研究進捗発表　2018年9月

LDAを用いたAmazonのレビューデータのデータマイニング

B8EM1016　富田優(とみたゆう)

アウトライン

* ①はじめに
* ②前回のおさらい
* ③使用したデータ
* ④モデル適用の結果
* ⑤考察

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

○今日発表する内容

5月に紹介した「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」という論文をもとに進めている研究の発表

前回のおさらい①

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

前回のおさらい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

前回のおさらい③
～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

前回のおさらい④

○やりたいこと

　レビューデータをもっと有効活用して、消費者の商品に対する判断基準を理解し、さらなる購買につなげたい

○具体的な手法

　LDA(トピックモデル)を用いてトピック分布と単語分布を推定

○回帰分析

　商品のレーティングを、トピック分布、単語分布、価格、消費者の属性に回帰する

○どの評判要因がレーティングに影響しているのか分析

実験データ

Amazon.comのレビューデータ		
分類	Office Products	Musical instruments
レビュー数	53,258	10,261
総語彙数	73,104	18,928

実験データ

	asin	helpful.0	helpful.1	overall	reviewText	reviewTime	reviewerID	reviewerName	summary	unixReviewTime
	<chr>	<dbl>	<dbl>	<dbl>	<chr>	<chr>	<chr>	<chr>	<chr>	<dbl>
1	B00000JBLH	3	4	5	I bought m~	09 3, 2004	A32T2H815~	ARH	A soli~	1094169600
2	B00000JBLH	7	9	5	"WHY THIS ~	12 15, 20~	A3MAFS04Z~	"Let it Be ~	Price ~	1197676800
3	B00000JBLH	3	3	2	I have an ~	01 1, 2011	A1F1A0QQP~	Mark B	Good f~	1293840000
4	B00000JBLH	7	8	5	I've start~	04 19, 20~	A49R5DBXX~	R. D Johnson	One of~	1145404800
5	B00000JBLH	0	0	5	For simple~	08 4, 2013	A2XRMQA6P~	Roger J. Bu~	Still ~	1375574400
6	B00000JBLH	10	12	5	While I do~	01 23, 20~	A2JFOHC9W~	scott_from_~	Every ~	1011744000
7	B00000JBLH	3	4	5	I've had a~	01 17, 20~	A38NELQT9~	W. B. Halper	A work~	1168992000
8	B00000JBLH	0	0	5	Bought thi~	11 14, 20~	AA8M6331N~	ZombieMom	Fast s~	1384387200
9	B00000JBLU	3	3	5	This is a ~	12 7, 2010	A25C2M3QF~	Comdet	Nice d~	1291680000
10	B00000JBLU	0	0	5	I love thi~	12 2, 2013	A1RTVWTWZ~	"Hb \"Black~	Love I~	1385942400

モデルの設定

* サンプリング方法：ギブスサンプリング
* トピック数の判断基準：パープレキシティ
* ハイパーパラメータ：α＝？　β＝？
* サンプリング回数　？

Perplexity

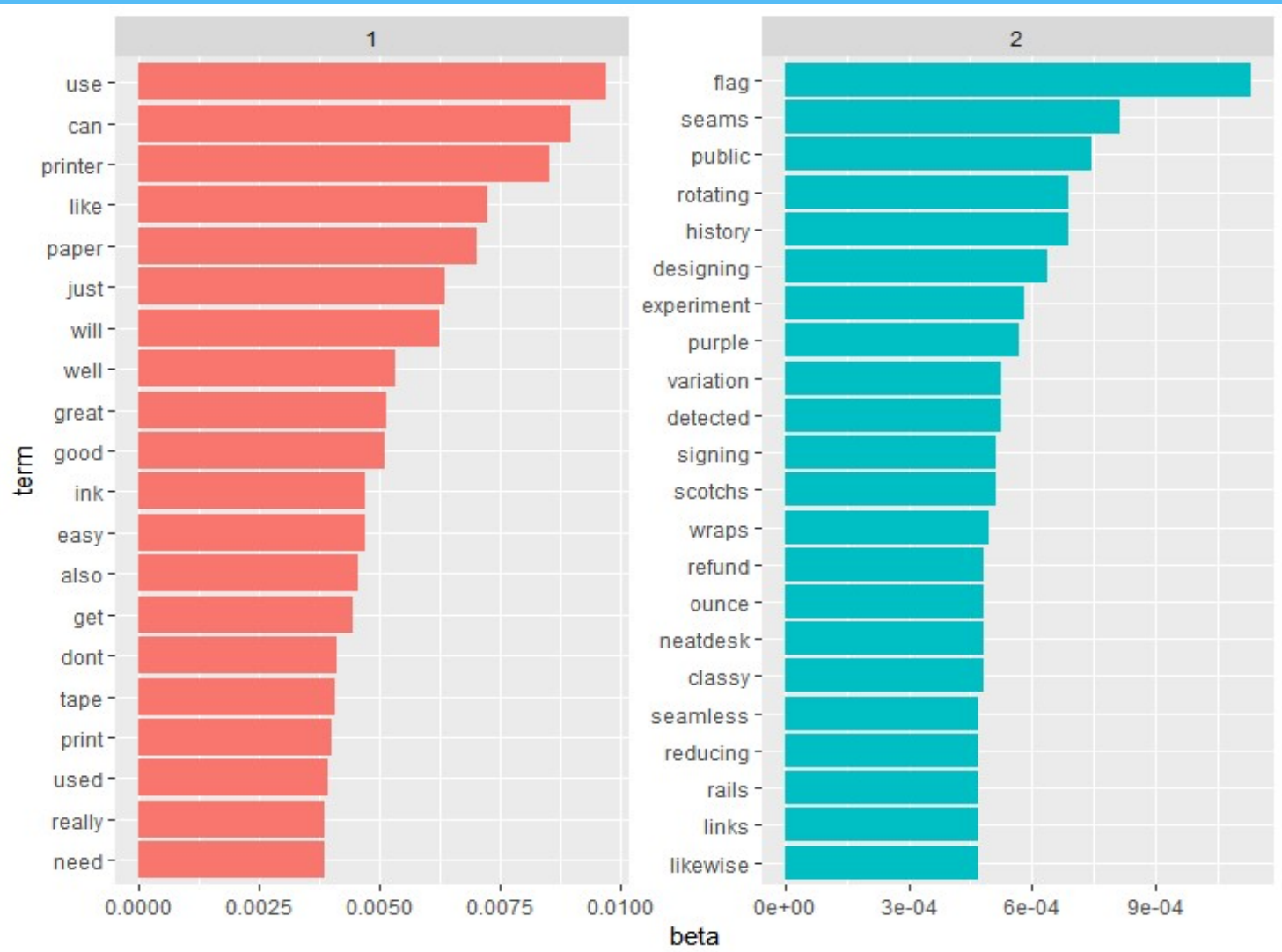
* *Perplexityは分岐数または選択肢の数を表している*
* *モデルによって単語の候補をどれだけ絞り込めるか*
* *より絞り込める方がよい→値が低い方がよい*
* $L\left(\boldsymbol{w^{test}}\middle|\boldsymbol{M}\right)=\sum_{d=1}^{M}\sum_{w_{d,i}\in w_d^{test}}\log p\left(w_{d,i}\middle|M\right)$
* $PPL\left(\boldsymbol{w^{test}}\middle|\boldsymbol{M}\right)=exp\left\{-\frac{L\left(\boldsymbol{w^{test}}\middle|\boldsymbol{M}\right)}{\sum_{d=1}^{M}n_d^{test}}\right\}$
* $\boldsymbol{w^{test}}$**:テストデータの単語集合**
* $\boldsymbol{M}$ **:学習されたモデル**
* $w_{d,i}$**:ドキュメント**d**の**i**番目の単語**

パープレキシティ

トピック数	オフィス製品	楽器
2	933849.5	96999.52
3	1004840	100062.7
4	1052383	103362.1
5	1080513	106244.8
6	1105529	108141.5
7	1136654	109806.3
8	1146791	111919.5
9	1162816	112349.6
10	1176841	113067.9

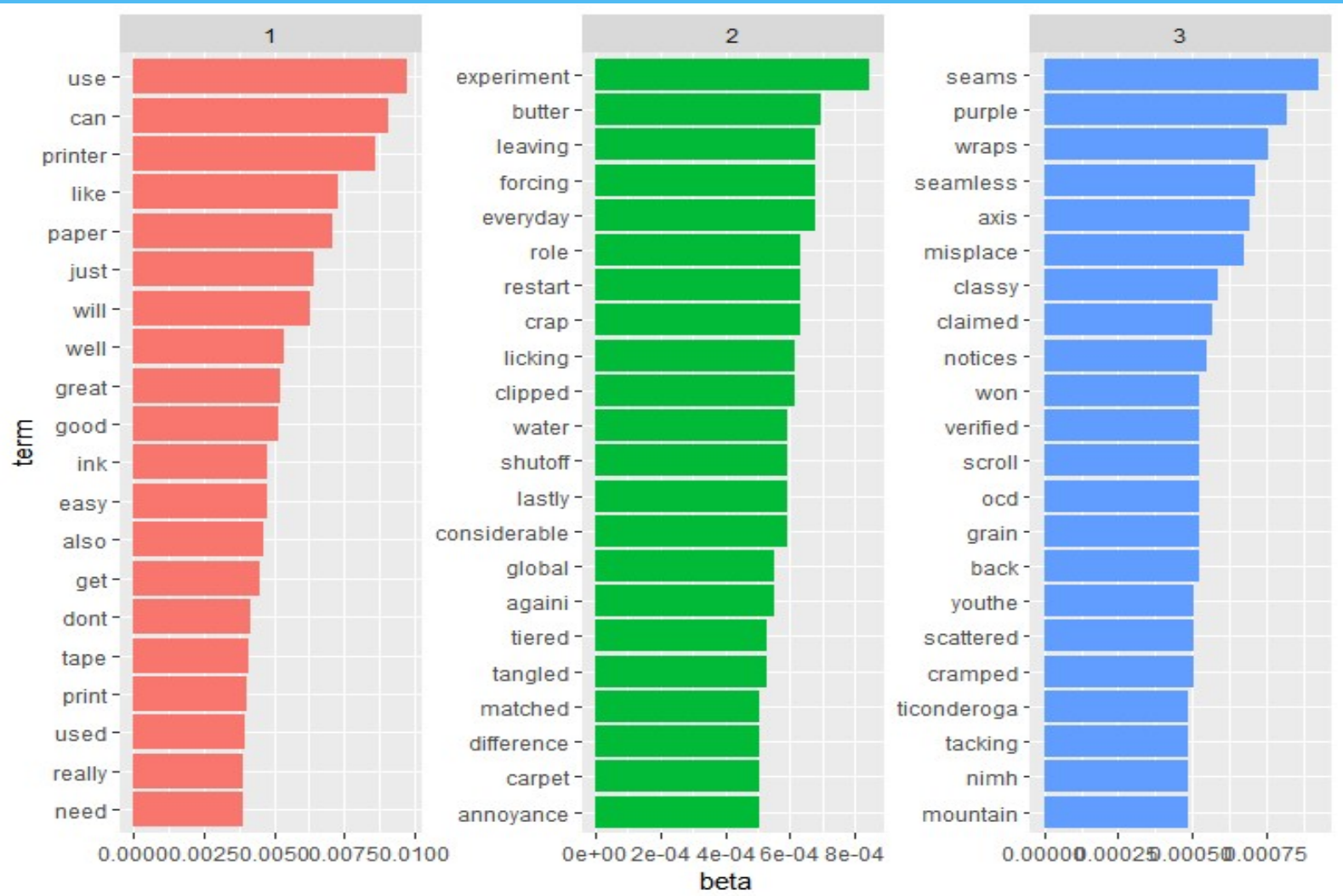
モデル適用の結果

各トピックの単語出現確率の上位20単語:オフィス製品



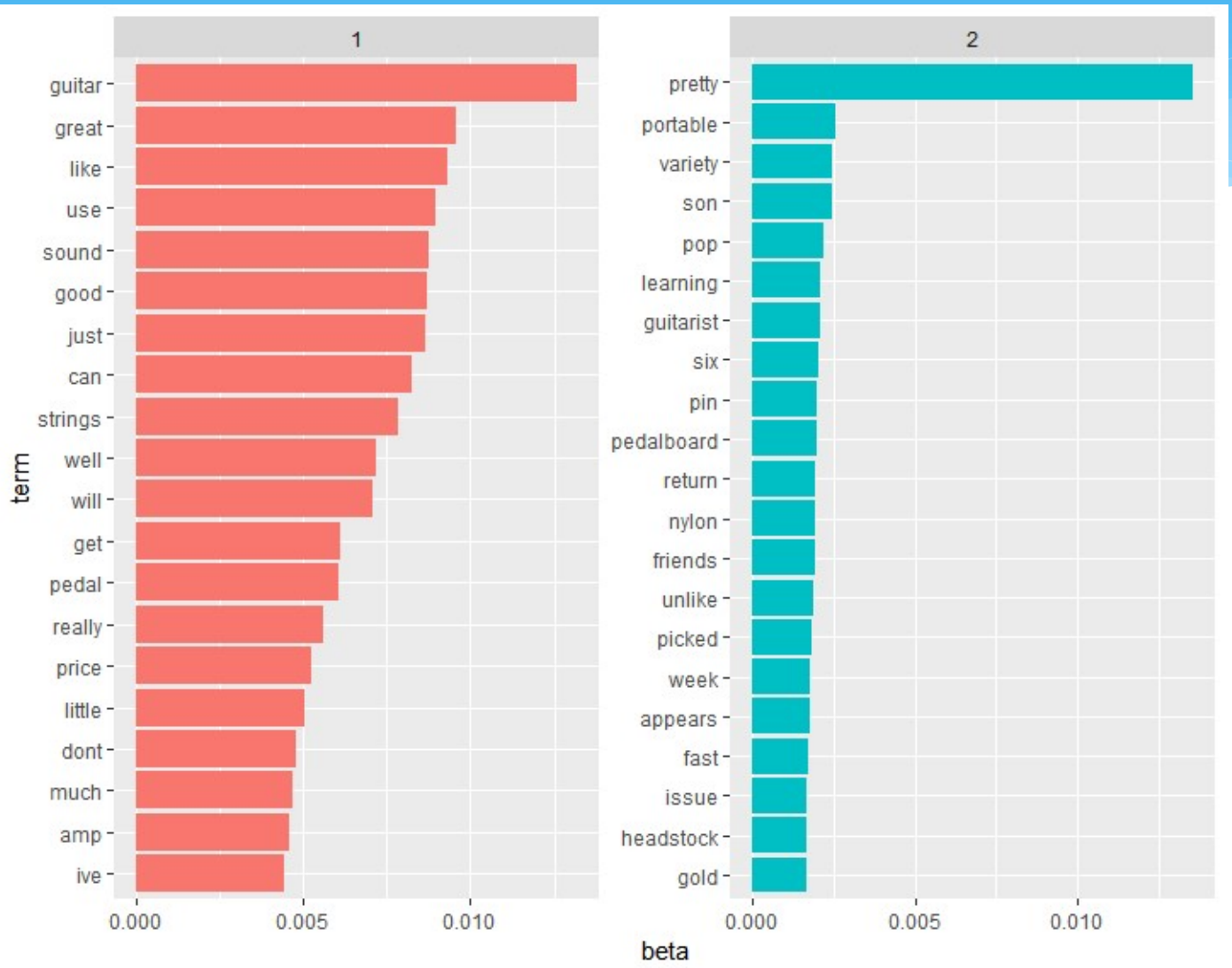
モデル適用の結果

各トピックの単語出現確率の上位20単語:オフィス製品



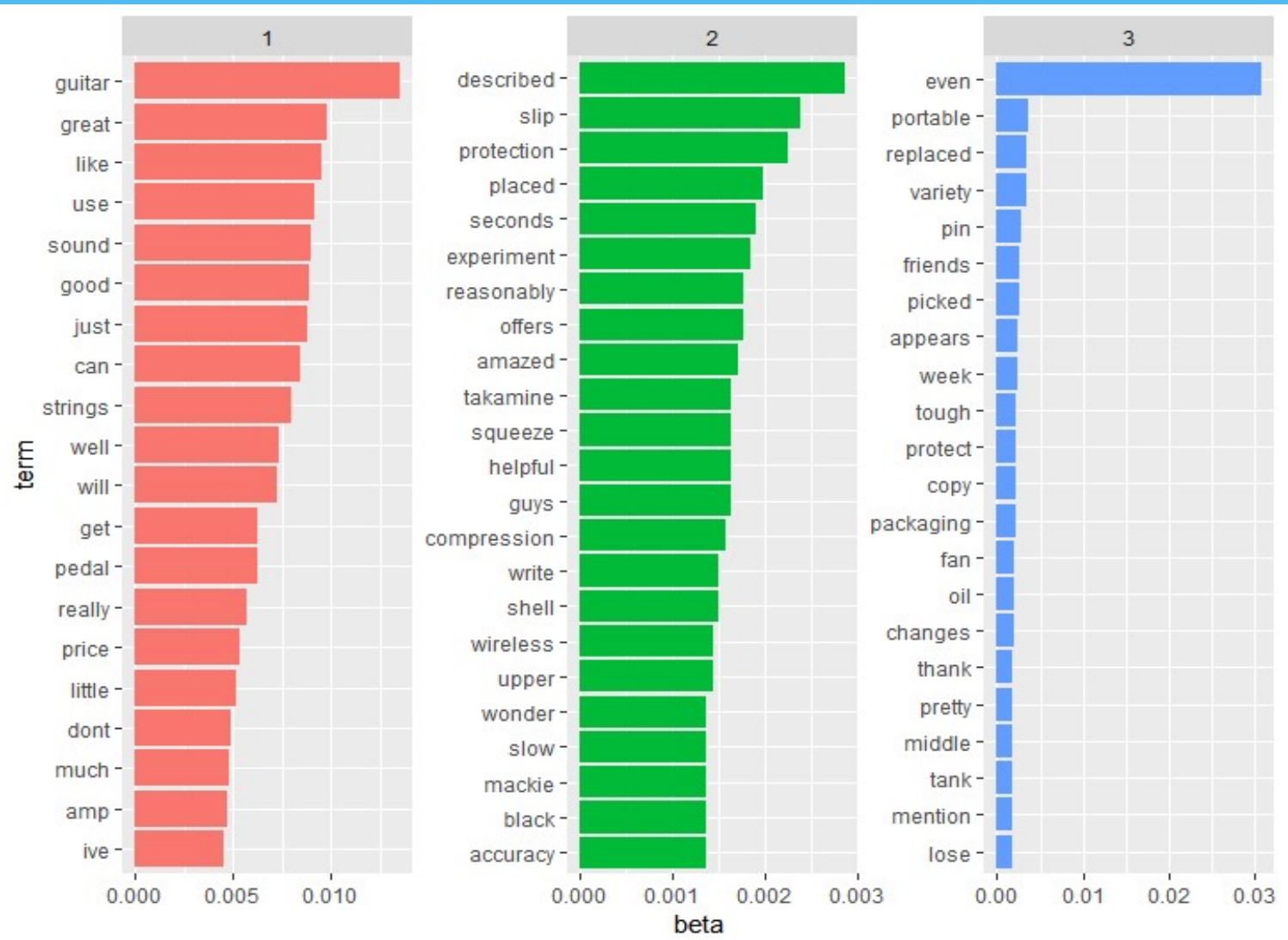
モデル適用の結果

各トピックの単語出現確率の上位20単語:楽器



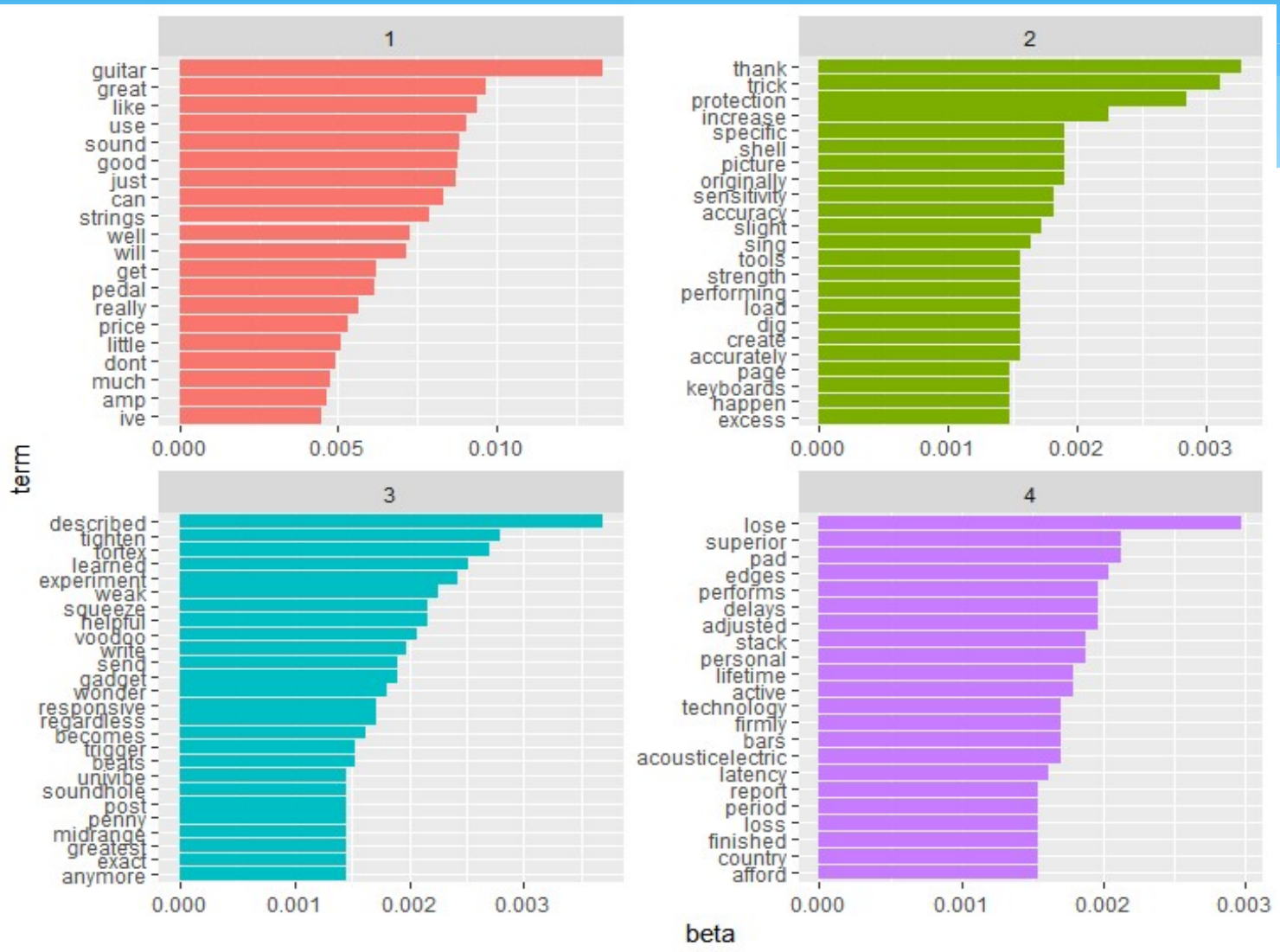
モデル適用の結果

各トピックの単語出現確率の上位20単語:楽器



モデル適用の結果

各トピックの単語出現確率の上位20単語:楽器



考察

- トピック数が2と考えられる理由
  1. データに対して合わないモデル
     - tf-idf, word to vectorも考えられる
  2. 前処理が不十分
     - 名詞と形容詞のみでやるべき？
  3. その他
- Perplexityがモデルの評価に合わない可能性
  1. パープレキシティではなく階層ディリクレ過程を用いる
- 極性をみるべきか
- ノイズ(やらせ)が混入

参考

1. Ldaのモデル選択におけるperplexityの評価(東京農工大学工学部情報工学科2年 森尾 学 2016/02/01)
2. トピックモデルによる統計的潜在意味解析(奥村2013)
3. 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(月岡2013et al)