Иерархическая мультимодальная тематическая модель коллекции научно-популярных текстов

Ефимова Ирина

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Научный руководитель д.ф.-м.н., профессор РАН К. В. Воронцов

15 июня 2017

Цель исследования

Имеется коллекция текстовых документов, где каждый документ тегирован: редакторами ресурса приписано некоторое количество ключевых слов или фраз.

Цель исследования: разработать методику построения двухуровневой тематической иерархии с автоматическим именованием тем

- задача построения первого уровня иерархии;
- задача построения второго уровня иерархии;
- задача автоматического именования тем.

Задача построения тематических моделей

Дано:

- W^m — словарь токенов m-й модальности, $m \in M$
  $W = W^1 \sqcup \cdots \sqcup W^M$ — объединённый словарь,
- D – коллекция текстовых документов $d = \{w_1, ..., w_{n_d}\}$,
- n_d – длина документа d, n_{dw} – частота термина w в d.

Предположения:

- каждый термин $w \in W$ в $d \in D$ имеет тему $t \in T$;
- $D \times W \times T$ – дискретное вероятностное пространство;
- Коллекция – это i.i.d. выборка $(d_i, w_i, t_i)_{i=1}^n \sim p(d, w, t)$;
- d_i, w_i – наблюдаемые, темы t_i – скрытые;
- гипотеза условной независимости: $p(w|d, t) = p(w|t)$.

Задача построения тематических моделей

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

$\phi_{wt} = p(w|t)$ – вероятности терминов w в каждой теме t,
$\theta_{td} = p(t|d)$ – вероятности тем t в каждом документе d.

Интерпретация: стохастическое матричное разложение

$$F = \Phi\Theta$$

- $F = \bigcup_{m \in M} F^m, \quad \Phi = \bigcup_{m \in M} \Phi^m, \quad \Theta = (\theta_{td})_{T \times D}$;
- $F^m = \{p(w|d)\}_{W^m \times D}, m \in M$ — матрицы наблюдаемых вероятностей для каждой модальности;
- $\Phi^m = \{\phi_{wt}\}_{W^m \times T}, m \in M, \ \phi_{wt} = p(w|t)$ — матрицы терминов тем.

ARTM: аддитивная регуляризация тематических моделей

Максимизация log правдоподобия с регуляризаторами $R_i(\Phi, \Theta)$:

$$\sum_{m \in M} \eta_m \sum_{d \in D} \sum_{w \in W^m} n_{dw} \ln \sum_t \phi_{wt} \theta_{td} + \sum_i \tau_i R_i(\Phi, \Theta) \to \max_{\Phi, \Theta}$$

EM-алгоритм: метод простой итерации для системы уравнений

$$\text{E-шаг:} \quad p_{tdw} \equiv p(t|d,w) = \underset{t \in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right)$$

$$\text{M-шаг:} \quad \begin{cases} \phi_{wt} = \underset{w \in W^m}{\text{norm}}\left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right), n_{wt} = \sum_{d \in D} \lambda_{m(w)} n_{dw} p_{tdw} \\ \theta_{td} = \underset{t \in T}{\text{norm}}\left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right), \quad n_{td} = \sum_{w \in d} \lambda_{m(w)} n_{dw} p_{tdw} \end{cases}$$

где $\underset{t \in T}{\text{norm}}(x_t) = \frac{\max\{x_t, 0\}}{\sum_{s \in T} \max\{x_s, 0\}}$ — операция нормировки вектора.

Иерархическая модель hARTM

Пусть построено $l \geqslant 1$ уровней иерархии.

Φ_l, T_l – параметры модели l;
Φ_{l+1}, T_{l+1} – параметры модели $l+1$.

Связь между уровнями:

$$\Phi_l \approx \Phi_{l+1}\Psi,$$

где $\Psi = \{p(t_{l+1}|t_l)\}_{T_{l+1} \times T_l}$ – **матрица перехода**.

Межуровневый регуляризатор:

$$R(\Phi, \Psi) = \sum_{t_l \in T_l} \sum_{w \in W} n_{wt_l} \ln \sum_{t_{l+1} \in T_{l+1}} \phi_{wt_{l+1}} \psi_{t_{l+1}t_l} \to \max_{\Phi, \Psi}.$$

Эквивалентно добавлению в коллекцию $|T_l|$ псевдодокументов, представленных матрицей $\{n_{wt_l}\}_{W \times T_l}$.

Иерархический регуляризатор разреживания

Гипотеза разреженности:

$\tilde{\psi}_{t_{l+1}} = \{p(t_l|t_{l+1})\}_{t_l \in T_l}$ далеки от распределения $\gamma = \{\frac{1}{|T_l|}\}_{t_l \in T_l}$.

$$\sum_{t_l \in T_{l+1}} KL(\gamma, \tilde{\psi}_{t_{l+1}}) \to \max_{\Psi}.$$

что эквивалентно:

$$R(\Psi) = \sum_{t_l, t_{l+1}} \frac{1}{|T_l|} \ln p(t_l|t_{l+1}) = \frac{1}{|T_l|} \sum_{t_l, t_{l+1}} \frac{\psi_{t_{l+1}t_l} p(t_l)}{\sum_{t_{l'}} \psi_{t_{l+1}t_{l'}} p(t_{l'})} \to \min_{\Psi},$$

где $p(t_l)$ вычисляется по Θ_l.

Также позволяет исключить случаи:

$$p(t_{l+1}|t_l) \to \frac{1}{|T_l|}, \quad \forall t_l \in T_l.$$

Постановка задачи

Пусть G — словарь модальности тегов,
Φ^g — подматрица Φ, соответствующая модальности тегов G,
W^1 — словарь модальности слов.

Предлагается на каждом уровне вводить **фоновые темы**:
S_l и B_l – множество предметных и фоновых тем уровня l,
$T_l = S_l \bigcup B_l$.

Предположение: среди G находятся названия тем разных уровней иерархии $S_l \subset G \ \forall l$.

Постановка задачи

Задачи:

1. Число тем $|S_1|$ и их названия заданы экспертами. Построить тематическую модель, в которой $G \in \{W^m | m \in M\}$.
2. Построить второй уровень иерархической системы, в которой $G \in \{W^m | m \in M\}$ и есть межуровневый регуляризатор.
3. Необходимо каждой теме $t \in S$ поставить в соответствие название из множества тегов G, наилучшим образом описывающее данную тему t.

Задача построения первого уровня иерархии

Классификация документов $d \in D$ по заданным темам $S_1 \in G$.
Пусть $G_1 = S_1$.

1. Обычно $|G| \ll |W^1|$, поэтому предлагается модальность тегов учитывать с большим весом ($\frac{\eta_g}{\eta_1} \approx 10^k$, где $k = \frac{|W^1|}{|G|}$);
2. Инициализация Φ^g:

$$\phi_{gt} = \begin{cases} z, z > 0, & \text{если } g = t, g \in G_1 \\ 0, & \text{если } g \neq t, g \in G_1 \\ \text{rand}(0,1), & \text{если } g \in G \setminus G_1 \end{cases}$$

Выполняется перенормировка: $\sum_{g \in G} \phi_{gt} = 1 \ \forall t \in S_1$.

Задача построения второго уровня иерархии

Идея: темы B_1 не являются родительскими для тем из S_2.

- темы B_1 полностью переносятся на второй уровень — $B_{21} \subset B_2$.
- оставшиеся темы из B_2 являются фоновыми темами плоской модели, описывающей второй уровень иерархии.

Предлагается ввести регуляризатор разреживания матрицы Ψ:

$$\begin{cases} \psi_{t_2 t_1} = 0, & \text{если } t_1 \neq t_2 \left[\begin{array}{l} t_2 \in B_2,\ t_1 \in T_1 \\ t_2 \in T_2,\ t_1 \in B_1 \end{array}\right. \\ \psi_{t_2 t_1} = 1, & \text{если } t_1 = t_2,\ t_2 \in B_2,\ t_1 \in B_1 \end{cases}$$

Предлагается столбцы Φ_2, соответствующие B_{21}, проинициализировать столбцами Φ_1, которые соответствуют B_1.

Признаки

- Именование тем S предлагается производить на основе модальности имён-кандидатов G.
- В качестве названий тем l-го уровня предлагается рассматривать только те теги $g \in G$, которые не были выбраны для именования тем $1, ..., l-1$ уровней иерархии.
- Предлагается формировать универсальный набор признаков $R_i(t, g) \in [0, 1]$ по матрицам Φ^g и Θ, которые ранжируют теги G для каждой темы $t \in S$.

Признаки

Признак 1 формируется по матрице Φ^g, оценивает насколько часто тег $g \in G$ встречается в теме $t \in S$.

$$R_1(t,g) = \left(\frac{p(g|t)}{p(g^*|t)}\right)^{\gamma_1} = \left(\frac{\phi_{gt}}{\phi_{g^*t}}\right)^{\gamma_1},$$

где $g^* = \arg\max\limits_{g \in G} p(g|t)$.

Признак 2 формируется по матрице Θ, оценивает долю документов темы $t \in S$, имеющих тег $g \in G$.

$$R_2(t,g) = \left(\frac{\sum\limits_{d \in D} \theta_{td}\,[g \in G_d]}{\sum\limits_{d \in D} \theta_{td}\,[g^* \in G_d]}\right)^{\gamma_2},$$

где $g^* = \arg\max\limits_{g \in G} \frac{\sum\limits_{d \in D} \theta_{td}[g \in G_d]}{\sum\limits_{d \in D} \theta_{td}}$.

Признаки

Признак 3 оценивает вероятность выделения темы $t \in S$ для тега $g \in G$,

$$R_3(t,g) = \left(\frac{p(t|g)}{p(t^*|g)}\right)^{\gamma_3},$$

$$p(t|g) = \frac{\phi_{gt}}{n_g} \sum_{d \in D} n_d \theta_{td},$$

где n_g — число вхождений тега $g \in G$ в коллекцию D, n_d — длина документа $d \in D$ в тегах.

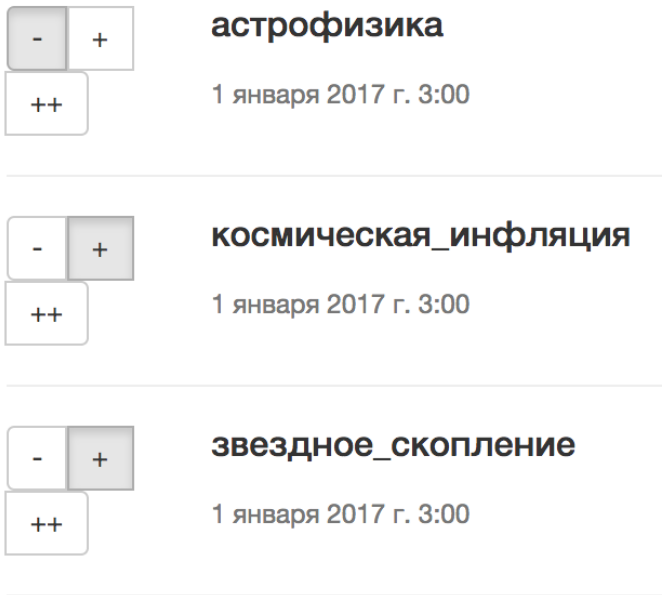
$$R_3(t,g) = \left(\frac{\phi_{gt} \sum\limits_{d \in D} n_d \theta_{td}}{\phi_{gt^*} \sum\limits_{d \in D} n_d \theta_{t^* d}}\right)^{\gamma_3},$$

где $t^* = \arg\max\limits_{t \in S} \dfrac{\phi_{gt}}{n_g} \sum\limits_{d \in D} n_d \theta_{td}$.

Оценка качества именования тем

Тема интерпретируема, если человек понимает о чем эта тема и может дать ей краткое именование.

Слова: звезда, вселенная, масса, телескоп, наблюдение, расстояние, излучение, гравитация, астроном, наблюдать, космология, миллиард, образоваться, размер, небо, скопление, температура, примерно, орех, плотность, маленький, обнаружить, гравитационный, спектр, видеть, измерить, плазма, измерение, обсерватория, масштаб, гелий, астрономический, сверхновый, светимость, излучать, неоднородность, пыль, хокинг, горячее, амплитуда, галактический, кеплер, параметр, гигантский, облако, стадия, плотный, сжатие, далёкий, распределение,



Задача асессора: отметить какие теги из G_t подходят в качестве названия темы t (поставить $+$).

Оценка качества именования тем

A – множество асессоров, $I(g,a) = [\text{асессор } a \text{ поставил } + \text{ для } g]$, $g_t^* = \arg\max\limits_{g \in G_t} R(g,t)$ — имя, выбранное моделью для темы t.

Средняя доля асессоров, согласных с именем, выбранным моделью:

$$MK = \frac{1}{|S|}\sum_{t\in S}\frac{1}{|A|}\sum_{a\in A} I(g_t^*, a).$$

Согласованность асессоров:

$$C = \frac{1}{|S|}\sum_{t\in S}\frac{1}{|A|}\sum_{a\in A}\frac{\sum\limits_{g\in G_t} I(g,a)\frac{1}{|A|-1}\sum\limits_{a'\in A\setminus a} I(g,a')}{\sum\limits_{g\in G_t} I(g,a)}.$$

Описание данных

- Использовалась коллекция статей научно-популярного интернет-журнала ПостНаука.
- В коллекции $|D| = 3404$ и имеются модальности слов ($|W^1| = 19186$), авторов ($|W^a| = 859$), биграмм ($|W^2| = 11442$), триграмм ($|W^3| = 464$) и тегов ($|G| = 930$).
- Было выбрано $|S_1| = 20$, $S_1 \in G$, $|B_1| = 1$, $|B_2| = 2$.
- Для модальностей авторов и тегов были введены фиктивные автор и тег соответственно. B_1 и B_2 для данных модальностей содержит только фиктивные токены.
- Модель строилась алгоритмом hARTM, реализованном в библиотеке BIGARTM.

Метрики качеств

Перплексия:

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

Интерпретация перплексии:

- если распределение $p(w|d) = \frac{1}{|W|}$ равномерное, то $\mathcal{P} = |W|$
- мера различности или неопредленности слов в тексте

Разреженность модели измеряется долей $\mathcal{R}_{\Phi^m}$ и $\mathcal{R}_{\Theta}$ нулевых элементов в частях матриц Φ^m $m \in M$ и Θ, соответствующим предметным темам S.

Ошибки первого и второго рода

False Positive Rate (FPR) — доля пар (d, t): тема t присутствует в d, но соответствующий ей тег $g = t$ не приписан документу d.

$$FPR = \frac{\sum_d \sum_{t \in S \setminus G_d} \left[\frac{\theta_{td}}{1-b(d)} \geqslant k \right]}{\sum_d |S \setminus G_d|}.$$

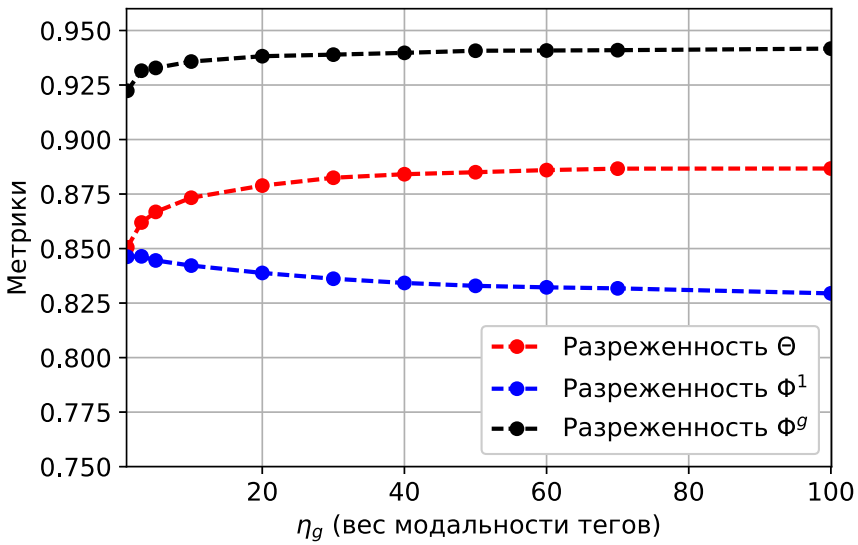
False Negative Rate (FNR) — доля пар (d, t): тег g приписан документу d, а соответствующая ей тема $t = g$ в d отсутствует.

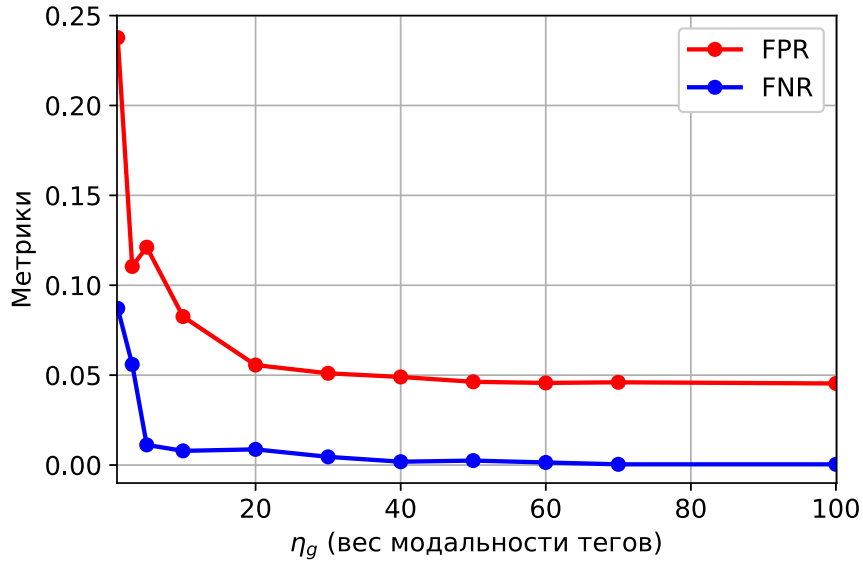
$$FNR = \frac{\sum_d \sum_{t \in G_d \cap S} \left[\frac{\theta_{td}}{1-b(d)} < k \right]}{\sum_d |G_d \cap S|}$$

Для формализации присутствия и отсутствия темы вводится порог k.

Выбор веса модальности тегов

Графики зависимости значений метрик качеств от веса модальности тегов G для первого уровня иерархической модели, в которую включены модальности слов и тегов.





Вывод: Для баланса в качестве η_g было выбрано значение 60.

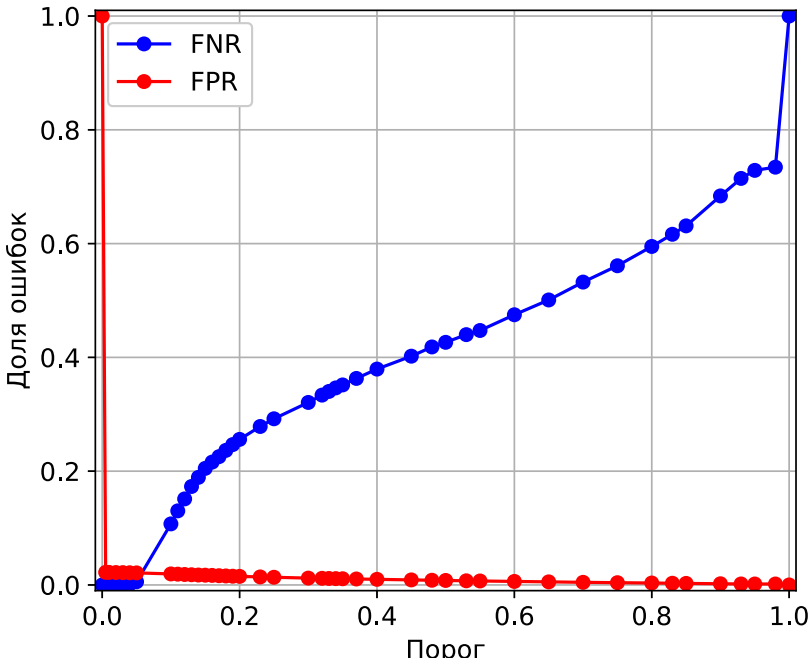
Изменение метрик при добавлении регуляризаторов

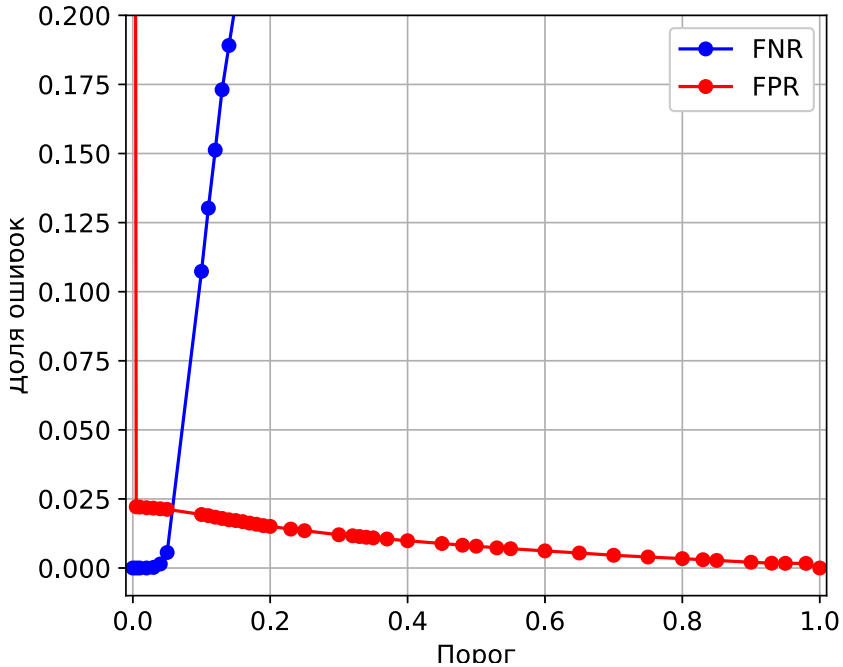
Регуляризотор	**Вес**	$\mathcal{R}_\Phi$	$\mathcal{R}_\Theta$	$\mathcal{P}$	FPR	FNR
Модальность тегов	60	0.832	0.886	4074	0.046	0.002
Сглаживание $t \in B$ в Θ	500	0.841	0.893	4347	0.046	0.000
Модальность авторов	100	0.837	0.896	4243	0.043	0.001
Модальность биграмм	1	0.836	0.898	3001	0.042	0.001
Модальность триграмм	1	0.836	0.898	2925	0.043	0.000
Разреживание $t \in S$ в Θ	-30	0.830	0.927	2970	0.020	0.000
Декоррелирование слов	0.2	0.921	0.925	3632	0.020	0.000
Декоррелирование триграмм	0.1	0.921	0.925	3632	0.020	0.000

Таблица: Значения метрик для различных моделей (модели расположены в порядке добавления регуляризоторов).

Ошибки первого и второго рода

Графики зависимости ошибок первого и второго рода (FPR и FNR) от порога для конечной модели первого уровня

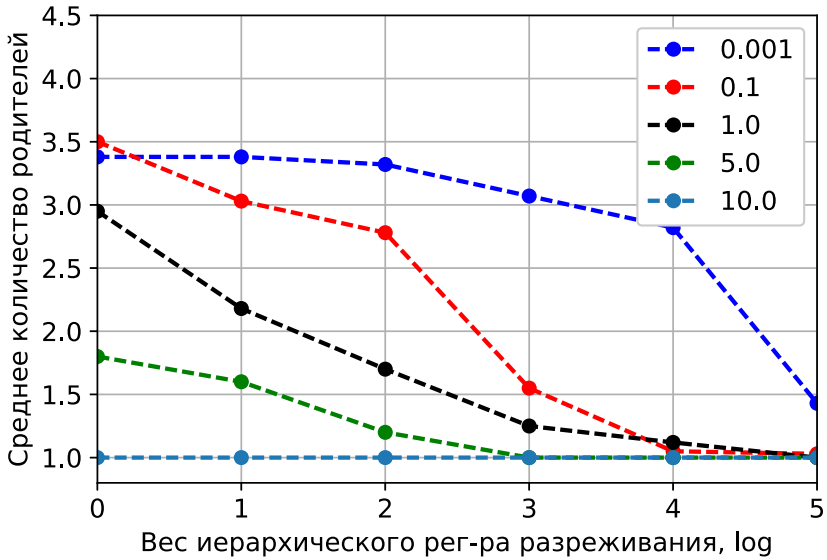




Вывод: при k от 0 до 0.05 модель практически не допускает ошибок второго рода FNR, приэтом и значения ошибок первого рода FPR невелико≈ 0.025.

Иерархические регуляризаторы

График зависимости среднего количества родителей для тем второго уровня от веса иерархического регуляризатора разреживания при различных значениях веса межуровнего регуляризатора.



Вывод:при значении веса межуровнего регуляризатора, равном 10, при любом весе регуляризатора разреживания многодольный граф вырождается в дерево.

Именование тем

Согласованность 3-х асессоров $C = 0.54$

Признак	**MK**
R_1	0.6
R_2	0.62
R_3	0.32
$R_1 R_2$	0.62

Средняя доля асессоров, согласных с именем, выбранным моделью.

Заключение

Предложена методика построения двухуровневой тематической иерархии с автоматическим именованием тем:

- Предложенный метод позволяет классифицировать документы первого уровня иерархии с хорошей точностью (на коллекции Постнаука FPR = 0.02, FNR = 0.00);
- Предложены рекомендации для автоматического построения хорошо интерпретируемой двухуровневой тематической иерархии;
- Предложенный алгоритм именования тем выдет названия, которые хорошо согласованы с именами, выбранными асессорами.