映像アーカイブと Folksonomy

—— Topic Models / Latent Dirichlet Allocation の応用 ——

古　賀　　豊

1　映像アーカイブにおける双方向的な情報

人々は，映像アーカイブの映像を見て，さまざまな意見や感想を抱く。また，人によっては，その映像に関して，有益な情報（メタデータ）を持っていることもある。このような映像の利用者が生み出す情報も，映像アーカイブの重要な要素となるはずである。オンライン・ネットワーク上に構築される今日の映像アーカイブでは，利用者からの情報を収集・集積する技術的な方策は，すでに整備されている。映像アーカイブは，そこに収録される映像とともに，それを利用する人が生み出す情報——双方向的に生成されていく多様な情報——をあわせて集積することにより，より豊かなものとなっていくはずである。

一方，このような利用者が生み出すさまざまな情報をさして，従来，CGM (Consumer-Generated Media), UGC (User-Generated Contents), social tagging, 集合知といったさまざまな用語が使われてきた。このような用語（ここでは，とりあえず，folksonomy (Vander 2007) という語で代表させておく）で呼ばれる状況は，これまでにない新しい事態であるがゆえに，解決しなければならない問題を抱えている。つまり，さまざまな分野で大量の情報が溢れる今日的な状況においては，そのような情報を適切に分類・整理し，また，類似した（類似していない）文書を見つけ，あるいは，そのような情報が持っているテーマ・論点を把握するといった用途に役立つ道具が必要となる。

そのような道具の一つとして，ここで採り上げるのが，Topic Models——特に，その代表的な技法である LDA (Latent Dirichlet Allocation) である。

2 分析技法の検討：Topic Models / LDA (Latent Dirichlet Allocation)

2-1 Topic Models とは

Topic Models は，膨大な情報に対処するための分析技法として，主に，自然言語処理／機械学習の分野で開発されてきたものである。

まず，この領域のきっかけとなり，かつ，よく知られているのが，LSA/LSI (Latent Semantic Analysis/Indexing) という技法である (Deerwester et al. 1990; Landauer et al. 1998)。その後，この LSA/LSI に確率の概念を導入した PLSA/PLSI (Probabilistic Latent Semantic Analysis/Indexing) が発案され (Hofmann 1999)，それを，洗練・一般化させたものが LDA であると言うことができる (Blei, Ng and Jordan 2003; Blei 2012)。

今日では，この LDA が Topic Models と呼ばれる技法の代表的，かつ，基礎的なものと考えられている (石黒勝彦 2013; 持橋大地 2013; 佐藤一誠 2012)。

LDA とは，乱暴に要約すれば，次のように言うことができる。まず，「トピック」と呼ばれる潜在的なテーマを想定し，そのトピックに基づき，各文書が生成されると考える（その場合，1 つの文書に複数のトピックが存在することも想定される)。そして，そのトピックが，対象となる文書，および，その文書内で用いられる語に，どの程度，関わるのかを確率的に推定する技法である。

2-2 さまざまな Topic Models

前述の通り，LDA は，Topic Models の中でもっとも基礎的な技法と考えられているが，その展開や拡張の方向について，簡単に整理しておきたい。

Topic Models の展開・拡張を考える上で，大きく分類すれば，次の 3 つに分けることができる。

- トピックの扱いに関するもの
- 文書の扱いに関するもの
- 語の扱いに関するもの

2-2-1 トピックの扱いに関するもの

まず，トピックの扱いに関するものであるが，基本的な LDA モデルは，各トピックは独立であることを仮定しており，また，分析前の時点で，その数を決定し

ておく必要がある。

このような問題点への対応として，データからトピック数を推定し，さらに，トピックの階層構造も推定可能とするモデル (Blei, Griffiths and Jordan 2010) や，各トピック間の相関関係を想定するモデル (Blei and Lafferty 2007) が考案されている。

2-2-2 文書の扱いに関するもの

次に，文書の扱いに関してであるが，基本的な LDA モデルでは，文書間の関係を考慮しないが，文書間の関係を取り入れたモデルもいくつか考案されている。

まずあげられるのが，文書の時系列的な構造を考慮に入れた Dynamic Topic Model というモデル (Blei and Lafferty 2006) である。また，文書の中には，学術論文（引用）や，html 文書など，それぞれの文書同士でリンク構造を持つものが存在する。そのような文書構造を扱うものとして，Relational Topic Model (Chang and Blei 2010) が考案されている。

2-2-3 語の扱いに関するもの

最後に，語の扱いに関してであるが，LDA では，LSA などと同様に，その入力データとして，文書に含まれる語の順序情報／文法的情報を考慮に入れない。いわゆる “bag of words” と呼ばれるものであるが，一方，人が行う実際のテキストの解釈においては，当然ながら，語順や文法的情報は重要な要素となってくる。

これは，自然言語処理として見ると，大きな問題点であるが，一方で，語の順序情報／文法的情報を対象としない利点も指摘されている。つまり，文法的情報を考慮に入れないということは，多言語に容易に拡張可能であり，さらには，対象として，もはや語に限られる必要はなく，画像処理，音声処理の分野への応用も行われつつある。

3 分析の実際：Wikipedia「明治・幕末期の写真家」記事の分析

3-1 利用するパッケージ

ここでは，LDA を用いた分析の実際の例を示すことにしたい。

まず，LDA を用いた分析を行うために，比較的，容易に利用できるアプリケーション／パッケージとして，次のものをあげることができる。

- R (R Core Team 2014) 上で動作する lda パッケージ (Chang 2012)
- R (ibid.) 上で動作する topicmodels パッケージ (Grün and Hornik 2011)
- Python (Python Software Foundation 1990-2015) 上で動作する gensim パッケージ (Řehůřek and Sojka 2010)

ここでは，R の lda パッケージを利用して分析を行うことにする。

3-2 分析対象

今回，採り上げる分析対象は，Wikipedia 日本語版の「明治・幕末期の写真家」というカテゴリーに含まれている 41 記事（図表 1 を参照）である。

まず，各記事の内容（文章）に対して，形態素解析エンジンである mecab (Kudo and NTT 2008) を利用して，形態素解析を行い，その結果から，下記の語を入力データとして採用した（つまり，接続詞，副詞，連体詞，助動詞や助詞などは省かれていることに注意）。

- 品詞分類が「名詞」である語のうち，品詞細分類が「自立」および「固有名詞」である語
- 品詞分類が「形容詞」である語のうち，品詞細分類が「自立」である語
- 品詞分類が「動詞」である語のうち，品詞細分類が「一般」，「形容動詞語幹」，「サ変接続」，のいずれかである語
- 上記の語のうち，2 文字以上の語（1 文字の語は，解釈が困難な場合があるため，除外）

その結果，得られた入力データの総語数は 10010，語数は 1564 であった。このうち，出現頻度の高い語を，その順に示したものが，図表 2 である。

3-3 分析結果 (1)

LDA では，事前にトピック数を決定する必要があるため，ここでは，トピック数を 5 として，分析を実行した。

結果として得られたデータのうち，トピック内で上位に位置する語をトピック毎に示したものが，図表 3〜5 である。これらは，各トピックの意味内容を反映し

たものとみなすことができる。

これを見ると，例えば，Topic 1 の「箱館/函館」,「領事」,「ロシア」といった語は，ほぼ Topic 1 にしか出現しておらず，この Topic の意味内容を示すものとして考えることができる。

一方，Topic 2 の「写真」や「する」といった語は，ほかの Topic においても上位に位置しており，つまり，複数のトピックにわたって使用されている語であることがわかる。(Topic 内の順位の決定に，単純に各語に割り当てられた確率を用いると，出現頻度の高い語が上位に位置してしまうため，ここでは図表 5 の注記に示した数値を用いている。この式の機能・意図については，Blei and Lafferty, 2009, p.75 を参照のこと。)

3-4 分析結果 (2)

次に，各文書でもっとも比率の高いトピックを，その文書が属するトピックとみなし，トピック毎に，その比率の多い順に並べたものが，図表 6〜9 である。

これを見ると,「ヨシフ・ゴシケーヴィチ」の記事はほぼ Topic 1 のみで構成されており，同様に,「アドルフォ・ファルサーリ」の記事はほぼ Topic 2 のみ，そして,「金丸源三」,「淡海槐堂」,「島隆」,「中島仰山」,「鈴木真一」,「島霞谷」,「横山松三郎」といった記事はほぼ Topic 5 のみで構成されていることがわかる。

一方,「市来四郎」の記事は Topic 1 と Topic 5,「玉村康三郎」の記事は Topic 2 と Topic 5 というように，2 つの Topic で構成されているものもあれば,「下岡蓮杖」の記事のように，すべての Topic に関わっているものも存在していることがわかる。

3-5 分析の総括

このように各々の語や文書が 1 つの Topic のみにグループ分けされるわけではないという点は，従来，よく用いられてきたクラスター分析などの分類方法とは異なる LDA の特徴であり，言い換えれば，より多面的なデータの見方が可能になる技法と言うことができる。

また，ここでは簡単な実例を示したにすぎないが，LDA をはじめとするここであげた技法が，特にその真価を発揮するのは，人力では対応が難しい大規模デー

タを対象としたときであり，今後，その大規模データを対象に，本格的な分析を試みる予定である。

図表 1　分析対象となる記事項目一覧

- ヨシフ・ゴシケーヴィチ
- ウィリアム・K・バートン
- エイベル・ガウワー
- 市来四郎
- 大鳥圭介
- アドルフォ・ファルサーリ
- ハーバート・ポンティング
- 鹿嶋清兵衛
- ライムント・フォン・シュティルフリート
- 玉村康三郎
- 日下部金兵衛
- 富重利平
- フェリーチェ・ベアト
- 亀井茲明
- ピエール・ロシエ
- オリン・フリーマン
- 内田九一
- 田本研造
- ヨハネス・ポンペ・ファン・メーデルフォールト
- ヤン・カレル・ファン・デン・ブルーク
- アントニウス・ボードウィン
- 堀江鍬次郎
- 前田玄造
- 川崎道民
- 古川俊平
- 上野彦馬
- 金丸源三
- 淡海槐堂
- 島隆
- 中島仰山
- 鈴木真一
- 島霞谷
- 横山松三郎
- 鵜飼玉川
- 江崎礼二
- 木津幸吉
- 下岡蓮杖
- 武林盛一
- 臼井秀三郎
- ジョン・ウィルソン
- 小川一真

（注記：項目の順番は，図表 6〜9 の掲載順と対応させてある。）

図表 2　出現頻度の高い語

語	出現頻度	総語数に占める比率 % ( ) 内は累積比率	10 を底とする対数表示 0　10　100　1000
する	896	8.9510 (　8.9510)	
写真	647	6.4635 (15.4146)	
明治	178	1.7782 (17.1928)	
撮影	166	1.6583 (18.8511)	
日本	159	1.5884 (20.4396)	
なる	157	1.5684 (22.0080)	
ある	87	0.8691 (22.8771)	
学ぶ	57	0.5694 (23.4466)	
横浜	56	0.5594 (24.0060)	
長崎	55	0.5495 (24.5554)	
ファルサーリ	54	0.5395 (25.0949)	
東京	48	0.4795 (25.5744)	
技術	41	0.4096 (25.9840)	
時代	40	0.3996 (26.3836)	
スタジオ	38	0.3796 (26.7632)	
医学	38	0.3796 (27.1429)	
外国	38	0.3796 (27.5225)	
脚注	37	0.3696 (27.8921)	
江戸	35	0.3497 (28.2418)	
経歴	33	0.3297 (28.5714)	
生まれる	33	0.3297 (28.9011)	
元年	33	0.3297 (29.2308)	
作品	32	0.3197 (29.5504)	
開業	31	0.3097 (29.8601)	
箱館	31	0.3097 (30.1698)	
オランダ	30	0.2997 (30.4695)	
研究	30	0.2997 (30.7692)	
鹿嶋清兵衛	30	0.2997 (31.0689)	
最初	29	0.2897 (31.3586)	
目次	29	0.2897 (31.6484)	
下岡蓮杖	28	0.2797 (31.9281)	
幕末	28	0.2797 (32.2078)	
文久	27	0.2697 (32.4775)	
安政	27	0.2697 (32.7473)	
日本人	27	0.2697 (33.0170)	
上野	27	0.2697 (33.2867)	
行う	27	0.2697 (33.5564)	

図表 3　各トピックにおける上位語と分布 (1)

Topic 1 の上位 7 語

語	topic	頻度	$\beta_{w,k}$
箱館	1	31	0.0232 (0.1435)
	2	0	0.0000 (0.0000)
	3	0	0.0000 (0.0000)
	4	0	0.0000 (0.0000)
	5	0	0.0000 (0.0000)
領事	1	20	0.0149 (0.0873)
	2	0	0.0000 (0.0000)
	3	0	0.0000 (0.0000)
	4	0	0.0000 (0.0000)
	5	0	0.0000 (0.0000)
ロシア	1	17	0.0127 (0.0726)
	2	0	0.0000 (0.0000)
	3	0	0.0000 (0.0000)
	4	0	0.0000 (0.0000)
	5	0	0.0000 (0.0000)
バートン	1	16	0.0119 (0.0677)
	2	0	0.0000 (0.0000)
	3	0	0.0000 (0.0000)
	4	0	0.0000 (0.0000)
	5	0	0.0000 (0.0000)
函館	1	17	0.0127 (0.0628)
	2	0	0.0000 (0.0000)
	3	1	0.0005 (0.0008)
	4	0	0.0000 (0.0000)
	5	0	0.0000 (0.0000)
東京	1	21	0.0157 (0.0590)
	2	1	0.0004 (0.0000)
	3	0	0.0000 (0.0000)
	4	0	0.0000 (0.0000)
	5	26	0.0106 (0.0356)
英国	1	13	0.0097 (0.0459)
	2	0	0.0000 (0.0000)
	3	1	0.0005 (0.0008)
	4	0	0.0000 (0.0000)
	5	0	0.0000 (0.0000)

Topic 2 の上位 7 語

語	topic	頻度	$\beta_{w,k}$
ファルサーリ	1	0	0.0000 (0.0000)
	2	54	0.0206 (0.1260)
	3	0	0.0000 (0.0000)
	4	0	0.0000 (0.0000)
	5	0	0.0000 (0.0000)
写真	1	1	0.0007 (-0.0026)
	2	232	0.0887 (0.1188)
	3	221	0.1027 (0.1528)
	4	20	0.0139 (-0.0071)
	5	173	0.0704 (0.0781)
横浜	1	2	0.0015 (0.0020)
	2	46	0.0176 (0.0671)
	3	0	0.0000 (0.0000)
	4	0	0.0000 (0.0000)
	5	8	0.0033 (0.0069)
鹿嶋清兵衛	1	0	0.0000 (0.0000)
	2	30	0.0115 (0.0646)
	3	0	0.0000 (0.0000)
	4	0	0.0000 (0.0000)
	5	0	0.0000 (0.0000)
する	1	110	0.0822 (-0.0011)
	2	327	0.1250 (0.0508)
	3	194	0.0901 (0.0072)
	4	85	0.0589 (-0.0204)
	5	180	0.0732 (-0.0094)
スタジオ	1	0	0.0000 (0.0000)
	2	29	0.0111 (0.0491)
	3	0	0.0000 (0.0000)
	4	0	0.0000 (0.0000)
	5	9	0.0037 (0.0122)
彩色	1	0	0.0000 (0.0000)
	2	19	0.0073 (0.0383)
	3	0	0.0000 (0.0000)
	4	0	0.0000 (0.0000)
	5	0	0.0000 (0.0000)

図表 4 各トピックにおける上位語と分布 (2)

Topic 3 の上位 7 語

語	topic	頻度	$\beta_{w,k}$
写真	1	1	0.0007 (-0.0026)
	2	232	0.0887 (0.1188)
	3	221	0.1027 (0.1528)
	4	20	0.0139 (-0.0071)
	5	173	0.0704 (0.0781)
撮影	1	0	0.0000 (0.0000)
	2	34	0.0130 (0.0203)
	3	99	0.0460 (0.1299)
	4	3	0.0021 (-0.0006)
	5	30	0.0122 (0.0183)
戦争	1	2	0.0015 (0.0040)
	2	0	0.0000 (0.0000)
	3	19	0.0088 (0.0391)
	4	0	0.0000 (0.0000)
	5	0	0.0000 (0.0000)
イギリス	1	1	0.0007 (0.0016)
	2	0	0.0000 (0.0000)
	3	18	0.0084 (0.0378)
	4	0	0.0000 (0.0000)
	5	0	0.0000 (0.0000)
熊本	1	0	0.0000 (0.0000)
	2	0	0.0000 (0.0000)
	3	14	0.0065 (0.0337)
	4	0	0.0000 (0.0000)
	5	0	0.0000 (0.0000)
明治	1	8	0.0060 (0.0117)
	2	0	0.0000 (0.0000)
	3	26	0.0121 (0.0322)
	4	0	0.0000 (0.0000)
	5	144	0.0586 (0.2485)
日下部金兵衛	1	0	0.0000 (0.0000)
	2	0	0.0000 (0.0000)
	3	13	0.0060 (0.0310)
	4	0	0.0000 (0.0000)
	5	0	0.0000 (0.0000)

Topic 4 の上位 7 語

語	topic	頻度	$\beta_{w,k}$
医学	1	0	0.0000 (0.0000)
	2	0	0.0000 (0.0000)
	3	0	0.0000 (0.0000)
	4	38	0.0263 (0.1658)
	5	0	0.0000 (0.0000)
長崎	1	0	0.0000 (0.0000)
	2	0	0.0000 (0.0000)
	3	7	0.0033 (0.0069)
	4	46	0.0319 (0.1406)
	5	2	0.0008 (0.0006)
オランダ	1	0	0.0000 (0.0000)
	2	0	0.0000 (0.0000)
	3	0	0.0000 (0.0000)
	4	30	0.0208 (0.1270)
	5	0	0.0000 (0.0000)
化学	1	0	0.0000 (0.0000)
	2	0	0.0000 (0.0000)
	3	0	0.0000 (0.0000)
	4	25	0.0173 (0.1033)
	5	0	0.0000 (0.0000)
伝習	1	0	0.0000 (0.0000)
	2	0	0.0000 (0.0000)
	3	0	0.0000 (0.0000)
	4	24	0.0166 (0.0986)
	5	0	0.0000 (0.0000)
幕末	1	0	0.0000 (0.0000)
	2	0	0.0000 (0.0000)
	3	0	0.0000 (0.0000)
	4	19	0.0132 (0.0601)
	5	9	0.0037 (0.0120)
医師	1	2	0.0015 (0.0027)
	2	0	0.0000 (0.0000)
	3	0	0.0000 (0.0000)
	4	18	0.0125 (0.0493)
	5	1	0.0004 (0.0002)

図表 5　各トピックにおける上位語と分布 (3)

Topic 5 の上位 7 語

語	topic	頻度	$\beta_{w,k}$
明治	1	8	0.0060 (0.0117)
	2	0	0.0000 (0.0000)
	3	26	0.0121 (0.0322)
	4	0	0.0000 (0.0000)
	5	144	0.0586 (0.2485)
写真	1	1	0.0007 (-0.0026)
	2	232	0.0887 (0.1188)
	3	221	0.1027 (0.1528)
	4	20	0.0139 (-0.0071)
	5	173	0.0704 (0.0781)
下岡 蓮杖	1	0	0.0000 (0.0000)
	2	0	0.0000 (0.0000)
	3	0	0.0000 (0.0000)
	4	0	0.0000 (0.0000)
	5	28	0.0114 (0.0641)
天保	1	0	0.0000 (0.0000)
	2	0	0.0000 (0.0000)
	3	0	0.0000 (0.0000)
	4	0	0.0000 (0.0000)
	5	20	0.0081 (0.0436)
学ぶ	1	0	0.0000 (0.0000)
	2	1	0.0004 (-0.0003)
	3	1	0.0005 (-0.0003)
	4	22	0.0152 (0.0444)
	5	33	0.0134 (0.0375)
仰山	1	0	0.0000 (0.0000)
	2	0	0.0000 (0.0000)
	3	0	0.0000 (0.0000)
	4	0	0.0000 (0.0000)
	5	17	0.0069 (0.0362)
文久	1	0	0.0000 (0.0000)
	2	0	0.0000 (0.0000)
	3	0	0.0000 (0.0000)
	4	6	0.0042 (0.0144)
	5	21	0.0085 (0.0358)

注記：$\beta_{w,k}$ 欄の（　）内の数値は，下記の式で算出されたものである。各語のトピック内の順位は，この値によって決められる。(Blei and Lafferty 2009, p.75; Chang 2012, Reference manual p.25)

$$\beta_{w,k}\left(\log\beta_{w,k}-\frac{1}{K}\sum_{k=1}^{K}\log\beta_{w,k}\right)$$

w は各単語，k は各トピック，K はトピックの総数を表す。

図表 6　文書毎のトピックの分布 (1)

	Topic 1 ⟶	
箱館/領事/ロシア/バートン/函館	289 (0.8353)	250 (0.7599)
ファルサーリ/写真/横浜/鹿嶋清兵衛	0 (0.0000)	34 (0.1033)
写真/撮影/戦争/イギリス/熊本	17 (0.0491)	45 (0.1368)
医学/長崎/オランダ/化学/伝習	31 (0.0896)	0 (0.0000)
明治/写真/下岡蓮杖/天保/学ぶ	9 (0.0260)	0 (0.0000)
	ゴシケーヴィチ	バートン

箱館/領事/ロシア/バートン/函館	124 (0.7337)	67 (0.5982)
ファルサーリ/写真/横浜/鹿嶋清兵衛	15 (0.0888)	0 (0.0000)
写真/撮影/戦争/イギリス/熊本	0 (0.0000)	0 (0.0000)
医学/長崎/オランダ/化学/伝習	30 (0.1775)	1 (0.0089)
明治/写真/下岡蓮杖/天保/学ぶ	0 (0.0000)	44 (0.3929)
	エイベル・ガウワー	市来四郎

	Topic 2 ⟶	
箱館/領事/ロシア/バートン/函館	226 (0.4612)	20 (0.0177)
ファルサーリ/写真/横浜/鹿嶋清兵衛	0 (0.0000)	1030 (0.9107)
写真/撮影/戦争/イギリス/熊本	109 (0.2224)	68 (0.0601)
医学/長崎/オランダ/化学/伝習	79 (0.1612)	13 (0.0115)
明治/写真/下岡蓮杖/天保/学ぶ	76 (0.1551)	0 (0.0000)
	大鳥圭介	ファルサーリ

箱館/領事/ロシア/バートン/函館	43 (0.1000)	37 (0.0555)
ファルサーリ/写真/横浜/鹿嶋清兵衛	287 (0.6674)	442 (0.6627)
写真/撮影/戦争/イギリス/熊本	100 (0.2326)	10 (0.0150)
医学/長崎/オランダ/化学/伝習	0 (0.0000)	0 (0.0000)
明治/写真/下岡蓮杖/天保/学ぶ	0 (0.0000)	178 (0.2669)
	ポンティング	鹿嶋清兵衛

箱館/領事/ロシア/バートン/函館	0 (0.0000)	0 (0.0000)
ファルサーリ/写真/横浜/鹿嶋清兵衛	85 (0.6159)	36 (0.5625)
写真/撮影/戦争/イギリス/熊本	47 (0.3406)	0 (0.0000)
医学/長崎/オランダ/化学/伝習	6 (0.0435)	0 (0.0000)
明治/写真/下岡蓮杖/天保/学ぶ	0 (0.0000)	28 (0.4375)
	シュティルフリート	玉村康三郎

	Topic 3 ⟶	
箱館/領事/ロシア/バートン/函館	0 (0.0000)	0 (0.0000)
ファルサーリ/写真/横浜/鹿嶋清兵衛	71 (0.4765)	0 (0.0000)
写真/撮影/戦争/イギリス/熊本	52 (0.3490)	274 (0.8012)
医学/長崎/オランダ/化学/伝習	0 (0.0000)	0 (0.0000)
明治/写真/下岡蓮杖/天保/学ぶ	26 (0.1745)	68 (0.1988)
	日下部金兵衛	富重利平

図表 7　文書毎のトピックの分布 (2)

箱館/領事/ロシア/バートン/函館	1 (0.0012)	23 (0.1117)
ファルサーリ/写真/横浜/鹿嶋清兵衛	256 (0.2946)	0 (0.0000)
写真/撮影/戦争/イギリス/熊本	608 (0.6997)	133 (0.6456)
医学/長崎/オランダ/化学/伝習	4 (0.0046)	0 (0.0000)
明治/写真/下岡蓮杖/天保/学ぶ	0 (0.0000)	50 (0.2427)
	フェリーチェ・ベアト	亀井茲明
箱館/領事/ロシア/バートン/函館	12 (0.1081)	0 (0.0000)
ファルサーリ/写真/横浜/鹿嶋清兵衛	0 (0.0000)	38 (0.3762)
写真/撮影/戦争/イギリス/熊本	60 (0.5405)	45 (0.4455)
医学/長崎/オランダ/化学/伝習	39 (0.3514)	0 (0.0000)
明治/写真/下岡蓮杖/天保/学ぶ	0 (0.0000)	18 (0.1782)
	ピエール・ロシエ	オリン・フリーマン
箱館/領事/ロシア/バートン/函館	0 (0.0000)	60 (0.2500)
ファルサーリ/写真/横浜/鹿嶋清兵衛	0 (0.0000)	0 (0.0000)
写真/撮影/戦争/イギリス/熊本	54 (0.4154)	99 (0.4125)
医学/長崎/オランダ/化学/伝習	22 (0.1692)	19 (0.0792)
明治/写真/下岡蓮杖/天保/学ぶ	54 (0.4154)	62 (0.2583)
	内田九一	田本研造
	Topic 4 ⟶	
箱館/領事/ロシア/バートン/函館	9 (0.0164)	0 (0.0000)
ファルサーリ/写真/横浜/鹿嶋清兵衛	0 (0.0000)	18 (0.1538)
写真/撮影/戦争/イギリス/熊本	31 (0.0565)	0 (0.0000)
医学/長崎/オランダ/化学/伝習	509 (0.9271)	99 (0.8462)
明治/写真/下岡蓮杖/天保/学ぶ	0 (0.0000)	0 (0.0000)
	メーデルフォールト	ブルーク
箱館/領事/ロシア/バートン/函館	22 (0.1350)	0 (0.0000)
ファルサーリ/写真/横浜/鹿嶋清兵衛	21 (0.1288)	4 (0.0339)
写真/撮影/戦争/イギリス/熊本	0 (0.0000)	31 (0.2627)
医学/長崎/オランダ/化学/伝習	120 (0.7362)	83 (0.7034)
明治/写真/下岡蓮杖/天保/学ぶ	0 (0.0000)	0 (0.0000)
	ボードウィン	堀江鍬次郎
箱館/領事/ロシア/バートン/函館	0 (0.0000)	0 (0.0000)
ファルサーリ/写真/横浜/鹿嶋清兵衛	12 (0.1071)	0 (0.0000)
写真/撮影/戦争/イギリス/熊本	27 (0.2411)	5 (0.0833)
医学/長崎/オランダ/化学/伝習	70 (0.6250)	31 (0.5167)
明治/写真/下岡蓮杖/天保/学ぶ	3 (0.0268)	24 (0.4000)
	前田玄造	川崎道民

図表 8　文書毎のトピックの分布 (3)

箱館/領事/ロシア/バートン/函館	30 (0.1172)	0 (0.0000)
ファルサーリ/写真/横浜/鹿嶋清兵衛	35 (0.1367)	0 (0.0000)
写真/撮影/戦争/イギリス/熊本	78 (0.3047)	65 (0.3250)
医学/長崎/オランダ/化学/伝習	113 (0.4414)	75 (0.3750)
明治/写真/下岡蓮杖/天保/学ぶ	0 (0.0000)	60 (0.3000)
	古川俊平	上野彦馬
	Topic 5 ⟶	
箱館/領事/ロシア/バートン/函館	0 (0.0000)	12 (0.0909)
ファルサーリ/写真/横浜/鹿嶋清兵衛	0 (0.0000)	0 (0.0000)
写真/撮影/戦争/イギリス/熊本	0 (0.0000)	1 (0.0076)
医学/長崎/オランダ/化学/伝習	1 (0.0263)	0 (0.0000)
明治/写真/下岡蓮杖/天保/学ぶ	37 (0.9737)	119 (0.9015)
	金丸源三	淡海槐堂
箱館/領事/ロシア/バートン/函館	0 (0.0000)	2 (0.0060)
ファルサーリ/写真/横浜/鹿嶋清兵衛	4 (0.0645)	34 (0.1015)
写真/撮影/戦争/イギリス/熊本	3 (0.0484)	0 (0.0000)
医学/長崎/オランダ/化学/伝習	0 (0.0000)	7 (0.0209)
明治/写真/下岡蓮杖/天保/学ぶ	55 (0.8871)	292 (0.8716)
	島隆	中島仰山
箱館/領事/ロシア/バートン/函館	0 (0.0000)	0 (0.0000)
ファルサーリ/写真/横浜/鹿嶋清兵衛	0 (0.0000)	5 (0.0685)
写真/撮影/戦争/イギリス/熊本	0 (0.0000)	0 (0.0000)
医学/長崎/オランダ/化学/伝習	13 (0.1429)	7 (0.0959)
明治/写真/下岡蓮杖/天保/学ぶ	78 (0.8571)	61 (0.8356)
	鈴木真一	島霞谷
箱館/領事/ロシア/バートン/函館	30 (0.0664)	0 (0.0000)
ファルサーリ/写真/横浜/鹿嶋清兵衛	8 (0.0177)	0 (0.0000)
写真/撮影/戦争/イギリス/熊本	45 (0.0996)	17 (0.2024)
医学/長崎/オランダ/化学/伝習	0 (0.0000)	0 (0.0000)
明治/写真/下岡蓮杖/天保/学ぶ	369 (0.8164)	67 (0.7976)
	横山松三郎	鵜飼玉川
箱館/領事/ロシア/バートン/函館	0 (0.0000)	24 (0.3380)
ファルサーリ/写真/横浜/鹿嶋清兵衛	0 (0.0000)	0 (0.0000)
写真/撮影/戦争/イギリス/熊本	34 (0.1921)	0 (0.0000)
医学/長崎/オランダ/化学/伝習	19 (0.1073)	0 (0.0000)
明治/写真/下岡蓮杖/天保/学ぶ	124 (0.7006)	47 (0.6620)
	江崎礼二	木津幸吉

図表 9　文書毎のトピックの分布 (4)

	下岡蓮杖	武林盛一
箱館/領事/ロシア/バートン/函館	25 (0.0504)	13 (0.2031)
ファルサーリ/写真/横浜/鹿嶋清兵衛	81 (0.1633)	0 (0.0000)
写真/撮影/戦争/イギリス/熊本	38 (0.0766)	11 (0.1719)
医学/長崎/オランダ/化学/伝習	41 (0.0827)	0 (0.0000)
明治/写真/下岡蓮杖/天保/学ぶ	311 (0.6270)	40 (0.6250)

	臼井秀三郎	ジョン・ウィルソン
箱館/領事/ロシア/バートン/函館	3 (0.0435)	15 (0.1685)
ファルサーリ/写真/横浜/鹿嶋清兵衛	22 (0.3188)	22 (0.2472)
写真/撮影/戦争/イギリス/熊本	0 (0.0000)	0 (0.0000)
医学/長崎/オランダ/化学/伝習	3 (0.0435)	9 (0.1011)
明治/写真/下岡蓮杖/天保/学ぶ	41 (0.5942)	43 (0.4831)

	小川一真
箱館/領事/ロシア/バートン/函館	2 (0.0112)
ファルサーリ/写真/横浜/鹿嶋清兵衛	57 (0.3202)
写真/撮影/戦争/イギリス/熊本	45 (0.2528)
医学/長崎/オランダ/化学/伝習	0 (0.0000)
明治/写真/下岡蓮杖/天保/学ぶ	74 (0.4157)

参考文献

Blei, D. M., A. Ng and M. Jordan, 2003. “Latent Dirichlet allocation”, *Journal of Machine Learning Research*, vol. 3, pp.993–1022.

Blei, D. M. and J. D. Lafferty, 2006. “Dynamic topic models”, *International Conference on Machine Learning*, ACM.

Blei, D. M. and J. D. Lafferty, 2007. “correlated topic model of Science”, *Annals of Applied Statistics*, 1(1), pp.17–35.

Blei, D. M. and J. D. Lafferty, 2009. “TOPIC MODELS”, in A. Srivastava and M. Sahami, eds., *Text Mining: Classification, Clustering, and Applications*, CRC Press, pp.71–95.

Blei, D. M., T. Griffiths and M. Jordan, 2010. “The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies” *Journal of the ACM*, vol. 57, no. 2, pp.1–30.

Blei, D. M., 2012. “Probabilistic Topic Models”, *Communications of the ACM*, 55(4), pp.77–84.

Chang, J. and D. M. Blei, 2009. “Relational Topic Models for Document Networks”, *Artificial intelligence and statistics*. pp.81–88.

Chang, J. and D. M. Blei, 2010. “Hierarchical relational models for document networks”, *Annals of Applied Statistics*, 4(1).

Chang, J., 2012. “lda: Collapsed Gibbs Sampling Methods for Topic Models”, R package version 1.3.2. (http://CRAN.R-project.org/package=lda)

Deerwester, S., et al., 1990. “Indexing by latent semantic analysis”, *Journal of the American Society for Information Science*, 41, pp.391–407.

Grün B. and K. Hornik, 2011. “topicmodels: An R Package for Fitting Topic Models”, *Journal of Statistical Software*, 40(13), pp.1–30. (http://www.jstatsoft.org/v40/i13/)

Hofmann, T., 1999. “Probabilistic Latent Semantic Analysis”, Uncertainity in Articial Intelligence, UAI’99, Stockholm.

Kudo, T and NTT, 2008. “MeCab: Yet Another Part-of-Speech and Morphological Analyzer”, version 0.996.

Landauer, T. K., et al., 1998. “Introduction to Latent Semantic Analysis”, *Discourse Processes*, 25, pp.259–284.

The Python Software Foundation, 1990–2015, *Python*, https://www.python.org/

R Core Team, 2014. *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing. (http://www.R-project.org)

Řehůřek, Radim and Petr Sojka, 2010. “Software Framework for Topic Modelling with Large Corpora”, *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp.45–50. (http://radimrehurek.com/gensim/)

Vander Wal, T., 2007. “Folksonomy Coinage and Definition”, (http://vanderwal.net/folksonomy.html)

石黒勝彦, 2013. 「統計数理研究所 H24 年度公開講座『確率的トピックモデル』2 日目講義資料」(http://www.ism.ac.jp/ daichi/lectures/ISM-2012-TopicModels-ishiguro.pdf)

持橋大地, 2013. 「統計数理研究所 H24 年度公開講座『確率的トピックモデル』1 日目講義資料」(http://www.ism.ac.jp/ daichi/lectures/ISM-2012-TopicModels-daichi.pdf)

佐藤一誠, 2012. 「私のブックマーク: Latent Topic Model (潜在的トピックモデル)」, 『人工知能学会誌』, Vol.27, No.3. (http://www.ai-gakkai.or.jp/my-bookmark_vol27-no3/)