ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ «ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»
ФАКУЛЬТЕТ КОМПЬЮТЕРНЫХ НАУК

Ходырева Виктория Константиновна

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

АВТОМАТИЧЕСКОЕ ИМЕНОВАНИЕ ТЕМ В ВЕРОЯТНОСТНОМ ТЕМАТИЧЕСКОМ МОДЕЛИРОВАНИИ

AUTOMATIC LABELING OF TOPIC MODELS

по направлению подготовки 01.04.02 Прикладная математика и информатика
образовательная программа « Науки о данных »

Студент:

_____

В.К. Ходырева

Научный руководитель:

д.ф.-м.н., доцент, профессор РАН

_____

К.В. Воронцов

Москва, 2019

Содержание

1 Введение.

Тематическое моделирование – это технология статистического анализа текстов для автоматического выявления тематики в больших коллекциях документов. Данное направление активно развивается с конца 90-х годов и имеет множество применений: выявление трендов в новостных потоках, патентных базах, архивах научных публикаций, информационный поиск, классификация и кластеризация документов, тегирование веб-страниц, рубрикация коллекций изображений, музыки, видео и многое другое. (1)

Наибольшее применение в тематическом моделировании находят вероятностные модели. Они осуществляют «мягкую» кластеризацию, описывая каждую тему дискретным распределением на множестве терминов, а каждый документ - дискретным распределением на множестве тем. Предполагается, что коллекция документов - выборка троек (документ, термин, тема), порожденная смесью таких распределений, причём темы здесь являются скрытыми переменными. Построить вероятностную тематическую модель — значит решить задачу восстановления исходных распределений по известной коллекции.

Для интерпретации тем и представления их в информационно-поисковых и рекомендательных сервисах необходимо решать задачу автоматического именования тем. В результате построения тематической модели, каждая тема представляется набором терминов и их вероятностями принадлежать данной теме. Однако, этого недостаточно для полного понимания смысла темы, а также того, в чём заключается её отличие от остальных тем. Один из лучших вариантов представления смысла темы - её именование, которое заключается в выборе ограниченного набора наиболее полно описывающего её слов или фраз.

Потенциальных приложений автоматического именования тем - разнообразное множество. В частности, необходимость именования возникает на этапе интерпретации результатов моделирования,

когда необходимо выделить значимые темы и отбросить шумовые. Полученные осмысленные названия тем также можно использовать для визуализации коллекции документов на плоскости. (2) Наконец, именование тем подходит для выявления трендов в динамических тематических моделях.

До недавнего времени существовало два основных способа именования тем. Первый заключался в использовании топ-токенов, полученных из тематической модели или отобранных по частоте. Однако, с помощью отдельных слов, лишенных контекста, сложно понять, о чём идёт речь в теме. Это наглядно продемонстрировал Mei et al. (2007) (3). Второй подход заключается в ручной разметке тем, и он имеет массу недостатков. Во-первых, человеческая оценка субъективна и может быть неточной или непонятной широкому кругу людей. Во-вторых, ручная разметка - затратный процесс, требующий времени и усилий специалистов со знанием предметной области. Иногда тематическую модель нужно перестраивать несколько раз за день, например, если речь идёт о динамически меняющихся данных. В таком случае скорость разметки может быть существенным ограничением.

В данной работе будут рассмотрены актуальные подходы к именованию тем в вероятностном тематическом моделировании, в частности, с использованием сторонних баз знаний. Также будут сформулированы основные требования к кандидатам в названия тем, предложен алгоритм именования верхнеуровневых тем иерархической тематической модели. Идея алгоритма заключается в именовании тем новой тематической модели на основе переноса имён из существующей размеченной иерархии. Для получения такой иерархии предлагается брать за основу категории и относящиеся к ним документы Википедии и строить на них иерерхическую тематическую модель. Были проведены эксперименты на русскоязычном корпусе новостей.

2 Обзор литературы.

В актуальных исследованиях на тему автоматического именования тем выделяется два направления: первый опирается на извлечение фраз-кандидатов непосредственно из корпуса с последующим их ранжированием, второй основан на использовании именованных иерархий и онтологий. Первый подход лучше применим для именования тем на нижних уровнях иерархических тематических модели, где описываются более конкретные события. Второй подход, наоборот, позволяет получать более абстрактные понятия в качестве названий, что больше подходит при именовании тем на верхних уровнях.

Методы извлечения названий тем из корпуса:

Первым предложил автоматический подход к именованию тем Mei et al. (2007) (3) Он отметил, что принятые на практике методы использования нескольких самых частотных слов эмпирического распределения или ручная генерации именований не позволяет достичь приемлемого качества. Он предложил переформулировать задачу как оптимизационную задачу минимизации дивергенции Кульбака-Лейблера (KL) между кандидатом в название и темой. Согласно данному подходу, сначала формируется список кандидатов из статистически значимых биграмм или именных групп, выделенных из документов коллекции. Далее все кандидаты ранжируются по KL дивергенции для каждой темы, и в качестве названий выбираются кандидаты с максимальным рангом.

В работе 2011 года Lau et al. (6) предпринял попытку обогатить топ-слова темы заголовками Википедии. Для получения нужных заголовков использовался поиск по Википедии по запросам, составленным из топ-слов тем. Авторы также использовали всевозможные подстроки заголовков, являющиеся именными группами и добавляли их в списки кандидатов. Ранжирование происходило в процессе обучения с учителем методом опорных векторов (SVR) с использованием лексических мер схожести: взаимная информация (PMI), t-

тест Стьюдента, критерий Пирсона, коэффициент Дайса, логарифм правдоподобия, также лексических признаков и рейтингов поисковых систем.

Наконец, в более поздних работах (9), (10) использовались методы глубокого обучения. Bhatia применял word2vec и doc2vec с целью представления кандидатов и тем в одном латентном пространстве.

Методы, опирающиеся на внешние данные:

В последние годы возрастает интерес к исследованиям, связанным с использованием внешних баз знаний. В создании таких баз задействованы сообщества людей, а подлинность информации проверяются несколькими экспертами. Позитивные результаты этой интеграции можно наблюдать в таких областях, как поиск информации, классификация, визуализация знаний [8]. Идею использования сторонних баз знаний для именования тем впервые применил Magatti в своей работе [5], где использовалась иерархия Google Directory (gDir). Позже в [8] авторы искали названия среди концептов DBpedia, а в работах [6] и [9] по топ словам темы делались запросы в Википедию, чтобы потом использовать заголовки статей и названия категорий в качестве кандидатов в названия тем.

В 2009 году Magatti представил метод на основе размеченных вручную иерархий, полученных с использованием сервиса Google Directory. (5) Для выбора наиболее подходящей из размеченных тем использовался ряд мер схожести тем, после чего использовались правила, по которым из дерева тем выбирались лучшие названия. Применялся алгоритм ALOT, который выбирал и переиспользовал названия самых близких тем.

В 2012 году Mao (7) предложил использовать информацию об отношениях между темами в иерархической модели. В частности, рассматривались отношения брат-брат и отец-сын. В статье делается ряд предположений: 1) слова, часто встречающиеся в потомках, более вероятно окажутся подходящими для именования вершины; 2) чем ближе тема находится к корню, тем более общими должны быть названия; 3) слова, часто встречающиеся только в одном из

братьев, должны доминировать над теми, которые встречаются в многих братьях одновременно.

В 2013 году Hulpus (8) был предложен ещё один метод, использующий на этот раз структурированный источник данных DBpedia. Метод основан на построении семантического графа понятий, извлекаемых из DBpedia с помощью топ-слов темы. Применяя графовые алгоритмы измерения центральности, узлы графа ранжируются, и узел с максимальным рангом может претендовать на название темы. Преимущество данного алгоритма в том, что он не требует предварительной обработки и может работать on-line.

Стоит отметить,что во всех перечисленных работах в качестве вероятностной тематической модели были использованы модели PLSA или LDA. В данной работе будет рассмотрена модель ARTM.

3 Вероятностное тематическое моделирование.

3.1 Задача тематического моделирования

Введем ряд необходимых понятий. Пусть имеется коллекция текстовых документов D и словарь уникальных слов коллекции W. Каждый документ d из D представлен последовательностью слов $\{w_1, w_2, \ldots, w_{n_d}\}$, где n_d - общее число слов в d. Слова в одном документе могут повторяться. Обозначим число вхождений слова w в документ d за n_{dw}.

Для упрощения модели принимается гипотеза "мешка слов согласно которой порядок слов в каждом документе считается неважным, и документ представляется набором уникальных слов и их частот.

Предположим, что существует конечное число тем T, и каждое слово w в документе d связано с одной из тем $t \in T$. Так как сами темы непосредственно не наблюдаются, они являются скрытыми переменными.

Таким образом, имеется набор троек (d, w, t) и предполагается,

что существует некоторое распределение $p(d, w, t)$ на множестве $D \times W \times T$, которое требуется восстановить.

Также вводится гипотеза условной независимости. Согласно данной гипотезе, появление слов, относящихся к теме t, не зависит от выбора документа и описывается общим распределением $p(w|t)$:

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

Используя формулу полной вероятности, и гипотезу условной независимости, получаем:

$$p(w|d) = \sum_{t \in T} p(w|t, d)p(t|d) = \sum_{t \in T} p(w|t)p(t|d)$$

Таким образом, получается, что для написания очередного слова в документе сначала выбирается тема из распределения $p(t|d)$, а затем само слово из распределения $p(w|t)$, зависящего только от темы.

Задача построения тематической модели заключается в восстановлении неизвестных условных распределений $p(w|t)$ для каждой темы, а также определении оптимального числа тем.

Задачу можно сформулировать и в матричном виде как восстановление стохастических матриц:

$$\Phi = (\phi_{wt})_{W \times T} = (p(w|t))_{W \times T} - \text{матрица слова-темы}$$

$$\Theta = (\theta_{td})_{T \times D} = (p(t|d))_{T \times D} - \text{матрица темы-документы}$$

$$F = (p(w|d))_{W \times D}$$

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

$$F = \Phi\Theta$$

Для оценки Φ и Θ максимизируют логарифм правдоподобия выборки: найденное распределение $p(w|d)$ приближают к реальному, заданному частотами $\hat{p}(w|d) = \frac{n_{dw}}{n_d}$:

$$L(\Phi, \Theta) = \log \prod_{d \in D, w \in d} p(w, d)^{n_{dw}} = \log \prod_{d \in D, w \in d} p(w|d)^{n_{dw}} \underbrace{p(d)^{n_{dw}}}_{\text{const}} =$$

$$= \sum_{d \in D} \sum_{d \in D} n_{dw} \log \sum_{t \in T} \phi_{wt} \theta_{td} \longrightarrow \max_{\Phi, \Theta} (1)$$

при ограничениях неотрицательности и нормировки (необходимо для получения стохастических матриц):

$$\sum_{w \in W} \phi_{wt} = 1, \ \phi_{wt} \geq 0$$
$$\sum_{t \in T} \theta_{td} = 1, \ \theta_{td} \geq 0$$

Описанный метод - метод вероятностного латентного семантического анализа (PLSA) (Hofmann T. Probabilistic latent semantic analysis // UAI. — 1999.)

3.2 EM-алгоритм

Задача (1) решается применением итерационного EM-алгоритма, в процессе которого последовательно выполняются две операции:

1. E-шаг: оценка скрытых переменных по приближению Φ, Θ на текущем шаге.

$$p(t|d,w) = \frac{p(w,t|d)}{p(w|d)} = \frac{p(w,t)p(t|d)}{p(w|d)} = \frac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{ts}}$$

2. M-шаг: Пересчет Φ, Θ по фиксированным скрытым переменным.

$$n_{tdw} = n_{dw} p(t|d,w)$$

$$n_{wt} = \sum_{d \in D} n_{tdw}; \ n_t = \sum_{w \in W} n_{wt}; \ \phi_{wt} = \frac{n_{wt}}{n_t}$$

$$n_{td} = \sum_{w \in d} n_{tdw}; \ n_d = \sum_{t \in T} n_{td}; \ \theta_{td} = \frac{n_{td}}{n_d}$$

Процесс продолжается вплоть до сходимости.

Для больших коллекций существует более эффективная модификация - online алгоритм, в процессе которого происходит несколько обновлений матрицы Φ за один проход по коллекции.

3.3 Иерархическая тематическая модель

До сих пор рассматривалась только плоская тематическая модель. Ещё один вид тематических моделей с более сложной структурой - иерархические тематические модели. Они представляют собой древовидные графы, моделирующие отношения гипонимии и гиперонимии между темами.

Существует несколько подходов к построению иерархии: восходящий или нисходящий, также модель может строится для каждой отдельной темы родительского уровня или целиком на уровень. В данной работе будет рассматриваться иерархия, построенная сверху вниз, где каждый уровень - плоская модель.

Предположим, что родительская матрица Φ^p уже построена. S - множество дочерних тем, T - множество родительских тем. Подразумевается, что $|S| > |T|$. Приблизим Φ^p матричным разложением $\Phi\Psi$:

$$p(w|t) = \sum_{s \in S} p(w|s)p(s|t) = \sum_{s \in S} \phi_{ws}\psi_{st}$$

В качестве меры близости распределений используем дивергенцию Кульбака-Лейблера, тогда данное выражение можно записать в виде регуляризатора:

$$R(\Phi, \Psi) = \tau \sum_{w \in W} \sum_{t \in T} n_{wt} \log \sum_{s \in S} \phi_{ws}\psi_{st}$$

Полученное выражение похоже на формулу правдоподобия модели (1), поэтому вместо добавления регуляризатора можно добавить во входную матрицу n_{dw} |T| псевдодокументов, отвечающих столбцам родительской матрицы Φ^p, так, что $n_{d'w} = \lambda\phi_{wt},\ d' = t$. В процессе EM-алгоритма нужные распределения для дочерних тем Ψ будут находиться в соответствующих столбцах матрицы Θ.

3.4 Аддитивная регуляризация

Задача построения вероятностной тематической модели является некорректно поставленной по Адамару, так как искомое матричное

разложение $\Phi\Theta$ может быть найдено не единственным способом. В таком случае EM-алгоритм будет неустойчивым. Решением проблемы может быть введение дополнительных ограничений в функционал качества. Введем понятие регуляризаторов $R(\Phi, \Theta)$. Максимизируем комбинацию логарифма правдоподобия и регуляризаторов.

$$L(\Phi, \Theta) = \sum_{d \in D} \sum_{d \in D} n_{dw} \log \sum_{t \in T} \phi_{wt} \theta_{td} + R(\Phi, \Theta) \longrightarrow \max_{\Phi, \Theta}$$

$$R(\Phi, \Theta) = \sum_{i=1}^{k} \tau_i R_i(\Phi, \Theta)$$

при ограничениях неотрицательности и нормировки:

$\sum_{w \in W} \phi_{wt} \in \{0, 1\},\ \phi_{wt} \geq 0$

$\sum_{t \in T} \theta_{td} \in \{0, 1\},\ \theta_{td} \geq 0$

τ_i - неотрицательные коэффициенты регулиризации, являющиеся гиперпараметрами модели.

3.5 Типы регуляризаторов

Регуляризатор сглаживания/разреживания

Слова в темах принято разделять на предметные и фоновые. Согласно гипотезе разреженности, предметные слова составляют лексическое ядро темы, в которое входит относительно небольшая доля словаря. Как следствие, предметные темы (соответствующие столбцы матрицы Φ) должны иметь разреженное распределение.

Фоновые слова - это общеупотребимая лексика, стоп-слова, оставшиеся после этапа предобработки. Они имеют значимые вероятности встретиться во многих темах, но не несут полезной информации, зашумляют темы. Для того, чтобы справиться с данной проблемой, вводятся фиктивные фоновые темы, цель которых - собрать в себя все неинформативные слова. Распределение по словам в таких темах должно быть сглаженным.

Кроме того, каждый документ скорее всего принадлежит к небольшому количеству тем, поэтому и столбцы матрицы Θ должны быть

разрежены по темам.

Объединяя вышеупомянутые ограничения на столбцы Φ и Θ, получаем регуляризатор сглаживания/разреживания:

$$R(\Phi, \Theta) = \sum_{t \in T} \sum_{w \in W} \beta_{wt} \log(\phi_{wt}) + \sum_{d \in D} \sum_{t \in T} \alpha_{td} \log(\theta_{td})$$

β_{wt}, α_{td} - коэффициенты регуляризации. При отрицательных коэффициентах происходит сглаживание, а при положительных - разреживание.

Регуляризатор декоррелирования

Для того, чтобы темы отличались друг от друга, соответствующие вектора распределений должны стремиться к ортогональным. Для этого вводится регуляризатор декоррелирования, в котором минимизируются попарные скалярные произведения столбцов матрицы Φ.

$$R(\Phi, \Theta) = -\frac{\tau}{2} \sum_{t' \in T} \sum_{t \in T \setminus \{t'\}} < \phi_t, \phi_{t'} >$$

Сглаживающий/разреживающий иерархический регуляризатор

Для того, чтобы каждой родительской теме в иерархической тематической модели соответствовало небольшое количество дочерних тем, матрица Ψ разреживается.

$$R(\Psi) = \sum_{s \in S} \sum_{t \in T} \gamma_{st} \log \psi_{st}$$

4 Именование тем.

4.1 Требования к названиям тем

В статье Mei et al (3) были перечислены основные свойства, которыми должно обладать название темы:

1. Название должно передавать содержание темы;

2. Название должно максимально полно покрывать содержание темы;

3. Название должно быть понятно человеку, быть семантически и грамматически корректным;

4. Название позволяет отличать тему от всех остальных тем.

Так как темы представлены распределениями над словами, то минимальной смысловой единицей является слово. Помимо отдельных слов можно использовать частотные колокации, n-граммы, целые предложения, будем называть их фразами. Однако, для полного покрытия темы может оказаться, что одной фразы недостаточно. Например, если тема разбивается на смысловые блоки. Тогда может потребоваться несколько не пересекающихся по смыслу фраз, каждая их которых будет отвечать за отдельный смысловой блок. В данной работе для упрощения будем считать, что каждой теме соответствует название, состоящее из одной фразы.

4.2 Википедия как источник кандидатов в названия

Википедия является мультидисциплинарной коллекцией документов с широким спектром тем и может быть полезна при именовании тем. Однако, использование Википедии в большей степени обосновано для крупных тем на верхних уровнях иерархии в иерархических тематических моделях.

Как правило, для верхних уровней лучше подходят абстрактные понятия: 'Биология', 'Политика', 'Культура', 'Страны мира', 'История кино'. Нижние уровни, наоборот, уточняют и конкретизируют информацию верхних уровней, поэтому чем ниже тема в иерархии, тем длиннее и конкретнее её название: 'Песни и роли народного артиста России Олега Анофриева', 'Мобильные бригады детских врачей выезжают в Тверскую область', 'Бен Арфа объявил об уходе из ПСЖ'. Кроме того, на нижних уровнях названия тем чаще будут содержаться в заголовках и текстах документов. И наоборот,

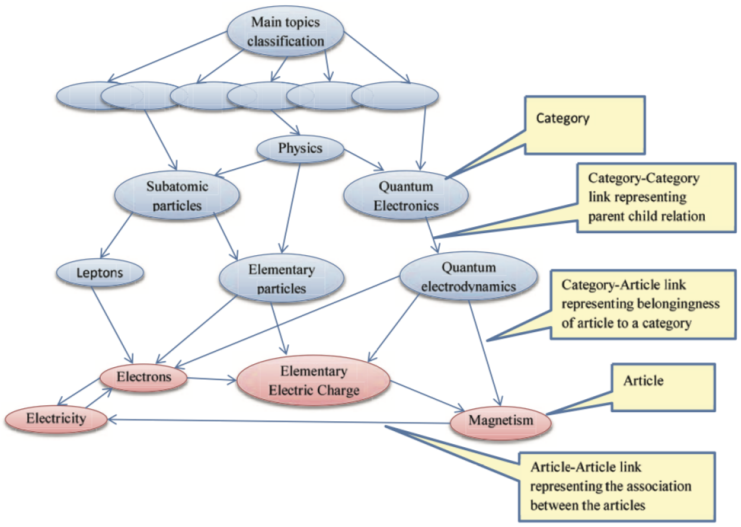


Рис. 1: Структура Википедии

для верхних уровней чаще возникает ситуация, когда подходящее название сложно выделить из контекста. Например, в текстах про культуру может не упоминаться или редко упоминаться само слово 'культура'. В таком случае есть смысл обратиться к Википедии.

В Википедии является источником разнообразной мета информации. Из публичных дампов Википедии[1] . можно извлечь тексты и заголовки статей, названия ссылок с других страниц Википедии. Более того, можно получить информацию о связях между категориями и ссылок между статьями. Существует несколько методов доступа к получению данных из Википедии, которые не требуют скачивания всей базы данных. Например, MediaWiki API[2] - это веб-сервис, который предоставляет доступ к некоторым вики-функциям, таким как аутентификация, операции с страницами и поиск. Вся эта мета информация может пригодиться при именовании тем.

Структура Википедии изображена на рисунке 1. Основые элементы этой структуры - статьи, категории и связи между ними.

Категории - это специальные страницы, которые используются для группировки статей. Например, категория «Физика» использу-

[1]https://dumps.wikimedia.org

[2]https://www.mediawiki.org

ется для группировки статей, связанных с физикой. Каждая категория может иметь несколько родительских категорий и несколько дочерних категорий. Статьям также может быть присвоено несколько категорий, связанных с ней. (11) Таким образом, структура Википедии представляет собой ориентированный циклический граф.

5 Метод именования тем с помощью Википедии.

В данном разделе будет описан метод именования верхних уровней иерархической тематической модели, построенной на заданном корпусе документов с помощью иерархической тематической модели, построенной на Википедии. Метод мотивирован изначальной иерархической структурой Википедии, где в качестве псевдо-тем выступают категории. Идея заключается в попытке построить на Википедии такую иерархическую тематическую модель, которая будет приближена к именованному дереву категорий, но лишена его недостатков: цикличности, неоднозначности. Более того, каждая тема будет иметь представление в виде дискретного распределения на множестве терминов.

Полученные темы Википедии намного проще именовать, так как нужные названия можно заимствовать из ближайших категорий. Далее, используя метод сопоставления тем в двух иерархиях, который будет описан ниже, можно выбрать в качестве кандидата в название темы название соответствующей темы Википедии.

5.1 Постановка задачи

Пусть имеется иерархическая тематическая модель $(\{\Phi^l_{wiki}\}, \{\Theta^l_{wiki}\}, \{\Psi^{l,l+1}_{wiki}\})$ с множеством тем T'_{wiki} и числом уровней L_{wiki}, где $l \in \overline{0, L_{wiki} - 1}$, построенная на документах Википедии, и иерархическая тематическая модель $(\{\Phi^k_c\}, \{\Theta^k_c\}, \{\Psi^{k,k+1}_c\})$ с множеством тем T'_c и числом уровней L_c, где $k \in \overline{0, L_c - 1}$, построенная на некотором пользовательском корпусе.

Предполагается, что среди тем T'_c существует подмножество тем

T_c, а среди тем T'_{wiki} - подмножество T_{wiki} такие, что можно построить соответствие между T_c и T_{wiki} (many-to-one соответствие).

Тогда задача поиска соответсвия между иерархиями заключается в поиске матрицы переходов $A_{T_{wiki} \times T_c}$, $A_{i,j} \in \{0, 1\}$, причем если $A_{i,j} = 1$, будем говорить, что тема t_i пользовательского корпуса соответствует теме t_j из Википедии.

Введём ряд дополнительных матриц:

$P_{T_c \times T_c}$ - матрица связей между темами внутри пользовательской иерархии.

Пусть $level(i), level(j)$ - уровни, на которых в иерархии находятся темы t_i и t_j. Вспомним, что в матрице Ψ содержатся условные вероятности перехода между темами соседних уровней. Последовательно перемножая матрицы между уровнями $level(i), level(j)$ получаем аппроксимацию вероятности принадлежать теме t_j при условии принадлежности теме t_i на более высоком уровне. Тогда матрица P задаётся следующим образом: если $level(i) \geq level(j)$, то $P[i, j] = 0$. Иначе $P[i, j] = (\prod\limits_{level(i) \leq m \leq level(j)} \Psi_c^m)[i, j]$.

Аналогичным образом зададим матрицу $Q_{T_{wiki} \times T_{wiki}}$ - матрица связей между темами внутри иерархии для Википедии.

Наконец, пусть $S_{T_{wiki} \times T_c}$ - матрица попарной близости тем для двух иерархий: $S_{i,j} = sim(t_i, t_j)$, где sim - некоторая функция близости.

При поиске матрицы A будем руководствоваться рядом предположений:

1. Каждой теме из T_c соответствует не более одной темы из T_{wiki}

2. Сумма близостей между соответствующими темами должна быть максимальна: $\sum\limits_{t_i \in T_c, t_j \in T_{wiki}} sim(t_i, t_j) \to max$

3. Связи между темами не должны сильно измениться после преобразования. Если темы t_{i1}, t_{i2} перешли в темы t_{j1}, t_{j2}, то $P_{i1,i2} - Q[j1, j2] \to min$ для всех таких пар. Более того, если две темы

в T_c находились в отношении отец-сын, то после преобразования отношение должно сохраниться.

5.2 Оптимизационная задача

Матрицу A будем искать в процессе решения оптимизационной задачи. Распишем функцию потерь:

$Loss(F) = -\sum_{i,j} A \otimes S + \lambda||P - AQA^T||_2^2 +$
$+\gamma \sum_{i,j} \mathbb{1}(P = 0) \otimes (AQA^T)$,

$A = softmax(F)$: $\mathrm{A}_{i,j} = \frac{e^{F_{i,j}}}{\sum_k e^{(F_{i,k})}}$, $F \in \mathbb{R}^{T_c \times T_{wiki}}$

$\otimes$ - поэлементное умножение (умножение Адамара).

$||\cdot||_2$ - евклидова норма.

$\mathbb{1}$ - индикатор, то есть $\mathbb{1}(P = 0)$ - матрица, имеющая размерность $T_c \times T_c$, в которой на i, j позиции стоит 1, если $P_{i,j} = 0$, 0 иначе.

Матрица AQA^T имеет размерность $T_c \times T_c$, в ней содержатся вероятности переходов между темами Википедии после применения преобразования A.

Остановимся подробнее на каждом слагаемом:

1. $-\sum_{i,j} A \otimes S$ - суммарная близость между сопоставленными темами. Так как решаем задачу минимизации, ставим минус.

2. $||P - AQA^T||_2^2$ - норма матрицы, в которой на i, j позиции стоит $P_{i,j} - Q_{i',j'}$, где i', j' - индексы тем второй иерархии, полученные после сопоставления.

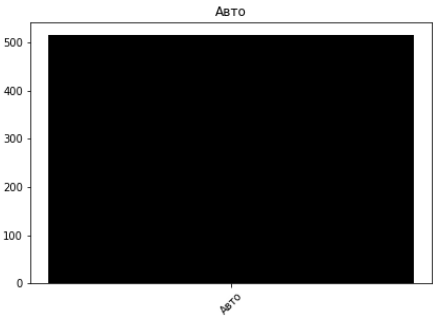
3. $\sum_{i,j} \mathbb{1}(P = 0) \otimes (AQA^T)$ - взвешенная сумма "неправильных"отношений во второй иерархии, неправильное отношение это когда сын становится отцом.

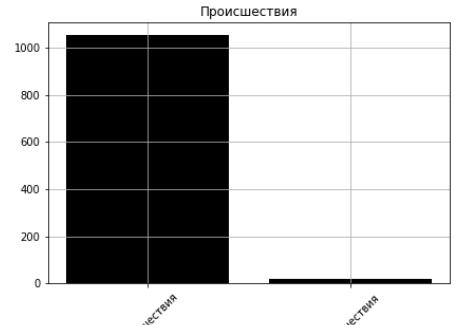
Будем искать матрицу A приближенно методом градиентного спуска. Для начала инициализируем матрицу F случайно. Так как в итоге в матрице A в каждой строке должна стоять только одна единица, будем на каждой итерации брать функцию Softmax от F.

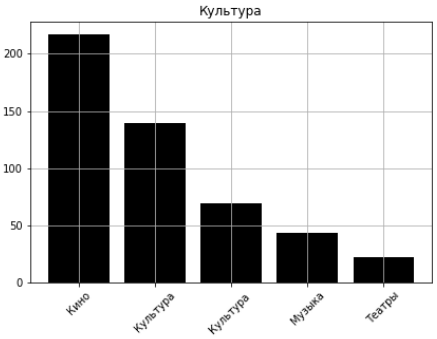
6 Эксперимент.

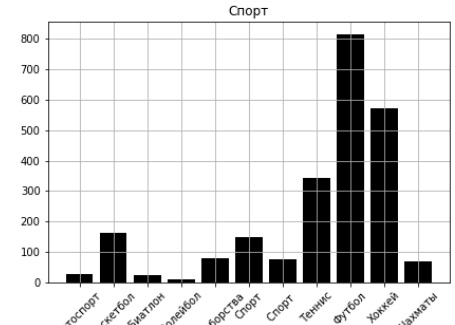
6.1 Описание данных

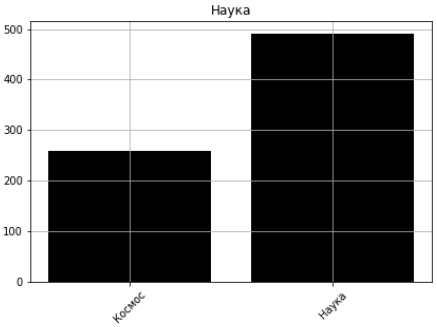
Для эксперимента была выбрана новостная коллекция документов на русском языке. Каждый документ содержит заголовок, основной текст, размеченные асессорами категории. После удаления документов, содержащих небольшую долю латинских символов, а также слишком коротких документов, размер коллекции составил 11127 документов. Дополнительно асессорская разметка категорий была перепроверена и дополнена. Были выделены категории двух уровней, распределение документов по которым можно наблюдать на рисунке (2). Видно, что имеется сильный дисбаланс классов на втором уровне иерархии, могут возникнуть трудности при выделении мелких тем.

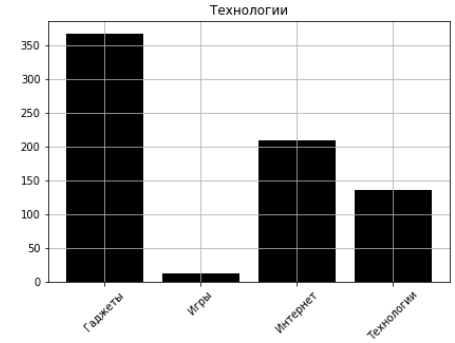

Авто
0
100
200
300
400
500
Авто

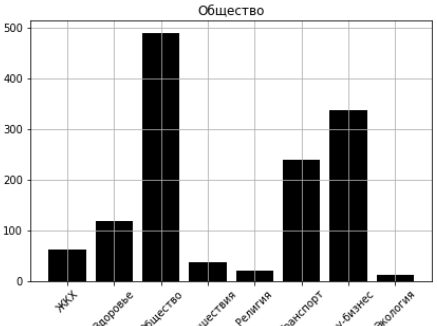

Происшествия
0
200
400
600
800
1000
Происшествия
Происшествия

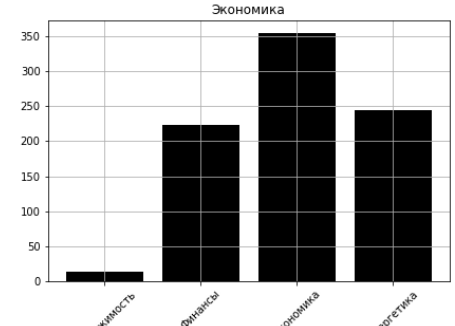

Культура
0
50
100
150
200
Кино
Культура
Культура
Музыка
Театры

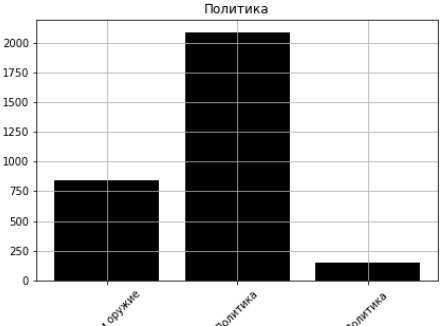

Спорт
0
100
200
300
400
500
600
700
800
Автоспорт
Баскетбол
Биатлон
Волейбол
Единоборства
Спорт
Спорт
Теннис
Футбол
Хоккей
Шахматы

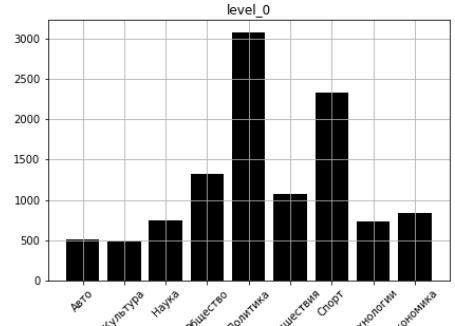

Наука
0
100
200
300
400
500
Космос
Наука


Технологии
0
50
100
150
200
250
300
350
Гаджеты
Игры
Интернет
Технологии


Общество
0
100
200
300
400
500
ЖКХ
Здоровье
Общество
Путешествия
Религия
Транспорт
Шоу-бизнес
Экология


Экономика
0
50
100
150
200
250
300
350
Недвижимость
Финансы
Экономика
Энергетика


Политика
0
250
500
750
1000
1250
1500
1750
2000
Армия и оружие
Политика
Политика


level_0
0
500
1000
1500
2000
2500
3000
Авто
Культура
Наука
Общество
Политика
Происшествия
Спорт
Технологии
Экономика

Статьи Википедии были получены из актуального дампа. С помощью MediaWiki было построено дерево категорий следующим образом: было выбрано 12 категорий верхнего уровня ('Информация', 'Общество', 'Природа', 'Техника', 'Человек', 'Транспорт', 'Спорт', ' Политика', 'Культура', 'Происшествия', 'Экономика', 'Наука'), далее обходом в ширину извлекались подкатегории и относящиеся к ним документы. Категории, которые ранее уже были достигнуты, не рассматривались повторно, также обход останавливался, когда глубина дерева достигала 6. В итоге для каждой статьи был получен список категорий - совокупность путей, по которым эта статья была найдена. Общее число документов Википедии - 1100327.

6.2 Предварительная обработка данных

Перед построением тематической модели данные были предобработаны. Сперва были исключены документы, содержащие большое количество иностранных слов и html разметки и других элементов верстки. Оставшиеся документы были разбиты на токены, лемматизированы, из них были исключены знаки препинания, цифры и стоп-слова русского языка, содержащиеся в библиотеке NLTK. С помощью библиотеки RNNMorph[3] . была произведена автоматическая морфологическая разметка, после чего исключены все второстепенные части речи. (Список оставленных частей речи: 'ADJ', 'ADV', 'INTJ', 'NOUN', 'PROPN', 'VERB').

6.3 Тематическая модель

На новостном корпусе была построена иерархическая тематическая модель с двумя модальностями (текст статьи и её заголовок) и тремя уровнями иерархии 9 - 36 - 100 тем. Дополнительно на каждом уровне была выделена одна фоновая тема. Для первых двух уровней использовалось обучение с учителем - для этого создавался сглаживающий регуляризатор для матрицы Θ. В качестве истинных тем использовались категории асессоров.

[3]https://github.com/IlyaGusev/rnnmorph

На документах Википедии была построена иерархическая тематическая модель с двумя модальностями (текст статьи и список идентификаторов категорий, к которым она относится, включая родительские категории) и тремя уровнями иерархии 20 - 100 - 300 тем (аналогично, добавлялись фоновые темы на каждый уровень). Перед началом построения модели на Википедии необходимо было уменьшить объем словаря, чтобы модель не занимала излишне много памяти. Объем словаря новостной коллекции $|W_c| = 40000$ слов. Поскольку в дальнейшем темы обеих моделей необходимо будет сравнивать, в словаре Википедии W_{wiki} были оставлены все слова, встречающиеся в W_c. Помимо этого, было добавлено 50000 частотных слов и 20000 идентификаторов категорий.

Для обеих моделей использовались одинаковые регуляризаторы. Экспериментально было выявлено, что небольшие изменения гиперпараметров не сильно влияют на качество. Для каждой модальности и каждого уровня добавлялось 3 регуляризатора с одинаковыми коэффициентами для всех модальностей:

1. Сглаживающий регуляризатор фоновых тем Φ, $\tau = 0.9$ $\gamma = 0$
2. Декоррелирующий регуляризатор Φ, $\tau = 0.01$ $\gamma = 0$
3. Разреживающий регуляризатор Ψ, $\tau = 2$

Коэффициенты регуляризации были подобраны опытным путём.

6.4 Оценка качества моделей для новостного корпуса

Для оценки качества модели на новостном корпусе помимо перплексии использовались следующие метрики качества кластеризации: Гомогенность, Adjusted Rand Index, Индекс Fowlkes-Mallows, Adjusted Mutual Information.

На графиках обучения (рисунки 2 - 4) указаны две метрики для 0-1 уровней: перплексия и гомогенность. Видно, что достаточно небольшого числа итераций EM-алгоритма для достижения лучшего качества.

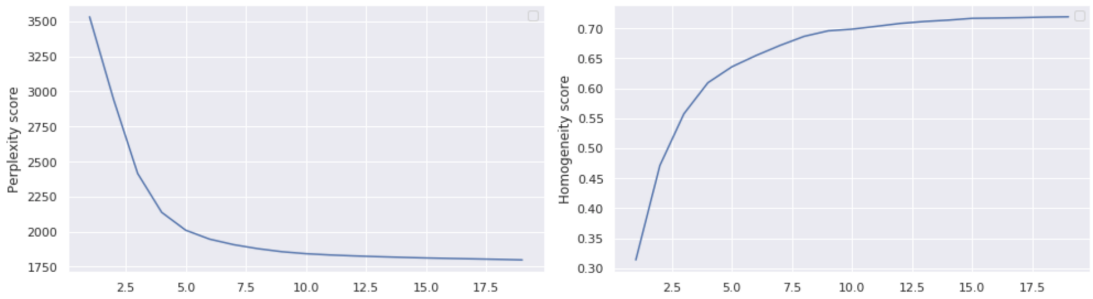


Рис. 2: Графики обучения. Новостной корпус. Уровень 0.

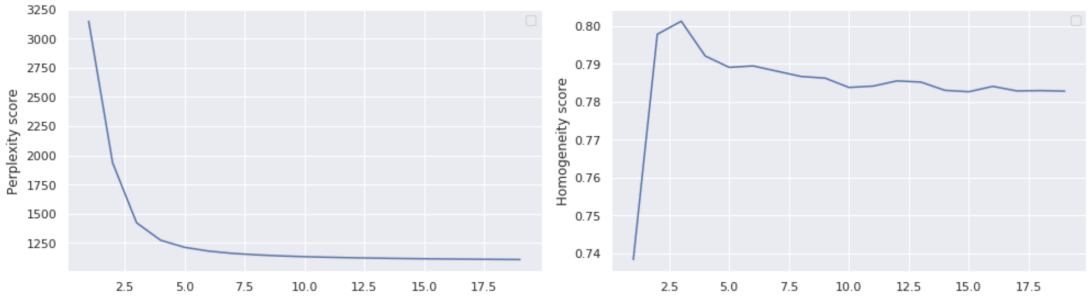


Рис. 3: Графики обучения. Новостной корпус. Уровень 1.

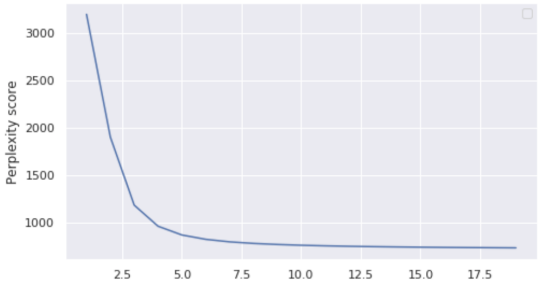


Рис. 4: Графики обучения. Новостной корпус. Уровень 2.

level news	level wiki	count1	count2
0	0	4	1
0	1	4	8
0	2	1	3
1	0	3	0
1	1	21	13
1	2	12	20

Таблица 1: Частоты переходов между уровнями

6.5 Результат именования

В качестве функции близости было выбрано косинусное расстояние между столбцами матриц Φ_c, Φ_{wiki}. Результат применения алгоритма для новостных тем нулевого уровня можно найти в Приложении 1. В приложении 2 - результат для первых тем первого уровня.

В таблице 1 представлены частоты событий "тема уровня <level news> перешла в тему уровня <level wiki> Википедии". count1 - алгоритм, в котором использовалась фунция потерь без последнего солагаемого, count2 - учитывались все 3 слагаемых.

Оценка качества алгоритма именования рассчитывалась следующим образом: для топ-10 выделенных категорий считалось количество тех, которые подходят в качестве названий.

Средняя оценка для тем уровня 0 (9 тем): 5/10.

Число тем с неправильно выбранной темой Википедии: 2.

Средняя оценка для тем уровня 0 (36 тем): 5,55/10.

Число тем с неправильно выбранной темой Википедии: 6.

Гистограмма распределения оценок для обоих уровней показана на рисунке 5

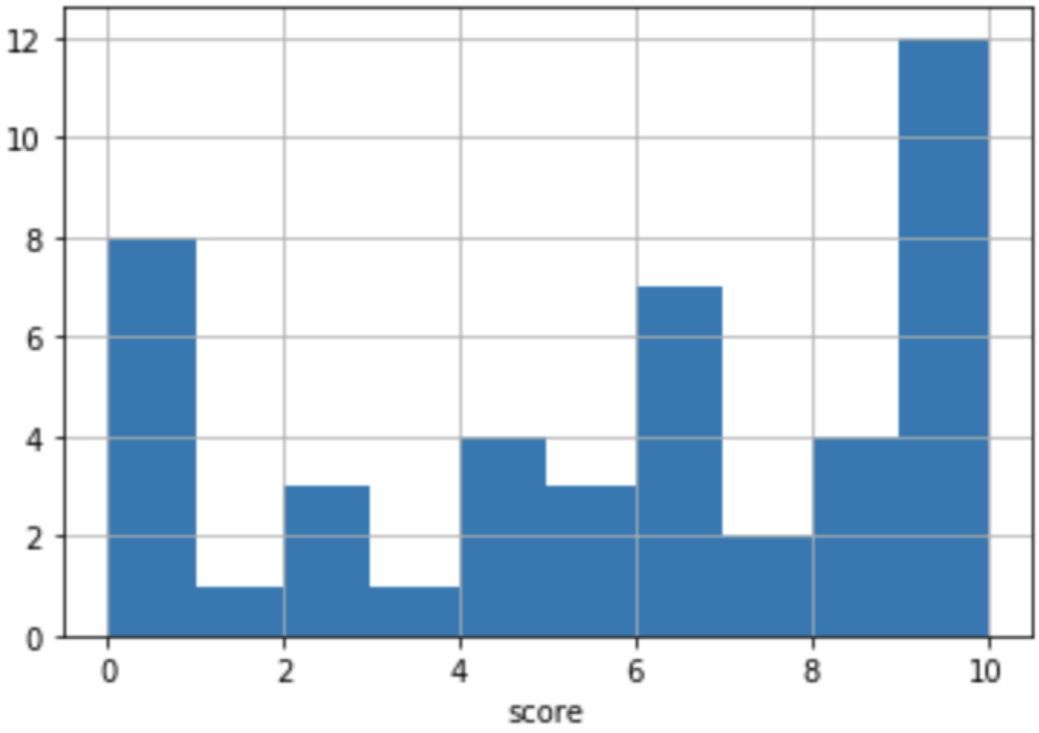


Рис. 5: Гистограмма оценок для первых двух уровней.

7 Заключение.

В данной работе предложен способ автоматического именования тем для иерархической модели ARTM. Подход основан на построении дополнительной иерархической модели на корпусе Википедии с использованием двух модальностей: текстов статей и списков категории, к которым они относятся. Согласно выдвинутой идее, среди множеств тем обеих иерархий существует два подмножества, между которыми можно выстроить соответствие (many-to-one) и в качестве названий тем заимствовать топ категорий Википедии, относящихся к соответствующей теме. Исходя из идеи сохранения структуры иерархии и связей между темами были сформулированы свойства, которым должна обладать матрица перехода между подмножествами. На основе выдвинутых предположений была составлена функция потерь. Метод был опробован на корпусе русскоязычных новостей.

Было продемонстрировано, что предложенный алгоритм умеет

выдавать названия, которые можно использовать для верхних уровней иерархий в тематических моделях. В дальнейших исследованиях планируется комбинировать данных подход с методами именования нижних уровней иерархии, разработать методику оценки качества именования посредством асессорской оценки.

8 Приложение 1.

level news	topic news	level wiki	topic wiki	top tokens news	top tokens wiki	wiki categories	score
0	0	1	40	новый, компания, автомобиль, модель, машина, также, смартфон, кроссовер	автомобиль, двигатель, модель, машина, литр, переть, мощность, задний	Техника, Транспорт, Техника по странам, Транспортные средства по странам, Транспорт по годам, Структура, Транспортная терминология, Теория автомобиля, Компоновка автомобиля, Транспорт в XX веке	6
0	1	1	94	матч, команда, клуб, чемпионат, игра, турнир, сезон, счёт	команда, клуб, матч, место, лига, кубок, группа, чемпионат, финал, один	Спорт по годам, Спорт, Спортивные соревнования, Спортивные клубы по годам основания, Футбольные клубы по годам основания, Спорт по странам и годам, Спорт по странам, Соревнования по видам спорта, Футбольные клубы по странам, Спортивные организации	6
0	2	1	93	военный, март, российский, также, президент, страна, заявить, лидер	армия, войско, сила, война, человек, сражение, военный, отряд, под	Политика, Общество, Общественные науки, Социальные проблемы, Военная история, Военное дело, Военная история по странам, Конфликт, Гражданское общество, События по странам	8
0	3	1	0	человек, март, дело, пожар, произойти, суд, сообщить, сообщать	женщина, время, посол, также, отношение, мочь	Общество, Человек, Общество по странам, Социальные проблемы, Люди, Преступность по странам, Сексуальность человека, Персоналии по времени, Категории по людям, Общественные деятели	4
0	4	1	0	март, апрель, строительство, также, проект, решение, работа, северный	женщина, время, посол, также, отношение, мочь	Общество, Человек, Общество по странам, Социальные проблемы, Люди, Преступность по странам, Сексуальность человека, Персоналии по времени, Категории по людям, Общественные деятели	0
0	5	1	69	фильм, ребёнок, жизнь, человек, известный, актёр, также	сын, дочь, семья, брак, отец, умереть, категория, брат	Умершие, Смерть, Персоналии по датам смерти, Природа, Люди, Персоналии по времени, Человек, Персоналии по датам рождения, Персоналии по месту погребения, Дворянство Германии	6
0	6	0	10	российский, страна, дипломат, дело, заявить, пользователь, британский, также, представитель	посол, человек, время, суд, война, также, сила	Общество, Политика, Общественные науки, Общество по странам, Социальные проблемы, История, Время, События по странам, Политика по странам, События	5
0	7	1	74	рубль, миллион, миллиард, компания, цена, рынок, доллар, март	миллион, предприятие, производство, промышленность, миллиард, рубль, рынок, товар	Экономика, Экономика по отраслям, Экономика по местоположению, Экономика по годам, Предприятия по годам основания, История экономики, Экономика по городам, Экономика по странам, Бизнес, Предприятия по городам	10
0	8	1	0	учёный, человек, специалист, исследование, система, новый, время	женщина, время, посол, также, отношение, мочь	Общество, Человек, Общество по странам, Социальные проблемы, Люди, Преступность по странам, Сексуальность человека, Персоналии по времени, Категории по людям, Общественные деятели	0

9 Приложение 2.

	level news	topic news	level wiki	topic wiki	top tokens news	top tokens wiki	wiki categories	score
9	1	0	1	40	автомобиль, новый, модель, кроссовер, машина, компания, также, получить	автомобиль, двигатель, модель, машина, литр, переть, мощность, задний	Техника, Транспорт, Техника по странам, Транспортные средства по странам, Транспорт по годам, Структура, Транспортная терминология, Теория автомобиля, Компоновка автомобиля, Транспорт в XX веке	8
0	1	1	1	39	законопроект, лицо, организация, банк, система, правительство, фонд, закон, гражданин	суд, право, дело, закон, власть, орган, судебный, уголовный	Политика, Политика по странам, Общество, Общество по странам, Государство, Общество по историческим государствам, Право по историческим государствам, Общественные науки, Права человека, Государственное устройство по странам	6
1	1	2	2	41	матч, чемпионат, команда, сборная, клуб, футболист, игра, сборный	матч, гол, сборная, против, забить, игрок, фк, футболист, клуб	Спорт по странам, Спорт, Сборные по странам, Спорт по городам, Футболисты по странам, Персоналии по времени, Персоналии по датам рождения, Люди, Человек, Незавершённые статьи о футболистах	5
2	1	3	2	202	президент, март, страна, заявить, дело, власть, отношение, также	президент, вице-президент, президентский, срок, республиканский, сенат, сенатор, губернатор, буш, глава	Политика, Политика по странам, Политики по странам, Персоналии:Политика, О политиках, Государственные деятели по странам, Главы государств, Президенты по странам, Политики США, Политики по векам	10
3	1	4	1	3	спортсмен, виза, мир, олимпийский, борьба, кубок, российский, спорт, отставка	олимпийский, игра, категория, мир, чемпионат, чемпион, летний, спорт, призёр	Спорт, Спортивные соревнования, Спортсмены по спортивным соревнованиям, Участники Олимпийских игр, Виды спорта, Категории видов спорта по странам, Спорт по странам, Спортсмены по видам спорта по странам, Спортивные звания, Чемпионы	4
4	1	5	2	54	человек, дело, произойти, март, пожар, результат, полиция, сообщить	дело, убийство, обвинение, расследование, суд, посол, заявить	Общество по странам, Общество, Преступность по странам, Человек, Люди, Персоналии по времени, Персоналии по датам рождения, Общественные деятели по странам, Преступники по странам, Люди по роду занятий	2
5	1	6	1	97	военный, система, боевой, армия, российский, комплекс, техника, также	война, военный, боевой, советский, сила, армия, войско, бой	Персоналии по датам смерти, Природа, Смерть, Умершие, Военное дело, Техника, Персоналии:Техника, Общественные науки, Персоналии:Авиация, Участники Второй мировой войны	4
6	1	7	2	82	срок, средство, право, также, гражданин, изменение, документ, изделие	власть, конституция, реформа, новый, орган, закон, полномочие, стать	Политика, Политика по странам, Государственное устройство по странам, Общество, Политики по странам, Государство, Государственное устройство исторических государств, Политическая история, Общество по странам, Исчезнувшие организации	6
7	1	8	2	107	движение, апрель, транспорт, март, дорога, пассажир, улица, участок, самолёт	маршрут, транспорт, автобус, вагон, до, трамвай, транспортный, движение	Транспорт, Транспорт по странам, Социальная инфраструктура, Общественный транспорт, Транспорт по городам, Дороги, География транспорта, Общество, Транспорт по городам России, Типы дорог	10

Список литературы

[1] К. В. Воронцов *Вероятностное тематическое моделирование: обзор моделей и аддитивная регуляризация* 2018, http://www.machinelearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf

[2] D. Newman, T. Baldwin, L. Cavedon, S. Karimi, D. Mar- tinez, and J. Zobel. 2010a. *Visualizing document col- lections and search results using topic mapping.*, Journal of Web Semantics, 8(2-3):169–175.

[3] Qiaozhu Mei, Xuehua Shen, Chengxiang Zhai *Automatic Labeling of Multinomial Topic Models*, KDD'07, August 12–15, 2007.

[4] El-Kishky, Ahmed, et al. *Scalable topical phrase mining from text corpora.* Proceedings of the VLDB Endowment 8.3 (2014): 305-316.

[5] Magatti, D., S. Calegari, D. Ciucci, and F. Stella. *Automatic labeling of topics.*, Proceedings of the International Conference on Intelligent Systems Design and Applications, pages 1227–1232, Pisa, Italy, 2009.

[6] Jey Han Lau, Karl Grieser, David Newman, Timothy Baldwin *Automatic Labelling of Topic Models*, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 1536–1545, 2011.

[7] Xian-Ling Mao, Zhao-Yan Ming, Zheng-Jun Zha, Tat-Seng Chua, Hongfei Yan, Xiaoming Li *Automatic Labeling Hierarchical Topics*, CIKM'12, October 29-November 2, 2012

[8] I. Hulpus, C. Hayes, M. Karnstedt, and D. Greene *Unsupervised graph-based topic labelling using dbpedia*, Proceedings of the sixth ACM international conference on Web search and data mining. ACM, 2013, pp. 465–474.

[9] Shraey Bhatia, Jey Han Lau, and Timothy Baldwin. *Automatic labelling of topics with neural embeddings.*, Proceedings of the 26th

International Conference on Computational Linguistics (COLING 2016), pages 953–963, Osaka, Japan, 2016

[10] I Sorodoc, JH Lau, N Aletras, T Baldwin. *Multimodal Topic Labelling.*, Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 701-706, 2017

[11] B. Ramakrishna Bairi, Mark Carman, Ganesh Ramakrishnan. *On the Evolution of Wikipedia: Dynamics of Categories and Articles.*, Wikipedia, a Social Pedia: Research Challenges and Opportunities: Papers from the 2015 ICWSM Workshop, pages 6-10.