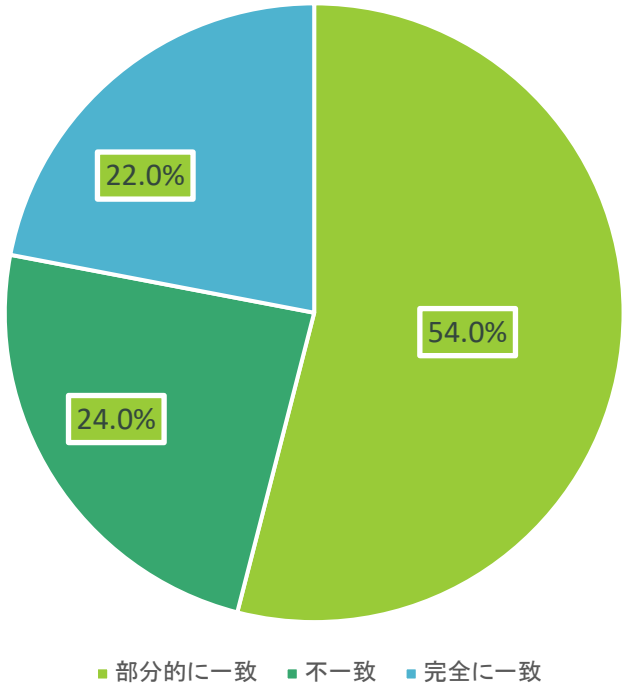


図6　両方の分析結果の一致具合（「部屋」を除く）

読み取れること

・テキストデータには「部屋」がほぼずべての項目で出現したが、主成分分析では全く出現しなかった。

・「部屋」抜きだと図6のように75%が少しでも一致していた。

考察、まとめ

・約７５％のテキストレビューが部分的にでも同じようなことを書いてあり、テキストと数字による評価は比較的近くなるという傾向が見られた。

・テキストマイニングと主成分分析において「部屋」の扱われ方が異なったのは、顧客の部屋に対する評価が良し悪しに関わらず同じようなものだったからだと考えられる。

・主成分分析より、宿泊施設に対する評価は元々の評価項目の中の１つか２つの要素に依存しやすいことがわかった。

参考文献

[1]経済産業省「令和2年度産業経済研究委託事業（電子商取引に関する市場調査）」

https://www.meti.go.jp/policy/it_policy/statistics/outlook/210730_new_hokokusho.pdf

（最終閲覧日：2022年12月1日）

[2]総務省「28年版白書　情報資産（レビュー（口コミ）等）」

https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h28/html/nc114230.html

（最終閲覧日：2022年12月1日）

[3]観光庁「旅行・観光消費動向調査2022年4-6月期（速報）」

https://www.mlit.go.jp/common/001498059.pdf

（最終閲覧日：2022年12月1日）

参考文献

[4]統計科学研究所「主成分分析」

https://statistics.co.jp/reference/software_R/statR_9_principal.pdf

（最終閲覧日：2022年12月1日）

Appendix

Alkano使用時の様子