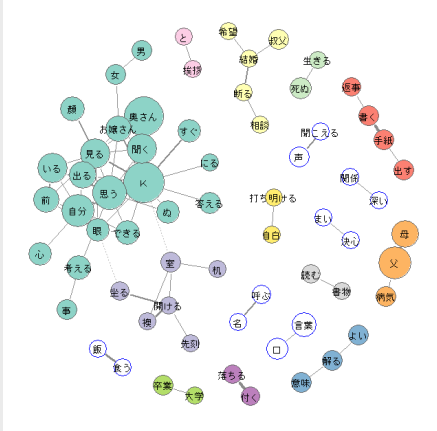


不満情報のテキストマイニングと ビジネス情報の抽出に関する研究

研究背景

マーケティング分析において
テキストマイニングは頻繁に活用される傾向

EX) 共起ネットワーク



ワードクラウド



研究目的

「あなたの不満を買い取ります」
データを用いた
テキストマイニングによる
マーケティング分析

不満調査データセット

投稿情報
• 5,248,820レコード
• 9フィールド
• 2015/3/18～2017/3/12

投稿者情報
• 182,242レコード
• 9フィールド
• 2015/3/18～2017/3/12

分析手法

フリーソフトウェアKH Coderを
文書データの分析に使用

1. 前処理

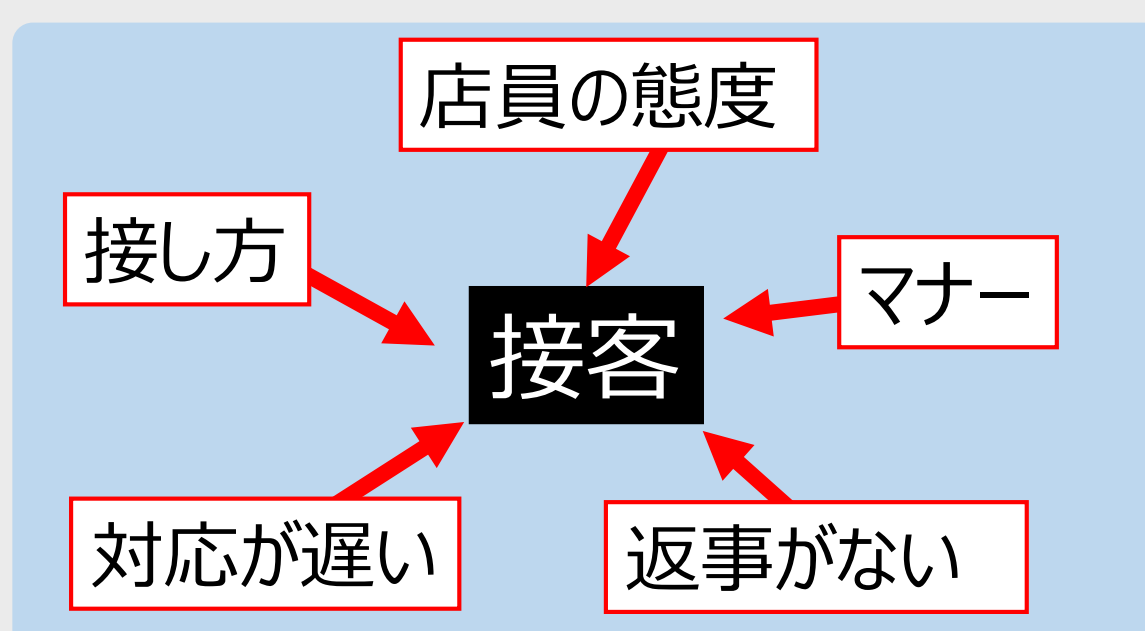
- 形態素解析
- 無視語の設定
- 複合語の設定

2. 抽出語を用いた統計的分析

- コーディングルール作成
- コレスポネンス分析
- クロス集計

コーディングルールの作成

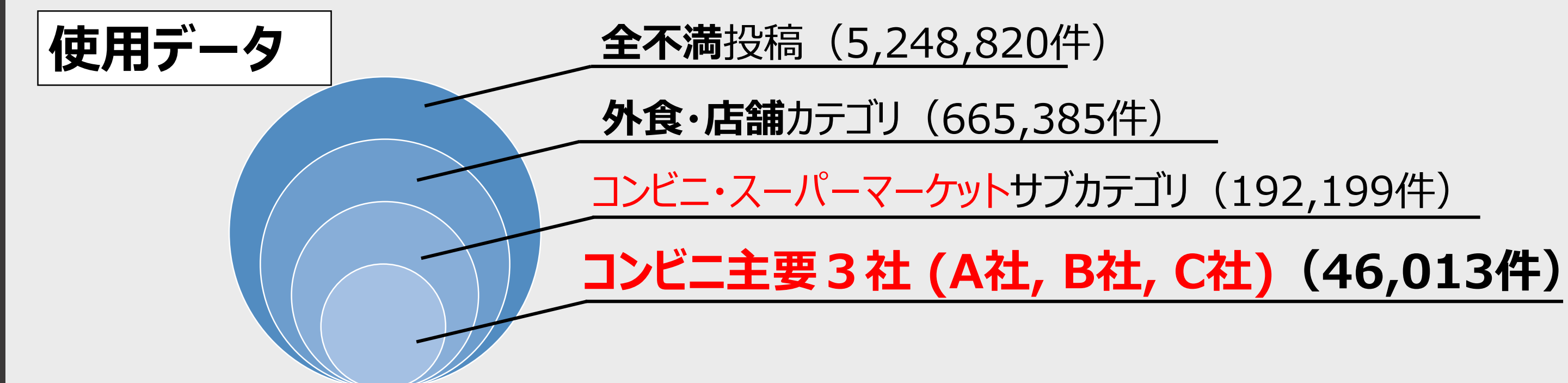
コーディングルールに則り、
一つの**キーワード**に集約



論理演算子を使った記述が可能

分析結果

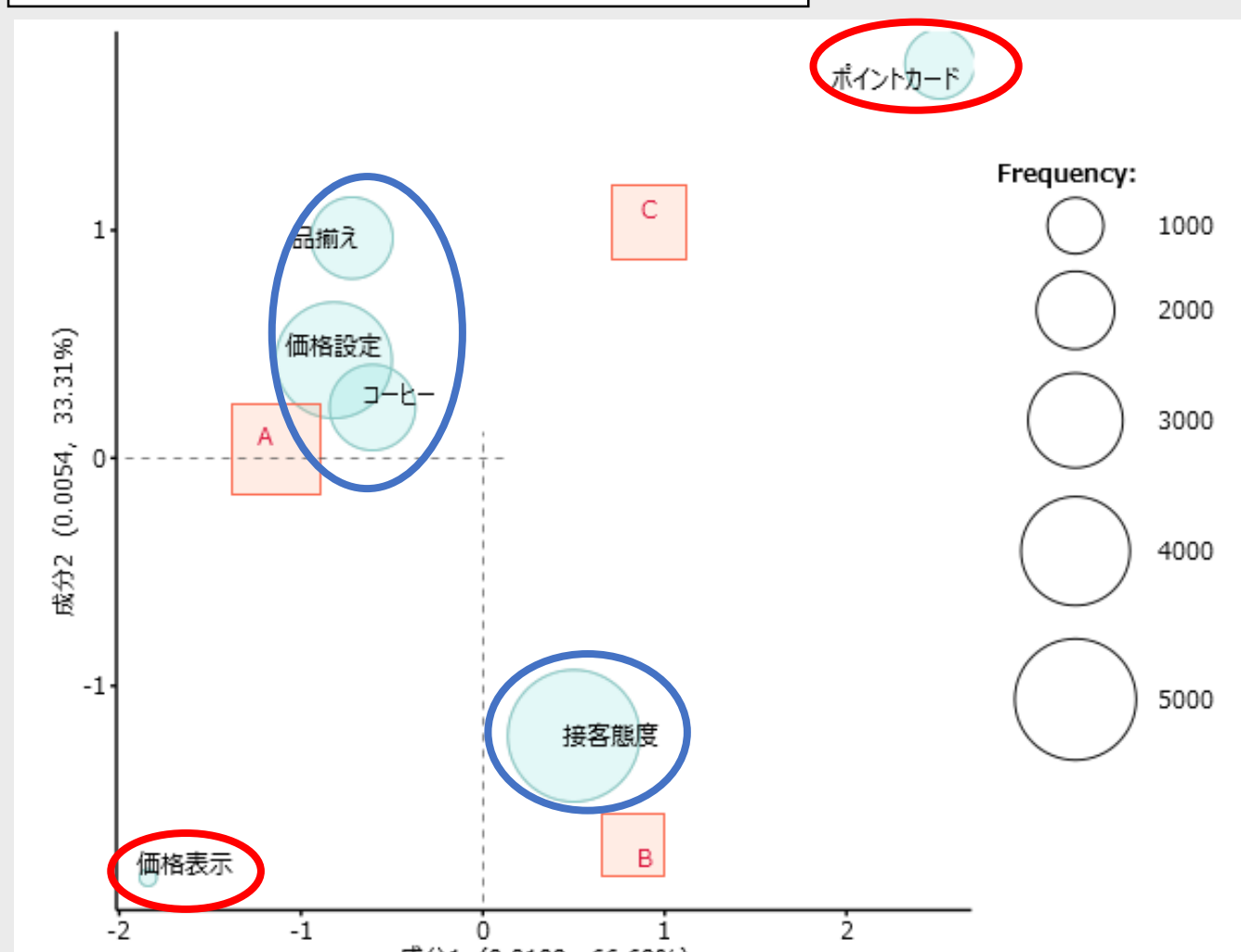
外食・店舗カテゴリ



コーディングルール

1. 接客態度
接客 | 態度 | 店員 | タメ | 不愛想 | 対応 | 表情
2. コーヒー
コーヒー | ミルク | 'カフェ' | 珈琲
3. 価格表示
(価格 | 値段) & 表示
4. ポイントカード
ポイント | カード | Point
5. 価格設定
(価格 | 値段 | 料金 | '料') | (高い | たかい | 安
6. 品揃え
売り切れ | '品揃え' | (品 & 少ない) | Sold | 品切れ

コレスポネンス分析



分割表による独立性の検定

	*接客態度	*コーヒー	*価格表示	*ポイントカード	*価格設定	*品揃え	ケース数
A社	2,434 (11.53%)	1,145 (5.42%)	36 (0.17%)	488 (2.31%)	2,194 (10.39%)	1,048 (4.96%)	21,110
B社	1,561 (15.45%)	494 (4.89%)	16 (0.16%)	347 (3.43%)	880 (8.71%)	396 (3.92%)	10,102
C社	1,954 (13.20%)	807 (5.45%)	16 (0.11%)	746 (5.04%)	1,503 (10.15%)	763 (5.16%)	14,801
合計	5,949 (12.93%)	2,446 (5.32%)	68 (0.15%)	1,581 (3.44%)	4,577 (9.95%)	2,207 (4.80%)	46,013
カイ2乗値	94.816**	4.676	2.397	195.224**	22.629**	22.466**	

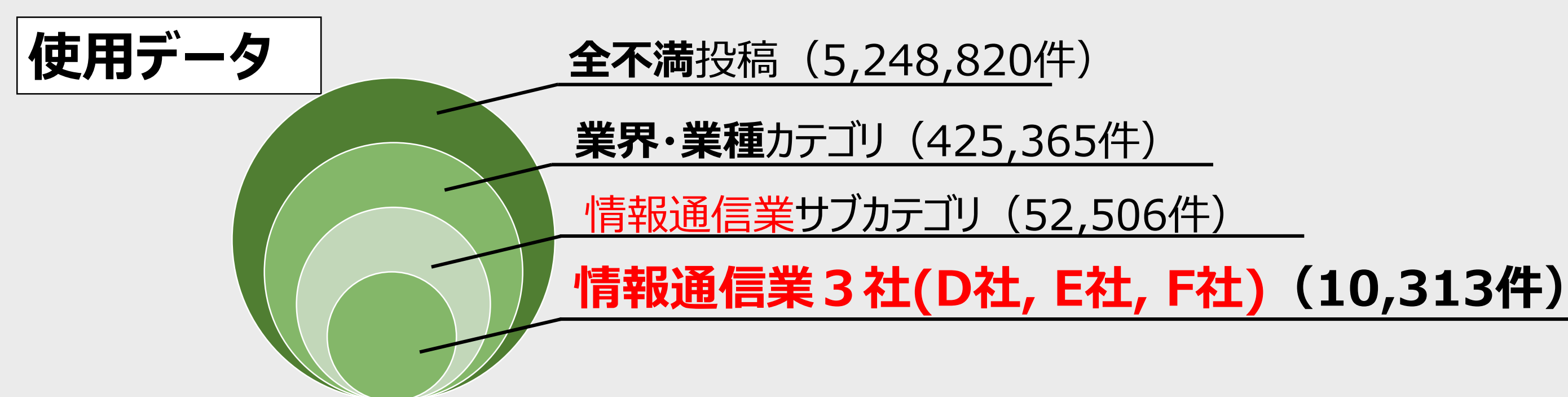
帰無仮説 H_0 : 「比較対象」と「コーディング」は独立 (関連がない)

考察

- 「接客態度」「ポイントカード」「価格設定」「品揃え」に関して3社間で統計的な有意差が認められた
- 「接客態度」は、全体に占める割合が多く重要な不満の要因

B社は「接客態度」の不満を率先して解消する必要がある

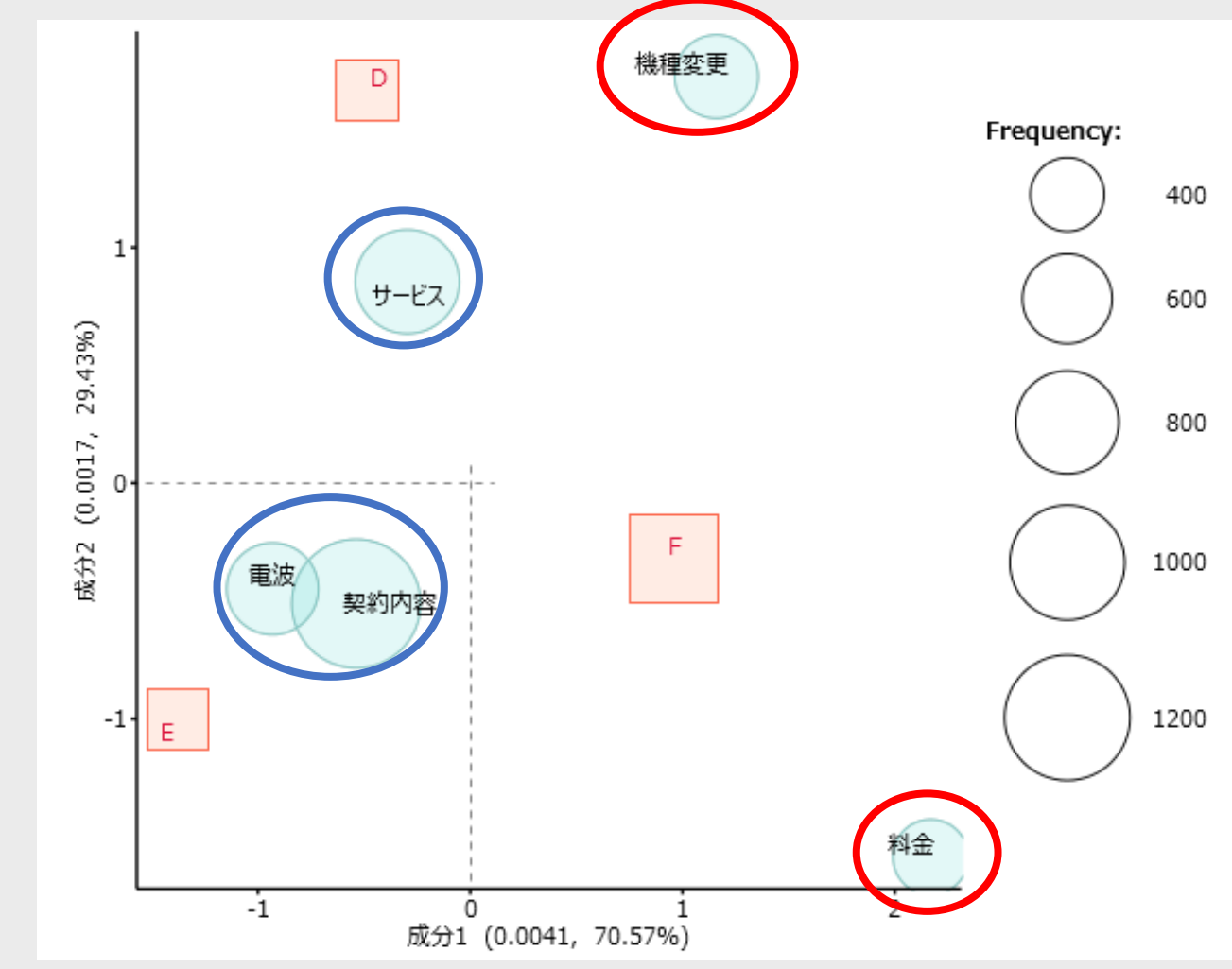
情報通信業サブカテゴリ



コーディングルール

1. 料金
(料金 | 携帯代 | 価格 | 値段) & ((安い & 高い) & 高い) & 高い
2. 契約内容
契約プラン | 契約 | 解約 | プラン
3. 機種変更
機種変更 | '機種変' | (機種 & 交換)
4. サービス
サービス | (店員 & 対応) | (対応 & (悪い | 窓口
5. 電波
電波 | 通信状態 | 電波状態 | 不安定 | 通信 | ((つな
6. 料金
料金

コレスポネンス分析



分割表による独立性の検定

	*料金	*契約内容	*機種変更	*サービス	*電波	ケース数
D社	86 (3.32%)	310 (11.97%)	140 (5.41%)	218 (8.42%)	157 (6.06%)	2,589
E社	94 (3.69%)	347 (13.62%)	110 (4.32%)	208 (8.17%)	178 (6.99%)	2,547
F社	239 (4.62%)	610 (11.78%)	269 (5.20%)	393 (7.59%)	296 (5.72%)	5,177
合計	419 (4.06%)	1,267 (12.29%)	519 (5.03%)	819 (7.94%)	631 (6.12%)	10,313
カイ2乗値	8.626*	5.681	3.766	1.857	4.819	

帰無仮説 H_0 : 「比較対象」と「コーディング」は独立 (関連がない)

考察

- 「料金」に関しての不満割合に3社間で統計的な有意差が認められた
- F社の「料金」に関する不満は、D社と比較して高い

全体でも不満の大きな要因となっているF社の「料金」の不満解消が求められる

トピックモデルを応用したコーディングルール作成自動化の検証



LDAによるトピック抽出

TOPIC ID	要素1	要素2	要素3	要素4	要素5
TOPIC:0	0.023*"変更"	0.022*"機種"	0.018*"充電"	0.015*"メール"	0.014*"携帯"
TOPIC:1	0.018*"通信"	0.017*"機種"	0.016*"携帯"	0.016*"契約"	0.016*"プラン"
TOPIC:2	0.020*"機種"	0.020*"対応"	0.015*"..."	0.014*"ショップ"	0.012*"アプリ"
TOPIC:3	0.030*"料金"	0.020*"契約"	0.019*"携帯"	0.017*"ショップ"	0.017*"解約"
TOPIC:4	0.053*"電話"	0.025*"解約"	0.022*"携帯"	0.017*"契約"	0.011*"機種"

考察

- トピック数を減らすとトピックの判断が比較的容易になった
- トピック名の推測が極めて困難
- トピック分類の際の各トピックの文章数が均等になるため、意味的分類とは程遠い

今後の課題

- 不満調査データの**そのほかのカテゴリ**のマーケティング分析
- コーディングルールの見直し
- LDAなどの手法を用いた**コーディングルール作成の自動化**の検討

謝辞 本研究では、株式会社 Insight Tech が国立情報学研究所の協力により研究目的で提供している「不満調査データセット」を利用した。