INTERNET STUDIES LAB

NATIONAL RESEARCH UNIVERSITY

Восстановление плотности многомерных распределений в виде смеси распределений на основе тематического моделирования

Sergei Koltcov

Outline of presentation – part 1

- Введение в проблематику тематического моделирования.
- Обзор математических принципов тематического моделирования:
  А. Метод максимизации функции правдоподобия при помощи E-M алгоритма: Модель ‘Probabilistic Latent Semantic Analysis, PLSA’. Модель ‘Latent Dirichlet allocation’.
  Б. Метод оценки математического ожидания при помощи Gibbs sampling.
- Краткая характеристика возможностей и ограничений тематического моделирования
- Проблема выбора числа тем.
- Проблема стабильности тематического моделирования. Социологический аспект проблематики стабильности ТМ. Математические основания проблематики ТМ

Outline of presentation – part 2

Социологический анализ результатов ТМ.

- **1**. Discovering Health Topics in Social Media Using Topic Models.
- **2**. Cross-Lingual Topic Alignment in Time Series Japanese / Chinese News.
- **3**. Cross-Cultural Analysis of Blogs and Forums with Mixed-Collection Topic Models.
- **4**. LiveJournal Libra!': The political blogosphere and voting preferences in Russia in 2011–2012.
- **5**. Мониторинг мэжэтнической напряженности по материалам соц. сетей. (ЛИНИС 2015-2017)

Введение в проблематику тематического моделирования

Социология: «народная» повестка дня и онлайновое общественное мнение на основе пользовательского контента в социальных сетях. Соотнесение его с демографическими характеристиками, событиями и внешними повестками (напр., в СМИ)

Исследования СМИ: тематическая структура медиа-контента. Различие между изданиями, цензура, динамика роста и упадка внимания в новостным поводам.

Наукометрия: изучение структуры научного знания и его эволюции. Составление обзоров литературы.

Семейство моделей в области тематического моделирования

Year/Type	Basic PDPTMs	Inter-Document Correlated PDPTMs	Intra-Document Correlated PDPTMs	Temporal PDPTMs	Supervised PDPTMs
1999	PLSA				
2000					
2001		**PLSA**→A Joint Probabilistic Model			
2002	A probabilistic Approach				
2003	LDA, A Topic Model				**LDA** → Corr-LDA
2004	Discrete PCA	**LDA** →Mixed Membership Models, **LDA** →Author-Topic Model, **LDA**→ **Author-Topic Model** →ART			
2005			**LDA** → HMM-LDA		**LDA** →LLDA
2006		**LDA** → PAM, **LDA** → CTM, **LDA** →Statistical Entity-Topic Models	Bigram Topic Model, **PLSA** → CPLSA	**LDA** → TOT, **LDA** → DTM, **(PAM, TOT)** → Continuous Time Model	
2007		**LDA** → GWN-LDA, **LDA** → Citation Influence Model	**LDA** →HTMM, **LDA** → TNG	MTTM	**LDA** → sLDA
2008		**LDA** →LTHM, **LDA** → **Author-Topic Model** → ACT Model **(A Joint Probabilistic Model, LDA)** → Link-PLSA-LDA		**LDA** → **DTM** →cDTM	
2009		**LDA** → Generalized LDA, **LDA** → **Author-Topic Model** → **ACT** à Generalized ACT		**LDA** → **Author-Topic Model** → TAT	

Основные направления тематического моделирования

Методом максимизации функции правдоподобия

Probabilistic latent semantic analysis

Hofmann T. Probabilistic latent semantic analysis. 1999

Latent Dirichlet allocation

Blei D M, Latent Dirichlet allocation. 2003

Регуляризация TM на основе PLSA, Воронцов К, 2008, **BIGARTM**

Метод Монте – Карло

Gibbs sampling

Griffiths T L, Steyvers M. Finding scientific topics. 2004

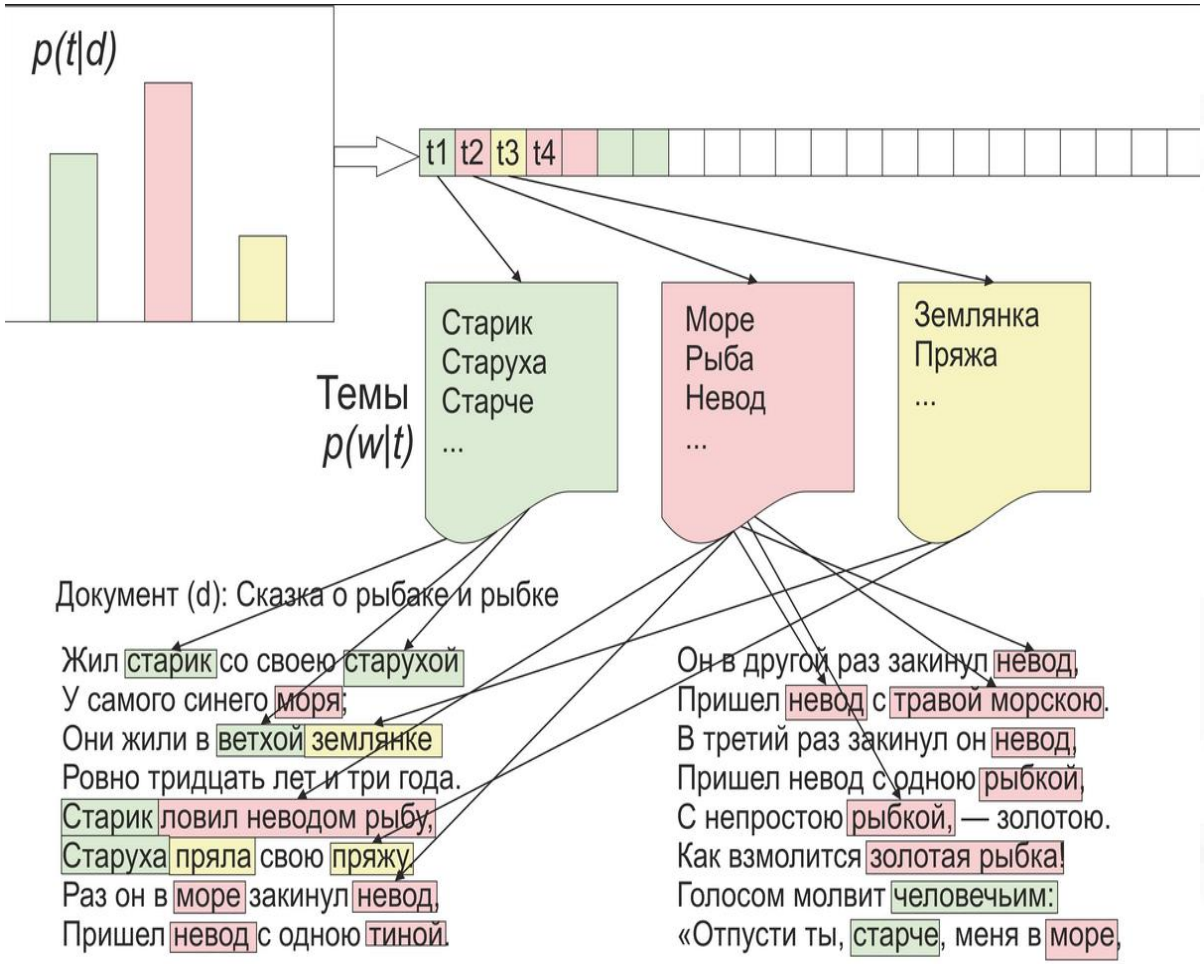
Регуляризация TM на основе Gibbs sampling

Newman D., Improving topic coherence with regularized topic models, 2011

LINIS: Stable Topic Modeling with Local Density Regularization, 2016

TOPICMINER

Вероятностная постановка задачи тематического моделирования



Тематическая модель (topic model) — модель коллекции текстовых документов, которая определяет, к каким темам относится каждый документ коллекции. Алгоритм построения тематической модели получает на входе коллекцию текстовых документов. На выходе для каждого документа выдаётся числовой вектор, составленный из оценок степени принадлежности данного документа каждой из тем.

Вероятностная постановка задачи тематического моделирования

Пусть D — коллекция текстовых документов, W — множество (словарь) всех уникальных слов. Каждый документ $d \in D$ представляет собой последовательность терминов $w_1,...,w_{nd}$ из словаря W. Мы также предполагаем, что существует конечное множество тем T, и каждое вхождение слова w в документ d связано с некоторой темой $t \in T$. Коллекция документов рассматривается как случайная и независимая выборка троек (w_i,d_i,t_i), $i = 1,...,n$ из дискретного распределения $p(w,d,t)$ на конечном вероятностном пространстве $\mathbf{W \times D \times T}$. Слова w и документы d являются наблюдаемыми переменными, тема $t \in T$ является латентной (скрытой) переменной. Предполагается, что порядок терминов в документах не важен для выявления тематики (bag of words). Порядок документов в коллекции также не имеет значения.

Вероятностная постановка задачи тематического моделирования

Вероятность p(w|d) появления терминов w в документах d можно выразить через произведение распределений p(w|t) и p(t|d). Согласно формуле полной вероятности и гипотезе условной независимости имеем следующее выражение:

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d) = \sum_{t \in T} \phi_{wt}\theta_{td}$$

где p(w| t) – распределения слов по темам, p(t | d) – распределение документов по темам. Построить тематическую модель данных означает решить обратную задачу, в которой необходимо найти множество скрытых тем T, т.е. множество одномерных условных распределений p(w|t) ≡ φ(w,t) для каждой темы t, которые составляют матрицу **Φ** (распределение слов по темам) и множество одномерных распределений p(t|d) ≡ θ(t,d) (матрица **Θ**, распределение документов по темам) для каждого документа d на основе наблюдаемых переменных d и w.

Распределение слов по темам

Words with high probability

	11	12	13	14	15	16
1	поддерживать: 0.004491	мировоззрение: 0.010559	мир: 0.007165	еврей: 0.016119	компания: 0.006241	инвалид: 0.017142
2	фотография: 0.003396	религия: 0.010006	известный: 0.006156	гитлер: 0.008104	больная: 0.006241	февраль: 0.015248
3	точка: 0.003396	цивилизация: 0.010006	список: 0.005147	еврейский: 0.006323	заболевание: 0.005470	полицейский: 0.013354
4	япония: 0.002300	сущность: 0.007795	премия: 0.005147	ротшильд: 0.005432	ну: 0.004700	ла-пас: 0.013354
5	массимо: 0.002300	информация: 0.007795	из: 0.005147	мирова: 0.005432	врач: 0.004700	боливия: 0.011459
6	развертываться: 0.002300	развитие: 0.007795	нобелевский: 0.005147	а: 0.004542	таблетка: 0.004700	на: 0.010512
7	избыток: 0.002300	весь: 0.007242	кандидат: 0.004138	сша: 0.004542	пример: 0.003929	reuters: 0.009565
8	уникальный: 0.002300	существовать: 0.007242	включать: 0.004138	сталин: 0.004542	диета: 0.003929	mercado: 0.009565
9	барби: 0.002300	вид: 0.006136	комитет: 0.003128	война: 0.004542	препарат: 0.003929	david: 0.009565
10	манекен: 0.002300	парадигма: 0.006136	средства: 0.003128	россия: 0.004542	сахар: 0.003929	фотография: 0.007671
11	армия: 0.002300	вы: 0.005584	лист: 0.003128	мировая: 0.004542	профилактика: 0.003929	путь: 0.006724
12	род: 0.002300	мир: 0.005584	номинантов: 0.003128	клан: 0.003651	пациент: 0.003929	набрасываться: 0.005777
13	отсюда: 0.002300	теория: 0.005584	лонг: 0.003128	советский: 0.003651	медицина: 0.003929	костыль: 0.004830
14	египет: 0.001205	доказывать: 0.004478	мэннинг: 0.002119	высота: 0.003651	лечение: 0.003929	перегораживать: 0.003883
15	разогревать: 0.001205	теорема: 0.004478	wikileaks: 0.002119	правительство: 0.00	пилить: 0.003929	автомобиль: 0.003883
16	возбуждение: 0.001205	сознание: 0.004478	ильин: 0.002119	мафия: 0.003651	лекарство: 0.003929	плаз: 0.003883
17	получасы: 0.001205	идеальный: 0.004478	дискуссия: 0.002119	финансовый: 0.0036	выглядеть: 0.003159	патрульный: 0.002936
18	lionel: 0.001205	единственный: 0.003925	юлий: 0.002119	воля: 0.003651	страхов: 0.003159	правительство: 0.002936
19	кормилица: 0.001205	простой: 0.003925	брэдли: 0.002119	рокфеллеров: 0.0036	строительство: 0.003159	направляться: 0.002936
20	агонизировать: 0.001205	поток: 0.003925	сталкиваться: 0.002119	действовать: 0.0027	относиться: 0.003159	площадь: 0.002936
21	явственно: 0.001205	черная: 0.003925	воображать: 0.002119	задача: 0.002761	почечный: 0.003159	лишь: 0.002936
22	ратчадамри: 0.001205	система: 0.003925	той: 0.002119	бомбить: 0.002761	болезнь: 0.003159	де: 0.002936
23	брендинга: 0.001205	разумный: 0.003925	отмечать: 0.002119	снова: 0.002761	народ: 0.003159	разбивать: 0.002936
24	куртка: 0.001205	основа: 0.003925	втора: 0.002119	катастрофа: 0.00276	диабет: 0.003159	армас: 0.001989
25	пограничник: 0.001205	определение: 0.003925	становиться: 0.002119	состоять: 0.002761	классический: 0.002388	мурильо: 0.001989

Распределение документов по темам

Documents with high probability

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	48: 0,421482	84: 0,146154	34: 0,106557	57: 0,596774	31: 0,452247	76: 0,435268	27: 0,236979	15: 0,906034	7: 0,391587	50: 0,563670	45: 0,140244	81: 0,077778	34: 0,139344	4: 0,350147
2	68: 0,064103	47: 0,042763	69: 0,080508	29: 0,145320	71: 0,087209	97: 0,323529	34: 0,090164	21: 0,748333	6: 0,339714	88: 0,218333	95: 0,083929	84: 0,069231	39: 0,054245	34: 0,057377
3	67: 0,054124	81: 0,033333	46: 0,060345	37: 0,077320	41: 0,069588	92: 0,306250	56: 0,087963	86: 0,100000	5: 0,325306	2: 0,206537	96: 0,062319	99: 0,058442	46: 0,043103	46: 0,043103
4	86: 0,047368	45: 0,030488	86: 0,047368	42: 0,066832	87: 0,059677	91: 0,292526	81: 0,055556	26: 0,087838	3: 0,051165	82: 0,048387	72: 0,053867	87: 0,043548	71: 0,040698	56: 0,041667
5	92: 0,043750	83: 0,030201	78: 0,041985	34: 0,040984	77: 0,055556	64: 0,284483	45: 0,042683	56: 0,078704	39: 0,049528	77: 0,043210	55: 0,053030	56: 0,041667	41: 0,038660	60: 0,034545
6	34: 0,040984	46: 0,025862	98: 0,041667	86: 0,036842	99: 0,045455	85: 0,253817	29: 0,036946	33: 0,049043	82: 0,048387	45: 0,042683	37: 0,046392	67: 0,038660	26: 0,033784	80: 0,028846
7	79: 0,033333	56: 0,023148	55: 0,037879	81: 0,033333	66: 0,034247	96: 0,242029	95: 0,030357	34: 0,040984	25: 0,044479	83: 0,030201	93: 0,039474	72: 0,031768	99: 0,032468	27: 0,028646
8	47: 0,029605	97: 0,022876	37: 0,036082	27: 0,023438	81: 0,033333	60: 0,241818	35: 0,026814	47: 0,036184	9: 0,043534	71: 0,029070	86: 0,036842	79: 0,023810	84: 0,023077	47: 0,023026
9	11: 0,026786	80: 0,022436	40: 0,035865	55: 0,022727	74: 0,023504	61: 0,240809	89: 0,026699	81: 0,033333	77: 0,043210	11: 0,026786	26: 0,033784	52: 0,022124	3: 0,022879	39: 0,021226
10	46: 0,025862	39: 0,021226	56: 0,023148	45: 0,018293	47: 0,023026	63: 0,224138	53: 0,019481	92: 0,031250	42: 0,042079	47: 0,023026	67: 0,028351	50: 0,020599	80: 0,022436	11: 0,020833
11	26: 0,020270	53: 0,019481	89: 0,021845	71: 0,017442	52: 0,022124	93: 0,223684	30: 0,017949	35: 0,023659	66: 0,029680	74: 0,019231	90: 0,027778	64: 0,020115	53: 0,019481	50: 0,020599
12	99: 0,019481	71: 0,017442	77: 0,018519	63: 0,017241	58: 0,022013	83: 0,218121	63: 0,017241	69: 0,021186	4: 0,029087	36: 0,018519	32: 0,027273	96: 0,018841	63: 0,017241	43: 0,018719
13	63: 0,017241	63: 0,017241	45: 0,018293	90: 0,016667	35: 0,020505	95: 0,205357	82: 0,016129	45: 0,018293	78: 0,026718	27: 0,018229	63: 0,017241	92: 0,018750	49: 0,016432	63: 0,017241
14	82: 0,016129	90: 0,016667	71: 0,017442	82: 0,016129	63: 0,017241	79: 0,195238	74: 0,014957	41: 0,018041	46: 0,025862	30: 0,017949	82: 0,016129	71: 0,017442	82: 0,016129	82: 0,016129
15	58: 0,015723	60: 0,016364	63: 0,017241	98: 0,013889	50: 0,016854	86: 0,194737	98: 0,013889	71: 0,017442	84: 0,023077	42: 0,017327	76: 0,015625	63: 0,017241	95: 0,016071	87: 0,014516
16	98: 0,013889	82: 0,016129	60: 0,016364	56: 0,013889	82: 0,016129	94: 0,183019	93: 0,013158	63: 0,017241	99: 0,019481	63: 0,017241	64: 0,014368	29: 0,017241	98: 0,013889	64: 0,014368
17	93: 0,013158	9: 0,015948	82: 0,016129	52: 0,013274	28: 0,015823	75: 0,159898	68: 0,012821	61: 0,016544	53: 0,019481	65: 0,016883	79: 0,014286	48: 0,016960	56: 0,013889	98: 0,013889
18	89: 0,012136	86: 0,015789	8: 0,013980	20: 0,013235	98: 0,013889	69: 0,148305	42: 0,012376	100: 0,016484	92: 0,018750	40: 0,014768	98: 0,013889	82: 0,016129	62: 0,013393	93: 0,013158
19	91: 0,011598	11: 0,014881	25: 0,013804	93: 0,013158	62: 0,013393	67: 0,146907	66: 0,011416	82: 0,016129	45: 0,018293	79: 0,014286	56: 0,013889	38: 0,015487	52: 0,013274	68: 0,012821
20	76: 0,011161	64: 0,014368	20: 0,013235	68: 0,012821	93: 0,013158	66: 0,139269	76: 0,011161	11: 0,014881	63: 0,017241	98: 0,013889	62: 0,013393	91: 0,014175	93: 0,013158	72: 0,012431
21	81: 0,011111	98: 0,013889	93: 0,013158	69: 0,012712	27: 0,013021	72: 0,122928	40: 0,010549	64: 0,014368	43: 0,016223	93: 0,013158	52: 0,013274	31: 0,014045	27: 0,013021	78: 0,011450
22	74: 0,010684	52: 0,013274	68: 0,012821	10: 0,012521	68: 0,012821	78: 0,110687	96: 0,010145	79: 0,014286	33: 0,015550	68: 0,012821	41: 0,012887	98: 0,013889	68: 0,012821	81: 0,011111

INTERNET STUDIES LAB

Восстановление матриц Φ и Θ методом максимизации функции правдоподобия при помощи E-M алгоритма (PLSA – модель Хофмана).

Формулировка задачи максимума правдоподобия выглядит следующим образом. Заменив p(w|t) и p(t|d) на матрицы ϕ_{wt}, θ_{td} и прологарифмировав выражение, что бы избавится от произведения, получим выражение, в котором необходимо подбирать значения матриц, так что бы это выражение было максимальным.

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d) = \sum_{t \in T} \phi_{wt}\theta_{td}$$

$$\sum_{t \in T} \theta_{td} = 1 \qquad \sum_{w \in W} \phi_{wt} = 1$$

$$L(\phi, \theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} ln \sum_{t \in T} \phi_{wt}\theta_{td} \rightarrow max$$

где n_{dw} – число термина w в документе d.

Поиск максимального значения функции L(**Φ**, **Θ**) осуществляется при помощи так называемого EM-алгоритм.

Hoffmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Mach. Learn. 42(1), 177–196 (2001)

Алгоритм нахождения неизвестных распределений (E-M алгоритм)

Частотные оценки условных вероятностей. Вероятности, связанные с наблюдаемыми переменными d и w, можно оценивать по выборке как частоты (здесь и далее выборочные оценки вероятностей p будем обозначать через $\hat{p}$):

$$\hat{p}(d,w) = \frac{n_{dw}}{n}, \qquad \hat{p}(d) = \frac{n_d}{n}, \qquad \hat{p}(w) = \frac{n_w}{n}, \qquad \hat{p}(w \mid d) = \frac{n_{dw}}{n_d},$$

n_{dw} — число вхождений термина w в документ d;

$n_d = \sum_{w \in W} n_{dw}$ — длина документа d в терминах;

$n_w = \sum_{d \in D} n_{dw}$ — число вхождений термина w во все документы коллекции;

$n = \sum_{d \in D} \sum_{w \in d} n_{dw}$ — длина коллекции в терминах.

Алгоритм нахождения неизвестных распределений (E-M алгоритм)

Вероятности, связанные со скрытой переменной t, также можно оценивать как частоты, если рассматривать коллекцию документов как выборку троек (d, w, t):

$$\hat{p}(t) = \frac{n_t}{n}, \qquad \hat{p}(w \mid t) = \frac{n_{wt}}{n_t}, \qquad \hat{p}(t \mid d) = \frac{n_{dt}}{n_d}, \qquad \hat{p}(t \mid d, w) = \frac{n_{dwt}}{n_{dw}},$$

n_{dwt} — число троек, в которых термин w документа d связан с темой t;

$n_{dt} = \sum\limits_{w \in W} n_{dwt}$ — число троек, в которых термин документа d связан с темой t;

$n_{wt} = \sum\limits_{d \in D} n_{dwt}$ — число троек, в которых термин w связан с темой t;

$n_t = \sum\limits_{d \in D} \sum\limits_{w \in d} n_{dwt}$ — число троек, связанных с темой t.

Алгоритм нахождения неизвестных распределений (E-M алгоритм)

Для решения задачи в PLSA применяется итерационный процесс, в котором каждая итерация состоит из двух шагов — E (expectation) и M (maximization) Перед первой итерацией выбирается начальное приближение параметров ϕ_{wt} , θ_{td}. Таким образом, мы задаем начальные вероятности:

$$\varphi_{wt} = \frac{n_{wt}}{n_t}, \qquad \theta_{td} = \frac{n_{dt}}{n_d},$$

E-шаг.

На E-шаге по текущим значениям параметров ϕ_{wt}, θ_{td} с помощью формулы Байеса вычисляются условные вероятности p(t | d, w) всех тем $t \in T$ для каждого термина $w \in d$ в каждом документе d. Это значит, что генерируем тему с вероятностью p(t |d), затем для данной темы, берем слово и ему присваиваем сгенерированную тему с вероятностью p(w|t), то есть, получается что у данного слово в данной теме пересчитываем вероятность:

$$p(t \mid d, w) = \frac{p(w \mid t)p(t \mid d)}{p(w \mid d)} = \frac{\varphi_{wt}\theta_{td}}{\sum \varphi_{ws}\theta_{sd}}.$$

Воронцов К.В. Лекции по тематическому моделированию. Voron-2013-ptm.pdf.

Алгоритм нахождения неизвестных распределений (E-M алгоритм)

M шаг.
На M-шаге, наоборот, по условным вероятностям тем p(t | d, w) вычисляется новое приближение параметров ϕ_{wt}, θ_{td}.

Просуммировав $n\hat{}_{dwt}$ по документам d и по терминам w, получим оценки $n\hat{}_{wt}$, $n\hat{}_{dt}$, $n\hat{}_{t}$, и через них, — частотные оценки условных вероятностей ϕ_{wt}, θ_{td}:

$$\varphi_{wt} = \frac{\hat{n}_{wt}}{\hat{n}_t}, \qquad \hat{n}_t = \sum_{w \in W} \hat{n}_{wt},$$

$$\theta_{td} = \frac{\hat{n}_{dt}}{\hat{n}_d}, \qquad \hat{n}_d = \sum_{t \in T} \hat{n}_{dt},$$

Таким образом гоняя E и M шаги можно рассчитать p(t | d) – вероятность принадлежности документа к теме, p(w | t) – вероятность принадлежности слова к теме. Число итерация является параметром, который определяется в эксперименте.

Воронцов К.В. Лекции по тематическому моделированию. Voron-2013-ptm.pdf.

Latent Dirichlet Allocation (LDA)

Идея Блея заключается в следующем. Он предположил, что столбцы матриц ϕ_{wt} и строки θ_{td} в модели PLSA порождены распределениями Дирихле. Каждое распределение характеризуется своим параметром α (для слов) и β (для документов), то есть все параметры α и β различны. Также в модели предполагается, что вероятность встретить каждую тему подчиняется мультиномиальному распределению:

Мульт. распр. : $P(\mathrm{x} \mid \mathrm{p}) = \frac{n!}{\Pi_{i=1}^{K} x_i!} \Pi_{i=1}^{K} p_i^{x_i}$

Дирихле распр. : $P(\mathrm{p};\alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} p_i^{\alpha_i - 1}$

Данные предположения существенно упрощает байесовский вывод модели за счет свойств сопряжённости распределений Дирихле с мультиномиальным распределением. Модель PLSA с учетом функций Дирихле трансформируется в следующую модель LDA:

$$p(\mathrm{z}, \mathrm{w} \mid \alpha, \beta) = \mathrm{p}(\mathrm{w} \mid \mathrm{z}, \beta) \cdot \mathrm{p}(\mathrm{z} \mid \alpha) = \prod_{k=1}^{K} \frac{B(\Psi_k + \beta)}{B(\beta)} \cdot \prod_{d=1}^{D} \frac{B(\Omega_d + \alpha)}{B(\alpha)}$$

Определение матриц ϕ_{wt} и θ_{td} в модели LDA в определяется при помощи E-M алгоритма

Итоговые формулы расчета матриц имеют следующий вид:

$$\theta_{td} = \frac{n_{td} + \alpha_t}{n_d + \alpha_0} \qquad \phi_{td} = \frac{n_{wt} + \beta_w}{n_t + \beta_0}$$

Особенности алгоритмов восстановления плотности распределения методами Монте - Карло

Математическим ожиданием дискретной случайной величины называется сумма произведений ее возможных значений на соответствующие им вероятности:

$$M(X) = x_1p_1 + x_2p_2 + \ldots + x_np_n$$

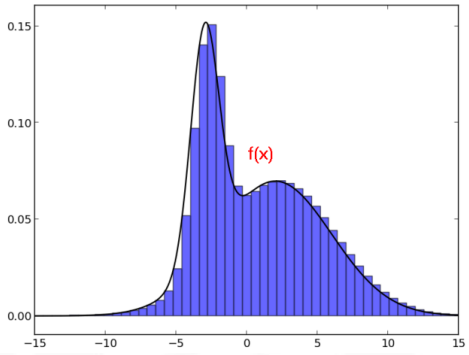
$$M(X) = \int_{-\infty}^{+\infty} xf(x)dx.$$

где x – случайная величина, p – вероятность реализации случайной величины

Вместо x можно использовать функцию φ(x),
Соответственно мат. ожидание M[φ(x)] будет:

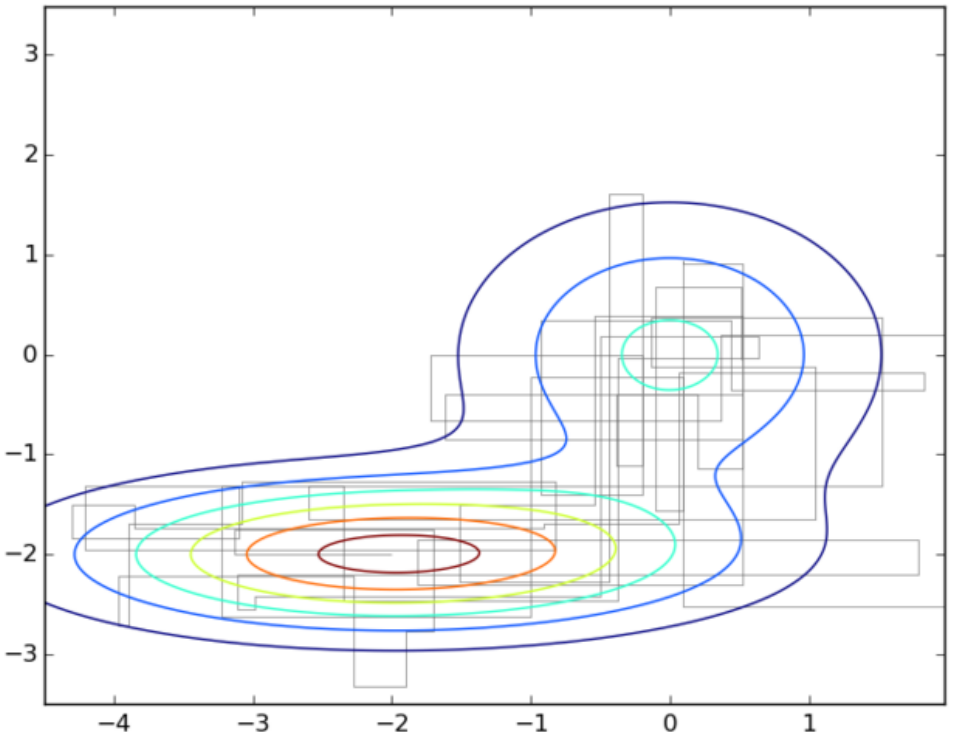
$$M[\varphi(X)] = \int_{-\infty}^{\infty} \varphi(x)f(x)dx$$

Подобный интеграл сложно найти аналитически, однако можно рассчитать при помощи алгоритма Метрополиса – Гастингса, или сэмплирования Гиббса.



Алгоритм сэмплирования Гиббса

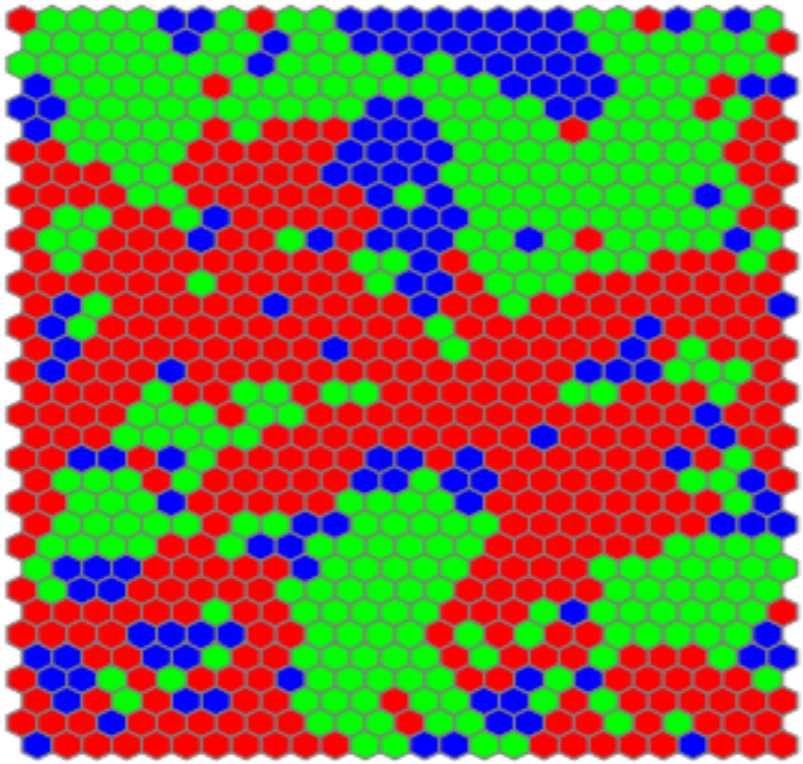
Идея сэмплирования по Гиббсу совсем простая: предположим, что мы находимся в очень большой размерности, вектор **x** очень большой, и нам сложно выбирать весь сэмпл сразу (то есть по всем осям), не получается. Давайте попробуем выбирать сэмпл не весь сразу, а покомпонентно. Это особенно удобно, если каждая компонента не зависит друг от друга, то есть многомерный вектор это произведение множества одномерных функций.



$$p(x_i|x_1^{t+1},\ldots,x_{i-1}^{t+1},x_{i+1}^{t},\ldots,x_n^{t})$$

Альтернативный вариант тематического моделирования

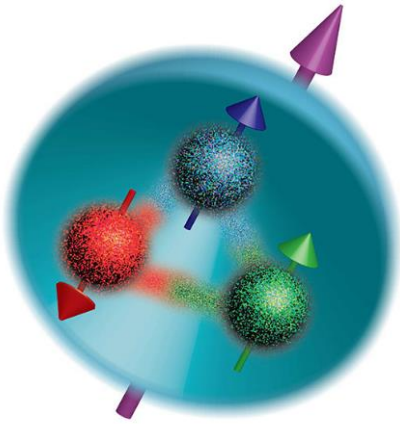
В статистической физике существует модель Поттса.



A configuration of the Q = 3 Potts model

Это дискретный ансамбль элементов, где каждый элемент представляет собой узел решетки. Каждый узел может находится в одном из K состояний. Целью исследования модели Поттса является получение распределения узлов по состояниям и оценка распределения узлов по энергии.

Важно: в модели Поттса состояние узла зависит от ближайшего окружения.



Critical exponents of domain walls in the two-dimensional Potts model
Jérôme Dubail, Journal of Physics A: Mathematical and Theoretical, 2010

INTERNET STUDIES LAB

Gibbs sampling на основе модели Поттса

Термины: **D** – пространство документов, **W** – пространство слов, **Z** – пространство тем. Темы являются скрытыми параметрами, которые должны быть найдены. Причем оценка основана на двух вещах: 1. Оценка вероятности производится как математическое ожидание. 2. В качестве функций используются мультиномиальные функции и функции Дирихле.

$$p(w \mid z, \beta) = \int p(w \mid z, \Phi)\, p(\Phi \mid \beta)\, d\Phi \qquad p(z \mid \alpha) = \int p(z \mid \Theta)\, p(\Theta \mid \alpha)\, d\Theta$$

Θ, Φ : матрица слова – темы и документы – темы.
Эти матрицы могут быть найдены двумя способами.

Идея Griffiths и Steyvers заключается в применении модели Поттса для тематического моделирования. Каждый узел решетки рассматривается в виде слова, а вся решетка рассматривается как документ. Номер спинового состояние рассматривается как номер темы. Отличие между TM и моделью Поттса состоит в том, что в тематическом моделировании используется множество документов. Вероятности распределения слов и документов можно оценить при помощи мат. ожидания, при условии знания подинтегральных функций.

$$P(z_i = j \mid w_i = m, z_{-i}, w_{-i}) \approx \frac{C^{WT}_{m,j} + \beta}{\sum_{m'} C^{WT}_{m',j} + V\beta} \cdot \frac{C^{DT}_{d,j} + \alpha}{C^{DT}_{d,j'} + \alpha T}$$

T. Griffiths, M. Steyvers. Finding Scientific Topics // Proceedings of the National Academy of Sciences. 2004. Vol. 101 (Suppl. 1). P. 5228–5335.

Свернутое семплирование по Гиббсу.

$$p(w_i) = \sum_{j=1}^{T} P(w_i|z_i = j)P(z_i = j) \cong \frac{C_{mj}^{wt} + \beta}{\sum_m C_{mj}^{wt} + V\beta} \frac{C_{dj}^{dt} + \alpha}{\sum_m C_{dj}^{dt} + \alpha T}$$

$C_{m,j}^{WT}$ - Матрица; в каждой ячейке находится число сколько раз слово **w** было связанно с темой **t**,

$C_{d,j}^{DT}$ - Матрица; в каждой ячейке находится число сколько раз слово **w** в документе **d** связанно с темой **t**.

$\sum_{m'} C_{m',j}^{WT} = n_t$ - Вектор; в каждой ячейке находится общее число слов - связанно с темой **t**,

$C_{d,j'}^{DT} = n_d$ Дина документа **d** словах.

Результат моделирования:

1. Распределение слов по темам.

$$\theta_{dj} = \frac{C_{d,j}^{DT} + \alpha}{C_{d,j'}^{DT} + T\alpha}$$

2. Распределение документов по темам.

$$\phi_{m,j} = \frac{C_{m,j}^{WT} + \beta}{\sum_{m'} C_{m',j}^{WT} + V\beta}$$

T. Griffiths, M. Steyvers. Finding Scientific Topics // Proceedings of the National Academy of Sciences. 2004. Vol. 101 (Suppl. 1). P. 5228–5335.

Алгоритм семплирования по Гиббсу

На входе: коллекция документов D, число тем |T|,
L - число итераций, коэффициенты β, α, N - число уникальных слов;
Initialization: φ(w,t)=1/N, θ(t,d)=1/T;

Внешний цикл по документам (i)
Внутренний цикл по словам в текущем документе. Длина цикла i.
Перебираем слово за словом. Для каждого слова выбираем номер темы в соответствии с распределением:

$$p(w_i) = \frac{C_{mj}^{wt} + \beta}{\sum_m C_{mj}^{wt} + V\beta} \frac{C_{dj}^{dt} + \alpha}{\sum_m C_{dj}^{dt} + \alpha T}$$

Обновление счетчиков.
Конец внутреннего цикла.
Конец внешнего цикла

Расчет матрицу по формулам

$$\theta_{dj} = \frac{C_{d,j}^{DT} + \alpha}{C_{d,j'}^{DT} + T\alpha} \qquad \phi_{m,j} = \frac{C_{m,j}^{WT} + \beta}{\sum_{m'} C_{m',j}^{WT} + V\beta}$$

T. Griffiths, M. Steyvers. Finding Scientific Topics // Proceedings of the National Academy of Sciences. 2004. Vol. 101 (Suppl. 1). P. 5228–5335.

Краткая характеристика возможностей и ограничений тематического моделирования

Достоинства:

1. Тематическое моделирование - один из немногих алгоритмов, который может работать с миллионами документов.
2. ТМ не зависит от языка.
3. Пользователь может работать только с наиболее вероятностными словами и документами в темам.
4. Хорошо работает на больших текстах.

Недостатки:

1. Существует проблема выбора числа тем (проблема присуще всем классификационным схемам). Эта проблема пока полностью не решена.
2. Метод обладаем определенной нестабильностью. На данный момент предложено несколько схем которые уменьшают проблему стабильности.
3. ТМ плохо работает на коротких текстах. Предложены некоторые решения этой проблемы.

Расстояние Ку́льбака — Ле́йблера, Perplexity

Расстояние Ку́льбака — Ле́йблера :

$$K = 0.5 \sum_{k}^{W} \Phi_k^1 \log\left(\frac{\Phi_k^1}{\Phi_k^2}\right) + 0.5 \sum_{k}^{W} \Phi_k^2 \log\left(\frac{\Phi_k^2}{\Phi_k^1}\right)$$

где Φ^1 - первое распределение, Φ^2 – второе распределение.
Если К=0 тогда два распределения идентичны, если К=max, то два распределения не идентичны.
Нормализованная мера КЛ:

$$Kn = \left(1 - \frac{K}{Max}\right) * 100$$

Perplexity:

n — длина коллекции в словах. n_{dw}- число вхождений слова w в документ d. p(w|d) – распределение слова в документе.

$$P = \exp\left(-\frac{1}{n} \sum_{d \in D} \sum_{w \in d} n_{dw} \ln p(w|d)\right),$$

Коэффициент Жаккара

Пусть у нас есть два множества X и Y. Тогда можно рассчитать следующие параметры:

a – множество видов X, отсутствующих в Y;

b – множество видов Y, отсутствующих в X;

c – множество видов, общих для X и Y;

Тогда коэффициентов Жаккара называется следующая комбинация:

$$Jk=|c/(a+b-c)|.$$

Коэффициент равен 1 в случае полного совпадения двух множеств и равны 0, если множества совершенно различны.

Проблема выбора числа тем в кластерном анализе

Finding the number of clusters in a data set : An information theoretic approach
CATHERINE A. SUGAR AND GARETH M. JAMES

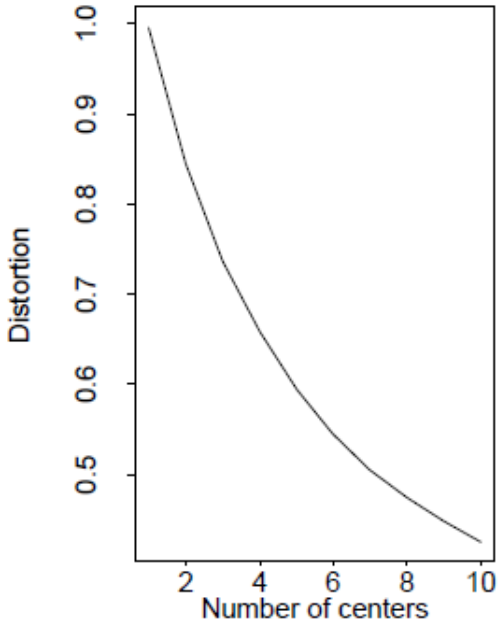
1. Определение минимального искажения (distortion). В качестве минимального искажения берется минимальное значение внутрикластерной дисперсии, встречающееся в данном кластерном решении.

2. Коэффициент трансформации. В качестве коэффициента также можно взять величину Y=1/K, где K – число кластеров.

3. **Transformed distortion.** Данная величина рассчитывается следующим образом:

$$D_t(K) = d^Y(K)$$

4. Расчет скачков ('Jumps'). Оценка поведения функции **transformed distortion** основана на оценке скачков функции, которые происходят при изменении числа.

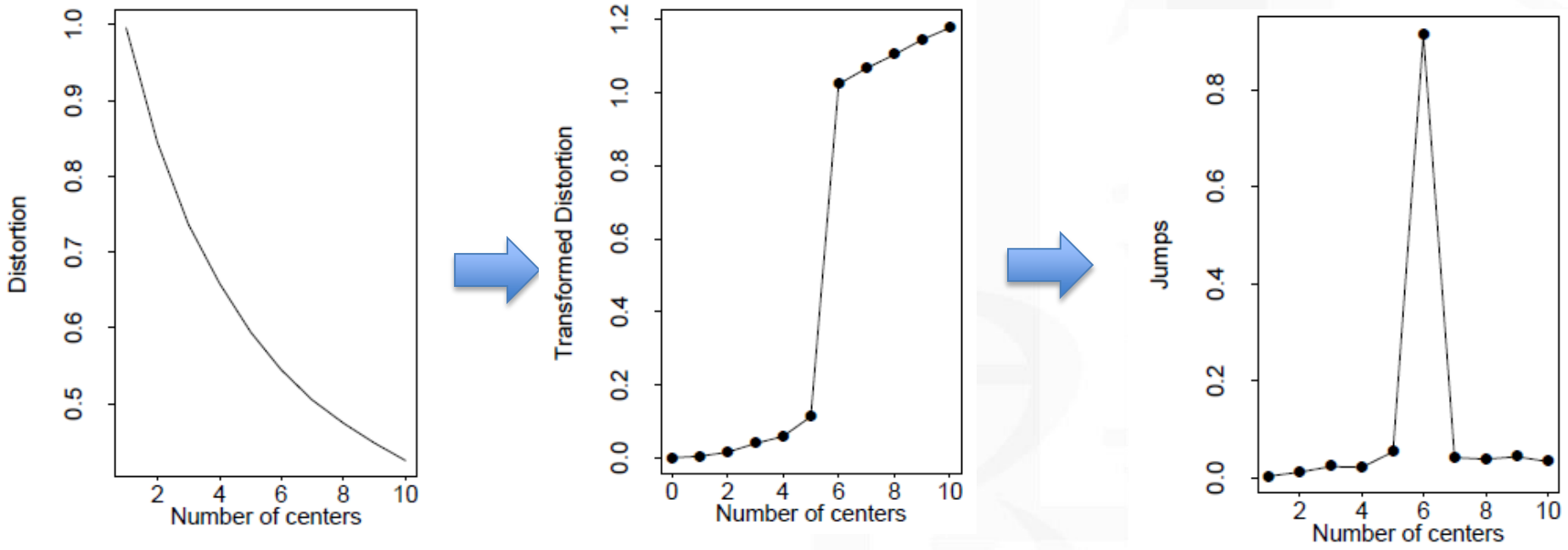


$$J_K = d_K^{-Y} - d_{K-1}^{-Y}$$

Авторы доказывают, что, при определенном выборе параметра *Y* (степень трансформации), построенная для функционала кривая будет иметь резкий скачок в том месте, которое соответствует действительному количеству кластеров.

Проблема выбора числа тем в кластерном анализе

Finding the number of clusters in a data set : An information theoretic approach
CATHERINE A. SUGAR AND GARETH M. JAMES

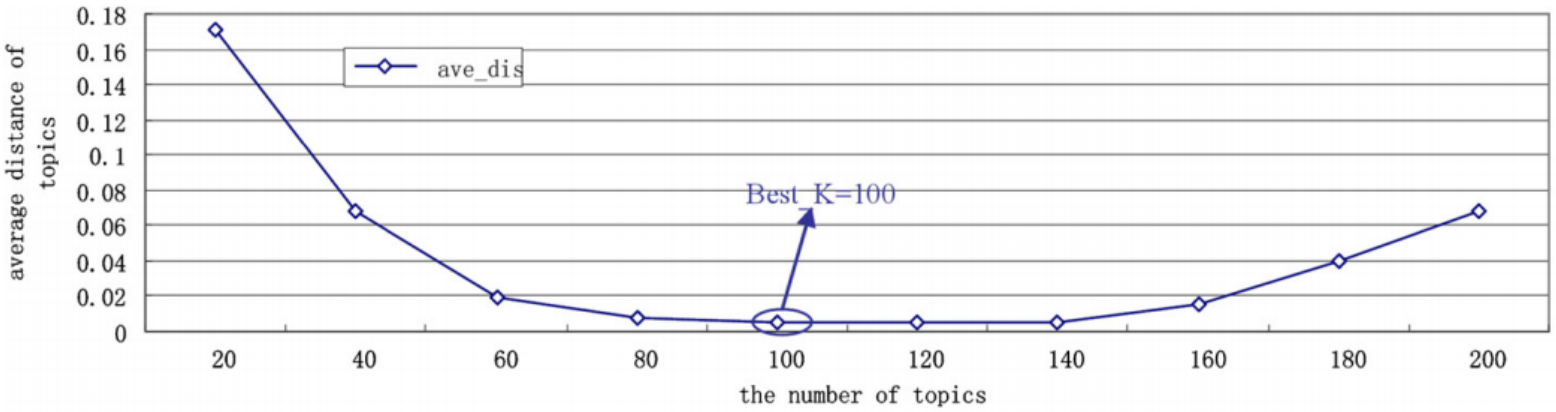


Проблема выбора числа тем в тематическом моделировании

A density-based method for adaptive LDA model selection
Juan Cao, Tian Xia,Jintao Li, Yongdong Zhang, Sheng Tang

cosine distance:

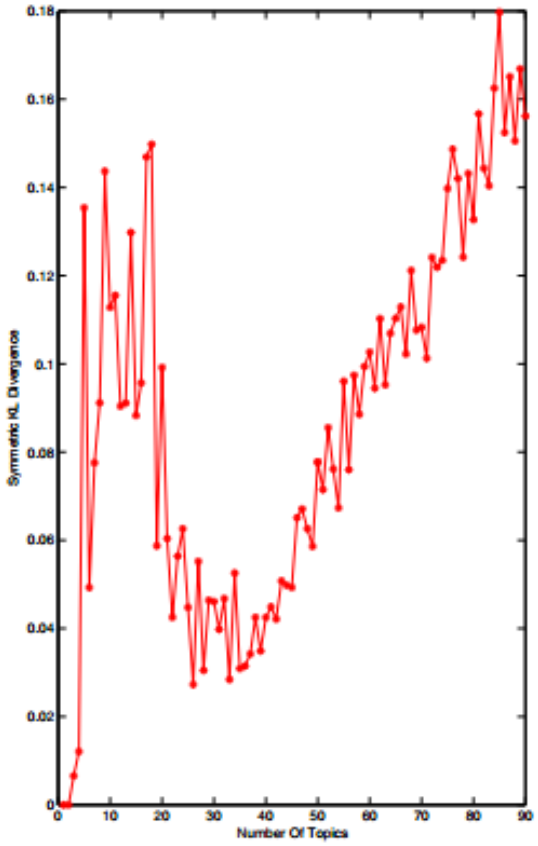
$$\mathrm{corre}(T_i, T_j) = \frac{\sum_{v=0}^{V} T_{iv} T_{jv}}{\sqrt{\sum_{v=0}^{V} (T_{iv})^2} \sqrt{\sum_{v=0}^{V} (T_{jv})^2}}$$



В данной работе авторы рассматривали тему как семантический кластер. Соответственно можно рассчитать сходства между кластерами по косинусной мере. Авторы постулировали, что минимум такой меры соответствует оптимальному числу тем.

Проблема выбора числа тем в тематическом моделировании

On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. R. Arun, V. Suresh, C.E. Veni Madhavan, and M. Narasimha Murty



$$M = U * \Sigma * V'$$

Авторы предлагают разложить матрицы Term-Topics и Doc-Topics при помощи SVD разложения. После этого можно рассчитать близость между векторами, содержащими сингулярные величины.

$$K = KL(C_{M1}||C_{M2}) + KL(C_{M2}||C_{M1})$$

C_{M1} distribution of singular values of Topic-Word matrix M1

C_{M2} Normalized distribution of singular values of Doc-Topic matrix M2

Проблема стабильности ТМ. Неоднозначность матричного разложения

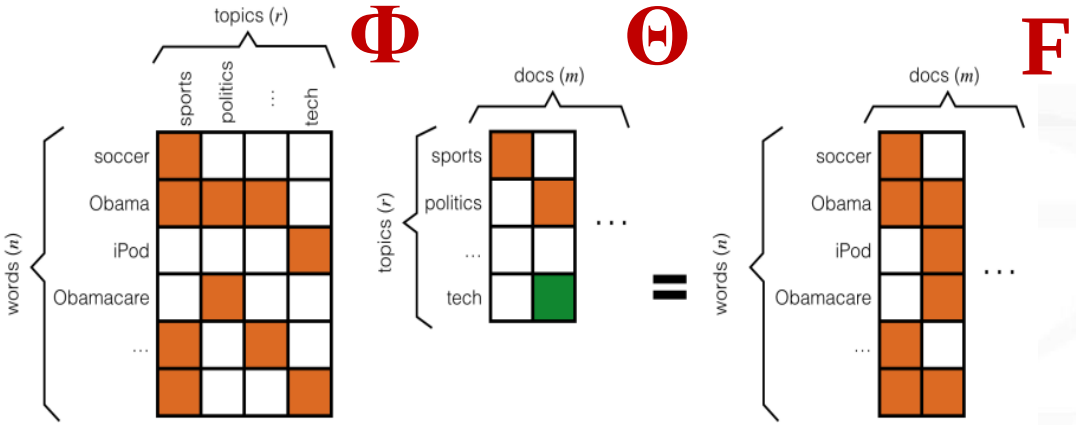
$$F[documents \times words] = \Theta[documents \times topics] \cdot \Phi[topics \times words]$$

Матрица F датасет (строки – документы, колонки список уникальных слов). Большой датасет может быть представлен в виде произведения двух относительно небольших матриц Φ and Θ. **Однако**:

$$F = \Theta \cdot \Phi = (\Theta \cdot R) \cdot (R^{-1}\Phi) = \Theta' \cdot \Phi'$$

Матрица может быть представлена в виде комбинации различных матриц, но такого же размера.

Это значит, что исходному набору слов и документов могут быть сопоставлены разные тематические решения, но имеющие одинаковое количество тем. Содержимое матриц и могут отличатся при разных матрицах преобразования.

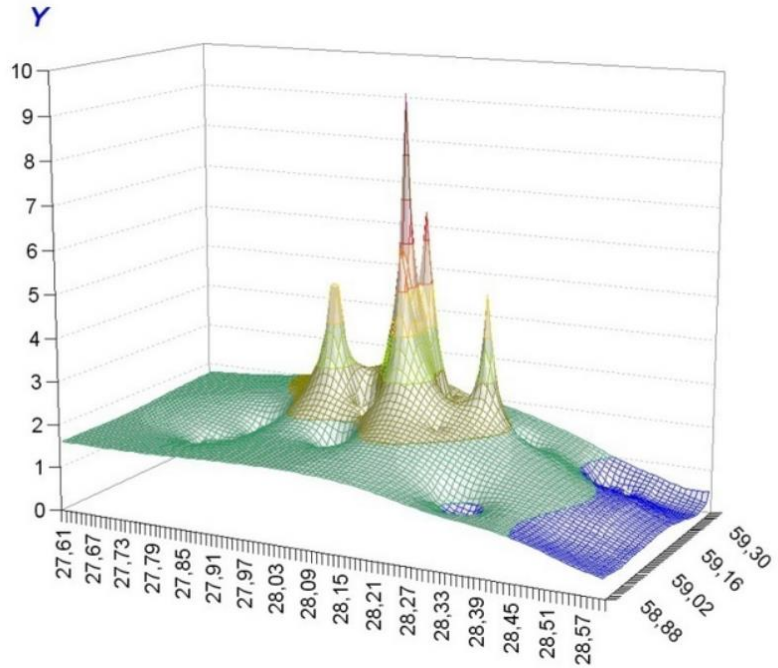


Проблема стабильности ТМ. Множество локальных минимумов и максимумов

$$\theta_{\mathrm{m,k}} = \int p(\mathrm{t}\mid \mathrm{d})\, \mathrm{p}(\theta_d, \alpha) = \frac{n(k;m)+\alpha_k}{(\sum_{k=1}^{K}\Omega(d;\mathrm{k})+\alpha_k)}$$

$$\phi_{\mathrm{k},t} = \frac{n(\mathrm{t};\mathrm{k})+\beta_t}{(\sum_{t'=1}^{V} n(\mathrm{t}';\mathrm{k})+1+\beta_{t'})}$$

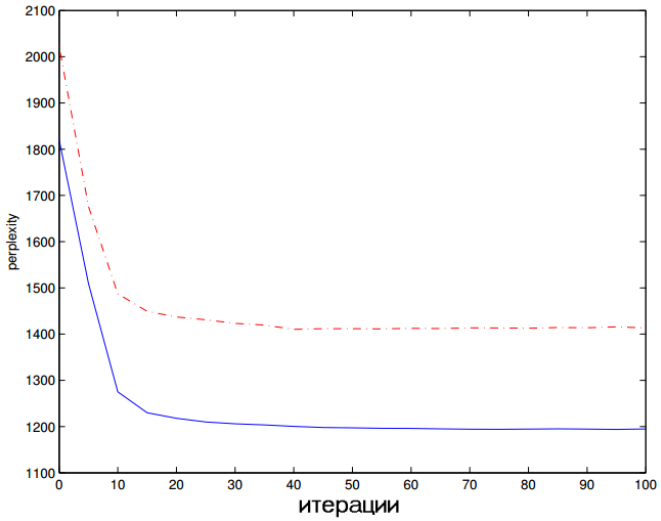
Под интегралом стоит произведение функций Дирихле. Итоговое подинтегральное выражение имеет множество локальным максимумов и минимумов.

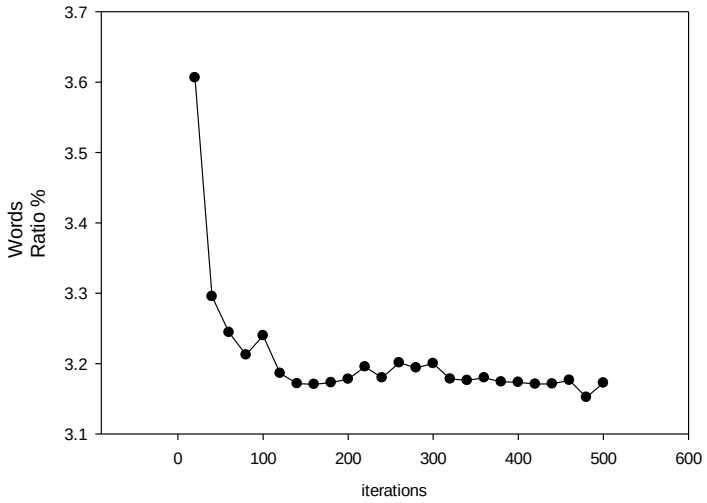


Так как расчет интегралов основан на методе Монте - Карло, и сводится к подсчету счетчиков, то в итоге значения счетчиков могут существенно отличатся при разных запусках процедуры сэмплирования. Это связанно, с тем что в ходе сэмплирования не все минимумы могут быть обойдены, и процесс сэмплирования может приводит к тому, что итоговые матрицы распределений слов и документов по темам будут отражать какой то один (или несколько) минимум.

Проблема стабильности – как измерять сходимость тематического моделирования

LDA (Gibbs sampling) – исходное распределение матрицы **Φ** (terms - topics) = **1/N**, матрица **Θ** (topics - doc) = 1/T. В ходе тематического моделирования происходит перераспределение вероятностей таким образом, что равновесное состояние становится сильно не равновесным. Характер неравновесности можно показать на основе того как меняется величина перплексити в ходе моделирования.





Words ratio: величина равная проценту числа элементов в матрице **Φ** (terms - topics), чьи вероятности выше величины 1/N. В начальный момент word_ratio=1. По истечению определенного числа итераций, word_ratio≅const(T). Данная величина характеризует меру разряженности матрицы **Φ.**

Проблема стабильности – как измерять сходства между темами

$$K = 0.5 \sum_{k}^{W} \Phi_k^1 \log\left(\frac{\Phi_k^1}{\Phi_k^2}\right) + 0.5 \sum_{k}^{W} \Phi_k^2 \log\left(\frac{\Phi_k^2}{\Phi_k^1}\right) \qquad Kn = \left(1 - \frac{K}{Max}\right) * 100$$

	A	B	C	D	E	F	G	H
1		Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
2	Topic 1	74.282	63.838	59.917	48.993	53.586	76.325	73.314
3	Topic 2	88.674	80.838	70.908	71.478	76.009	87.92	86.829
4	Topic 3	84.176	75.395	64.271	64.999	72.031	82.857	81.664
5	Topic 4	88.728	81.018	70.969	71.508	76.271	87.986	87.146
6	Topic 5	88.581	81.581	70.435	70.688	75.922	87.739	87.144
7	Topic 6	88.468	80.605	68.769	70.022	79.408	87.373	86.567
8	Topic 7	88.441	80.438	70.388	71.674	76.302	87.847	86.907
9	Topic 8	85.94	77.203	63.889	65.844	73.586	84.55	86.108
10	Topic 9	88.482	80.499	70.156	70.169	76.559	87.823	86.842
11	Topic 10	88.601	80.81	70.185	72.039	76.314	88.216	87.257
12	Topic 11	88.475	80.682	70.065	70.508	76.762	87.573	86.588
13	Topic 12	88.517	80.726	70.33	71.037	76.116	87.639	86.731
14	Topic 13	88.575	80.797	70.158	71.163	76.636	87.574	86.684

Оценка сходства между темами

$$Kn = \left(1 - \frac{K}{Max}\right) * 100$$

Level 90 - 93% (and more) means that first 50 words are almost identical.

Level about 85%: topics are completely different.

Similarity 0.935

USA	0.04734	USA	0.03567
American	0.02406	American	0.01804
Syria	0.02082	Syria	0.01758
Obama	0.01374	country	0.01495
weapon	0.01343	war	0.01361
war	0.01309	military	0.01246
president	0.01169	weapon	0.01084
UN	0.01018	Russia	0.01004
military	0.01014	Obama	0.00996
country	0.01005	president	0.0096
chemical	0.00944	UN	0.00869
Syrian	0.00851	international	0.00769

Similarity 0.854

USA	0.04734	water	0.01758
American	0.02406	help	0.01296
Syria	0.02082	city	0.01262
Obama	0.01374	far	0.01199
weapon	0.01343	house	0.01064
war	0.01309	east	0.0104
president	0.01169	region	0.00945
UN	0.01018	dam	0.0091
military	0.01014	flood	0.00904
country	0.01005	resident	0.00839
chemical	0.00944	injured	0.00714
Syrian	0.00851	FRS	0.00698

Проблема стабильности – как сравнивать тематические модели

BG	BH	BI	BJ	BK	BL	BM	BN
30 - 88.860	37 - 0.3514	31 - 86.245	150 - 0.3158	32 - 90.019	15 - 0.2270	33 - 87.539	15 - 0.1765
выбор	выбор	originally	c	белая	теперь	сон	теперь
избиратель	участок	фото	at	негр	говорить	ребенок	говорить
участок	наблюдател	at	originally	место	эта	спать	эта
кандидат	голос	this	e	можно	наш	когда	наш
партия	избирательн	please	posted	девать	узбекистан	чтобы	узбекистан
голос	комиссия	entry	via	очень	оставаться	час	оставаться
наблюдател	голосование	published	w	автобус	хотеть	время	хотеть
путин	бюллетень	фотография	there	быль	оказываться	может	оказываться
россия	уик	x	entry	кто	алмаз	ночь	алмаз
комиссия	член	the	this	хотеть	будет	хороший	будет
результат	избиратель	of	comments	мень	где	мочь	где
голосовани	фальсификац	comments	please	просто	сейчас	можно	сейчас
избиратель	нарушение	v	published	чтобы	лишь	начинать	лишь
депутат	результат	read	function	мой	бывший	будет	бывший
член	протокол	any	metrika	даже	сегодня	дитя	сегодня
шеин	председател	leave	by	наш	когда	она	когда
бюллетень	голосовать	rest	app	будет	мочь	минута	мочь
уик	подсчет	новость	of	время	ряд	во	ряд
давать	проголосова	a	read	она	система	месяц	система

Пути преодоления проблемы нестабильности ТМ

Тематическое моделирование относится к классу некорректно поставленных задач. Некорректность задачи заключается в том, что при одних и тех же исходных данных результаты тематического моделирования могут отличатся друг от друга. Возможное решение этой проблемы следует искать при помощи теории регуляризации, предложенного Тихоновым А.Н.

Суть регуляризации заключается либо в до определении априорной информации либо в сужении класса функций. Применении дополнительной априорной информации играет ключевую роль в теории регуляризации, чем большей априорной информацией мы обладаем, тем более устойчивые алгоритмы могут быть использованы при решении некорректно поставленной задачи. Применение процедуры регуляризации к тематическому моделированию заключается в разумном введении ограничений на матрицы, что позволяет сузить множество решений.

Регуляризационные тематические модели: Quadratic Regularizer

Newman D., Bonilla E. V., Buntine W. L. Improving topic coherence with regularized topic models.

Авторы предлагают ввести процедуру регуляризации, которая заключается в использовании информации о связях между словами. Эта информация должна браться извне (например из датасета) и выражается в виде ковариационной матрицы C размера WxW, где W - количество уникальных слов. Связь между матрицей Ф и матрицей C можно выразить следующим образом:

$$p(\phi_t \mid C) \propto (\phi_t^T \, C \, \phi_t)^{\nu}$$

где v – параметр регуляризации.

максимум логарифма правдоподобия будет выглядеть следующим образом:

$$L = \sum_{i=1}^{W} N_{ii} \log(\phi_{i|t}) + \nu \log(\phi_t^T C \phi_t)$$

процедура обновления матрицы

$$\phi_{w|t} \approx \frac{1}{N_t + 2\nu} \left(N_{wt} + 2\nu \cdot \frac{\phi_{w|t} \sum_{i=1}^{W} C_{iw} \phi_{i|t}}{\phi_t^T C \phi_t} \right)$$

Регуляризационные тематические модели: Convolved Dirichlet Regularizer

Newman D., Bonilla E. V., Buntine W. L. Improving topic coherence with regularized topic models.

Другой вариант регуляризации, предложенный авторами основан на идеи что матрица Ф зависит от матрицы C, которая в свою очередь содержит в себе зависимость между каждой парой уникальных слов.

Вероятность слова w в теме z выглядит так: $$p(\mathrm{w} \mid \mathrm{z} = \mathrm{t},) = \prod_{i=1}^{W} (\sum_{j=1}^{W} C_{ij} \psi_{\mathrm{j|t}})^{N_{it}}$$

Максимум логарифма правдоподобия будет: $$L = \sum_{i=1}^{W} N_{it} \log \sum_{j=1}^{W} C_{ij} \psi_{jt} + \sum_{j=1}^{W} (\gamma - 1) \log \psi_{jt}$$

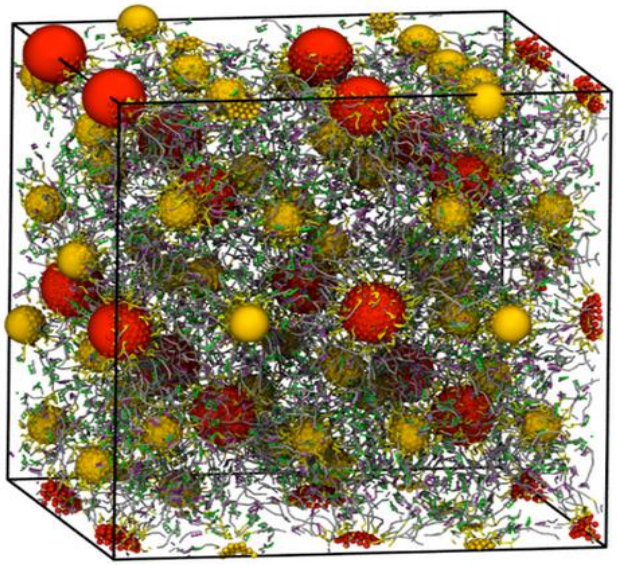
Обновление матрицы: $$\psi_{\mathrm{w|z}} \propto \sum_{i=1}^{W} \frac{N_{it} C_{iw}}{\sum_{j=1}^{W} C_{ij} \psi_{\mathrm{j|t}}} + \gamma$$

Недостатком данного подходя является во-первых, огромный размер матрицы C, во-вторых, содержимое матрицы C заранее не известно.

Semi-Supervised Latent Dirichlet Allocation (Gibbs sampling)

Следующая идея регуляризации основано на том, что можно зафиксировать набор слов по отношению к заданным номерам тем. Соответственно, когда происходит процедура сэмплирования, то для таких слов номера тем не меняются. Для остальных слов производится обычное сэмплирование Гиббса.

$$p(\mathrm{z}, \mathrm{w}, \alpha, \beta) \propto \begin{cases} z = t & \text{Начальное распределение слов (anchor words)} \\ q(\mathrm{z}, \mathrm{w}, \alpha, \beta) & \text{Gibbs sampling} \end{cases}$$



Модель SLDA ведет себя как процесс кристаллизации, в котором anchor words являются центрами кристаллизации. Так как фиксированные слова принадлежат определенным темам, то слова, которые часто совместно встречаются с этими словам будут формировать темы. Соответственно такие темы будут более стабильными.

Недостатком такого подхода является необходимость знания данного набора слов. Однако в ряде случаев это оправданно.

GRANULATED LDA (based on Gibbs sampling)

Гранулированный вариант LDA основан на идеи, что каждый документ можно рассматривать как гранулированную поверхность (по аналогии с моделью Поттса). Каждая гранула – набор слов, которые локализированы внутри одной гранулы, соответственно все слова внутри гранулы могут принадлежат одной теме.

DOCUMENT

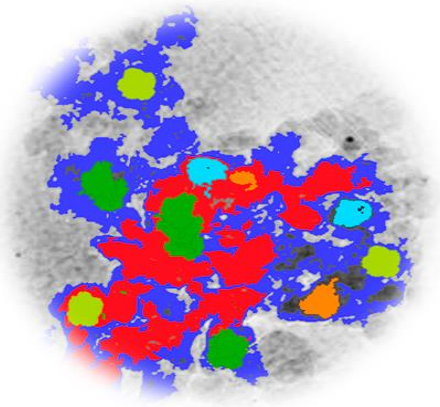
The central theme of ethnic nationalists is that «nations are defined by a shared heritage, which usually includes a common language, a common faith, and a common ethnic ancestry».[2] It also includes ideas of a culture shared between members of the group, and with their ancestors, and usually a shared language; however it is different from purely cultural definitions of «the nation» (which allow people to become members of a nation by cultural accimilation) and a purely linguistic definitions (which see «the nation» as all speakers of a specific language). Herodotus is the first who stated the main characteristic of ethnicity, with his famous account of what defines Greek identity, where he lists kinship language, cults and customs.

The central political tenet of ethnic nationalism is that ethnic groups can be identified unambiguously, and that each such group is entitled to self-determination.

The outcome of this right to self-determination may vary, from calls for self-regulated administrative bodies within an already-established society, to an autonomous entity separate from that society, to a sovereign state removed from that society. In international relations, it also leads to policies and movements for irredentism to claim a common nation based upon ethnicity.

KEYWORDS

ethnic	7
common language	3
culture	2
self-determination	2
society	3

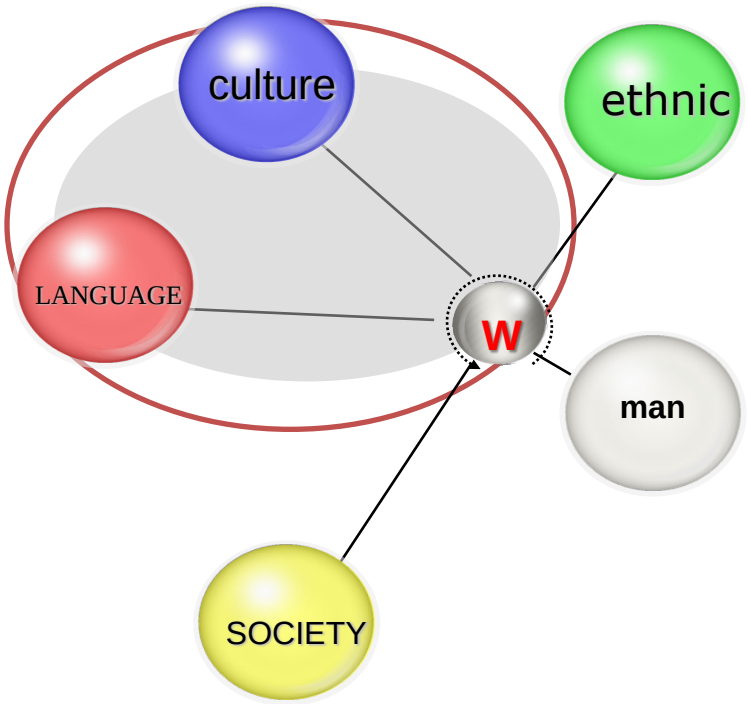


INTERNET STUDIES LAB

GRANULATED LDA: Algorithm

Entrance: коллекция документов D, число тем |T|, число итераций, размер гранулы

Initialization: инициализация матриц φ(w,t), θ(t,d) d ∈ D, w ∈ W , t ∈ T;



Запуск внешнего цикла(i) по документам

Запуск внутренного цикла (j) по словам

1. Генерируем случайное число k. Max value of k is number of words in documents i.
2. Выбираем слово под номером **k** из документа **i**.
3. Рассчитываем номер темы **t** для слова **к**.
4. Присваиваем всем словам внутри гранулы один и тот номер темы.

Конец внутреннего цикла.

Конец внешнего цикла.

$$\theta_{dj} = \frac{C_{d,j}^{DT} + \alpha}{C_{d,j'}^{DT} + T\alpha} \qquad \phi_{m,j} = \frac{C_{m,j}^{WT} + \beta}{\sum_{m'} C_{m',j}^{WT} + V\beta}$$

Sergei Koltcov, Stable Topic Modeling with Local Density Regularization, Lecture Notes in Computer Science series Vol. 9934. Switzerland : Springer, 2016.

Результаты сравнения стабильности некоторых моделей.

Topic model	Topic quality metrics		Topic stability metrics	
	coherence	tf-idf coherence	stable topics	Jaccard
pLSA	-237.38	-126.08	54	0.47
pLSA + Φ sparsity reg., $\alpha = 0.5$	-230.90	-126.38	9	0.44
PLSA + Θ sparsity reg., $\beta = 0.2$	-240.80	-124.09	87	0.47
LDA, Gibbs sampling	-207.27	-116.14	77	0.56
LDA, variational Bayes	-254.40	-106.53	111	0.53
SLDA	-208.45	-120.08	84	0.62
GLDA, $l = 1$	-183.96	-125.94	195	0.64
GLDA, $l = 2$	-169.36	-122.21	195	0.71
GLDA, $l = 3$	-163.05	-121.37	197	0.73
GLDA, $l = 4$	-161.78	-119.64	200	0.73

RESULT: (1) регуляризация может существенно влиять на результаты тематического моделирования. Регуляризация может как улучшать так и убивать стабильность тематического моделирования.

(модель LDA является регуляризованной версией pLSA, где регуляризация заключается в факте добавления информации о Dirichlet функциях).

Термодинамический подход к проблеме определения числа кластеров на основе тематического моделирования

1. В рассматриваемой информационной термодинамической системе общее количество слов и документов является константой, то есть, изменение объема отсутствует.
2. 2. Под темой понимается состояние (аналог направления спина), которое может принимать каждое слово и документ в коллекции.
3. 3. Информационная термодинамическая система является открытой и обменивается только энергией с внешней средой за счет изменения температуры. В данном подходе под температурой системы понимается число тем (или кластеров), которое задается извне.
4. Мерой неравновесности такой информационной системы можно использовать разность свободных энергий: $\Lambda_F=F(T)-F_0$, где F_0 – свободная энергия начального состояния, $F(T)$ – свободная энергия при заданной величине T.

Кольцов С, Письма в журнал технической физики, 2017. Принято в печать

Термодинамический подход к проблеме определения числа кластеров на основе тематического моделирования

Введем функцию плотности состояния следующим образом:

$\boldsymbol{\rho(E) = \frac{\sum_{ij}^{NT} N(E)_{nt}}{NT}}$, где $N(E)_{nt}$ - где $N(E)_{nt}$ – число состояний с энергией E, n,t – суммирование по всем словам и темам. Соответственно, энтропия данной системы выражается через плотность состояний следующим образом: $\boldsymbol{S(E) = \ln(\rho(E))}$

Энергия одного состояния можно выразить в виде: $\boldsymbol{\varepsilon_{nt} = -\ln(p_{nt})}$, где n – номер слова в списке уникальных слов, t –номер темы, p_{nt} - вероятность слова n в теме t.

В ходе тематического моделирования происходит переход к сильно неравновесному состоянию, которое характеризуется тем, что одна часть состояний имеет высокую вероятность $P_{nt} > 1/N$, а другая – низкую $P_{nt} < 1/N$ вероятность, близкую к нулю.

Кольцов С, Письма в журнал технической физики, 2017. Принято в печать

Термодинамический подход к проблеме определения числа кластеров на основе тематического моделирования

Неравновесную свободную энергию тематической модели можно представить в следующем виде:

$$\Lambda_F = F(T) - F_0 = (E(T) - E_0) - (S(T) - S_0) \cdot T = -\ln\left(\frac{\sum_{t=1}^{T}\sum_{n=1}^{N} P_{nt}}{T}\right) - T \cdot \ln\left(\frac{N_{k1}}{N \cdot T}\right)$$

где N_{k1} – число состояний, в которых $P_{nt}>1/N$, $(N \cdot T)$ – общее число всех состояний, T – число тем (варьируемый параметр), N – размер словаря уникальных слов, E_0 S_0,– энергия и энтропия системы при начальном равномерном распределении (flat distribution). Величины N_{k1} и P_{nt} подсчитываются для каждой тематической модели при вариации свободного параметра T, таким образом, величина Λ_F является функцией от T. Однако удобнее исследовать нормализованную неравновесную энергию, то есть $\Psi = \frac{\Lambda_F}{T}$.

Кольцов С, Письма в журнал технической физики, 2017. Принято в печать

Термодинамический подход к проблеме определения числа кластеров на основе тематического моделирования

Мера информации представляется как энтропия, взятая с обратным знаком, то есть максимум энтропии соответствует минимуму информации. Поэтому поиск оптимального числа тем в complex system сводится к поиску минимума энтропии.

Энтропию Реньи, в рамках термодинамического формализма $\boldsymbol{S^R_{q=1/T}} = \frac{\ln(\sum_k p_k^q)}{q-1}$, можно выразить через нормализованную свободную энергию следующим образом: $\boldsymbol{S^R_{q=1/T}} = \frac{T \cdot \Psi}{T-1} = \frac{F}{T-1}$, q=1/T. При этом, температура T рассматривается как формальный параметр, который можно менять в ходе вычислительного эксперимента.

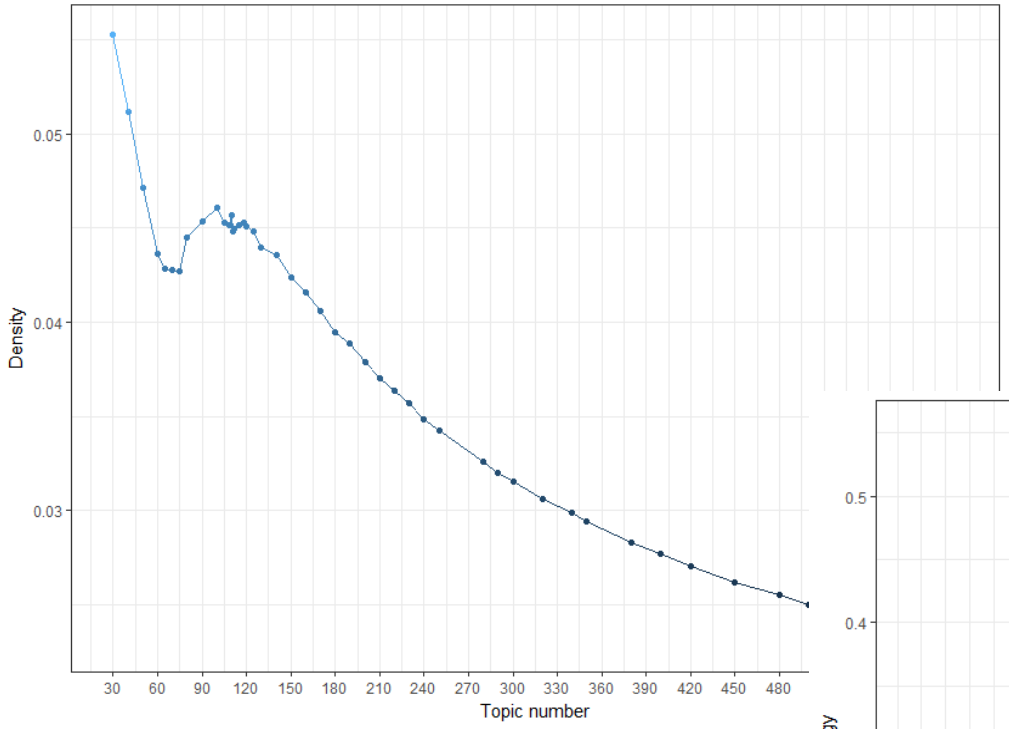
В свою очередь, энтропия Тсаллиса, записанную в виде $\boldsymbol{S^{Ts}_{q=1/T}} = \frac{1-\sum_k p_k^q}{q-1}$, также можно выразить через энтропию Re´nyi: $\boldsymbol{S^{Ts}_q} = \frac{e^{(q-1)\cdot S^R_\beta}-1}{q-1}$ и через свободную энергию.

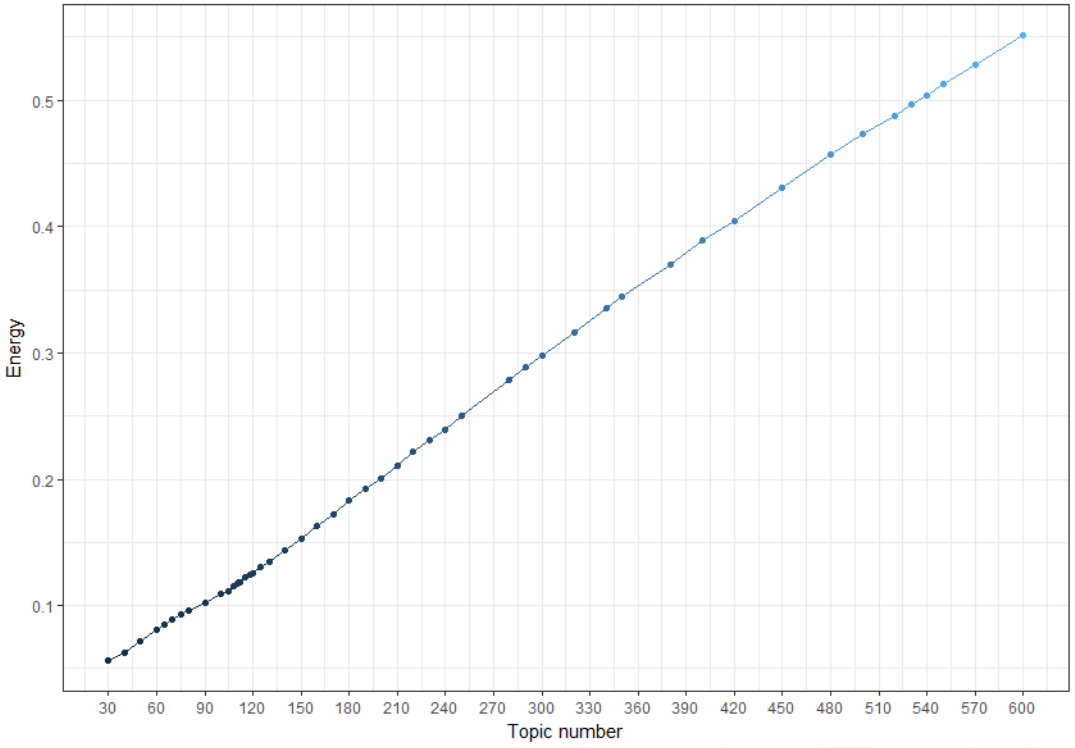
Термодинамический подход к проблеме определения числа кластеров на основе тематического моделирования

В рамках данного исследования был проведен следующий набор вычислительных экспериментов. На наборе данных фиксированного размера: 101481 постов из социальной сети ‘Живой журнал’ и N=172939 уникальных слов, проводилось тематическое моделирование на основе LDA с сэмплированием Гиббса, при разном количестве тем. Число тем варьировалось в диапазоне: T=[30; 500].

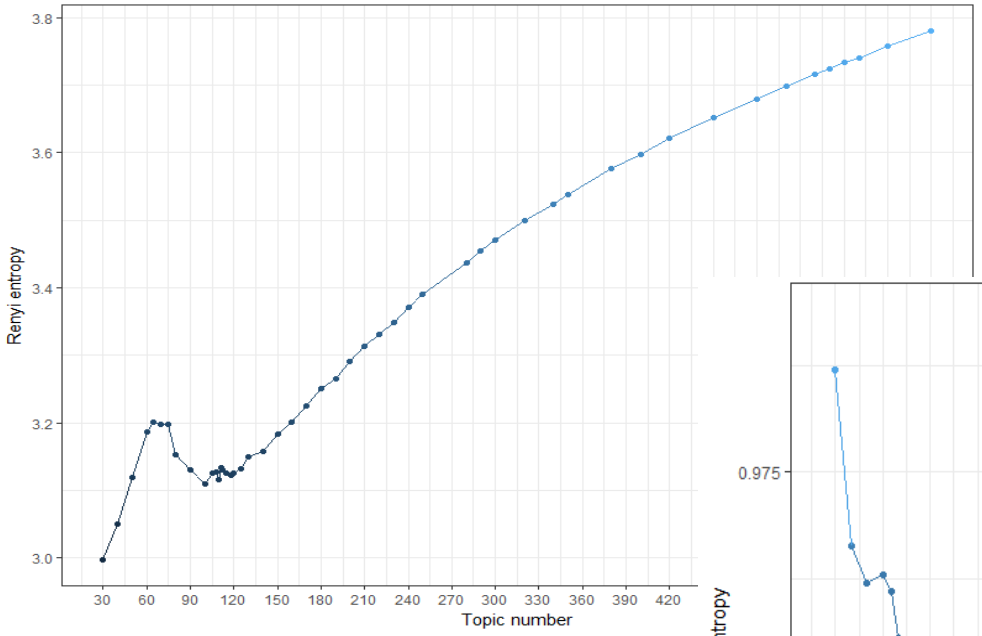
В каждом эксперименте, измерялось количество микросостояний, чьи вероятности больше величины 1/N, где N число слов в датасете. Таким образом, на основании формулы $\rho(E) = \frac{\sum_{nt}^{NT} N(E)_{nt}}{NT}$ рассчитывалась функция плотности состояний от числа тем, которая усреднялась по трем запускам. Также рассчитывалась энергия каждого тематического решения в соответствии с формулой $E(T) - E_0$. Величина энергии также усреднялась по трем запускам. Величина свободной неравновесной энергии рассчитывалась на основе усредненных значений плотности состояний и внутренней энергии.

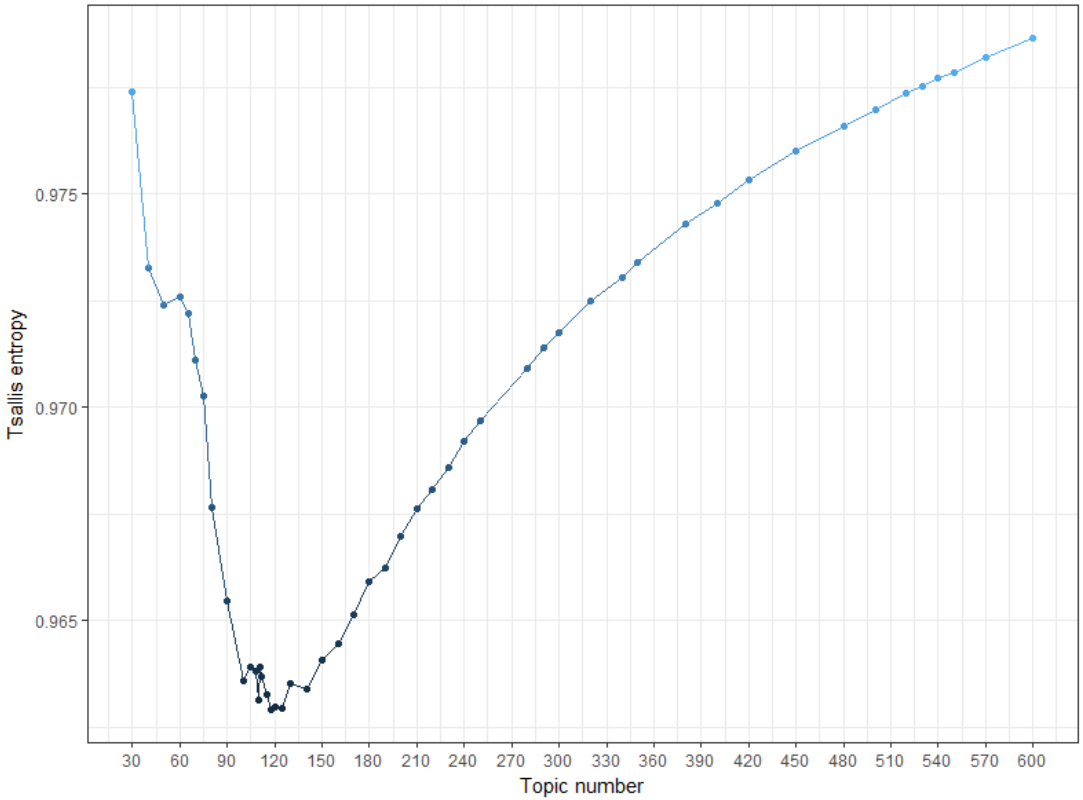
Термодинамический подход к проблеме определения числа кластеров на основе тематического моделирования





Термодинамический подход к проблеме определения числа кластеров на основе тематического моделирования





Discovering Health Topics in Social Media Using Topic Models. Michael J. Paul , Mark Dredze

Цитата из статьи

We used two Twitter datasets from different time periods. The first is a collection of over 2 billion tweets from May 2009 to October 2010. We used this dataset in earlier experiments which were used to inform our current data collection process. The second collection comes from Twitter's streaming API starting in August 2011 until February 2013, a daily average of 4 million tweets. We select all tweets that match any of 269 health keywords as well as 1% of public tweets.

Цитата из статьи

We collected 20,000 keyphrases related to illnesses, symptoms, and treatments from two websites. We added "sick" and "doctor" and removed spurious keywords. These keyphrases were used for our health filter and to identify symptom and treatment words as described below.

We additionally collected articles concerning 20 health issues from WebMD:[29] allergies, anxiety, asthma, back pain, breast cancer, COPD, depression, diabetes, ear infection, eye health, flu, foot injuries, heartburn, irritable bowel syndrome, migraine, obesity, oral health, skin health, and sleep disorders.

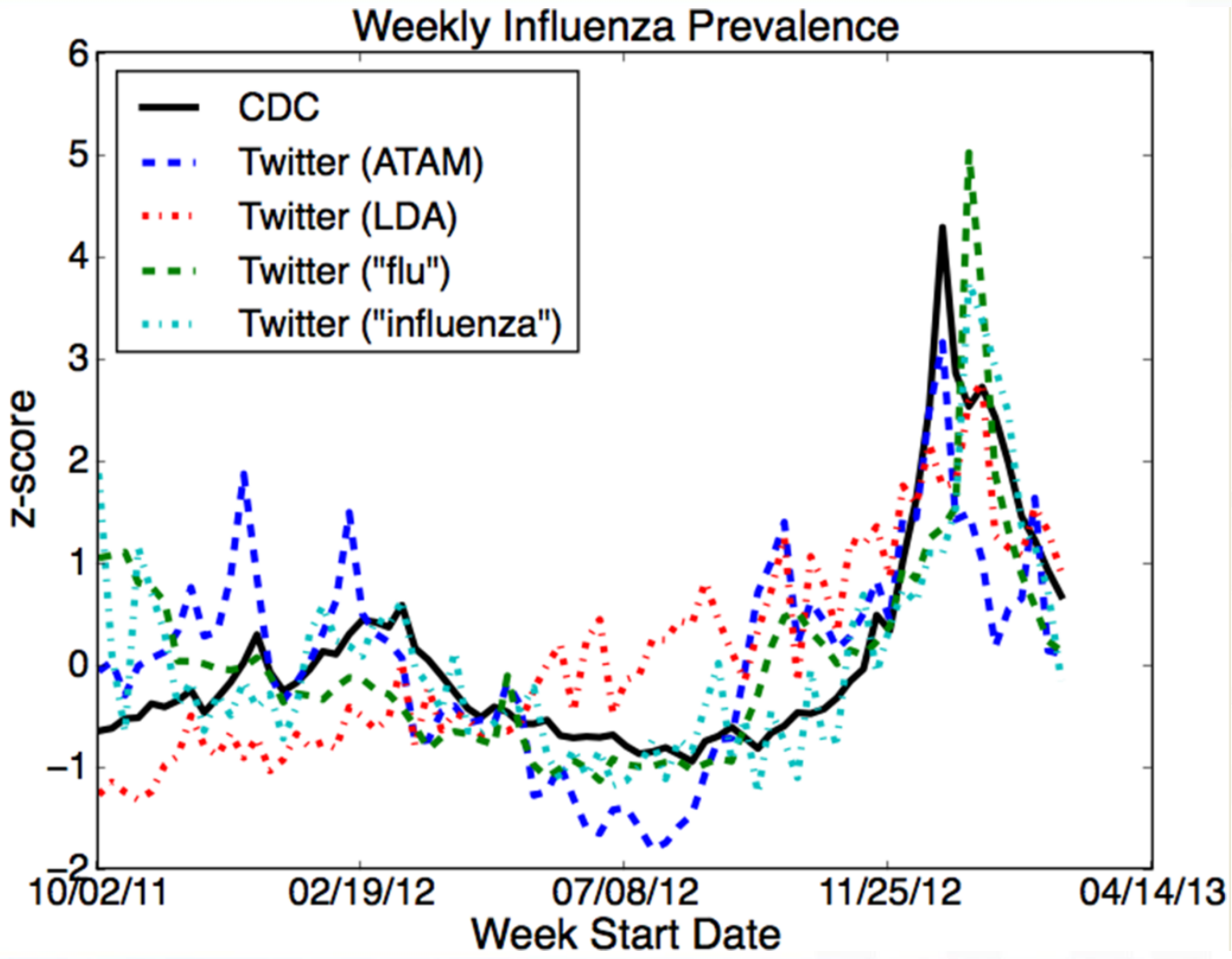
Discovering Health Topics in Social Media Using Topic Models. Michael J. Paul , Mark Dredze

Non-Ailment Topics						
TV & Movies	**Games & Sports**	**School**	**Conversation**	**Family**	**Transportation**	**Music**
watch	killing	ugh	ill	mom	home	voice
watching	play	class	ok	shes	car	hear
tv	game	school	haha	dad	drive	feelin
killing	playing	read	ha	says	walk	lil
movie	win	test	fine	hes	bus	night
seen	boys	doing	yeah	sister	driving	bit
movies	games	finish	thanks	tell	trip	music
mr	fight	reading	hey	mum	ride	listening
watched	lost	teacher	thats	brother	leave	listen
hi	team	write	xd	thinks	house	sound

Ailments						
	Influenza-like Illness	**Insomnia & Sleep Issues**	**Diet & Exercise**	**Cancer & Serious Illness**	**Injuries & Pain**	**Dental Health**
General Words	better	night	body	cancer	hurts	dentist
	hope	bed	pounds	help	knee	appointment
	ill	body	gym	pray	ankle	doctors
	soon	ill	weight	awareness	hurt	tooth
	feel	tired	lost	diagnosed	neck	teeth
	feeling	work	workout	prayers	ouch	appt
	day	day	lose	died	leg	wisdom
	flu	hours	days	family	arm	eye
	thanks	asleep	legs	friend	fell	going
	xx	morning	week	shes	left	went
Symptoms	sick	sleep	sore	cancer	pain	infection
	sore	headache	throat	breast	sore	pain
	throat	fall	pain	lung	head	mouth
	fever	insomnia	aching	prostate	foot	ear
	cough	sleeping	stomach	sad	feet	sinus
Treatments	hospital	sleeping	exercise	surgery	massage	surgery
	surgery	pills	diet	hospital	brace	braces
	antibiotics	caffeine	dieting	treatment	physical	antibiotics
	fluids	pill	exercises	heart	therapy	eye
	paracetamol	tylenol	protein	transplant	crutches	hospital

INTERNET STUDIES LAB

Discovering Health Topics in Social Media Using Topic Models. Michael J. Paul , Mark Dredze



The weekly rate P(*a*) for the ATAM ailment we identified as “influenza-like illness” correlated strongly with the CDC ILI data.

Cross-Lingual Topic Alignment in Time Series Japanese / Chinese News,

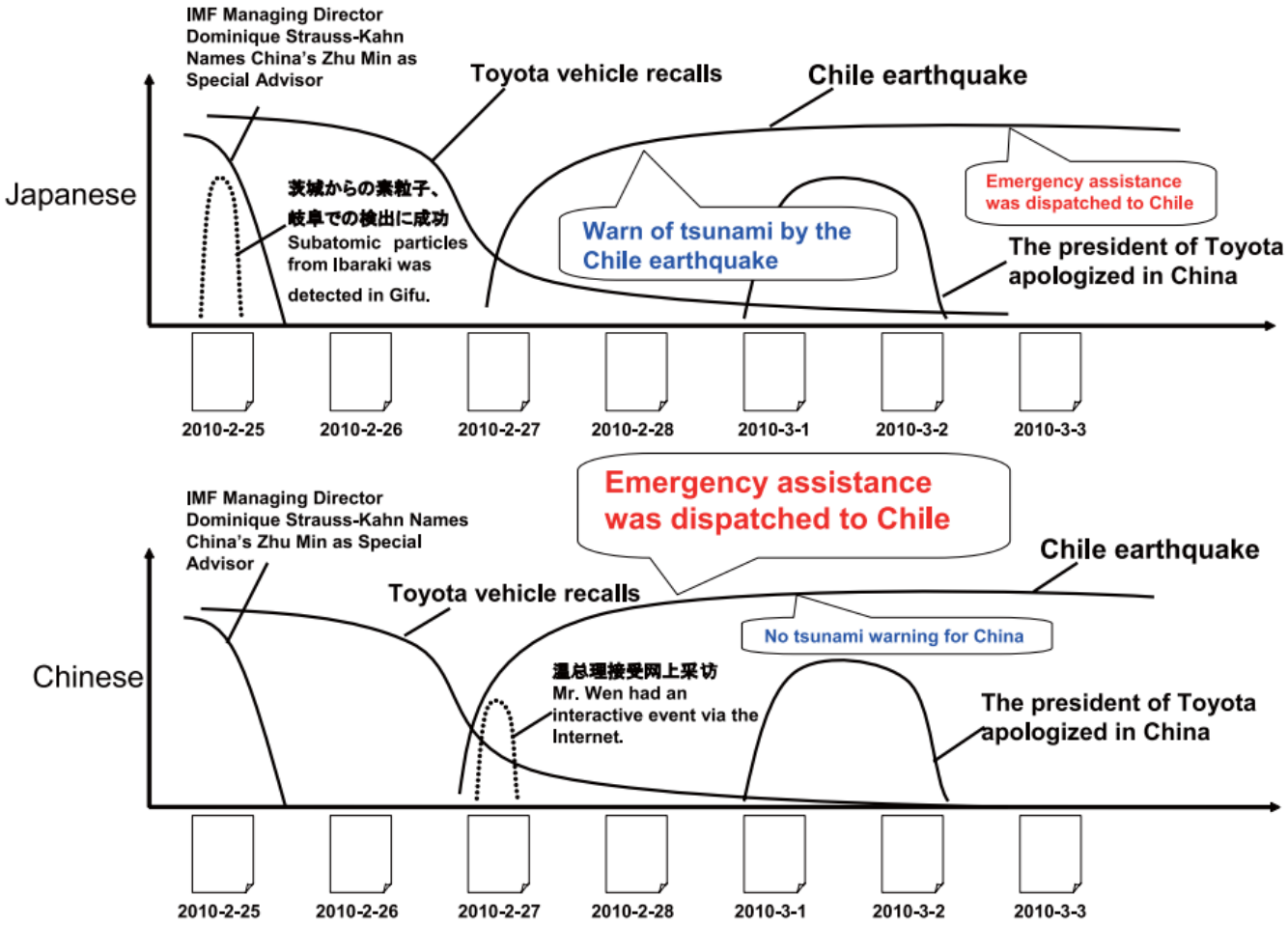
Shuo Hu Yusuke Takahashi Liyi Zheng Takehito Utsuro

1. Взяли китайские и японские новости
2. Построили тематические модели и посмотрели изменения тем во времени.
3. Нашли соответствия между парами японских китайских слов, обучив алгоритм на идентичных статьях википедии
4. Нашли тексты новостей, содержащих не менее заданного количества пар слов. Через них соотнесли японские темы с китайскими
5. Получили сравнительную картину изменения повестки дня в китайских и японских новостях

http://www.aclweb.org/anthology/Y12-1054

Cross-Lingual Topic Alignment in Time Series Japanese / Chinese News,

Shuo Hu Yusuke Takahashi Liyi Zheng Takehito Utsuro



http://www.aclweb.org/anthology/Y12-1054

Cross-Cultural Analysis of Blogs and Forums with Mixed-Collection Topic Models. Michael Paul and Roxana Girju

In the that experiment we consider data about three destination countries. Using the data provided by lonelyplanet.com, we crawled 1,108 threads from the UK forum, 1,112 from the India forum, and 1,046 from the Singapore forum.

weather time day going rain summer month high days thanks		
UK	**India**	**Singapore**
wind	leh	hot
waterproof	monsoon	humid
ending	road	humidity
rolling	manali	heat
walkers	ladakh	degree
rochdale	trekking	equator
layers	trek	sweat
snow	season	bring
footwear	rains	rain
ankle	monsoons	umbrella

Perspective of Locals			**Perspective of Tourists**		
food add chicken recipe cooking taste rice recipes sugar soup			food eat restaurant restaurants tea cheap meal eating cafe drink		
UK	**India**	**Singapore**	**UK**	**India**	**Singapore**
food	recipe	coffee	chips		
wine	recipes	cup	haggis		
restaurant	powder	oil	fish		
coffee	indian	comments	respectability		
cheese	salt	fried	decent		
soup	tsp	add	veggie		
eat	rice	restaurant	pudding		
chef	masala	rice	photoblog		
english	oil	tea	sausages		
drink	coriander	seafood	sandwiches		

Preferred by Locals			**Preferred by Tourists**		
recipe bowl lemon tomato simple spring spoon vanilla stir pour			street cheap couple yeah crowd old road floor run locals		
UK	**India**	**Singapore**	**UK**	**India**	**Singapore**
food	indian	cup	pubs	mother	quay
healthy	recipes	comments	music	ate	coast
shop	cup	tea	lane	tree	parkway
favorite	chicken	mins	brick	party	reasonably
wine	minutes	pot	fish	fields	air
icing	kitchen	note	jazz	base	sultan
coffee	mustard	nice	pints	rock	tum
leeds	fried	salt	dancing	toilet	views
duck	ginger	tarts	arms	bottled	plenty
extra	salt	fish	recommend	olive	rochester

LiveJournal Libra!': The political blogosphere and voting preferences in Russia in 2011–2012, O. Koltsova, A. Shcherbak

- Тематическое моделирование использовалось для вычленения постов на политические темы из всего множества постов 2000 популярных блогеров Живого Журнала за предвыборный период 2011-2012.
- Затем наиболее вероятные в этих темах посты кодировались вручную на «про-правительственные» и «оппозиционные» и отслеживалась динамика настроений в сети.
- Количество постов сравнивалась с предвыборными рейтингами.

LiveJournal Libra!': The political blogosphere and voting preferences in Russia in 2011–2012, O. Koltsova, A. Shcherbak

Topic 000	25397	Topic 001	29420
ukraine	1003	nemtsov	905
president	611	putin	477
russian	532	navalny	
unkrainian	440	sobchak	286
lukashenko	234	opposition	
head	229	conversation	209
minister	211	bolotny	204
viktor	180	bojen	200
union	178	the people	197
yanukovich	175	come out	196
state	170	sakharov	195
timoshenko	166	boris	184
vladimir	164	give a speech	171
claim	163	kurginyan	166
belarus	163	street	163
council	161	come	161
nuclear	161	xenia	158
security	158	leader	152
moscow	147	освистывать	152
belorussian	147	to name	139

INTERNET STUDIES LAB

LiveJournal Libra!': The political blogosphere and voting preferences in Russia in 2011–2012, O. Koltsova, A. Shcherbak

1,0 Из опубликованных ФСБ телефонных разговоров Немцова с Яшиным, Пархоменко, Пономаревым, Панюшкиным и другими я понял только то, что у Бориса Ефимовича какие то недопонимания с Евгенией Чириковой и что он реально не хотел подставлять людей под омоновские дубинки.

0,94 Из чужого твиттера: один чувак отнес Навальному LJUSERnavalny в изолятор мандарины, чурчхелы, сулугуни и лаваш. И сопроводил это запиской: "от Кавказа который хватит кормить"

0,88 А чо, теперь самые главные революционеры - это собчак, канделака и боженарынска? Суркофф - замечательный разводчик всё-таки!

0,87 Белые ленточки. Путин. Выборы. Беспредел. Митинг на Болотной. Посмотрим. Набигут ли спамеры.

0,83 После того как выложили записи Немцова - надо точно идти 24-го на митинг. Пусть все знают, что я трусливое офисное хомячье! Ура, товарищи!

LiveJournal Libra!': The political blogosphere and voting preferences in Russia in 2011–2012, O. Koltsova, A. Shcherbak

	ПОСТЫ О ВЫБОРАХ (%)	ПОСТЫ О ПОЛИТИКЕ (%)
Рейтинг КПРФ	**0.872****	**0.571***
	(p=0.000)	(p=0.041)
Рейтинг Единой России	-0.172	0.091
	(p=0.574)	(p=0.768)
Рейтинг Справедливой России	*0.551*	0.321
	(p=0.051)	(p=0.285)
Рейтинг ЛДПР	0.220	0.027
	(p=0.471)	(p=0.931)
Рейтинг В.Путина	-0.254	-0.054
	(p=0.402)	(0.860)
Рейтинг Г.Зюганова	**0.680***	0.462
	(p=0.011)	(p=0.112)
Рейтинг Д.Медведева	-0.278	-0.108
	(p=0.358)	(p=0.724)
Рейтинг В.Жириновского	0.106	-0.155
	(p=0.730)	(p=0.613)

LiveJournal Libra!': The political blogosphere and voting preferences in Russia in 2011–2012, O. Koltsova, A. Shcherbak

- Апелляция к общественному мнению в онлайне способна приносить определенные выгоды политикам в оффлайне
- «Рецепт успеха» для оппозиции оказался крайне прост: главное даже не «о чем» писать, а «сколько» писать. Любая информация о ходе российских выборов играла на руку оппозиции. Больше постов о выборах - выше предвыборные рейтинги оппозиционных партий.

Мониторинг межэтнической напряженности по материалам соц. сетей. (ЛИНИС 2015-2017)

ЗАДАЧИ

- С помощью тематического моделирования находить актуальные темы и релевантные тексты;
- Оценивать тональность текстов этно-релевантных тем для выявления потенциально проблемных тем;
- Отслеживать изменения во времени в региональном разрезе.

- Первая проблема ground truth:
- Что такое тема про этничность?
- Данная тема плохо определяется через ключевые слова и поэтому требует тематического моделирования для выявления.

ЭТАП 1:
Стандартный LDA на Живом Журнале

республика	украина	японский	век	олень
чечня	украинский	япония	народ	чум
кавказ	киев	японец	земля	праздник
чеченский	украинец	токио	история	стадо
чеченец	янукович	остров	русская	ханты
грозный	одесса	самурай	племя	охотник
дагестан	киевский	сакура	татарин	рог
кавказский	тимошенко	префектура	русь	оленевод
рамзан	грн	страна	европа	север
кадыр	ес	артур	город	якутия

(+) Темы хорошо интерпретируемы
(-) Темы о страх и регионах
(-) Темы о внешних этнических группах
(-) Темы о крупных этнических группах
(-) Темы совсем не стабильны

ЭТАП 2. SEMI-SUPERVISED LDA & pLSA на Живом Журнале

Слово или список слов закрепляется за темой или за диапазоном тем.

ЭТНОНИМЫ	КВАЗИЭТНОНИМЫ	ЭТНОФОЛИЗМЫ
бельгиец	азиат	хачик
бербер	европеец	чурка
болгар	африканец	чучмек
босниец	сибиряк	азер
бразилец	горец	укр
булгар	помор	татарва
бурят	кавказец	япошка
вайнах	южанин	китаёз
варяг	козак	узкоглазый
великоросс	варяг	черножопый

(+) появились темы о более редких этнических группах
(-)сложный отбор этнонимов
(-) темы все равно не стабильны
(-) темы все равно в основном про страны и регионы
(-) нельзя оценить вес тем в сравнении с неэтничными темами

ЭТАП 3: SEMI-SUPERVISED pLSA & LDA на ВКОНТАКТЕ

ВООБЩЕ НИЧЕГО НЕ РАБОТАЕТ ☹

- Короткие тексты не моделируются;

- Word2vec не помогает, т.к. появляются темы – списки этнонимов;
- Помогает сливание текстов одного автора, однако тогда нельзя отслеживать динамику;
- Сливание текстов одного региона за одну дату объединяет вместе тексты разных людей. Трудно ожидать в них цельности.

(Моделирование проводилось на случайной выборке из 5 млн текстов на все темы)

ЭТАП 4: SEMI-SUPERVISED pLSA & LDA на обогащенной выборке

Приобретена выборка в почти 4 млн постов за два года из всех русскоязычных социальных сетей, отобранная по списку этнонимов

ТМ работает лучше за счет того, что выборке есть заметная доля длинных текстов.

Ограничения:

- Не известно, отражают ли результаты такого ТМ тематику, содержащуюся в коротких текстах
- Отбор по ключевым словам отчасти обесценивает преимущества тематического моделирования

обзор программного обеспечения

C/C++/C#		**R/R studio**	
lda-c	David Blei (variational)	Topicmodels	Bettina Grün, C
hlda	Hierarchical LDA, Blei		Many LDA models
	http://www.cs.columbia.edu/~blei/topicmodeling_software.html	lda	Jonathan Chang Collapsed Gibbs sampling (greedy)
Java	Andrew McCallum, NLP	**Python**	Rehurek & Sojka
Mallet	Toolkit	Gensim	Online LDA (greedy)
	http://mallet.cs.umass.edu		https://radimrehurek.com/gensim/about.html
Matlab	Marc Steyvers (Gibbs		
MTMT (Matlab topic modeling toolkit)	sampling, C-matlab) http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm	**C+Delphi** TopicMiner (GUI!)	Sergei Koltcov Topic modeling, preprocessing & testing toolkit https://linis.hse.ru/en/soft-linis/

TopicMiner: обзор программного обеспечения

1. TopicMiner является 64 битной версией под Windows 7, 8, 10.
2. TopicMiner является полностью интерфейсной программой и не требует умения программирования.

Программа состоит из следующих компонент.

1. Модуль препроцессинга: A. Производит лематизацию русского (mystem.exe) и английского (porter.exe). B. Производит удаление стоп слов. C. Производит сасчет частот.
2. Модуль корректировки метаданных.
3. Модуль сэмплирования Гиббса (реализованы модели LDA, SLDA, ISLDA, GLDA).
4. Модуль BigArtm (офлайновый алгоритм).
5. Модуль сравнения тематических решений между собой.

Дополнительное ПО.
Модуль поиска стабильных тем на основе результатов сравнения тематических решений между собой.
ПО выложено на сайте лаборатории Интернет исследований:
https://linis.hse.ru/soft-linis/
https://linis.hse.ru/en/soft-linis/

NATIONAL RESEARCH UNIVERSITY

Thank you for your attention!

Room 216, building 2, 55 Sedova St., St.Petersburg, Russia
Laboratory for Internet Studies
www.linis.hse.ru