Московский Государственный Университет имени М.В. Ломоносова
Факультет вычислительной математики и кибернетики
Фоминская Галина
кафедра ММП, группа 317
2018

Тематическое моделирование в мультиязычном поиске

1 Введение

В данной работе был проведен анализ работы тематической модели в задаче мультиязычного поиска. EM-алгоритм был реализован на языке python. Эксперименты были проведены на сгенерированных данных и на датасете data news parallel. В качестве мер качества использовались топ-слова тем, топ-документы тем, а также метрика Average precision at n.

2 Постановка задачи

Рассматривается задача поиска переводов документа по мультиязычной коллекции документов. Будем решать её с помощью тематической модели коллекции, используя тематические профили документов в качестве векторных представлений.

3 Эксперименты

3.1 Проверка работы алгоритма

Перед запуском алгоритма на реальных данных проверим на сгенерированных данных, увеличивается ли правдоподобие. На рисунке 1 приведен график зависимости логарифма правдоподобия от номера итерации. Видно, что правдоподобие неубывает.

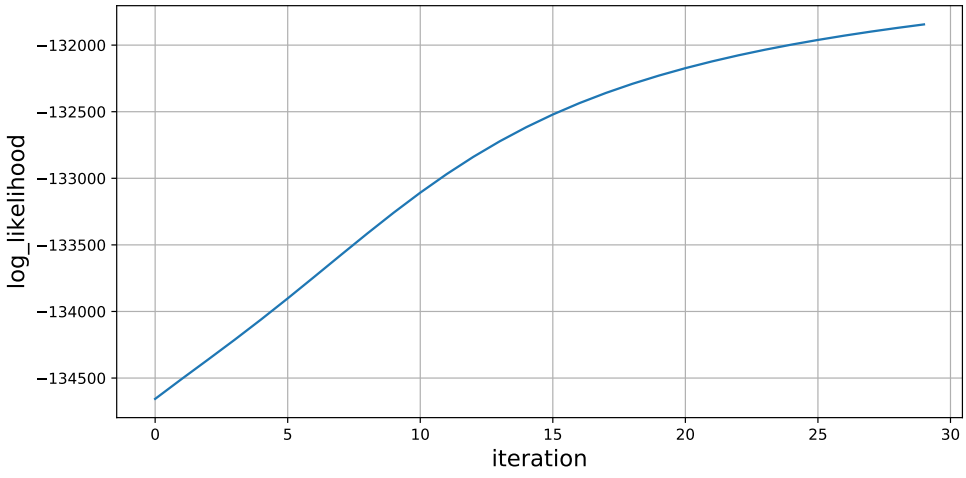

Рис. 1: Правдоподобие модели на сгенерированных данных

3.2 Модель без регуляризации

Перед тем, как решать конечную задачу, проверим качество работы модели без регуляризаторов. Построим модель с 50 темами на русскоязычной части модели и посмотрим на топ-слова и на топ-документы полученных тем. В таблице 1 приведены топ-слова пяти полученных тем. Видно, что в каждом столбце можно выделить общую тематику. В таблице 2 приведены начала топ 5 документов для первой темы. Видно, что в целом все тексты о правах человека и демократии.

topic 1	**topic 2**	**topic 3**	**topic 4**	**topic 5**	**topic 6**
право	рост	политический	ставка	финансовый	вид
страна	экономический	правительство	процентный	банк	культура
демократия	экономика	министр	актив	рынок	новый
система	ввп	лидер	облигация	система	история
принцип	расход	власть	долг	капитал	современный

Таблица 1: top words

В защиту банковского дела Швейцарии ЖЕНЕВА. Лидеры «большой двадцатки» только что объявили, что «эра тайны банковских операций завершилась», и пригрозили судебными разбирательствами «юрисдикциям, не идущим на сотрудничество, в том числе «налоговым оазисам» . . .
Миф о справедливых отношениях между поколениями ОКСФОРД: "Зачем что-то делать для потомков? извечный риторический вопрос, за которым обычно следует продолжение: "в конце концов, они никогда ничего не делали для нас". По общему мнению, это снимает с нас обязательства перед будущими поколениями . . .
Демократия приходит второй Демократия постепенно распространяется по всему миру. От Ближнего Востока до Латинской Америки и Азии многие автократии постепенно двигаются в сторону более демократичных и подотчетных форм правления . . .
Преступления и наказания истории БУЭНОС АЕРОС: Какова цена правосудия? Расследования истрических нарушений прав человека - предмет повседневной политики в странах Латинской Америки, Восточной Европы и Африки . . .
Являются ли права человека универсальными? Даже сейчас, когда в нашем мире происходит процесс глобализации, открытым все еще остается вопрос о том, является ли понятие «права человека» по существу западной концепцией . . .

Таблица 2: top documents for topic 1

3.3 Регуляризация на основе переводов слов

Здесь мы построим двуязычную модель. Будем отталкиваться от предположения, что тематики слова и его переводов должны быть близки друг к другу. Чтобы промоделировать это, введем регуляризатор, сближающий $p(t|w)$ слова w одного языка и частотной оценки $\frac{n_{tu}}{n_u}$ для его перевода u из другого языка.

Выведем формулу регуляризатора и градиента:

$$R = -\sum_w \sum_{u \in P(w)} \sum_t \frac{n_{tu}}{n_u} \log \left(\frac{n_{tu} n_w}{n_u \phi_{wt} n_t} \right),$$

$$\frac{\partial R}{\partial \phi_{wt}} = \sum_{u \in P(w)} \frac{n_{tu}}{n_u \phi_{wt}},$$

где $P(w)$ — множество переводов слова w.

Теперь необходимо подобрать коэффициент регуляризации. На рисунке 2 приведены графики зависимости логарифма правдоподобия от итерации для значений $\tau = 1, 10, 50, 100, 500$. В таблице 3 приведены значения логарифма правдоподобия для этих моделей для обучающей и тестовой частей документов, а так же значения метрики Ap@n. Видно, что при значении $\tau = 500$ получается наименьшее значение правдоподобия, но при этом значение Ap@n наилучшее из полученных. Это объясняется тем, что в мультиязычной коллекции с точки зрения обычного PLSA слова из разных языков должны быть в разных темах, поэтому когда слова перемешиваются, правдоподобие падает. Но при этом можно решать задачу поиска переводов документов, так как теперь модель настраивается на то, что слова из разных языков, хотя и не встречаются никогда в одних и тех же документах, имеют схожую тематику. В таблицах 4, 5, 6, 7, 8 приведены топ 10 слов в темах полученных моделей. Видно, что в моделях с маленькими значениями регуляризации топ слова тем состоят из слов только одного из языков. В то же время для больших значений τ в темах почти для каждого из топ слов в топ слова попал его перевод. Это согласуется с нашими ожиданиями.

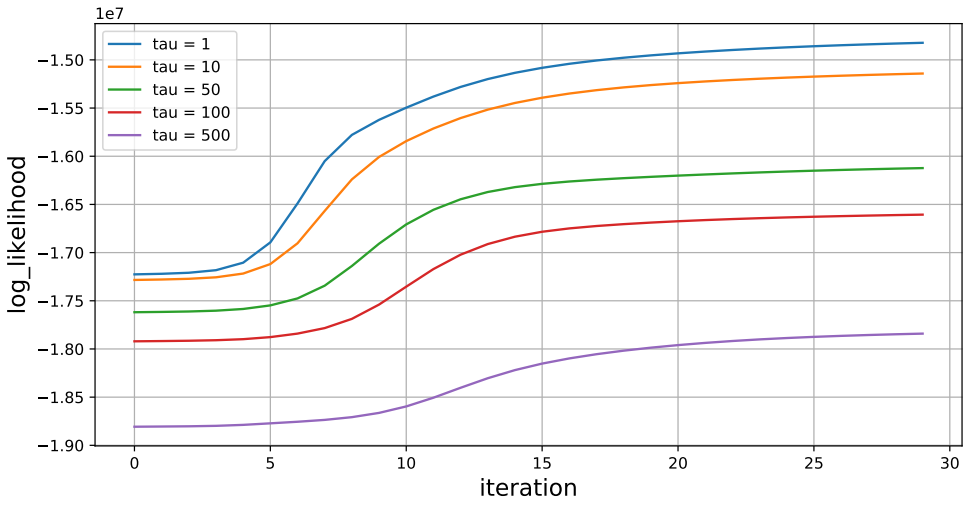

Рис. 2: Правдоподобие модели для разных значений τ

tau	final train log-likelihood	test log-likelihood	Ap@n
1	**-14822829**	**-15044356**	36.3
10	-15141701	-15364491	45.8
50	-16123002	-16257985	199.4
100	-16606099	-16679872	236.5
500	-17840785	-17912228	**976.1**

Таблица 3: tau selection

topic 1	topic 2	topic 3	topic 4	topic 5	topic 6
страна	energy	израиль	debt	health	european
африка	oil	израильский	eurozone	disease	europe
ребенок	climate	палестинский	greece	people	eu
сельский	carbon	палестинец	bank	research	germany
развитие	price	хамас	crisis	ha	country
население	change	мирный	fiscal	science	union
бедный	emission	арабский	country	life	member
миллион	food	государство	greek	human	state
фермер	world	газ	ecb	world	ha
помощь	gas	египет	government	aid	euro

Таблица 4: top words for $\tau = 1$

topic 1	topic 2	topic 3	topic 4	topic 5	topic 6
energy	country	ирак	женщина	рост	european
company	world	иракский	ребенок	экономический	eu
climate	people	саудовский	интернет	экономика	europe
financial	africa	саддам	образование	доход	union
world	development	аравия	de	уровень	member
carbon	health	война	школа	страна	country
global	ha	войско	мужчина	рабочий	state
oil	million	американский	работа	налог	ha
change	aid	аль	la	расход	germany
investment	global	суннит	сеть	высокий	euro

Таблица 5: top words for $\tau = 10$

3.4 Регуляризация на основе переводов слов

Рассмотрим другой подход к поиску переводов документов. Здесь мы предполагаем, что у нас нет словаря переводов, но есть корпус параллельных текстов. Будем отталкиваться от предположения, что тематики документа и его переводов должны быть близки друг к другу. Чтобы промоделировать это, введем регуляризатор, сближающий $p(t|d)$ документа d одного языка и частотной оценки $\frac{n_{ts}}{n_s}$ для его перевода s из другого языка. У такого подхода по сравнению с предыдущим есть ряд преимуществ. Здесь мы сближаем именно тематики документов, а значит

topic 1	topic 2	topic 3	topic 4	topic 5	topic 6
japan	вода	наука	болезнь	climate	суд
japanese	город	питание	расстройство	carbon	justice
морской	water	продукт	patient	emission	обвинение
abe	река	исследование	disorder	energy	court
японский	dam	вирус	лечение	change	prisoner
флот	дамба	diet	mental	warming	судебный
самолет	склон	обучение	врач	global	trial
успокаивать	river	ученый	пациент	coal	приговор
корабль	suburb	генетический	больной	gas	torture
судно	building	study	заболевание	heat	правосудие

Таблица 6: top words for $\tau = 50$

topic 1	topic 2	topic 3	topic 4	topic 5	topic 6
nbsp	israel	вознаграждение	european	mortgage	malaria
chicago	palestinian	profitability	europe	repayment	africa
lying	israeli	лицо	eu	monthly	ghana
ложь	hamas	компенсация	political	скидка	donor
envy	settlement	pay	union	discount	искоренение
зависть	arab	derivative	member	financier	уганда
слабый	peace	клиент	state	жилищный	малярия
story	gaza	покупатель	germany	sink	charity
невидимый	palestine	payoff	party	residential	ангола
слепой	палестинский	compensation	wa	holder	angola

Таблица 7: top words for $\tau = 100$

topic 1	topic 2	topic 3	topic 4	topic 5	topic 6
chinese	islamic	scientific	буш	wa	animal
china	мусульманский	ученый	bush	приговор	культурный
communist	muslim	universe	prisoner	ha	caste
император	fundamentalist	наука	суд	обвинение	каста
emperor	религиозный	научный	warrant	тюрьма	культура
тайваньский	prince	science	justice	government	вирус
тайвань	christian	system	подсудимый	law	язык
китайский	исламский	открытие	обвинение	осуждать	млекопитающее
материк	сторонник	psychiatric	torture	обращение	education
taiwanese	фанатизм	психиатрический	impunity	утверждение	происхождение

Таблица 8: top words for $\tau = 500$

в одной теме будут не просто слова - переводы друг друга, а слова из разных языков, которые часто употребляются в одном смысловом контексте.

Выведем формулу регуляризатора и градиента:

$$R = -\sum_d \sum_t \frac{n_{ts}}{n_s} \log\left(\frac{n_{ts}}{n_s \theta_{td}}\right),$$

$$\frac{\partial R}{\partial \theta_{td}} = \frac{n_{ts}}{n_s \theta_{td}},$$

где s — перевод документа d.