Санкт-Петербургский государственный университет

ХАЛИУЛЛИНА Лия Рауфовна

Выпускная квалификационная работа

Вероятностное моделирование в классификации коллекции документов

Уровень образования: магистратура

Направление: 01.04.02 «Прикладная математика и информатика»

Основная образовательная программа: ВМ.5504 «Исследование операций и системный анализ»

Научный руководитель:
профессор кафедры
МТИСР, доктор
технических наук,
Буре В. М.

Рецензент: аналитик,
ООО «Эпам Системз»,
Староверова К. Ю.

Санкт-Петербург

2020 год

Содержание

Введение 3
Постановка задачи 5
Глава 1. Вероятностное тематическое моделирование 6
1.1. Основные понятия 6
1.2. Модель PLSA 7
1.3. Модель LDA 9
1.4. Аддитивная регуляризация тематических моделей 12
1.4.1. Общий подход 12
1.4.2. Разновидности регуляризаторов 13
1.5. Метрики для оценки качества модели 18
Глава 2. Предлагаемый алгоритм 20
3.1. Предварительная обработка текста 22
3.2. Создание обучающей выборки и обучение классификатора 25
3.3. Оценка качества построенной модели и выводы 30
Заключение 33
Список литературы 34
Приложение 36

Введение

Быстрый рост потоков информации ставит не только вопрос её хранения, но и задачу её систематизации и анализа. При работе с текстами подобный анализ помогает извлечь необходимые сведения о настроении, актуальности, тематике, а также является необходимой ступенью перед последующими действиями вроде поиска, сравнения или категоризации. Большой популярностью пользуются различные статистические методы обработки текста, в частности, тематическое моделирование. Вероятностная тематическая модель (probabilistic topic model) коллекции документов представляет каждый документ в виде дискретного распределения вероятностей тем, а каждую тему – в виде дискретного распределения вероятностей слов (терминов). Построение вероятностной тематической модели можно также описать как задачу одновременной кластеризации (би-кластеризации) документов и слов по одному и тому же множеству кластеров, называемых темами. Особенностью подобного моделирования является осуществление «нечеткой кластеризации» (soft clustering), то есть документ может принадлежать нескольким темам [2].

Тематические модели актуальны для решения множества задач анализа текста: информационный поиск, тематическая сегментация текстов, выявление трендов в новостных публикациях, обнаружение текстового спама, а также классификация и категоризация документов, что и будет раскрыто подробнее в данной работе.

Задача классификации чаще всего предполагает построение алгоритма отнесения каждого документа лишь к одному классу (однозначная классификация). Вероятностная тематическая модель же способна выявить принадлежность документа к нескольким классам (многозначная классификация). К примеру, в медицинских исследованиях зачастую используют статистические методы, и при стандартной классификации текст

с подобным исследованием будет отнесен к условному классу «медицина», так как характеризующие этот класс слова будут преобладать. Но при классификации с помощью вероятностной тематической модели этот текст будет также определен к классу, скажем, «статистический анализ», а может и к нескольким другим. Более того, будет выявлена вероятность принадлежности документа к этим классам. Таким образом, какому-либо классу будут соответствовать и документы, для которых этот класс не является основным.

Классификация также предполагает наличие обучающей выборки, которая составляется вручную экспертом (или несколькими) и по возможности является достаточно информативной, чтобы метод в дальнейшем смог правильно распределять новые документы по имеющимся классам. Создание обучающей выборки занимает много времени, особенно при условии отнесения каждого документа к нескольким классам, а также ограничено знаниями эксперта, ведь он может слабо разбираться в какой-либо области, значит не исключена изначальная неточность в определении классов этой области и принадлежности к ним документов. К тому же общее количество классов также может оказаться относительно небольшим, и тогда скорее всего будут сформулированы лишь обобщенные темы. Поэтому в задачах классификации по большому количеству маленьких классов имеется проблема наличия подходящей обучающей выборки.

Постановка задачи

Целью данной работы является построение вероятностной тематической модели для многозначной классификации коллекции документов по небольшим классам, удовлетворяющей условиям: отсутствие готовой качественной обучающей выборки и отнесение документа к классу, даже если этот класс не формирует основную тематику документа.

В качестве документов рассматриваются выпускные квалификационные работы студентов, но это также могут быть научные статьи или любые другие документы. Сделанный выбор не влияет на результаты и объясняется исключительно личным удобством.

Для достижения данной цели ставятся следующие задачи:

1) исследовать существующие вероятностные тематические модели и способы их оценки;
2) выбрать наиболее подходящие для поставленной цели модели и подобрать параметры;
3) построить алгоритм и оценить его работу.

Глава 1. Вероятностное тематическое моделирование

1.1. Основные понятия

Пусть D – множество текстовых документов, W – словарь употребляемых слов коллекции. Предполагается, что порядок слов в документе не имеет значения (гипотеза «мешка слов»), порядок документов в коллекции также не важен. Каждый документ $d \in D$ представляется последовательностью n_d слов $d = w_d = (w_{1_d} \dots w_{n_d})$, $W = \bigcup_1^{|D|} w_d$, то есть документ можно представить как мультимножество: $d \subset W$, в котором $w \in d$ повторяется n_{dw} раз.

Считается, что каждое слово w в документе d связано с темой t из некоторого конечного множества T. Коллекцию документов D можно представить в виде последовательности троек (w_i, d_i, t_i), $i = 1 \dots n$, где выборка каждой тройки случайно и независимо порождена из распределения $p(w, d, t)$ на конечном множестве $W \times D \times T$. Документы d_i и слова w_i являются наблюдаемыми переменными, темы t_i не известны и являются латентными (скрытыми) переменными [1].

Также предполагается, что появление слова w в документе d зависит от темы t, но не от документа (гипотеза условной независимости):

$$p(w|d,t) = p(w|t)\,.$$

Вероятностная тематическая модель коллекции получается из формулы полной вероятности распределения слов в документе коллекции и гипотезы условной независимости и описывает процесс порождения документов коллекции по известным распределениям слов в темах $\varphi_{wt} = p(w|t)$ и тем в документе $\theta_{td} = p(t|d)$:

$$p(w|d) = \sum_{t \in T} p(w|t,d)\, p(t|d) = \sum_{t \in T} p(w|t)\, p(t|d) = \sum_{t \in T} \varphi_{wt}\, \theta_{td}\,. \quad (1)$$

Построение вероятностной тематической модели является обратной задачей, то есть это подбор параметров φ_{wt} и θ_{td} по имеющейся коллекции.

Можно рассмотреть равенство (1) в матричном виде. Так как число тем $|T|$ обычно существенно меньше количества документов $|D|$ и объема словаря $|W|$, решение задачи сводится к поиску приближения к заданной матрице частот слов в документах $P = \left(\hat{p}(w|d)\right)_{W\times D}$, $\hat{p}(w|d) = \frac{n_{dw}}{n_d}$ в виде произведения двух неизвестных стохастических матриц: $P \approx \Phi\Theta$, где матрица слова-темы $\Phi = (\varphi_{wt})_{W\times T}$, а матрица темы-документы $\Theta = (\theta_{td})_{T\times D}$, и ранг разложения не должен превышать $|T|$. Поиск этих матриц производится методом максимизации логарифма правдоподобия коллекции:

$$\ln L(\Phi,\Theta) = \sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td} \to \max_{\Phi,\Theta} \quad . \qquad (2)$$

Таким образом, тематическое вероятностное моделирование выявляет тематическую кластерную структуру коллекции, определяет тематику входящих в нее документов и дает описание каждой темы на естественном языке с помощью распределений $p(w|t)$.

Дополнительно вводится оператор norm. Этот оператор преобразует заданный вектор $(x_i)_{i\in I}$ в вектор вероятностей $(p_i)_{i\in I}$ дискретного распределения через обнуление отрицательных элементов и нормировки:

$$p_i = \mathrm{norm}_{i\in I}(x_i) = \frac{\max\{0, x_i\}}{\sum_{j\in I}\max\{0, x_j\}} \quad . \qquad (3)$$

Соответственно, если $x_i \le 0$ для всех $i \in I$, то оператор norm выдаст нулевой вектор.

1.2. Модель PLSA

Модель PLSA (probabilistic latent semantic analysis), или вероятностный латентно-семантический анализ, - первая вероятностная тематическая модель, которая была предложена Hofmann T. [8] в 1999 году.

Модель документа представляется в виде равенства (1) при ограничениях неотрицательности и нормировки столбцов φ_t и θ_d:

$$\sum_{w \in W} \varphi_{wt} = 1; \quad \varphi_{wt} \geq 0; \qquad \sum_{t \in T} \theta_{td} = 1; \quad \theta_{td} \geq 0\,. \tag{4}$$

Для решения задачи максимизации логарифма правдоподобия (2) при условиях (4) обычно используется EM-алгоритм. EM-алгоритм представляет собой двухшаговый итерационный процесс. Сначала выбираются начальные приближения параметров φ_{wt} и θ_{td}, обычно это произвольные положительные нормированные векторы.

$$n_{tdw} = n_{dw} p_{tdw}\,, \tag{5}$$

$$\varphi_{wt} = \frac{n_{wt}}{n_t}, \quad \theta_{td} = \frac{n_{td}}{n_d}\,, \tag{6}$$

$$n_{wt} = \sum_{d \in D} n_{tdw}\,,\; n_t = \sum\nolimits_{w \in W} n_{wt},\;\; n_{td} = \sum\nolimits_{w \in W} n_{tdw}\,,\; n_d = \sum\nolimits_{t \in T} n_{tdw}\,. \tag{7}$$

На E-шаге (expectation) по формуле (5) вычисляется вспомогательная переменная n_{tdw} – оценка числа употреблений слова w темы t в документе d, где p_{tdw} выражает тематику слова w в документе d по формуле Байеса:

$$p(t|d,w) = \frac{p(w,t|d)}{p(w|d)} = \frac{p(w|t)p(t|d)}{p(w|d)} = \frac{\varphi_{wt}\theta_{td}}{\sum_s \varphi_{ws}\theta_{sd}} = p_{tdw}\;\;.$$

На M-шаге (maximization) пересчитываются параметры модели по формулам (6)-(7), где n_{td} – оценка числа слов темы t в документе d, а n_{wt} – оценка числа употреблений слова w темы t во всей коллекции.

Формулы (6) можно записать следующим образом:

$$\varphi_{wt} \propto n_{wt}, \quad \theta_{td} \propto n_{td}\,,$$

где знак пропорциональности $\propto$ означает, что для получения распределения выражение справа нужно нормировать (оператор norm введен в параграфе 1.1. формулой (3)).

1.3. Модель LDA

Модель LDA (latent Dirichlet allocation), или латентное размещение Дирихле, был предложен Blei D.M. et al. [4] в 2003 году. Эта модель является самой используемой в вероятностном тематическом моделировании, поскольку она получает более интерпретируемые темы.

LDA основывается на предположении, что параметры (Φ, Θ) являются случайными переменными и подчиняются априорному распределению $p(\Phi, \Theta; \gamma)$ с неслучайным вектором гиперпараметров γ, и в качестве априорного распределения берется распределение Дирихле с параметрами $\beta \in \mathbb{R}^{|W|}$ и $\alpha \in \mathbb{R}^{|T|}$ соответственно:

$$\mathrm{Dir}(\theta_d; \alpha) = \frac{\Gamma(\alpha_0)}{\prod_t \Gamma(\alpha_t)} \prod_t \theta_{td}^{\alpha_t - 1},$$

$$\alpha_t > 0, \quad \alpha_0 = \sum_t a_t, \quad \theta_{td} > 0, \quad \sum_t \theta_{td} = 1;$$

$$\mathrm{Dir}(\varphi_t; \beta) = \frac{\Gamma(\beta_0)}{\prod_w \Gamma(\beta_w)} \prod_w \varphi_{wt}^{\beta_t - 1},$$

$$\beta_w > 0, \quad \beta_0 = \sum_w \beta_w, \quad \varphi_{wt} > 0, \quad \sum_w \varphi_{wt} = 1.$$

где $\Gamma(z)$ – гамма-функция. Параметр α отвечает за выраженность тем в документах. Чем меньше будет α, тем более разреженным окажется вектор распределения. Параметр β определяет разреженность вектора, описывающего распределение слов в теме.

Распределение Дирихле порождает векторы дискретных распределений слов в темах $\varphi_{wt} = p(w|t)$, которые становятся центрами тематических кластеров. Порождаемые векторы распределений могут быть и разреженными, и плотными. Чем меньше β_w, тем ближе к нулю условные вероятности φ_{wt} в векторах φ_t. Предположение о разреженности распределений $\varphi_{wt} = p(w|t)$ выводит к естественному предположению о существовании семантического

ядра для каждой темы. Под семантическим ядром понимаются слова, характеризующую определенную тему и, как следствие, имеющие в ней большие вероятности. Количество слов в семантическом ядре считают небольшим.

Предположения о разреженности применимы и для распределений тем в документе $\theta_{td} = p(t|d)$ и приводят к естественному предположению, что каждый документ относится к небольшому количеству тем.

Предположение о подчинении неизвестных параметров (Φ, Θ) априорному распределению $p(\Phi, \Theta; \gamma)$ приводит к принципу максимума логарифма апостериорной вероятности:

$$\sum_{d \in D} \sum_{w \in W} n_{dw} \ln \sum_{t \in T} \varphi_{wt} \theta_{td} + \ln p(\Phi, \Theta; \gamma) \to \max_{\Phi, \Theta, \gamma} . \qquad (8)$$

Для обучения модели LDA чаще всего используют путь, называемый байесовским выводом, которое заключается в вычислении совместного апостериорного распределения для параметров модели и скрытых переменных $p(Z, \Phi, \Theta | \mathrm{X}, \alpha, \beta)$ в явном виде, то есть вместо точечной оценки параметров строятся их распределения. Теоретически такой подход дает информацию о устойчивости параметров и позволяет строить доверительные и интервальные оценки, но на практике получаемые апостериорные распределения чаще переводят в точечные оценки. Наиболее популярными являются вариационный байесовский вывод [11] и сэмплирование Гиббса [13]. Причем оба подхода приводят к EM-подобным алгоритмам. Сэмплирование по Гиббсу обычно проще расширить на модификации LDA, вариационный подход же быстрее и часто стабильнее.

В вариационном байесовском выводе находят приближение апостериорного распределения в виде разложения на множители (оператор norm считается формулой (3)):

$$q(\mathrm{Z},\Phi,\Theta)=\prod_{i=1}^{n}q_i(t_i)\prod_{t\in T}q_t(\varphi_t)\prod_{d\in D}q_d(\theta_d)\,,$$

$$\ln q_t(\varphi_t)=\ln \mathrm{Dir}\left(\varphi_t|\tilde{\beta}_t\right),\qquad \ln q_d(\theta_d)=\ln \mathrm{Dir}(\theta_d|\tilde{\alpha}_d)\,,$$

$$q_i(t)=\mathrm{norm}_{t\in T}\left(\frac{n_{w_it}+\beta_{w_i}-0.5}{n_t+\beta_0-0.5}\cdot\frac{n_{td_i}+\alpha_t-0.5}{n_{d_i}+-0.5}\right).$$

Формула $q_i(t)$ похожа на E-шаг в общем EM-алгоритме для модели LDA: $p(t|d_i,w_i)=\mathrm{norm}_t\left(\varphi_{w_it}\theta_{td_i}\right)$ (шаги общего EM-алгоритма сформулированы в параграфе 1.3), а M-шаг полностью совпадает, если полагать $q_i(t)=\ p(t|d_i,w_i)$:

$$n_{wt}=\sum_{i=1}^{n}[w_i=w]q_i(t),\qquad n_{td}=\sum_{i=1}^{n}[d_i=d]q_i(t)\ \ .$$

По окончании итераций искомые параметры модели можно оценить через математическое ожидание апостериорных распределений Дирихле:

$$\varphi_{wt}=\mathrm{norm}_{w\in W}(n_{wt}+\beta_w),\qquad \theta_{td}=\mathrm{norm}_{t\in T}(n_{td}+\alpha_t)\,.$$

Модель LDA также актуальна для задач классификации [10].

Имеется некоторая обучающая выборка документов d, каждому из которых присвоена метка классов подмножества $C_d\subset C$. Задача состоит в классифицировании новых документов по классам из C. Используется следующая вероятностная модель классификации:

$$\hat{C}_d=\left\{c\in C\mid p(c|d)=\sum_{t\in T}\varphi_{ct}\theta_{td}\geq\gamma_c\right\}.$$

Коэффициенты модели $\varphi_{ct}=p(c|t)$ и пороги γ_c обучаются по выборке документов с известными C_d. Признаковое описание нового документа θ_d вычисляется тематической моделью только по его терминам.

1.4. Аддитивная регуляризация тематических моделей

Как упоминалось в параграфе 1.1., построение вероятностной тематической модели сводится к задаче стохастического матричного разложения вида $P \approx \Phi\Theta$. В общем случае множество разложений не единственно, так как если $\Phi\Theta$ является решением задачи, то $(\Phi S)(S^{-1}\Theta)$ тоже является решением для всех невырожденных матриц S, при которых матрицы ΦS и $S^{-1}\Theta$ тоже стохастические. То есть эта задача некорректно поставлена. Существует общий поход для решения некорректно поставленных задач, называемый регуляризацией [3]. Она заключается в добавлении дополнительных критериев, регуляризаторов, учитывающих специфику решаемой задачи.

1.4.1. Общий подход

Аддитивная регуляризация тематических моделей (ARTM) [12] основана на линейной комбинации логарифма правдоподобия (2) и нескольких регуляризаторов $R_i(\Phi, \Theta), i = 1 \dots k$:

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \varphi_{wt}\theta_{td} + \sum_{i=1}^{k} \tau_i R_i(\Phi, \Theta) \to \max_{\Phi,\Theta} \quad . \tag{9}$$

при условиях (4), где τ_i – неотрицательные коэффициенты регуляризации. В результате скаляризации, вектора критериев оказались преобразованы в один скалярный критерий: $R(\Phi, \Theta) = \sum_{i=1}^{k} \tau_i R_i(\Phi, \Theta)$.

Таким образом, ARTM представляет собой PLSA с дополнительными критериями. Важно отметить, что ARTM не является ещё одной моделью или ещё одним методом. Это общий подход к построению и комбинированию тематических моделей.

Задача (4), (9) решается с помощью EM-алгоритма.

$$n_{tdw} = n_{dw} p_{tdw}\,, \qquad p_{tdw} = \mathrm{norm}_{t \in T}(\varphi_{wt}\theta_{td})\,; \tag{10}$$

$$\varphi_{wt} = \text{norm}_{w \in W}\left(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right), \qquad n_{wt} = \sum_{d \in D} n_{dw} p_{tdw} \ ; \tag{11}$$

$$\theta_{td} = \text{norm}_{t \in T}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right), \quad n_{td} = \sum_{w \in W} n_{dw} p_{tdw} \ . \tag{12}$$

Вычисления по формуле (10) дают E-шаг алгоритма, а оценивание параметров по формулам (11) и (12) дают M-шаг алгоритма (оператор norm вычисляется по формуле (3)). При $R(\Phi, \Theta) = 0$ формулы (11)-(12) переходят в формулы M-шага для PLSA (6)-(7). То есть модель становится чистой PLSA без регуляризаторов.

Разные варианты EM-алгоритма рассматриваются в [1].

Регуляризатор R может слишком разреживать модель, и тогда могут появиться вырожденные темы или документы. Тема t считается вырожденной, если $n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}} \leq 0$ для всех $w \in W$. Аналогично документ d считается вырожденным, если $n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}} \leq 0$ для всех $t \in T$. Документ может оказаться вырожденным, если, например, он слишком короткий для определения тематики. Вырожденные темы и документы сокращаются из модели.

1.4.2. Разновидности регуляризаторов

В ARTM все требования формализуются в виде регуляризаторов R_i, принимающих вид гладких функций для удобства вычислений на M-шаге, и настраиваются с помощью коэффициентов τ_i, которые для каждой задачи приходится подбирать экспериментально, что является существенной проблемой. На практике коэффициенты τ_i могут изменяться в ходе итераций. Более того, некоторые регуляризаторы рекомендуют включать лишь на поздних этапах, другие советуют использовать лишь на начальных итерациях, а некоторые и вовсе могут нейтрализовать друг друга, поэтому их следует применять поочередно.

В основном критерии накладываются на темы в вероятностной тематической модели: они должны быть интерпретируемы, различны, разрежены и т.д.

Дивергенция Кульбака-Лейблера (Kullback–Leibler divergence; KL-дивергенция) [9] является одним из важнейших инструментов конструирования регуляризаторов. KL-дивергенция представляет собой несимметричную функцию расстояния между двумя дискретными распределениями $P = (p_i)_{i=1}^n$ и $Q = (q_i)_{i=1}^n$ с совпадающими носителями, $\{i: p_i > 0\} = \{i: q_i > 0\}$, при этом обычно под первым аргументом (распределение P) подразумевают истинное распределение, а по вторым (распределение Q) – проверяемое:

$$KL(P \parallel Q) \equiv KL(p_i \parallel q_i) = \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i} = H(P,Q) - H(P)\,,$$

где $H(P,Q) = -\sum_i p_i \ln q_i$ и $H(P) = -\sum_i p_i \ln p_i$ являются кросс-энтропией (также называют перекрестной) пары распределений (P,Q) и энтропией распределения P соответственно.

Несколько важных свойств дивергенции Кульбака-Лейблера:

- $KL(P \parallel Q) \geq 0$, причем $KL(P \parallel Q) = 0 \Leftrightarrow p_i \equiv q_i$.
- Несимметричность: $KL(P \parallel Q) \neq KL(Q \parallel P)$. При этом если $KL(P \parallel Q) \leq KL(Q \parallel P)$, то распределение P вложено в распределение Q сильнее, чем Q в P. То есть KL-дивергенцию можно считать мерой вложенности двух распределений.
- Если P – эмпирическое распределение, а $Q(\alpha)$ – параметрическая модель, то минимизация KL-дивергенции эквивалентна минимизации кросс-энтропии и максимизации правдоподобия:

$$KL\big(P \parallel Q(\alpha)\big) = \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i(\alpha)} \to \min_{\alpha} \quad \Leftrightarrow \quad \sum_{i=1}^{n} p_i \ln q_i(\alpha) \to \max_{\alpha} \quad .$$

- Максимизация правдоподобия (2) эквивалентна минимизации взвешенной суммы KL-дивергенций между эмпирическими распределениями $\hat{p}(w|d) = \frac{n_{dw}}{n_d}$ и модельными $p(w|d)$ по всем документам $d \in D$:

$$\sum_{d \in D} n_d KL_w \left(\frac{n_{dw}}{n_d} \parallel \sum_{t \in T} \varphi_{wt} \theta_{td} \right) \to \min_{\Phi, \Theta} \quad ,$$

где весом документа d является его длина n_d. Если веса n_d убрать, то все документы буду искусственно приведены к одинаковой длине.

Однако, в силу своей несимметричности, KL-дивергенция не всегда подходит. Поэтому также будет использоваться дивергенция Йенсена-Шеннона (Jensen-Shannon divergence) [6]. Дивергенция Йенсена-Шеннона основана на дивергенции Кульбака-Лейблера, с некоторыми заметными и полезными различиями, например, она симметричная. Идея этой дивергенции заключается в том, что расстояние между двумя распределениями не может сильно отличаться от среднеарифметического значения расстояний до их среднего распределения.

$$JSD(P \parallel Q) = \frac{1}{2} KL(P \parallel M) + \frac{1}{2} KL(Q \parallel M), \qquad (13)$$

где $M = \frac{1}{2}(P + Q)$. При вычислении логарифма KL-дивергенции с основанием 2 дивергенция Йенсена-Шеннона принимает значения из диапазона [0, 1]. Если же использовать натуральный логарифм, $0 \leq JSD(P \parallel Q) \leq \ln(2)$. Чем ближе значение к нулю, тем лучше аппроксимация исходного распределения, т.е. тем меньше было потеряно информации.

- *Регуляризатор модели LDA.*

Для модели LDA в 1.3. условие порождения параметров модели распределением Дирихле по сути своей является регуляризатором. В соответствии с (8), регуляризатор LDA с точностью до константы равен логарифму априорного распределения, то есть распределения Дирихле:

$$R(\Phi,\Theta) = \ln \prod_{t\in T} \mathrm{Dir}(\varphi_t;\beta) \prod_{d\in D} \mathrm{Dir}(\theta_d;\alpha) + \mathrm{const} =$$

$$= \sum_{t\in T}\sum_{w\in W} (\beta_w - 1)\ln \varphi_{wt} + \sum_{d\in D}\sum_{t\in T} (\alpha_t - 1)\ln \theta_{td}\,. \qquad (14)$$

При $\beta_w = 1, \alpha_t = 1$ формулы М-шага (11)-(12) переходят в (6)-(7), а модель LDA становится PLSA [7]. При $\beta_w > 1,\ \alpha_t > 1$ регуляризатор сглаживает, то есть большие вероятности становятся еще больше, а малые – еще меньше, но не достигают нуля. При $0 < \beta_w < 1,\ 0 < \alpha_t < 1$ регуляризатор разреживает, то есть малые вероятности могут обнулиться.

Таким образом, LDA можно представить как ARTM с единственным регуляризатором вида (14).

Этот же регуляризатор LDA можно расписать через KL-дивергенции:

$$R(\Phi,\Theta) = |W| \sum_{t\in T} KL_w\left(\frac{1}{|W|} \parallel \varphi_{wt}\right) - \beta_0 \sum_{t\in T} KL_w\left(\frac{\beta_w}{\beta_0} \parallel \varphi_{wt}\right) +$$

$$+ |T| \sum_{d\in D} KL_t\left(\frac{1}{|T|} \parallel \theta_{td}\right) - \alpha_0 \sum_{d\in D} KL_t\left(\frac{\alpha_t}{\alpha_0} \parallel \theta_{td}\right)\quad .$$

- *Регуляризатор разреживания/сглаживания.* [12]

Предполагается, что семантическое ядро каждой темы состоит из небольшого количества слов, и каждый документ можно отнести к небольшому количеству тем. Тогда большая часть вероятностей φ_{wt} и θ_{td} обнуляются. Разреженность тем необходима для их интерпретируемости и позволяет сократить время вычислений для построения модели и ее дальнейшего использования.

Данный регуляризатор выводится из регуляризатора модели LDA (14) после снятия ограничений на неотрицательность параметров α_t и β_w, а также задания этих параметров индивидуально для каждого элемента матриц Φ и Θ:

$$R(\Phi,\Theta) = \sum_{t\in T}\sum_{w\in W} \beta_{wt}\ln \varphi_{wt} + \sum_{d\in D}\sum_{t\in T} \alpha_{td}\ln \theta_{td}\,. \qquad (15)$$

Применение этого регуляризатора дает следующие формулы М-шага в ЕМ-алгоритме (получаются из формул (11)-(12) при использовании (15)):

$$\varphi_{wt} = norm_{w \in W}(n_{wt} + \beta_{wt});\ \theta_{td} = norm_{t \in T}(n_{td} + \alpha_{td})\ .$$

Для сглаживания подбираются положительные параметры α_{td} и β_{wt}, для разреживания – отрицательные.

Формулу (15) сглаживающего или разреживающего регуляризаторов также можно получить через минимизацию или максимизацию KL-дивергенций между распределениями φ_t и θ_d и заданными дискретными распределениями α и β.

При таком регуляризаторе нет необходимости в априорных распределениях Дирихле.

- *Регуляризатор декорреляции тем.*[12]

Чем меньше в модели дублирующихся или похожих тем, тем она информативнее. Уменьшение корреляции между вектор-столбцами разных тем приводит к повышению различности тем, следовательно, улучшает их интерпретируемость. Таким образом необходимо минимизировать сумму попарных скалярных произведений столбцов матрицы Ф $\langle \varphi_t, \varphi_s \rangle = \sum_w \varphi_{wt}\varphi_{ws}$. Полученный регуляризатор принимает вид:

$$R(\Phi) = -\frac{\tau}{2}\sum_{t \in T}\sum_{s \in T \setminus t}\sum_{w \in W}\varphi_{wt}\varphi_{ws}\ . \qquad (16)$$

Формула М-шага (получается из формулы (11) при использовании (16)):

$$\varphi_{wt} = norm_{w \in W}\left(n_{wt} - \tau\varphi_{wt}\sum_{s \in T \setminus t}\varphi_{ws}\right).$$

В результате этого регуляризатора в каждой строке вероятности φ_{wt} наиболее значимых тем слова w увеличиваются, а вероятности остальных тем уменьшаются и даже могут обнулиться. Соответственно, получается еще и эффект разреживания.

Основные рекомендации использования регуляризаторов: декоррелирование и сглаживание включать сразу, разреживание – после 10-20 итераций, когда параметры модели начнут сходиться.

1.5. Метрики для оценки качества модели.

В оценивании построенной тематической модели нет четкого понятия ошибки, а критерии качества кластеризации вроде среднего расстояния между кластерами не особо эффективны при нечеткой кластеризации. К тому же от получаемых в ходе построения тематической модели распределений φ_{wt} и θ_{td} требуется соответствие выбранным условиям вроде разреженности, различности или интерпретируемости тем.

Следующие метрики являются наиболее распространенными критериями:

- *Перплексия* (perplexity) – мера несоответствия модели $p(w|d)$ словам w, документам d из коллекции D. Она отслеживается через правдоподобие (2):

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

$$n = \sum_{d\in D}\sum_{w\in W} n_{dw}\ , p(w|d) = \sum_{t\in T} \varphi_{wt}\theta_{td}\ .$$

Чем меньше величина перплексии, тем лучше модель. Однако ее численные значения неочевидно интерпретируются.

- *Разреженность и различность тем*. Разреженность модели измеряется долей нулевых элементов в матрицах Φ и Θ.

Предполагается, что любая интерпретируемая тема содержит лексическое ядро, под которым понимается множество слов, отличающих эту тему от остальных, соответственно они с большей вероятностью употребляются в этой теме и редко употребляются в других: $W_t =$

$\{w \in W \mid p(t|w) > \delta\}$. Затем по ядру определяются показатели интерпретируемости тем:

- Чистота темы – суммарная вероятность слов ядра. Чем выше значение показателя, тем лучше данная тема описывается своим ядром.

$$pur_t = \sum_{w \in W_t} p(t|w) = \sum_{w \in W_t} \varphi_{wt} .$$

- Контрастность темы – средняя вероятность встретить слова ядра в данной теме. Чем выше значение показателя, тем однозначнее угадывается тема по своему ядру.

$$Con\ (t) = \frac{1}{|W|} \sum_{w \in W_t} p(t|w) .$$

Глава 2. Предлагаемый алгоритм

Удовлетворяющая поставленной цели модель классификации должна соответствовать следующим критериям:

- Реализуется многозначная классификация, то есть документ может быть отнесен к нескольким классам.
- Классификация выполняется с помощью вероятностного тематического моделирования.
- Классы модели небольшие.

Подразумевается, что слова, характеризующие какой-либо класс, описывают относительно небольшую интерпретируемую тему (сравнение идет с тематикой всей коллекции). К примеру, класс «физика» будет считаться обобщенным даже в коллекции с большой разнообразностью тем. Классы же «механика», «термодинамика», «электродинамика» и другие, по факту формирующие обобщенный класс «физика», относительно него являются небольшими. И если при этом документы, относящиеся к этим классам, не формируют большинство коллекции, то эти классы удовлетворяют критерию.

При увеличении количества классов, соответствующие им темы будут сужаться. Следовательно, необходимо установить такое число тем, при котором темы не будут считаться обобщенными, и при этом сохранят свою интерпретируемость.

- Если сравнительно небольшая доля документа соответствует тематике какого-то класса, то документ все равно будет к нему отнесен.

Этот критерий имеет значение на этапе дальнейшего использования построенной модели. Например, при сфокусированном тематическом поиске по построенной модели среди релевантных документов окажутся даже те, в которых искомая тема не является основной, но тем не менее фигурирует в тексте. Более подробно в параграфе 3.3. текущей главы.

- Отсутствует готовая качественная обучающая выборка.

Обычно для проверки методов классификации используют какую-нибудь коллекцию документов из открытых порталов. Несомненным достоинством такого подхода является уже имеющаяся разметка классов на документах всей коллекции, поэтому вопрос формирования обучающей выборки и проверки правильности работы метода решается тривиально: имеющиеся документы делят на обучающую и контрольную выборки в соотношении 9 к 1 и используют метрики оценки качества вроде F-меры или площади под кривой ошибок (ROC AUC), которые основаны на сравнении правильной метки и метки, полученной в результате метода.

Если же коллекция не имеет меток принадлежности к классам, или же классы, по которым есть разметка, являются слишком обобщенными и, следовательно, не подходят для обучения, появляется задача формирования качественной обучающей выборки. То есть нужно не только сформулировать небольшие классы, но также каждый документ выборки отнести к нескольким подобным классам. Учитывая, что объем выборки должен быть достаточно большим для наиболее успешного обучения модели, ее формирование экспертом, как это обычно происходит, требует слишком много времени и усилий.

В данной работе предлагается следующий подход для решения этой задачи (более подробно в параграфе 3.2. данной главы): отбирается часть имеющейся коллекции, на которой строится вероятностная тематическая модель, решающая задачу мягкой кластеризации. Далее остается установить порог вероятности принадлежности документа к теме. В результате получаем пары $\langle \text{документ}, \text{вектор тем} \rangle$, которые можно использовать в качестве обучающей выборки для классификатора модели по всей коллекции.

Для немаркированной коллекции выбираются случайные документы, для промаркированной желательно сохранить соотношение количеств документов промаркированных классов, как в исходной коллекции.

При подобном подходе можно построить модель по достаточно небольшим классам, однако выявленные классы не будут иметь названий, понятных для человека. Они будут характеризоваться некоторым набором слов, который не всегда можно назвать интерпретируемым. В параграфе 3.3 текущей главы рассказано подробнее, почему этот недостаток не является серьезной проблемой.

Сформулируем этапы выполнения алгоритма построения вероятностной тематической модели классификации с использованием предложенного подхода:

1. Предварительная обработка текста.
2. Создание обучающей выборки и обучение классификатора.
3. Оценка качества построенной модели

Алгоритм был реализован на коллекции из 1208 документов. Каждый документ представляет собой выпускную квалификационную работу студента, написанную на русском языке. Коллекция собрана в частном порядке в основном из открытых репозиториев университетов РФ и не имеет маркировок принадлежности к готовым классам.

Стоит отметить, что подобный выбор документов обусловлен лишь персональным предпочтением и в целом не влияет на результаты. Небольшая разница есть лишь на этапе подготовки текстов к анализу (пункт 6 в параграфе 3.1. данной главы).

Реализация алгоритма проводилась на языке Python 3.7.6.

3.1. Предварительная обработка текста.

Документы коллекции содержатся в формате pdf. Для извлечения текста из pdf файлов был использован инструмент XpdfReader (функция pdftotext) [15]. Полученные текстовые файлы могут содержать некорректно конвертированные слова, опечатки. Да и в целом перед последующим использованием и анализом все тексты нуждаются в предварительной

обработке, которая состоит из следующих этапов (результаты применения этапов к имеющейся коллекции в Таблице 1):

1) Перевод всех слов в нижний регистр.

2) Токенизация текста (использовался токенизатор из библиотеки nltk).

3) Удаление символов не русского языка с помощью регулярных выражений python (используется модуль re). Регулярные выражения (regular expression) – специальные шаблоны для поиска подстрок в тексте. С их помощью текст можно избавить от объектов, не имеющих смысловую нагрузку: формулы, числа, гиперссылки и прочее.

4) Удаление некорректно конвертированных слов. К ним можно отнести слова длиной больше двадцати символов и слова, в которых встречаются три и более подряд идущих одинаковых символов. Подобные ошибки вероятнее всего появились в результате склеивания после неправильного конвертирования.

5) Удаление стоп-слов (их список stopwords взят из библиотеки nltk). Стоп-словами называют слова, не несущие смысловой нагрузки, и их удаление осуществимо без потери важной информации текста. К стоп-словам относят: союзы и союзные слова, предлоги, местоимения, междометия, цифры, знаки препинания, вводные слова и выражения, ряд некоторых существительных, глаголов и наречий.

6) Удаление фраз, присущих типу документов рассматриваемой коллекции. Так как данные обрабатываемые документы являются научными работами студентов, в них ожидаемо появление шаблонов, некоторые из которых могут употребляться часто, но при этом не иметь стоящей смысловой нагрузки. В Приложении перечислен список таких слов.

Для обработки других типов текстов список будет различаться или вообще отсутствовать.

7) Проведение процедуры стемминга (взят стеммер SnowballStemmer из nltk). Стемминг обозначает выделение основы слова, то есть ее неизменяемой части (не всегда совпадает с морфологическим корнем).

На практике при построении вероятностной тематической модели семантические ядра полученных в результате кластеризации тем могут быть плохо интерпретируемы. Из-за процедуры стемминга появились дополнительные проблемы понимания некоторых получившихся основ слов, что, соответственно, лишь ухудшило интерпретируемость тем. Поэтому вместо стемминга в данной работе применялась лемматизация, то есть приведение слова к его нормальной форме (использовался лемматизатор mystem из библиотеки pymystem3). К тому же, использование лемматизатора также выполняет пункт 2) (разделение текста на слова).

Следует также отметить, что в научных работах присутствуют списки использованной литературы. Статьи английского авторства удаляются в результате пункта 3), русские же названия и имена авторов остаются. Названия использованных для написания работы статей соответствуют тематикам этой работы, поэтому в целом не мешают определению собственно тематик. Имена авторов же, забегая вперед, образуют отдельную тему, по факту характеризующую все документы в той или иной степени.

Таблица 1. Поэтапные результаты предварительной обработки текстов коллекции

Выполненные этапы обработки	Общее количество слов в коллекции, W_{all}	Количество уникальных слов коллекции (словарь), W
1)-3)	4 146 098	530 461
4)	4 028 730	431 896
5)	2 516 146	427 212
6)	2 187 423	427 152
7)	2 187 423	109 734

В среднем 3,4 тысячи слов было токенизировано для каждого документа. Словарь W составляет ~12,8% от общего количества слов в коллекции W_{all}. Следует учитывать, что одно и то же слово в разных формах будет считаться разными словами до этапа лемматизации.

Почти 3% W_{all} оказались некорректно конвертированы, это примерно 97 слов в каждом документе, при этом ~16% от общего количества таких слов повторялись в разных документах, то есть ошибочное конвертирование для некоторых фраз и слов происходило нередко.

~36,5% W_{all} удалились как стоп-слова, при этом набор уникальных стоп-слов оказался в сотни раз меньше.

~8% W_{all} составляли шаблонные слова, указанные в Приложении. Список состоит из всего 60 слов, поэтому W сократился не сильно.

Но после лемматизации W резко уменьшился в почти 4 раза, а W_{all} осталось неизменным, поскольку ничего не удалялось.

В конечном итоге W_{all} сократилось на ~47%, то есть в одном документе теперь осталось в среднем ~1,8 тысяч слов.

3.2. Создание обучающей выборки и обучение классификатора

Для формирования обучающей выборки из данной коллекции в 1208 документов случайным образом были отобраны 100 документов.

Сначала состоялась попытка создать маркировку вручную, в результате которой документы выборки оказались распределены по 50 классам, в среднем каждый документ был отнесен к 2-3 классам. Однако некоторые классы все равно получились достаточно обобщенными в силу отсутствия необходимых знаний предметной области для более узкой формулировки.

Экспертный подход создания качественной обучающей выборки, классификация по которой будет соответствовать условиям поставленной

цели, в принципе реализуем. Однако он требует больших усилий и имеет высокий риск неудачи.

Для решения задачи мягкой кластеризации на полученной выборке в 100 документов использовался подход ARTM, то есть решалась задача (9), с применением регуляризаторов.

Использовалась BigARTM v0.8.0 – open source библиотека для тематического моделирования, основывающаяся на подходе ARTM. BigARTM рекомендуется использовать на Python 2.7, однако есть и возможность использовать Python 3. [14].

Входные данные должны соответствовать гипотезе «мешка слов», то есть вся предварительная обработка текста должна быть сделана заранее с использованием других библиотек (что и описано в предыдущем параграфе). Есть два варианта формата входных данных:

- Vowpal Wabbit – представляется через текстовый файл, в котором каждый документ отображается через одну строчку, каждое слово документа является отдельным токеном текстового формата.
- UCI Bag-of-words – представляется через «мешок слов» – файл, который содержит набор строк вида $\langle docID \;\; wordID \;\; count \rangle$, где $docID$ – номер документа коллекции, $wordID$ – номер слова из словаря, $count$ – количество использований слова с указанным номером в документе с указанным номером.

После предварительной обработки отобранные 100 документов были приведены к первому из перечисленных форматов. Следующий шаг – создание словаря по полученным на вход данным. После, наконец, можно создать саму модель.

В BigARTM реализованы онлайн и оффлайн EM-алгоритмы обучения моделей. Онлайн EM-алгоритм применяется при обработке больших коллекций в потоковом режиме. В нем лишь один раз проходятся по

коллекции, при этом реализуется многократное итерирование по документу. Коллекция разбивается на пакеты, и каждый пакет обрабатывается при своей фиксированной матрице Φ (то есть она обновляется через определенное количество обработанных документов), а матрицу Θ нет необходимости хранить: тематическую модель можно сразу использовать в процессе. Оффлайн EM-алгоритм применяется при обработке небольших коллекций. В нем реализуются многократное итерирование по коллекции и однократный проход по документу. Матрица Φ обновляется в конце каждой итерации по коллекции, и появляется необходимость хранить матрицу Θ.

Для построения модели применялся оффлайн EM-алгоритм с 33 итерациями по коллекции и количеством тем 200.

Использовались следующие регуляризаторы (в комбинациях):

a) регуляризатор декорреляции тем (16) с параметром $\tau = 10^7$ добавлялся после 2 итераций;

b) регуляризатор разреживания матрицы Θ (15) с параметром $\tau = -1{,}5$ добавлялся после 18 итерации.

Model 1 – чистая модель PLSA без регуляризаторов.

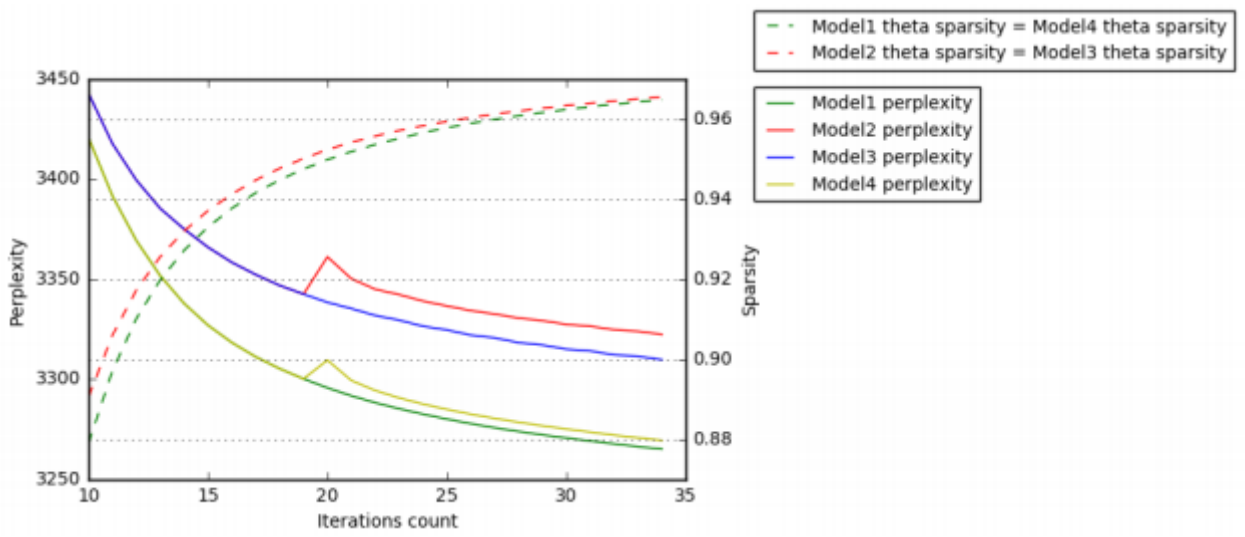
Model 2 – модель с добавлением обоих регуляризаторов a) и b).

Model 3 – модель с добавлением лишь a).

Model 4 – модель с добавлением лишь b).

В качестве оценок качества модели высчитывались перплексия и разреженность матрицы Θ в виде чистоты тем. Изменение их значений показано на Рис.1.

Рис.1. Изменение оценок качества моделей 1-4.



Для каждой модели набор полученных тем оценивался экспертом по спискам топ-10 характеризующих тему слов (Таблица 2). Оценки поделены на 3 условные категории: хорошо (можно понять, что объединяет слова темы), средне (можно догадаться, что объединяет слова темы, используя некоторые дополнительные знания или исключив одно-два слова) и плохо (понять, что объединяет слова темы, сложно или невозможно) интерпретируемые темы.

Примеры:

Хорошо интерпретируемая тема: [планета, звезда, солнечный, система, астероид, падение, созвездие, пояс, млечный, путь]

Средне интерпретируемая тема: [война, армия, ссср, немецкий, фронт, сирия, сталин, орден, теракт, офицер]

Плохо интерпретируемая тема: [мина, синий, бар, татарин, татарский, узнать, минь, бер, син, тег]

Можно отметить высокое качество модели 2 (использование обоих регуляризаторов) и по оценкам качества модели, и по оцениванию интерпретируемости полученных тем.

Таблица 2. Результаты оценивания полученных наборов тем.

Интерпрети-руемость тем	Модель 1	Модель 2	Модель 3	Модель 4
Хорошо	50%	63%	62%	58%
Средне	16%	18%	10%	14%
Плохо	34%	19%	28%	28%

Модель 2 сформировала достаточно качественную вероятностную тематическую модель на множестве выбранных 100 документов. В качестве обучающей выборки модели классификации по всей коллекции передаются общий список тем $topicsList$ и пары $\langle docID \ \ docTopics \rangle$, где $docID$ – номер документа выборки, а $docTopics$ – список тем, к которым относится документ с указанным номером, сформированный из полученного распределения тем в документе: $docTopics = \{p(t|d) \mid p(t|d) \geq \lambda\} \subset topicsList$. λ – некоторый порог значения вероятности, в данной работе $\lambda = 0{,}05$.

Вследствие введенного порога, если среди всех тем была такая, что имела слишком малую вероятность (меньше λ) описания документов выборки, то есть $t_j : p(t_j|d) < \lambda$, для всех $d \in D_{\text{обуч.}}$ ($D_{\text{обуч.}}$ – множество документов обучающей выборки), то эта тема не вошла ни в одну из переданных классификатору пар. Значит, классификатор не сможет обучиться относить к этой теме документы, тогда эту тему можно удалить из $topicsList$.

В данном случае из 200 тем в $topicsList$ вошли 196, а каждый документ выборки в среднем оказался описан 6-7 темами.

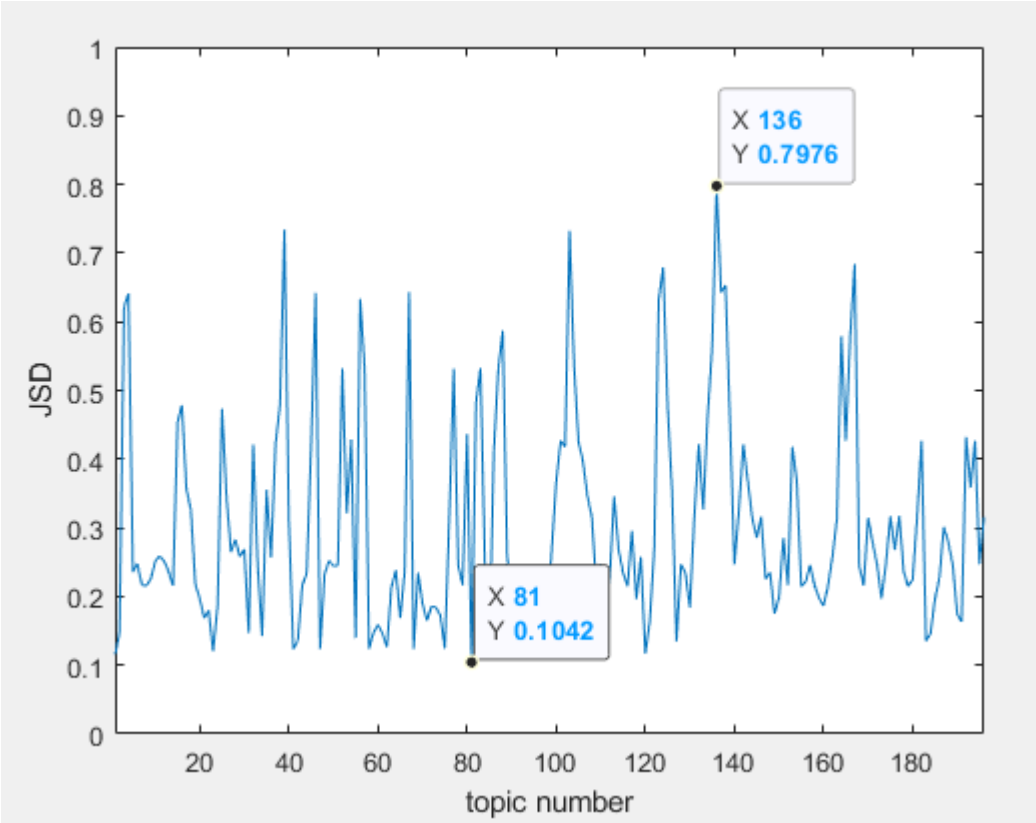
Для построения классификатора была взята модель LDA, а ее можно представить как частный случай ARTM. Таким образом, была реализована множественная классификация через подход ARTM. Для обучения классификатора было сделано 4 итерации по выборке, на каждой итерации по 25 проходов по каждому документу.

3.3. Оценка качества построенной модели и выводы

В результате обучения классификатора было получено приближение распределения слов в темах. При этом мы фактически имеем приближаемое распределение, ведь оно выведено при построении вероятностной тематической модели на множестве обучающей выборки, и классификатор стремится приблизиться именно к этой созданной модели.

Значит, можно посчитать значения (на рис.2) дивергенции Йенсена-Шеннона по формуле (13) между распределением φ_{wt} и его аппроксимацией $\hat{\varphi}_{wt}$, полученной из модели классификации по всей коллекции. Логарифм KL-дивергенции вычислялся с основанием 2.

Рис. 2. Значения дивергенции Йенсена-Шеннона.



Рассмотрим темы, для которых дивергенция Йенсена-Шеннона приняла максимальное и минимальное значения. Так как чем ближе значение дивергенции к нулю, тем лучше, максимальное значение обозначает худшее приближение темы, а минимальное – лучшее. В таблице 3 приведены топ-10

характеризующих эти темы слов из каждого распределения. Справа рассматривается тема с $\min JSD = 0{,}1042$, слева –тема с $\max JSD = 0{,}7976$.

Таблица 3. Топ-10 характеризующих слов тем с максимальным и минимальным значениями дивергенции Йенсена-Шеннона.

Topic_136, φ_{wt}	Topic_136, $\hat{\varphi}_{wt}$	Topic_81, φ_{wt}	Topic_81, $\hat{\varphi}_{wt}$
Минута	День	Акция	Акция
Модель	Ждать	Облигация	Облигация
День	Улыбка	Ценный	Биржа
Растение	Чистый	Биржа	Ценный
Сделать	Растение	Купон	Купон
Реакция	Сажать	Цена	Цена
Теплый	Теплый	Торговать	Торговать
Сажать	Модель	Контракт	Контракт
Руки	Сделать	Вексель	Продавать
Паутина	Руки	Продавать	Премия

Тема с максимальным значением JSD, к тому же, является одной из плохо интерпретируемых. Топ-10 характеризующих рассматриваемую тему слов для данных двух распределений не просто различаются, но также имеют разброс в вероятностях определений темы.

Тема с минимальным значением JSD, наоборот, интерпретируется хорошо, топ-10 характеризующих ее слов для данных распределений почти полностью совпадают.

В целом, тем, для которых $JSD > 0{,}5$, относительно немного. Более того, для большинства тем $0{,}1042 < JSD < 0{,}3$, что является признаком хорошего приближения. Следовательно, обучение классификатора привело к созданию качественной построенной по всей коллекции вероятностной тематической модели.

Таким образом, предложенный в начале главы алгоритм был реализован. Создание обучающей выборки через построение вероятностной тематической модели по выборке показало себя приемлемым подходом.

Однако, существенным отличием от классификаций с готовыми обучающими выборками является отсутствие названий классов, их многочисленность и, в некоторых случаях, плохая интерпретируемость. Впрочем, классификации вроде описанной в данной работе очень удобны для некоторых задач, и тогда этот недостаток обращается в преимущество, в частности, речь идет о сфокусированном информационном поиске. В качестве запроса задается семантическое ядро интересующей темы. Это может быть не только набор характеризующих слов, но и какой-нибудь фрагмент текста. Запрос сначала, конечно же, подвергается предварительной обработке (параграф 3.1.), затем ищутся пересечения между словами запроса и словарем коллекции. Словам из пересечения соответствуют темы коллекции, а темам – документы. Среди выявленных документов нужно будет найти действительно релевантные и ранжировать их. Релевантные документы обычно составляют достаточно малую часть коллекции, поэтому такую задачу, образно говоря, можно назвать «классификацией иголок в стоге сена» [5]. Использование классификаторов позволяет ограничить поиск необходимой информации относительно небольшим подмножеством документов.

Вспомним пример из введения. Пусть имеется документ об исследовании в медицине, в котором используются статистические методы для решения промежуточных задач. Тема «статистический анализ» в таком документе будет покрывать лишь относительно небольшую часть тематики документа. Использование классификатора вроде описанного в этой работе присвоит такому документу метку класса, соответствующего теме «статистический анализ», пусть эта тема и далеко не основная для документа. Таким образом, в дальнейшем при поиске релевантных документов по теме «статистический анализ» упомянутый документ также будет выявлен.

Заключение

В данной работе рассмотрены некоторые вероятностные тематические модели. Выделена проблема отсутствия готовой качественной обучающей выборки для множественной классификации по относительно большому количеству относительно небольших классов. Предложено решение в виде создания обучающей выборки путем мягкой кластеризации через вероятностно тематическую модель. Описан алгоритм построения вероятностной тематической модели множественной классификации коллекции документов с обучением на выборке, созданной в результате предложенного решения.

Описанный в работе алгоритм показал хорошую работу, и модель можно считать достаточно качественной.

Список литературы

[1] Воронцов К.В., Потапенко А.А. Модификации EM-алгоритма для вероятностного тематического моделирования // Машинное обучение и анализ данных. — 2013. — Т. 1, № 6. — С. 657–686.

[2] Коршунов Антон, Гомзин Андрей. Тематическое моделирование текстов на естественном языке // Труды Института системного программирования РАН, 2012. Т. 23. С. 215–244.

[3] Тихонов А.Н., Арсенин В.Я. Методы решения некорректных задач. — М.: Наука, 1986.

[4] Blei D.M., Ng A.Y., Jordan M.I. Latent Dirichlet allocation // Journal of Machine Learning Research. — 2003. — Vol. 3. — Pp. 993–1022.

[5] Bodrunova S., Koltsov S., Koltsova O., Nikolenko S.I., Shimorina A. Interval semisupervised LDA: Classifying needles in a haystack // MICAI (1) / Ed. by F.C. Espinoza, A.F. Gelbukh, M. Gonzalez-Mendoza. — Vol. 8265 of Lecture Notes in Computer Science.— Springer, 2013.— Pp. 265–274.

[6] Fuglede, B., Topsoe F. Jensen-Shannon divergence and Hilbert space embedding // Proceedings of the International Symposium on Information Theory, 2004. IEEE. p. 30.

[7] Girolami M., Kaban A. On an equivalence between PLSI and LDA // SIGIR'03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval. — 2003. — Pp. 433–434.

[8] Hofmann T. Probabilistic latent semantic indexing // Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. — New York, NY, USA: ACM, 1999. — Pp. 50–57.

[9] *Kullback S., Leibler R.A.* On information and sufficiency // The Annals of Mathematical Statistics. 1951. V.22. № 1. P. 79-86.

[10] Rubin T.N., Chambers A., Smyth P., Steyvers M. Statistical topic models for multi-label document classification // Machine Learning. — 2012. — Vol. 88, no. 1-2. — Pp. 157–208.

[11] Teh Y.W., Newman D., Welling M. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation // NIPS. – 2006. – Pp. 1353-1360.

[12] Vorontsov K. V., Potapenko A. A. Additive regularization of topic models // Machine Learning, Special Issue on Data Analysis and Intelligent Optimization with Applications. — 2015. — Vol. 101, no. 1. — Pp. 303–323.

[13] Wang Y. Distributed Gibbs sampling of latent dirichlet allocation: The gritty details, 2008.

[14] BigARTM – open source library. URL: http://bigartm.org

[15] XpdfReader — a free PDF toolkit. URL: http://xpdfreader.com

Приложение

Список шаблонных фраз, присущих выпускным квалификационным работам студентов (часть из них, возможно, включена в список стоп-слов):

Выпускная	Вывод	Решается
Квалификационная	Приложение	Предполагается
Бакалавр	Обзор	Упоминается
Бакалавриат	Постановка	Является
Магистр	Задача	Описывается
Магистратура	Цель	Также
Аспирант	Поставлена	Можно
Аспирантура	Формула	Например
Руководитель	Формулировка	Скажем
Рецензент	Предположение	Называемый
Профессор	Гипотеза	Разумеется
Доцент	Условие	Итак
Преподаватель	Критерий	Однако
Ассистент	Вводится	Тогда
Студент	Выводится	Соответственно
Содержание	Вычисляется	Часто
Введение	Заключается	Обычно
Глава	Принимается	Вероятно
Параграф	Представляется	Вообще
Заключение	Считается	Лишь