Московский Государственный Университет
имени М. В. Ломоносова

На правах рукописи

Нокель Михаил Алексеевич

МЕТОДЫ УЛУЧШЕНИЯ ВЕРОЯТНОСТНЫХ ТЕМАТИЧЕСКИХ МОДЕЛЕЙ ТЕКСТОВЫХ КОЛЛЕКЦИЙ НА ОСНОВЕ ЛЕКСИКО-ТЕРМИНОЛОГИЧЕСКОЙ ИНФОРМАЦИИ

Специальность 05.13.11 – математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

ДИССЕРТАЦИЯ
на соискание учёной степени
кандидата физико-математических наук

Научный руководитель:
кандидат физико-математических наук,
Лукашевич Наталья Валентиновна

Москва – 2015

Содержание

Введение

В настоящее время в связи с бурным развитием сети Интернет наблюдается обилие электронной неструктурированной информации, представленной текстами на естественных языках. Всё более востребованной становится задача автоматической обработки таких текстов с целью извлечения структурированных данных, которые затем используются при решении различного рода проблем: извлечения фактических данных, поиска информации и т.п. [61].

Под **обработкой текстов на естественном языке** обычно понимается теоретически обоснованный набор вычислительных методов анализа и представления естественных текстов на нескольких уровнях лингвистического анализа с целью достижения качества, соответствующего обработке вручную, для последующего применения в различных задачах и приложениях [35].

Одной из важных задач автоматической обработки текстов является кластеризация текстов (автоматическое выделение групп похожих документов), и, в частности, выделение тематик, обсуждаемых в коллекциях. Для выявления скрытых тем в текстовых коллекциях в последнее время всё чаще применяются **статистические тематические модели** (далее просто **тематические модели**) [3]. Это современный инструмент анализа текстов, определяющий, какие темы присутствуют в каждом документе коллекции, и какие слова задают каждую такую тему. При этом темы представляются в виде дискретных распределений на множестве слов, а документы – в виде дискретных распределений на множестве тем [3].

Тематические модели осуществляют «нечёткую» кластеризацию слов и документов по кластерам-темам, означающую, что слово или документ могут быть отнесены сразу к нескольким темам с различными вероятностями. При этом синонимы с большой вероятностью окажутся в одних и тех же темах, поскольку часто употребляются в рамках одних и тех же документов. В то же время омонимы (разные по значению, но одинаковые по написанию слова)

попадут в различные темы, так как употребляются в различных контекстах.

Фактически каждая тема задаётся списком часто встречающихся в одних и тех же контекстах слов. Так, например, в коллекциях исследовательских работ темы будут соответствовать различным теориям, методам и алгоритмам. В коллекциях новостей темы могут соответствовать событиям, компаниям, процессам и т.д. Примером такой темы может служить следующая (для наглядности вероятности слов опущены): *денежный, деньги, обращение, масса, факторинг, средство, функция, оборот, факторинговый, товар* ...

На данный момент тематические модели успешно применяются в информационном поиске [93], разрешении морфологической неоднозначности [23], многодокументном аннотировании [91], машинном переводе [37], категоризации и кластеризации документов [101], обнаружении спама [10]. Также на их основе были достигнуты значительные успехи в выявлении трендов в научных публикациях и новостных потоках [19], обработке и визуализации аудио- и видеосигналов [48], разработке рекомендательных систем [97] и многих других задачах.

Однако, несмотря на значительный успех, тематические модели не лишены недостатков. Одним из них является использование модели «мешка слов», в которой каждый документ представляется в виде множества несвязанных между собой слов. Данная модель не учитывает порядок слов и основывается на гипотезе независимости появлений слов друг от друга в текстах. Это предположение оправдано с точки зрения вычислительной эффективности, но оно далеко от реальности. Так, некоторые слова меняют свой смысл при объединении в словосочетания: например, словосочетание *«точка зрения»* плохо связана со своими компонентами *«точка»* и *«зрение»*.

Таким образом, актуальной является задача улучшения качества тематических моделей за счёт добавления в них подходящих словосочетаний и многословных выражений (в частности, терминов) и учёта связей между ними и образующими их словами.

При этом само по себе понятие «термин» не формализовано. Одним из наиболее распространённых определений является следующее. **Термин** – слово (или словосочетание), являющееся точным обозначением определённого понятия какой-либо специальной области науки, техники, искусства, общественной жизни и т.п. [4]. Так, в банковской предметной области терминами будут такие слова и словосочетания, как *«банк»*, *«кредит»*, *«инвестиционный фонд»*, *«ипотечный кредит»*, *«лизинг»*, а в общественно-политической – *«соглашение»*, *«перемирие»*, *«сельскохозяйственный сектор»*. В отличие от общеупотребительных слов термины в рамках некоторой конкретной предметной области всегда однозначны и лишены эмоциональной окраски. Тем не менее, следует отметить, что термины и общеупотребительные слова могут переходить друг в друга. Кроме того, терминологичность слов и словосочетаний полностью зависит от рассматриваемой предметной области. Так, слово *«акт»* является термином в банковской области и не является таковым в общественно-политической.

Исторически словари терминов составлялись вручную экспертами. Впоследствии такие словари могут применяться при разработке различных баз знаний, для улучшения качества информационного поиска [78], машинного перевода [94], автоматического реферирования [43] и в других задачах. Однако привлечение экспертов – очень трудоёмкий и дорогостоящий процесс, и полученные словари обладают низкой полнотой покрытия, поскольку не содержат постоянно появляющихся новых терминов. Поэтому актуальной является задача автоматического извлечения терминов для различных предметных областей.

На настоящий момент большинство применяющихся на практике методов используют для извлечения терминов множество критериев, основывающихся на статистических и лингвистических признаках. Однако ни один из таких признаков не является определяющим [77], и фактически из текстов извлекается довольно большой список слов и словосочетаний, являющихся лишь кандидатами в термины, которые затем должны быть проанализированы и подтверждены экспертами по предметной области текстов. Важно поэтому задействовать как

можно больше признаков, описывающих кандидаты с разных сторон, и провести в ходе извлечения ранжирование кандидатов, так, чтобы в начале итогового списка стояли слова и словосочетания, действительно являющиеся терминами, для минимизации ручных затрат на проверку кандидатов.

Кроме того, на текущий момент традиционно используемые для извлечения терминов статистические признаки почти не отражают тот факт, что большинство терминов относятся к той или иной тематике, обсуждаемой в текстах коллекции. Поэтому актуальной является задача использования тематической информации при извлечении терминов.

Цель диссертационной работы

Целью данной диссертационной работы является разработка методов и программных средств интеграции лексико-терминологической информации в тематические модели, выбирающих наиболее подходящие единицы из слов и словосочетаний для интеграции и улучшения качества. Разрабатываемые программные средства и полученные модели должны удовлетворять следующим требованиям:

- Более высокое по сравнению с существующими методами качество тематических моделей;
- Независимость от языков и предметных областей текстовых коллекций.
- Более высокое по сравнению с существующими методами качество извлечённых из текстовых коллекций терминов.

Для достижения этой цели были поставлены и решены следующие **задачи**:

1. Исследование и разработка методов построения тематических моделей, учитывающих словосочетания и связи между ними и образующими их словами;

2. Разработка и реализация методов извлечения терминов на основе информации, получаемой из тематических моделей.

Основные положения, выносимые на защиту:

1. Предложен и реализован новый метод построения тематических моделей, учитывающий словосочетания и улучшающий характеристики качества тематических моделей, включая интерпретацию тем экспертами, что полезно для организации человеко-машинных интерфейсов в информационных системах. Для предложенного метода приведено теоретическое обоснование;

2. Предложен и реализован новый итеративный метод добавления словосочетаний в тематические модели, улучшающий меру соответствия тематических моделей словам и словосочетаниям текстовых коллекций (перплексию). Для предложенных методов приводятся теоретические оценки вычислительной сложности;

3. Предложены новые признаки для извлечения терминов, основанные на тематических моделях. Показано, что использование тематической информации улучшает качество извлечения терминов для включения их в базы знаний и терминологические ресурсы;

4. Разработан и выложен в открытый доступ программный комплекс по построению тематических моделей с использованием лексико-терминологической информации.

Научная новизна

Научная новизна настоящей диссертационной работы заключается в том, что предложены новые алгоритмы построения тематических моделей, позволяющие добавлять словосочетания и учитывать сходство между ними и образу-

ющими их словами. Кроме того, предложены новые признаки для извлечения терминов, основанные на тематической информации. Применимость данных алгоритмов обоснована теоретически, а также численно, на основе тестирования интеграции словосочетаний в тематические модели. Разработанные модели и алгоритмы не зависят от языка и предметной области текстовых коллекций и могут применяться в различных задачах информационного поиска.

Методы исследования

В данной диссертационной работе применялись аналитические методы теории вероятностей, математической статистики, теории обучения машин и теории алгоритмов.

Практическая значимость

На основе предложенных методов спроектирована и разработана многомодульная программная система со следующими функциональными возможностями:

- построения тематических моделей с добавлением словосочетаний, учитывающих сходство между ними и образующими их словами;
- извлечения однословных и двусловных терминов из текстов предметной области;
- автоматических оценок качества построенных тематических моделей и извлечённых терминов.

Таким образом, разработанная система может быть использована как для автоматического пополнения существующих терминологических ресурсов (словарей, тезаурусов), так и для подготовки дополнительной входной информации для других систем обработки текстов на естественном языке.

Результаты научных исследований, представленных в диссертации, были получены в рамках гранта РФФИ №14-07-00383.

Апробация работы

Основные результаты работы докладывались на следующих конференциях и научных семинарах:

1. Международной конференции по компьютерной лингвистике «Диалог» (Московская область, 30 мая – 3 июня 2012 г.);

2. Летней школе по информационному поиску RUSSIR (Ярославль, 6–10 августа 2012 г.; Казань, 16–20 сентября 2013 г.; Нижний Новгород, 18–22 августа 2014 г.; Санкт-Петербург, 24–28 августа 2015 г.);

3. Международной конференции по информационному поиску ECIR (Москва, 24–27 марта 2013 г.);

4. Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (Ярославль, 14–17 октября 2013 г.; Дубна, 13–16 октября 2014 г.);

5. Международной конференции «Terminology and Artificial Intelligence» (Париж, 28–30 октября 2013 г.);

6. Международной научной конференции студентов, аспирантов и молодых учёных «Ломоносов» (Москва, 7–11 апреля 2014 г.);

7. Международной конференции «Nordic Conference on Computational Linguistics» (Вильнюс, 11–13 мая 2015 г.);

8. Семинаре по многословным выражениям MWE на конференции «North American Chapter of the Association for Computational Linguistics – Human Language Technologies» (Денвер, 31 мая – 5 июня 2015 г.).

Кроме того, результаты обсуждались на семинаре лаборатории анализа информационных ресурсов НИВЦ МГУ и на регулярном семинаре ACM SIGMOD в Москве.

Публикации

Основные результаты по теме диссертации изложены в 12 печатных работах, в том числе в 4 статьях в журналах из списка ВАК [1, 7–9], 3 статьях, входящих в базу SCOPUS [20, 70, 74], 1 – в тезисах докладов [6], и 4 – в других изданиях [71–73, 75].

Личный вклад

Личный вклад автора заключается в выполнении основного объёма теоретических и экспериментальных исследований, изложенных в диссертационной работе, включая разработку теоретических моделей, методик и проведение исследований, анализ и оформление результатов в виде публикаций и научных докладов. В работах [8,73,74] Н. В. Лукашевич принадлежит постановка задачи и обсуждение результатов её решения. В работах [1, 20, 71] вклад Е. И. Большаковой ограничен обсуждением результатов, Н. В. Лукашевич – постановкой задачи. В работах [9, 72, 75] Н. В. Лукашевич принадлежит постановка задачи, привлечение лингвистов для получения экспертных оценок тем, выявленных в текстовых коллекциях, и обсуждение результатов.

Объём и структура работы

Диссертация состоит из введения, четырёх глав, заключения и двух приложений. Полный объём диссертации составляет 137 страниц с 14 рисунками и 30 таблицами, объём приложений – 22 страницы. Список литературы содержит 101 наименование.

1 Анализ предметной области

Данная глава посвящена описанию основных подходов, применяемых в задачах построения тематических моделей и автоматического извлечения терминов. Особое внимание уделяется рассмотрению существующих способов добавления словосочетаний и многословных выражений в тематические модели. Кроме того, в данной главе описываются существующие признаки ранжирования слов и словосочетаний-кандидатов в термины. Целью данной главы является анализ достоинств и недостатков существующих методов интеграции словосочетаний в тематические модели и автоматического извлечения терминов.

1.1 Тематический анализ текстовых коллекций

С конца 1990-х годов появился и стал успешно развиваться новый вид анализа документов – статистический с применением тематических моделей. Тематические модели предназначены для выявления скрытых тем в текстовых коллекциях. Они определяют, к каким темам относится каждый документ коллекции, и какие слова формируют каждую тему. При этом каждая тема представляется в виде некоторого дискретного распределения на множестве слов [3].

Тематические модели осуществляют «нечёткую» кластеризацию слов и документов по кластерам-темам. Это означает, что слово или документ могут относиться сразу к нескольким темам с различными вероятностями. При этом слова, часто встречающиеся в одних и тех же контекстах, с большой вероятностью попадут в одну и ту же тему, а слова, употребляющиеся в различных контекстах, распределятся между разными темами.

Тематические модели, как правило, используют модель «мешка слов», в которой каждый документ рассматривается как множество несвязанных между собой слов. При этом перед выделением тем текстовая коллекция обычно подвергается предобработке, где обычно проводится морфологический анализ документов с целью определения начальной формы слов (*леммы*) и их ча-

стей речи. Данный процесс необходим, поскольку темы, как правило, задаются именными группами, а роль служебных слов в этом мала. Поэтому в качестве слов, участвующих в образовании тем, обычно оставляют *прилагательные*, *существительные*, *глаголы* и *наречия*.

На сегодняшний день разработано достаточно большое количество алгоритмов построения тематических моделей. Исторически первыми появились методы, основанные на традиционной кластеризации текстов [45]. Они основываются на методах «жёсткой» кластеризации, рассматривающих каждый документ как разреженный вектор в пространстве слов большой размерности [81]. После окончания работы алгоритма кластеризации каждый получившийся кластер рассматривается как отдельная тема, содержащая в себе слова с вероятностями, вычисленными по следующей формуле:

$$P(w|t) = \frac{TF(w|t)}{\sum\limits_{w \in t} TF(w|t)}, \tag{1}$$

где $TF(w|t)$ – частотность слова w в теме-кластере t.

В качестве алгоритма кластеризации может выступать любой из известных. На данный момент выделяются следующие основные группы.

1.1.1 Алгоритм K-Средних и его модификации

Это наиболее популярный метод кластеризации [60], разбивающий n объектов на k кластеров, относя каждое наблюдение к кластеру с ближайшим центром. При кластеризации текстов в качестве наблюдений выступают вектора слов $(x_1, \ldots, x_n)$. Алгоритм пытается разбить их на k множеств $S_1, \ldots, S_k$, минимизируя суммарное отклонение точек кластеров от центров этих кластеров:

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x \in S_i} d(x, \mu), \tag{2}$$

где μ – центр точек в кластере S_i, $d(x, \mu)$ – функция расстояния между объектами x и μ.

Несмотря на то, что проблема нахождения такого разбиения является NP-трудной, существуют эффективные эвристические алгоритмы, применяющиеся на практике и сходящиеся к локальному оптимальному значению. Одним из таких широко использующихся алгоритмов является алгоритм Ллойда [57].

Согласно этому алгоритму вначале необходимо провести инициализацию центров строящихся кластеров. После этого, получив центры начальных кластеров $m_1^{(1)}, \ldots, m_k^{(1)}$, алгоритм итеративно повторяет следующие шаги (до прекращения изменений в кластерах или до заданного числа итераций):

1. **Шаг присваивания**. Все объекты разбиваются на кластеры в соответствии с тем, какой из центров оказался ближе по заданной метрике. Таким образом определяются объекты, относящиеся к каждому из кластеров $S_i^{(t)}$:

$$S_i^{(t)} = \{x_p : d(x_p, m_i^{(t)}) \leq d(x_p, m_j^{(t)}), \forall j : 1 \leq j \leq k\} \tag{3}$$

2. **Шаг обновления**. Для каждого кластера пересчитывается его центр:

$$m_i^{t+1} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j \tag{4}$$

В качестве меры близости между двумя объектами могут использоваться различные метрики, наиболее популярными из которых являются следующие:

- *Квадрат Евклидова расстояния.* В этом случае получается оригинальный алгоритм *K-Средних*:

$$d(x, y) = \sum_k (x_k - y_k)^2 \tag{5}$$

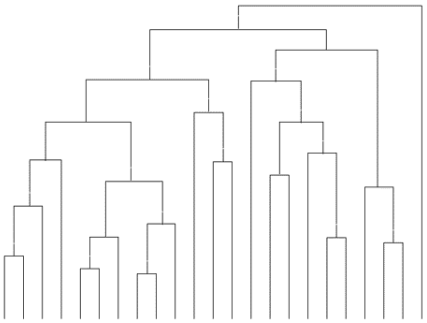
- *Косинусное расстояние.* В этом случае все векторы x_i, представляющие объекты, нормализуются к единичной гиперсфере. В результате получается алгоритм *Сферический K-Средних* [100]:

$$d(x, y) = 1 - \frac{\sum\limits_{k}(x_k \times y_k)}{\sqrt{\sum\limits_{k}(x_k)^2} \times \sqrt{\sum\limits_{k}(y_k)^2}} \tag{6}$$

1.1.2 Иерархические алгоритмы кластеризации

Под иерархическими алгоритмами кластеризации понимаются методы, строящие иерархию кластеров, обычно представляемую в виде *дендрограммы* – дерева, построенного по матрице мер близости (см. Рисунок 1).

Рис. 1: Пример дендрограммы при иерархической кластеризации



Существуют две стратегии построения иерархической кластеризации [50]:

- **Агломеративная**. Этот метод заключается в построении дендрограммы снизу вверх. Вначале каждый объект относится к своему кластеру. Затем кластеры объединяются с получением более высоких уровней иерархии;
- **Дивизимная**. Этот метод заключается в построении дендрограммы сверху вниз. Вначале все объекты относятся к одному кластеру. Затем кластеры разбиваются с получением более низких уровней иерархии.

Сложность дивизимных алгоритмов равна $O(2^n)$, где n – число объектов, что делает эти алгоритмы неприменимыми на практике. Поэтому обычно используются агломеративные алгоритмы со сложностью $O(n^3)$. В них процесс слияния наиболее близких кластеров и пересчёта расстояний между ними и остальными продолжается, пока не останется заданное число кластеров.

Для определения того, какие из кластеров надо объединить, необходима мера различия объектов. Обычно для этого выбирают подходящую меру расстояния между объектами и критерий определения наиболее похожих кластеров.

При алгомеративной кластеризации текстовых документов наиболее часто используется квадрат Евклидова расстояния между документами x и y:

$$d(x, y) = \sum_i (x_i - y_i)^2 \tag{7}$$

В качестве критериев определения наиболее похожих кластеров обычно используются следующие [50]:

- *Полное связывание.* Наиболее близкие кластеры – это кластеры с наименьшим максимальным парным расстоянием между объектами. Расстояние $D(X, Y)$ между кластерами X и Y вычисляется следующим образом:

$$D(X, Y) = \max\{d(x, y) : x \in X, y \in Y\} \tag{8}$$

- *Одиночное связывание.* Наиболее близкие кластеры – это кластеры с наименьшим минимальным парным расстоянием между объектами. Расстояние $D(X, Y)$ между кластерами X и Y вычисляется следующим образом:

$$D(X, Y) = \min\{d(x, y) : x \in X, y \in Y\} \tag{9}$$

- *Среднее связывание.* Наиболее близкие кластеры – это кластеры с наименьшим средним парным расстоянием между объектами. Расстояние $D(X, Y)$ между кластерами X и Y вычисляется следующим образом:

$$D(X, Y) = \frac{1}{|X| \times |Y|} \sum_{x \in X} \sum_{y \in Y} d(x, y), \tag{10}$$

где $|X|$ и $|Y|$ – число объектов в кластерах X и Y соответственно.

Ограничением всех тематических моделей, основанных на методах «жёсткой» кластеризации является отнесение документа лишь к одной теме, в то время как почти в любом документе затрагивается несколько различных тем.

1.1.3 Неотрицательная матричная факторизация

Помимо «жёстких» алгоритмов кластеризации существуют методы, выполняющие «нечёткую» кластеризацию, относящие один и тот же объект к разным кластерам с различными вероятностями. Одним из таких алгоритмов является алгоритм NMF (неотрицательной матричной факторизаци) [96].

Принимая на вход неотрицательную матрицу V, которая получается записыванием векторов документов по столбцам, алгоритм ищет такие неотрицательные матрицы W и H меньшей размерности, что $V \approx W \times H$ по некоторой метрике. В качестве таких метрик обычно рассматриваются следующие [54]:

- *Квадрат Евклидова расстояния* между двумя матрицами A и B:

$$||A - B||^2 = \sum_{i,j}(A_{ij} - B_{ij})^2 \tag{11}$$

- *Расстояние Кульбака-Лейблера* между двумя матрицами A и B:

$$D(A|B) = \sum_{i,j}\left(A_{ij}\log\frac{A_{ij}}{B_{ij}} - A_{ij} + B_{ij}\right) \tag{12}$$

Для нахождения приближённого разложения $V \approx WH$ существует несколько способов, наиболее популярным из которых является мультипликативный итерационный алгоритм, предложенный в работе [54]. Согласно данному алгоритму вначале матрицы W и H заполняются случайным образом. После чего итеративно (до сходимости или до заданного числа итераций) повторяются мультипликативные правила обновления матриц W и H.

Для минимизации Евклидова расстояния используются следующие правила обновления матриц H и W [54]:

$$H_{ij} \leftarrow H_{ij}\frac{(W^TV)_{ij}}{(W^TWH)_{ij}}, \qquad W_{ij} \leftarrow W_{ij}\frac{(VH^T)_{ij}}{(WHH^T)_{ij}} \tag{13}$$

Для минимизации расстояния Кульбака-Лейблера применяются следующие правила обновления матриц W и H [54]:

$$H_{ij} \leftarrow H_{ij} \frac{\left(\sum\limits_k \frac{W_{ki} V_{kj}}{(WH)_{kj}} \right)}{\sum\limits_k W_{ki}}, \qquad W_{ij} \leftarrow W_{ij} \frac{\left(\sum\limits_k \frac{H_{jk} V_{ik}}{(WH)_{ik}} \right)}{\sum\limits_k H_{jk}} \tag{14}$$

В результате работы алгоритма в матрице W получается распределение слов по кластерам, а в матрице H распределение векторов слов (документов) по кластерам. Нормируя соответствующие величины для каждого слова (документа), можно получить вероятности принадлежности слов (документов) кластеру.

Дальнейшим развитием идей «нечёткой» кластеризации стали вероятностные методы выявления тем, рассматривающие каждый документ в виде смеси нескольких тем, а каждую тему – в виде некоторого распределения над словами. При этом порождение слов происходит по следующему правилу:

$$P(w|d) = \sum_t P(w|t)P(t|d), \tag{15}$$

где $P(w|t)$ и $P(t|d)$ – скрытые распределения слов по темам и тем по документам, а $P(w|d)$ – наблюдаемое распределение слов по документам.

При известных распределениях $P(w|t)$ и $P(t|d)$ порождение слов в документах происходит согласно Алгоритму 1.

Задача построения тематической модели заключается в восстановлении скрытых распределений $P(w|t)$ и $P(t|d)$ по известной коллекции документов D. Самыми известными представителями данной категории моделей являются Латентное Размещение Дирихле (**LDA**) [18], использующее априорное распределение параметров Дирихле, и метод Вероятностного Латентного Семантического Анализа (**PLSA**) [46], не использующий никаких априорных распределений. При описании данных методов будут использоваться следующие обозначения:

- w – слово в текстовой коллекции;

Алгоритм 1: Алгоритм порождения коллекции текстов с помощью тематической модели

Вход: распределения $P(w|t)$ и $P(t|d)$

Выход: коллекция документов D

```
for d ∈ D do
    Задать длину документа n_d
    for i = 1, ..., n_d do
        Сэмплировать тему t из распределения P(t|d)
        Сэмплировать слово w из распределения P(w|t)
        Добавить в документ d коллекции D слово w
```

- d – документ в текстовой коллекции;
- n_{dw} – частотность слова w в документе d;
- D – текстовая коллекция;
- T – множество выделяемых тем в текстовой коллекции D;
- W – множество уникальных слов в текстовой коллекции D (словарь);
- $\Phi = \{\phi_{wt}\} = \{P(w|t)\}$ – матрица скрытых распределений $P(w|t)$;
- $\Theta = \{\theta_{td}\} = \{P(t|d)\}$ – матрица скрытых распределений $P(t|d)$.

1.1.4 Метод Вероятностного Латентного Семантического Анализа

Метод PLSA решает поставленную задачу восстановления скрытых распределений $P(w|t)$ и $P(t|d)$ методом максимума правдоподобия [46]:

$$\log \prod_{d \in D} \prod_{w \in d} P(w|d)^{n_{dw}} = \sum_{d \in D} \sum_{w \in d} n_{dw} \log \sum_{t \in T} \phi_{wt} \theta_{td} \to \max_{\Phi, \Theta} \tag{16}$$

при ограничениях неотрицательности и нормировки:

$$\phi_{wt} = P(w|t) \geq 0; \qquad \sum_{w \in W} \phi_{wt} = 1; \qquad \theta_{td} = P(t|d) \geq 0; \quad \sum_{d \in D} \theta_{td} = 1 \tag{17}$$

Поскольку тематическая модель зависит от нескольких скрытых переменных, для нахождения оценок максимального правдоподобия параметров Φ и Θ используется *EM-алгоритм* (*Expectation-Maximization*) [31]. Это итеративный алгоритм, каждая итерация которого состоит из двух шагов, повторяющихся до сходимости или до заданного числа итераций:

1. *E-шаг* (*Expectation-шаг*). На данном шаге вычисляется ожидаемое значение функции правдоподобия, при этом скрытые переменные рассматриваются как наблюдаемые. В рассматриваемой задаче условные вероятности $P(t|d,w)$ для всех тем t, документов d и слов w вычисляются через скрытые параметры ϕ_{wt} и θ_{td} по формуле Байеса:

$$P(t|d,w) = \frac{P(w,t|d)}{P(w|d)} = \frac{P(w|t)P(t|d)}{P(w|d)} = \frac{\phi_{wt}\theta_{td}}{\sum\limits_{s \in T} \phi_{ws}\theta_{sd}} \tag{18}$$

2. *M-шаг* (*Maximization-шаг*). На данном шаге находится оценка максимального правдоподобия, тем самым увеличивается ожидаемое правдоподобие. В рассматриваемой задаче частотные оценки условных вероятностей вычисляются путём суммирования счётчика $n_{dwt} = n_{dw}P(t|d,w)$:

$$\phi_{wt} = \frac{\sum\limits_{d \in D} n_{dwt}}{\sum\limits_{d \in D} \sum\limits_{w \in d} n_{dwt}}; \qquad \theta_{td} = \frac{\sum\limits_{w \in d} n_{dwt}}{\sum\limits_{w \in W} \sum\limits_{t \in T} n_{dwt}} \tag{19}$$

В работе [33] теоретически обосновано, что алгоритм *NMF*, минимизирующий *расстояние Кульбака-Лейблера*, эквивалентен алгоритму *PLSA*. Поэтому в дальнейшем эти два алгоритма отождествляются друг с другом.

К основным недостаткам тематической модели PLSA относят следующие:

- Число параметров модели ($DT + WT$) слишком велико и растёт линейно по числу документов, что способствует переобучению модели;
- При добавлении нового документа d невозможно вычислить распределение $p(t|d)$ для этого документа, не перестраивая всю модель заново.

1.1.5 Метод Латентного Размещения Дирихле

Для решения описанных выше недостатков PLSA была предложена новая модель – метод Латентного Размещения Дирихле (LDA) [18]. В данном методе для борьбы с переобучением модели вводится байесовская регуляризация, основанная на использовании априорного распределения Дирихле.

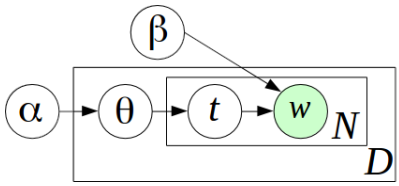
Метод LDA предполагает, что векторы документов $\theta_d = (\theta_{td}) \in \mathbf{R}^T$ и векторы тем $\phi_t = \phi_{wt} \in \mathbf{R}^W$ порождаются распределением Дирихле с параметрами $\alpha \in \mathbf{R}^T$ и $\beta \in \mathbf{R}^W$ (они именуются гиперпараметрами) [18]:

$$\begin{aligned} Dir(\theta_d; \alpha) &= \frac{(\alpha_0)}{\prod\limits_w (\alpha_t)} \prod_t \theta_{td}^{\alpha_t - 1}, \ \alpha_t > 0, \ \alpha_0 = \sum_t \alpha_t, \ \theta_{td} > 0, \ \sum_t \theta_{td} = 1 \\ Dir(\phi_t; \beta) &= \frac{(\beta_0)}{\prod\limits_w (\beta_w)} \prod_w \phi_{wt}^{\beta_w - 1}, \ \beta_0 = \sum_w \beta_w, \ \phi_{wt} > 0, \ \sum_w \phi_{wt} = 1, \end{aligned} \tag{20}$$

где (x) – гамма-функция.

Графическая схема модели LDA представлена на Рисунке 2.

Рис. 2: Тематическая модель LDA



Модель LDA позволяет хорошо описывать тематические структуры текстовых коллекций [3]. Чем меньше значение α_0, тем сильнее различаются документы θ_d. Чем меньше значение β_0, тем сильнее различаются темы ϕ_t. Та-

ким образом, гиперпараметры позволяют моделировать тематические структуры различной степени выраженности.

Известно, что тематическую модель LDA также можно строить с помощью EM-алгоритма, использующегося при построении тематической модели PLSA, с другими формулами M-шага, приводящими к сглаживанию [3]:

$$\phi_{wt} = \frac{\sum\limits_{d \in D} n_{dwt} + \beta_w}{\sum\limits_{d \in D} \sum\limits_{w \in d} n_{dwt} + \beta_0} \qquad \theta_{td} = \frac{\sum\limits_{w \in d} n_{dwt} + \alpha_t}{\sum\limits_{w \in W} \sum\limits_{t \in T} n_{dwt} + \alpha_0} \tag{21}$$

1.1.6 Критерии оценки качества тематических моделей

Наиболее известным критерием является **Перплексия**, используемая для оценки качества различных языковых моделей [28]. Перплексия представляет собой меру несоответствия модели $P(w|d)$ словам w, наблюдаемым в документах коллекции, и определяется через логарифм правдоподобия:

$$Perplexity(D) = \exp\left(-\frac{1}{n} \sum_{d \in D} \sum_{w \in d} n_{dw} \log P(w|d)\right), \tag{22}$$

где n – число всех рассматриваемых слов в текстовой коллекции, D – множество всех документов в коллекции, n_{dw} – частотность слова w в документе d, $P(w|d)$ – вероятность появления слова w в документе d.

Чем меньше значение перплексии, тем лучше модель предсказывает появление слов w в документах коллекции D. Поскольку известно, что перплексия, вычисленная по той же самой коллекции, может давать оптимистически заниженные оценки [18], чаще применяется метод вычисления контрольной перплексии [14]. Текстовая коллекция при этом разбивается на две части: обучающую D, по которой строится модель, и контрольную D', по которой вычисляется мера. При этом параметры ϕ_{wt} оцениваются только по обучающей коллекции D. После обучения параметры ϕ_{wt} фиксируются, а каждый контрольный документ $d \in D'$ случайно разбивается на две половины: по первой из них оцениваются параметры θ_{td}, а по второй – вычисляется контрольная перплексия. При этом

новые слова, отсутствующие в обучающей коллекции, но попавшие в контрольные документы, игнорируются. Хотя на данный момент существуют работы, утверждающие, что перплексию не стоит применять для оценки качества тематических моделей [24, 69], данная мера по-прежнему широко используется.

Другим традиционным способом оценки качества тематических моделей являются **экспертные оценки** [69]. Экспертам предоставляются полученные темы в виде списков слов и словосочетаний, упорядоченных по убыванию степени принадлежности, и они решают, является ли каждая такая тема в какой-то степени осмысленной и интерпретируемой. Индикатором такой темы служит возможность дать ей некоторое обобщённое название. Таким образом, перед экспертами ставится задача классификации всех предоставленных тем на два класса в зависимости от того, можно ли дать теме некоторое название или нет. В Таблице 1 приведены примеры согласованной темы с названием, данным экспертами, и несогласованной темы, которой невозможно дать никакого названия.

Таблица 1: Примеры согласованной и несогласованной темы

Верхняя часть списка слов из темы	**Название темы**
Быть, человек, люди, год, когда, время, женщина	–
Предприятие, лизинг, имущество, объект, лизинговый	*Лизинг*

Поскольку экспертная оценка тем является дорогостоящей операцией, в последнее время были предприняты попытки предложить способ автоматизации оценки качества тематических моделей, несвязанный с перплексией и коррелирующий с мнениями экспертов. Так, недавно было показано, что можно автоматически оценивать **согласованность тем** с точностью, почти совпадающей с экспертами [63, 69]. Поскольку темы предоставляются экспертам в виде списка первых N слов, согласованность тем оценивает то, насколько данные слова соответствуют рассматриваемой теме. При этом тема считается *согласованной*, если наиболее частые в этой теме слова неслучайно часто совместно

встречаются в документах коллекции. Newman в работе [69] предложил способ вычисления данной меры исходя из меры взаимной информации (**TC-PMI**):

$$TC\text{-}PMI = \frac{1}{|T|}\sum_{t=1}^{|T|}\sum_{j=2}^{10}\sum_{i=1}^{j-1}\log\frac{P(w_i w_j)}{P(w_i)\times P(w_j)} \tag{23}$$

где T – множество полученных тем, $(w_1, w_2, \ldots, w_{10})$ – первые 10 слов в рассматриваемой теме t, $P(w_i)$ и $P(w_j)$ – вероятности слов w_i и w_j соответственно, а $P(w_i w_j)$ – вероятность словосочетания $w_i w_j$.

Существуют разные способы подсчёта вероятностей слов и словосочетаний для вычисления данной меры. Так, в работе [69] для избежания переобучения используется внешний корпус текстов – Википедия. При этом вероятности слов и словосочетаний вычисляются путём деления частотностей в одних и тех же контекстных окнах на число всех контекстных окон. В работе [63] для вычисления меры *TC-PMI* используется тот же самый корпус текстов, а вероятности вычисляются путём деления числа документов, в которых встретилось слово или словосочетание, на число всех документов в коллекции.

Чем выше значение *TC-PMI*, тем лучше согласованы темы. Данная мера показывает высокую корреляцию с оценками экспертов [69]. При этом она рассматривает только первые 10 слов из каждой темы, поскольку они, как правило, предоставляют достаточно информации для формирования предмета темы и отличительных черт одной темы от другой. Согласованность тем становится всё более широко используемым методом оценки качества. Так, в работе [87] также было показано, что данная мера сильно коррелирует с оценками экспертов. А в работе [13] она используется как один из способов оценки качества.

Также существует и другой вариант вычисления данной меры на основе логарифма условной вероятности (*TC-LCP*), предложенный в работе [63]. Он оценивает вероятность менее частотного слова при условии более частотного:

$$TC\text{-}LCP(t) = \sum_{j=2}^{10}\sum_{i=1}^{j-1}\log\frac{P(w_i, w_j)}{P(w_i)} \tag{24}$$

Однако в работе [53] было показано, что этот вариант работает значительно хуже, чем *TC-PMI*.

1.2 Интеграция словосочетаний в тематические модели

Все описанные выше тематические модели работают только со словами, основываясь на модели «мешка слов». Данная модель не учитывает порядок слов и основывается на независимости появлений слов друг от друга в тексте. Это предположение оправдано с точки зрения вычислительной эффективности, но оно далеко от реальности. Так, некоторые слова меняют свой смысл при объединении в словосочетания: например, словосочетание *«точка зрения»* плохо связано со своими компонентами *«точка»* и *«зрение»*.

В последнее время появились работы, в которых исследуется вопрос добавления словосочетаний в тематические модели для улучшения их качества [44, 90, 92]. На текущий момент существуют два основных подхода к решению проблемы: создание единой унифицированной вероятностной тематической модели и предварительное извлечение словосочетаний для последующего добавления.

Большинство существующих работ посвящено именно первому подходу.

1.2.1 Биграммная Тематическая Модель

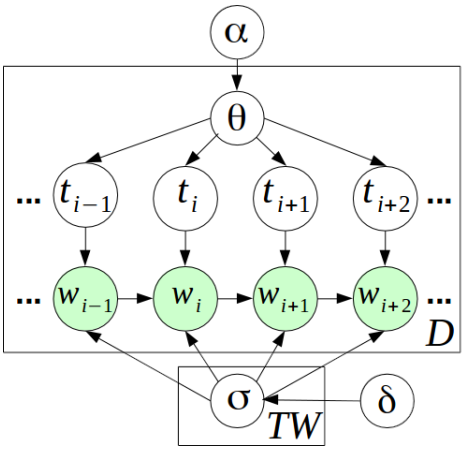
Первая попытка уйти от гипотезы «мешка слов» была предпринята в работе [90], где была предложена *Биграммная Тематическая Модель* (*БТМ*). В БТМ вводится понятие «порядка слов», близкое к иерархической языковой модели Дирихле [59], в которой вероятность появления слова w_i в тексте зависит только от непосредственно предшествующего слова w_{i-1}:

$$P(w_i|w_{i-1}) = \frac{n_{ii-1} + \delta_{w_i}}{n_{i-1} + \delta_0}, \qquad \delta_0 = \sum_i \delta_{w_i} \tag{25}$$

где $\{\delta_{w_i}\}$ – гиперпараметры модели, n_{i-1} – частотность слова w_{i-1}, n_{ii-1} – частотность словосочетания $w_i w_{i-1}$.

Графическая модель БТМ представлена на Рисунке 3.

Рис. 3: Биграммная Тематическая Модель



Процесс порождения текстовой коллекции при этом следующий:

1. Для каждой темы t и для каждого слова w построить распределение σ_{tw} из априорного распределения Дирихле δ;

2. Для каждого документа d построить распределение θ_d из априорного распределения Дирихле α. Затем для каждого слова w_i в документе d:

   (а) Сэмплировать тему t из распределения θ_d;

   (б) Сэмплировать слово w_i из распределения $\sigma_{tw_{i-1}}$.

Стоит отметить, что БТМ работает только со словосочетаниями из двух слов и никак не учитывает в образовании тем отдельные слова. Кроме того, предположение данной модели о том, что появление слова в тексте зависит только от предшествующего слова, по-прежнему далеко от реальности.

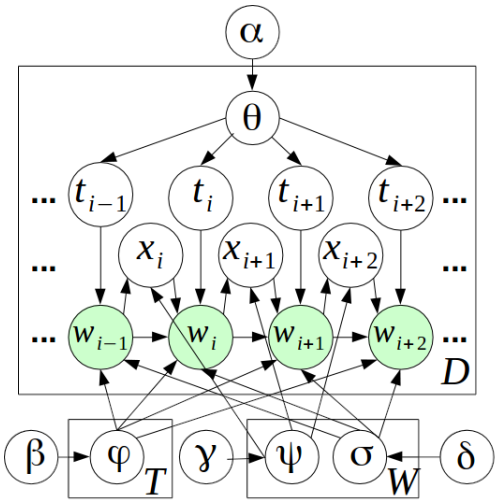
1.2.2 Модель Словосочетаний LDA

Модель Словосочетаний LDA, представленная в работе [44], расширяет Биграммную Тематическую Модель за счёт введения дополнительных переменных x_i, которые для каждого слова в документе указывают, составляет ли

данное слово с предыдущим словосочетание или нет. Если $x_i = 1$, то слово w_i считается частью словосочетания и порождается из распределения, зависящего от предыдущего слова: $P(w_i|w_{i-1}, x_i = 1)$. Если же $x_i = 0$, то слово w_i порождается из распределения $P(w_i|t, x_i = 0)$. При этом значение x_i выбирается в зависимости от предыдущего слова w_{i-1} из распределения $P(x_i|w_{i-1})$.

Графическая модель представлена на Рисунке 4.

Рис. 4: Модель Словосочетаний LDA



Процесс порождения текстовой коллекции при этом следующий:

1. Для каждой темы t построить распределение ϕ_t из априорного распределения Дирихле β;

2. Для каждого слова w построить распределение ψ_w из априорного Бета-распределения γ;

3. Для каждого слова w построить распределение σ_w из априорного распределения Дирихле β;

4. Для каждого документа d построить распределение θ_d из априорного распределения Дирихле α. Затем для каждого слова w_i в документе d:

   (a) Сэмплировать x_i из распределения $\psi_{w_{i-1}}$;

(б) Сэмплировать тему t из распределения θ_d;

(в) Если $x_i = 1$, то сэмплировать слово w_i из распределения $\sigma_{w_{i-1}}$. Иначе сэмплировать w_i из распределения ϕ_t.

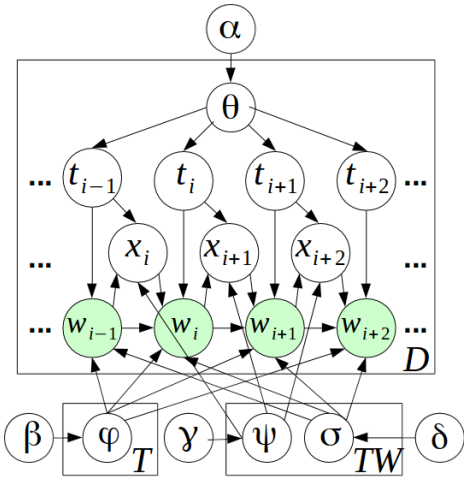
При этом в Модели Словосочетаний LDA темы по-прежнему задаются отдельными словами, а словосочетания – вероятностями перехода от слова к слову вне зависимости от темы.

1.2.3 N-граммная Тематическая Модель

В работе [92] представлена *N-граммная Тематическая Модель*, усложняющая предыдущие модели для предоставления возможности формирования словосочетаний в текстах в зависимости от контекста.

Графическая модель представлена на Рисунке 5.

Рис. 5: N-граммная Тематическая Модель



Процесс порождения текстовой коллекции при этом следующий:

1. Для каждой темы t построить распределение ϕ_t из априорного распределения Дирихле β;

2. Для каждой темы t и каждого слова w построить распределение ψ_{tw} из априорного Бета-распределения γ;

3. Для каждой темы t и каждого слова w построить распределение σ_{zw} из априорного распределения Дирихле δ;

4. Для каждого документа d построить распределение θ_d из априорного распределения Дирихле α. Затем для каждого слова w_i в документе d:

   (а) Сэмплировать x_i из распределения $\psi_{t_{i-1},w_{i-1}}$;

   (б) Сэпмлировать тему t_i из распределения θ_d;

   (в) Если $x_i = 1$, то сэмплировать слово w_i из распределения $\sigma_{t_i w_{i-1}}$. Иначе сэмплировать w_i из распределения ψ_{t_i}.

Стоит отметить, что из N-граммной Тематической Модели можно получить Биграммную Тематическую Модель путём обращения всех переменных x_i в 1 и Модель Словосочетаний LDA путём сужения зависимости σ только от непосредственного предшествующего слова.

1.2.4 Тематическая Модель Слово-Символ

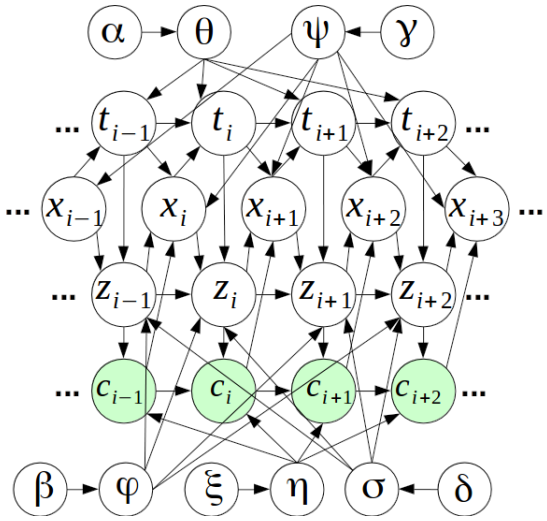
Во всех описанных выше тематических моделях используется предположение о том, что тема каждого словосочетания определяется темой слов, его образующих. В работе [47] предложена Тематическая Модель *Слово-Символ*, уходящая от этого предположения. Данная модель разрабатывалась для китайского языка и представляет собой усложнение N-граммной Тематической Модели. В ней разделяются темы слов и темы символов. При порождении китайского символа вначале выбирается тема для слова, а затем тема для символа.

Графическая модель представлена на Рисунке 6.

В данной модели присутствуют 4 типа переменных: последовательность символов c, последовательность тем символов z, последовательность тем слов t и последовательность переменных-индикаторов x.

Процесс порождения текстовой коллекции при этом следующий:

Рис. 6: Тематическая Модель Слово-Символ



1. Для каждого документа d построить распределение θ_d из априорного распределения Дирихле α;

2. Для каждой темы слов t построить распределение ϕ_t из априорного распределения Дирихле β;

3. Для каждой темы слов t и каждой темы символов z построить распределение σ_{tz} из априорного распределения Дирихле δ;

4. Для каждой темы слов t, каждой темы символов z и каждого символа c построить распределение ψ_{tzc} из априорного Бета-распределения γ;

5. Для каждой темы символов z построить распределение η_z из априорного распределения Дирихле ξ;

6. Для каждого символа c_i в документе d:

   (а) Сэмплировать x_i из распределения $\psi_{t_{i-1}z_{i-1}c_{i-1}}$;

   (б) Если $x_i = 0$, то сэмплировать тему слова t_i из распределения θ_d. Иначе $t_i = z_{t-1}$;

   (в) Если $x_i = 0$, то сэмплировать тему символа z_i из распределения ϕ_{t_i}. Иначе сэмплировать z_i из распределения $\sigma_{t_i z_{i-1}}$;

(г) Сэмплировать символ c_i из распределения η_{z_i}.

Тематическая Модель Слово-Символ разрабатывалась и оказалась наиболее пригодной именно для китайского языка.

1.2.5 Предварительное извлечение словосочетаний

Несмотря на то, что все описанные выше тематические модели, строящие единую унифицированную модель, имеют теоретически элегантное обоснование, у них очень много параметров для настройки, что делает их интересными только с теоретической точки зрения и ограничивает возможность их применения на практике. Так, число параметров у Биграммной Тематической Модели равно W^2T, у Модели словосочетаний LDA – $W^2T + W^2$, у N-граммной Тематической Модели – W^NT, в то время как у LDA – WT, у PLSA – $WT + DT$, где W – размер словаря (т.е. число уникальных слов и словосочетаний в коллекции), D – число документов в текстовой коллекции, T – число тем, N – размер N-грамм.

Ко второму типу методов, добавляющих словосочетания в тематические модели, относится подход, предложенный в работе [53]. На этапе предобработки авторы извлекают все встретившиеся в коллекции текстов словосочетания, после чего упорядочивают их в соответствии с ассоциативной мерой $T\text{-}Score$:

$$T\text{-}Score(xy) = \frac{TF(xy) - \frac{TF(x) \times TF(y)}{|W|}}{\sqrt{TF(xy)}}, \tag{26}$$

где $TF(xy)$ – частотность словосочетания xy, $TF(x)$ и $TF(y)$ – частотности слов x и y соответственно, $|W|$ – число различных слов в коллекции.

После ранжирования извлечённых словосочетаний по данной мере авторы заменяют в документах отдельные слова лучшими словосочетаниями, добавляя их в словарь коллекции (при этом рассматривается 1000 лучших словосочетаний). В работе [53] показано, что стандартная мера качества тематических моделей (перплексия) ухудшается, но улучшается согласованность тем [69].

Таким образом, для улучшения качества тематических моделей актуальным является разработка метода добавления словосочетаний, не усложняющего исходные модели. Такой метод может выявлять более согласованные темы для улучшения их интерпретируемости и может использоваться в других задачах для улучшения качества (в частности, в задаче извлечения терминов).

1.3 Терминологический анализ текстовых коллекций

Задача терминологического анализа текстовых коллекций состоит в анализе специфичной лексики, использующейся в текстах некоторой конкретной предметной области, и автоматическом извлечении терминов с целью создания специализированных терминологических ресурсов: тезаурусов, онтологий и словарей. Традиционно процесс извлечения терминов из текстов коллекции некоторой конкретной предметной области включает в себя несколько этапов.

Вначале осуществляется отбор из текстов коллекции по определённым лингвистическим признакам слов и словосочетаний – кандидатов в термины. Для этой цели проводится морфологический анализ текстов с целью определения начальной формы слов (*леммы*) и их частей речи. Данный процесс необходим, поскольку большинство терминов задаются именными группами (т.е. словосочетаниями, в которых существительное является главным словом) [22].

После извлечения кандидатов в термины вычисляются различные признаки, характеризующие терминологичность слов и словосочетаний, и осуществляется их ранжирование согласно данным признакам. В течение длительного времени исследователи пытались найти наилучший признак, однако ни один из них так и не стал определяющим [77]. На данный момент выделяются следующие группы используемых на практике признаков.

1.3.1 Признаки, основанные на частотности

Основным предположением признаков данной группы является то, что термины, как правило, встречаются в текстовой коллекции гораздо чаще других слов. При описании признаков будут использованы следующие обозначения:

- $TF(w)$ – частотность слова или словосочетания w в коллекции;
- $|W|$ – общее число слов или словосочетаний в коллекции;
- $DF(w)$ – число документов, содержащих слово или словосочетание w;
- $|D|$ – число документов в текстовой коллекции;
- $TF(w|d)$ – частотность слова или словосочетания w в документе d;
- $|W_d|$ – общее число слов в документе d.

Самыми простыми признаками этой группы являются **Частотность** (**TF**) и **Документная частотность** (**DF**). Более сложным признаком является **Term Frequency - Inversed Document Frequency** (**TF-IDF**), поощряющий слова и словосочетания, встречающиеся часто в малом числе документов [86]:

$$TF\text{-}IDF(w) = TF(w) \times \log \frac{|D|}{DF(w)} \tag{27}$$

Также была предложена модификация данного признака – **Term Frequency – Residual Inverse Dcoument Frequency** (**TF-RIDF**) [25], использующая модель Пуассона для предсказания терминологичности, основываясь на том, что распределение терминов в коллекции отличается от распределения Пуассона сильнее, чем распределение остальных слов:

$$TF\text{-}RIDF(w) = TF(w) \times \left(\log \frac{|D|}{DF(w)} - \left(- \log \left(1 - e^{-\frac{TF(w)}{|D|}} \right) \right) \right) \tag{28}$$

Следующим признаком в данной группе является **Domain Consensus** [67], основанный на энтропии. Данная мера поощряет слова или словосочетания, часто встречающиеся в различных документах текстовой коллекции:

$$Domain\ Consensus(w) = -\sum_{d \in D} \left(\frac{TF(w|d)}{|W_d|} \times \log \frac{TF(w|d)}{|W_d|} \right) \tag{29}$$

Последними признаками, рассматриваемыми в данной группе, являются признаки, использующиеся при кластеризации документов. Для успешного решения данной задачи важно уметь оценивать вклад каждого слова или словосочетания при вычислении сходства между документами. В последнее время было придумано несколько новых мер, которые тоже целесообразно использовать для рассматриваемой задачи автоматического извлечения терминов.

Так, в работе [56] был предложен новый признак **Term Contribution** (**TC**), штрафующий общеупотребительные слова и словосочетания, часто встречающиеся в коллекции, но равномерно распределённые по документам:

$$TC(w) = \sum_{d_i, d_j \in D : d_i \neq d_j} \left(TF(w|d_i) \times \log \frac{|D|}{DF(w)} \right) \times \left(TF(w|d_j) \times \log \frac{|D|}{DF(w)} \right) \tag{30}$$

В работе [32] был предложен признак **Term Variance Quality** (**TVQ**). Данная мера штрафует слова и словосочетания, встречающиеся во многих документах коллекции ровно один раз и представляет собой дисперсию над документами, в которых слово или словосочетание встречается хотя бы один раз:

$$TVQ(w) = \sum_{d \in D} TF(w|d)^2 - \frac{1}{|D|} \left(\sum_{d \in D} TF(w|d) \right)^2 \tag{31}$$

Заключительным признаком, относящимся к данной группе, является признак **Term Variance** (**TV**) [56]. Слова и словосочетания, встречающиеся в малом количестве документов или равномерно распределённые по документам в текстовой коллекции, будут иметь низкие значения данной меры:

$$TV(w) = \sum_{d \in D} \left(TF(w|d) - \frac{TF(w)}{|D|} \right) \quad (32)$$

1.3.2 Признаки, использующие контрастную коллекцию

Следующей группой признаков, применяющихся в задаче автоматического извлечения терминов, являются признаки, использующие *контрастную коллекцию*. Под такой коллекцией обычно понимается коллекция более общей предметной области. Хорошими примерами контрастной коллекции являются новостные потоки, национальные корпуса русского[1] и английского[2] языков.

Основная идея данных признаков заключается в том, что частотности терминов в целевой коллекции обычно больше, чем в контрастной. При описании признаков будут дополнительно использоваться следующие обозначения:

- $TF_r(w)$ – частотности слова (словосочетания) w в контрастной коллекции;
- $|W_r|$ – число слов в контрастной коллекции;
- $DF_r(w)$ – число документов контрастной коллекции, содержащих слово или словосочетание w;
- $|D_r|$ – число документов в контрастной коллекции;
- $|C|$ – число всех текстовых коллекций (включая контрастные);
- $|C_w|$ – число коллекций, содержащих слово или словосочетание w.

Базовым признаком данной группы является **Странность** (**Weirdness**) [11], вычисляющаяся как отношение относительных частотностей слова или словосочетания в целевой и контрастной коллекциях:

$$Weirdness(w) = \frac{TF(w)}{|W|} \Big/ \frac{TF_r(w)}{|W_r|} \quad (33)$$

[1] `http://www.ruscorpora.ru`

[2] `http://www.natcorp.ox.ac.uk`

На похожем принципе основан и признак **Relevance** [78]. Низкие значения данной меры будут иметь низкочастотные слова и словосочетания. А высокочастотные же – напротив, будут иметь высокие значения при условии, что они не являются высокочастотными в контрастной коллекции и не встречаются в слишком маленьком числе документов в целевой коллекции:

$$Relevance(w) = 1 - \frac{1}{\log_2\left(2 + \frac{TF(w) \times DF(w)}{TF_r(w)}\right)} \tag{34}$$

В рассматриваемую группу также попадает и мера *TF-IDF*, в которой, в отличие от прошлого раздела, IDF вычисляется по контрастной коллекции (в дальнейшем такая мера будет именоваться как **TF-IDF Reference**).

Помимо этой меры к данной группе относятся и несколько её модификаций. Первой такой модификацией является мера **Contrastive Weight** (**CW**) [16], основывающаяся на наблюдении, что общеупотребительные слова должны быть одинаково распределены в целевой и контрастной коллекциях:

$$CW(w) = \log TF(w) \times \log\left(\frac{|W| + |W_r|}{TF(w) + TF_r(w)}\right) \tag{35}$$

Предложенные исследователями признаки *Domain Tendency* (DT) и *Domain Prevalence* (DP), являющиеся незначительными модификациями *Weirdness* и *Contrastive Weight*, образуют вместе меру **Discriminative Weight** (**DW**) [95], штрафующую слова и словосочетания, часто встречающиеся в целевой коллекции, если они более специфичны в контрастной коллекции:

$$\begin{gathered} DW(w) = DP(w) \times DT(W), \text{ где } DT(w) = \log_2\left(\frac{TF(w) + 1}{TF_r(w) + 1} + 1\right) \text{ и} \\ DP(w) = \log_{10}(TF(w) + 10) \times \log_{10}\left(\frac{|W| + |W_r|}{TF(w) + TF_r(w)}\right) \end{gathered} \tag{36}$$

Следующей модификацией меры TF-IDF является **KF-IDF** [52], поощряющий слова и словосочетания, встречающиеся чаще остальных в целевой коллекции, и редко встречающиеся в контрастных коллекциях:

$$KF\text{-}IDF(w) = DF(w) \times \log\left(\frac{|C|}{|C_w|} + 1\right) \tag{37}$$

Последним признаком в данной группе является **Loglikelihood** [42], поощряющий слова и словосочетания, у которых относительная частотность в целевой коллекции больше, чем в контрастной:

$$Loglikelihood(w) = 2 \times \left(TF(w) \times \log\frac{TF(w)}{TF^{exp}(w)} + TF_r(w) \times \log\frac{TF_r(w)}{TF_r^{exp}(w)}\right), \text{ где}$$
$$TF^{exp}(w) = |W| \times \frac{TF(w) + TF_r(w)}{|W| + |W_r|} \text{ и } TF_r^{exp}(w) = |W_r| \times \frac{TF(w) + TF_r(w)}{|W| + |W_r|} \tag{38}$$

1.3.3 Контекстные признаки

Признаки из данной группы соединяют в себе информацию о частотностях слов и словосочетаний с данными о контекстах их употребления в текстовой коллекции. Под *контекстом* обычно понимается множество слов, с которыми данное слово или словосочетание употребляется в документах. При описании признаков будут дополнительно использоваться следующие обозначения:

- P_w – множество всех фраз, содержащих слово или словосочетание w;
- C_w – множество всех контекстных слов для слова или словосочетания w;
- $|W_c|$ – число контекстных для слова или словосочетания w слов;
- $TF_w(c)$ – частотность слова c в качестве контекстного для слова или словосочетания w;
- $F_{max}(w)$ – максимальная частотность фразы, содержащей слово или словосочетание w;
- P_w^N – множество N самых частотных фраз, содержащих слово или словосочетание w;

- $l_{token}(w)$ и $r_{token}(w)$ – суммы частотностей слов, расположенных в текстах непосредственно слева и справа от слова или словосочетания w;
- $l_{type}(w)$ и $r_{type}(w)$ – число уникальных контекстных слов, расположенных в текстах непосредственно слева и справа от слова или словосочетания w;
- $|w|$ – длина словосочетания w в словах.

Одним из наиболее базовых признаков в этой группе является признак **C-Value**, изначально предложенный для извлечения многословных терминов [12]. Эта мера поощряет длинные кандидаты в термины и штрафует словосочетания, часто встречающиеся в составе объемлющих именных групп:

$$C\text{-}Value(w) = \begin{cases} \log_2 |w| \times \left(TF(w) - \frac{\sum\limits_{p \in P_w} TF(p)}{|P_w|} \right), & \text{есть объемлющие фразы с } w \\ \log_2 |w| \times TF(w), & \text{иначе} \end{cases} \tag{39}$$

Данная мера была впоследствии обобщена и на однословные термины [64]:

$$C\text{-}Value(w) = TF(w) - \frac{\sum\limits_{p \in P_w} TF(p)}{|P_w|} \tag{40}$$

Наиболее известной модификацией *C-Value* является мера **NC-Value** [39], добавляющая контекстную информацию в *C-Value* и определяющая, насколько независимо используется слово или словосочетание в текстовой коллекции:

$$\begin{gathered} NC\text{-}Value(w) = \frac{1}{|W|} \times MC\text{-}Value(w) \times cweight(w), \text{ где} \\ cweight(w) = \sum_{c \in C_w} weight(c) + 1 \text{ и } weight(c) = \frac{1}{2} \left(\frac{|W_c|}{|W|} + \frac{\sum\limits_{e \in W_c} TF(e)}{TF(c)} \right) \end{gathered} \tag{41}$$

Также существует и другой вариант меры **NC-Value** [40]:

$$NC\text{-}Value(w) = 0.8 \times C\text{-}Value(w) + 0.2 \times \sum_{c \in C_w} TF(c) \quad (42)$$

Следующим признаком является **LR** [65], основанный на предположении, что некоторые слова чаще используются в качестве терминологических единиц, и фразы, содержащие такие единицы, скорее окажутся терминами. В работе были предложены два варианта данной меры – **Token-LR** и **Type-LR**:

$$Token\text{-}LR(w) = \sqrt{l_{token}(w) \times r_{token}(w)} \text{ и } Type\text{-}LR(w) = \sqrt{l_{type} \times r_{type}} \quad (43)$$

Поскольку оба варианта признака **LR** рассматривают только контекстные слова и не рассматривают сами кандидаты в термины, в работе [65] был предложен и другой признак для избавления от этого недостатка – **FLR**. Аналогично LR используются два варианта данной меры – **Token-FLR** и **Type-FLR**:

$$Token\text{-}FLR(w) = TF(w) \times Token\text{-}LR(w) \text{ и } Type\text{-}FLR(w) = TF(w) \times Type\text{-}LR(w) \quad (44)$$

Дополнительно в группу контекстных признаков попадают признаки, рассматривающие слова и словосочетания в контексте объемлющих фраз. Первой такой мерой является **Insideness** [5], находящая слова и словосочетания, представляющие собой сокращённые части настоящих терминов:

$$Insideness(w) = \frac{F_{max}(w)}{TF(w)} \quad (45)$$

Также выделяется группа признаков **SumN**, предложенная в работе [5], где N – число наиболее частотных фраз, содержащих рассматриваемое слово или словосочетание. Данные признаки проверяют, насколько слово или словосочетание продуктивно в образовании фраз из предметной области:

$$SumN(w) = \frac{\sum_{p \in P_w^N} TF(p)}{N \times TF(w)} \quad (46)$$

1.3.4 Ассоциативные меры

Признаки из данной группы предназначены только для извлечения многословных терминов (в частности, двусловных). Под *ассоциативными мерами* понимаются математические критерии, определяющие силу связи между составными частями фраз, основываясь на частотах встречаемости отдельных слов и словосочетаний целиком. При описании признаков из данной группы будут дополнительно использоваться следующие обозначения:

- $\overline{x}$ – любое слово, отличное от x;
- $r(w)$ и $l(w)$ – число различных слов, встретившихся непосредственно справа (соответственно слева) от слова w.

Наиболее известной ассоциативной мерой является **Взаимная Информация** (**MI**), показывающая разницу между совместными появлениями в текстах словосочетаний целиком и независимыми появлениями отдельных слов, образующих рассматриваемые словосочетания [26]:

$$MI(xy) = \log \frac{|W| \times TF(xy)}{TF(x) \times TF(y)} \tag{47}$$

Так, если слова x и y, образующие словосочетание xy встречаются в текстах независимо друг от друга, то $MI(xy) = 0$. В то же время, чем больше значение меры MI, тем более достоверной является связь между этими словами.

Мера *Взаимной Информации* рассматривает только случаи совместного появления слов, игнорируя случаи отсутствия одного из слов. Для преодоления данной проблемы была предложена мера **Дополненной Взаимной Информации** (**Augmented MI**) [98], учитывающая число появлений в текстах одного из слов при условии отсутствия другого. Данная мера определяется как отношение вероятности быть словосочетанием к вероятности не быть таковым:

$$Augmented\ MI(xy) = \log \frac{|W| \times TF(xy)}{TF(x\overline{y}) \times TF(\overline{x}y)} \tag{48}$$

Поскольку у *Взаимной Информации* нет верхней границы, были предприняты попытки по её нормализации, в частности для улучшения работы с низкочастотными словами. Так была предложена мера **Нормализованной Взаимной Информации** [21]. В качестве верхней границы выступает случай, когда два слова встречаются в текстах всегда вместе:

$$Normalized\ MI(xy) = \frac{\log \frac{|W| \times TF(xy)}{TF(x) \times TF(y)}}{-\log \frac{TF(xy)}{|W|}} \tag{49}$$

В данной диссертационной работе принята терминология из работы [30], в которой для извлечения терминов предлагается использовать **Настоящую Взаимную Информацию** (**True MI**), отличающуюся от *Взаимной Информации* умножением на частотность словосочетания xy::

$$True\ MI(xy) = TF(xy) \times \log \frac{TF(xy)}{TF(x) \times TF(y)} \tag{50}$$

Также была предложена ещё одна модификация *Взаимной Информации* для лучшего учёта низкочастотных словосочетаний – **Кубическая Взаимная Информация** (**Cubic MI**) [27]:

$$Cubic\ MI(xy) = \log \frac{TF(xy)^3}{|W| \times TF(x) \times TF(y)} \tag{51}$$

Помимо различных вариантов *Взаимной Информации* в данную группу попадают и другие меры. Так, в работе [58] была предложена **Симметричная Условная Вероятность** (**Symmetrical Conditional Probability**), которая проверяет на корреляцию слова x и y, умножая условные вероятности встретить в тексте каждое из слов при условии встречи другого:

$$SCP(xy) = \frac{TF(xy)^2}{TF(x) \times TF(y)} \tag{52}$$

В теории информации для определения степени взаимосвязанности двух признаков также применяется **Коэффициент Сёренсена** (**DC**). Для извлечения терминов его впервые было предложено использовать в работе [84]:

$$DC(x,y) = \frac{2 \times TF(x,y)}{TF(x) + TF(y)} \tag{53}$$

В отличие от *Взаимной Информации Коэффициент Сёренсена* не ранжирует вверх низкочастотные кандидаты [84]. Впоследствии был предложен **Модифицированный Коэффициент Сёренсена** (**Modified DC**) [51], улучшающий качество извлечения двусловных терминов. Для слов, имеющих высокие частоты совместной встречаемости, значения данной меры будут высокими:

$$Modified\ DC(xy) = \log\left(TF(xy)\right) \times \frac{2 \times TF(xy)}{TF(x) + TF(y)} \tag{54}$$

Следующей ассоциативной мерой является **Gravity Count** [29], оценивающая, насколько часто справа от первого слова в словосочетании встречается второе, и наоборот:

$$Gravity\ Count(xy) = \log \frac{TF(xy)r(x)}{TF(x)} + \log \frac{TF(xy)l(y)}{TF(y)} \tag{55}$$

Daille в работе [27] предложила использовать несколько ассоциативных мер из теории информации для задачи извлечения двусловных терминов. Первой такой мерой стал **Простой Коэффициент Соответствия** (**SMC**), подсчитывающий число совместных и раздельных появлений двух слов в текстах:

$$SMC(xy) = \frac{TF(xy) + TF(\overline{x}, \overline{y})}{|W|} \tag{56}$$

Следующей мерой, адаптированной в работе [27] к задаче извлечения терминов, стал **Коэффициент Кульчинского** (**Kulczinsky Coefficient**). Эта мера варьируется от 0 до 1, и в случае, когда одно из слов встречается только с другим, становится больше 0.5:

$$Kulczinsky\ Coefficient(x,y) = \frac{TF(x,y)}{2}\left(\frac{1}{TF(x)} + \frac{1}{TF(y)}\right) \tag{57}$$

Следующим коэффициентом стал **Коэффициент Юла** (**Yule Coefficient**), варьирующийся от -1 до 1. Он равен 0 для независимых слов и +1, если

слова встречаются в текстах всегда рядом друг с другом. Если же слова никогда не встречаются вместе, то данный коэффициент принимает значение -1:

$$YUC(xy) = \frac{TF(xy) \times TF(\overline{xy}) - TF(x\overline{y}) \times TF(\overline{x}y)}{TF(xy) \times TF(\overline{xy}) + TF(x\overline{y}) \times TF(\overline{x}y)} \tag{58}$$

Стоит отметить, что в работе [27] упоминаются ещё несколько ассоциативных мер, но они выводятся из уже рассмотренных выше. Ещё одной мерой является **Коэффициент Жаккара** (**Jaccard Coefficient**) [49]. Применительно к задаче извлечения терминов он вычисляется как отношение совместной встречаемости двух слов к суммарной встречаемости отдельных слов в текстах:

$$Jaccard\ Coefficient(xy) = \frac{TF(xy)}{TF(x\overline{y}) + TF(\overline{x}y)} \tag{59}$$

Следующей ассоциативной мерой является **T-Score**, проверяющей два слова на независимость появлений в текстах:

$$T\text{-}Score(xy) = \frac{TF(xy) - \frac{TF(x) \times TF(y)}{|W|}}{\sqrt{TF(xy)}} \tag{60}$$

Также в задаче извлечения многословных терминов может применяться и критерий **Хи-квадрат** (**Chi-Square**), использующийся для проверки появлений двух слов, образующих словосочетание, на независимость:

$$Chi\text{-}Square(xy) = \frac{\left(TF(xy) - \frac{TF(x) \times TF(y)}{|W|}\right)^2}{TF(x) \times TF(y)} \tag{61}$$

Для проверки гипотезы о том, что слова образуют многословное выражение или нет, используются и параметрические тесты. Одним из широко известных тестов является **Отношение Логарифмического Правдоподобия** (**LogLikelihood Ratio**), адаптированное для извлечения терминов в работе [36]. Данная мера сравнивает максимальные значения двух функций правдоподобия, полученных из двух гипотез – образуют ли слова целое словосочетание или нет:

$$LLR(xy) = 2 \times \left(TF(xy) \times \log \frac{|W| \times TF(xy)}{TF(x) \times TF(y)} + TF(x\overline{y}) \times \log \frac{|W| \times TF(x\overline{y})}{TF(x) \times TF(\overline{y})} + \right.$$
$$\left. TF(\overline{x}y) \times \log \frac{|W| \times TF(\overline{x}y)}{TF(\overline{x}) \times TF(y)} + TF(\overline{xy}) \times \log \frac{|W| \times (TF(\overline{xy})}{TF(\overline{x}) \times TF(\overline{y})} \right) \tag{62}$$

1.3.5 Гибридные признаки

В последнее время стали появляться новые признаки для многословных терминов, объединяющие в себе идеи ассоциативных мер с идеями других категорий признаков для лучшего ранжирования словосочетаний-кандидатов. Так, в проекте **TermExtractor** (**TE**) [82] была предложена мера, заключающаяся в линейном смешивании трёх рассмотренных ранее признаков: *Weirdness*, *Domain Consensus* и модификации *Modified Dice Coefficient – Lexical Cohesion*:

$$TE(xy) = \frac{1}{3} \left(Weirdness(xy) + DomainConsensus(xy) + LexicalCohesion(xy) \right) \tag{63}$$

Также был предложен признак **TC-Value** [89], модифицирующий определение *C-/NC-Value* путём замены частотностей, фигурирующих в определении *C-Value*, на выражение *T-Score* для улучшения ранжирования словосочетаний:

$$TC\text{-}Value(w) = \log_2 |w| \times \left(F(w) - \frac{\sum\limits_{p \in P_w} F(p)}{|P_w|} \right), \text{ где}$$
$$F(a) = \begin{cases} TF(a), \text{ если } \min TS(a) \leq 0 \\ TF(a) \times \log \left(2 + \min TS(a)\right), \text{ если } \min TS(a) > 0 \end{cases} \tag{64}$$

В работе [89] была предложена и гибридная мера **NTC-Value**, являющаяся объединением *TC-Value* и *NC-Value*:

$$NC\text{-}Value(w) = 0.8 \times TC\text{-}Value(w) + 0.2 \times \sum_{c \in C_w} TF(c) \tag{65}$$

1.3.6 Критерии оценки качества систем извлечения терминов

Однако несмотря на обилие разнообразных признаков, ни один из них так и не стал определяющим [77], и фактически из текстов извлекается довольно большой список слов и словосочетаний, являющихся кандидатами в термины, которые затем должны быть подтверждены экспертом по предметной области.

В отличие от ручной проверки извлечённых терминов на практике часто используют так называемые *«золотые стандарты»* – уже существующие, разработанные экспертами терминологические ресурсы (словари, тезаурусы). При этом, если слово или словосочетание присутствует в таком ресурсе, то данный кандидат считается подтверждённым термином. Конечно, здесь присутствует проблема неполноты покрытия предметной области.

При описании мер, использующихся для оценки качества извлечённых терминов, будут использоваться следующие обозначения:

- TP (*True Positive*) – число слов и словосочетаний-кандидатов, правильно классифицированных как термины;
- FP (*False Positive*) – число слов и словосочетаний-кандидатов, ошибочно классифицированных как термины;
- FN (*False Negative*) – число слов и словосочетаний-кандидатов, ошибочно классифицированных как не термины;
- n – число первых слов и словосочетаний-кандидатов в списке извлечённых терминов.

В исследовательских работах для оценки качества извлечённых терминов используются следующие меры:

- **Точность** (**Precision**) [62] – доля кандидатов, правильно классифицированных как термины, среди всего списка кандидатов:

$$Precision = \frac{TP}{TP + FP} \tag{66}$$

Также часто используется её модификация: **Precision@n** – точность на уровне первых n кандидатов, т.е. доля кандидатов среди первых n штук, правильно классифицированных как термины.

- **Полнота** (**Recall**) [62] – доля кандидатов, правильно классифицированных как термины, среди всех терминов из списка кандидатов:

$$Recall = \frac{TP}{TP + FN} \tag{67}$$

Также часто используется и её модификация: **Recall@n** – полнота на уровне первых n кандидатов, т.е. доля кандидатов, правильно классифицированных как термины, среди всех терминов из первых n кандидатов.

- **Средняя точность** (**AvP**) [62] на уровне первых n кандидатов:

$$AvP@n = \frac{1}{m} \sum_{k=1}^{n} \left(r_k \times \left(\frac{1}{k} \sum_{1 \le i \le k} r_i \right) \right), \tag{68}$$

где $r_i = 1$, если i-й кандидат является термином, и $r_i = 0$ в противном случае, а m – число терминов среди кандидатов.

Преимущество последней меры перед остальными заключается в том, что она отражает тот факт, что чем больше терминов находится наверху результирующего списка извлечённых слов и словосочетаний-кандидатов, тем выше значение AvP. Стоит отметить, что основной задачей систем извлечения терминов является построение именно такого списка кандидатов.

1.3.7 Применение методов машинного обучения

Поскольку любая система автоматического извлечения терминов строит в итоге большой список слов и словосочетаний-кандидатов, который впоследствии просматривается экспертом, необходимо минимизировать ручные затраты на эту проверку. Поэтому важно предложить как можно больше признаков,

описывающих термины с различных сторон и провести в ходе извлечения ранжирование кандидатов таким образом, чтобы в начале итогового списка стояли слова и словосочетания, с наибольшей вероятностью являющиеся терминами.

Для поиска наилучшей комбинации различных признаков, используемых для извлечения терминов, в исследовательских работах последних лет привлекаются методы машинного обучения [15, 34, 38, 77, 88, 99], позволяющие изучить комбинирование большого количества признаков.

Первой такой попыткой стала работа [88], где комбинирование нескольких статистических признаков и признаков, полученных с помощью тезауруса EuroWordNet, осуществлялось с помощью алгоритма машинного обучения AdaBoost. В работе показано улучшение качества извлекаемых терминов по сравнению с индивидуальными признаками.

Развивая идеи применения машинного обучения к задаче автоматического извлечения многословных терминов, авторы в работе [15] осуществили комбинирование 13 различных статистических мер с помощью генетического алгоритма обучения с учителем ROGER. В работе также показана устойчивость результатов на двух различных предметных областях (биологии и резюме) и языках (английском и французском).

В работе [77] показано, что комбинация из более, чем 80 статистических признаков, применяемая для извлечения словосочетаний нескольких типов из текстов на чешском языке, даёт суммарный выигрыш в 20% точности по сравнению с результатами извлечения на основе одного наилучшего признака.

В работе [99] проведено сравнение 5 алгоритмов извлечения терминов, способных работать с однословными и многословными терминами и не применяющими никаких порогов по частности. Также авторы предлагают использовать комбинированный метод на основании метода голосования, который хорошо работает на коллекции текстов из Википедии общей тематики и не очень хорошо – на специфичном биологическом корпусе Genia узкой тематики.

В работе [38] применяется система автоматического вывода правил Ripper

для извлечения терминов из шведских патентных корпусов. При этом используются не только статистические, но и лингвистические признаки.

В работе [34] применяется комбинирование многочисленных признаков с помощью логистической регрессии из текстов двух различных предметных областей: широкой области естественных наук и технологий и более узкой банковской области. Применяемые признаки вычисляются из трёх различных источников: из самой целевой текстовой коллекции, из поисковых систем (таких, как Яндекс) и предметно-ориентированного существующего тезауруса.

Стоит заметить, что почти во всех работах рассматривалось только извлечение многословных терминов, поскольку таких терминов оказывается более 85% [66]. Разработанные при этом методы существенно зависят от предметных областей текстов и размеров текстовых коллекций. Так, результаты работы метода [77] зависят от предметной области коллекции и типов извлекаемых словосочетаний. Предложенный в работе [99] алгоритм голосования демонстрирует хорошее качество извлечения на коллекции текстов общей тематики и более низкое – на коллекции узкой тематики. Тем самым важно не только предложить некоторую комбинацию признаков, но и проверить её в других условиях – например, на коллекции другого языка и другой предметной области.

Таким образом, для качественного решения задачи автоматического извлечения терминов актуальным является предложение и комбинирование с помощью машинного обучения новых признаков, позволяющих лучше ранжировать извлекаемые слова и словосочетания-кандидаты – в частности, признаков, основанных на тематиках, обсуждаемых в текстах предметной области, для чего полезно проведение тематического анализа текстовой коллекции.

1.4 Выводы к первой главе

В данной главе проведен обзор методов тематического и терминологического анализа данных. Каждая из данных задач востребована на практике.

Анализ предметной области показал, что традиционные тематические модели базируются на модели «мешка слов», которая далека от реальности и не учитывает порядок слов и их зависимость друг от друга в текстах. В последнее время появились два подхода, позволяющие добавлять словосочетания в тематические модели: создание единой унифицированной тематической модели и предварительное извлечение словосочетаний для последующего добавления.

К достоинствам первых методов относится теоретически элегантное обоснование. Существенным же их недостатком является большое количество параметров, которые сложно настраивать на реальных данных. К достоинствам методов второго типа относится то же самое количество параметров, что и в исходных моделях, а также возможность использования с различными тематическими моделями. К недостаткам же существующих таких методов относится ухудшение основной меры качества тематических моделей – перплексии.

Таким образом, для качественного решения задачи построения тематических моделей необходимо разработать метод, не усложняющий исходные модели и улучшающий все целевые меры качества. Кроме того, такие модели смогут использоваться и в других прикладных задачах для улучшения качества (в частности, в задаче извлечения терминов).

На данный момент для извлечения терминов из текстовых коллекций были предложены различные признаки ранжирования слов и словосочетаний-кандидатов, однако ни один из них так и не стал определяющим. При этом в последнее время для повышения качества всё чаще применяются методы машинного обучения. Поэтому актуальной задачей является разработка и комбинирование с помощью методов машинного обучения новых признаков, позволяющих полнее описать характеристики извлекаемых терминов, в частности, признаков, основанных на тематиках, обсуждаемых в текстах предметной области, для чего полезно применение тематических моделей.

2 Тематические модели: учёт сходства между словами и словосочетаниями

В данной главе предлагается новый алгоритм построения тематических моделей PLSA-SIM, являющийся модификацией алгоритма PLSA. Предлагаемый алгоритм позволяет добавлять словосочетания и учитывать сходство между ними и образующими их словами. Для выбора и последующего включения словосочетаний в тематические модели изучается возможность применения различных ассоциативных мер. Также предлагается новый итеративный алгоритм без учителя PLSA-ITER, позволяющий добавлять наиболее подходящие словосочетания. Помимо автоматического выбора словосочетаний рассматривается возможность интеграции в предложенные методы ручных терминологических ресурсов, разработанных экспертами.

2.1 Модель учёта словосочетаний в определении тематической структуры текстов

Оригинальные тематические модели (*PLSA* и *LDA*) используют модель «мешка слов», не учитывающую порядок слов и предполагающую независимость появлений слов в текстах. Более того, словосочетания обычно тоже добавляются как «чёрные ящики» без связей с остальными словами [53].

Под **словосочетанием** в рамках данной диссертационной работы будет пониматься сочетание двух слов. Процесс добавления словосочетаний в тематические модели при этом обычно следующий. Вначале они добавляются в словарь коллекции, после чего в каждом документе слова, образующие добавляемые словосочетания, заменяются этими словосочетаниями [53]. Таким образом, предположение модели «мешка слов» выполняется.

Данное предположение упрощает выкладки, но далеко от реальности, поскольку в документах есть много слов и словосочетаний, связанных между со-

бой по смыслу: например, словосочетания, содержащие одно общее слово (такие, как *бюджетный – бюджетные расходы – бюджетные доходы – бюджетные средства*). Стоит отметить, что такие словосочетания не только имеют одинаковые слова, но многие из них обладают также семантической и тематической близостью. В то же время у других словосочетаний, содержащих общие слова (например, у идиом), могут быть значительные семантические различия. Таким образом, основная идея предлагаемой модели состоит в том, что если словосочетания, содержащие общие слова, встречаются часто в одних и тех же документах коллекции, то их целесообразно относить к одним и тем же темам [9]:

$$\begin{gathered}\exists d_i \in D : |D \setminus \{d_i\}| << |D|,\ \forall w \in W : w_1, \ldots, w_k \in S_w, \\ n_{d_i w_1} > 0, \ldots, n_{d_i w_k} > 0 \Longrightarrow r(w_1, \ldots, w_k, t),\end{gathered} \tag{69}$$

где $d_i \in D$ – документы, $w \in W$ – слова, $w_1, \ldots, w_k$ – слова и словосочетания, S_w – множество похожих слов и словосочетаний, $n_{d_i w_k}$ – частотность w_k в документе d_i, $r(w_1, \ldots, w_k, t)$ – отношение принадлежности $w_1, \ldots, w_k$ к темам t.

В результате реализации данной модели был предложен новый алгоритм PLSA-SIM, являющийся модификацией исходного алгоритма PLSA [9, 72, 75]. При его описании будут использоваться следующие обозначения [3]:

- D – текстовая коллекция;
- T – множество полученных тем;
- W – словарь коллекции (множество уникальных слов и словосочетаний в коллекции документов D);
- $\Phi = \{\phi_{wt} = P(w|t)\}$ – распределение слов (словосочетаний) w по темам t;
- $\Theta = \{\theta_{td} = P(t|d)\}$ – распределение тем t по документам d;

- $S = \{S_w\}$ – множества похожих слов и словосочетаний, где S_w – множество слов и словосочетаний, похожих на w [9]:

$$S_w = \{w, \bigcup_v wv, \bigcup_v vw\}, \tag{70}$$

где w – лемматизированное слово, wv и vw – лемматизированные словосочетания, содержащие w. В Таблице 2 приведены примеры таких множеств (центральное слово выделено курсивом).

Таблица 2: Множества похожих слов и словосочетаний в алгоритме PLSA-SIM

Множество похожих слов и словосочетаний
Кластеризация; *кластеризация* текст; *кластеризация* документ
Недвижимость; рынок *недвижимость*; строительство *недвижимость*
Мобильный; *мобильный* устройство; *мобильный* телефон

- n_{dw} и n_{ds} – частотности слов (словосочетаний) w и s в документе d;
- n_d – длина документа d (общее число слов (словосочетаний) в документе);
- $\hat{n}_{wt}$ – оценка частотности слова (словосочетания) w в теме t;
- $\hat{n}_{td}$ – оценка частотности темы t в документе d;
- $\hat{n}_t$ – оценка частотности темы t в коллекции документов D;
- $P(t|d,w)$ – условная вероятность отнесения вхождения слова (словосочетания) w в документе d к теме t.

При описании алгоритма PLSA-SIM (см. Алгоритм 2) за основу был взят EM-алгоритм для модели PLSA с внесённым E-шагом внутрь M-шага для избежания хранения трёхмерной матрицы $P(t|d,w)$ [3]. Единственная модификация

касается строчки 7, где в рассмотрение добавляются предварительно вычисленные множества похожих слов и словосочетаний. Тем самым вес подобных слов и словосочетаний увеличивается в каждом документе коллекции.

Алгоритм 2: Алгоритм PLSA-SIM

Вход: коллекция документов D, число тем T, множества похожих слов и словосочетаний S, начальные приближения Φ и Θ

Выход: распределения Φ и Θ

1 **while** *не выполнится критерий остановки* **do**
2 **for** $d \in D$, $w \in W$, $t \in T$ **do**
3 $\hat{n}_{wt} = 0$, $\hat{n}_{td} = 0$, $\hat{n}_t = 0$
4 **for** $d \in D$, $w \in W$ **do**
5 **for** $t \in T$ **do**
6 $P(t|d,w) = \frac{\phi_{wt}\theta_{td}}{\sum\limits_{s \in T} \phi_{ws}\theta_{sd}}$
7 $\hat{n}_{wt}, \hat{n}_{td}, \hat{n}_t += \left(n_{dw} + \sum\limits_{s \in S_w} n_{ds}\right) P(t|d,w)$
8 **for** $d \in D$, $w \in W$ **do**
9 $\phi_{wt} = \frac{\hat{n}_{wt}}{\hat{n}_t}$
10 **for** $d \in D$, $t \in T$ **do**
11 $\theta_{td} = \frac{\hat{n}_{td}}{n_d}$

При этом, если похожие слова и словосочетания встречаются в одних и тех же документах, то алгоритм PLSA-SIM старается их отнести к одним и тем же темам, предполагая, что такие слова и словосочетания обладают тематической близостью. Однако, если же похожие слова и словосочетания не встречаются вместе, исходный алгоритм PLSA не модифицируется. Предполагается, что такие слова и словосочетания обладают семантическими различиями.

Стоит отметить, что алгоритм PLSA-SIM не увеличивает число парамет-

ров оригинального алгоритма – оно остаётся равным $WT+DT$ (см. раздел 1.1.4).

Для предложенного алгоритма также справедлива следующая теорема, отражающая основную идею предложенного алгоритма: похожие слова и словосочетания, часто встречающиеся в рамках одних и тех же документов, имеют максимальную вероятность среди выявленных тем.

Теорема 1. *Пусть имеется коллекция текстов D со словарём W. Пусть u – самое частотное слово в коллекции D:*

$$\forall w \in W \setminus \{u\} : \sum_{d \in D} n_{du} > \sum_{d \in D} n_{dw}$$

Тогда при добавлении любых словосочетаний вида uv_j $(j = 1, \ldots, l)$ в словарь W так, что $n_{du} > \sum\limits_{j=1}^{l} n_{duv_j}$, и построении тематической модели алгоритмом PLSA-SIM с одной темой t будут выполнены следующие неравенства:

$$\forall j = 1, \ldots, l \ \forall w \in W \setminus \{u, v_1, \ldots, v_l\} :$$

$$P(uv_j|t) \geq P(u|t),$$

$$P(uv_j|t) > P(w|t) \ \text{и} \ P(u|t) > P(w|t)$$

Доказательство. Пусть распределения $P(w|t)$ и $P(t|d)$ инициализированы, где $w \in W$ – произвольный элемент словаря, $t \in T$ – выделяемая тема, $d \in D$ – произвольный документ коллекции. Рассмотрим первую итерацию EM-алгоритма:

- *E-шаг.* Рассмотрим произвольный элемент словаря $w \in W$. Учитывая, что выделяется одна тема, получим:

$$P(t|d, w) = \frac{P(w|t)P(t|d)}{\sum\limits_{s} P(w|s)P(s|d)} = \frac{P(w|t)P(t|d)}{P(w|t)P(t|d)} = 1.$$

  Значит, при следующих итерациях никакой модификации на E-шаге производиться не будет, и алгоритм сойдётся за одну итерацию.

- *M-шаг.* Отметим, что имеется одно множество похожих слов и словосочетаний $S = \{u, uv_1, \ldots, uv_l\}$. После добавления словосочетаний uv_j $(j = 1, \ldots, l)$ частотность слова u будет равна $n_{du} - \sum\limits_{j=1}^{l} n_{duv_j}$, частотности слов v_j будет равна $n_{dv_j} - n_{duv_j}$, частотности же остальных слов w_i не изменятся. После модификации частот в алгоритме PLSA-SIM получим:

$$n'_{du} = n_{du} - \sum_{j=1}^{l} n_{duv_j} + \sum_{j=1}^{l} n_{duv_j} = n_{du}$$

$$n'_{duv_j} = n_{duv_j} + n_{du} - \sum_{j=1}^{l} n_{duv_j} + \sum_{k \neq j} n_{duv_k} + n_{dv_j} - n_{duv_j} = n_{du} + n_{dv_j} - n_{duv_j}$$

$$n'_{dv_j} = n_{dv_j} - n_{duv_j} + n_{duv_j} = n_{dv_j}$$

$$n'_{dw_i} = n_{dw_i}$$

Тогда, учитывая, что $\forall w \in W : P(t|d, w) = 1$, получим, что:

$$\sum_{w \in W} \sum_{d \in D} (n_{dw} + \sum_{s \in S} n_{ds}) P(t|d, w) = \sum_{d \in D} (n'_{du} + \sum_{j=1}^{l} (n'_{duv_j} + n'_{dv_j}) + \sum_{i=1}^{m} n'_{dwi}) =$$

$$\sum_{d \in D} (l n_{du} + \sum_{j=1}^{l} (2n_{dv_j} - n_{duv_j}) + \sum_{i=1}^{m} n_{dwi});$$

$$P(u|t) = \frac{\sum\limits_{d \in D} (n_{du} + \sum\limits_{s \in S} n_{ds}) P(t|d, u)}{\sum\limits_{w \in W} \sum\limits_{d \in D} (n_{dw} + \sum\limits_{s \in S} n_{ds}) P(t|d, w)} \propto \sum_{d \in D} n'_{du} \propto \sum_{d \in D} n_{du};$$

$$P(uv_j|t) = \frac{\sum\limits_{d \in D} (n_{duv_j} + \sum\limits_{s \in S} n_{ds}) P(t|d, uv_j)}{\sum\limits_{w \in W} \sum\limits_{d \in D} (n_{dw} + \sum\limits_{s \in S} n_{ds}) P(t|d, w)} \propto \sum_{d \in D} n'_{duv_j} \propto \sum_{d \in D} (n_{du} + n_{dv_j} - n_{duv_j});$$

$$P(v_j|t) = \frac{\sum\limits_{d \in D} (n_{dv_j} + \sum\limits_{s \in S} n_{ds}) P(t|d, v_j)}{\sum\limits_{w \in W} \sum\limits_{d \in D} (n_{dw} + \sum\limits_{s \in S} n_{ds}) P(t|d, w)} \propto \sum_{d \in D} n'_{dv_j} \propto \sum_{d \in D} n_{dv_j};$$

$$P(w_i|t) = \frac{\sum\limits_{d\in D}(n_{dw_i} + \sum\limits_{s\in S} n_{ds})P(t|d,w_i)}{\sum\limits_{w\in W}\sum\limits_{d\in D}(n_{dw} + \sum\limits_{s\in S} n_{ds})P(t|d,w)} \propto \sum_{d\in D} n'_{dw_i} \propto \sum_{d\in D} n_{dw_i}$$

Поскольку $\forall j = 1,\ldots,l : n_{dv_j} \geq n_{duv_j}$ (т.к. частотность словосочетания не может быть больше частотности образующего его слова) и, учитывая, что $\forall w \in W \setminus \{u\} : n_{du} > n_{dw}$, получим, что:

$$P(uv_j|t) = \frac{\sum\limits_{d\in D}(n_{du} + n_{dv_j} - n_{duv_j})}{\sum\limits_{d\in D}(ln_{du} + \sum\limits_{j=1}^{l}(2n_{dv_j} - n_{duv_j}) + \sum\limits_{i=1}^{m} n_{dwi})} \geq$$

$$\geq \frac{\sum\limits_{d\in D} n_{du}}{\sum\limits_{d\in D}(ln_{du} + \sum\limits_{j=1}^{l}(2n_{dv_j} - n_{duv_j}) + \sum\limits_{i=1}^{m} n_{dwi})} = P(u|t) >$$

$$> \frac{\sum\limits_{d\in D} n_{dv_j}}{\sum\limits_{d\in D}(ln_{du} + \sum\limits_{j=1}^{l}(2n_{dv_j} - n_{duv_j}) + \sum\limits_{i=1}^{m} n_{dwi})} = P(v_j|t);$$

$$P(uv_j|t) = \frac{\sum\limits_{d\in D}(n_{du} + n_{dv_j} - n_{duv_j})}{\sum\limits_{d\in D}(ln_{du} + \sum\limits_{j=1}^{l}(2n_{dv_j} - n_{duv_j}) + \sum\limits_{i=1}^{m} n_{dwi})} \geq$$

$$\geq \frac{\sum\limits_{d\in D} n_{du}}{\sum\limits_{d\in D}(ln_{du} + \sum\limits_{j=1}^{l}(2n_{dv_j} - n_{duv_j}) + \sum\limits_{i=1}^{m} n_{dwi})} = P(u|t) >$$

$$> \frac{\sum\limits_{d\in D} n_{dw_i}}{\sum\limits_{d\in D}(ln_{du} + \sum\limits_{j=1}^{l}(2n_{dv_j} - n_{duv_j}) + \sum\limits_{i=1}^{m} n_{dwi})} = P(w_i|t).$$

Полученные неравенства доказывают данную теорему.

□

Замечание 1. Доказанная теорема не накладывает никаких ограничений на встречаемость тех или иных слов или словосочетаний в документах – их набор в каждом конкретном документе может быть произвольным.

Замечание 2. В последнее время появились работы, в которых предлагается использовать различные регуляризаторы, позволяющие получать темы с нужными характеристиками (разреженные, лучше согласованные и т.д.) [3, 68]. Общий подход заключается во введении дополнительных регуляризаторов $R_i(\Phi, \Theta)$ $(i = 1, \ldots, n)$ и в максимизации линейной комбинации правдоподобия L и регуляризаторов R_i с коэффициентами регуляризации τ_i:

$$R(\Phi, \Theta) = \sum_{i=1}^{n} \tau_i R_i(\Phi, \Theta), \qquad L(\Phi, \Theta) + R(\Phi, \Theta) \to \max_{\Phi, \Theta} \tag{71}$$

Решение данной задачи приводит к следующим формулам М-шага [3]:

$$\phi_{wt} = \frac{(\hat{n}_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}}(\Phi, \Theta))_+}{\sum\limits_{u \in W} (\hat{n}_{ut} + \phi_{ut} \frac{\partial R}{\partial \phi_{ut}}(\Phi, \Theta))_+}, \qquad \theta_{td} = \frac{(\hat{n}_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}}(\Phi, \Theta))_+}{\sum\limits_{s \in T} (\hat{n}_{sd} + \theta_{sd} \frac{\partial R}{\partial \theta_{sd}}(\Phi, \Theta))_+}, \tag{72}$$

Однако данные формулы применяются ко всем словам и словосочетаниям сразу и не учитывают ситуации, когда словосочетания, содержащие общие слова, не встречаются в рамках одних и тех же документов и имеют значительные семантические различия (например, идиомы). Поэтому данный подход не сводится к предложенному в рамках данной диссертационной работы.

2.2 Итеративная модель учёта словосочетаний в определении тематической структуры текстов

Основной идеей итеративной модели учёта словосочетаний в определении тематической структуры текстов является то, что для добавления в тематические модели наиболее подходящие словосочетания выбираются исходя из вида верхушек списков слов, образующих эти темы [9]. Для этой цели в каждой те-

ме из первых слов составляются все возможные словосочетания, которые затем добавляются в тематическую модель при обучении:

$$W = \{u_1, \ldots, u_n, u_{1t_1}u_{2t_1}, \ldots, u_{it_k}u_{jt_k}\}, \tag{73}$$

где W – словарь коллекции, u_i – слова, $u_{it_k}u_{jt_k}$ – добавляемые в модель словосочетания, составляемые из верхушек темы t_k.

Так, если в верхней части некоторой темы окажутся слова *«ценный»* и *«бумага»*, то в тематическую модель добавляется лемматизированное словосочетание *«ценный бумага»*. Для реализации данной модели был предложен новый итеративный алгоритм выбора словосочетаний *PLSA-ITER* [7, 9, 70].

При описании предлагаемого алгоритма будут использоваться следующие дополнительные обозначения:

- B – множество всех словосочетаний в коллекции документов D;
- B_A – множество всех словосочетаний, добавленных в модель;
- B_i – множество словосочетаний, добавленных в модель на i-й итерации;
- S_i – множество потенциальных кандидатов на похожие слова и словосочетания на i-й итерации;
- $(u_1, u_2, \ldots, u_{10})$ – первые 10 слов в теме t;
- $TF(u_iu_j)$ – частотность словосочетания u_iu_j.

На каждой итерации алгоритм PLSA-ITER (см. Алгоритм 3) добавляет в множество кандидатов в похожие слова и словосочетания первые 10 слов из каждой темы. Также в это множество и в саму модель добавляются все словосочетания, образующиеся с помощью этих слов. При этом анализируются только первые 10 слов в темах, поскольку одной из целевых мер качества является согласованность тем, использующая именно это множество (см. раздел 1.1.6).

Алгоритм 3: Итеративный алгоритм PLSA-ITER

Вход: коллекция документов D, число тем $|T|$, множество словосочетаний B

Выход: полученные темы T

Запуск оригинального алгоритма PLSA на коллекции документов D
$B_A = \emptyset$
while *не выполнится критерий остановки* **do**
    $S_i = \emptyset$, $B_i = \emptyset$
    for $t \in T$ **do**
        $S_i = S_i \bigcup \{u_1, u_2, \ldots, u_{10}\}$
        for $u_i, u_j \in (u_1, u_2, \ldots, u_{10})$ **do**
            if $u_i u_j \in B$ *и* $u_j u_i \in B$ *и* $TF(u_i u_j) > TF(u_j u_i)$ **then**
                $B_i = B_i \bigcup \{u_i u_j\}$
    Создание множества похожих слов и словосочетаний S из $S_i \bigcup B_i$
    $B_A = B_A \bigcup B_i$
    Запуск PLSA-SIM с множествами S и B_A

Стоит отметить, что алгоритм PLSA-ITER не увеличивает число параметров оригинального алгоритма – оно остаётся равным $WT+DT$ (см. раздел 1.1.4).

2.3 Уровень согласия между экспертами

Как следует из раздела 1.1.6, одним из критериев оценки качества тематических моделей являются экспертные оценки, получаемые при классификации тем на 2 класса в зависимости от того, можно ли той или иной теме дать некоторое обобщённое название (класс «+») или нет (класс «–»). Для получения таких оценок были приглашены двое экспертов. Для определения уровня согласия между экспертами используется коэффициент Каппа [85]:

$$\kappa = \frac{P(a) - P(e)}{1 - P(e)}, \tag{74}$$

где $P(a)$ – относительное наблюдаемое согласие между экспертами, а $P(e)$ – вероятность случайного согласия между экспертами.

При составленной таблице сопряжённости (см. Таблица 3) указанные вероятности вычисляются следующим образом:

		Эксперт 2	
		Число +	**Число –**
Эксперт 1	**Число +**	a	b
	Число –	c	d

Таблица 3: Таблица сопряжённости для вычисления коэффициента Каппа

$$\begin{aligned} P(a) &= \frac{a+b}{a+b+c+d}, \\ P(e) &= \frac{a+b}{a+b+c+d} \times \frac{a+c}{a+b+c+d} + \frac{c+d}{a+b+c+d} \times \frac{b+d}{a+b+c+d} \end{aligned} \tag{75}$$

Известно, что, если оба эксперта отнесли все объекты к одному и тому же классу, значение коэффициента Каппа оказывается неопределённым, поскольку в этом случае $P(e) = 1$ [85]. Такие случаи исключались из рассмотрения.

2.4 Текстовые коллекции и предобработка

Для тестирования предлагаемых алгоритмов использовались текстовые коллекции различных языков и предметных областей:

- Для английской части были выбраны:
  - Английская часть корпуса параллельных текстов Europarl, составленная из речей заседаний Европарламента[3] и содержащая 54 млн.

[3] http://www.statmt.org/europarl

слов в 9672 документах;

  - Английская часть корпуса параллельных текстов JRC-Acquiz, представляющая собой статьи из законодательства Евросоюза за период с 1950 по 2005 годы[4] и содержащая 45 млн. слов в 23545 документах.
  - Архив исследовательских работ по компьютерной лингвистике ACL Anthology[5], содержащий 42 млн. слов в 10921 документе.

- Для русской части была взята подборка статей из электронных банковских журналов (таких, как Аудитор, РБК и другие), содержащая 18.5 млн. слов в 10422 документах.

На стадии предобработки проводился морфологический анализ документов. Для английских корпусов использовались средства Stanford Core NLP[6], для русского – собственный морфологический анализатор. Все слова были лемматизированы, т.е. приведены к начальной форме. В качестве слов, образующих темы, рассматривались *существительные*, *прилагательные*, *наречия* и *глаголы*, поскольку другие слова не играют особой роли в данном процессе. Также слова, встретившиеся в коллекции менее 5 раз, исключались из рассмотрения. Кроме того, были извлечены все встретившиеся словосочетания в формах:

- *Существительное + Существительное*, *Прилагательное + Существительное*, *Существительное + предлог of + существительное* – для английских коллекций;

- *Существительное + Существительное в родительном падеже*, *Прилагательное + Существительное* – для русской коллекции.

Рассматривались только такие словосочетания, поскольку темы, как правило, образуются с помощью существительных и именных групп [92].

[4] http://ipsc.jrc.ec.europa.eu/index.php?id=198

[5] http://acl-arc.comp.nus.edu.sg

[6] http://nlp.stanford.edu/software/corenlp.shtml

2.5 Интеграция словосочетаний с помощью алгоритма PLSA-SIM

На первом этапе тестирования сравнивались результаты работы предложенного алгоритма PLSA-SIM с оригинальным алгоритмом PLSA. Для этой цели были извлечены все словосочетания, встретившиеся в коллекции не менее 5 раз. Для последующего упорядочивания словосочетаний применялись все **16** ассоциативных мер, описанные в разделе 1.3.4: *Взаимная Информация (MI), Дополненная MI, Нормализованная MI, Настоящая MI, Кубическая MI, Коэффициент Сёренсена (DC), Модифицированный DC, Симметричная Условная Вероятность, Коэффициент Простого Соответствия, Коэффициент Кульчинского, Коэффициент Юла, Коэффициент Жаккара, Хи-Квадрат, Отношение Логарифмического Правдоподобия, T-Score* и *Gravity Count*. В качестве простой меры упорядочивания словосочетаний была выбрана *частотность* (*TF*).

В соответствии с результатами, представленными в работе [53], в тематические модели добавлялись 1000 лучших словосочетаний для каждой меры. Стоит отметить, что при тестировании число тем фиксировалось равным 100.

В качестве критериев оценки полученных тем использовались *Перплексия* и *TC-PMI*, описанные в разделе 1.1.6. Как следует из доказанной выше теоремы, в предлагаемом алгоритме похожие слова и словосочетания с большой вероятностью окажутся среди первых в полученных темах. Тем самым происходит неявная максимизация меры *TC-PMI*, поскольку такие слова и словосочетания склонны встречаться в одних и тех же документах. Поэтому данная мера была модифицирована для учёта первых 10 непохожих слов в каждой теме (в дальнейшем такая мера будет обозначаться как **TC-PMI-nSIM**).

При тестировании используются все **17** описанных выше мер упорядочивания словосочетаний на всех текстовых коллекциях с целью сравнить качество следующих трёх алгоритмов:

1. Оригинальный алгоритм PLSA;

2. Алгоритм PLSA с 1000 лучших словосочетаний как «чёрные ящики»;

3. Алгоритм PLSA-SIM с 1000 лучших словосочетаний.

В соответствии с результатами тестирования все рассматриваемые меры распределились по двум группам:

1. В первую группу попали **11** ассоциативных мер: *Взаимная Информация (MI), Дополненная MI, Нормализованная MI, Коэффициент Сёренсена, Симметричная Условная Вероятность, Коэффициент Простого Соответствия, Коэффициент Кульчинского, Коэффициент Юла, Коэффициент Жаккара, Хи-квадрат* и *Отношение Логарифмического Правдоподобия*. При добавлении лучших словосочетаний по любой из данных мер любым рассматриваемым способом все меры качества остаются на том же самом уровне, что и в случае оригинального алгоритма. Это объясняется тем, что меры данной группы упорядочивают вверх специфичные низкочастотные словосочетания, не влияющие на работу тематических моделей. В Таблице 4 представлены результаты добавления 1000 лучших по мере MI словосочетаний (как самой широко известной в группе).

2. Во вторую группу мер попали **6** мер: *Частотность (TF), Кубическая MI, Настоящая MI, Модифицированный DC, T-Score* и *Gravity Count*. При добавлении лучших словосочетаний по любой из данных мер как «чёрные ящики» перплексия ухудшается, но улучшается согласованность тем. Стоит отметить, что в работе [53] авторы в качестве меры упорядочивания словосочетаний рассматривали *T-Score* из этой группы. Таким образом, представленные выше результаты согласуются с результатами, описанными в работе [53]. Однако при добавлении 1000 лучших словосочетаний по любой из мер данной группы в предложенный алгоритм PLSA-SIM значительно улучшаются все меры качества. Это объясняется тем, что меры

Таблица 4: Результаты добавления 1000 лучших по мере MI словосочетаний в тематические модели

Корпус	**Модель**	**Перплексия**	**TC-PMI**	**TC-PMI-nSIM**
Банковский	*PLSA*	1724.2	86.1	86.1
	PLSA + словосочетания	1714.1	84.2	84.2
	PLSA-SIM + словосочетания	1715.4	84.1	84.1
Europarl	*PLSA*	1594.3	53.2	53.2
	PLSA + словосочетания	1584.6	55	55
	PLSA-SIM + словосочетания	1591.3	55.2	55.2
JRC-Acquiz	*PLSA*	812.1	67	67
	PLSA + словосочетания	815.4	66.3	66.3
	PLSA-SIM + словосочетания	815.6	66.4	66.4
ACL Anthology	*PLSA*	2134.7	74.8	74.8
	PLSA + словосочетания	2138.1	75.5	75.5
	PLSA-SIM + словосочетания	2144.8	75.8	75.8

данной группы упорядочивают вверх частотные типичные словосочетания. В Таблице 5 представлены результаты добавления 1000 лучших по мере *TF* словосочетаний (как самой широко известной в группе).

Таблица 5: Результаты добавления 1000 самых частотных словосочетаний в тематические модели

Корпус	**Модель**	**Перплексия**	**TC-PMI**	**TC-PMI-nSIM**
Банковский	*PLSA*	1724.2	86.1	86.1
	PLSA + словосочетания	2251.8	98.8	98.8
	PLSA-SIM + словосочетания	**1450.6**	**156.5**	**102.6**
Europarl	*PLSA*	1594.3	53.2	53.2
	PLSA + словосочетания	1993.5	57.3	57.3
	PLSA-SIM + словосочетания	**1431.6**	**127.7**	**84.7**
JRC-Acquiz	*PLSA*	812.1	67	67
	PLSA + словосочетания	1038.9	72	72
	PLSA-SIM + словосочетания	**743.7**	**108.4**	**76.9**
ACL Anthology	*PLSA*	2134.7	74.8	74.8
	PLSA + словосочетания	2619.3	73.7	73.7
	PLSA-SIM + словосочетания	**1806.4**	**152.7**	**87.8**

Таким образом, при добавлении 1000 лучших словосочетаний, упорядоченных по любой из мер из второй группы, в предложенный алгоритм PLSA-SIM значительно улучшается качество тематических моделей по всем целевым мерам независимо от языка и предметной области.

Помимо автоматических оценок качества тематических моделей использовались также и экспертные. Экспертам-лингвистам были выданы темы, полученные по одним и тем же коллекциям с помощью следующих трёх алгоритмов:

1. Оригинальный алгоритм PLSA;
2. Алгоритм PLSA с добавленными как «чёрные ящики» 1000 самых частотных словосочетаний;
3. Алгоритм PLSA-SIM с 1000 самых частотных словосочетаний.

Перед экспертами была поставлена задача классификации тем на 2 класса в зависимости от того, можно ли той или иной теме дать некоторое обобщённое название (класс «+») или нет (класс «–»). В Таблице 6 представлены результаты экспертных оценок для всех текстовых коллекций, кроме архива работ ACL Anthology, поскольку для правильной разметки тем в этой коллекции требуется наличие специальных знаний в области компьютерной лингвистики.

Для определения уровня согласия между экспертами был посчитан коэффициент Каппа (см. раздел 2.3), усреднённый для каждой из коллекций по сравниваемым моделям. При этом случаи, когда оба эксперта относили все объекты к одному классу, были исключены из рассмотрения. Полученные значения коэффициента Каппа представлены в Таблице 7.

Как видно из Таблицы 6 при добавлении 1000 самых частотных словосочетаний в алгоритм PLSA-SIM число тем, которым может быть выдано некоторое обобщённое название, увеличивается по сравнению с оригинальным алгоритмом PLSA для всех текстовых коллекций. Тем самым повышается интерпретируемость тем экспертами. Также следует отметить, что добавление словосочетаний как «чёрные ящики» не увеличивает число таких тем. Данный результат также подтверждает то, что предложенный алгоритм улучшает качество тематических моделей независимо от языка и предметной области.

Таблица 6: Экспертная оценка тем для алгоритмов PLSA и PLSA-SIM

Корпус	**Модель**	**Эксперт 1**		**Эксперт 2**	
		Число +	**Число −**	**Число +**	**Число −**
Банковский	*PLSA*	93	7	92	8
	PLSA + словосочетания	92	8	95	5
	PLSA-SIM + словосочетания	**96**	**4**	**97**	**3**
Europarl	*PLSA*	98	2	99	1
	PLSA + словосочетания	96	4	99	1
	PLSA-SIM + словосочетания	**100**	**0**	**100**	**0**
JRC-Acquiz	*PLSA*	92	8	91	9
	PLSA + словосочетания	94	6	97	3
	PLSA-SIM + словосочетания	**99**	**1**	**100**	**0**

Таблица 7: Коэффициент Каппа при тестировании алгоритма PLSA-SIM

Корпус	**Коэффициент Каппа, κ**
Банковский	59.8%
Europarl	60.0%
JRC-Acquiz	60.2%

В Таблице 8 представлены первые 10 слов и словосочетаний из одной случайно выбранной темы из трёх текстовых корпусов для оригинального ал-

горитма PLSA и предложенного алгоритма PLSA-SIM с добавленными 1000 самых частотных словосочетаний. В рамках одной и той же текстовой коллекции представлены темы с одинаковыми названиями, данными обоими экспертами.

Таблица 8: Первые 10 слов и словосочетаний из тем, полученных алгоритмами PLSA и PLSA-SIM с добавленными 1000 самых частотных словосочетаний

Банковский		**Europarl**		**JRC-Acquiz**	
PLSA	*PLSA-SIM*	*PLSA*	*PLSA-SIM*	*PLSA*	*PLSA-SIM*
Бумага	Ценный бумага	Financial	Economic crisis	Animal	Animal
Ценный	Бумага	Crisis	Financial crisis	Bovine	Bovine animal
Облигация	Облигация	Have	European economy	Have	Meat
Выпуск	Выпуск облигация	European	Time of crisis	Slaughter	Animal health
Рынок	Сделка	Market	Crisis	Health	Have
Акция	Выпуск	Need	Current crisis	Disease	Number of animal
Эмитент	Сделка РЕПО	Regulation	Economic recovery	State	Bovine
Размещение	Эмитент	System	European project	Member	Meat product
Эмиссия	РЕПО	Supervision	Financial market	Veterinary	Test
Обращение	Вторичный рынок	Agency	Financial	Embryo	Slaughter

Список первых 10 слов и словосочетаний из всех тем, полученных на бан-

ковском корпусе алгоритмами PLSA и PLSA-SIM с добавлением 1000 самых частотных словосочетаний приведён в приложениях А и Б соответственно.

2.6 Интеграция словосочетаний с помощью алгоритма PLSA-ITER

Предложенный алгоритм PLSA-ITER был также апробирован на описанных выше текстовых коллекциях. В Таблице 9 представлены результаты работы алгоритма PLSA-ITER после первой итерации и результаты работы алгоритма PLSA-SIM с добавленными в него 1000 самых частотных словосочетаний.

Таблица 9: Результаты работы алгоритма PLSA-ITER после первой итерации и алгоритма PLSA-SIM с добавленными 1000 самых частотных словосочетаний

Коллекция	**Модель**	**Перплексия**	**TC-PMI**	**TC-PMI-nSIM**
Банковская	*PLSA-SIM*	1450.6	156.5	102.6
	PLSA-ITER	1499.6	138.3	97.3
Europarl	*PLSA-SIM*	1431.6	127.7	84.7
	PLSA-ITER	1303.6	98.6	59.3
JRC-Acquiz	*PLSA-SIM*	743.7	108.4	76.9
	PLSA-ITER	786.9	92.5	68.2
ACL	*PLSA-SIM*	1806.4	152.7	87.8
	PLSA-ITER	1949.4	119.6	77.1

Как видно из результатов, представленных в Таблице 9, заметно ухудшение мер качества в результате работы итеративного алгоритма. Это связано с тем, что кандидаты в похожие слова и словосочетания отбираются очень тщательно, и в результате таких множеств образуется очень мало. Для нахождения большего числа похожих слов и словосочетаний из образующихся кандидатов использовались *стеммеры*, т.е. алгоритмы, пытающиеся найти основы слов:

- Для английской части исследования были выбраны два стеммера:
  - Широко известный стеммер Портера [80]. Данный алгоритм, применяя последовательно ряд правил, отсекает окончания и суффиксы, основываясь на особенностях английского языка. Так, стеммер Портера приведёт к одинаковой основе *«fish»* слова *«fishing»* и *«fish»*.
  - Более жадный стеммер Ланкастерского университета [76]. Данный алгоритм отличается от стеммера Портера тем, что применяющиеся правила более жадно отсекают окончания и суффиксы. Так, данный стеммер в отличие от стеммера Портера приведёт слова *«Europe»* и *«European»* к общей основе *«europ»*.

- Для русской части исследования был выбран единственный широко известный стеммер Snowball[7]. Данный алгоритм представляет собой модификацию стеммера Портера для русского языка. Так, стеммер Snowball приведёт к одинаковой основе *«тайн»* слова *«тайна»* и *«тайный»*.

Для того чтобы учесть стеммеры в итеративном алгоритме, было модифицировано определение множеств похожих слов и словосочетаний $S = \{S_w\}$:

$$S_w = \{w,\ \bigcup_u u,\ \bigcup_{u,v} uv : stem(u) = stem(w) \text{ или } stem(v) = stem(w)\} \qquad (76)$$

где w и u – лемматизированные слова, uv – лемматизированное словосочетание, $stem(u)$ – основа слова u, получающаяся в результате работы стеммера. В Таблице 10 приведены примеры множеств похожих слов и словосочетаний по тому или иному стеммеру (центральные слова выделены курсивом).

В Таблице 11 представлены результаты работы итеративного алгоритма PLSA-ITER со стеммерами после первой итерации и результаты работы алгоритма PLSA-SIM с добавленными в него 1000 самых частотных словосочетаний.

[7] `http://snowball.tartarus.org/algorithms/russian/stemmer.html`

Таблица 10: Примеры множеств похожих слов и словосочетаний, получающихся по разным стеммерам

Стеммер	**Множество похожих слов и словосочетаний**
Snowball	*Тайна*, банковский *тайна*, *тайный*
	Право, *право* собственность, *правый*, *правый* сторона
Портер	*Fish*, *fish* agreement, *fishing*, *fishing* agreement
	Alcohol, use of *alcohol*, *alcoholic*, *alcoholic* product
Ланкастер	*Budget*, *budget* year, *budgetary*, *budgetary* year
	Culture, european *culture*, *cultural*, *cultural* Europe

Как видно из результатов, представленных в Таблице 11, использование стеммера Snowball для русского языка и стеммера Ланкастерского университета для английского языка в алгоритме PLSA-ITER приводит к дальнейшему улучшению качества тематических моделей по перплексии с незначительным падением уровня согласованности тем.

Замечание. Стоит заметить, что в алгоритме PLSA-ITER зафиксирован параметр – число первых слов в темах, из которых строятся добавляемые словосочетания и множества похожих слов и словосочетаний, равное 10. Было также проведено тестирование, варьирующее данный параметр. В результате было установлено, что значения согласованности тем *TC-PMI* и *TC-PMI-nSIM* не зависят от данного параметра, в то время как с точки зрения перплексии оптимальное значение находится на уровне 5-7 первых слов в темах.

Также были получены экспертные оценки тем, выявленных первой итерацией алгоритма PLSA-ITER со стеммером Портера для английского языка и стеммером Snowball – для русского. В Таблице 12 приведены результаты экспертных оценок для всех коллекций, кроме архива работ ACL Anthology, поскольку для правильной разметки тем в этой коллекции требуются специальные знания в области компьютерной лингвистики (для сравнения приведены

Таблица 11: Результаты работы алгоритмов PLSA-ITER со стеммерами после первой итерации и PLSA-SIM с 1000 самых частотных словосочетаний

Коллекция	**Модель**	**Перплексия**	**TC-PMI**	**TC-PMI-nSIM**
Банковская	*PLSA-SIM*	1450.6	156.5	102.6
	PLSA-ITER + Snowball	**1265.1**	137.6	96.7
Europarl	*PLSA-SIM*	1431.6	127.7	84.7
	PLSA-ITER + Портер	1293.8	99.6	61.2
	PLSA-ITER + Ланкастер	**1077.7**	105	55.2
JRC-Acquiz	*PLSA-SIM*	743.7	108.4	76.9
	PLSA-ITER + Портер	777.7	90.8	68.2
	PLSA-ITER + Ланкастер	**736.5**	94.5	68.6
ACL	*PLSA-SIM*	1806.4	152.7	87.8
	PLSA-ITER + Портер	1853.7	123.6	76.2
	PLSA-ITER + Ланкастер	**1772.1**	121.3	76.5

результаты алгоритма PLSA-SIM с 1000 самых частотных словосочетаний).

Для определения уровня согласия между экспертами был посчитан коэффициент Каппа (см. раздел 2.3), усреднённый для каждой из коллекций по сравниваемым моделям. При этом случаи, когда оба эксперта относили все объекты к одному классу, были исключены из рассмотрения. Полученные значения коэффициента Каппа представлены в Таблице 13. Стоит отметить, что для кол-

Таблица 12: Экспертная оценка тем для алгоритмов PLSA-ITER и PLSA-SIM

Корпус	**Модель**	**Эксперт 1**		**Эксперт 2**	
		Число +	**Число −**	**Число +**	**Число −**
Банковский	*PLSA-SIM*	96	4	97	3
	PLSA-ITER	96	4	97	3
Europarl	*PLSA-SIM*	100	0	100	0
	PLSA-ITER	100	0	100	0
JRC-Acquiz	*PLSA-SIM*	99	1	100	0
	PLSA-ITER	96	4	99	1

лекции Europarl посчитать данный коэффициент невозможно в силу того, что в обеих сравниваемых моделях оба эксперта дали названия всем темам.

Таблица 13: Коэффициент Каппа при тестировании алгоритма PLSA-ITER

Корпус	**Коэффициент Каппа κ**
Банковский	85.2%
Europarl	–
JRC-Acquiz	62.0%

Как видно из Таблицы 12 в результате работы алгоритма PLSA-ITER число тем, которым можно дать название, почти не изменяется по сравнению с алгоритмом PLSA-SIM с добавленными 1000 самых частотных словосочетаний. Тем самым интерпретируемость тем экспертами остаётся на высоком уровне.

В Таблице 14 представлены результаты первых итераций итеративного алгоритма PLSA-ITER со стеммерами Snowball и Ланкастерского университета вместе с результатами оригинального алгоритма PLSA.

Как видно из результатов, представленных в Таблице 14, после первой итерации наблюдается существенное улучшение качества полученных тем. Сто-

Таблица 14: Результаты первых итераций алгоритма PLSA-ITER

Коллекция	**Итерация**	**Перплексия**	**TC-PMI**	**TC-PMI-nSIM**
Банковская	*0 (PLSA)*	1724.2	86.1	86.1
	1	1265.1	137.6	96.7
	2	1257.1	133.5	95
	3	1259.8	134.5	95.7
Europarl	*0 (PLSA)*	1594.3	53.2	53.2
	1	1077.7	105	55.2
	2	1210.8	92.1	55.2
	3	1242.9	80.1	53.2
JRC-Acquiz	*0 (PLSA)*	812.1	67	67
	1	736.5	94.5	68.6
	2	751.9	94.9	67
	3	749.6	99.5	67.7
ACL	*0 (PLSA)*	2134.7	74.8	74.8
	1	1772.1	121.3	76.5
	2	1775.5	139.3	81
	3	1767.6	144.6	83

ит заметить, что на следующих итерациях результаты начинают колебаться вокруг примерно тех же самых уровней перплексии и согласованности тем.

В Таблице 15 представлены первые 10 слов и словосочетаний из двух случайно выбранных тем рассматриваемых коллекций для алгоритма PLSA-ITER со стеммерами Ланкастерского университета и Snowball после первой итерации.

Стоит отметить, что алгоритмы PLSA-SIM и PLSA-ITER были апробированы и на основе алгоритма LDA, и результаты оказались аналогичными.

Таблица 15: Первые 10 слов и словосочетаний из тем, полученных алгоритмом PLSA-ITER после первой итерации

Банковский		**Europarl**		**JRC-Acquiz**	
Система банк	Страховой компания	European budget	Fishery agreement	Community producer	Fishing vessel
Платёжный система	Страховой	Budgetary	Fish stock	Import of product	Fishing
Система расчёт	Страхование	Budget	Fishing	Community market	Fishery
Система	Страховой случай	Commission budget	Fish	Community industry	Fish
Банковский система	Договор страхование	Budgetary policy	Fishing agreement	Producer	Vessel
Система платёж	Компания	Financial year	Fishing fleet	Sale of product	Fishing area
Развитие система	Страхование жизнь	Budget year	Fishery	Production	Fish stock
Работа система	Страховой выплата	Have	Have	Export price	Board fishing
Платёжный	Страховой взнос	Budgetary year	European commission	Import price	Fishing license
Расчёт	Страховщик	European fund	Committee	Community	Board vessel

2.7 Интеграция терминов в тематические модели

Помимо автоматического выбора словосочетаний также рассматривалась возможность интеграции в предложенные методы словосочетаний из ручных

терминологических ресурсов, разработанных экспертами (так называемых *«золотых стандартов»*). Под **термином** в рамках данной диссертационной работы будет пониматься эталонное слово или сочетание двух слов, выделенное экспертом, содержащееся в соответствующем «золотом стандарте».

В качестве «золотых стандартов» были выбраны:

- Для русского банковского корпуса – тезаурус, разработанный вручную для Центрального Банка Российской Федерации. Данный тезаурус включает в себя 15628 терминов, относящихся к сфере банковской активности, денежной политики и макроэкономики;
- Для английского корпуса Europarl – официальный многопрофильный тезаурус Европейского Союза Eurovoc[8], предназначенный для ручного индексирования заседаний Европарламента. Его английская версия включает в себя 15161 термин.

В Таблице 16 представлены результаты интеграции 1000 самых частотных терминов в алгоритм PLSA-SIM вместе с результатами первой итерации алгоритма PLSA-ITER со стеммерами, добавляющего словосочетания-термины (для сравнения приведены и результаты оригинального алгоритма).

Как видно из Таблицы 16, предложенные алгоритмы позволяют интегрировать в тематические модели словосочетания из ручных терминологических ресурсов, улучшая качество по сравнению с оригинальным алгоритмом.

2.8 Выводы ко второй главе

В данной главе был предложен новый алгоритм PLSA-SIM, добавляющий похожие слова и словосочетания в тематические модели и учитывающий сходство между словосочетаниями и образующими их словами. При тестировании была выделена группа мер, упорядочивающих словосочетания, добавление

[8]http://eurovoc.europa.eu/drupal

Таблица 16: Результаты интеграции терминов в предложенные алгоритмы

Коллекция	**Модель**	**Перплексия**	**TC-PMI**	**TC-PMI-nSIM**
Банковская	*PLSA*	1724.2	86.1	86.1
	PLSA-SIM + термины	1475.2	**151.8**	**102.9**
	PLSA-ITER + термины	**1267.6**	134.9	96
Europarl	*PLSA*	1594.3	53.2	53.2
	PLSA-SIM + термины	1522.4	**133.9**	**84.8**
	PLSA-ITER + термины	**1193.5**	97.4	66.6

верхних частей списков которых приводит к существенному улучшению качества тематических моделей по всем целевым мерам.

Кроме того, был предложен новый итеративный алгоритм PLSA-ITER, позволяющий более тщательно отбирать словосочетания для последующего добавления. Тестирование показывает, что использование жадных стеммеров для определения множеств похожих слов и словосочетаний в предложенном алгоритме позволяет ещё больше улучшить качество тематических моделей по основной мере – перплексии.

Также была показана возможность интеграции словосочетаний из ручных терминологических ресурсов, разработанных экспертами, в предложенные модели с улучшением качества по сравнению с оригинальным алгоритмом.

3 Применение тематических моделей в задаче автоматического извлечения терминов

Данная глава посвящена исследованию возможности применения тематической информации в задаче автоматического извлечения терминов. Для этого предлагаются новые признаки терминологичности слов и словосочетаний-кандидатов, основанные на тематических моделях, и разрабатываются комбинированные модели извлечения терминов. Главными целями являются нахождение наилучших индивидуальных признаков и определение вклада предложенных тематических признаков, а также комбинаций признаков, осуществляющих ранжирование так, чтобы в начале итогового списка стояли кандидаты, с наибольшей вероятностью являющиеся терминами.

3.1 Модели извлечения терминов из текстов предметной области

Автоматическое извлечение терминов представляет собой процедуру упорядочивания множества слов и словосочетаний S из текстовой коллекции D так, чтобы эталонные термины T_e оказались в начале списка. Под **термином** в рамках данной диссертационной работы будет пониматься слово или сочетание из двух слов, присутствующее в соответствующем «золотом стандарте» (т.е. в существующем, разработанном экспертами терминологического ресурсе).

Итак, пусть имеется X – множество описаний слов и словосочетаний, встречающихся в текстах некоторой предметной области, и 2 целевых класса $\{0, 1\}$, соответствующие тому, что рассматриваемое слово (словосочетание) является термином или нет. При этом на объектах обучающей выборки $X^m = \{x_1, \ldots, x_l\}$ известен правильный порядок $i \prec j$ на парах $(i, j) \in \{1, \ldots, l\}$. Требуется построить ранжирующую функцию $a : X \to \{0, 1\}$, сохраняющую правильный порядок на парах (i, j):

$$i \prec j \Longrightarrow a(x_i) \prec a(x_j) \tag{77}$$

и максимизирующую число эталонных терминов в начале списка:

$$a^* = \arg\max_a |\{t_e \in T_e | \forall s \in S \setminus T_e : a(s) \prec a(t_e)\}| \tag{78}$$

В качестве описаний слов и словосочетаний используются различные признаки. **Признаком** называется отображение: f: $X \to D_f$, где D_f – множество допустимых значений признака. Если заданы признаки $f_1, \ldots, f_n$, то вектор $x = (f_1(x), \ldots, f_n(x))$ называется признаковым описанием объекта $x \in X$. Признаковые описания, как правило, отождествляются с самими объектами. Все предложенные признаки распределяются по следующим группам [1, 71]:

1. **Признаки, основанные на частотности**. Основным предположением признаков данной группы является то, что термины, как правило, встречаются в текстовой коллекции гораздо чаще остальных слов:

$$TF(t) >> TF(w), \tag{79}$$

где t – термин, w – слово или словосочетание, не являющееся термином, $TF(w)$ и $TF(t)$ – частотность слова или словосочетания w (соответственно термина t) в коллекции.

2. **Признаки, использующие контрастную коллекцию**. Под такой коллекцией обычно понимается коллекция более широкой предметной области (например, национальные корпуса различных языков). Основная идея признаков из этой группы заключается в том, что частотности терминов в целевой коллекции должны быть существенно больше, чем в контрастной:

$$TF(t) >> TF_r(t), \tag{80}$$

где t – термин, $TF(t)$ и $TF_r(t)$ – частотность термина t в целевой и контрастной коллекциях соответственно.

3. **Контекстные признаки**. Данные признаки соединяют в себе информацию о частотностях слов и словосочетаний с данными о контекстах их употребления в текстовой коллекции. Под *контекстом* обычно понимаются слова, вместе с которыми встречается слово или словосочетание.

4. **Ассоциативные меры**. Признаки из данной группы предназначены только для извлечения многословных терминов. Под ассоциативными мерами понимаются математические критерии, определяющие силу связи между составными частями фраз, используя частотности словосочетаний и образующих их слов. Для вычисления ассоциативных признаков используется таблица сопряжённости для словосочетаний xy (см. Таблица 17).

Таблица 17: Таблица сопряжённости для словосочетаний xy

	x	$\overline{x} \neq x$
y	$TF(x, y)$	$TF(\overline{x}, y)$
$\overline{y} \neq y$	$TF(x, \overline{y})$	$TF(\overline{x}, \overline{y})$

5. **Гибридные признаки**. В последнее время стали появляться новые признаки, объединяющие в себе идеи ассоциативных мер с идеями других категорий признаков для лучшего ранжирования кандидатов.

Однако ни один из предложенных признаков не стал определяющим [77], и фактически из текстов извлекается большой список кандидатов, которые затем должны быть подтверждены экспертом по предметной области. Важно поэтому дополнять список используемых признаков, что позволит получать в начале списка больше кандидатов, являющихся терминами.

На текущий момент все описанные выше статистические признаки слабо отражают тот факт, что большинство терминов относятся к той или иной **тематике**, обсуждаемой в рамках текстов предметной области. Одной из немногих работ, в которой предлагается использовать тематическую информацию для извлечения терминов, стала работа [55], в которой вводится новая тематическая модель I-SWB для последующего извлечения терминов. Авторы утверждают, что словосочетание является термином, если образующие его слова принадлежат одной из следующих семантических групп:

- Слова, общие для всей предметной области;
- Слова, относящиеся к некоторой тематике в предметной области;
- Слова, специфичные для малого числа документов и отражающие некоторые характеристики предметной области.

Отталкиваясь от данной работы, было выдвинуто предположение, что множество всех терминов предметной области $Term$ обычно представимо в виде:

$$Term = Term_{general} \bigcup \left(\bigcup_{t \in T} Term_t \right), \tag{81}$$

где $Term_{general}$ – множество общих терминов, не относящихся к какой-либо тематике, T – множество всех тематик в текстовой коллекции, $Term_t$ – множество терминов, относящихся к тематике t.

При этом, как правило, мощность множества общих терминов намного меньше мощности множеств терминов, относящихся к тематикам, а сами эти множества не пересекаются. То есть выполняются следующие условия:

$$\forall t \in T: \ |Term_{general}| << |Term_t| \text{ и } Term_{general} \bigcap Term_t = \emptyset \tag{82}$$

Сами множества $Term_t$ могут тоже пересекаться, поскольку некоторые термины могут относиться сразу к нескольким тематикам в рамках текстов

предметной области. Для выявления же самих тематик, как правило, применяются подходы, основанные на тематическом моделировании [18].

Для проверки гипотезы о принадлежности терминов тематикам в данной главе будут рассмотрены все тематические модели, описанные в разделе 1.1:

- Основывающиеся на методах кластеризации текстов [45]: метод *K-средних* [60] и его модификации, методы *агломеративной кластеризации* [50] и метод неотрицательной матричной факторизации (NMF) [54];
- Вероятностные: метод латентного распределения Дирихле (LDA) [18] и метод вероятностного семантического анализа ($PLSA$) [46].

Все тематические модели определяют, к каким темам и с какой вероятностью относится каждый документ коллекции, и какие слова с какой вероятностью формируют каждую найденную тему. Таким образом, результатом работы любой тематической модели является нахождение распределения слов и словосочетаний по темам $P(w|t)$ и тем по документам $P(t|d)$.

Основная задача данной главы заключается в исследовании возможности использования тематической информации для повышения качества извлечения терминов. Для этой цели вначале в текстовой коллекции выделяются темы, а затем к ним применяются новые признаки, вычисляемые по построенным тематическим моделям. При этом вначале тематические модели будут исследованы с точки зрения задачи извлечения однословных и двусловных терминов с целью выбора наилучшей. Затем будет осуществлено сравнение признаков, посчитанных для лучшей тематической модели, с остальными признаками с целью определения вклада тематической информации в рассматриваемой задаче.

Стоит отметить, что в большинстве работ либо вообще не рассматривается задача извлечения однословных терминов [15, 34, 77, 88], поскольку более 85% всех терминов образуют многословные выражения [22], либо строится единая модель извлечения всех терминов [38, 99]. Важной особенностью же данного

исследования является то, что модели извлечения однословных и двусловных терминов исследуются отдельно, поскольку они могут быть различными.

3.2 Признаки, использующие тематическую информацию

Основной идеей всех признаков, использующих полученную с помощью тематической модели информацию, является тот факт, что термины относятся к той или иной тематике, обсуждаемой в рамках текстовой коллекции. В рамках данной диссертационной работы были предложены модификации некоторых хорошо известных признаков, вычисляемые по построенным тематическим моделям [8, 20, 74]. При их описании будут использованы следующие обозначения:

- $P(w|t)$ – условная вероятность слова (словосочетания) w в теме t;
- $DF(w)$ – число тем, содержащих слово (словосочетание) w с $P(w|t) > \epsilon$;[9]
- T – множество тем, полученных из текстовой коллекции.

Первым признаком, использующим тематическую информацию, является **Частотность** (**TF**). Значение признака тем больше, чем больше суммарная вероятность слова или словосочетания принадлежать различным темам:

$$TF(w) = \sum_{t \in T} P(w|t) \tag{83}$$

Также основную идею отражает и **Максимальная Частотность** (**Maximum TF**). Данный признак упорядочивает в верх списка слова и словосочетания, имеющие максимальную вероятность в какой-то из полученных тем:

$$Maximum\ TF(w) = \max_{t \in T} P(w|t) \tag{84}$$

[9] $\epsilon = 10^{-300}$

Для расчёта по полученным из текстовой коллекции темам был также модифицирован и признак **TF-IDF**, поощряющий слова и словосочетания, встречающиеся часто в малом числе тем:

$$TF\text{-}IDF(w) = TF(w) \times \log \frac{|T|}{DF(w)} \tag{85}$$

Следующим модифицированным признаком стал **Domain Consensus**, основанный на энтропии. Данная мера поощряет слова и словосочетания, часто встречающиеся в различных темах:

$$DomainConsensus(w) = -\sum_{t \in T} (P(w|t) \times \log P(w|t)) \tag{86}$$

В работе [17] было предложено визуализировать найденные темы с помощью меры **Term Score** (**TS**). Данная мера является расширением меры *TF-IDF* и принимает низкие значения для слов и словосочетаний, имеющих высокую вероятность принадлежности всем найденным темам:

$$TS(w) = \sum_{t \in T} TS(w|t), \text{ где } TS(w|t) = P(w|t) \times \log \frac{P(w|t)}{\left(\prod\limits_{t \in T} P(w|t)\right)^{\frac{1}{|T|}}} \tag{87}$$

Также рассматривается и признак **Maximum Term Score** (**Maximum TS**), представляющий собой максимальное значение *Term Score* среди всех тем:

$$Maximum\ TS(w) = \max_{t \in T} TS(w|t) \tag{88}$$

Последним признаком, использующим тематическую информацию, стал признак **TS-IDF**, комбинирующий идеи признака *TF-IDF* и *Term Score*:

$$TS\text{-}IDF(w) = TS(w) \times \log \frac{|T|}{DF(w)} \tag{89}$$

3.3 Прочие признаки кандидатов в термины

Помимо описанных выше признаков были предложены и несколько новых, не использующих тематическую информацию. Во-первых, были представлены следующие лингвистические признаки с булевским значением:

- **Неоднозначность** (**Ambiguity**) определяет, имеет ли кандидат в термины более одного варианта нормализации;
- **Новизна** (**Novelty**) фиксирует отсутствие слова или словосочетание в морфословаре (т.е. его новизну);
- **Специфичность** (**Specificity**) фиксирует, присутствует ли слово или словосочетание в контрастной коллекции текстов или нет;
- **Типы терминов-кандидатов** (*Существительное*, *Прилагательное* – для однословных терминов, *Существительное + Существительное* и *Прилагательное + Существительное* – для двусловных терминов).

Кроме того, некоторые признаки использовались для дополнительной фильтрации извлечённых кандидатов в термины. В качестве таких признаков для отбора тех подмножеств кандидатов, в которых плотность терминов априори выше, чем в исходном множестве, были взяты падеж главного слова-существительного и регистр первой буквы слова. По ним были отобраны соответственно:

- Кандидаты, в которых главное существительное стоит в именительном падеже – поскольку такая форма характерна для подлежащих, а подлежащие часто несут важную информацию для предметной области;
- Кандидаты, начинающиеся с заглавной буквы – поскольку они, скорее всего, представляют именованные сущности рассматриваемой области;

- Кандидаты, начинающиеся с заглавной буквы, но не стоящие первыми в предложениях текстов – для исключения случаев, когда заглавная буква слова свидетельствует только о начале предложения текста.

Для указанных подмножеств слов рассматривались шесть признаков: *Term Frequency*, *Document Frequency*, *TF-IDF* (два варианта: по целевой и по контрастной коллекции соответственно), *TF-RIDF* и *Domain Consensus*.

Также было сделано предположение, что термины скорее всего встречаются в текстах рядом с наиболее частотными словами. На основе этого предположения предложен признак **NearTermsFreq**, вычисляемый как число вхождений данного слова или словосочетания в контекстное окно для нескольких самых частотных кандидатов в термины, в качестве которых были взяты первые 10 самых частотных слов и словосочетаний-кандидатов.

Ещё один признак **NearTermsFreq-IDF** соединяет в себе предложенный признак и меру *TF-IDF*, вычисляемую по контрастной коллекции:

$$NearTermsFreq\text{-}IDF(w) = NearTermsFreq(w) \times \log \frac{|D|}{DF(w)} \qquad (90)$$

Кроме того, был предложен новый контекстный признак – *Modified Gravity Count*, основанный на ассоциативной мере *Gravity Count* [29]. Данный признак для словосочетания xy рассчитывается следующим образом:

$$Modified\ Gravity\ Count(xy) = \log \left(\frac{TF(xy) \times l(x)}{TF(x)} + \frac{TF(xy) \times r(y)}{TF(y)} \right), \qquad (91)$$

где $l(x)$ – количество различных слов, встретившихся слева от x, $r(y)$ – количество различных слов, встретившихся справа от y. При этом $l(x)$ и $r(y)$ рассчитывались только для слов, входящих в состав объемлющих именных групп.

Последними же двумя признаками стали номер позиции первого вхождения кандидата в термины и длина кандидата в словах.

3.4 Комбинирование признаков кандидатов в термины

Для комбинирования признаков использовался **метод градиентного бустинга**. Данный метод был впервые предложен в работе [41] и на текущий момент считается одним из наиболее универсальных алгоритмов машинного обучения. Метод градиентного бустинга (см. Алгоритм 4) относится к семейству алгоритмов, строящих ансамбли классификаторов, основываясь на следующей идее: каждый последующий классификатор минимизирует суммарную ошибку, которую даёт объединение всех предыдущих классификаторов. Данный алгоритм отличается от остальных тем, что использует метод градиентного спуска для построения ансамбля, минимизирующего ошибку на обучающей выборке, заданную дифференцируемой неотрицательной функцией потерь (штрафа) $L(y, y')$. Финальный алгоритм классификации ищется в виде композиции:

$$F(x) = \sum_{m=1}^{M} \gamma_m h_m(x), \gamma_m \in R, \tag{92}$$

где $h_m(x)$ – классификатор, добавленный на m-ом шаге, γ_m – вес $h_m(x)$.

При этом использовалась реализация над деревьями решений в библиотеке scikit-learn для языка Python[10]. Поскольку у данного метода много параметров для настройки, все параметры, кроме числа и максимальной глубины деревьев, были зафиксированы. Указанные же два параметра настраивались в каждом случае отдельно методом скользящего контроля по четырём блокам. Данный метод разбивает исходную выборку случайным образом на четыре равные непересекающиеся части, при этом каждая часть по очереди становится контрольной выборкой, а обучение проводится по остальным трём.

3.5 Проверка статистической значимости результатов

Для проверки статистической значимости результатов использовался непараметрический односторонний критерий знаковых рангов Вилкоксона, приме-

[10] `http://scikit-learn.org`

Алгоритм 4: Алгоритм градиентного бустинга

Вход: обучающая выборка $\{(x_i, y_i)\}$, функция потерь $L(y, F(x))$, число итераций M

Выход: алгоритм классификации $F_M(x)$

1 Инициализация композиции константной функцией:

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{l} L(y_i, \gamma)$$

2 **for** $m = 1, \ldots, M$ **do**

3 **for** $i = 1, \ldots, l$ **do**

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$$

4 Обучить классификатор $h_m(x)$ по обучающей выборке $\{(x_i, r_{im})\}_{i=1}^{n}$

5 Вычислить вес $h_m(x)$, решая задачу одномерной оптимизации:

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i))$$

6 Обновить композицию: $F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$

няемый для проверки различий между двумя выборками [83] (см. Алгоритм 5). В описании алгоритма используются следующие обозначения:

- $\{x_{1,i}\}$ и $\{x_{2,i}\}$ – выборки измерений;
- M_D – разница медиан между парами выборок $\{x_{1,i}\}$ и $\{x_{2,i}\}$;
- N_r – размер выборок после удаления пар, для которых верно $|x_{2,i} - x_{1,i}| = 0$;
- $z_{critical}$ – пороговое значение *z-score*, по которому принимается решение о справедливости той или иной гипотезы;
- $W_{critical,N_r}$ – пороговое значение статистики критерия знаковых рангов Вилкоксона, по которому принимается решение о справедливости той или иной гипотезы.

Алгоритм 5: Односторонний критерий знаковых рангов Вилкоксона

Вход: Выборки $\{x_{1,i}\}$ и $\{x_{2,i}\}$, гипотезы – H_0: $M_D < 0$ и H_1: $M_D \geq 0$

Выход: Принимаемая гипотеза

1 Вычислить $|x_{2,i} - x_{1,i}|$ и $sgn(x_{2,i} - x_{1,i})$ $(i = 1, \ldots, N)$

2 Удалить пары, для которых верно: $|x_{2,i} - x_{1,i}| = 0$

3 Упорядочить пары, начиная с минимального значения $|x_{2,i} - x_{1,i}|$

4 Присвоить парам ранги R_i, начиная с 1

5 Вычислить статистику критерия знаковых рангов Вилкоксона W:

$$W = \left| \sum_{i=1}^{N_r} (sgn(x_{2,i} - x1, i) * R_i) \right|$$

6 **if** $N_r \geq 10$ **then**

7 Вычислить значение *z-score*:

$$z = \frac{W - \frac{1}{2}}{\sigma_w}, \text{ где } \sigma_w = \sqrt{\frac{N_r(N_r + 1)(2N_r + 1)}{6}}$$

8 **if** $z > z_{critical}$ **then**

9 Отвергнуть гипотезу H_0 в пользу H_1

10 **else**

if $W \geq W_{critical,N_r}$ **then**

11 Отвергнуть гипотезу H_0 в пользу H_1

При этом использовалась реализация в библиотеке stats для языка R[11]. Данный метод возвращает *p-value* – вероятность ошибки первого рода. Если *p-value* меньше заданного уровня значимости (обычно равного 0.05), то нулевая гипотеза отвергается в пользу альтернативной. Иначе она принимается.

[11] `https://stat.ethz.ch/R-manual/R-patched/library/stats/html/wilcox.test.html`

3.6 Текстовые коллекции и предобработка

Для тестирования признаков, предложенных в данной главе, использовались коллекции разных языков и предметных областей (см. раздел 2.4):

- Коллекция банковских русскоязычных текстов (примерно 18.5 млн слов);
- Английская часть корпуса текстов Europarl[12] (примерно 54 млн. слов).

Для подтверждения терминологичности кандидатов в термины использовались следующие «золотые стандарты» (см. раздел 2.7).

- Для русской части исследования – тезаурус, разработанный вручную для Центрального Банка Россиийской Федерации (15628 терминов);
- Для английской части исследования – официальный многопрофильный тезаурус Европейского Союза Eurovoc[13] (15161 термин).

При этом слово или словосочетание-кандидат считается термином, если оно содержится в соответствующем тезаурусе. Все признаки кандидатов в термины рассчитывались для 5000 самых частотных слов и словосочетаний.

На этапе предобработки проводился морфологический анализ текстовых коллекций. Для корпуса Europarl использовались средства Stanford Core NLP[14], а для банковского – собственный морфологический анализатор. Все слова были лемматизированы. В качестве кандидатов в термины рассматривались:

- Для английской коллекции – слова и словосочетания в формах: *Существительное*, *Существительное + Существительное*, *Прилагательное + Существительное*, *Существительное + of + Существительное*;

[12] `http://www.statmt.org/europarl/`

[13] `http://eurovoc.europa.eu/drupal`

[14] `http://nlp.stanford.edu/software/corenlp.shtml`

- Для русской коллекции – слова и словосочетания в формах: *Существительное*, *Прилагательное*, *Существительное + Существительное в родительном падеже*, *Прилагательное + Существительное*.

Отбор слов и словосочетаний-кандидатов был осуществлён именно по таким лингвистическим формам, потому что они покрывают большую часть однословных и двусловных терминов [22].

Для улучшения результатов извлечения терминов для русской коллекции проводилась фильтрация морфологических омонимов. Во-первых, из рассмотрения исключались те варианты нормализации слов, словоформы которых не согласуются в тексте с соседними словами. Так, для словоформы *банке* из словосочетания *в центральном банке* отбиралась только нормальная форма *банк* (но не *банка*). Во-вторых, удалялись слова, нормальные формы которых совпадали с нормальными формами слов других частей речи, так как маловероятно, что они окажутся терминами в данном контексте. Так, исключалось слово *том* в словосочетании *в том* из-за возможной словоформы *том* местоимения *то*.

Для улучшения результатов извлечения терминов на данном этапе для английской коллекции был вручную составлен список из стоп-слов (таких, как *other*, *another*, *that*, *this*, *those*, *etc.*). Из рассмотрения исключались все слова из этого списка и словосочетания, содержащие такие слова.

3.7 Выбор лучшей тематической модели для извлечения терминов

Для исследования были выбраны все рассмотренные в разделе 1.1 тематические модели: *K-Средних*, *Сферический K-Средних*, *Иерархическая кластеризация* с полным, одиночным и средним связыванием, *NMF*, минимизирующий Евклидово расстояние, *PLSA* и *LDA*. В качестве базовой была взята «тематическая» модель, рассматривающая каждый документ как отдельную тему.

В Таблицах 18 и 19 представлены лучшие признаки для каждой модели для 5000 самых частотных слов и словосочетаний для исследуемых корпусов.

Таблица 18: Средняя точность *AvP*@5000 лучших тематических признаков

Модель	**Банковский корпус**		**Корпус Europarl**	
	Лучший признак	**AvP**	**Лучший признак**	**AvP**
Базовая	TS-IDF	40.7	Maximum TS	36.7
K-Средних	TF-IDF	36.8	TF-IDF	32.2
Сферический K-Средних	Term Score	32.5	TS-IDF	29.8
Иерархическая, одиночное	TF-IDF	36.9	Maximum TS	40
Иерархическая, полное	TF-IDF	38.3	Maximum TS	39
Иерархическая, среднее	TF-IDF	38.7	Maximum TS	39
NMF	Term Score	46.9	Term Score	43.3
PLSA	**TS-IDF**	**49.8**	**Maximum TS**	**45.7**
LDA	Maximum TS	44.9	Maximum TS	43

Как видно из результатов, представленных в Таблицах 18 и 19, лучшее качество независимо от языка и предметной области даёт тематическая модель **PLSA**. Так, лучшими признаками для извлечения однословных терминов оказались **TS-IDF** для банковского корпуса и **Maximum TS** для корпуса Europarl с приростом качества относительно лучших признаков базовой модели в **22.6%** и **24.9%** соответственно. А лучшим признаком для извлечения двусловных терминов для обоих языков стал **Term Score** с приростом качества относительно лучших признаков базовой модели в **3.3%** и **20.8%** соответственно.

Помимо вычисления отдельных признаков осуществлялось их комбинирование для каждой модели с помощью метода градиентного бустинга. Результаты комбинирования признаков представлены в Таблице 20.

Как видно из результатов, представленных в Таблице 20, тематическая

Таблица 19: Средняя точность *AvP*@5000 лучших тематических признаков

Модель		**Банковский корпус**		**Корпус Europarl**	
		Лучший признак	**AvP**	**Лучший признак**	**AvP**
Базовая		TS-IDF	47.8	Maximum TF	34.6
K-Средних		Domain Consensus	42.6	Domain Consensus	37.3
Сферический K-Средних		Term Frequency	42.6	Term Frequency	31.8
Иерархическая	*одиночное*	Domain Consensus	44.8	Term Score	39.1
	полное	Domain Consensus	45.5	Term Score	37.5
	среднее	Domain Consensus	45.4	Term Score	38.5
NMF		Domain Consensus	46.5	Term Frequency	41.1
PLSA		**Term Score**	**49.4**	**Term Score**	**41.8**
LDA		Term Score	49.3	Term Score	41.7

модель **PLSA** снова даёт наилучшее качество. Для однословных терминов прирост относительно базовой модели составил **36.4%** для банковского корпуса и **17.7%** – для корпуса Europarl. А для двусловных терминов данный прирост составил **21.2%** для банковского корпуса и **22.2%** – для корпуса Europarl.

Таким образом, лучшей тематической моделью для извлечения терминов независимо от языка и предметной области оказалась модель **PLSA**.

3.8 Вклад тематических признаков в модель извлечения терминов

Для изучения вклада тематической информации в задачу извлечения терминов результаты предложенных признаков, посчитанных для лучшей тематической модели PLSA, сравнивались с остальными признаками для исследуемых корпусов для 5000 самых частотных слов и словосочетаний. В качестве призна-

Таблица 20: Средняя точность AvP@5000 комбинирования тематических признаков

Модель		**Средняя точность**			
		Однословные термины		**Двусловные термины**	
		Банковский корпус	**Корпус Europarl**	**Банковский корпус**	**Корпус Europarl**
Базовая		37.4	38.5	44.3	37.4
K-Средних		41.8	33.3	44.8	40
Сферический K-Средних		38.9	31.1	47.2	37.3
Иерархическая	*одиночное*	40.2	41.8	45.5	38.6
	полное	42.3	40.6	45.2	38.1
	среднее	44.1	40.1	45.9	38.3
NMF		50	43.9	51	44.4
PLSA		**51**	**45.3**	**53.7**	**45.7**
LDA		45.9	45	52.2	43.7

ков, не использующих тематическую информацию, были взяты все, представленные в разделах 1.3 и 3.3. Всего рассматривалось, включая предложенные тематические признаки, **67** признаков для однословных терминов и **86** признаков для двусловных терминов. Лучшие признаки каждой из упомянутых выше групп для обоих корпусов приведены в Таблицах 21 и 22.

Как видно из Таблиц 21 и 22, независимо от языка и предметной области лучшими признаками для извлечения однословных терминов оказались тематические (**TS-IDF** для банковского корпуса и **Maximum TS** для корпуса Europarl с приростом качества относительно остальных признаков в **7.7%** и **16%** соответственно). Для извлечения двусловных же терминов лучшим признаком стал предложенный в разделе 3.3 признак **Modified Gravity Count**

Таблица 21: Средняя точность *AvP*@5000 лучших признаков для однословных терминов

Группа признаков	**Банковский Корпус**		**Корпус Europarl**	
	Лучший признак	**AvP**	**Лучший признак**	**AvP**
Основанные на частотности	Term Variance	45.9	Term Variance	39.4
Использующие контрастный корпус	Contrastive Weight	42	Contrastive Weight	30.3
Контекстные	Token-FLR	34.3	C-Value	31.3
Тематические	**TS-IDF**	**49.8**	**Maximum TS**	**45.7**
Прочие	TF (слова с большой буквы)	39.3	TF-RIDF (подлежащие)	38.5

с приростом качества относительно остальных признаков в **0.9%** и **1.5%** для банковского корпуса и корпуса Europarl соответственно.

Также на Рисунках 7 и 8 представлены графики средней точности в зависимости от числа наиболее частотных кандидатов в термины для отдельных признаков и комбинации всех признаков с помощью метода градиентного бустинга. В качестве индивидуальных признаков рассматривались лучшие тематические и широко известные базовые: *Weirdness*, *TF-IDF* и *C-Value* для однословных терминов и *TF-IDF*, *C-Value* и *MI* для двусловных терминов.

Рисунки 7 и 8 ещё раз показывают, что лучшими признаками для извлечения однословных терминов являются тематические (*TS-IDF* и *Maximum TS*), а для извлечения двусловных терминов – признак *Modified Gravity Count*. В то же время комбинирование всех признаков методом градиентного бустинга даёт существенный прирост качества по сравнению с индивидуальными признаками.

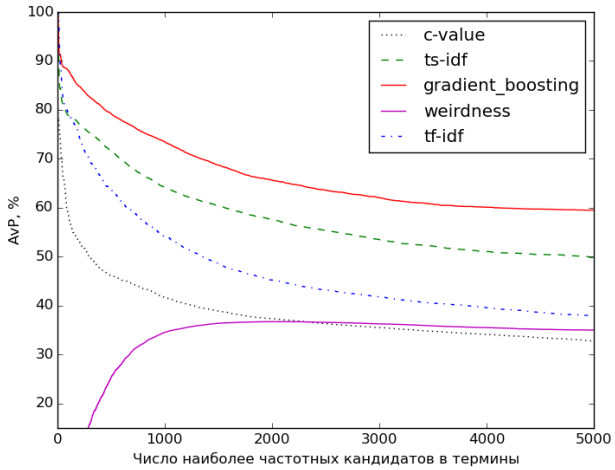
Замечание. В близкой работе [55] приводится сравнение результатов

Таблица 22: Средняя точность *AvP*@5000 лучших признаков для двусловных терминов

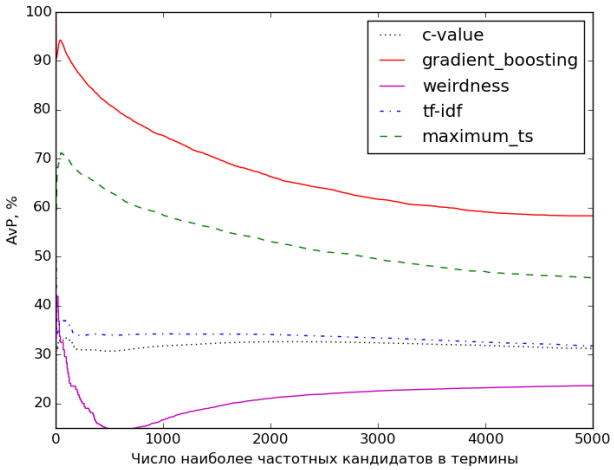
Группа признаков	**Банковский корпус**		**Корпус Europarl**	
	Лучший признак	**AvP**	**Лучший признак**	**AvP**
Основанные на частотности	TF-RIDF	53.3	Term Variance Quality	42.2
Использующие контрастный корпус	Loglikelihood	48.1	Loglikelihood	36.4
Контекстные	**MGCount**	**53.8**	**MGCount**	**43.6**
Тематические	Term Score	49.4	Maximum TF	40.8
Ассоциативные меры	Gravity Count	50.5	Gravity Count	41.6
Прочие	NearTermsFreq	45.8	TF-RIDF (подлежащие)	40.7

Рис. 7: Сравнение признаков слов для исследуемых корпусов

Банковский корпус



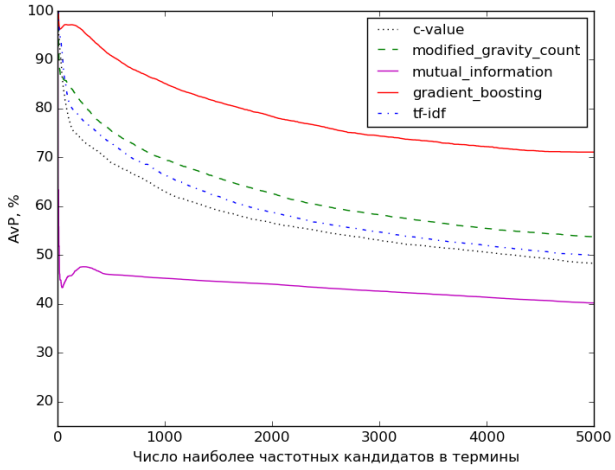
Корпус Europarl



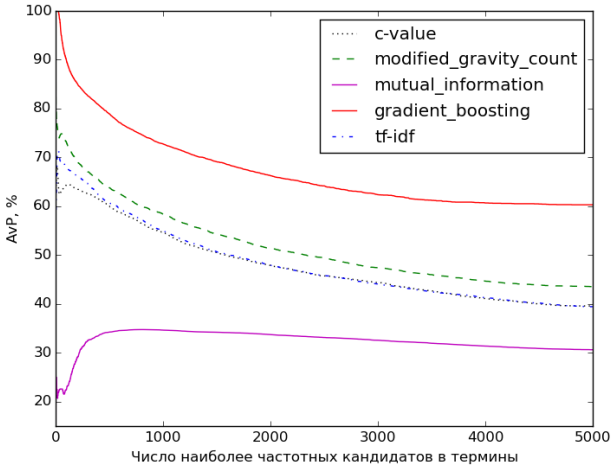
работы тематической модели I-SWB в задаче извлечения терминов с откры-

Рис. 8: Сравнение признаков словосочетаний для исследуемых корпусов

Банковский корпус



Корпус Europarl



той системой TerMine[15], основанной на признаках C-Value/NC-Value, и индивидуальным признаком *TF-IDF*. В качестве целевой меры используется мера *Precision@n* (см. раздел 1.3.6). Авторы показывают прирост качества на уровне первых 500 слов примерно в 12% независимо от предметной области. Для сравнения с этими данными составлялись единые унифицированные списки, содержащие однословные и двусловные кандидаты, извлечённые с помощью базовых признаков *TF-IDF*, *C-Value*, *NC-Value* и лучших тематических признаков (*TS-IDF* и *Term Score* для банковского корпуса; *Maximum TS* и *Maximum TF* для корпуса Europarl). В результате для банковского корпуса лучшие тематические признаки показали прирост качества на уровне первых 500 слов в **33%**, а для корпуса Europarl – в **54%**. Таким образом, новые предложенные признаки показывают значительно больший прирост качества, чем в работе [55].

Для оценки же вклада тематических признаков в общую модель извлечения терминов модель, учитывающая тематические признаки (для базовой и лучшей тематической модели), сравнивалась с моделью, не использующей их. Результаты сравнения для обоих корпусов на уровне 5000 самых частотных кандидатов в термины приведены в Таблице 23.

[15] `http://www.nactem.ac.uk/software/termine/`

Таблица 23: Средняя точность AvP@5000 моделей извлечения терминов с тематическими признаками и без них

Модель	**Однословные термины**		**Двусловные термины**	
	Банковский корпус	**Корпус Europarl**	**Банковский корпус**	**Корпус Europarl**
Без тематических признаков	57.3	58.5	70.8	60.0
С тематическими признаками	**59.0**	**58.7**	**71.6**	**60.3**

Для проверки статистической значимости результатов использовался односторонний критерий знаковых рангов Вилкоксона (см. раздел 3.5). Для каждой модели вычислялись значения средней точности на разных уровнях самых частотных кандидатов (до 5000 включительно). Проверялась гипотеза H_0 о том, что модель без тематических признаков работает лучше модели с тематическими признаками. Полученные значения *p-value* представлены в Таблице 24.

Таблица 24: Статистическая значимость вклада тематических признаков в задачу извлечения терминов

Модель извлечения	**Банковский корпус**	**Корпус Europarl**
Однословные термины	$p\text{-}value < 2.2 * 10^{-16}$	$p\text{-}value = 0.02251$
Двусловные термины	$p\text{-}value < 2.32 * 10^{-11}$	$p\text{-}value = 0.01605$

Как видно из Таблицы 24, полученные значения *p-value* меньше уровня значимости, равного 0.05. Поэтому гипотеза H_0 была отвергнута в пользу альтернативной. Таким образом, было показано, что тематические модели вносят дополнительную информацию в процесс извлечения терминов.

В Таблице 25 представлены первые 10 кандидатов, извлечённых с помо-

щью моделей с тематическими признаками (термины выделены курсивом).

Таблица 25: Примеры извлечённых слов и словосочетаний-кандидатов

№	Однословные термины		Двусловные термины	
	Банковский корпус	**Корпус Europarl**	**Банковский корпус**	**Корпус Europarl**
1	*Налоговый*	*Brazil*	*Государственный регистрация*	*Public service*
2	*Аудит*	*Croatia*	*Федеральный бюджет*	*Health care*
3	*Валютный*	*Cuba*	*Банковский вклад*	*United nation*
4	*Валюта*	*Georgia*	*Саморегулируемый организация*	*Code of conduct*
5	*Банковский*	*Syria*	*Финансовый отчётность*	*Regional policy*
6	Риска	*Iceland*	*Доверительный управление*	*Rural development*
7	*Страхование*	*Taiwan*	*Иностранный валюта*	*Social service*
8	*Актив*	*Israel*	*Налоговый период*	*Economic policy*
9	Млрд	*Egypt*	*Некоммерческий организация*	*Natural resource*
10	*Акция*	*Afghanistan*	*Информационный технология*	*Exchange rate*

3.9 Унифицированная модель извлечения терминов

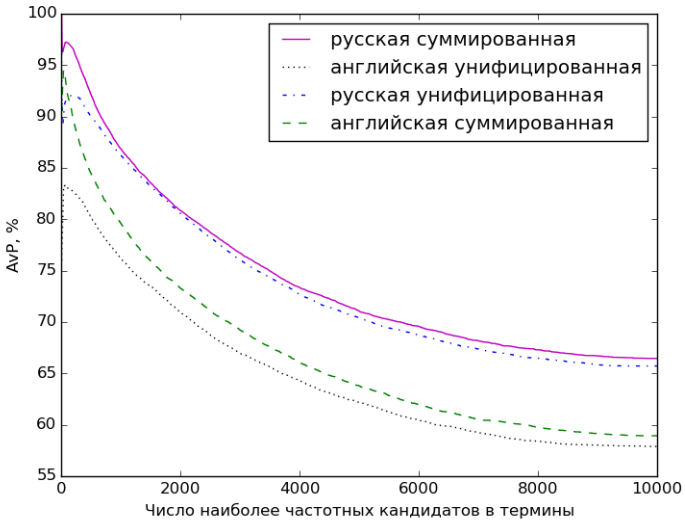
Помимо разработки отдельных моделей извлечения терминов изучался вопрос создания единой унифицированной модели, извлекающей одновременно однословные и двусловные термины [73]. Для её создания использовались все признаки, кроме ассоциативных и гибридных, применимые для извлечения

однословных и двусловных терминов. На выходе такая модель выдаёт унифицированный список слов и словосочетаний-кандидатов.

Для сравнения качества извлечённых унифицированной моделью терминов отдельно обучались модели для извлечения однословных и двусловных терминов. Результирующие списки слов и словосочетаний-кандидатов объединялись в один в соответствии с вероятностями, полученными методом градиентного бустинга. Таким образом получался суммированный список кандидатов в термины, который сравнивался с унифицированным.

На Рисунке 9 приведены графики зависимости средней точности от числа наиболее частотных кандидатов в термины для обоих корпусов, показывающие, что суммированный список незначительно превосходит унифицированный.

Рис. 9: Сравнение унифицированной модели извлечения с суммированной



В Таблице 26 представлены первые 10 кандидатов, полученных унифицированными моделями для обоих корпусов (термины выделены курсивом).

Стоит отметить, что унифицированные модели включают в себя слишком много различных признаков. Возможно, некоторые из них являются избыточными и только усложняют обучение моделей. Для исключения таких признаков применялся жадный пошаговый алгоритм *Add* [2].

Таблица 26: Примеры кандидатов, извлечённых унифицированными моделями

№	**Банковский корпус**	**Корпус Europarl**
1	*Банковский*	*Canada*
2	*Финансовый отчётность*	Green paper
3	*Денежно-кредитный политика*	*Belarus*
4	*Валютный операция*	*Cuba*
5	*Иностранный валюта*	White paper
6	*Доверительный управление*	*Budgetary control*
7	*Бухгалтерский учёт*	*Social service*
8	Ассоциация	*Brazil*
9	*Корпоративный управление*	*Serbia*
10	*Ипотечный кредит*	*Romania*

Данный алгоритм начинает свою работу с пустого множества признаков, а затем на каждом шагу добавляет признак, максимизирующий среднюю точность, пока есть улучшение между итерациями. В результате была найдена комбинация из **13** признаков (всего – **67**) для обеих коллекций (см. Таблицу 27).

Из того, что в таблице 27 присутствуют представители всех групп признаков, следует, что каждая группа признаков является существенной для унифицированных моделей извлечения терминов независимо от языка и предметной области. Кроме того, следует отметить, что наборы признаков для обеих моделей получились достаточно похожими друг на друга.

3.10 Применение тематических моделей, полученных алгоритмом PLSA-SIM, для извлечения терминов

На последнем этапе тестирования в задаче извлечения однословных терминов были апробированы модели, получаемые предложенным в разделе 2.1

Таблица 27: Результат отбора признаков для унифицированных моделей

№	**Банковский корпус**	**Корпус Europarl**
1	Modified Gravity Count	Modified Gravity Count
2	C-Value	C-Value
3	Sum10	Sum10
4	NearTermsFreq	NearTermsFreq
5	TF-IDF (слова с заглавной буквы)	TF-IDF (слова с заглавной буквы)
6	TF-RIDF (подлежащие)	TF-RIDF (подлежащие)
7	Прилагательные	Прилагательные
8	TF-IDF (PLSA)	Maximum TF (PLSA)
9	Relevance	Discriminative Weight
10	Type-LR	Token-LR
11	DF (подлежащие)	TF-IDF (подлежащие)
12	Номер первого вхождения	Новизна
13	TS-IDF (базовая модель)	Term Variance Quality

алгоритмом PLSA-SIM. Для данной цели были выбраны тематические модели, построенные алгоритмом PLSA-SIM с 1000 самых частотных словосочетаний, показавшие в разделе 2.5 улучшение по всем целевым мерам качества.

По построенным моделям вычислялись все тематические признаки, предложенные в разделе 3.2: *Term Frequency (TF), Maximum TF, TF-IDF, Domain Consensus, Term Score (TS), Maximum TS, TS-IDF*. Однако для учёта влияния добавленных в модель словосочетаний использующаяся в данных признаках вероятность принадлежности слова w теме t $P(w|t)$ была заменена на $\hat{P}(w|t)$, вычисляемую согласно признаку *C-Value* (см. раздел 1.3.3):

$$\hat{P}(w|t) = P(w|t) - \frac{\sum\limits_{p \in P_w} P(p|t)}{|P_w|}, \tag{93}$$

где w – слово, p – словосочетание, $P(w|t)$ и $P(p|t)$ – вероятности слова w и словосочетания p принадлежности теме t, P_w – множество всех добавленных в тематическую модель словосочетаний, содержащих слово w.

В Таблице 28 приведены значения средней точности лучших предложенных признаков вместе с лучшими индивидуальными признаками (см. раздел 3.8) для исследуемых коллекций на уровне 5000 самых частотных слов.

Таблица 28: Средняя точность *AvP*@5000 лучших признаков, посчитанных для модели PLSA и PLSA-SIM

Корпус	**Признак**	**AvP@5000**
Русский	Maximum Term Score (PLSA-SIM)	47.7
	TS-IDF (PLSA)	**49.4**
Английский	**Term Score (PLSA-SIM)**	**46.1**
	Maximum Term Score (PLSA)	45.7

Как видно из Таблицы 28, предложенные признаки, посчитанные для модели PLSA-SIM, необязательно являются лучшими. Поэтому для оценки вклада предложенных признаков в общую модель извлечения однословных терминов модель извлечения терминов, учитывающая данные признаки, сравнивалась с моделью, не использующей их. Результаты сравнения для обоих корпусов на уровне 5000 самых частотных слов приведены в Таблице 29.

Таблица 29: Средняя точность *AvP*@5000 моделей извлечения однословных терминов с признаками, посчитанными по модели PLSA-SIM, и без них

Модель	**Русский корпус**	**Английский корпус**
Без признаков по PLSA-SIM	59.0	58.7
С признаками по PLSA-SIM	**59.9**	**58.9**

Для проверки статистической значимости результатов использовался од-

носторонний критерий знаковых рангов Вилкоксона (см. раздел 3.5). Для каждой сравниваемой модели вычислялись значения средней точности на разных уровнях самых частотных слов (до 5000 включительно). Проверялась гипотеза H_0 о том, что модель без признаков, посчитанных по модели PLSA-SIM, работает лучше модели с данными признаками. В результате были получены следующие значения:

- Для русской части исследования $p\text{-}value < 2.2 * 10^{-16}$;
- Для английской части исследования $p\text{-}value = 0.002411$.

Поскольку полученные значения $p\text{-}value$ меньше уровня значимости, равного 0.05, гипотеза H_0 была отвергнута в пользу альтернативной. Таким образом, было показано, что признаки, посчитанные по модели PLSA-SIM, вносят дополнительную информацию в процесс извлечения однословных терминов.

3.11 Выводы к третьей главе

В данной главе представлены результаты исследования возможности применения тематических моделей для улучшения качества извлечения терминов.

Были предложены отдельные комбинированные модели извлечения однословных и двусловных терминов. Было показано, что значимость разных признаков извлечения терминов в этих моделях существенно различается. В процессе моделирования был введён новый контекстный признак извлечения терминов Modified Gravity Count, показавший наилучшее качество извлечения для двусловных терминов независимо от языка и предметной области.

Также были предложены новые признаки, основанные на тематических моделях. При тестировании было показано, что применение таких признаков в рассматриваемой задаче улучшает качество извлечения терминов. При этом наибольший рост качества выявлен при извлечении однословных терминов.

4 Система построения вероятностных тематических моделей на основе лексико-терминологической информации

В рамках данной диссертационной работы был разработан программный комплекс по построению вероятностных тематических моделей с использованием лексико-терминологической информации и применению тематических моделей для извлечения терминов. Данный комплекс выложен в открытый доступ[16] и включает в себя следующие условно-независимые пакеты программ:

- Пакет программ построения тематических моделей с возможностью добавления словосочетаний и похожих слов;
- Пакет программ извлечения однословных и двусловных терминов, использующий тематическую информацию.

Данные пакеты программ могут взаимодействовать друг с другом по принципу конвейера (результаты работы одного пакета могут передаваться другому на вход) в любой последовательности – либо снабжая пакет программ извлечения терминов тематическими моделями, либо, наоборот, отдавая двусловные термины для добавления в тематические модели.

4.1 Общее описание программного комплекса

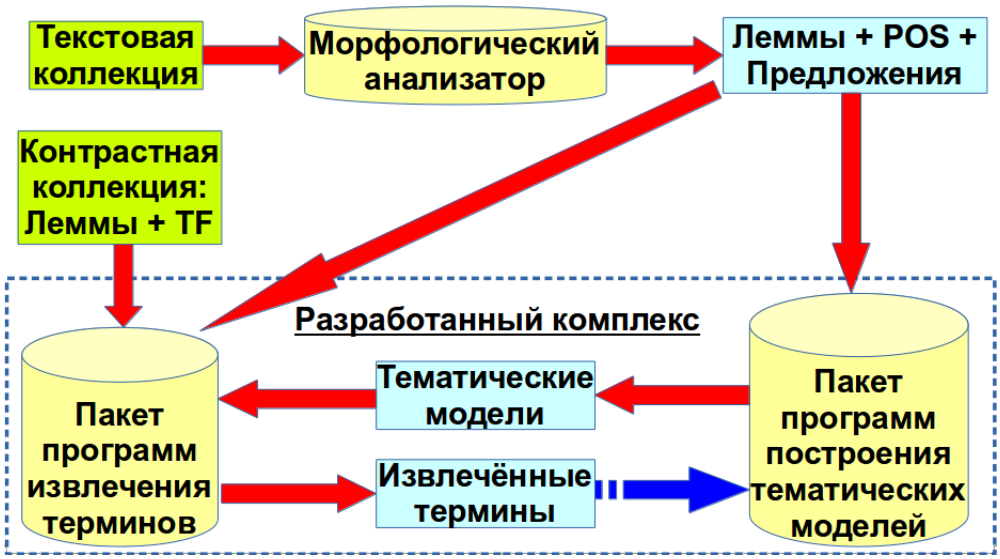
4.1.1 Архитектурная схема

На Рисунке 10 представлена общая архитектурная схема разработанного программного комплекса (штрихованная стрелка означает необязательную связь, сплошная – обязательную). В качестве входных данных выступает тек-

[16] `https://bitbucket.org/Meister17/dissertation`

стовая коллекция некоторой предметной области, которая проходит предварительную обработку внешним модулем – морфологическим анализатором.

Рис. 10: Архитектурная схема программного комплекса



Результаты работы морфологического анализатора (леммы слов, части речи, границы предложений) передаются на вход двум пакетам программ. Пакет программ построения тематических моделей способен строить тематические модели методом PLSA, LDA, PLSA-SIM и PLSA-ITER. Также на вход можно подать список словосочетаний для добавления в строящиеся модели (в качестве такого списка могут выступать извлечённые вторым пакетом двусловные термины). Кроме того, на вход пакету программ извлечения терминов поступает также контрастная коллекция (в виде лемм слов и частотностей), а также результат работы пакета программ построения тематических моделей.

В качестве среды разработки использовался текстовый редактор Sublime Text Editor версии[17]. Основные части модулей написаны на языке C++ с использованием библиотеки Boost[18]. Также использовался язык Python и библиотека scikit-learn[19] для машинного обучения методом градиентного бустинга. Суммарное количество написанного лично автором строк кода – 12000.

[17] http://www.sublimetext.com/

[18] http://www.boost.org/

[19] http://scikit-learn.org/stable/

4.1.2 Внешний модуль морфологического анализатора

Первым этапом обработки поданной на вход программному комплексу текстовой коллекции является морфологический анализ. Основными результатами работы морфологического анализатора являются:

1. Токенизация текстовой коллекции. В результате входной поток разбивается на отдельные токены, соответствующие словам и знакам препинания.

2. Лемматизация токенов. Каждому найденному токену сопоставляется лемма – начальная форма слова. Данный процесс необходим, поскольку существует большое число словоформ одних и тех же слов, которые должны единообразно обрабатываться любыми алгоритмами обработки текстов.

3. Определение частей речи. Помимо осуществления лемматизации слов в текстовой коллекции морфологический анализатор также определяет части речи найденных лемм. Данный процесс необходим для последующего извлечения слов и словосочетаний (в виде именных групп).

4. Определение границ предложений. Поскольку из текстовой коллекции выделяются не только слова, но и знаки препинания, морфологический анализатор также способен определять границы предложений. Данная информация используется модулем извлечения терминов для определения слов, с которых начинаются предложения в текстах.

5. Определение словарных лемм. Для проверки на наличие слова в словаре (см. раздел 3.3) использовался морфословарь анализатора для русского языка и библиотека PyEnchant для английского языка[20].

В результате работы морфологического анализатора получается структурированное описание текстовой коллекции: набор лемм слов с указанием их частей речи и разбиением по предложениям.

[20] `http://pythonhosted.org/pyenchant/`

Для морфологического анализа использовались внешние модули:

- Для русского языка – модуль, разработанный в Лаборатории Анализа Информационных Ресурсов НИВЦ МГУ им. М. В. Ломоносова.
- Для английского языка – Stanford Core NLP[21], разработанный в Стэнфордском университете.

Примеры обработки русского и английского текста морфологическими анализаторами представлены на Рисунках 11 и 12.

Рис. 11: Пример работы морфологического анализатора для русского языка

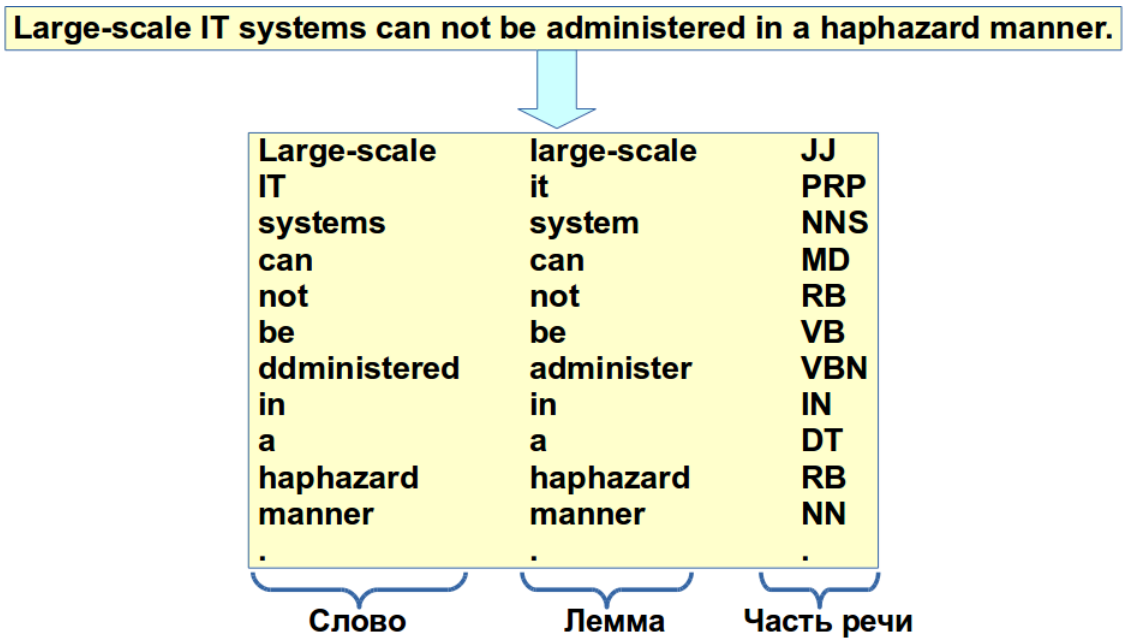
Исследование и разработка модуля морфологического анализатора не является предметом данной диссертационной работы.

4.2 Пакет программ построения тематических моделей

Данный пакет программ реализует возможности построения тематических моделей, используя алгоритмы *PLSA*, *LDA*, *PLSA-SIM* и *PLSA-ITER*. На

[21] http://nlp.stanford.edu/software/corenlp.shtml

Рис. 12: Пример работы морфологического анализатора для английского языка



вход пакету программ поступает текстовая коллекция, прошедшая обработку модулем морфологического анализатора. Архитектурная схема работы пакета программ представлена на Рисунке 13 (штрихованная стрелка обозначает необязательную связь, сплошная – обязательную).

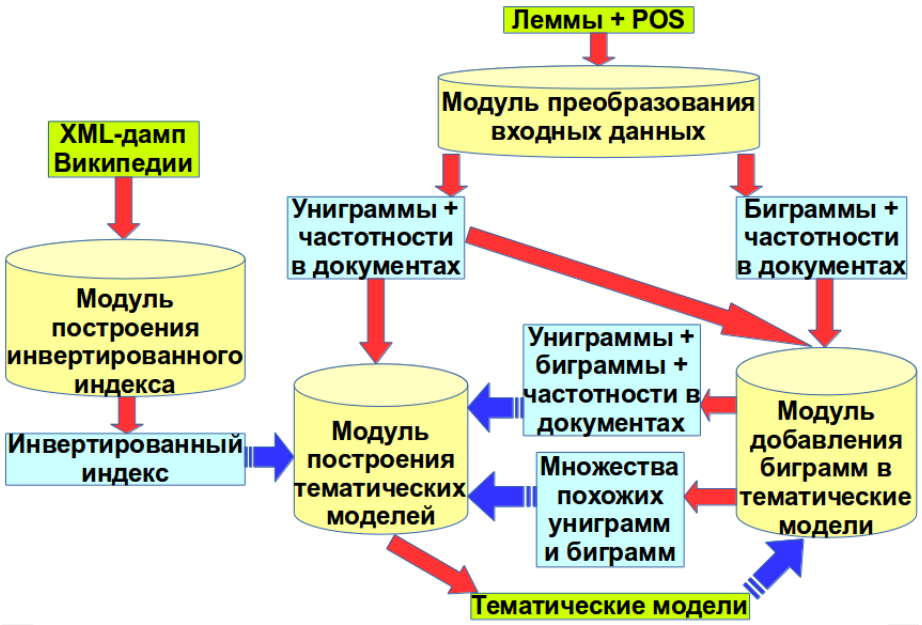
Пакет программ построения тематических моделей состоит из четырёх условно-независимых модулей: модуля преобразования входных данных, модуля добавления словосочетаний в тематические модели, модуля построения инвертированного индекса и модуля построения тематических моделей.

4.2.1 Модуль преобразования входных данных

Данный модуль преобразует входные данные в формат, удобный для дальнейшего построения тематических моделей. Модуль получает на вход набор лемм с проставленными частями речи, полученный в результате работы модуля морфологического анализатора.

В качестве слов, участвующих в образовании тем, по умолчанию рассмат-

Рис. 13: Схема пакета программ построения тематических моделей



риваются только *существительные*, *прилагательные*, *глаголы* и *наречия*, поскольку другие слова не играют особой роли в данном процессе. В то же время имеется возможность выбрать конкретные части речи наряду с языком текстовой коллекции. Кроме того, есть возможность указать минимальный порог частотности, ниже которого слова и словосочетания не рассматриваются при образовании тем. По умолчанию, слова и словосочетания, встретившиеся в текстовой коллекции менее 5 раз исключаются из рассмотрения.

Также модуль извлекает все встретившиеся словосочетания в формах:

- *Существительное + Существительное*, *Прилагательное + Существительное*, *Существительное + предлог of + Существительное* – для английских коллекций;
- *Существительное + Существительное в родительном падеже*, *Прилагательное + Существительное* – для русских коллекций.

Рассматриваются только такие словосочетания, поскольку темы, как правило, образуются с помощью существительных и именных групп.

В результате работы данного модуля образуется набор входных данных для последующего построения тематических моделей: набор слов и словосочетаний вместе с их частотностями в документах. Данный набор представляется для слов и словосочетаний в виде двух файлов:

- Словарь – файл, содержащий все слова (или словосочетания) в лексикографическом порядке;
- Файл с частотностями слов (или словосочетаний) в документах коллекции. Причём i-я строчка этого файла соответствует одному документу d_i, который в свою очередь представляется в следующем виде:

$$d_i = \{(id_w, TF(w|d_i)) : TF(w|d_i) > 0\}, \tag{94}$$

где w – слово (или словосочетание), id_w – индекс w в словаре, $TF(w|d_i)$ – частотность w в документе d_i.

4.2.2 Модуль добавления словосочетаний в тематические модели

Данный модуль предназначен для добавления словосочетаний в тематические модели, а также формирования множеств похожих слов и словосочетаний для предложенных алгоритмов PLSA-SIM (см. раздел 2.1) и PLSA-ITER (см. раздел 2.2). Модуль получает на вход результат работы модуля преобразования входных данных: словари слов и словосочетаний вместе с их частотностями в документах. Также на вход могут подаваться уже построенные тематические модели. Это может происходить в случае работы алгоритма PLSA-ITER.

Вначале отбираются словосочетания для последующего добавления в тематические модели. По умолчанию отбираются самые частотные словосочетания, однако есть возможность предоставить свой список словосочетаний. Стоит отметить, что в случае подачи на вход построенных тематических моделей при работе алгоритма PLSA-ITER отбор осуществляется иначе: словосочетания выбираются путём всевозможных комбинаций из первых 10 слов в каждой теме.

После отбора словосочетаний образующие их слова в каждом документе заменяются этими словосочетаниями. Кроме того, осуществляется формирование множеств похожих слов и словосочетаний на основании идеи в методах PLSA-SIM (см. раздел 2.1) или PLSA-ITER со стеммерами (см. раздел 2.2).

В результате работы данного модуля образуется набор слов и добавленных словосочетаний в том же формате, что и в результате работы модуля преобразования входных данных (см. раздел 4.2.1), а также множества похожих слов и словосочетаний для последующего построения тематических моделей.

4.2.3 Модуль построения инвертированного индекса

Данный модуль предназначен для построения инвертированного индекса Википедии. Это необходимо для подсчёта мер *TC-PMI* (см. раздел 1.1.6) и *TC-PMI-nSIM* (см. раздел 2.5) для оценки качества тематических моделей. Данные меры вычисляют вероятности слов и словосочетаний, учитывая число документов, в которых они встретились во внешнем корпусе (Википедии).

Модуль построения инвертированного индекса получает на вход XML-дамп Википедии[22]. После чего осуществляется разбор этого файла с помощью внешней библиотеки Wikipedia Extractor[23], написанной на языке Python. Результат обработки подаётся на вход внешнему модулю морфологического анализа (см. раздел 4.1.2). После чего осуществляется построение инвертированного индекса, в котором напротив каждой леммы записываются номера документов (статей), в которых данная лемма встретилась.

В результате работы данного модуля образуется инвертированный индекс поданного на вход XML-дампа Википедии, в котором каждой лемме сопоставлены номера статей, в которых она встретилась.

[22] `http://en.wikipedia.org/wiki/Wikipedia:Database_download`

[23] `http://medialab.di.unipi.it/wiki/Wikipedia_Extractor`

4.2.4 Модуль построения тематических моделей

Данный модуль предназначен для построения самих тематических моделей с помощью одного из алгоритмов: *PLSA*, *LDA*, *PLSA-SIM* и *PLSA-ITER*. На вход данный модуль получает словарь слов, возможно с добавленными словосочетаниями, вместе с их частотностями в документах в формате, описанном в разделе 4.2.1. Также в случае запуска алгоритмов PLSA-SIM и PLSA-ITER на вход подаются множества похожих слов и словосочетаний.

В результате работы данного модуля строятся тематические модели. Также есть возможность указать, какие именно автоматические оценки качества следует посчитать: *Перплексию*, *TC-PMI* (см. раздел 1.1.6) или *TC-PMI-nSIM* (см. раздел 2.5). Для подсчёта последних двух мер на вход необходимо также подать предварительно построенный инвертированный индекс внешней коллекции – Википедии (см. раздел 4.2.3). Стоит отметить, что в случае алгоритма PLSA-ITER проводится только одна очередная итерация. После чего тематические модели снова подаются на вход модулю добавления словосочетаний.

4.2.5 Вычислительная сложность алгоритмов PLSA-SIM и PLSA-ITER

Переданный словарь слов и словосочетаний с их частотностями в документах представляется в памяти в виде терм-документной матрицы, строки которой соответствуют документам в коллекции, столбцы – словам и словосочетаниям, а в ячейках записана частотность слов и словосочетаний в каждом документе. Известно, что вычислительная сложность алгоритмов построения тематических моделей PLSA и LDA равна $O(NTi)$, где N – число ненулевых элементов терм-документной матрицы, T – число тем, i – число итераций EM-алгоритма [18, 46]. Пусть K – число добавленных в тематическую модель словосочетаний, W – размер словаря, D – число документов в коллекции. Тогда для вычислительной сложности предложенных алгоритмов PLSA-SIM и PLSA-

ITER справедливо следующее утверждение.

Утверждение 1. *Вычислительные сложности алгоритма PLSA-SIM и одной итерации PLSA-ITER совпадают и равны* $O(NTi + NK + WK \ln K + DK \ln K)$.

Доказательство. Предложенные алгоритмы отличаются от PLSA тем, что добавляют словосочетания и учитывают сходство между ними и образующими их словами (см. разделы 2.1 и 2.2). Данные алгоритмы содержат следующие фазы:

- *Фаза извлечения словосочетаний* с подсчётом их частотностей. Извлечение слов и словосочетаний осуществляется одновременно по регулярным выражениям (см. раздел 4.2.1). Так как такие выражения представляются в виде конечных автоматов, на извлечение каждого словосочетания требуется $O(1)$ времени. Для подсчёта частотностей словосочетания хранятся в виде хеш-таблицы. Поэтому на данную фазу требуется $O(N)$ времени;

- *Фаза добавления словосочетаний*, где отбираются словосочетания, и образующие их слова в документах заменяются на данные словосочетания. В алгоритме PLSA-SIM список словосочетаний для добавления передаётся на входе либо строится путём отбора K самых частотных словосочетаний. Для оптимизации такого отбора используются бинарные кучи с суммарным временем работы $O(W + K \ln W)$. Затем отобранные словосочетания сохраняются в памяти в виде хеш-множества для быстрого поиска. В алгоритме PLSA-ITER отбор словосочетаний происходит путём построения из первых 10 слов в каждой теме всевозможных словосочетаний. Сохранив предварительно все словосочетания с частотностями в виде хеш-таблицы за $O(W)$ времени, данная операция займёт $O(T)$ времени. Таким образом, для данного алгоритма отбор происходит за $O(W + T)$ времени.

  Затем надо заменить в документах слова, образующие добавляемые словосочетания, на сами словосочетания. Для этого в каждом документе просматриваются входящие в него словосочетания и среди них ищутся ото-

бранные. При условии наличия такого словосочетания оно добавляется в словарь, а у образующих его слов уменьшается частотность. На данную операцию тратится $O(1)$ времени при просмотре каждого словосочетания в документе. Таким образом, общие временные затраты данной фазы для алгоритма PLSA-SIM составляют $O(W + K \ln W + N) = O(N + K \ln W)$ (т.к. $W < N$), а для алгоритма PLSA-ITER – $O(W + T + N) = O(N + T)$;

- *Фаза построения множеств похожих слов и словосочетаний.* Данные множества строятся в виде хеш-таблицы, ключами которой являются либо сами слова (для алгоритма PLSA-SIM), либо результат работы стеммера (для алгоритма PLSA-ITER), а значениями – словосочетания, содержащие ключ. На данную фазу потребуется $O(W)$ времени (т.к. число похожих слов и словосочетаний не больше размера всего словаря);

- *Фаза построения тематических моделей методами PLSA-SIM или PLSA-ITER*, где модифицируются частотности похожих слов и словосочетаний и запускаются EM-алгоритмы. Для модификации частотностей надо в каждом документе найти элементы из каждого множества похожих слов и словосочетаний, после чего каждому из них присвоить суммарную частотность всех элементов из найденного множества в документе. Для этого надо просмотреть каждое слово из каждого документа на предмет наличия для него похожих словосочетаний и найти пересечение всех словосочетаний из данного документа с найденными похожими словосочетаниями.

  Для оптимизации времени поиска и пересечения множества похожих слов и словосочетаний сохраняются в виде хеш-таблицы, ключами которой являются слова, а значениями – отсортированные векторы похожих словосочетаний. На такое представление тратится $O(K \ln K)$ времени для каждого слова, т.к. мощность каждого множества похожих слов и словосочетаний не превосходит числа добавленных словосочетаний. При этом суммарно для всех слов потребуется $O(WK \ln K)$ времени. Кроме того, в

каждом документе все словосочетания также сортируются и составляется хеш-множество всех слов в документе. На сортировку требуется $O(K \ln K)$ времени для каждого документа, т.к. число уникальных словосочетаний в документе не превосходит числа добавленных словосочетаний, а для построения хеш-множества всех слов в документе – $O(d)$, где d – длина документа. Суммарно для всех документов требуется $O(DK \ln K)$ на сортировку и $O(N)$ на построение хеш-множеств слов, т.к. суммарные длины документов равны N. Далее на поиск похожих словосочетаний для каждого слова тратится $O(1)$ времени, а на построение пересечения двух отсортированных векторов – линейное время от их суммарной длины, то есть $O(K)$, т.к. каждый из векторов по длине не превосходит числа добавленных словосочетаний. Если пересечение не пусто, то считается его суммарная частотность в документе и присваивается каждому из элементов пересечения. Это осуществляется тоже за линейное время $O(K)$. Таким образом, суммарно для всех вхождений слов в документы на построение пересечений и модификации частотностей требуется $O(NK)$ времени.

Далее запускается оригинальный алгоритм PLSA с временными затратами $O(NTi)$. Таким образом, время выполнения данной фазы равно $O(W + WK \ln K + DK \ln K + N + NK + NTi) = O(NK + WK \ln K + DK \ln K + NTi)$.

Складывая временные затраты всех фаз и учитывая, что $W < N$ и $T < N$, суммарная вычислительная сложность алгоритма PLSA-SIM получится равной $O(N + N + K \ln W + W + WK \ln K + DK \ln K + NK + NTi) = O(NTi + NK + K \ln W + WK \ln K + DK \ln K)$. Поскольку $K \ln W < WK \ln K$ (т.к. $WK \ln K > K \ln W$ и $W \ln K > \ln W$ для любого числа добавляемых словосочетаний и размера словаря), сложность PLSA-SIM равна $O(NTi + NK + WK \ln K + DK \ln K)$. Для одной итерации алгоритма PLSA-ITER при этом получится: $O(N+N+T+W+WK \ln K + DK \ln K + NK + NTi) = O(NTi + NK + WK \ln K + DK \ln K)$. □

Стоит отметить, что на реальных данных выполнены условия: $W << N$ и

$D << N$. Тогда оказывается справедлива следующая теорема.

Теорема 2. *При условиях* $W = O\left(\frac{N}{\ln T}\right)$ *и* $D = O\left(\frac{N}{\ln T}\right)$ *вычислительные сложности алгоритмов PLSA-SIM и одной итерации PLSA-ITER совпадают с вычислительной сложностью алгоритмов PLSA и LDA и равны* $O(NTi)$.

Доказательство. Из доказанного выше утверждения следует, что вычислительные сложности алгоритмов PLSA-SIM и одной итерации PLSA-ITER совпадают и равны $O(NTi + NK + DK \ln K + WK \ln K)$.

Покажем, что для алгоритма PLSA-ITER верно, что $K = O(T)$. Действительно, на каждой итерации данный метод добавляет все словосочетания, образующиеся из первых 10 слов в каждой теме. При этом, если из двух слов можно составить два словосочетания, то выбирается наиболее частотное из них. Т.е. на каждой теме добавляется не более $\frac{10\times(10-1)}{2} = 45$ словосочетаний. Тогда всего будет добавлено $K \leq 45T$ словосочетаний. Т.е. $K = O(T)$.

Для алгоритма PLSA-SIM число добавляемых словосочетаний фиксировано до запуска алгоритма, т.е. $K = const$. Тогда тем более $K = O(T)$.

Исходя из условий теоремы $W = O\left(\frac{N}{\ln T}\right)$ и $D = O\left(\frac{N}{\ln T}\right)$ и найденных оценок для K получим, что $O(NK) = O(NT)$, $O(WK \ln K) = O\left(\frac{N}{\ln T} T \ln T\right) = O(NT)$ и $O(DK \ln K) = O\left(\frac{N}{\ln T} T \ln T\right) = O(NT)$.

Таким образом, вычислительные сложности алгоритмов PLSA-SIM и одной итерации PLSA-ITER оказываются равными $O(NTi+NK+DK \ln K+WK \ln K) = O(NTi + NT + NT + NT) = O(NTi)$. □

Доказанная теорема подтверждает, что предложенные алгоритмы не увеличивают вычислительную сложность оригинальных алгоритмов.

В Таблице 30 представлены результаты сравнения времени работы реализации алгоритма PLSA-SIM с 1000 самых частотных словосочетаний и реализации оригинального алгоритма LDA на языке C++ GibbsLDA++ [79]. Сравнение проводилось на следующем оборудовании: Intel Core i3-2375M 1.5 Ггц, ОЗУ 4 Гб, Ubuntu 15.04. Стоит отметить, что оригинальная реализация алгоритма

LDA[24] не участвовала в сравнении в виду слишком медленной скорости работы (примерно сутки на тестовых коллекциях). Также не представлено время работы одной итерации алгоритма PLSA-ITER, поскольку его вычислительная сложность совпадает со сложностью PLSA-SIM (см. Утверждение 1).

Таблица 30: Сравнение реализаций алгоритма PLSA-SIM и GibbsLDA++

Метод	**Банковский**	**Europarl**	**JRC-Acquiz**	**ACL**
GibbsLDA++	93 мин.	202 мин.	179 мин.	168 мин.
PLSA-SIM	13 мин.	12 мин.	15 мин.	22 мин.

Стоит отметить, что основное ускорение времени работы предложенной в рамках данной диссертационной работы реализации происходит за счёт того, что каждое слово в документах рассматривается на каждой итерации один раз (см. разделы 2.1 и 2.2). В реализации же GibbsLDA++ каждое слововхождение просматривается в отдельности [79].

4.3 Пакет программ извлечения терминов

Данный пакет программ реализует возможности извлечения однословных и двусловных терминов из текстовой коллекции некоторой предметной области. На вход данному пакету программ поступает текстовая коллекция, прошедшая предварительную обработку модулем морфологического анализатора. Архитектурная схема работы модуля представлена на Рисунке 14.

Пакет программ извлечения терминов состоит из трёх условно-независимых модулей: модуля извлечения кандидатов в термины, модуля вычисления признаков и модуля машинного обучения.

[24] `http://www.cs.princeton.edu/~blei/lda-c/index.html`

Рис. 14: Архитектурная схема пакета программ извлечения терминов

4.3.1 Модуль извлечения кандидатов в термины

Данный модуль извлекает слова и словосочетания-кандидаты в термины, а также объемлющие именные группы для последующего вычисления признаков терминологичности. Модуль получает на вход набор лемм с проставленными частями речи и границами предложений, полученный в результате работы модуля морфологического анализатора.

В качестве кандидатов в термины извлекаются:

- Для английского языка – слова и словосочетания в формах: *Существительное*, *Существительное + Существительное*, *Прилагательное + Существительное*, *Существительное + предлог of + Существительное*;
- Для русского языка – слова и словосочетания в формах: *Существительное*, *Прилагательное*, *Прилагательное + Существительное*, *Существительное + Существительное в родительном падеже*.

Отбор кандидатов осуществляется по таким лингвистическим формам, так как они покрывают большую часть однословных и двусловных терминов.

Для найденных кандидатов в термины вычисляются и некоторые «базовые» признаки, необходимые для последующего расчёта остальных признаков:

- **Прочие**: *Прилагательное*, *Существительное*, *Новизна*, *Многозначность*, *Номер первого вхождения в документы*, *NearTermsFreq*, а также признаки для подлежащих, слов с большой буквы и слов с большой буквы, с которых не начинаются предложения (см. раздел 3.3);
- **Основанные на частотности**: *TF*, *DF*, *Domain Consensus*, *Term Variance*, *Term Contribution*, *Term Variance Quality* (см. раздел 1.3.1);
- **Тематические признаки**, посчитанные для «базовой» тематической модели: *TF*, *Maximum TF*, *Term Score*, *Maximum Term Score* (см. раздел 3.2).

Помимо кандидатов в термины также извлекаются именные группы длиной не более трёх слов, описывающиеся регулярным выражением *(Прилагательное|Существительное)*$^+$ *Существительное*. Для каждой такой именной группы также подсчитывается её частотность в текстовой коллекции.

Для улучшения результатов извлечения терминов для русских коллекций проводится фильтрация морфологических омонимов. Во-первых, удаляются те варианты нормализации слов, словоформы которых не согласуются в тексте с соседними словами. Так, для словоформы *банке* из словосочетания *в центральном банке* отбирается только нормальная форма *банк* (но не *банка*). Во-вторых, удаляются слова, нормальные формы которых совпадают с нормальными формами слов других частей речи, так как маловероятно, что они окажутся терминами в данном контексте. Так, слово *том* в словосочетании *в том* будет исключено из-за возможной словоформы *том* местоимения *то*.

Для улучшения результатов извлечения терминов для англоязычных текстовых коллекций предусмотрена возможность подачи на вход списка из стоп-слов. В таком случае из рассмотрения будут исключены все слова из этого списка и все словосочетания, содержащие такие слова.

Также присутствует возможность указания порога частотности, ниже которого кандидаты в термины не рассматриваются. Кроме того, можно указать размер контекстного окна для расчёта признака $NearTermsFreq$.

4.3.2 Модуль вычисления признаков

Данный модуль предназначен для расчёта всех признаков для извлечения терминов. На вход данный модуль принимает извлечённые однословные и двусловные кандидаты с посчитанными упомянутыми выше «базовыми» признаками, а также извлечённые объемлющие именные группы с частотностями. Кроме того, на вход также подаётся контрастная коллекция в виде лемм с частотностями для вычисления признаков, использующих данную коллекцию, и построенные тематические модели для вычисления тематических признаков.

Есть возможность задать пороги самых частотных кандидатов, для которых будут рассчитываться признаки и вычисляться средняя точность.

В результате работы данного модуля образуются списки извлечённых кандидатов в термины вместе с посчитанными признаками. Для каждого признака также выводится средняя точность на заданном пороге.

4.3.3 Модуль машинного обучения

Данный модуль получает на вход списки извлечёных кандидатов в термины с посчитанными признаками, а также «золотой стандарт» для подтверждения терминологичности слов и словосочетаний.

Модуль конвертирует полученные списки кандидатов в представление, необходимое для работы библиотеки scikit-learn[25]. После чего запускается машинное обучение методом градиентного бустинга с применением скользящего контроля. Существует возможность настройки параметров метода градиентного бустинга и числа блоков для скользящего контроля.

[25] `http://scikit-learn.org/stable/`

В результате работы модуля машинного обучения получается итоговый список извлечённых кандидатов в термины, состоящий только из слов или словосочетаний, или из их объединения, вместе со средней точностью.

4.4 Выводы к четвёртой главе

В данной главе приведено описание разработанного программного комплекса, реализующего модели и алгоритмы, предложенные в рамках данной диссертационной работы:

- Новые алгоритмы построения тематических моделей PLSA-SIM и PLSA-ITER, позволяющие добавлять словосочетания и учитывать сходство между ними и образующими их словами;
- Отдельные и комбинированные модели извлечения однословных и двусловных терминов, использующие множество различных признаков, в том числе и новые, основанные на тематических моделях.

Заключение

В ходе данной диссертационной работы были получены следующие результаты:

1. Предложен и реализован новый метод построения тематических моделей, учитывающий словосочетания и улучшающий характеристики качества тематических моделей, включая интерпретацию тем экспертами, что полезно для организации человеко-машинных интерфейсов в информационных системах. Для предложенного метода приведено теоретическое обоснование;

2. Предложен и реализован новый итеративный метод добавления словосочетаний в тематические модели, улучшающий меру соответствия тематических моделей словам и словосочетаниям текстовых коллекций (перплексию). Для предложенных методов приводятся теоретические оценки вычислительной сложности;

3. Предложены новые признаки для извлечения терминов, основанные на тематических моделях. Показано, что использование тематической информации улучшает качество извлечения терминов для включения их в базы знаний и терминологические ресурсы;

4. Разработан и выложен в открытый доступ программный комплекс по построению тематических моделей с использованием лексико-терминологической информации.

Дальнейшие перспективы развития исследований могут быть связаны с использованием предложенных моделей построения тематических моделей для более качественного решения различных задач информационного поиска, включая кластеризацию и классификацию текстов, многодокументное аннотирование, разрешение морфологической неоднозначности.

Кроме того, разработанные модели извлечения однословных и двусловных терминов могут применяться для автоматического пополнения существующих баз знаний и терминологических ресурсов (словарей, тезаурусов) различных предметных областей, которые, в свою очередь, могут использоваться для улучшения качества информационного поиска, машинного перевода, автоматического аннотирования и в других задачах.

Стоит отметить, что разработанный программный комплекс по построению тематических моделей с использованием лексико-терминологической информации выложен в открытый доступ, что существенно упрощает дальнейшие исследования.

Список литературы

1. Большакова Е. И., Лукашевич Н. В., Нокель М. А. Извлечение однословных терминов из текстовых коллекций на основе методов машинного обучения // Информационные технологии. – 2013. – № 7. – С. 31-37.

2. Воронцов К. В. Математические методы обучения по прецедентам (теория обучения машин) // Курс лекций ВМК МГУ и МФТИ. – 2011.

3. 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.

4. Лукашевич Н. В. Тезаурусы в задачах информационного поиска // Издательство Московского университета, 2011.

5. Лукашевич Н. В., Логачев Ю. М. Использованием методов машинного обучения для извлечения слов-терминов // Труды 12й Национальной Конференции по Искусственному Интеллекту с Международным Участием (КИИ'2010). – 2010.

6. Нокель М. А. Использование лингвистической информации в тематической модели PLSA // Сборник тезисов XXI Международной научной конференции студентов, аспирантов и молодых учёных «Ломоносов-2014». Секция «Вычислительная Математика и Кибернетика». – 2014. – С. 120-121.

7. Нокель М. А. Метод учёта структуры биграмм в тематических моделях // Вестник ВГУ, Серия: Системный анализ и информационные технологии. – 2014. – № 4. – С. 89-97.

8. Нокель М. А., Лукашевич Н. В. Тематические модели в задаче извлечения однословных терминов // Программная инженерия. – 2014. – № 3. – С. 34-40.

9. Нокель М. А., Лукашевич Н. В. Тематические модели: добавление биграмм и учет сходства между униграммами и биграммами // Вычислительные методы и программирование. – 2015. – Т. 16, № 2. – 2015. – С. 215-234.

10. Павлов А. С., Добров Б. В. Метод обнаружения массово порождённых неестественных текстов на основе анализа тематической структуры // Вычислительные методы и программирование. – 2011. – Т. 12. – С. 58-72.

11. Ahmad K., Gillam L., Tostevin L. University of Survey participation in TREC8: Weirdness indexing for logical document extrapolation and retrieval // Proceedings of the 8th Text Retrieval Conference. – 2007. – P. 717-724.

12. Ananiadou S. A Methodology for Automatic Term Recognition // Proceedings of the 15th International Conference on Computational Linguistics (COLING'94). – 1994. – P. 1034-1038.

13. Andrzejewski D., Butter D. Latent topic feedback for information retrieval // Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – 2011. – P. 600-608.

14. Asuncion A., Welling M., Smyth P., Teh Y. W. On smoothing and inference for topic models // Proceedings of the International Conference on Uncertainty in Artificial Intelligence. – 2009. – P. 27-34.

15. Azé J., Roche M., Kodratoff Y., Sebag M. Preference Learning in Terminology Extraction: A ROC-based approach // Proceedings of Applied Stochastic Models and Data Analysis. – 2005. – P. 209-219.

16. Basili R., Moschitti A., Pazienza M., Zanzotto F. A Contrastive Approach to Term Extraction // Proceedings of the 4th Terminology and Artificial Intelligence Conference (TIA). – 2001. – P. 119-128.

17. Blei D., Lafferty J. Topic Models // Text Mining: Classification, Clustering and Applications. – Chapman & Hall, 2009. – P. 71-89.

18. Blei D., Ng A., Jordan M. Latent Dirichlet Allocation // Journal of Machine Learning Research. – MIT Press, 2003. – No. 3. – P. 993-1002.

19. Bolelli L., Ertekin Ş., Giles C. L. Topic and Trend Detection in Text Collections Using Latent Dirichlet Allocation // ECIR Proceedings, Lecture Notes in Computer Science. – 2009. – Vol. 5478. – P. 776-780.

20. Bolshakova E., Loukachevitch N., Nokel M. Topic Models Can Improve Domain Term Extraction // ECIR Proceedings. Серия LNCS. – Издательство SPRINGER HEIDELBERG, 2013. – Т. 7814. – С. 684-687.

21. Bouma G. Normalized (Pointwise) Mutual Information in Collocation Extraction // Proceedings of the Biennal GSCL Conference. – 2009. – P. 31-40.

22. Bourigault D. Surface grammatical analysis for the extraction of terminological noun phrases // Proceedings of COLING-92. – 1992. – P. 977-981

23. Boyd-Grabber J., Blei D., Zhu X. A Topic Model for Word Sense Disambiguation // Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Language Learning. – 2007. – P. 1024-1033.

24. Chang J., Boyd-Grabber J., Wang C., Gerrich S., Blei D. Reading tea leaves: How human interpret topic models // Proceedings of the 24th Annual Conference on Neural Information Processing Systems. – 2009. – P. 288-296.

25. Church K., Gale W. Inverse Document Frequency IDF: A Measure of Deviation from Poisson // Proceedings of the Third Workshop on Very Large Corpora. – 1995. – P. 121-130.

26. Church K., Hanks P. Word Association Norms, Mutul Information, and Lexicography // Computational Linguistics. – 1990. – Vol. 16. – P. 22–29.

27. Daille B. Combined Approach for Terminology Extraction: Lexical Statistics and Linguistic Filtering // PhD Dissertation. – University of Paris, 1995.

28. Daud A., Li J., Zhou L., Muhammad F. Knowledge discovery through directed probabilistic topic models: a survey // Frontiers of Computer Science in China. – Higher Education Press, 2010. – Vol. 2, № 2. – P. 280-301.

29. Daudarvičius V., Marcinkevičienė R. Gravity Counts for the Boundaries of Collocations // Corpus Linguistics. – 2004. – Vol. 9, № 2. – P. 321-348.

30. Deane P. A Nonparametric Method for Extraction of Candidate Phrasal Terms // Proceedings of the 43rd Annual Meeting of the ACL. – 2005. – P. 605-613.

31. Dempster A. P., Laird N. M., Rubin D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm // Journal of the Royal Statistical Society. Series B. – 1977. – Vol. 39, № 1. – P. 1-38.

32. Dhillon I., Kogan J., Nicholas C. Feature Selection and Document Clustering // Springer-Verlag. – 2003. – P. 73-100.

33. Ding C., Li T., Peng W. On the equivalence between Non-negative Matrix Factorization and Probabilistic Latent Semantic Indexing // Computational Statistics and Data Analysis. – 2008. – № 52. – P. 3913-3927.

34. Dobrov B., Loukachevitch N. Multiple Evidence for Term Extraction in Broad Domains // Proceedings of RANLP 2011. – 2011. – P. 710-715.

35. Drake M. Encyclopedia of Library and Information Science // CRC Press, 2003.

36. Dunning T. Accurate Methods for the Statistics of Surprise and Coincidence // Computational Linguistics. – 1993. – Vol. 19, № 1. – P. 61-74.

37. Eidelman V., Boyd-Grabber J., Resnik P. Topic models for dynamic translation model adaptation // Proceedings of the 50th Annual Meeting of the Association of Computational Linguistics. – 2012. – Vol. 2. – P. 115–119.

38. Foo J., Merkel M. Using Machine Learning to Perform Automatic Term Recognition // Proceedings of the LREC 2010 Acquisition Workshop. – 2010. – P. 49-54.

39. Frantzi K., Ananiadou S. Automatic Term Recognition Using Contextual Cues // Proceedings of the IJCAI Workshop on Multilinguality in Software Industry: the AI Contribution. – 1997. – P. 73-80.

40. Frantzi K., Ananiadou S. The C-value/NC-value Domain-independent Method for Multi-word Term Extraction // Journal of Natural Language Processing. – 1999. – Vol. 6, № 3. – P. 145-179.

41. Friedman J. H. Greedy Function Approximation: A Gradient Boosting Machine // Annals of Statistics. – 2000. – Vol. 29. – P. 1189-1232.

42. Gelbukh A., Sidorov G., Lavin-Villa E., Chanona-Hernandez L. Automatic Term Extraction using Log-likelihood based Comparison with General Reference Corpus // Proceedings of the Natural Language Processing and Information Systems, and the 15th International Conference on Applications of Natural Language to Information Systems. – 2010. – P. 248-255.

43. Georgantopoulos B., Piperidis S. Term-based identification of sentences for text summarization // Proceedings of the 2nd International Conference on Language Resources and Evaluation. – 2000. – P. 1067-1070.

44. Griffiths T., Steyvers M., Tenenbaum J. Topics in semantic representation // Psychological Review. – American Psychological Association, 2007. – Vol. 114, № 2. – P. 211-244.

45. He Q., Chang K., Lim E., Banerjee A. Keep It Smile with Time: A Reexamination of Probabilistic Topic Detection Models // Proceedings of IEEE Transaction Pattern Analysis and Machine Intelligence. – 2010. – Vol. 32, № 10. – P. 1795–1808.

46. Hofmann T. Probabilistic Latent Semantic Indexing // Proceedings of the 22nd Annual Internationl SIGIR Conference on Research and Development in Information Retrieval. – 1999. – P. 50-57.

47. Hu W., Shimizu N., Nakagawa H., Shenq H. Modeling Chinese documents with topical word-character models // Proceedings of the 22nd International Conference on Computational Linguistics. – 2008. – P. 345-352.

48. Hyunh T., Fritz M., Schiele B. Discovery of activity patterns using topic models // Proceedings of the 10th International Conference on Ubiquitous Computing. – 2008. – P. 10-19.

49. Jaccard P. Distribution de la flore alpine dans le Bassin des Drances et dans quelques regions voisines // Bull. Soc. Vaudoise sci. Natur. – 1901. – V. 37, Bd. 140. – P. 241-272.

50. Johnson S. C. Hierarchical Clustering Schemes // Psychometrica. – 1967. – № 2. – P. 241-254.

51. Kitamura M., Matsumoto Y. Automatic Extraction of Word Sequence Correspondences in Parallel Corpora // Proceedings of the 4th Annual Workshop on Very Large Corpora. – 1996. – P. 79-87.

52. Kurz D., Xu F. Text Mining for the Extraction of Domain Relevant Terms and Term Collocations // Proceedings of the International Workshop on Computational Approaches to Collocations. – 2002.

53. Lau J. H., Baldwin T., Newman D. On Collocations and Topic Models // ACM

Transactions on Speech and Language Processing. – ACM Press, 2013. – Vol. 10, № 3. – P. 1-14.

54. Lee D. D., Seung H. S. Algorithms for Non-negative Matrix Factorization // Proceedings of NIPS. – 2000. – P. 556-562.

55. Li S., Li J., Song T., Li W., Chang B. A Novel Topic Model for Automatic Term Extraction // Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. – 2013. – P. 885-888.

56. Liu L., Kang J., YU J., Wang Z. A Comparative Study on Unsupervised Feature Selection Methods for Text Clustering // Proceedings of NLP-KE'05. – 2005. – P. 201-219.

57. Lloyd S. P. Least square quantization in PCM // Bell Telephone Laboratories Paper. – 1982.

58. Lopes J. G. P., Silva J. F. A Local Maxima Method and a Fair Dispersion Normalization for Extracting Multiword Units // Proceedings of the 6th Meeting on the Mathematics of Language. – 1999. – P. 369-381.

59. MacKay D. J. C., Peto L. C. B. A Hierarchical Dirichlet Language Model // Natural Language Engineering. – 1995. – Vol. 1. – P. 289-307.

60. MacQueen J. B. Some Methods for Classification and Analysis of Multivariate Observations // Proceedings of the 5th Berkeley Symposium on Mathematical Statisics and Probability. – University of California Press, 2009. – P. 281-297.

61. Manning C. D. Introduction to Information Retrieval // MIT Press, 2008.

62. Manning C. D., Schutze H. Foundations of Statistical Natural Language Processing // The MIT Press, 1999.

63. Mimno D., Wallach H., Talley E., Leenders M., McCallum A. Optimizing semantic coherence in topic models // Proceedings of EMNLP'2011. – 2011. – P. 262-272.

64. Nakagawa H., Mori T. A Simple but Powerful Automatic Term Extraction Method //Procedings of the 2nd International Workshop on Computational Terminology (COMPUTERM'02). – 2002. – P. 29-35.

65. Nakagawa H., Mori T. Automatic Term Recognition based on Statistics of Compound Nouns and their Components // Terminology. – 2003. – Vol. 9, № 2. – P. 201-219.

66. Nakagawa H., Mori T. Nested Collocation and Compound Noun for Term Recognition // Proceedings of the First Workshop on Computational Terminology COMPTERM'98. – 1998. – P. 64-70.

67. Navigli R., Velardi P. Semantic Interpretation of Terminological Strings // Proceedings of the 6th International Conference on Terminology and Knowledge Engineering (TKE). – 2002. – P. 95-100.

68. Newman D., Bonilla E. V., Buntine W. Improving Topic Coherence with Regularized Topic Models // Proceedings of NIPS. – 2011. – P. 1-9.

69. Newman D., Lau J., Grieser K., Baldwin T. Automatic evaluation of topic coherence // Proceedings of Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics. – 2010. – P. 100-108.

70. Nokel M. Topic models: Taking into account similarity between unigrams and bigrams // CEUR Workshop Proceedings. – 2014. – Vol. 1297. – P. 243–252.

71. Nokel M., Bolshakova E., Loukachevitch N. Combining Multiple Features For Single-Word Term Extraction // Компьютерная лингвистика и интеллекту-

альные технологии. По материалам конференции Диалог-2012. – 2012. – С. 490-501.

72. Nokel M., Loukachevitch N. A Method of Accounting Bigrams in Topic Models // Proceedings of North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL-HLT). – 2015. – P. 1-9.

73. Nokel M., Loukachevitch N. An Experimental Study of Term Extraction for Real Information-Retrieval Thesauri // Proceedings of the 10th International Conference on Terminology and Artificial Intelligence (TIA-2013). – 2013. – P. 69-76.

74. Nokel M., Loukachevitch N. Application of topic models to the task of single-word term extraction // CEUR Workshop Proceedings. – 2013. – Vol. 1108. – P. 52-60.

75. Nokel M., Loukachevtich N. Topic Models: Accounting Component Structure of Bigrams // Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA-2015). – 2015. – P. 145-152.

76. Paice C. D. Another Stemmer // SIGIR Forum. – 1990. – Vol. 24, № 3. – P. 56-61.

77. Pecina P., Schlesinger P. Combining Association Measures for Collocation Extraction // Proceedings of the COLING/ACL 2006. – ACL Press, 2006. – P. 651-658.

78. Peñas A., Verdejo F., Gonzalo J. Corpus-based terminology extraction applied to information access // Proceedings of Corpus Linguistics. – 2001. – P. 458-465.

79. Phan X.-H., Nguyen C.-T. GibbsLDA++: A C/C++ implementation of latent Dirichlet allocation (LDA) // Technical report. – 2007.

80. Porter M. F. An algorithm for suffix stripping // Program. – 1980. – Vol. 14, № 3. – P. 130-137.

81. Salton G. Automatic text processing: the transformation, analysis and retrieval of information by computer // Addison-Wesley. – 1989.

82. Sclano F., Velardi P. TermExtractor: a Web Application to Learn the Common Terminology of Interest Groups and Research Communities // Proceedings of the 9th Conference on Terminology and Artificial Intelligence TIA 2007. – 2007.

83. Siegel S. Non-parametric statistics for the behavioral sciences // McGraw-Hill. – New-York, 1956. – P. 75-83.

84. Smadja F, McKeown K., Hatzivassiloglou V. Translating Collocations for Bilingual Lexicons: A Statistical Approach // Computational Linguistics. – 1996. – Vol. 22, № 1. – P. 1-38.

85. Smeeton N. C. Early History of the Kappa Statistic // Biometrics. – 1985. – Vol. 41, № 3. – P. 795.

86. Sparck Jones K. A Statistical Interpretation of Term Specificity and its application in retrieval // Journal of Documentation. – 1972. – № 28. – P. 11-21.

87. Stevens K., Kegelmeyer P., Andrzejewski D., Butter D. Exploring topic coherence over many models and many topics // Proceedings of EMNLP-CoNLL'2012. – 2012. – P. 952-961.

88. Vivaldi J., Màrquez L., Rodríguez H. Improving Term Extraction by System Combination Using Boosting // Proceedings of ECML. – 2001. – P. 515-526.

89. Vu T., Aw A. Ti, Zhang M. Term Extraction Through Unithood and Termhood Unification // Proceedings of IJCNLP. – 2008. – P. 631-636.

90. Wallach H. Topic Modeling: Beyond Bag-Of-Words // Proceedings of the 23rd International Conference on Machine Learning. – 2006. – P. 977-984.

91. Wang D., Zhu S., Li T., Gong Y. Multi-Document Summarization using Sentence-Based Topic Models // Proceedings of the ACL-IJCNLP Conference Short Papers. – 2009. – P. 297-300.

92. Wang X., McCallum A., Wei X. Topical n-grams: Phrase and topic discovery, with an application to information retrieval // Proceedings of the 7th IEEE International Conference on Data Mining. – 2007. – P. 697-702.

93. Wei X., Croft B. LDA-based document models for ad-hoc retrieval // Proceedings of the 29th International ACM-SIGIR Conference on Research and Development in Information Retrieval. – 2006. – P. 178-185.

94. Wolf P., Bernardi U., Federmann C., Hunsicker S. From Statistical Term Extraction to Hybrid Machine Translation // Proceedings of the 15th Conference of the European Association for Machine Translation. – 2011. – P. 225-232.

95. Wong W., Liu W., Bennamoun M. Determining Termhood for Learning Domain Ontologies using Domain Prevalence and Tendency // Proceedings of the 6th Australasion Conference on Data Mining (AusDM). – 2007. – P. 47-54.

96. Xu W., Liu X., Gong Y. Document Clustering Based On Non-negative Matrix Factorization // Proceedings of SIGIR. – 2003. – P. 267-273.

97. Yeh J.-H., Wu M.-L. Recommendation based on latent topics and social network analysis // Proceedings of the 2nd International Conference on Computer Engineering and Applications. – IEEE Computer Society, 2010. – Vol. 1. – P. 209-213.

98. Zhang W., Yoshida T., Ho T., Tang X. Augmented Mutual Information for Multi-Word Term Extraction // International Journal of Innovative Computing, Information and Control. – 2008. – Vol. 8, № 2. – P. 543-554.

99. Zhang Z., Iria J., Brewster C., Ciravegna F. A Comparative Evaluation of Term

Recognition Algorithms // Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008). – 2008. – P. 2108-2113.

100. Zhong S. Efficient Online Spherical K-Means Clustering // Proceedings of IEEE-IJCNN. – 2005. – P. 3180-3185.

101. Zhou S., Li K., Liu Y. Text Categorization Based on Topic Model // International Journal of Computational Intelligence Systems. – Atlantis Press, 2009. – Vol. 2, № 4. – P. 398-409.

Приложение А

Список первых 10 слов из тем, полученных алгоритмом PLSA на банковском корпусе

ТЕМА 1	**ТЕМА 2**	**ТЕМА 3**
КАРТА 0.107	БЕЗОПАСНОСТЬ 0.029	БАНК 0.048
КАРТ 0.084	ИНТЕРНЕТ 0.016	БАНКА 0.028
ПЛАТЕЖНЫЙ 0.024	ИНФОРМАЦИОННЫЙ 0.015	БЫТЬ 0.022
БАНКОМАТ 0.015	ЗАЩИТА 0.015	НОВЫЙ 0.012
СИСТЕМА 0.013	ИНФОРМАЦИЯ 0.014	ГОД 0.011
ПЛАСТИКОВЫЙ 0.013	СИСТЕМА 0.013	РАЗВИТИЕ 0.01
ИСПОЛЬЗОВАНИЕ 0.013	ЭЛЕКТРОННЫЙ 0.012	БАЗЕЛЬ 0.01
БАНКОВСКИЙ 0.009	СЕТЬ 0.01	УЗКИЙ 0.009
КАРТОЧНЫЙ 0.009	ДОСТУП 0.01	УЖЕ 0.009
УСЛУГА 0.008	ДАТЬ 0.009	УЖ 0.009
ТЕМА 4	**ТЕМА 5**	**ТЕМА 6**
НЕФТЬ 0.014	СИСТЕМА 0.014	ЗАТРАТЫ 0.037
РОССИЯ 0.012	ЯВЛЯТЬСЯ 0.011	УЧЕТ 0.024
ГАЗ 0.01	ПРОЦЕСС 0.008	РАСХОД 0.018
НЕФТЯНОЙ 0.009	ВИД 0.007	ПРОДУКЦИЯ 0.018
СТРАНА 0.009	ДАТЬ 0.007	СЕБЕСТОИМОСТЬ 0.014
ДОБЫЧА 0.008	ФАКТОР 0.007	ПРОИЗВОДСТВО 0.012
КОМПАНИЯ 0.008	РЕСУРС 0.007	ЗАТРАТА 0.01
БЫТЬ 0.007	ДАННЫЙ 0.007	ПРЕДПРИЯТИЕ 0.009
ЗОНА 0.007	ПОНЯТИЕ 0.006	ОСНОВНОЙ 0.009
ПРОЕКТ 0.006	МЕТОД 0.006	ОСНОВНЫЙ 0.009
ТЕМА 7	**ТЕМА 8**	**ТЕМА 9**
БЮДЖЕТНЫЙ 0.041	ПРЕДПРИЯТИЕ 0.023	КЛИЕНТ 0.088
БЮДЖЕТ 0.033	ДЕЯТЕЛЬНОСТЬ 0.019	БАНК 0.05
ГОСУДАРСТВЕННЫЙ 0.026	ПРОЦЕСС 0.018	УСЛУГА 0.048
ПОЛИТИКА 0.025	АНАЛИЗ 0.018	БАНКА 0.028
ФИНАНСОВЫЙ 0.022	ОЦЕНКА 0.014	ПРОДУКТ 0.026
СРЕДСТВА 0.016	СИСТЕМА 0.013	ОБСЛУЖИВАНИЕ 0.023
СРЕДСТВО 0.016	ИНФОРМАЦИЯ 0.011	БАНКОВСКИЙ 0.016
РАСХОД 0.01	РЕЗУЛЬТАТ 0.011	ОФИС 0.014
ПРОГРАММА 0.01	ЦЕЛЬ 0.01	РОЗНИЧНЫЙ 0.012
ЦЕЛЬ 0.01	ЭТАП 0.009	НОВЫЙ 0.011

ТЕМА 10	**ТЕМА 11**	**ТЕМА 12**
РОССИЯ 0.034	СРЕДСТВА 0.046	ГОД 0.053
РОССИЙСКИЙ 0.017	СРЕДСТВО 0.046	БАНК 0.029
СНГ 0.016	РЕЗЕРВ 0.034	БАНКА 0.017
БЫТЬ 0.015	ОРГАНИЗАЦИЯ 0.028	РАБОТАТЬ 0.015
СТРАНА 0.014	ДОХОД 0.025	ДОЛЖНОСТЬ 0.014
УКРАИНА 0.014	ОБЯЗАТЕЛЬСТВО 0.021	УПРАВЛЕНИЕ 0.013
СОВЕТ 0.013	КРЕДИТНЫЙ 0.02	ДИРЕКТОР 0.013
ПРИНЯТЬ 0.012	ЧИСТЫЙ 0.019	НАЗНАЧИТЬ 0.011
РЕСПУБЛИКА 0.011	ПРОЦЕНТ 0.016	БЫТЬ 0.01
СОГЛАШЕНИЕ 0.011	СОБСТВЕННЫЙ 0.016	НАЗНАЧЕННЫЙ 0.01
ТЕМА 13	**ТЕМА 14**	**ТЕМА 15**
ЭКОНОМИКА 0.041	СИСТЕМА 0.075	БЫТЬ 0.024
ИНВЕСТИЦИОННЫЙ 0.038	ПЛАТЕЖНЫЙ 0.043	РОССИЯ 0.014
РАЗВИТИЕ 0.03	ПЛАТЕЖ 0.037	ГОД 0.011
ИНВЕСТИЦИЯ 0.025	РАСЧЕТ 0.033	РФ 0.009
ЭКОНОМИЧЕСКИЙ 0.023	ЭЛЕКТРОННЫЙ 0.022	ЗАКОНОПРОЕКТ 0.008
РЫНОК 0.018	БАНК 0.018	ГОСДУМА 0.007
ФИНАНСОВЫЙ 0.017	ЦЕНТРАЛЬНЫЙ 0.015	ПРАВИТЕЛЬСТВО 0.007
КАПИТАЛ 0.014	УЧАСТНИК 0.014	ЗАКОН 0.007
РЕСУРС 0.013	РАСЧЕТНЫЙ 0.011	МНЕНИЕ 0.007
РОСТ 0.013	БЫТЬ 0.008	ПРЕДСЕДАТЕЛЬ 0.006
ТЕМА 16	**ТЕМА 17**	**ТЕМА 18**
СУД 0.034	ПРЕДСЕДАТЕЛЬ 0.082	УПРАВЛЕНИЕ 0.029
ДЕЛО 0.015	ПРАВЛЕНИЕ 0.074	СТРАТЕГИЯ 0.019
СУДЕБНЫЙ 0.014	БАНК 0.038	ПРОЦЕСС 0.014
РФ 0.013	ЗАМЕСТИТЕЛЬ 0.034	СТРАТЕГИЧЕСКИЙ 0.014
РЕШЕНИЕ 0.013	ДИРЕКТОР 0.03	ОРГАНИЗАЦИЯ 0.014
ОРГАН 0.013	ОАО 0.017	ЦЕЛЬ 0.011
БЫТЬ 0.012	ЗАО 0.016	МЕНЕДЖМЕНТ 0.01
ЛИЦО 0.012	КБ 0.016	ДЕЯТЕЛЬНОСТЬ 0.008
АРБИТРАЖНЫЙ 0.012	АЛЕКСАНДР 0.015	БИЗНЕС 0.008
НАРУШЕНИЕ 0.011	НАЧАЛЬНИК 0.014	ЗАДАЧА 0.008
ТЕМА 19	**ТЕМА 20**	**ТЕМА 21**
ИНФЛЯЦИЯ 0.04	ФИНАНСОВЫЙ 0.079	БЫТЬ 0.022
ЦЕНА 0.024	РЕЙТИНГ 0.03	ЧЕЛОВЕК 0.012
ПОЛИТИКА 0.016	ИНФОРМАЦИЯ 0.018	ГОД 0.01
РОСТ 0.014	АГЕНТСТВО 0.017	ЛЮДИ 0.007
ЭКОНОМИЧЕСКИЙ 0.013	ОЦЕНКА 0.016	СТАТЬ 0.007
ЭКОНОМИКА 0.012	РЕЙТИНГОВЫЙ 0.011	КОМАНДА 0.006
ИНФЛЯЦИОННЫЙ 0.008	МОЧЬ 0.01	КОГДА 0.005
УРОВЕНЬ 0.006	ЯВЛЯТЬСЯ 0.009	ВРЕМЯ 0.005
ВОЗДЕЙСТВИЕ 0.006	ДЕЯТЕЛЬНОСТЬ 0.009	РЕБЕНОК 0.005
РЫНОК 0.005	АНАЛИЗ 0.008	ЖИЗНЬ 0.004

ТЕМА 22	**ТЕМА 23**	**ТЕМА 24**
СОТРУДНИК 0.048	СТОИМОСТЬ 0.025	НАЛОГОВЫЙ 0.098
РАБОТА 0.042	ЦЕНА 0.02	НАЛОГ 0.058
СПЕЦИАЛИСТ 0.027	ТОВАР 0.017	НАЛОГОПЛАТЕЛЬЩИК 0.028
ПЕРСОНАЛ 0.02	ДОХОД 0.012	НАЛОГООБЛОЖЕНИЕ 0.016
РУКОВОДИТЕЛЬ 0.018	ЯВЛЯТЬСЯ 0.01	УПЛАТА 0.015
ОБУЧЕНИЕ 0.017	ЭКОНОМИЧЕСКИЙ 0.01	ОРГАН 0.015
ПОДРАЗДЕЛЕНИЕ 0.012	ПРИБЫЛЬ 0.009	РФ 0.013
ПРОГРАММА 0.011	МОЧЬ 0.009	ОРГАНИЗАЦИЯ 0.013
ОТДЕЛ 0.01	РЫНОЧНЫЙ 0.008	ПРИБЫЛЬ 0.01
РАБОТНИК 0.009	ВРЕМЯ 0.007	ДЕКЛАРАЦИЯ 0.01
ТЕМА 25	**ТЕМА 26**	**ТЕМА 27**
БЫТЬ 0.009	МОСКВА 0.078	ГОД 0.016
ТЕАТР 0.007	ОАО 0.074	США 0.015
ДЕНЬ 0.005	ЗАО 0.033	БЫТЬ 0.014
ГОД 0.005	КБ 0.032	АМЕРИКАНСКИЙ 0.013
ВРЕМЯ 0.004	НПФ 0.03	СТРАНА 0.009
ЖИЗНЬ 0.003	ООО 0.028	СТАТЬ 0.007
НОВЫЙ 0.003	БАНК 0.026	МИР 0.007
ГДЕ 0.003	АКБ 0.022	МИРО 0.006
ПРАЗДНИК 0.003	ОБЛАСТЬ 0.019	РОССИЯ 0.006
СТАТЬ 0.003	САНКТ-ПЕТЕРБУРГ 0.012	УЖ 0.005
ТЕМА 28	**ТЕМА 29**	**ТЕМА 30**
БАНК 0.04	КРЕДИТ 0.023	МОЧЬ 0.017
МОСКВА 0.026	БАНК 0.022	ПРОБЛЕМА 0.013
РОССИЯ 0.017	БАНКА 0.013	БЫТЬ 0.012
БАНКА 0.017	РЫНОК 0.013	СЛУЧАЙ 0.011
ЛИЦЕНЗИЯ 0.016	КРЕДИТОВАНИЕ 0.011	РЕШЕНИЕ 0.009
ТРЕБОВАНИЕ 0.016	БЫТЬ 0.011	МОЖНО 0.009
КРЕДИТОР 0.016	ПОТРЕБИТЕЛЬСКИЙ 0.01	ДАТЬ 0.008
КОНКУРСНЫЙ 0.015	ДОЛГ 0.009	НЕОБХОДИМЫЙ 0.007
ООО 0.014	АГЕНТСТВО 0.008	СИСТЕМА 0.007
УПРАВЛЯЮЩИЙ 0.013	ГОД 0.008	ДАННЫЙ 0.007
ТЕМА 31	**ТЕМА 32**	**ТЕМА 33**
РЫНОК 0.074	СТАНДАРТ 0.046	БАНК 0.411
СДЕЛКА 0.036	МЕЖДУНАРОДНЫЙ 0.027	БАНКА 0.225
ИНСТРУМЕНТ 0.019	УЧЕТ 0.02	РОССИЯ 0.076
ФИНАНСОВЫЙ 0.017	ОТЧЕТНОСТЬ 0.019	КРЕДИТНЫЙ 0.047
КОНТРАКТ 0.016	ОРГАНИЗАЦИЯ 0.018	БАНКОВСКИЙ 0.036
УЧАСТНИК 0.014	ПРОФЕССИОНАЛЬНЫЙ 0.014	ОРГАНИЗАЦИЯ 0.021
ОПЕРАЦИЯ 0.014	ФИНАНСОВЫЙ 0.012	КОММЕРЧЕСКИЙ 0.011
БИРЖА 0.013	БЫТЬ 0.011	ЦЕНТРАЛЬНЫЙ 0.01
СРОЧНЫЙ 0.012	БУХГАЛТЕРСКИЙ 0.01	БАНКИЙ 0.009
ТОРГИ 0.01	ДЕЯТЕЛЬНОСТЬ 0.01	УЧРЕЖДЕНИЕ 0.009

ТЕМА 34	**ТЕМА 35**	**ТЕМА 36**
ГОД 0.05	СТРАНА 0.038	БЫТЬ 0.011
РОСТ 0.041	МИГРАНТ 0.014	ЧЕЛОВЕК 0.011
ТЕМП 0.02	РАЗВИВАЮЩИЙСЯ 0.013	ЖЕНЩИНА 0.01
ОБЪЕМ 0.017	РАЗВИВАТЬСЯ 0.012	МУЖЧИНА 0.006
ДОЛЯ 0.014	РАЗВИТИЕ 0.009	ЛЮДИ 0.005
РОССИЯ 0.013	ИНДИЯ 0.008	ЦВЕТ 0.005
СОСТАВИТЬ 0.012	ИНДИЙ 0.007	ОДЕЖДА 0.004
РУБЛЬ 0.012	БЫТЬ 0.006	МОЖНО 0.004
МЛРД 0.012	ФИНАНСОВЫЙ 0.006	МОЧЬ 0.004
РЫНОК 0.01	ГОД 0.006	СТИЛЬ 0.003
ТЕМА 37	**ТЕМА 38**	**ТЕМА 39**
ЭКОНОМИЧЕСКИЙ 0.018	ОТЧЕТНОСТЬ 0.037	СТРАНА 0.06
ЭКОНОМИКА 0.014	ФИНАНСОВЫЙ 0.03	США 0.017
ГОСУДАРСТВО 0.014	АКТИВ 0.027	ГЕРМАНИЯ 0.012
ОБЩЕСТВО 0.009	БУХГАЛТЕРСКИЙ 0.025	ГЕРМАНИЙ 0.011
СТРАНА 0.008	УЧЕТ 0.023	ФРАНЦИЯ 0.011
ПОЛИТИЧЕСКИЙ 0.008	СТОИМОСТЬ 0.02	ФРАНЦИЙ 0.01
РЕФОРМА 0.007	ОБЯЗАТЕЛЬСТВО 0.011	ВЕЛИКОБРИТАНИЯ 0.01
СИСТЕМА 0.007	ПРИБЫЛЬ 0.011	ЕВРОПА 0.009
БЫТЬ 0.007	ОРГАНИЗАЦИЯ 0.01	ЕВРОПЕЙСКИЙ 0.008
СОЦИАЛЬНЫЙ 0.007	ОТЧЕТНЫЙ 0.01	РАЗВИТЫЙ 0.008
ТЕМА 40	**ТЕМА 41**	**ТЕМА 42**
ВАЛЮТНЫЙ 0.052	ФЕДЕРАЛЬНЫЙ 0.056	ДОХОД 0.03
ИНОСТРАННЫЙ 0.033	ФЕДЕРАЦИЯ 0.054	ФОНД 0.015
РЕЗИДЕНТ 0.032	РОССИЙСКИЙ 0.053	БЮДЖЕТ 0.014
ОПЕРАЦИЯ 0.028	ЗАКОН 0.04	ЦЕНА 0.012
ВАЛЮТА 0.028	РФ 0.035	РАСХОД 0.012
НЕРЕЗИДЕНТ 0.024	ОРГАН 0.029	УРОВЕНЬ 0.012
СЧЕТ 0.018	ГОСУДАРСТВЕННЫЙ 0.022	ВВП 0.012
РФ 0.016	ДЕЯТЕЛЬНОСТЬ 0.02	ГОД 0.011
УПОЛНОМОЧЕННЫЙ 0.015	ЗАКОНОДАТЕЛЬСТВО 0.013	БЫТЬ 0.009
РОССИЙСКИЙ 0.015	АКТ 0.012	НАСЕЛЕНИЕ 0.008
ТЕМА 43	**ТЕМА 44**	**ТЕМА 45**
АССОЦИАЦИЯ 0.032	РЕГИОН 0.051	РЫНОК 0.031
РОССИЯ 0.026	РЕГИОНАЛЬНЫЙ 0.037	ЦЕНА 0.025
БАНК 0.024	РАЗВИТИЕ 0.036	РОСТ 0.023
БАНКОВСКИЙ 0.016	ОБЛАСТЬ 0.025	ГОД 0.015
КОНФЕРЕНЦИЯ 0.015	НАСЕЛЕНИЕ 0.012	БЫТЬ 0.013
РОССИЙСКИЙ 0.012	УРОВЕНЬ 0.012	КРИЗИС 0.01
СОВЕТ 0.012	ТЕРРИТОРИЯ 0.008	МОЧЬ 0.009
УЧАСТИЕ 0.012	РЕСПУБЛИКА 0.008	АКЦИЯ 0.008
МЕЖДУНАРОДНЫЙ 0.01	РОССИЯ 0.008	СПРОС 0.007
ФОРУМ 0.01	ЭКОНОМИКА 0.007	УРОВЕНЬ 0.007

ТЕМА 46	**ТЕМА 47**	**ТЕМА 48**
ГОСУДАРСТВЕННЫЙ 0.016	РЕГИОНАЛЬНЫЙ 0.023	СТАВКА 0.131
НАУКА 0.014	РЕГИОН 0.017	ПРОЦЕНТНЫЙ 0.076
ЗАЩИЩЕННЫЙ 0.012	БАНК 0.017	СРОК 0.024
ЗАЩИТИТЬ 0.012	БИЗНЕС 0.015	ПРОЦЕНТ 0.024
СИСТЕМА 0.012	РЫНОК 0.014	РЕФИНАНСИРОВАНИЕ 0.019
ЭКОНОМИЧЕСКИЙ 0.011	СРЕДНИЙ 0.014	УСЛОВИЕ 0.015
ЭКОНО 0.011	ГОД 0.012	УСЛОВИЯ 0.014
БИБЛИОГР 0.01	СРЕДНЕЕ 0.011	МОЧЬ 0.013
БИБЛИОГРА 0.01	ГРУППА 0.01	ГОДОВОЙ 0.012
НАЗВ 0.01	МАЛЫЙ 0.009	ЗАЕМ 0.011
ТЕМА 49	**ТЕМА 50**	**ТЕМА 51**
ВКЛАД 0.101	КОМПАНИЯ 0.03	БАНК 0.073
БАНК 0.059	БЫТЬ 0.017	СЧЕТ 0.068
СТРАХОВАНИЕ 0.047	БИЗНЕС 0.015	КЛИЕНТ 0.061
СИСТЕМА 0.035	РЫНОК 0.011	БАНКОВСКИЙ 0.06
ВКЛАДЧИК 0.03	ГОД 0.009	БАНКА 0.034
БАНКА 0.029	ПРОЕКТ 0.008	СРЕДСТВО 0.03
АГЕНТСТВО 0.015	ДОЛЛАР 0.008	СРЕДСТВА 0.03
ДЕПОЗИТ 0.014	ДИРЕКТОР 0.007	ОПЕРАЦИЯ 0.024
СТРАХОВОЙ 0.011	РОССИЯ 0.007	СЧЕТЫ 0.013
СРЕДСТВА 0.011	РБК 0.007	МОЧЬ 0.013
ТЕМА 52	**ТЕМА 53**	**ТЕМА 54**
АВТОМОБИЛЬ 0.011	КРЕДИТНЫЙ 0.135	КОНТРОЛЬ 0.057
БЫТЬ 0.008	ИСТОРИЯ 0.07	ВНУТРЕННИЙ 0.045
МАШИНА 0.008	БЮРО 0.043	УПРАВЛЕНИЕ 0.044
НЕДВИЖИМОСТЬ 0.006	ИНФОРМАЦИЯ 0.037	ОРГАНИЗАЦИЯ 0.033
ПЛОЩАДЬ 0.006	ЗАЕМЩИК 0.019	ДЕЯТЕЛЬНОСТЬ 0.018
ГОРОД 0.006	ДАТЬ 0.014	СОВЕТ 0.017
ГОД 0.006	ДАННЫЙ 0.013	ДИРЕКТОР 0.016
МОСКВА 0.006	ЗАКОН 0.013	СИСТЕМА 0.014
НОВЫЙ 0.006	ДАННЫЕ 0.011	КОРПОРАТИВНЫЙ 0.013
ТЫС 0.006	СУБЪЕКТ 0.009	СЛУЖБА 0.012
ТЕМА 55	**ТЕМА 56**	**ТЕМА 57**
БЫТЬ 0.011	ДАТЬ 0.023	ФОНД 0.047
РУССКИЙ 0.008	ДАННЫЙ 0.023	УПРАВЛЯТЬ 0.024
ГОД 0.007	ОЦЕНКА 0.021	УПРАВЛЯЮЩИЙ 0.023
ИСКУССТВО 0.006	ДАННЫЕ 0.019	КОМПАНИЯ 0.021
РУССКАЯ 0.005	ИССЛЕДОВАНИЕ 0.016	ИНВЕСТИЦИОННЫЙ 0.02
КОЛЛЕКЦИЯ 0.004	СТАТИСТИЧЕСКИЙ 0.014	УПРАВЛЕНИЕ 0.019
СТАТЬ 0.004	СТАТИСТИКА 0.013	УК 0.016
КАРТИНА 0.004	РЕЗУЛЬТАТ 0.01	ПИФ 0.016
ФИЛЬМ 0.004	АНАЛИЗ 0.01	РЫНОК 0.014
ВРЕМЯ 0.004	БЫТЬ 0.009	ФАКТОРИНГ 0.014

ТЕМА 58	**ТЕМА 59**	**ТЕМА 60**
БАНКОВСКИЙ 0.048	КУРС 0.042	ПЕРЕВОД 0.089
АССОЦИАЦИЯ 0.047	ВАЛЮТА 0.039	ДЕНЕЖНЫЙ 0.022
БАНК 0.033	ВАЛЮТНЫЙ 0.038	СИСТЕМА 0.022
АРБА 0.029	РУБЛЬ 0.025	ДЕНЬГИ 0.01
РОССИЙСКИЙ 0.019	ДОЛЛАР 0.022	БЫТЬ 0.01
РОССИЯ 0.014	СТРАНА 0.014	ПОЛУЧАТЕЛЬ 0.01
СИСТЕМА 0.014	БЫТЬ 0.013	БАНК 0.01
СООБЩЕСТВО 0.011	ЕВРО 0.013	СЧЕТ 0.009
РАЗВИТИЕ 0.01	РЕЗЕРВ 0.012	АДВОКАТ 0.008
СОВЕТ 0.009	ПОЛИТИКА 0.011	ПОРУЧЕНИЕ 0.008
ТЕМА 61	**ТЕМА 62**	**ТЕМА 63**
ОРГАНИЗАЦИЯ 0.136	ГОД 0.054	БУМАГА 0.071
КРЕДИТНЫЙ 0.128	МЛН 0.023	ЦЕННЫЙ 0.059
БАНКОВСКИЙ 0.068	ДОЛЛАР 0.017	РЫНОК 0.055
ФИЛИАЛ 0.048	КОМПАНИЯ 0.016	ОБЛИГАЦИЯ 0.045
РОССИЯ 0.022	РЫНОК 0.014	ВЫПУСК 0.026
ОПЕРАЦИЯ 0.018	МЛРД 0.013	ЭМИТЕНТ 0.02
ДЕЯТЕЛЬНОСТЬ 0.015	БЫТЬ 0.011	ИНВЕСТОР 0.02
ЧИСЛО 0.015	УЖЕ 0.009	РАЗМЕЩЕНИЕ 0.02
ЛИЦО 0.015	УЖ 0.009	АКЦИЯ 0.019
ТОМ 0.013	УЗКИЙ 0.009	ФОНДОВЫЙ 0.017
ТЕМА 64	**ТЕМА 65**	**ТЕМА 66**
БЫТЬ 0.053	ПРОИЗВОДСТВО 0.035	ДОГОВОР 0.05
ЕСТЬ 0.018	ПРЕДПРИЯТИЕ 0.028	РФ 0.016
ОЧЕНЬ 0.013	ПРОДУКЦИЯ 0.019	ОБЯЗАТЕЛЬСТВО 0.012
МОЧЬ 0.011	ОТРАСЛЬ 0.016	ЛИЦО 0.012
УЖ 0.01	ПРОМЫШЛЕННОСТЬ 0.011	ПРАВО 0.012
НУЖНЫЙ 0.01	ЭКОНОМИКА 0.011	ДОКУМЕНТ 0.011
ТОМ 0.01	ЭКОНОМИЧЕСКИЙ 0.01	ИСПОЛНЕНИЕ 0.011
ГОД 0.009	ИННОВАЦИОННЫЙ 0.01	СТОРОНА 0.01
УЖЕ 0.009	ПРОИЗВОДСТВЕННЫЙ 0.01	УСЛОВИЕ 0.01
БОЛЬШОЙ 0.009	ТЕХНОЛОГИЯ 0.009	СЛУЧАЙ 0.009
ТЕМА 67	**ТЕМА 68**	**ТЕМА 69**
НАСЕЛЕНИЕ 0.042	ДЕНЬГИ 0.076	ПЕНСИОННЫЙ 0.084
ДОХОД 0.018	ДЕНЕЖНЫЙ 0.073	ФОНД 0.037
ПОТРЕБИТЕЛЬСКИЙ 0.017	ОБРАЩЕНИЕ 0.019	ПЕНСИЯ 0.025
ЧАСТНЫЙ 0.014	СРЕДСТВО 0.015	ГОД 0.018
СБЕРЕЖЕНИЕ 0.013	МАССА 0.015	СРЕДСТВА 0.018
УСЛУГА 0.01	СРЕДСТВА 0.012	СРЕДСТВО 0.018
ЧЕЛОВЕК 0.009	ЭКОНОМИКА 0.009	НАКОПЛЕНИЕ 0.015
ГРАЖДАНИН 0.008	ФУНКЦИЯ 0.009	НАКОПЛЕНИЯ 0.015
ДОЛЯ 0.008	ОБОРОТ 0.008	СТРАХОВОЙ 0.012
ФИНАНСОВЫЙ 0.008	ЦЕНТРАЛЬНЫЙ 0.008	ВЗНОС 0.012

ТЕМА 70	**ТЕМА 71**	**ТЕМА 72**
ИПОТЕЧНЫЙ 0.079	РАСХОД 0.023	АУДИТОРСКИЙ 0.063
КРЕДИТ 0.066	СУММА 0.018	АУДИТ 0.055
КРЕДИТОВАНИЕ 0.033	РУБ 0.015	АУДИТОР 0.043
ИПОТЕКА 0.025	ОРГАНИЗАЦИЯ 0.014	ПРОВЕРКА 0.031
ЖИЛЬЕ 0.024	РФ 0.013	БУХГАЛТЕРСКИЙ 0.016
РЫНОК 0.02	ДОХОД 0.012	ОРГАНИЗАЦИЯ 0.014
БАНК 0.02	УЧЕТ 0.011	КОНТРОЛЬ 0.014
ЖИЛИЩНЫЙ 0.017	СТОИМОСТЬ 0.011	ОТЧЕТНОСТЬ 0.014
ЗАЕМЩИК 0.016	ДОГОВОР 0.01	ДЕЯТЕЛЬНОСТЬ 0.014
НЕДВИЖИМОСТЬ 0.012	НДС 0.01	ПРОВЕДЕНИЕ 0.01
ТЕМА 73	**ТЕМА 74**	**ТЕМА 75**
СИСТЕМА 0.044	ОБЩЕСТВО 0.049	БУМАГА 0.069
ДАТЬ 0.016	ДЕЯТЕЛЬНОСТЬ 0.028	ЦЕННЫЙ 0.066
ДАННЫЙ 0.016	ОРГАНИЗАЦИЯ 0.024	СДЕЛКА 0.028
ДАННЫЕ 0.016	УСТАВНЫЙ 0.022	ВЕКСЕЛЬ 0.02
ИНФОРМАЦИОННЫЙ 0.015	ЛИЦО 0.018	АККРЕДИТИВ 0.018
ДОКУМЕНТ 0.013	УЧАСТНИК 0.016	РЕПО 0.016
РЕШЕНИЕ 0.012	УСТАВНОЙ 0.016	ОБЯЗАТЕЛЬСТВО 0.011
ТЕХНОЛОГИЯ 0.01	АКЦИОНЕРНЫЙ 0.016	ДЕПОЗИТАРИЙ 0.011
ПРОГРАММНЫЙ 0.01	ОБЩИЙ 0.014	ОПЕРАЦИЯ 0.011
ИНФОРМАЦИЯ 0.01	СОБРАНИЕ 0.013	БЫТЬ 0.01
ТЕМА 76	**ТЕМА 77**	**ТЕМА 78**
РОССИЯ 0.03	БАНК 0.057	КРЕДИТ 0.093
ПОРЯДОК 0.021	РУБ 0.046	БАНК 0.076
ДОКУМЕНТ 0.018	МЛН 0.038	КРЕДИТНЫЙ 0.056
ИЗМЕНЕНИЕ 0.018	БАНКА 0.027	ЗАЕМЩИК 0.042
ОРГАНИЗАЦИЯ 0.016	МЛРД 0.026	БАНКА 0.038
УКАЗАНИЕ 0.015	ГОД 0.023	КРЕДИТОВАНИЕ 0.026
ФОРМА 0.015	РУБЛЬ 0.02	ЗАДОЛЖЕННОСТЬ 0.021
ДЕНЬ 0.013	СОСТАВИТЬ 0.016	ССУДА 0.02
ЛИЦО 0.013	ТЫС 0.011	АКТИВ 0.01
ПОЛОЖЕНИЕ 0.011	ОБЪЕМ 0.01	ПОРТФЕЛЬ 0.01
ТЕМА 79	**ТЕМА 80**	**ТЕМА 81**
СЧЕТ 0.121	РЕКЛАМА 0.029	БЫТЬ 0.032
СЧЕТЫ 0.053	РЕКЛАМНЫЙ 0.016	ГОСБАНК 0.015
СРЕДСТВА 0.05	БРЭНД 0.007	СССР 0.014
СРЕДСТВО 0.049	ЖУРНАЛ 0.007	ГОСУДАРСТВЕННЫЙ 0.012
УЧЕТ 0.028	БРЕНД 0.006	ГОД 0.01
ОПЕРАЦИЯ 0.028	СМИ 0.006	ОТДЕЛЕНИЕ 0.01
ДЕНЕЖНЫЙ 0.022	ИЗДАНИЕ 0.006	БЫЛЬ 0.01
РАСЧЕТ 0.02	АУДИТОРИЯ 0.006	УЧРЕЖДЕНИЕ 0.009
РАСЧЕТНЫЙ 0.016	ПРОДУКТ 0.006	КОНТОРА 0.007
ДОКУМЕНТ 0.016	ПОТРЕБИТЕЛЬ 0.005	ВОЙНА 0.006

ТЕМА 82	**ТЕМА 83**	**ТЕМА 84**
ЕВРОПЕЙСКИЙ 0.022	ПРАВО 0.052	БАНК 0.069
СТРАНА 0.021	ПРАВЫЙ 0.038	БАНКА 0.04
БЫТЬ 0.016	ЗАКОН 0.032	БЫТЬ 0.015
ПОЛЬША 0.011	ИМУЩЕСТВО 0.022	БАНКИР 0.012
НОВЫЙ 0.009	ЗАЛОГ 0.019	КЛИЕНТ 0.008
ЧЛЕН 0.008	ДОГОВОР 0.013	БАНКОВСКИЙ 0.007
ГОСУДАРСТВО 0.008	КРЕДИТОР 0.011	МОЧЬ 0.007
ПИИ 0.007	СОБСТВЕННОСТЬ 0.01	БИЗНЕС 0.006
ЕВРОСОЮЗ 0.007	ТРЕБОВАНИЕ 0.01	ДЕЛО 0.005
ЕВРО 0.007	ОБЕСПЕЧЕНИЕ 0.01	КОГДА 0.005
ТЕМА 85	**ТЕМА 86**	**ТЕМА 87**
РОССИЙСКИЙ 0.065	НАЛИЧНЫЙ 0.039	МАЛЫЙ 0.037
РОССИЯ 0.043	КАССОВЫЙ 0.024	ПРОЕКТ 0.034
ИНОСТРАННЫЙ 0.032	НАЛИЧНЫЕ 0.022	ПРЕДПРИЯТИЕ 0.023
КАПИТАЛ 0.019	БАНКНОТА 0.019	БИЗНЕС 0.023
СТРАНА 0.018	БАНКНОТ 0.018	РАЗВИТИЕ 0.021
РЫНОК 0.009	ДЕНЕЖНЫЙ 0.016	ПРОГРАММА 0.018
ИНВЕСТИЦИЯ 0.009	БАНК 0.014	ФИНАНСИРОВАНИЕ 0.017
ОТЕЧЕСТВЕННЫЙ 0.009	ВАЛЮТА 0.014	КРЕДИТОВАНИЕ 0.017
ЭКОНОМИКА 0.008	ОПЕРАЦИЯ 0.013	КРЕДИТ 0.015
ОБЪЕМ 0.008	ЧЕК 0.012	СРЕДСТВА 0.015
ТЕМА 88	**ТЕМА 89**	**ТЕМА 90**
РАБОТНИК 0.032	КОМПАНИЯ 0.138	ПОКАЗАТЕЛЬ 0.042
СОЦИАЛЬНЫЙ 0.029	БИЗНЕС 0.04	ОЦЕНКА 0.018
ТРУД 0.028	РЫНОК 0.022	ЗНАЧЕНИЕ 0.018
ТРУДОВОЙ 0.018	ФИНАНСОВЫЙ 0.017	КОЭФФИЦИЕНТ 0.018
ОБРАЗОВАНИЕ 0.017	РОССИЙСКИЙ 0.013	ДАННЫЙ 0.014
РАБОТА 0.011	РЕШЕНИЕ 0.012	РАСЧЕТ 0.014
СТУДЕНТ 0.01	УСЛУГА 0.011	ДАТЬ 0.014
РАБОТОДАТЕЛЬ 0.01	ТЕХНОЛОГИЯ 0.007	ВЕЛИЧИНА 0.012
РАБОЧИЙ 0.009	УПРАВЛЕНИЕ 0.007	МОДЕЛЬ 0.011
ОБРАЗОВАТЕЛЬНЫЙ 0.009	КОРПОРАТИВНЫЙ 0.006	АНАЛИЗ 0.01
ТЕМА 91	**ТЕМА 92**	**ТЕМА 93**
ДОЛЖНЫЙ 0.044	СТРАНА 0.027	СТРАХОВОЙ 0.104
ДОЛЖЕН 0.043	ВТО 0.013	СТРАХОВАНИЕ 0.075
БЫТЬ 0.027	КИТАЙ 0.013	КОМПАНИЯ 0.039
МОЧЬ 0.019	ЭКОНОМИЧЕСКИЙ 0.012	СТРАХОВЩИК 0.027
ПРИНЦИП 0.015	ТОРГОВЛЯ 0.012	РЫНОК 0.013
ТРЕБОВАНИЕ 0.013	РОССИЯ 0.011	ЖИЗНЬ 0.01
НЕОБХОДИМЫЙ 0.012	ИНТЕГРАЦИЯ 0.009	СЛУЧАЙ 0.009
СЛЕДОВАТЬ 0.01	МЕЖДУНАРОДНЫЙ 0.009	БЫТЬ 0.008
СООТВЕТСТВОВАТЬ 0.009	БЫТЬ 0.008	ПРЕМИЯ 0.008
ДАТЬ 0.008	СОТРУДНИЧЕСТВО 0.007	ВИД 0.007

ТЕМА 94	**ТЕМА 95**	**ТЕМА 96**
БЫТЬ 0.008	ОРГАНИЗАЦИЯ 0.028	БАНК 0.092
ОТЕЛЬ 0.006	ЗАКОН 0.022	БАНКА 0.035
ЧЕЛОВЕК 0.006	ОПЕРАЦИЯ 0.021	БАНКОВСКИЙ 0.029
МОЖНО 0.005	ЛИЦО 0.02	РЫНОК 0.025
ГОД 0.005	КРЕДИТНЫЙ 0.019	РОССИЙСКИЙ 0.021
САМОЛЕТ 0.004	ОТМЫВАНИЕ 0.016	ИНОСТРАННЫЙ 0.019
ВОДА 0.004	ПРЕСТУПНЫЙ 0.016	КАПИТАЛ 0.013
ГОРОД 0.004	ДОХОД 0.015	БЫТЬ 0.011
ВРЕМЯ 0.004	ИНФОРМАЦИЯ 0.014	КРУПНЫЙ 0.009
РБК 0.003	ЛЕГАЛИЗАЦИЯ 0.014	БИЗНЕС 0.008
ТЕМА 97	**ТЕМА 98**	**ТЕМА 99**
БАНКОВСКИЙ 0.118	КАПИТАЛ 0.062	МЛРД 0.06
СИСТЕМА 0.046	АКЦИЯ 0.055	ДОЛЛ 0.028
БАНК 0.043	КОМПАНИЯ 0.017	ДОЛ 0.019
ФИНАНСОВЫЙ 0.032	АКЦИОНЕР 0.016	ОБЪЕМ 0.018
НАДЗОР 0.032	СОБСТВЕННЫЙ 0.016	ИНВЕСТИЦИЯ 0.013
СЕКТОР 0.024	ДОЛЯ 0.015	БАЛАНС 0.013
РЕГУЛИРОВАНИЕ 0.013	БЫТЬ 0.014	США 0.011
БАНКА 0.012	СРЕДСТВО 0.011	АКТИВ 0.011
ОРГАН 0.012	СРЕДСТВА 0.011	ЛИЗИНГ 0.01
КАПИТАЛ 0.011	РАЗМЕР 0.01	ЛИЗИНГОВЫЙ 0.01
ТЕМА 100		
РИСК 0.173		
РИСКА 0.077		
УПРАВЛЕНИЕ 0.02		
ОПЕРАЦИОННЫЙ 0.015		
ПОТЕРЯ 0.013		
ОЦЕНКА 0.013		
КРЕДИТНЫЙ 0.012		
ОПЕРАЦИОННАЯ 0.009		
МОЧЬ 0.009		
КАПИТАЛ 0.007		

Приложение Б

Список первых 10 слов и словосочетаний из тем, полученных алгоритмом PLSA-SIM на банковском корпусе с добавлением 1000 самых частотных словосочетаний

ТЕМА 1	**ТЕМА 2**
НАЦИОНАЛЬНЫЙ БАНК 0.2	ИПОТЕЧНЫЙ КРЕДИТ 0.066
ПРЕДСЕДАТЕЛЬ БАНК 0.12	ИПОТЕЧНЫЙ БАНК 0.046
БАНК СТРАНА 0.106	ИПОТЕЧНЫЙ КРЕДИТОВАНИЕ 0.044
ЦЕНТРАЛЬНЫЙ БАНК 0.085	КРЕДИТ 0.036
БАНК РЕСПУБЛИКА 0.08	ИПОТЕЧНЫЙ 0.033
ДЕЯТЕЛЬНОСТЬ БАНК 0.033	ПОТРЕБИТЕЛЬСКИЙ КРЕДИТ 0.03
КОММЕРЧЕСКИЙ БАНК 0.032	ИПОТЕЧНЫЙ РЫНОК 0.026
ВСЕМИРНЫЙ БАНК 0.032	БАНК 0.017
ИНВЕСТИЦИОННЫЙ БАНК 0.012	КРЕДИТОВАНИЕ 0.017
ЦЕНТРАЛЬНЫЙ БАНКА 0.01	ВЫДАЧА КРЕДИТ 0.016
ТЕМА 3	**ТЕМА 4**
ДИРЕКТОР БАНК 0.168	БЫТЬ 0.021
АКЦИЯ БАНК 0.15	ГОСУДАРСТВЕННЫЙ БАНК 0.014
РАЗВИТИЕ БАНК 0.095	ГОСБАНК СССР 0.014
ДОЧЕРНИЙ БАНК 0.025	ГОСУДАРСТВЕННЫЙ 0.009
МЛН РУБ 0.021	ГОД 0.009
СОВЕТ ДИРЕКТОР 0.019	СССР 0.009
РОССИЙСКИЙ БАНК 0.018	ГОСБАНК 0.008
БАНК РОССИЯ 0.016	УЧРЕЖДЕНИЕ 0.006
МЛРД РУБ 0.015	БЫЛЬ 0.006
БАНКА 0.014	ОТДЕЛЕНИЕ 0.006
ТЕМА 5	**ТЕМА 6**
ФИНАНСОВЫЙ РЫНОК 0.085	БЫТЬ 0.03
ФИНАНСОВЫЙ УСЛУГА 0.079	ДОЛЖНЫЙ 0.018
ФИНАНСОВЫЙ 0.062	ДОЛЖЕН 0.017
ФИНАНСОВЫЙ ОРГАНИЗАЦИЯ 0.059	МОЧЬ 0.013
ФИНАНСОВЫЙ ИНСТИТУТ 0.053	ПРОБЛЕМА 0.009
ФИНАНСОВЫЙ УЧРЕЖДЕНИЕ 0.034	ДОЛЖНЫЙ СТАТЬ 0.009
ФИНАНСОВЫЙ СИСТЕМА 0.032	ДОЛЖЕН СТАТЬ 0.009
ФИНАНСОВЫЙ СЕКТОР 0.032	ВОПРОС 0.008
ФИНАНСОВЫЙ ОПЕРАЦИЯ 0.027	ТОМ 0.008
РЫНОК 0.018	НЕОБХОДИМЫЙ 0.007

ТЕМА 7	**ТЕМА 8**
КОМПАНИЯ 0.119	ЭКОНОМИЧЕСКИЙ 0.014
РОССИЙСКИЙ КОМПАНИЯ 0.085	ЭКОНОМИЧЕСКИЙ ТЕОРИЯ 0.012
ДЕЯТЕЛЬНОСТЬ КОМПАНИЯ 0.053	ЭКОНОМИЧЕСКИЙ СУБЪЕКТ 0.01
КРУПНЫЙ КОМПАНИЯ 0.047	ЭКОНОМИЧЕСКИЙ СИСТЕМА 0.009
ДИРЕКТОР КОМПАНИЯ 0.038	ЭКОНОМИЧЕСКИЙ НАУКА 0.008
ДОЧЕРНИЙ КОМПАНИЯ 0.03	ТЕОРИЯ 0.008
ФИНАНСОВЫЙ КОМПАНИЯ 0.027	ЭКОНОМИЧЕСКИЙ ДЕЯТЕЛЬНОСТЬ 0.007
ГРУППА КОМПАНИЯ 0.026	ЯВЛЯТЬСЯ 0.007
СТРАХОВОЙ КОМПАНИЯ 0.026	ДАТЬ 0.005
ИНОСТРАННЫЙ КОМПАНИЯ 0.024	ПРОЦЕСС 0.005
ТЕМА 9	**ТЕМА 10**
БЮДЖЕТНЫЙ СРЕДСТВО 0.034	УПРАВЛЕНИЕ БАНК 0.153
БЮДЖЕТНЫЙ СРЕДСТВА 0.034	УЧРЕЖДЕНИЕ БАНК 0.14
ФЕДЕРАЛЬНЫЙ БЮДЖЕТ 0.032	БАНК РОССИЯ 0.119
БЮДЖЕТНЫЙ СИСТЕМА 0.023	УПОЛНОМОЧЕННЫЙ БАНК 0.104
БЮДЖЕТ 0.022	БАНКА РОССИЯ 0.084
БЮДЖЕТНЫЙ 0.022	УПОЛНОМОЧЕННЫЙ БАНКА 0.047
БЮДЖЕТНЫЙ КОДЕКС 0.017	ЦЕНТРАЛЬНЫ БАНК 0.037
ОРГАН ВЛАСТЬ 0.017	БАНКИЙ РОССИЯ 0.036
БЮДЖЕТНЫЙ РАСХОД 0.017	БАНКА 0.022
ФЕДЕРАЛЬНЫЙ 0.015	СЕТЬ БАНК 0.022
ТЕМА 11	**ТЕМА 12**
КРЕДИТ БАНК 0.052	КРЕДИТНЫЙ ОРГАНИЗАЦИЯ 0.097
ОБЪЕМ КРЕДИТ 0.03	БАНК РОССИЯ 0.079
БАНК РОССИЯ 0.025	БАНКА РОССИЯ 0.061
ОБЩИЙ ОБЪЕМ 0.02	БАНКИЙ РОССИЯ 0.057
БАНКА РОССИЯ 0.013	ОРГАНИЗАЦИЯ 0.055
БАНК 0.012	БАНКОВСКИЙ ОПЕРАЦИЯ 0.03
МЛРД 0.012	СБЕРБАНК РОССИЯ 0.029
БАНКОВСКИЙ КРЕДИТ 0.011	УСТАВНОЙ КАПИТАЛ 0.027
КРЕДИТНЫЙ ОРГАНИЗАЦИЯ 0.011	БАНКОВСКИЙ ДЕЯТЕЛЬНОСТЬ 0.026
ОБЪЕМ 0.01	УСТАВНЫЙ КАПИТАЛ 0.026
ТЕМА 13	**ТЕМА 14**
ПЕНСИОННЫЙ ФОНД 0.082	ИНОСТРАННЫЙ КАПИТАЛ 0.047
ПЕНСИОННЫЙ НАКОПЛЕНИЕ 0.047	ИНОСТРАННЫЙ ИНВЕСТИЦИЯ 0.047
ПЕНСИОННЫЙ СТРАХОВАНИЕ 0.047	ИНОСТРАННЫЙ ИНВЕСТОР 0.034
ПЕНСИОННЫЙ НАКОПЛЕНИЯ 0.047	ИНОСТРАННЫЙ 0.025
ПЕНСИОННЫЙ ОБЕСПЕЧЕНИЕ 0.046	РОССИЙСКИЙ ПРЕДПРИЯТИЕ 0.022
ПЕНСИОННЫЙ 0.039	ИНВЕСТИЦИОННЫЙ БАНК 0.02
ПЕНСИОННЫЙ СИСТЕМА 0.034	КАПИТАЛ 0.02
ПЕНСИОННЫЙ РЕЗЕРВ 0.024	ПРЯМОЙ ИНВЕСТИЦИЯ 0.019
ФОНД 0.018	РОССИЙСКИЙ ЭКОНОМИКА 0.019
ПЕНСИЯ 0.014	ПРЯМАЯ ИНВЕСТИЦИЯ 0.018

ТЕМА 15	**ТЕМА 16**
КРЕДИТ 0.045	АУДИТОРСКИЙ ОРГАНИЗАЦИЯ 0.042
КРЕДИТНЫЙ 0.041	АУДИТОРСКИЙ ДЕЯТЕЛЬНОСТЬ 0.037
КРЕДИТНЫЙ ПОРТФЕЛЬ 0.038	АУДИТОРСКИЙ ПРОВЕРКА 0.032
ПРЕДОСТАВЛЕНИЕ КРЕДИТ 0.033	АУДИТОРСКИЙ 0.031
КРЕДИТНЫЙ РИСК 0.032	АУДИТ 0.028
КРЕДИТНЫЙ ДОГОВОР 0.026	АУДИТОР 0.021
ЗАЕМЩИК 0.026	АУДИТОРСКИЙ ЗАКЛЮЧЕНИЕ 0.021
КРЕДИТНЫЙ ОПЕРАЦИЯ 0.026	ПРОВЕДЕНИЕ АУДИТ 0.02
СУММА КРЕДИТ 0.026	ВНУТРЕННИЙ КОНТРОЛЬ 0.019
КРЕДИТНЫЙ ОРГАНИЗАЦИЯ 0.025	АУДИТОРСКИЙ УСЛУГА 0.019
ТЕМА 17	**ТЕМА 18**
ОСНОВНОЙ СРЕДСТВО 0.079	РОССИЙСКИЙ РЫНОК 0.091
ОСНОВНЫЙ СРЕДСТВО 0.078	РЫНОК 0.058
ОСНОВНОЙ СРЕДСТВА 0.077	ФОНДОВЫЙ РЫНОК 0.057
ОСНОВНЫЙ СРЕДСТВА 0.076	ФИНАНСОВЫЙ РЫНОК 0.044
ДЕНЕЖНЫЙ СРЕДСТВО 0.028	РАЗВИТИЕ РЫНОК 0.042
ДЕНЕЖНЫЙ СРЕДСТВА 0.028	РЫНОК КАПИТАЛ 0.038
СОБСТВЕННЫЙ СРЕДСТВО 0.023	РЫНОК АКЦИЯ 0.038
СОБСТВЕННЫЙ СРЕДСТВА 0.023	МЕЖДУНАРОДНЫЙ РЫНОК 0.037
СРЕДСТВО 0.014	УЧАСТНИК РЫНОК 0.036
СРЕДСТВА 0.014	ВНУТРЕННИЙ РЫНОК 0.033
ТЕМА 19	**ТЕМА 20**
ИНФОРМАЦИОННЫЙ СИСТЕМА 0.052	КРЕДИТНЫЙ ИСТОРИЯ 0.103
ИНФОРМАЦИОННЫЙ ТЕХНОЛОГИЯ 0.021	КРЕДИТНЫЙ БЮРО 0.068
СИСТЕМА 0.021	КРЕДИТНЫЙ ОРГАНИЗАЦИЯ 0.049
СИСТЕМА УПРАВЛЕНИЕ 0.019	КРЕДИТНЫЙ 0.048
АВТОМАТИЗИРОВАННЫЙ СИСТЕМА 0.018	КРЕДИТНЫЙ РИСК 0.035
ИНФОРМАЦИОННЫЙ 0.016	КРЕДИТНЫЙ РЫНОК 0.028
ИНФОРМАЦИОННЫЙ БЕЗОПАСНОСТЬ 0.015	ИНФОРМАЦИЯ 0.024
БЕЗОПАСНОСТЬ 0.013	ПОЛУЧЕНИЕ ИНФОРМАЦИЯ 0.023
ВНЕДРЕНИЕ СИСТЕМА 0.013	БЮРО 0.021
ОБЕСПЕЧЕНИЕ БЕЗОПАСНОСТЬ 0.012	ИСТОРИЯ 0.012
ТЕМА 21	**ТЕМА 22**
ПОЛОЖЕНИЕ БАНК 0.098	РОССИЙСКИЙ БАНК 0.182
БАНК РОССИЯ 0.069	КРУПНЫЙ БАНК 0.138
УКАЗАНИЕ БАНК 0.067	БОЛЬШИНСТВО БАНК 0.119
БАНКА РОССИЯ 0.06	РОССИЙСКИЙ БАНКА 0.083
ИНСТРУКЦИЯ БАНК 0.057	ЗАПАДНЫЙ БАНК 0.078
АКТ БАНК 0.052	БАНК 0.046
БАНКИЙ РОССИЯ 0.041	СРЕДНИЙ БАНК 0.042
ПИСЬМО БАНК 0.033	КОНКРЕТНЫЙ БАНК 0.041
ЦЕНТРАЛЬНЫЙ БАНК 0.027	КРУПНЫЙ БАНКА 0.037
УЧРЕЖДЕНИЕ БАНК 0.025	ОТЕЧЕСТВЕННЫЙ БАНК 0.031

ТЕМА 23	**ТЕМА 24**
ПРОШЛЫЙ ГОД 0.077	БАНКОВСКИЙ СЧЕТ 0.087
ГОД 0.06	СЧЕТ КЛИЕНТ 0.054
ПРОШЛОЕ ГОД 0.042	СЧЕТ 0.049
ТЕКУЩИЙ ГОД 0.037	БАНКОВСКИЙ СЧЕТЫ 0.039
КОНЕЦ ГОД 0.032	ОТКРЫТИЕ СЧЕТ 0.029
СЛЕДУЮЩИЙ ГОД 0.022	БАНК 0.027
НАЧАЛО ГОД 0.021	КОРРЕСПОНДЕНТСКИЙ СЧЕТ 0.026
ПОСЛЕДНИЙ ГОД 0.019	ДЕНЕЖНЫЙ СРЕДСТВО 0.024
ПРОШЕДШИЙ ГОД 0.016	ДЕНЕЖНЫЙ СРЕДСТВА 0.024
ПРЕДЫДУЩИЙ ГОД 0.013	РАСЧЕТНЫЙ СЧЕТ 0.022
ТЕМА 25	**ТЕМА 26**
СТОИМОСТЬ АКТИВ 0.034	РЕГИОНАЛЬНЫЙ БАНК 0.137
ФИНАНСОВЫЙ ОТЧЕТНОСТЬ 0.032	КРУПНЫЙ БАНК 0.092
ФИНАНСОВЫЙ РЕЗУЛЬТАТ 0.028	МЕСТНЫЙ БАНК 0.059
СТОИМОСТЬ 0.02	РЕГИОНАЛЬНЫЙ БАНКА 0.051
АКТИВ 0.019	МОСКОВСКИЙ БАНК 0.044
ФИНАНСОВЫЙ ПОЛОЖЕНИЕ 0.018	БАНК СТРАНА 0.043
ФИНАНСОВЫЙ 0.018	СРЕДНИЙ БАНК 0.039
РЫНОЧНЫЙ СТОИМОСТЬ 0.016	ЕВРОПЕЙСКИЙ БАНК 0.036
ОТЧЕТНОСТЬ 0.015	КРУПНЫЙ БАНКА 0.032
ФОРМА ОТЧЕТНОСТЬ 0.015	ОТДЕЛЬНЫЙ БАНК 0.019
ТЕМА 27	**ТЕМА 28**
СИСТЕМА УПРАВЛЕНИЕ 0.097	ГОСУДАРСТВЕННЫЙ ДОЛГ 0.018
УПРАВЛЕНИЕ 0.04	МЛРД 0.017
КАЧЕСТВО УПРАВЛЕНИЕ 0.028	ВНЕШНИЙ ДОЛГ 0.014
ПРОЦЕСС УПРАВЛЕНИЕ 0.025	ДОЛГ 0.013
СИСТЕМА 0.022	ДОХОД 0.012
КОРПОРАТИВНЫЙ УПРАВЛЕНИЕ 0.02	ГОСУДАРСТВЕННЫЙ 0.012
РЕЗУЛЬТАТ ДЕЯТЕЛЬНОСТЬ 0.017	ФОНД 0.011
ЭФФЕКТИВНЫЙ СИСТЕМА 0.017	ДОЛЛ 0.011
ОЦЕНКА КАЧЕСТВО 0.016	СТАБИЛИЗАЦИОННЫЙ ФОНД 0.01
ОЦЕНКА ЭФФЕКТИВНОСТЬ 0.015	МЛРД ДОЛЛ 0.009
ТЕМА 29	**ТЕМА 30**
ПОДРАЗДЕЛЕНИЕ БАНК 0.196	УСЛОВИЕ ДОГОВОР 0.039
ОРГАНИЗАЦИЯ РАБОТА 0.069	УСЛОВИЯ ДОГОВОР 0.037
ДЕЯТЕЛЬНОСТЬ БАНК 0.065	ДОГОВОР 0.034
СОТРУДНИК БАНК 0.035	ЗАКЛЮЧЕНИЕ ДОГОВОР 0.027
КРЕДИТНЫЙ ОРГАНИЗАЦИЯ 0.03	ИСПОЛНЕНИЕ ОБЯЗАТЕЛЬСТВО 0.017
РУКОВОДСТВО БАНК 0.02	ПРАВО ТРЕБОВАНИЕ 0.016
ОРГАНИЗАЦИЯ 0.019	ПРАВО 0.014
ВНУТРЕННИЙ КОНТРОЛЬ 0.019	СДЕЛКА 0.014
ВНУТРЕННИЙ ДОКУМЕНТ 0.019	КРЕДИТНЫЙ ДОГОВОР 0.014
ПОДРАЗДЕЛЕНИЕ 0.019	ОБЯЗАТЕЛЬСТВО 0.013

ТЕМА 31	**ТЕМА 32**
ВЕСТНИК БАНК 0.054	АКЦИЯ 0.046
БАНК РОССИЯ 0.042	ПАКЕТ АКЦИЯ 0.022
БАНК 0.031	РАЗМЕЩЕНИЕ АКЦИЯ 0.02
БАНКА РОССИЯ 0.026	АКЦИОНЕР 0.016
МОСКВА 0.019	КОМПАНИЯ 0.016
КОММЕРЧЕСКИЙ БАНК 0.019	СДЕЛКА 0.014
АКТ БАНК 0.018	ЦЕНА 0.011
ТРЕБОВАНИЕ КРЕДИТОР 0.016	КАПИТАЛ 0.011
ДИРЕКТОР БАНК 0.015	СОБСТВЕННЫЙ КАПИТАЛ 0.01
ООО КБ 0.013	УСТАВНЫЙ КАПИТАЛ 0.009
ТЕМА 33	**ТЕМА 34**
ЭКОНОМИЧЕСКИЙ РАЗВИТИЕ 0.031	РОССИЙСКИЙ ЭКОНОМИКА 0.027
РАЗВИТИЕ 0.027	ТЕМП РОСТ 0.025
СТРАТЕГИЯ РАЗВИТИЕ 0.025	СЕКТОР ЭКОНОМИКА 0.02
РАЗВИТИЕ СТРАНА 0.018	ЭКОНОМИЧЕСКИЙ РОСТ 0.02
ДАЛЬНЕЙШИЙ РАЗВИТИЕ 0.017	РОСТ ЭКОНОМИКА 0.018
НАПРАВЛЕНИЕ РАЗВИТИЕ 0.015	ЭКОНОМИКА РОССИЯ 0.018
РАЗВИТИЕ ЭКОНОМИКА 0.014	РАЗВИТИЕ ЭКОНОМИКА 0.017
ПЕРСПЕКТИВА РАЗВИТИЕ 0.013	ОСНОВНОЙ КАПИТАЛ 0.017
ЭКОНОМИЧЕСКИЙ ОТНОШЕНИЕ 0.013	РОСТ 0.017
ТЕНДЕНЦИЯ РАЗВИТИЕ 0.011	ЭКОНОМИКА 0.016
ТЕМА 35	**ТЕМА 36**
ПРАВО 0.021	СРЕДСТВО БАНК 0.148
БАНКОВСКИЙ ТАЙНА 0.02	СРЕДСТВА БАНК 0.148
ЗАЩИТА ПРАВО 0.018	ОБЯЗАТЕЛЬСТВО БАНК 0.068
ЗАКОН 0.017	СОБСТВЕННЫЙ СРЕДСТВО 0.032
СУД 0.017	СОБСТВЕННЫЙ СРЕДСТВА 0.031
ПРАВЫЙ 0.015	КРЕДИТ БАНК 0.023
АРБИТРАЖНЫЙ СУД 0.014	ПОЛОЖЕНИЕ БАНК 0.022
ПРАВОВОЙ 0.011	ОПЕРАЦИЯ БАНК 0.019
ФЕДЕРАЛЬНЫЙ ЗАКОН 0.011	СТОРОНА БАНК 0.017
СУД РФ 0.01	КОММЕРЧЕСКИЙ БАНК 0.017
ТЕМА 37	**ТЕМА 38**
РЕГИОН РОССИЯ 0.044	СУММА 0.038
РЕГИОН 0.042	ОБЩИЙ СУММА 0.027
РЕАЛИЗАЦИЯ ПРОЕКТ 0.029	РУБ 0.027
ОБЛАСТЬ 0.025	СРОК 0.022
ПРОЕКТ 0.024	ДЕНЬ 0.02
РЕГИОНАЛЬНЫЙ 0.021	РАБОЧИЙ ДЕНЬ 0.017
РАЗВИТИЕ РЕГИОН 0.017	БЫТЬ 0.016
ФЕДЕРАЛЬНЫЙ ОКРУГ 0.013	ДАТА 0.01
РАЗВИТИЕ 0.011	МЕСЯЦ 0.01
МОСКВА 0.01	РАЗМЕР 0.09

ТЕМА 39	**ТЕМА 40**
БАНКОВСКИЙ АССОЦИАЦИЯ 0.043	ТРУДОВОЙ ДОГОВОР 0.02
РЕГИОНАЛЬНЫЙ БАНК 0.037	ТРАНСПОРТНЫЙ СРЕДСТВО 0.018
РОССИЙСКИЙ БАНК 0.033	ТРАНСПОРТНЫЙ СРЕДСТВА 0.018
АССОЦИАЦИЯ 0.031	ОПЛАТА УСЛУГА 0.015
ПРЕЗИДЕНТ АССОЦИАЦИЯ 0.024	РАСХОД 0.014
БАНК РОССИЯ 0.023	ТРУДОВОЙ 0.014
БАНКОВСКИЙ СОВЕТ 0.021	РАБОТНИК 0.013
ЧЛЕН АССОЦИАЦИЯ 0.019	СОЦИАЛЬНЫЙ СТРАХОВАНИЕ 0.012
БАНКОВСКИЙ СООБЩЕСТВО 0.018	УСЛУГА 0.012
СОВЕТ АССОЦИАЦИЯ 0.017	РОССИЙСКИЙ ФЕДЕРАЦИЯ 0.01
ТЕМА 41	**ТЕМА 42**
ВАЛЮТНЫЙ 0.045	БЫТЬ 0.013
ВАЛЮТНЫЙ РЫНОК 0.044	США 0.011
ВАЛЮТНЫЙ ОПЕРАЦИЯ 0.038	ГОД 0.011
ВАЛЮТА 0.037	АМЕРИКАНСКИЙ 0.008
ИНОСТРАННЫЙ ВАЛЮТА 0.032	СТРАНА 0.008
ВАЛЮТНЫЙ КУРС 0.031	НОВЫЙ 0.006
ВАЛЮТНЫЙ РЕГУЛИРОВАНИЕ 0.031	СТАТЬ 0.006
ВАЛЮТНЫЙ КОНТРОЛЬ 0.029	ВЛАСТЬ 0.005
НАЦИОНАЛЬНЫЙ ВАЛЮТА 0.027	ПОЛИТИЧЕСКИЙ 0.005
ВАЛЮТНЫЙ ЗАКОНОДАТЕЛЬСТВО 0.02	ПОСОЛ 0.004
ТЕМА 43	**ТЕМА 44**
БАНК МОСКВА 0.144	ФИНАНСОВЫЙ ИНСТРУМЕНТ 0.053
МОСКОВСКИЙ БАНК 0.118	МЕЖДУНАРОДНЫЙ СТАНДАРТ 0.048
БАНК РЕКОНСТРУКЦИЯ 0.076	ФИНАНСОВЫЙ ОТЧЕТНОСТЬ 0.04
ЕВРОПЕЙСКИЙ БАНК 0.06	ФИНАНСОВЫЙ АКТИВ 0.037
ПРОМЫШЛЕННЫЙ БАНК 0.06	СТАНДАРТ 0.032
ИНВЕСТИЦИОННЫЙ БАНК 0.046	ФИНАНСОВЫЙ 0.025
БАНК 0.046	МЕЖДУНАРОДНЫЙ 0.023
РОССИЙСКИЙ БАНК 0.043	ОТЧЕТНОСТЬ 0.018
БАНКА МОСКВА 0.04	МЕЖДУНАРОДНЫЙ ПРАКТИКА 0.012
МЕЖДУНАРОДНЫЙ БАНК 0.035	ФИНАНСОВЫЙ РЫНОК 0.011
ТЕМА 45	**ТЕМА 46**
ЦЕННЫЙ БУМАГА 0.061	СИСТЕМА 0.023
БУМАГА 0.04	РЕШЕНИЕ 0.011
ИНВЕСТИЦИОННЫЙ ФОНД 0.026	КОМПАНИЯ 0.009
ОБЛИГАЦИЯ 0.026	ДАТЬ 0.008
ВЫПУСК ОБЛИГАЦИЯ 0.02	ДАННЫЙ 0.008
ФОНД 0.019	НОВЫЙ СИСТЕМА 0.008
УПРАВЛЯЮЩИЙ КОМПАНИЯ 0.017	ЭЛЕКТРОННЫЙ 0.008
ВЫПУСК 0.016	ДАННЫЕ 0.007
РЫНОК 0.013	РАБОТА 0.007
ДОВЕРИТЕЛЬНЫЙ УПРАВЛЕНИЕ 0.012	ДОКУМЕНТ 0.006

ТЕМА 47	**ТЕМА 48**
МИНФИН РОССИЯ 0.053	СИСТЕМА БАНК 0.113
ФНС РОССИЯ 0.025	БАНКОВСКИЙ ОПЕРАЦИЯ 0.046
ФЕДЕРАЛЬНЫЙ ОРГАН 0.017	ОПЕРАЦИЯ 0.044
РОССИЙСКИЙ ФЕДЕРАЦИЯ 0.016	ОПЕРАЦИЯ БАНК 0.041
ВНЕСЕНИЕ ИЗМЕНЕНИЕ 0.014	СОВЕРШЕНИЕ ОПЕРАЦИЯ 0.041
ФЕДЕРАЛЬНЫЙ ЗАКОН 0.014	ОСУЩЕСТВЛЕНИЕ ОПЕРАЦИЯ 0.032
ФЕДЕРАЛЬНЫЙ 0.014	ПРОВЕДЕНИЕ ОПЕРАЦИЯ 0.028
ИЗМЕНЕНИЕ 0.012	КАССОВЫЙ ОПЕРАЦИЯ 0.024
ФЕДЕРАЛЬНЫЙ СЛУЖБА 0.011	НАЛИЧНЫЙ ВАЛЮТА 0.014
ПОРЯДОК 0.011	ОБЪЕМ ОПЕРАЦИЯ 0.014
ТЕМА 49	**ТЕМА 50**
РИСК 0.063	ФИЛИАЛ БАНК 0.076
КРЕДИТНЫЙ РИСК 0.059	ОФИС БАНК 0.07
БАНКОВСКИЙ РИСК 0.053	ОБСЛУЖИВАНИЕ КЛИЕНТ 0.041
ОЦЕНКА РИСК 0.053	БАНКОВСКИЙ УСЛУГА 0.04
УРОВЕНЬ РИСК 0.036	БАНК 0.03
ОПЕРАЦИОННЫЙ РИСК 0.036	СЕТЬ БАНК 0.03
РЫНОЧНЫЙ РИСК 0.028	КЛИЕНТ 0.027
ВИД РИСК 0.028	БАНКОВСКИЙ ПРОДУКТ 0.025
РИСК ЛИКВИДНОСТЬ 0.027	БАНКОВСКИЙ ОБСЛУЖИВАНИЕ 0.023
РИСК БАНК 0.027	УСЛУГА 0.023
ТЕМА 51	**ТЕМА 52**
СИСТЕМА РОССИЯ 0.151	БЫТЬ 0.008
СИСТЕМА СТРАНА 0.101	ПРОИЗВОДСТВО 0.008
БАНКОВСКИЙ СИСТЕМА 0.08	ГОД 0.007
РАЗВИТИЕ СИСТЕМА 0.069	ЭКОНОМИЧЕСКИЙ ЗОНА 0.007
ФИНАНСОВЫЙ СИСТЕМА 0.055	НЕФТЬ 0.006
СИСТЕМА 0.031	ПРОДУКЦИЯ 0.006
СОЗДАНИЕ СИСТЕМА 0.023	ЗОНА 0.005
ОТДЕЛЬНЫЙ БАНК 0.021	ПРИРОДНЫЙ РЕСУРС 0.005
СОВЕРШЕНСТВОВАНИЕ СИСТЕМА 0.019	ПРИРОДНЫЙ РЕСУРСЫ 0.005
ЭФФЕКТИВНЫЙ СИСТЕМА 0.019	РОССИЯ 0.005
ТЕМА 53	**ТЕМА 54**
БУХГАЛТЕРСКИЙ УЧЕТ 0.075	РОСТ ЦЕНА 0.035
УЧЕТ 0.054	ЦЕНА 0.022
ДЕЯТЕЛЬНОСТЬ ОРГАНИЗАЦИЯ 0.046	УРОВЕНЬ ЦЕНА 0.02
БУХГАЛТЕРСКИЙ ОТЧЕТНОСТЬ 0.043	ЭКОНОМИЧЕСКИЙ РОСТ 0.019
БУХГАЛТЕРСКИЙ 0.037	ВЫСОКИЙ ЦЕНА 0.018
ОРГАНИЗАЦИЯ 0.024	ИНФЛЯЦИЯ 0.016
НАЛОГОВЫЙ УЧЕТ 0.02	УРОВЕНЬ ИНФЛЯЦИЯ 0.014
УПРАВЛЕНЧЕСКИЙ УЧЕТ 0.016	МИРОВОЙ ЦЕНА 0.014
БУХГАЛТЕРСКИЙ БАЛАНС 0.016	ТЕМП РОСТ 0.014
ОТЧЕТНОСТЬ 0.011	ТЕМП ИНФЛЯЦИЯ 0.013

ТЕМА 55	**ТЕМА 56**
ЧЕЛОВЕК 0.01	ДЕНЕЖНЫЙ 0.043
БЫТЬ 0.009	ДЕНЕЖНЫЙ ОБРАЩЕНИЕ 0.038
РЫНОК 0.008	ДЕНЬГИ 0.035
ДЕНЬГИ 0.007	НАЛИЧНЫЙ ДЕНЬГИ 0.034
РОССИЯ 0.007	ДЕНЕЖНЫЙ СРЕДСТВО 0.032
РЕКЛАМА 0.007	ДЕНЕЖНЫЙ МАССА 0.031
БОЛЬШОЙ 0.006	ДЕНЕЖНЫЙ СРЕДСТВА 0.03
БОЛЕЕ 0.006	НАЛИЧНЫЕ ДЕНЬГИ 0.03
ЛЮДИ 0.006	ДЕНЕЖНЫЙ ЕДИНИЦА 0.02
УСЛУГА 0.005	ДЕНЕЖНЫЙ РЫНОК 0.019
ТЕМА 57	**ТЕМА 58**
БАНК 0.051	СЧЕТ СРЕДСТВО 0.118
РАЗВИТИЕ БИЗНЕС 0.022	СЧЕТ СРЕДСТВА 0.104
БАНКА 0.022	ДЕНЕЖНЫЙ СРЕДСТВО 0.066
ГОД 0.018	ДЕНЕЖНЫЙ СРЕДСТВА 0.065
БИЗНЕС 0.012	СРЕДСТВО 0.056
РОЗНИЧНЫЙ БИЗНЕС 0.01	СРЕДСТВА 0.055
КОМПАНИЯ 0.01	ИСПОЛЬЗОВАНИЕ СРЕДСТВО 0.026
БЫТЬ 0.01	ИСПОЛЬЗОВАНИЕ СРЕДСТВА 0.025
КЛИЕНТ 0.01	СОБСТВЕННЫЙ СРЕДСТВО 0.022
НОВЫЙ 0.01	СОБСТВЕННЫЙ СРЕДСТВА 0.022
ТЕМА 59	**ТЕМА 60**
РУКОВОДИТЕЛЬ БАНК 0.189	УЧАСТИЕ БАНК 0.125
РУКОВОДСТВО БАНК 0.168	ПОРТФЕЛЬ БАНК 0.087
ОТДЕЛЬНЫЙ БАНК 0.13	РЕГИОНАЛЬНЫЙ БАНК 0.073
АМЕРИКАНСКИЙ БАНК 0.11	ЗАРУБЕЖНЫЙ БАНК 0.052
БАНК 0.041	МАЛЫЙ БИЗНЕС 0.052
СОТРУДНИК БАНК 0.028	СРЕДНИЙ БИЗНЕС 0.046
ЗАРУБЕЖНЫЙ БАНК 0.026	СРЕДНЕЕ БИЗНЕС 0.039
БАНКА 0.018	МАЛЫЙ 0.026
КОНКРЕТНЫЙ БАНК 0.013	МАЛЫЙ ПРЕДПРИЯТИЕ 0.025
БАНКОВСКИЙ БИЗНЕС 0.008	РОССИЙСКИЙ БАНК 0.017
ТЕМА 61	**ТЕМА 62**
ЭКОНОМИЧЕСКИЙ РАЗВИТИЕ 0.026	ПРОЦЕНТНЫЙ СТАВКА 0.032
РАЗВИТИЕ СТРАНА 0.025	СТАВКА 0.03
РАЗВИТИЕ ЭКОНОМИКА 0.02	ПОСЛЕДНИЙ ГОД 0.02
ВЫСОКИЙ УРОВЕНЬ 0.02	РОСТ 0.017
ДОХОД НАСЕЛЕНИЕ 0.018	ГОД 0.016
УРОВЕНЬ ДОХОД 0.017	ТЕМП РОСТ 0.014
УРОВЕНЬ РАЗВИТИЕ 0.017	УРОВЕНЬ 0.011
НИЗКИЙ УРОВЕНЬ 0.015	РОСТ ОБЪЕМ 0.01
УРОВЕНЬ 0.014	СТАВКА РЕФИНАНСИРОВАНИЕ 0.01
НАСЕЛЕНИЕ 0.014	ПРОЦЕНТНЫЙ 0.01

ТЕМА 63	**ТЕМА 64**
РЫНОЧНЫЙ ЭКОНОМИКА 0.028	РОССИЯ 0.027
НАЦИОНАЛЬНЫЙ ЭКОНОМИКА 0.021	ПРЕЗИДЕНТ РОССИЯ 0.024
ЭКОНОМИЧЕСКИЙ СИСТЕМА 0.02	РОССИЙСКИЙ 0.018
ЭКОНОМИКА 0.02	ВСТУПЛЕНИЕ РОССИЯ 0.015
МИРОВОЙ ЭКОНОМИКА 0.018	РОССИЙСКИЙ ФЕДЕРАЦИЯ 0.014
ЭКОНОМИКА РОССИЯ 0.015	СТРАНА СНГ 0.014
ЭКОНОМИЧЕСКИЙ 0.013	СТРАНА 0.011
ЭКОНОМИЧЕСКИЙ РОСТ 0.013	БЫТЬ 0.01
РОССИЙСКИЙ ЭКОНОМИКА 0.013	МЕЖДУНАРОДНЫЙ 0.008
ГОСУДАРСТВО 0.01	СОВЕТ 0.008
ТЕМА 65	**ТЕМА 66**
РАБОТА 0.021	ИНОСТРАННЫЙ БАНК 0.213
СОТРУДНИК 0.017	КРУПНЕЙШИЙ БАНК 0.153
РАБОТНИК 0.013	ИНОСТРАННЫЙ БАНКА 0.079
ОБУЧЕНИЕ 0.012	ДОЧЕРНИЙ БАНК 0.073
СПЕЦИАЛИСТ 0.012	МЕЖДУНАРОДНЫЙ БАНК 0.045
ПЕРСОНАЛ 0.011	ЗАРУБЕЖНЫЙ БАНК 0.038
ТРУД 0.011	БАНК РОССИЯ 0.026
ОРГАНИЗАЦИЯ 0.01	БАНК 0.022
ПРОГРАММА 0.009	ОТЕЧЕСТВЕННЫЙ БАНК 0.019
РЫНОК ТРУД 0.009	ИНОСТРАННЫЙ КАПИТАЛ 0.018
ТЕМА 67	**ТЕМА 68**
РОССИЙСКИЙ БАНК 0.14	ФИНАНСОВЫЙ ИНФОРМАЦИЯ 0.023
РОССИЙСКИЙ БАНКА 0.076	БАНКОВСКИЙ НАДЗОР 0.019
БАНК 0.043	ОРГАН НАДЗОР 0.018
РОССИЙСКИЙ РЫНОК 0.037	НАДЗОР 0.016
БАНКОВСКИЙ РЫНОК 0.034	СИСТЕМА КОНТРОЛЬ 0.015
БАНКА 0.023	ОРГАН 0.015
ИНОСТРАННЫЙ БАНК 0.019	НАДЗОРНЫЙ ОРГАН 0.014
РОССИЙСКИЙ 0.019	ОРГАН УПРАВЛЕНИЕ 0.012
РЫНОК 0.017	КОРПОРАТИВНЫЙ УПРАВЛЕНИЕ 0.011
РОССИЯ 0.011	БАНКОВСКИЙ ГРУППА 0.011
ТЕМА 69	**ТЕМА 70**
ПРЕДПРИЯТИЕ 0.055	ФИНАНСОВЫЙ АКАДЕМИЯ 0.018
ДЕЯТЕЛЬНОСТЬ ПРЕДПРИЯТИЕ 0.042	ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ 0.017
РОССИЙСКИЙ ПРЕДПРИЯТИЕ 0.015	ГОСУДАРСТВЕННЫЙ 0.013
ЗАТРАТЫ 0.013	НАУКА 0.011
ПРОМЫШЛЕННЫЙ ПРЕДПРИЯТИЕ 0.013	ГОСУДАРСТВЕННЫЙ РЕГУЛИРОВАНИЕ 0.011
ПРОДУКЦИЯ 0.012	ЭКОНОМИКА РОССИЯ 0.01
ПРОИЗВОДСТВО 0.011	ЭКОНОМИЧЕСКИЙ СИСТЕМА 0.01
ОБЪЕМ ПРОДАЖА 0.011	ЭКОНОМИЧЕСКИЙ 0.009
ОБЪЕМ ПРОИЗВОДСТВО 0.009	ЗАЩИТИТЬ 0.009
СРЕДНИЙ ПРЕДПРИЯТИЕ 0.009	ЗАЩИЩЕННЫЙ 0.009

ТЕМА 71	**ТЕМА 72**
БАНКОВСКИЙ ДЕЯТЕЛЬНОСТЬ 0.072	ОЦЕНКА 0.02
БАНКОВСКИЙ СИСТЕМА 0.072	ПОКАЗАТЕЛЬ 0.019
БАНКОВСКИЙ УСЛУГА 0.072	ДАННЫЙ 0.015
БАНКОВСКИЙ 0.07	ДАТЬ 0.015
БАНКОВСКИЙ СЕКТОР 0.065	ДАННЫЙ ПОКАЗАТЕЛЬ 0.013
БАНКОВСКИЙ БИЗНЕС 0.045	МЕТОД ОЦЕНКА 0.012
БАНКОВСКИЙ СФЕРА 0.037	АНАЛИЗ 0.01
БАНКОВСКИЙ ДЕЛО 0.032	ДАННЫЕ 0.009
БАНКОВСКИЙ НАДЗОР 0.032	ЗНАЧЕНИЕ 0.008
БАНКОВСКИЙ ОПЕРАЦИЯ 0.03	МЕТОД 0.008
ТЕМА 73	**ТЕМА 74**
БАНКОВСКИЙ КАРТА 0.078	РЫНОК РОССИЯ 0.058
БАНКОВСКИЙ КАРТ 0.068	УЧАСТНИК РЫНОК 0.048
КРЕДИТНЫЙ КАРТА 0.068	ФИНАНСОВЫЙ РЫНОК 0.044
КРЕДИТНЫЙ КАРТ 0.057	РЫНОК 0.042
КАРТА 0.051	ДЕНЕЖНЫЙ РЫНОК 0.031
ПЛАТЕЖНЫЙ КАРТА 0.04	КРЕДИТНЫЙ РЫНОК 0.031
КАРТ 0.04	РАЗВИТИЕ РЫНОК 0.028
ПЛАТЕЖНЫЙ КАРТ 0.036	МЕЖБАНКОВСКИЙ РЫНОК 0.028
ПЛАСТИКОВЫЙ КАРТА 0.032	ВАЛЮТНЫЙ РЫНОК 0.026
ПЛАТЕЖНЫЙ СИСТЕМА 0.023	СРОЧНЫЙ РЫНОК 0.025
ТЕМА 75	**ТЕМА 76**
РАЗВИТЫЙ СТРАНА 0.051	БЫТЬ 0.037
СТРАНА 0.051	ЕСТЬ 0.009
РАЗВИТОЙ СТРАНА 0.051	ЧЕЛОВЕК 0.008
СТРАНА МИР 0.033	КОГДА 0.008
СТРАНА МИРО 0.033	БАНК 0.007
ЭКОНОМИКА СТРАНА 0.031	ДЕНЬГИ 0.007
БОЛЬШИНСТВО СТРАНА 0.028	ОЧЕНЬ 0.007
РЯД СТРАНА 0.025	МОЧЬ 0.007
ЕВРОПЕЙСКИЙ СТРАНА 0.023	ГОВОРИТЬ 0.006
РАЗНЫЙ СТРАНА 0.018	ГОД 0.006
ТЕМА 77	**ТЕМА 78**
БАНК РФ 0.231	СИСТЕМА СТРАХОВАНИЕ 0.1
БАНК РОССИЯ 0.171	СТРАХОВАНИЕ ВКЛАД 0.074
ТРЕБОВАНИЕ БАНК 0.17	БАНКОВСКИЙ ВКЛАД 0.07
ЦЕНТРАЛЬНЫЙ БАНК 0.13	ВКЛАД 0.055
БАНКА РОССИЯ 0.109	БАНК 0.04
АКТ БАНК 0.037	ВКЛАД НАСЕЛЕНИЕ 0.022
ОБЯЗАТЕЛЬСТВО БАНК 0.021	УЧАСТНИК СИСТЕМА 0.021
БАНКА 0.017	БАНК РОССИЯ 0.021
БАНКИЙ РОССИЯ 0.012	БАНКА 0.019
БАНК 0.009	ОБЯЗАТЕЛЬНЫЙ СТРАХОВАНИЕ 0.018

ТЕМА 79	**ТЕМА 80**
ДАННЫЙ БАНК 0.337	ПРАВЛЕНИЕ БАНК 0.142
ДАННЫЙ БАНКА 0.164	ПРЕДСЕДАТЕЛЬ ПРАВЛЕНИЕ 0.048
КОНКРЕТНЫЙ БАНК 0.046	ЗАМЕСТИТЕЛЬ ПРЕДСЕДАТЕЛЬ 0.034
РУКОВОДСТВО БАНК 0.024	БАНК 0.022
БАНК 0.018	ПРЕДСЕДАТЕЛЬ СОВЕТ 0.02
ДАТЬ 0.018	ПРАВЛЕНИЕ КБ 0.016
ОБЯЗАТЕЛЬСТВО БАНК 0.017	БАНК РАЗВИТИЕ 0.016
БАНКА 0.017	ОАО 0.016
ДАННЫЙ 0.015	ЧЛЕН ПРАВЛЕНИЕ 0.015
ДАННЫЕ 0.01	ГЕНЕРАЛЬНЫЙ ДИРЕКТОР 0.014
ТЕМА 81	**ТЕМА 82**
СУММА НДС 0.021	СТРАХОВОЙ КОМПАНИЯ 0.08
ТОВАР 0.017	СТРАХОВОЙ 0.058
НДС 0.016	СТРАХОВАНИЕ 0.048
РФ 0.013	СТРАХОВОЙ РЫНОК 0.047
СУММА НАЛОГ 0.013	ДОГОВОР СТРАХОВАНИЕ 0.029
РАСХОД 0.012	СТРАХОВАНИЕ ЖИЗНЬ 0.027
УСЛУГА 0.012	СТРАХОВОЙ СЛУЧАЙ 0.025
ОКАЗАНИЕ УСЛУГА 0.012	СТРАХОВОЙ ОРГАНИЗАЦИЯ 0.022
СТОИМОСТЬ 0.011	КОМПАНИЯ 0.019
НАЛОГ 0.01	СТРАХОВЩИК 0.017
ТЕМА 83	**ТЕМА 84**
ПЛАТЕЖНЫЙ СИСТЕМА 0.055	КАПИТАЛ БАНК 0.274
СИСТЕМА 0.034	АКТИВ БАНК 0.168
СИСТЕМА БАНК 0.032	ДЕЯТЕЛЬНОСТЬ БАНК 0.108
УЧАСТНИК СИСТЕМА 0.025	БАНК 0.033
РАСЧЕТНЫЙ СИСТЕМА 0.025	СОБСТВЕННЫЙ КАПИТАЛ 0.026
ДЕНЕЖНЫЙ ПЕРЕВОД 0.024	ДОСТАТОЧНОСТЬ КАПИТАЛ 0.026
ПЛАТЕЖНЫЙ 0.022	КАПИТАЛ 0.026
ПЕРЕВОД 0.021	БАНКОВСКИЙ КАПИТАЛ 0.021
ЦЕНТРАЛЬНЫЙ БАНК 0.015	ЦЕНТРАЛЬНЫЙ БАНК 0.02
ПЛАТЕЖ 0.015	ПОРТФЕЛЬ БАНК 0.017
ТЕМА 85	**ТЕМА 86**
БАНК РОССИЯ 0.12	БАНК РАЗВИТИЕ 0.16
БАНКА РОССИЯ 0.062	БАНКОВСКИЙ КРЕДИТ 0.043
БАНК 0.054	БАНКА РАЗВИТИЕ 0.041
ЦЕНТРАЛЬНЫЙ БАНК 0.038	КРЕДИТНЫЙ РЕСУРС 0.03
БАНКОВСКИЙ СИСТЕМА 0.037	КРЕДИТНЫЙ РЕСУРСЫ 0.03
БАНКА 0.025	РОССИЙСКИЙ БАНК 0.028
БАНКОВСКИЙ СООБЩЕСТВО 0.023	КОММЕРЧЕСКИЙ БАНК 0.021
БАНКОВСКИЙ 0.022	РАЗВИТИЕ 0.016
БАНКОВСКИЙ СЕКТОР 0.017	КРЕДИТНЫЙ КООПЕРАТИВ 0.014
БАНКОВСКИЙ НАДЗОР 0.017	ПРОГРАММА РАЗВИТИЕ 0.012

ТЕМА 87	**ТЕМА 88**
ФЕДЕРАЛЬНЫЙ ЗАКОН 0.037	ДЕЯТЕЛЬНОСТЬ 0.045
ЗАКОН 0.027	ОСНОВНОЙ ПРОБЛЕМА 0.031
РФ 0.025	ВИД ДЕЯТЕЛЬНОСТЬ 0.03
ЮРИДИЧЕСКИЙ ЛИЦО 0.024	ОСНОВНЫЙ ПРОБЛЕМА 0.03
КОДЕКС РФ 0.024	СФЕРА ДЕЯТЕЛЬНОСТЬ 0.028
ПРАВИТЕЛЬСТВО РФ 0.023	ЭКОНОМИЧЕСКИЙ ДЕЯТЕЛЬНОСТЬ 0.024
ФИЗИЧЕСКИЙ ЛИЦО 0.021	ФИНАНСОВЫЙ ДЕЯТЕЛЬНОСТЬ 0.023
ЗАКОНОДАТЕЛЬСТВО РФ 0.021	ХОЗЯЙСТВЕННЫЙ ДЕЯТЕЛЬНОСТЬ 0.019
ЛИЦО 0.02	ОРГАНИЗАЦИЯ 0.019
РОССИЙСКИЙ ФЕДЕРАЦИЯ 0.012	ОСНОВНОЙ ЗАДАЧА 0.018
ТЕМА 89	**ТЕМА 90**
ФИНАНСОВЫЙ РЕСУРС 0.082	ПРАВООХРАНИТЕЛЬНЫЙ ОРГАН 0.01
ФИНАНСОВЫЙ РЕСУРСЫ 0.08	ОТМЫВАНИЕ 0.009
ФИНАНСОВЫЙ СИСТЕМА 0.057	ДЕНЕЖНЫЙ СРЕДСТВО 0.008
ФИНАНСОВЫЙ 0.047	ДЕНЕЖНЫЙ СРЕДСТВА 0.008
ФИНАНСОВЫЙ СЕКТОР 0.043	ОТМЫВАНИЕ ДЕНЬГИ 0.008
ФИНАНСОВЫЙ СОСТОЯНИЕ 0.043	БАНК 0.008
ФИНАНСОВЫЙ РЫНОК 0.041	ОПЕРАЦИЯ 0.008
ФИНАНСОВЫЙ УСТОЙЧИВОСТЬ 0.032	ПРЕСТУПНЫЙ ПУТЬ 0.008
ФИНАНСОВЫЙ ПОЛОЖЕНИЕ 0.028	ИНФОРМАЦИЯ 0.007
ФИНАНСОВЫЙ ПОТОК 0.027	ЛИЦО 0.007
ТЕМА 91	**ТЕМА 92**
БЫТЬ 0.007	КОММЕРЧЕСКИЙ БАНК 0.304
ГОД 0.004	КОММЕРЧЕСКИЙ БАНКА 0.106
ЧЕЛОВЕК 0.004	БАНК 0.093
ВРЕМЯ 0.003	ДЕЯТЕЛЬНОСТЬ БАНК 0.091
МОЖНО 0.003	ПОЛИТИКА БАНК 0.077
ЖИЗНЬ 0.003	ОПЕРАЦИЯ БАНК 0.07
ДОМ 0.003	ЦЕНТРАЛЬНЫЙ БАНК 0.057
РУССКИЙ 0.003	БАНКА 0.037
ЖЕНЩИНА 0.002	ЦЕНТРАЛЬНЫЙ БАНКА 0.018
ДЕНЬ 0.002	БАНКОВСКИЙ 0.007
ТЕМА 93	**ТЕМА 94**
ГОСУДАРСТВЕННЫЙ БАНК 0.198	ИНВЕСТИЦИОННЫЙ ПРОЕКТ 0.102
ЧАСТНЫЙ БАНК 0.128	ИНВЕСТИЦИОННЫЙ ДЕЯТЕЛЬНОСТЬ 0.088
ЦЕНТРАЛЬНЫЙ БАНК 0.028	ИНВЕСТИЦИОННЫЙ 0.087
БАНК 0.026	ИНВЕСТИЦИОННЫЙ ПРОЦЕСС 0.049
БАНКОВСКИЙ КРИЗИС 0.015	ИНВЕСТИЦИОННЫЙ РЕСУРС 0.046
ГОСУДАРСТВЕННЫЙ 0.011	ИНВЕСТИЦИОННЫЙ КЛИМАТ 0.034
БАНКА 0.011	ИНВЕСТИЦИОННЫЙ ФОНД 0.025
АГЕНТСТВО 0.01	ИНВЕСТИЦИЯ 0.025
КОЛЛЕКТОРСКИЙ АГЕНТСТВО 0.009	НПФ 0.022
РЕЙТИНГОВЫЙ АГЕНТСТВО 0.009	ПРОЕКТ 0.015

ТЕМА 95	**ТЕМА 96**
БЫТЬ 0.021	РАБОТА БАНК 0.212
ГОД 0.017	БИЗНЕС БАНК 0.144
БИЗНЕС 0.013	РИСК БАНК 0.091
ПОСЛЕДНИЙ ГОД 0.012	ДЕЯТЕЛЬНОСТЬ БАНК 0.073
РАБОТА 0.009	РАЗВИТИЕ БАНК 0.052
УЗКИЙ 0.009	БАНК 0.042
УЖ 0.009	БАНКА 0.029
УЖЕ 0.008	КОНКРЕТНЫЙ БАНК 0.022
РАБОТАТЬ 0.008	ОПЕРАЦИЯ БАНК 0.021
СЕГОДНЯ 0.007	РУКОВОДСТВО БАНК 0.019
ТЕМА 97	**ТЕМА 98**
ПОЛИТИКА БАНК 0.044	КЛИЕНТ БАНК 0.424
ОСНОВНОЙ ЦЕЛЬ 0.035	СОТРУДНИК БАНК 0.152
ОСНОВНЫЙ ЦЕЛЬ 0.034	СТОРОНА БАНК 0.122
БАНК РОССИЯ 0.03	БАНК 0.09
БАНКИЙ РОССИЯ 0.03	БАНКА 0.053
БАНКА РОССИЯ 0.027	КЛИЕНТ 0.026
ЭКОНОМИКА РОССИЯ 0.025	БАНКОВСКИЙ 0.01
ЦЕНТРАЛЬНЫЙ БАНК 0.017	КОРПОРАТИВНЫЙ КЛИЕНТ 0.003
ДЕНЕЖНО-КРЕДИТНЫЙ ПОЛИТИКА 0.017	СОТРУДНИК 0.003
ОПЕРАЦИЯ БАНК 0.013	УСЛУГА 0.002
ТЕМА 99	**ТЕМА 100**
КЛИЕНТ 0.034	НАЛОГОВЫЙ ОРГАН 0.046
ДАННЫЙ СЛУЧАЙ 0.019	НАЛОГОВЫЙ 0.045
ПОТЕНЦИАЛЬНЫЙ КЛИЕНТ 0.018	НАЛОГОВЫЙ КОДЕКС 0.034
МОЧЬ 0.018	НАЛОГОВЫЙ ЗАКОНОДАТЕЛЬСТВО 0.029
СЛУЧАЙ 0.015	НАЛОГОВЫЙ БАЗА 0.029
БОЛЬШИНСТВО СЛУЧАЙ 0.01	НАЛОГОВЫЙ ДЕКЛАРАЦИЯ 0.027
БЫТЬ 0.008	НАЛОГОВЫЙ ПЕРИОД 0.026
УСЛУГА 0.008	НАЛОГОВЫЙ УЧЕТ 0.025
ЧАСТНЫЙ КЛИЕНТ 0.007	НАЛОГОВЫЙ АГЕНТ 0.024
РЯД СЛУЧАЙ 0.006	УПЛАТА НАЛОГ 0.024