Санкт-Петербургский государственный университет
Филологический факультет
Кафедра математической лингвистики

Седова Анастасия Георгиевна

ТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ РУССКОЯЗЫЧНЫХ ТЕКСТОВ С ОПОРОЙ НА ЛЕММЫ И ЛЕКСИЧЕСКИЕ КОНСТРУКЦИИ

Выпускная квалификационная работа по направлению 45.03.02 «Лингвистика», образовательная программа «Прикладная, экспериментальная и математическая лингвистика (английский язык)»

Научный руководитель:
доц., к.ф.н. **Митрофанова О.А.**

Санкт-Петербург
2017

Оглавление

ВВЕДЕНИЕ

Данная работа посвящена активно развивающемуся в последние годы направлению вероятностного тематического моделирования, суть которого заключается в создании семантических моделей корпуса текстов на основе разновидностей нечеткой кластеризации лексики. ***Вероятностные тематические модели (probabilistic topic model)*** коллекций текстовых документов представляют текстовый документ как вероятностную смесь тем, каждая из которых является дискретным распределением на множестве терминов. Таким образом, тематическая модель выступает как средство обобщения, систематизации и смыслового поиска для больших текстовых коллекций. Особенно эффективно тематические модели используются для выявления скрытых структур и поиска неявных зависимостей в данных, поскольку они позволяют определять тематику текстов и служат для решения задач классификации и кластеризации документов (то есть, задач разделения документов на два или более взаимно исключающих класса), поиска похожих документов, выявления и анализа различных временных трендов (Митрофанова 2014).

Автоматическое определение тематики текстов активно применяется для разбиения текстов по группам на основе семантической близости содержания. С помощью тематических моделей решаются разнообразные актуальные задачи обработки естественного языка; например, задачи выявления научных интересов авторов, обнаружения скрытых ассоциативных связей между отдельными исследователями или группами людей, выявления тенденций в развитии научных направлений, определения эмоциональной окраски текстов, осуществления автоматического аннотирования и индексирования документов (то есть, поиска наиболее соответствующих запросу документов и их ранжирование

по данному запросу) и так далее. Кроме моделирования текстов, тематические модели широко используются для решения задач распознавания объектов и рукописного текста, кластеризации изображений и создания подписей для различных объектов, а также в других науках, например, в биоинформатике.

Традиционно тема представляется в виде номера темы и некоторого количества слов, вероятность принадлежности которых к данной теме наиболее высока (Нокель, 2015; Нокель, Лукашевич, 2015). Желаемое количество выделяемых тем, а также количество слов, представляющих данную тему, задается пользователем вручную. В дальнейшем в большинстве случаев пользователю предоставляется право самостоятельно интерпретировать данные, заложенные в выдаче.

В базовых алгоритмах тематического моделирования темы представлены исключительно униграммами. Это влечет за собой ухудшение точности и повышает сложность содержательной интерпретации выделяемых тем, особенно в случае некомпозиционных словосочетаний, значение которых не сводится к сумме значений входящих в них слов: например, *железная дорога* не сводится к значению слов *железная* и *дорога* соответственно (Нокель, Лукашевич 2015). Таким образом, добавление в темы расширение тем за счет *n*-грамм представляет собой ***актуальную исследовательскую задачу*.**

В последнее время было проведено несколько исследований и предложено несколько основывающихся на разных методах подходов к решению данной проблемы (Wallach 2006; Wang, McCallum, Wei 2007), однако многие из них снижают качество модели или же излишне усложняют её (Нокель, Лукашевич 2015). В данной работе была

предпринята попытка предложить новый метод, который бы действительно упрощал интерпретацию тем и повышал их точность.

Целью данного исследование является исследование существующих методов тематического моделирования, а также разработка алгоритма, позволяющего извлекать из корпуса текстов биграммы и триграммы и добавлять их в выделяемые темы наряду с униграммами.

Для достижения данной цели решаются следующие задачи:

1) исследование вероятностных тематических моделей, выбор модели, наиболее подходящей для целей данной работы;
2) исследование существующих методов добавления *n*-грамм в выделяемые темы;
3) разработка алгоритма для автоматического добавления биграмм, адаптированного для русскоязычных текстов и реализованного на языке программирования Python;
4) оценка работы предлагаемого алгоритма на двух русскоязычных корпусах: на корпусе специальных текстов по радиоэлектронике, ракетостроению и технике и на корпусе текстов на лингвистическую тематику.

Объектом исследования является тематическое моделирование русскоязычных текстов, ***предметом исследования*** – алгоритмы автоматического добавления биграмм в выделяемые темы. В работе используются ***методы*** статистического, лингвистического анализа данных и эксперимент. ***Материалом*** исследования является два русскоязычных корпуса специальных текстов: по радиоэлектронике, ракетостроению и технике и на лингвистическую тематику.

Данная работа состоит из введения, трех глав, заключения, списка литературы и приложений. В первой главе рассматриваются теоретические вопросы, связанные с выделением тем, проблемы тематического моделирования и основные виды тематических моделей. Во второй главе описаны существующие алгоритмы расширения тем с помощью биграмм. Третья глава посвящена теоретическому описанию предлагаемого алгоритма для русского языка. В четвертой главе обсуждаются и оцениваются полученные результаты работы алгоритма на материале двух русскоязычных корпусов.

1. ВЕРОЯТНОСТНОЕ ТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ

1.1. Основные понятия и термины.

Документ (document) – это объект, состоящий из множества слов, словосочетаний, специальных символов, таблиц, иллюстраций и т.д. Документ d можно записать в виде вектора слов $w_d = \{w_{d1} + \ldots + w_{dn}\}$, где w_{di}— i-е слово в документе d. Коллекция, состоящая из D документов, может быть представлена с помощью векторов слов $D = \{w_1 + \ldots + w_D\}$, где w_d — вектор слов документа d.

Лингвистический, или языковой корпус текстов (text corpora, corpus) – большая совокупность текстовых документов, представленных в электронном виде. Обязательным свойством корпуса текстов является структурированность и унифицированность (Захаров 2005). Корпуса текстов активно применяются для решения многих лингвистических задач, поскольку при достаточно большом объеме они обеспечивают полное отражения наиболее типичных употреблений слов в текстах, написанных на данном языке, и, таким образом, служат источником сведений о различных единицах языка и речи.

Модель мешка слов (bag of words) – модель для анализа текстов, которая учитывает только частоту слов, но не их порядок. Данная модель хорошо подходит для многих методом тематического моделирования, поскольку она позволяет обнаруживать неявные взаимосвязи между словами с учетом полисемии (Clark, Fox, Lappin 2013).

Тема (topic, latent topic), или скрытый паттерн (hidden pattern) – это дискретное вероятностное распределение в пространстве слов заданного словаря (Daud et al 2010). Формально говоря, темы являются

результатом би-кластеризации, то есть одновременной кластеризации и слов, и документов по их семантической близости. В Таблице 1 представлен пример трех тем и десяти самых употребительных слов в каждой из них, выделенных из корпуса Энрона[1], с помощью метода LDA (Darling 2011).

Таблица 1. Три темы, выделенные с помощью метода LDA из корпуса Энрона

"environment"	***"travel"***	***"fantasy football"***
emission	travel	game
environmental	hotel	yard
air	roundtrip	defense
permit	fares	allowed
plant	special	fantasy
facility	offer	point
unit	city	passing
epa	visit	rank
water	miles	against
station	deal	team

Внутри каждой темы слова текста распределяются с некоторыми вероятностями. В качестве примера в Таблице 2 представлены примеры двух тем (*искусство* и *образование*), по каждой из которых приведены 15 наиболее употребительных слов и соответствующие им вероятности (Daud et al 2010).

[1] Enron Email Dataset http://www.cs.cmu.edu/~enron/

Таблица 2. Примеры двух тем (*искусство* и *образование*).

Arts		Education	
Word	Probability	Word	Probability
New	0.03741	School	0.07344
Film	0.03626	Students	0.05702
Show	0.02753	Schools	0.04136
Music	0.02151	Education	0.02605
Movie	0.01854	Teachers	0.02465
Play	0.01124	High	0.02122
Musical	0.01109	Public	0.02026
Best	0.00989	Teacher	0.02006
Actor	0.00966	Bennett	0.01766
First	0.00899	Manigat	0.01746
York	0.00895	Namphy	0.01478
Opera	0.00870	State	0.0143
Theater	0.00854	President	0.01359
Actress	0.00817	Elementary	0.01219
Love	0.00806	Haiti	0.01211

Тематическим моделированием при этом называется восстановление вероятностных распределений всех тем в тексте, рассматриваемом как случайная независимая выборка слов (*мешок слов*), порожденная некоторыми темами. При этом нет прямой зависимости между объемом документа и числом порождаемых им тем. Порядок тем при различных запусках алгоритма может варьироваться, что обусловлено ***свойством неупорядоченности***, или ***перестановочности (exchangeability)*** тем.

Одной из важных подзадач в тематическом моделировании является определение оптимального числа тем, поскольку этот параметр устанавливается пользователем. Если это число недостаточно большое, то результаты будут слишком общие. Наоборот, завышение же числа приводит к невозможности разумной интерпретации.

Правила порождения единиц текста вероятностной смесью тем задает тематическая ***порождающая модель (generative model)***, описывающая вероятностный закон генерации случайных значений переменных. Основной идеей тематических моделей является использование скрытого слоя тематических переменных. Подразумевается, что данные являются воплощением некой случайной величины, распределенной в соответствии с порождающей моделью. В этом случае задачу тематического моделирования можно сформулировать как восстановление по наблюдаемым данным максимально реальных значений латентных переменных, определяющих вероятностную смесь тем.

Априорное распределение вероятностей (prior probability distribution, prior) неопределённой величины — распределение вероятностей этой величины, рассматриваемое в противоположность её условному распределению при некотором дополнительном условии (Математическая Энциклопедия. Ред. коллегия: И. М. Виноградов и др. 1997).

Апостериорная вероятность — условное распределение вероятностей какой-либо случайной величины при некотором условии, рассматриваемое в противоположность ее безусловному или априорному распределению (Математическая энциклопедия. Ред. коллегия: И. М. Виноградов и др, 1977).

Мультиномиальное распределение – это совместное распределение вероятностей случайных величин, каждая из которых есть число появлений одного из нескольких взаимоисключающих событий при повторных независимых испытаниях. Мультиномиальное распределение описывает эксперимент с k возможными исходами. Мультиноминальная функция

вероятности $f(y_1, \dots, y_k, p_1, \dots, p_k)$ равна вероятности получить j-й исход y_i раз, где $y_1 + \cdots + y_k = n$ (Большая советская энциклопедия 1969-1978).

Распределение Дирихле *Dir (α)* – это семейство непрерывных многомерных вероятностных распределений параметризованных вектором α неотрицательных вещественных чисел. Его функция вероятности $f(x_1, \dots, x_k, a_1, \dots, a_k)$ имеет k-мерный вектор неотрицательных вещественных параметров $\alpha = (\alpha_1, \dots, \alpha_k)$ и определяется как доверительная вероятность того, что вероятность каждого из k взаимоисключающих исходов равна x_i при условии, что каждое событие наблюдалось $\alpha_i - 1$ раз (Daud et al 2010).

1.2. Ориентированные вероятностные тематические модели

Ориентированные вероятностные тематические модели (directed probabilistic topic models, DPTM) – метод самообучения (unsupervised learning), основанный на Байесовских сетях. В рамках этого метода оценка близости документа к теме имеет вероятностный смысл и может интерпретироваться как доля содержимого документа, относящаяся к данной теме. Каждая тема описывается дискретным распределением на множестве терминов, а каждый документ — дискретным распределением на множестве тем. Предполагается, что коллекция документов — это последовательность терминов, выбранных случайно и независимо из смеси таких распределений, и ставится задача восстановления компонент смеси по выборке. В моделях подразумевается существование латентных взаимосвязей между некими объектами (например, авторами текстов, пользователями, журналами и т.д.), что оказывает влияние на словоупотребление. Более широкое распространение получили модели, допускающие наличие двух и более тем в документе. Данные модели

применяются для выявления тематики текстов в больших коллекциях документов.

Далее мы рассмотрим некоторые из основных вероятностных тематических моделей.

Вероятностный латентный семантический анализ (PLSA)

Первой статистически обоснованной вероятностной тематической моделью стал ***вероятностный латентный семантический анализ (probabilistic latent semantic analysis, PLSA)***, основанный на введении слоя скрытых переменных для описания тематик документов из корпуса текстов (Hofmann 1999).

На Рис. 1 представлено графическое изображение модели.

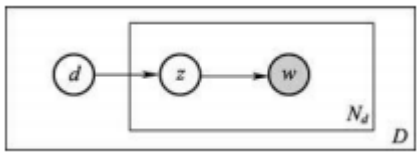

Рис. 1. Вероятностная модель PLSA

В качестве основополагающего предположения выступает идея обусловленности совместного появления пары (документ, слово) латентными переменными $z \in T = \{z_1, \dots, z_t\}$, где T – число тем в документе. Совместное распределение на парах $d \times w$ (документ × слово) определяется как смесь распределений:

$$p\,(d,w) = p(d)p(w|d), \text{где } p(w|d) = \sum_{z\in T} p(w|z)p(z|d).$$

Слова появляются в документе в зависимости от тем, но независимо друг от друга, как и пары (*d, w*), поскольку в данном методе используется модель *мешка слов*.

В рамках данного анализа каждый документ представляется числовым вектором, который задается числами, отображающими доли некой темы в документе. Соответственно, при рассмотрении бóльшего числа документов размер вектора линейно увеличивается. Поскольку модель не предоставляет информацию о законе распределения тем в документе, с помощью данной модели можно сформулировать лишь закон порождения слов конкретного документа, но не закон порождения документов. Также недостатком PLSA является склонность к переобучению из-за большого числа параметров, которые находятся в прямой зависимости от числа документов; это затрудняет работу модели на больших корпусах данных (Кольцов, Кольцова, Митрофанова, Шиморина 2014).

Модель латентного размещения Дирихле (LDA)

В качестве усовершенствованной версии модели PLSA была предложена ***модель латентного размещения Дирихле (latent Dirichlet allocation, LDA)***, которая также предполагает, что каждое слово в документе порождено некоторой латентной темой, определяющейся вероятностным распределением на множестве всех слов текста (Blei, Ng, Jordan 2003). Однако, в отличие от PLSA, данная модель в явном виде задает распределение слов в каждой теме, априорное распределение тем в документе, а также задает модель порождения не только слов, но и документов.

В рамках LDA (также как и в PLSA) считается, что в документе может быть выделено более одной темы; все выделенные темы считаются независимыми друг от друга. С ростом числа документов в коллекции число параметров вектора не увеличивается. LDA осуществляет ***мягкую кластеризацию*** (***soft clustering***; то есть, каждое слово и каждый документ

может относиться к нескольким темам одновременно с определенными вероятностями) и наилучшим образом подходит для описания кластерных структур (Воронцов 2013). Данная модель относится к методам *самообучения* или *обучения без учителя (unsupervised learning)*.

На Рис. 2 представлено графическое изображение модели.

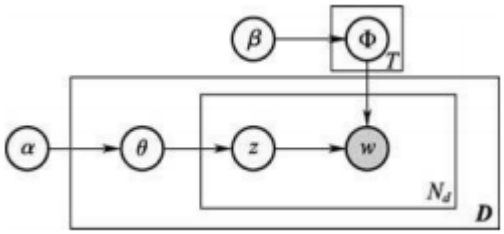

Рис. 2. Вероятностная модель LDA

На первом шаге для каждого документа d выбирается случайный вектор θ_d из распределения Дирихле с параметром α (обычно α принимается равным $\frac{50}{T}$). На втором шаге выбирается тема z_{di} из мультиномиального распределения с параметром θ_d. Наконец, согласно выбранной теме z_{di} выбирается слово w_{di} из распределения Φz_{di}, которое является распределением Дирихле с параметром β (обычно параметр $\beta = 0.1$) (Griffiths, Steyvers 2004).

Таким образом, порождающая модель слова w_i из документа d выражается следующим образом:

1. для $k = 1 \dots K$
    a. $\phi^{(k)} \sim Dirichlet\ (\beta)$
2. для каждого документа $d \in D$:
    a. $\theta_d \sim Dirichlet\ (\alpha)$
    b. для каждого слова $w_i \in d$:
        i. $z_i \sim Dirichle\ (\theta_d)$
        ii. $w_i \sim Dirichle\ \left(\phi^{(z_i)}\right)$,

где K – общее число скрытых тем, $\phi^{(k)}$ – распределение k-ой темы, θ_d – распределение всех тем документа, z_i – индекс темы для слова (Darling 2011). При этом каждое $\phi^{(k)}$ представляет собой матрицу $K \times V$, где $\phi_{i,j} = p(w_i|z_j)$. Данный алгоритм можно записать в виде формулы:

$$p(w|d, \theta, \Phi) = \sum_{z=1}^{T} p(w|z, \Phi_z) p(z|d, \theta_d).$$

В данной модели желаемое количество тем, равно как и количество слов, представляющих каждую тему, может быть задано пользователем. В рамках нашего исследования коэффициент подбирался исходя из результатов работы алгоритма, хотя стоит отметить, что существуют алгоритмы, предназначенные для автоматического определения наилучшего количества тем (Greene, O'Callaghan, Cunningham 2014).

Модели PLSA и LDA являются наиболее распространенными и эффективными алгоритмами для решения задач тематического моделирования. На их базе было создано множество других моделей, учитывающих другую информацию, позволяющую улучшить результаты формирования тем текста.

К примеру, большое количество документов, в особенности, научных текстов, связано цитированием. Совершенно естественно предположить, что если в одном документе приводится цитата из другого, то в них присутствует общая тема. В связи с этим была предложена ***совместная вероятностная модель (joint probabilistic model)***: модель-обобщение PLSA, учитывающая не только тематики, но и естественную связь между документами (Cohn, Hofmann 2001).

Достаточно распространенными сегодня являются модели, учитывающие информацию об авторе текста – например, ***автор-***

тематическая модель (author-topic model) (Steyvers, Smyth, Rosen-Zvi, Griffiths 2004). Позже появилось большое количество разновидностей данной модели, учитывающих, помимо самого сообщения и его автора, различные другие параметры. Например, ***модель автор-получатель (author-recipient-topic model)*** использует данные о получателе сообщения и о его связи с автором, что особенно актуально в связи с развитием сети Интернет и ростом популярности письменного общения в режиме онлайн (McCallum, Corrada-Emmanuel, Wang 2004).

В разное время было создано несколько моделей на основе LDA, учитывающие контекст слова. Особенно актуально применение подобных моделей в задачах снятия неоднозначности. В качестве примера подобных моделей можно назвать ***модель латентного размещения Дирихле с использованием словаря Word-Net (latent Dirichlet allocation with Word-Net, LDAWN)***, в рамках которой вместо порождения слов напрямую из темы, каждой теме был противопоставлен «путь» через иерархический словарь Word-Net, с помощью которого и порождалось необходимое слово (Boyd-Graber, Blei, Zhu 2007).

Во многих тематических моделях учитываются изменения, происходящие с тематикой текстов в корпусе с течением времени. Одной из таких моделей является ***динамическая тематическая модель (dynamic topic model, DTM)***, основанная на LDA, однако, в отличие от нее, рассматривающая эволюцию тем внутри корпуса текстов (выраженного в виде последовательности), и выделяющая интервалы увеличения и уменьшения популярности терминов (Blei, Lafferty 2006).

Отдельно стоит упомянуть модели, которые вовсе не используют модель *мешка слов*, а рассматривают также *контекст* слов наряду со скрытым тематическим слоем. Для этого активно применяются скрытые

марковские модели (Hidden Markov Models, HMM)[2], которые позволяют рассматривать текст как некоторую совокупность биграмм, *n*-грамм и даже предложений или абзацев. В качестве примера таких моделей можно назвать ***скрытую тематическую марковскую модель с применением латентного размещения Дирихле (Hidden Markov Model with Latent Dirichlet Allocation, HMM-LDA)***, в результате работы которой может быть построено совместное описание синтаксиса и семантики текста с помощью разбиения каждого предложения на функциональные слова и на термины (Griffiths, Steyvers, Tenenbaum 2007), и ***биграммную тематическую модель (bigram topic model)***. Подробнее о моделях такого рода речь пойдет в Главе 2.

Также на основе LDA, являющейся самой по себе методом *обучения без учителя*[3] *(unsupervised learning)*, было создано несколько моделей, в которых применен метод *обучения с учителем*[4] *(supervised learning)*,

---

[2] **Скрытая марковская модель** — статистическая вероятностная модель, состоящая из некоторого количества наблюдаемых и латентных переменных, при чем значение наблюдаемого вектора x_n, взятого в некоторый момент времени, зависит только от скрытого состояния в данный момент времени, которое, в свою очередь, зависит от скрытого состояния в предыдущий момент времени и так далее. Основной задачей при этом является выявление скрытых параметров на основе наблюдаемых (Gernot A. Fink 2007).

Марковский процесс — случайный процесс, в рамках которого «будущие» вероятности событий определяются только вероятностями наиболее ближайших к данному моменту событий. Иными словами, при фиксированном известном «настоящем» события, не наблюдается зависимости между «будущим» процесса и его «прошлым» (Bharucha-Reid 1960).

[3] **Обучение без учителя** (англ. Unsupervised learning, самообучение) — один из способов машинного обучения, при котором система самостоятельно обучается решать некоторую задачу. Основным отличием от обучения с учителем является то, что машине заранее не известны правильные ответы, а также не происходит никакого вмешательства со стороны ученого, проводящего эксперимент (Hastie et al. 2009)

[4] **Обучение с учителем** (англ. Supervised learning) — один из способов машинного обучения, при котором система принудительно обучается с непосредственным участием экспериментатора на примерах «стимул-реакция». На вход системе подается множество

требующий классификации объектов на основе обучающей выборки: например, ***обучаемая с учителем модель с применением латентного размещения Дирихле*** (***supervised latent Dirichlet allocation, sLDA)*** (Blei, McAuliffe 2007).

Помимо этого, существует целый ряд многоязычных тематических моделей, использующих в своей работе двуязычные и многоязычные параллельные корпуса. Одной из таких моделей является ***двуязычная тематическая модель с применением латентного распределения Дирихле (bilingual model with latent Dirichlet allocation, BiLDA)***. Одной из основополагающих идей данного метода можно назвать дистрибутивную гипотезу, гласящую, что лингвистические единицы, встречающиеся в схожих контекстах, имеют близкие значения (Harris 1954). Так, два слова считаются потенциальными переводными эквивалентами, если они часто встречаются в одной и той же межлингвистической теме и не встречаются в других; иными словами, их условное тематическое распределение совпадает (Vulic, De Smet, Moens 2011). Эта и подобные ей модели широко используются для поиска перевода единиц без дополнительного использования каких бы то ни было лингвистических ресурсов. Исследования (Vulic, De Smet, Moens 2011; Gaussier, Renders, Matveeva, Goutte, Déjean 2004), показали, что модели такого рода превосходят прочие методы, основанные на измерении схожести документов из двуязычных корпусов.

ситуаций и множество вероятных ответов (обучающая выборка). Задачей системы является построение алгоритма, определяющего связь между ответами и ситуациями и позволяющего получить точный ответ для любого стимула (Manning, Schuetze 1999).

Выводы к главе 1

Тематическое моделирование несомненно является одним из самых мощных инструментов структурирования информации, содержащейся в текстах на естественном языке, что позволяет облегчить их исследование и дальнейшую обработку. Тематические модели успешно применяются для решения задач снятия неоднозначности, информационного поиска, машинного перевода и многих других.

На сегодняшний день существует большое количество различных тематических моделей, учитывающих в своей работе различные характеристики текста, такие, как авторство текста и изменение интересующей автора тематики во времени. Большинство из них основано на модели PLSA и на её более усовершенствованной версии LDA.

Большинство тематических моделей используют в своей работе модель *мешка слов*, которая не предоставляет возможности учитывать связь слов в контексте, что, несомненно, является их недостатком. По этой же причине в большинстве моделей выделяемые темы представлены набором униграмм, что зачастую мешает интерпретации выделяемых тем и их точности. Для решения этой проблемы было предложено несколько алгоритмов, позволяющих добавить в темы различные *n*-граммы. Подробнее об этих алгоритмах речь пойдет в главе 2.

2. ТЕМАТИЧЕСКИЕ МОДЕЛИ, УЧИТЫВАЮЩИЕ *N*-ГРАММЫ

В данной главе подробно разбирается *n*-граммная модель языка как средство обработки текстов, написанных на естественном языке, и обсуждается целесообразность включения *n*-грамм в тематические модели. Также в ней приведены некоторые предложенные ранее алгоритмы, предполагающие расширение тематических моделей с помощью *n*-грамм, и рассмотрены их достоинства и недостатки.

2.1. Использование *n*-грамм в задачах автоматической обработки естественного языка

Как уже было упомянуто ранее, одним из наиболее распространенных способов представления документа в задачах компьютерной лингвистике является модель *мешка слов*, в рамках которой документ представляется в виде набора слов. Альтернативным способом является представление документа с помощью различных ***языковых моделей (language models, LMs)***, которые позволяют приписывать вероятности различным фрагментам текста (высказываниям, предложениям и так далее). Простейшей языковой моделью является ***n-граммная модель языка***.

Дадим формальное определение. Пусть задан некоторый конечный алфавит $V_T = \{w^i\}$, где w^i – отдельный символ. Тогда множество цепочек конечной длины, состоящих из символов алфавита V_T, называется ***языком на алфавите*** $\boldsymbol{V_T}$ и обозначается $L(V_T)$. Отдельную цепочку из языка $L(V_T)$ будем называть ***высказыванием*** на этом языке. Цепочка же длинной N называется ***n-граммой на алфавите*** V_T.

Например, если предположить, что алфавитом V_T являются слова русского языка, то в этом случае высказываниями будем считать фразы на

русском языке, а ***n-граммами*** – последовательности из *N* слов одной фразы (Бузикашвили, Крылова, Самойлов 2000). Здесь и далее мы будем использовать термин ***n-грамм*** именно в этом смысле.

Основным предположением, на котором базируются *n*-граммные языковые модели, является следующее: вероятность появления *n*-ого слова в предложении зависит от предыдущих *n−1* слов. Иными словами, вероятность появления слова w вычисляется по следующей формуле: $P(w|h)$, где *h* – слова, стоящие в предложении перед словом *w*.

Так, для вычисления вероятности появления слова w_i в предложении, в котором до рассматриваемого слова находится последовательность из *N* слов, записанная в виде вектора $w_1 \dots w_n$ или w_1^n, используется следующая формула:

$$P(w_1 \dots w_n) = P(w_1)P(w_2|w_1)P(w_3|w_1^2) \dots P(w_n|w_1^{n-1}) = \prod_{k=1}^{n} P\left(w_k|w_1^{k-1}\right)$$

Из данного равенства очевидно, что мы можем оценить совместную вероятность всей последовательности слов, перемножив ряд условных вероятностей. Однако стоит отметить, что у нас нет возможности сосчитать точную вероятность появления слова при условии длинной последовательности стоящих перед ним слов – а именно $P(w_n|w_1^{n-1})$. Именно поэтому особенностью *n*-грамм модели является отказ от вычисления условной вероятности появления всех слов предложения и аппроксимация лишь до нескольких слов, находящихся непосредственно перед интересующим нас словом.

Это положение используют многие вероятностные тематические модели, основанные на *n*-граммах. Так, например, ***биграммная модель языка (the bigram model)*** при вычислении условной вероятности появления

слова учитывает лишь предыдущее слово в контексте, допуская марковское предположение: $P(w_n|w_1^{n-1}) \sim P(w_n|w_{n-1})$. В таком случае условная вероятность вычисляется по формуле: $P(w_n|w_{n-1})$. Аналогично ***триграммная языковая модель*** учитывает два предыдущих слова. Теоретически возможно построение *n*-грамм и более высоких порядков, однако зачастую они дают худшие результаты, чем приведенные выше. Исследования показывают, что, как правило, наибольшей эффективности позволяет добиться одновременное использование униграмм и биграмм, реже – триграмм (Shannon 1948).

Таким образом, общее уравнение аппроксимации *n*-граммы для предсказания вероятности появления следующего слова в предложении можно записать следующим образом:

$$P(w_n|w_1^{n-1}) \sim P(w_n|w_{n-N+1}^{n-1}).$$

N-граммная модель языка, как и многие другие статистические модели, нуждается в обучении на тренировочном корпусе. Стоит отметить, что модели такого рода очень чувствительны к тренировочной выборки: в зависимости от специфики текстов, на которых будет проходить обучение модели, в дальнейшем она будет выделять в текстах *n*-граммы разного рода.

Помимо лингвистики, *n*-граммы применяются во многих науках, например, в теоретической математике, биологии. В области обработки естественного языка *n*-граммы особенно успешно используются для категоризации текста, а также для более эффективного получения знаний из текстовых данных. На практике модели такого рода повсеместно используются для решения задач машинного перевода, проверки правописания, моделирования текстов на естественном языке, распознания речи (Baker 1990; Jelinek 1990), выявления плагиата и многих других (Jurafsky Daniel & Martin H. James 2014).

Далее мы рассмотрим предложенные на данный момент способы добавления *n*-грамм в тематические модели.

2.2. Обзор предложенных ранее методов автоматического включения *n*-грамм в тематические модели

Все алгоритмы автоматического добавления *n*-грамм в тематические модели можно разделить на две большие группы с точки зрения метода извлечения словосочетаний из текстов:

1. представляющие собой унифицированную вероятностную тематическую модель и, таким образом, выделяющие *n*-граммы на этапе выделения тем;
2. предварительно выделяющие *n*-граммы на этапе предобработки текста.

Более распространенным является первый подход.

2.2.1. Унифицированные вероятностные тематические модели

Биграммная тематическая модель (bigram topic model)

Одной из первых моделей, объединяющей латентные темы и статистические методы выделения *n*-грамм, была ***биграммная тематическая модель,*** являющаяся иерархической порождающей вероятностной моделью (Wallach 2006). В рамках данной модели в документах выделяются темы, состоящие исключительно из биграмм. В качестве основополагающего используется предположение о зависимости появления слова w_i зависит исключительно от слова w_{i-1}, стоящего непосредственно перед интересующим нас словом:

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

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

Графическое изображение биграммной тематической модели представлено на Рис. 3:

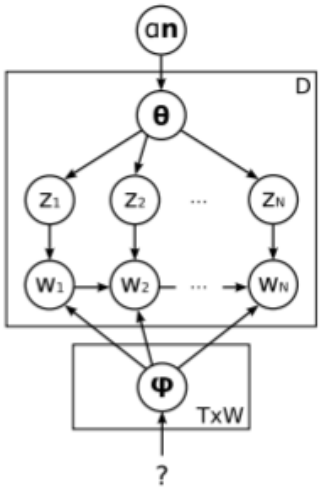

Рис. 3. Биграммная тематическая модель

В рамках данной модели на первом этапе для каждой темы *z* и для каждого слова *w* вычисляется распределение φ_{zw} из априорного распределения Дирихле δ. Затем для каждого документа *d* стоится распределение θ_d из априорного распределения Дирихле α_n. После этого для каждого слова w_N в документе *d* семплируются темы *z* из распределения θ_d и слово w_i из распределения $\varphi_{zw_{N-1}}$.

Скрытая тематическая марковская модель с применением латентного размещения Дирихле (Hidden Markov Model with Latent Dirichlet Allocation, HMM-LDA)

В данной модели строится совместное описание синтаксиса и семантики текста с помощью разбиения каждого предложения на *функциональные слова*, которые порождаются с помощью скрытой марковской модели, (таким образом, описываются локальные закономерности), и на *термины*, генерируемые тематической моделью LDA (так дается глобальное тематическое описание документа).

Графическое изображение модели представлено на Рис. 4:

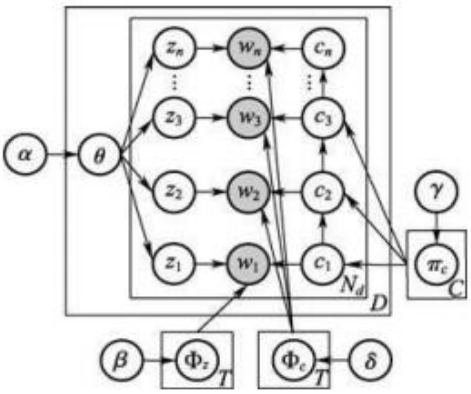

Рис. 4. Модель HMM-LDA

Модель состоит из последовательности переменных-слов $w = (w_1, \dots, w_n)$, тематических переменных $z = (z_1, \dots, z_n)$ и последовательности бинарных классификаций $c = (c_1, \dots, c_n)$, указывающих, образуют ли данное слово и предыдущее словосочетание. Значение c_n выбирается, основываясь на предыдущем слове w_{i-1}, исходя из распределения $P(x_i|w_{i-1})$. Если $c_i = 1$, то слова w_{i-1} и w_i образуют словосочетание и слово w_1 анализируется в семантическом аспекте, т. е. на основании тематического распределения Φ_z: $P(w_i|w_{i-1}, \mathrm{c}_{i=1})$. Если же $c_i \neq 1$, то слово порождается из распределения Φ_C: $P(w_i|t, c_i = 0)$. Каждому

документу *d* соответствует распределение в пространстве тем θ_d и матрица вероятностей переходов для описания переходов между классами c_{i-1} и c_i в марковской цепи строится на основании распределения $\pi_{s_{i-1}}$.

Распределения строятся следующим образом:

1. распределение для тем Φ_z строится для каждой темы *z* из априорного распределения Дирихле *β;*
2. распределение классов Φ_C строится из априорного распределения Дирихле *δ;*
3. распределение для слов π_s строится из априорного распределения Дирихле *β.*

Процесс порождения текстовой коллекции может быть записан следующий образом:

1. Семплирование распределения θ_d из априорного распределения Дирихле α для документа *d;*
2. Для каждого слова w_i в документе *d:*
   a. выявление темы z_i из распределения θ_d;
   b. вычисление c_i из $\pi^{(s_{i-1})}$;
   c. в случае, если $c_i = 1$, определение w_i исходя из Φ_{z_i}. Иначе: определение w_i исходя из Φ_{c_i}.

Апостериорное распределение тем z_i может быть выписано в следующем виде:

$$p(z_i|z_{i-1}, c, w) \propto p(z_i|z_{i-1})p(w_i|z, c, w_{i-1})$$

$$\propto \begin{cases} n_{zi}^{(di)} + \alpha, & c_i \neq 1 \\ (n_{zi}^{(di)} + \alpha)\,\dfrac{n_{wi}^{(zi)} + \beta}{n_{wi}^{(zi)} + V\beta}, & c_i = 1 \end{cases}$$

где $n_{zi}^{(di)}$ — число слов в документе d_i, относящихся к теме z_i, а $n_{wi}^{(zi)}$ — общее число слов, относящихся к теме z. Подсчёт слов производится лишь для слов i, для которых $c_i = 1$ (Griffiths, Steyvers, Blei, Tenenbaum 2005) .

N-граммная тематическая модель (topical n-gram model, TNG)

Одной из главных особенностей данной модели является предоставление возможности принятия решения, являются ли стоящие рядом слова биграммой, на основании их ближайшего контекста (Wang, McCallum, Wei 2007). Для этого в модель добавляется дополнительный уровень сложности, позволяющий автоматическое формирование биграмм с опорой на контекст.

Графически модель можно представить следующим образом (см Рис. 5):

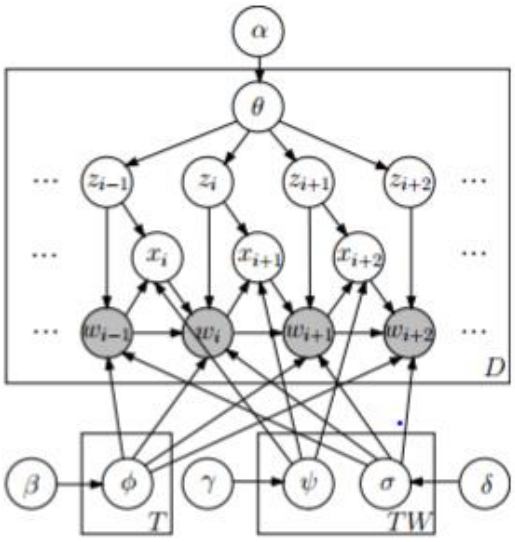

Рис. 5. *N*-граммная тематическая модель

Процесс порождения текстовой коллекции можно разложить на следующие этапы:

1. построение распределения φ_z для каждой темы из априорного распределения Дирихле β;

2. построение распределения ψ_{zw} для каждой темы и для каждого слова из априорного β-распределения γ;
3. построение распределения σ_{zw} из априорного распределения Дирихе δ;
4. построение распределения θ_d из априорного распределения Дирихле α;
5. Для каждого слова w_i в документе d:
   a. семплирование x_i из распределения $\psi_{z_{i-1},w_{i-1}}$;
   b. семплирование темы z_i из распределения θ_d;
   c. если $x_i = 1$, то семплирование слова w_i исходя из $\sigma_{z_i w_{i-1}}$;
      Иначе: семплирование слова w_i исходя из ψ_{t_i};
6. для каждого слова w_i в документе *d:*
   a. выявление темы z_i из распределения θ_d;
   b. вычисление c_i из $\pi^{(s_{i-1})}$;
   c. в случае, если $c_i = 1$, определение w_i исходя из Φ_{z_i}. Иначе: определение w_i исходя из Φ_{c_i}.

Тематическая модель «Слово-Символ» (Topical word-character model, TWC)

Все описанные выше модели разрабатывались в первую очередь для европейских языков и с учетом их специфики. Однако существуют также модели, созданные на основе корпусов азиатских текстов. Одной из таких моделей является модель «Слово-Символ» (Hu et al. 2008).

Существенным отличием данного алгоритма от всех остальных тематических моделей, учитывающих биграммы, является *не*использование предположения о выводимости темы n-граммы из тем образующих её униграмм. В рамках модели «Слово-Символ» рассматриваются два разных

списка тем: список тем слов корпуса и список тем символов. Подобное разграничение слов и символов обуславливается в первую очередь особенностями китайского языка, использующего иероглифическую письменность. Так, при порождении некоторого символа китайского языка происходит выбор на двух этапах: на первом выбирается тема слова, на втором – тема символа.

Графическое изображение модели, предложенное авторами, представлено ниже (см. Рис. 6).

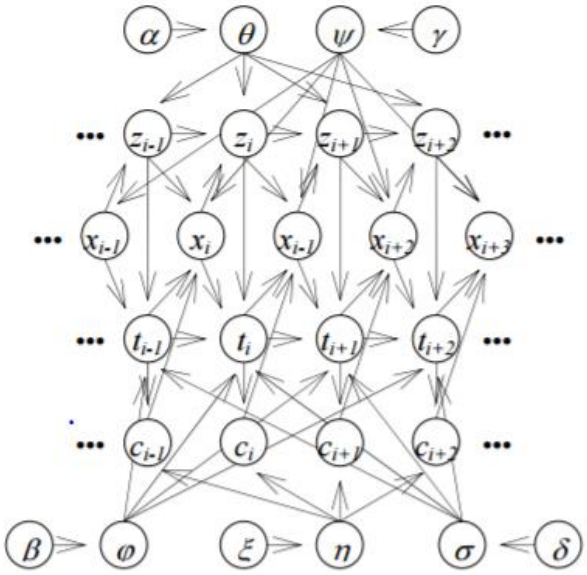

Рис. 6. Графическое изображение модели "Слово-Символ"

Процесс порождения текстовой коллекции сводится к следующим этапам:

1. построение распределения θ_d для каждого документа из априорного распределения Дирихле α;
2. построение распределения φ_t для каждой темы слов из априорного распределения Дирихле β;
3. построение распределения σ_{tw} для каждой темы слов t и каждой темы символов z из априорного распределения Дирихе δ;

4. построение распределения ψ_{tzc} для каждой темы слов t, темы символов z и каждого символа c из априорного β-распределения γ;
5. построение распределения η_z для каждой темы символов z из априорного распределения Дирихле ξ;
6. для каждого символа c_i в документе d:
    a. семплирование x_i из распределения $\psi_{t_{i-1}z_{i-1}c_{i-1}}$;
    b. если $x_i = 0$:
        i. семплирование темы слова t_i из распределения θ_d;
        ii. семплирование темы символа z_i из распределения φ_{t_i};
    c. если $x_i \neq 0$:
        i. $t_i = z_{t-1}$;
        ii. семплирование z_i из распределения $\sigma_{t_i z_{i-1}}$;
    d. семплирование символа c_i из распределения η_{z_i}.

2.2.2. Предварительное извлечение словосочетаний

Все рассмотренные выше модели являлись унифицированными: иными словами, извлечение n-грамм происходило одновременно с собственно процессом тематического моделирования. Однако это не единственный подход. Значительно менее распространенным является метод, позволяющий извлечь n-граммы на этапе предобработки корпуса, и лишь после этого построить вероятностные тематические модели – уже, соответственно, на корпусе с включенными в него n-граммами.

Далее в разделе будут рассмотрены некоторые алгоритмы, использующие в своей работе этот подход.

Алгоритм, предполагающий упорядочивание биграмм в соответствии с мерой T-score

Одним из таких методов является алгоритм, предложенный в работе (Lau, Baldwin, Newman 2013). В данном алгоритме коллокации выявляются на этапе предварительной обработки текста, упорядочиваются в соответствии с ассоциативной мерой ***T-Score***:

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

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

Далее наиболее удачные полученные словосочетания объединяются в один токен и добавляются в корпус, заменяя униграммы. Таким образом, в процессе собственно построения тематической модели (в данном случае, LDA) и построения модели *мешка слов*, они рассматриваются наряду с другими униграммами как токены.

PLSA-SIM

PLSA-SIM является усовершенствованной версией одной из базовых моделей PLSA (см. стр. 13) и также использует в работе модель *мешка слов*. Идея, которая легла в основу данного алгоритма, следующая: в любых текстах существует большое количество слов и коллокаций, семантически и лексически близких: например, *бюджетный, бюджетные расходы, бюджетные средства, бюджетные доходы* (Нокель, Лукашевич 2015). В рамках данного алгоритма при выделении тем подобные словосочетания относятся к одной теме. Если же подобные слова и словосочетания никогда

не встречаются в рамках одного документа, то для них выполняется стандартный алгоритм PLSA.

PLSA-ITER

Дальнейшим усовершенствованием алгоритма PLSA-SIM является итеративный алгоритм PLSA-ITER, в рамках которого рассматриваются самые частотные униграммы, представляющие темы, и из них составляются биграммы. В качестве примера авторами приводится биграмма *ценный бумага*, которая может быть составлена, если в некой теме среди первых n слов окажутся униграммы *ценный* и *бумага*. В (Нокель, Лукашевич 2015) рассматриваются первые 10 униграмм, из которых далее образуются биграммы и добавляются в тематические модели. Помимо этого, в данной модели, как и в предыдущей, учитывалась частеречная принадлежность слов: в выделении тем участвовали только существительные, прилагательные, глаголы и наречия, а в формировании биграмм для русского языка следующие коллокации: *существительное + существительное в родительном падеже, прилагательное + существительное.*

2.3. Сравнение двух подходов к выделению *n*-грамм

Несомненным достоинством унифицированных тематических моделей является их логическое теоретическое обоснование. Однако к их минусам можно отнести большое количество параметров, нуждающихся в настройке (Lau, Baldwin, Newman 2013). Например, число параметров у Биграммной Тематической Модели равно W^2T, в то время как у базовой модели LDA – WT, где W – размер словаря (т.е. число уникальных слов и словосочетаний корпуса), T – число выделенных тем.

Подход, предполагающий предварительное выделение биграмм в текстовых коллекциях, возможно, не имеет такого изящного теоретического обоснования, однако позволяет строить алгоритмы, являющиеся гораздо более простыми в применении. В первую очередь это достигается за счет того, что количество настраиваемых параметров в данных моделях равен их количеству в исходных моделях (как правило, LDA или PLSA). Недостатком данного подхода можно назвать повышение перплексии, что ведет к ухудшению обобщающей способности выявленной модели.

Выводы к главе 2

В данной главе был сделан обзор существующих методов для автоматического добавления *n*-грамм в тематические модели.

Все предложенные ранее методы можно разделить на две неравные группы: методы, предполагающие выделение биграмм в процессе построения вероятностной тематической модели, и методы, выполняющие поиск биграмм на уровне предварительной обработки текста. Алгоритмов, использующих первый метод, существует гораздо больше, нежели чем базирующихся на втором методе, несмотря на их очевидную сильную сторону, заключающуюся в простоте применения. Однако на данный момент не существует алгоритма выделения тематических моделей с биграммами, являющегося определяющим и явно превосходящего другие по эффективности и простоте использования.

В следующей главе представлен адаптированный для русского языка алгоритм автоматического выделения *n*-грамм в текстовых коллекциях и их последующее добавление в тематические модели и описаны проведенные эксперименты на корпусах русскоязычных текстов.

3. ТЕОРЕТИЧЕСКОЕ ОПИСАНИЕ ЭКСПЕРИМЕНТА ПО АВТОМАТИЧЕСКОМУ ДОБАВЛЕНИЮ БИГРАММ В ТЕМАТИЧЕСКИЕ МОДЕЛИ

Предложенный нами алгоритм относится ко второй группе подходов, изложенных нами на стр. 24, а именно предполагающих предварительное выделение *n*-грамм на этапе предобработки текста. Решение придерживаться именно такому подходу было принято в связи с желанием избежать излишнего усложнения модели. Также нашей целью было не создание новой тематической модели, а разработка алгоритма, который впоследствии можно бы было имплементировать в любую уже существующую модель.

Основная идея алгоритма заключается в предварительном выделении биграмм в корпусе текстов с помощью модуля Phrases, затем – построении модели и, наконец, еще одной процедуры выделения двусловных сочетаний в уже сформированных темах. На вход алгоритму подается корпус текстов, результатом работы системы является список выделенных в корпусе текстов тем, представленных не только униграммами, но и биграммами. В данной главе подробно рассматриваются все этапы работы алгоритма.

3.1. Предварительная обработка корпуса текстов

На этапе предварительной обработки текстов из них удаляются нетекстовые символы и сокращения (в рамках данного эксперимента было принято решение об удалении всех слов, длина которых составляет менее 3 символов). Помимо этого, из текстов исключаются слова, входящие в список стоп-слов на основе словарей служебных слов и оборотов НКРЯ (см. Приложение 1), а также 98 наиболее частотных глаголов и отвлеченных существительных (например, *использовать*, *позволять, наличие, отсутствие* и так далее – полный список приведен в Приложении 2).

Далее проводится лемматизация текстов и автоматическое разрешение морфологической неоднозначности с помощью морфологического анализатора русского языка MyStem 3.0 [5] (Segalovich 2003). Данный анализатор использует в своей работе словарь часто употребляемых русских слов. Для слов, не входящие в данный словарь, программа формирует морфологические гипотезы (Segalovich, Maslov 1998). В результате работы анализатора для каждой словоформы из корпуса текстов предлагается список возможных лемм, из которого случайным образом выбирается одна (как правило, первая из предложенных).

3.2. Выделение биграмм с помощью использования модуля Phrases

На первом этапе работы алгоритма в исследуемом корпусе выявляются биграммы. Для этого привлекается модуль Phrases, входящий в состав библиотеки Gensim[6]. Данный модуль, используя модель *мешка слов*, автоматически выявляет в документах наиболее часто встречающиеся многословные словосочетания. В наших экспериментах в качестве документов было решено рассматривать предложения из корпуса, поскольку наибольший интерес вызывает совместная встречаемость слов именно в одном предложении.

Для корректной работы алгоритма и повышения точности получаемых результатов важно правильно подобрать следующие параметры:

- min_count: минимальное количество встреваемости слова в корпусе текстов для того, чтобы оно рассматривалось в ходе работы алгоритма;

[5] https://tech.yandex.ru/mystem/ (дата последнего обращения 27.04.2017)
[6] https://radimrehurek.com/gensim/ (дата последнего обращения 27.04.2017)

- threshold: параметр, основанный на совместной встречаемости слов и отвечающий за принятие решения о формировании биграммы. Слова «a» и «b» считаются биграммой, если:

$$\frac{(cnt(a,b) - \text{min\_count}) * N}{cnt(a) * cnt(b)} > threshold,$$

где N – общий размер словаря;

- max_vocab_size: максимальный размер словаря;

После выделения биграммы образующие её униграммы объединяются в корпусе текстов символом, заданным пользователем (параметр *delimiter*), и, таким образом, а дальнейшем рассматриваются как одна лемма.

Алгоритм выделения биграмм был реализован на языке программирования Python[7], общий размер кода составил 480 строк.

3.3. Построение тематической модели корпуса текстов с выделенными в них биграммами.

На втором этапе работы алгоритма строится тематическая модель экспериментального корпуса. Нами было принято решение использовать для этого вероятностную тематическую модель латентного размещения Дирихле (Blei, Ng, Jordan 2003). Данная модель достаточно легко реализуется на языке программирования Python[8]; она также включена в различные библиотеки, например, в Gensim[9]. Однако нами был выбран пакет для анализа данных Scikit-Learn[10], поскольку он обеспечивает

[7] https://www.python.org/
[8] https://www.python.org/
[9] https://radimrehurek.com/gensim/
[10] http://scikit-learn.org/stable/

лучшую точность результатов благодаря более оптимальному выбору параметров. Данная библиотека, разработанная для языка программирования, включает в себя большое количество различных расширений, позволяющих реализовать различные алгоритмы машинного обучения (Pedregosa et al 2011).

На вход алгоритму подается корпус. Для точности анализа корпус текстов разбивается на документы приблизительно по 3500 символов каждый (оптимальное число символов было установлено экспериментальным путем). Вначале текстовая коллекция преобразуется в матрицу с помощью инструмента *CountVectorizer*. Основными параметрами для данного инструмента являются:

- max_df: параметр, позволяющая отсеять стоп-слова, специфичные для данного корпуса. Все слова, встретившиеся суммарно в текстах большее число раз, чем заданный параметр, не учитываются при работе алгоритма;
- min_df: параметр, позволяющий отсеять редко встречающиеся в корпусе текстов слова, которые неважны при построении тем. Все слова, встретившиеся суммарно в текстах меньшее число раз, чем заданный параметр, не учитываются при работе алгоритма.

Далее непосредственно строится тематическая модель. Наиболее важными являются следующие параметры: количество итераций алгоритма; количество выделяемых тем; количество слов, представляющих каждую тему.

Стоит сказать несколько слов о формате выдачи тем. Базовая модель LDA не предполагает автоматическое именование тем (*topic labeling*) для облегчения интерпретации. На сегодняшний день было предложено

достаточно больше количество алгоритмов, позволяющих автоматически приписывать выделенным темам *метки*, то есть, слова и словосочетания, наиболее точно и ёмко выражающие общее содержание темы (Mei, Shen, Zhai 2007; Hindle et al., 2013; Cano Basave, He, Xu, 2014; Nolasco, Oliveira, 2016; Magatti et al., 2009; Aletras, 2014). Один из них был проведен на корпусе, которые мы используем для тестирования предложенного нами алгоритма, а именно на корпусе текстов на лингвистическую тематику (Mirzagitova, Mitrofanova 2016). Однако в рамках нашего исследования мы сознательно не назначаем метки темы, дабы не отклоняться от основных задач данного исследования.

Отрывок кода приведен ниже (Рис. 7).

```
In [ ]: # Use tf (raw term count) features for LDA.
        print("Extracting tf features for LDA...")
        tf_vectorizer = CountVectorizer(max_df=0.8, min_df=2, max_features=n_features)
        tf = tf_vectorizer.fit_transform(corpusWithBigramsForTM)
        print("It's done")

        print("Fitting LDA models with tf features, " "n_samples=%d and n_features=%d..." % (n_samples, n_features))
        lda = LatentDirichletAllocation(n_topics=n_topics, max_iter=50, learning_method='online', learning_offset=50., random_state=0)
        t0 = time()
        lda.fit(tf)
        print("done in %0.3fs." % (time() - t0))

        print("\nTopics in LDA model:")
        tf_feature_names = tf_vectorizer.get_feature_names()
        print_top_words(lda, tf_feature_names, n_top_words)

        listOfTopTopics = join_Topics(lda, tf_feature_names, n_top_words)
```

Рис. 7. Отрывок из кода алгоритма LDA

На третьем этапе полученные темы заново обрабатываются с помощью модуля Phrases, что позволяет выделить еще некоторое количество биграмм. Переобучение модели при этом не происходит.

Данный алгоритм был протестирован на двух корпусах текстов:

1) корпус специальных текстов по радиоэлектронике, ракетостроению и технике;
2) корпус русскоязычных специальных текстов на лингвистическую тематику.

Результаты эксперимента приведены в следующей главе.

Выводы к главе 3

Итак, в главе был представлен и подробно описан предлагаемый нами алгоритм для автоматического выделения биграмм из корпуса текстов и последующего внедрения их в тематические модели.

Алгоритм устроен следующим образом: на вход подается корпус текстов, проводится стемминг и снятие морфологической неоднозначности, из текстов выделяются биграммы, соответствующие униграммы заменяются на них в текстах со знаком «_», проводится тематическое моделирование и, наконец, в полученных моделях проводится вторичное выделение биграмм. Все шаги рассматриваются в данной главе более подробно.

Технически алгоритм был реализован на языке программирования Python и проверен на двух корпусах русскоязычных текстов. Результаты приведены и проанализированы в следующей главе.

4. ПРАКТИЧЕСКАЯ РЕАЛИЗАЦИЯ АЛГОРИТМА АВТОМАТИЧЕСКОГО ДОБАВЛЕНИЯ БИГРАММ НА МАТЕРИАЛЕ КОРПУСОВ РУССКОЯЗЫЧНЫХ ТЕКСТОВ

Для проведения эксперимента были выбраны два корпуса русскоязычных текстов:

1. корпус специальных текстов по радиоэлектронике, ракетостроению и технике (Дубовик 2017);
2. корпус русскоязычных специальных текстов на лингвистическую тематику из лингвистического энциклопедического словаря (ЛЭС) под редакцией В.Н. Ярцевой и энциклопедии «Кругосвет» [11] (Mirzagitova, Mitrofanova 2016).

Причин выбора корпуса текстов именно научного стиля несколько. Во-первых, специфическими чертами подобных текстов является максимальная точность в использовании слов и лаконичность, что несомненно упростит интерпретацию тематических моделей, повысит их точность и снизит долю шума. Во-вторых, их характерной чертой является повышенное содержание терминов - «*слов или сочетаний из двух слов, присутствующее в соответствующем «золотом стандарте*» (т.е. в существующем, разработанном экспертами терминологического ресурсе)» (Нокель 2016). Термины всегда существуют в рамках терминосистемы, обладают фиксированным терминологическим значением и, как правило, обозначают какой-то предмет, явление или, в крайнем случае, процесс. С одной стороны, это также улучшит процесс тематического моделирования, поскольку темы документов достаточно очевидны и легко выделяемы. В другой стороны, частая воспроизводимость терминов в рамках одного

[11] http://www.krugosvet.ru

текста облегчает выделение биграмм. Также не стоит упускать из виду то, что, как правило, термины формируют существительные и глаголы, и словосочетания именно с этими частями речи являются классическими и наиболее частотными биграммами.

4.1. Предварительная обработка корпуса текстов

После прохождения этапа предварительной обработки текстов, включающем в себя удаление нетекстовых символов, сокращений, стоп-слов, наиболее частотных слов, а также лемматизацию и автоматическое разрешение морфологической неоднозначности, объемы корпусов оказались следующими: (см. Таблица 3).

Таблица 3. Объемы экспериментальных корпусов до и после предварительной обработки

Объем корпуса до предварительной обработки		*Объем корпуса после предварительной обработки*	
корпус текстов по радиоэлектронике, ракетостроению и технике	корпус текстов на лингвистическую тематику	корпус текстов по радиоэлектронике, ракетостроению и технике	корпус текстов на лингвистическую тематику
526 648 словоформ	1 333 546 словоформ	216 613 леммы	1 246 590 лемм

4.2. Выделение биграмм

Первый этап работы алгоритма заключался в выделении биграмм в корпусах текстов с помощью модуля Phrases (см. на стр. 36). Для этого оптимальными были приняты следующие параметры: min_count = 2, threshold = 5.0, max_vocab_size = 40000000, delimiter = '_'. Процесс обучения происходил непосредственно на самих корпусах текстов. Отрывок из кода приведен ниже (Рис. 8).

```
In [*]: #составление списка предложений с объединенными биграммами
        #(список списков строк [предложение[слова]])
        listWithBigrams = []
        phrases = Phrases(listOfLists, min_count=1, threshold=5.0)
        bigram = Phraser(phrases)
        listWithBigrams = Phraser[listOfLists]
```

Рис. 8 Отрывок кода алгоритма, выделяющего биграммы в корпусе текстов

В процессе работы алгоритма в корпусе текстов было образовано 14 542 биграмм для корпуса текстов по радиоэлектронике, ракетостроению и технике и 187 008 биграмм для корпуса текстов на лингвистическую тематику. В Таблица 4 представлено абсолютное количество выделенных биграмм для обоих корпусов текстов, а также их соотношение с общим объемом корпусов.

Таблица 4. Абсолютное количество выделенных биграмм и их соотношение с объемом корпусов

Корпус текстов по радиоэлектронике, ракетостроению и технике		*Корпус текстов на лингвистическую тематику*	
Количество выделенных биграмм	Соотношение с общим объемом корпуса	Количество выделенных биграмм	Соотношение с общим объемом корпуса
14 542 биграмм	13,4%	187 008 биграмм	30,0%

Сразу же бросается в глаза разница в процентном соотношении выделенных биграмм относительно двух корпусов: в корпусе текстов на лингвистическую тематику было выделено значительно больше биграмм, чем в корпусе текстов по радиоэлектронике, ракетостроению и технике. Причин тому несколько. Прежде всего, роль играет объем корпуса, поскольку от него напрямую зависит основной параметр, отвечающий за принятие решения о формировании биграммы (*threshold*): при увеличении объема число биграмм также увеличивается. Однако основной причиной можно назвать лексическую специфику собранных корпусов. Корпус

текстов на лингвистическую тематику более однородный; лексическое наполнение входящих в него текстов более однообразно, наблюдается бóльшее количество высокочастотных терминов. Напротив, корпус по радиоэлектронике, ракетостроению и технике представляет собой тематически более разнородную выборку текстов и изобилует низкочастотными терминами. Общая частота употребления терминов в корпусе снижается за счет лексического разнообразия, и поэтому меньшее количество униграмм проходит порог формирования биграммы. Помимо этого, в соответствии с установленными нами параметрами униграммы, встречающиеся в текстах менее двух раз, вовсе не рассматриваются при выделении биграмм; поэтому, вероятно, некоторое количество потенциальных биграмм, которые были образованы из низкочастотных терминов, не попали в конечный список.

Затем соответствующие униграммы корпуса были объединены символом «_» и, таким образом, заменены одной леммой. Примеры состава корпуса до предварительной обработки, после первично обработки и после работы алгоритма по выделению биграмм приведены ниже (см. Таблица 5 и Таблица 6).

Таблица 5. Примеры корпуса текстов по радиоэлектронике, ракетостроению и технике на разных этапах обработки

Отрывок корпуса до предварительной обработки
Попутно подтверждается и изложенное в 4.4 об асимптотической нормальности ОМП, ибо при qys>\ ошибка X — Xo линейно связана с величиной Z'(Xo), которая может считаться гауссовской в силу нормальности Z(X) в рассматриваемых условиях. Как выясняется, разработанная в 4.7 методика расчета потенциальной точности, т. е. дисперсий ОМП, оказывается удовлетворительной только при условии, что превышение сигнала над шумом настолько велико, что наблюдатель вправе полагать разброс X относительно Xo полностью укладывающимся в пределы линейного участка производной ФН, смещенной в точку X = X0. Для этого прежде всего необходимо, чтобы побочные (шумовые) выбросы на 4.7, в не превосходили основного пика, обусловленного ФН сигнала.
Отрывок корпуса до предварительной обработки, перед запуском алгоритма по выделению биграмм
попутно излагать подтверждаться считаться асимптотический линейно нормальность омп условие гауссовский величина связанный сила выясняться расчет точность потенциальный методика удовлетворительный точка полностью предел вправе шум производная дисперсия условие превышение укладываться смещать разброс наблюдатель настолько сигнал линейный участок омп большой побочный шумовой выброс необходимо обусловливать пик сигнал основный превосходить
Отрывок корпуса после работы алгоритма по выделению биграмм
попутно излагать подтверждаться считаться асимптотический линейно нормальность омп условие ***гауссовский_величина*** связанный сила выясняться ***расчет_точность*** потенциальный методика удовлетворительный точка полностью предел вправе шум ***производная_дисперсия условие_превышение*** укладываться смещать разброс наблюдатель ***настолько_сигнал*** линейный участок омп большой ***побочный_выброс*** шумовой необходимо обусловливать пик сигнал основный превосходить

Анализируя примеры из корпуса, можно заметить, что в результате работы алгоритма было образованно несколько правильных биграмм, хоть и не стоящих в данном отрывке в непосредственной близости: например, *гауссовский_величина*, *побочный_выброс*, *расчет_точность*. С другой стороны, остальные обобщенные униграммы в данном отрывке оказываются образованными случайно и биграммами не являются.

Таблица 6. Примеры корпуса текстов на лингвистическую тематику на разных этапах обработки

Отрывок корпуса до предварительной обработки
Кодифицированная норма часто отстаёт от реально сложившейся. Орфоэпия складывается одновременно с формированием национального языка, когда расширяется сфера действия устной речи, развиваются новые формы публичной речи. В разных национальных языках процесс становления орфоэпических норм проходит по-разному. Орфоэпические нормы могут пройти несколько этапов, прежде чем стать нормами национальными языка. Так, основные особенности русской произносительной нормы сформировались в 1- й половине 17 в. как особенности московского говора и лишь во 2- й половине 19 в. окончательно сложились как нормы национального языка.
Отрывок корпуса до предварительной обработки, перед запуском алгоритма по выделению биграмм
Кодифицированный часто отставать реально норма складываться речь язык новый устный формирование национальный публичный одновременно действие развиваться форма складываться сфера орфоэпия расширяться язык разный национальный становление норма орфоэпический проходить разному процесс язык национальный норма орфоэпический проходить этап русский особенность сформировываться основной норма произносительный половина московский особенность половина говор национальный язык норма окончательно складываться
Отрывок корпуса после работы алгоритма по выделению биграмм
кодифицированный ***часто_отставать*** реально ***норма_складываться*** речь язык новый ***устный_формирование национальный_публичный*** одновременно действие развиваться форма складываться ***сфера_орфоэпия*** расширяться язык разный национальный становление ***норма_орфоэпический*** проходить разному процесс язык национальный ***норма_орфоэпический*** проходить этап русский особенность сформировываться основной ***норма_произносительный*** половина московский особенность половина говор ***национальный_язык*** норма окончательно складываться

Проанализировав результат работы алгоритма на материале второго корпуса, можно также заметить некоторые правильно сформированные биграммы, например, *норма_произносительный, норма_орфоэпический, национальный_язык.* Относительно некоторых (например, *норма_складываться*), примечательно то, что в исходном документе униграммы, входящие в состав биграммы, не только разнесены, но даже не связаны друг с другом; однако можно наблюдать множество контекстов, в

которых данные слова появляются совместно, например, «*…Современная литературная норма складывается с конца 19 в…»*. Биграмма *национальный_публичный* выделилась, вероятно, на основе синонимических отношений, биграмма *формирование_устный* - на основе совместной встречаемости этих слов в контекстах вроде «*…на протяжении многих веков участвовали в формировании устной нормы английского языка…»*. Наконец, биграмма *сфера_орфоэпия* была выделена, по всей видимости, ошибочно.

4.3. Построение тематической модели на основании корпуса с выделенными биграммами.

Второй этап работы алгоритма заключается в построении тематических моделей экспериментальных корпусов с помощью вероятностной тематической модели латентного размещения Дирихле и состоит из двух подэтапов: преобразование текстовой коллекции в матрицу с помощью инструмента *CountVectorizer* и непосредственное построение тематической модели на её основе (см. на стр. 37).

Эмпирическим путем были установлены параметры, обеспечивающие наиболее точный результат работы *CountVectorizer* и *LDA*: max_df = 0.6, min_df = 2, количество итераций алгоритма – 200, количество тем – 20, количество слов, представляющих каждую тему – 10 первых слов. Последний параметр неслучайно был выбран именно таким: исследования показали, что именно 10 первых слов содержат в себе 30% информации о теме, распределенной в других словах, что является достаточным для достаточно полного представления темы (Lau, Newman, Karimi, Baldwin 2010). При построении тематической модели не учитывались слова, встретившиеся менее, чем в двух документах, а также высокочастотные слова – в данном случае, содержащиеся более чем в 80% документов.

На третьем этапе полученные темы были заново обработаны с помощью модуля Phrases, что позволило выделить еще некоторое количество биграмм. Полученные конечные результаты представлены ниже (см.Таблица 7 и Таблица 8.).

4.4. Конечный результат работы алгоритма для корпуса текстов по радиоэлектронике, ракетостроению и технике

Таблица 7. Результаты тематического моделирования на корпусе текстов по радиоэлектронике, ракетостроению и технике, содержащем биграммы, а также последующее дополнительное выявление биграмм в темах (в рамку взяты биграммы, выделенные в уже выявленных темах)

№ темы	*Список первых 10 слов из тем*
1	вектор множество элемент **пространство_линейный** оператор пример расстояние образовывать состоять
2	оценка вероятность правило наблюдение гипотеза задача решение средний правдоподобие оптимальный
3	система скорость рлс дальность цель измерение объект антенна точность координата
4	исз потребитель точка доплеровский момент положение измерение шкала пересечение поверхность
5	код суммарный канал измерение сдвиг система разностный частота устройство информация
6	последовательность код символ состояние сигнальный расстояние путь **скорость_код** сверточный пример
7	активный **обзор_рлс** информация система эффективность **ширина_спектр** мощность дальность радиотехнический
8	канал передача связь пользователь система характеристика частота цифровой скорость полоса
9	фаза схема **огибать_омп** фильтр начальный амплитуда шум детектор частота
10	вероятность условие величина случайный **случайный_величина** определение **фурье_преобразование** событие результат
11	фильтр коэффициент алгоритм последовательность эквалайзер оценка линейный уравнение модель характеристика
12	**суммарный_канал**, детектор, антенна, фаза, амплитуда, устранение, измеритель, характеристика
13	мощность импульс потеря выходной энергия шум длительность входной отношение достигаться
14	дальность рлс спектр **радиолокационный_цель** точность антенна частота заданный объект
15	система измерение помеха измеритель вектор обработка радиотехнический фильтрация комплексный способ
16	код источник бит вход кодирование кодовый **кодовый_слово** символ канал уровень
17	различение цель дискриминатор **проверка_гипотеза** дальность условие оптимальный импульс характеристика рлс

№ темы	*Список первых 10 слов из тем*
18	частота спектр импульс время частотный модуляция огибать амплитуда фаза полоса
19	распределение процесс случайный дисперсия нормальный момент средний вероятность характеристика выражение
20	генератор частота опорный **потребитель_исз** шкала измерять скорость уравнение изменение

Проанализировав полученные результаты, можно заметить, что большинство выделенных биграмм действительно образуют логичные словосочетания, такие, как: *линейное пространство*, *ширина спектра*, *случайная величина*, *преобразование Фурье*, *суммарный канал*, *радиолокационная цель*, *кодовое слово*, *проверка гипотезы*. Остальные же выделенные биграммы вполне объяснимы: например, биграмма *потребитель_исз* был выделен, вероятно, вследствие того, что слова *исз* (искусственный спутник Земли) и *потребитель* (в значении *исследователь*, *наблюдатель*) часто встречаются в таких схожих контекстах, как *расстояние между потребителем и ИСЗ*, *скорость ИСЗ относительно потребителя* и так далее. Также слова *обзор* и *рлс* (радиолокационные станции), хоть и не встречаются в тексте стоящими рядом, во многих контекстах встречаются в непосредственной близости: например, *РЛС дальнего обзора, РЛС ближнего обзора*, *обзорные РЛС* и т.п.

4.5. Конечный результат работы алгоритма для корпуса текстов по лингвистике

Таблица 8. Результаты тематического моделирования на корпусе текстов по лингвистике, содержащем биграммы, а также последующее дополнительное выявление биграмм в темах
(в рамку взяты биграммы, выделенные в уже выявленных темах)

№ темы	*Список первых 10 слов из тем*
1	система форма основа глагол гласный ряд согласный тип диалект группа
2	местоимение лицо число человек числительный личный класс группа указательный **число_местоимение**
3	имя **форма_падеж** число род система предлог русский морфологический прилагательный
4	значение форма глагол тип русский выражение отношение грамматический функция вид
5	китайский **латинский_письменность** слог романский диалект тон время часть французский
6	предложение логический синтаксис вещь универсальный семантика анализ предмет событие психологический
7	литературный диалект русский современный арабский форма языковой разный социальный национальный
8	английский немецкий французский современный новый германский греческий форма период изменение
9	знак письмо буква система **алфавит_письменность** согласный гласный звук форма
10	строй мышление **влияние_оказывать** след эпоха звук соответствие русский характер создавать
11	морфема значение форма грамматический тип часть термин морфологический словоформа правило
12	словарь значение лексика русский лингвистический семантический **словарный_толкование** лексический **словарный_статья**
13	человек говорить языковой речевой случай мир система выражение речь отношение
14	русский значение тип случай правило форма фонетический **ударение_позиция** разный
15	семья группа время диалект говорить история языковой исследование современный число
16	единица признак фонема морфема звук разный речь общий свойство уровень
17	предложение глагол конструкция синтаксический подлежащее порядок дополнение сказуемое субъект тип
18	текст анализ перевод дискурс год средство ряд термин автор структура
19	языковой лингвистический исследование система теория лингвистика изучение языкознание развитие работа
20	объект состояние предмет термин языкознание служить средство стиль сравнение изучение

Результаты, полученные на втором корпусе текстов, отличаются от полученных на первом в основном тем, что изначально в темах выделилось

значительно меньше биграмм: *влияние_оказывать* и *словарный_статья*, и обе данные биграммы являются верными. Общее количество выделенных в темах биграмм также меньше, чем в предыдущем случае, однако значительная их часть являются достаточно точными: из них можно образовать логичные словосочетания *латинская письменность*, *алфавитная письменность*, *падежная форма*, «*словарное толкование*». Биграмм *ударение_позиция* также нельзя назвать случайным: составляющие его униграммы часто встречаются в одном предложении, например, «… фиксированное ударение ориентируется на крайние позиции в слове – либо на его начало, либо на конец…» или даже в непосредственной близости: «…Особенность фонетики собственно алюторского диалекта – противопоставление по долготе в системе гласных, …, динамическое позиционное ударение…». Униграммы *число* и *местоимение*, по всей видимости, были объединены также на основании частой совместной встречаемости в одном предложении (несложно представить такие контексты, описывающие формы местоимений); однако нельзя утверждать, что они образуют верную биграмму.

4.6. Оценка результатов работы предложенного алгоритма автоматического добавления биграмм в тематические модели

Далее перед нами встала задача оценки результаты работы алгоритма, а именно принятие решения о правильности выделенных биграмм. Стоит отдельно оговорить, что правильно выделенной биграммой мы считаем последовательность слов, обладающих синтаксическим и семантическим единством (Choueka 1998).

Для оценки результатов работы предложенного алгоритма было решено провести психолингвистический эксперимент.

Мы выбрали вариант постановки эксперимента с двумя аннотаторами, которыми стали студенты четвертого курса кафедры математической лингвистики Санкт-Петербургского государственного университета. Такой выбор экспертов обуславливается тем, что они знакомы с основами терминоведения и имеют навыки работы со специальными текстами. Аннотаторам был предложен для оценки список биграмм, выделенных ранее в темах для обоих корпусах (12 биграмм для одного корпуса и 8 биграмм для другого), и дано следующее задание:

> Оцените, пожалуйста, какие из приведенных ниже биграмм являются правильно выделенными и образуют логичные словосочетания русского языка по следующей шкале:
>
> 0 затрудняюсь ответить
>
> 1 данная биграмма является правильной
>
> 2 данная биграмма является частично правильной
>
> 3 данная биграмма является неправильной

Результаты оценивания приведены ниже в Таблице 9.

Таблица 9. Результаты оценки биграмм

Биграмма	***Оценка первого эксперта***	***Оценка второго эксперта***	***Средняя оценка***	
пространство_линейный	1	1	Биграмма	Корпус текстов по радиоэлектронике, ракетостроению и технике
обзор_рлс	0	0	?	
огибать_омп	0	0	?	
фурье_преобразование	1	1	Биграмма	
суммарный_канал	1	1	Биграмма	
радиолокационный_цель	1	1	Биграмма	
ширины_спект	1	2	Частично правильная/правильная биграмма	
случайный_величина	1	1	Биграмма	
кодовый_слово	1	1	Биграмма	
проверка_гипотеза	1	1	Биграмма	
потребитель_исз	0	1	?	
скорость_код	2	3	Частично правильная/неправильная биграмма	
число_местоимение	1	1	Биграмма	Корпус текстов по лингвистике
форма_падеж	1	1	Биграмма	
латинский_письменность	1	1	Биграмма	
алфавит_письменность	1	1	Биграмма	
влияние_оказывать	1	1	Биграмма	
словарный_толкование	1	1	Биграмма	
словарный_статья	1	1	Биграмма	
ударение_позиция	1	1	Биграмма	

В результате работы алгоритма на материале корпуса текстов по радиоэлектронике, ракетостроению и технике было выделено 20 тем, каждая из которых была представлена 10 единицами. Общее количество униграмм - 189, общее количество биграмм – 12. Экспертная оценка

практически полностью совпала с нашей и также показала, что количество однозначно правильных биграмм равняется 8. Таким образом, точность эксперимента составила 66,7 %.

В результате работы алгоритма на материале корпуса текстов по лингвистики было также выделено 20 тем, каждая из которых была представлена 10 единицами. Общее количество униграмм – 189, общее количество биграмм – 8. В данном случае экспертная оценка отличается от сделанной нами: эксперты определили все 8 результатов как верные биграммы, в то время как мы положительно оценили лишь 7 из них. Таким образом, точность эксперимента составила 100% при сравнении с экспертными данными или 87,5% при сравнении с условным эталоном.

Для оценки степени согласованности между двумя экспертами нами было решено применить коэффициент Каппа Коэна, поскольку он является наиболее удобным для такого рода оценок (Gwet 2016: Powers 2012). Общая формула коэффициента следующая:

$$k = \frac{P_0 - P_e}{1 - P_e},$$

где P_0 – наблюдаемая согласованность между экспертами, P_e – ожидаемая вероятность случайной согласованности (Cohen 1960). Максимальное значение коэффициента Каппа Коэна равняется 1; в этом случае между экспертами не существует разногласий. Если же он равняется 0, то наблюдаемое распределение вероятнее всего носит случайный характер.

Существуют разные мнения, какой показатель можно считать свидетельством достаточно надежной степени согласованности; наиболее часто применяется следующая шкала (Fleiss 1981):

- $k > 0{,}75$ – высокая степень согласованности;
- $0{,}40 < k < 0{,}75$ – достаточная степень согласованности;
- $k < 0{,}40$ – низкая степень согласованности.

В рамках данного исследования коэффициент Каппа Коэна был вычислен с помощью утилиты ReCal[12], находящейся в свободном доступе. Результаты приведены ниже в Таблице 10.

Таблица 10. Результат оценки согласия экспертов

Процент согласия	Общее количество предоставленных для анализа бирамм	Количество совпавших оценок	Количество расхождений	Коэффициент Каппа Коэна
85%	20	17	3	0.571

Таким образом, коэффициент Каппа Коэна показывает, что степень согласованности между экспертами достаточно высокая, чтобы считать их оценку достоверной.

[12] http://dfreelon.org/

Выводы к главе 4

В данной главе разработанный нами алгоритм был применен на двух корпусах текстов: на корпусе текстов по радиоэлектронике, ракетостроению и технике и на корпусе текстов по лингвистике. Полученные результаты показали достаточно высокую точность: 73% и 87,5% соответственно. Абсолютное количество выделенных биграмм для корпуса текстов по лингвистике было больше, однако в конечные темы их попало меньше по сравнению с первым корпусом текстов.

В перспективе планируется усовершенствовать выделение биграмм, используя частеречную разметку корпуса. В большинстве своем правильно выделенные темы формируются именно из существительных и именных групп (Wang, McCallum, Wei 2007), поэтому в дальнейшем планируется формировать биграммы в корпусе текстов преимущественно в соответствии с моделями *существительное + существительное в родительном падеже*, *существительное + прилагательное*; также не исключено добавление коллокаций *существительное + глагол*, поскольку зачастую такие словосочетания также являются характерными для специальных текстов. Наряду с этим, ставится задача полного исключения формирования таких ошибочных сочетаний, как, например, *существительное + наречие* или *прилагательное + глагол*. Также планируется приведение биграмм из лемматизированной формы к согласованным словосочетаниям путем их повторного поиска в корпусе текстов и замены на исходные формы.

ЗАКЛЮЧЕНИЕ

Итак, в данной работе был изучен такой современный инструмент для обработки естественного текста, как тематическое моделирование.

Тематическое моделирование – это «*способ построения модели текстовой коллекции, отражающий переход от совокупности документов, совокупности слов и документах коллекции к набору тем, характеризующих текстовую коллекцию*» (Митрофанова 2014). Иными словами, построение тематической модели помогает лучше понять глубинную семантику текстовой коллекции, что, в свою очередь, значительно облегчает дальнейшую работу с текстом, их кластеризацию и категоризацию.

На сегодняшний момент создано и успешно применяется большое количество различных тематических моделей. Их значительная часть основана на двух базовых алгоритмах – LDA и PLSA. Каждая из моделей помогает решить разные задачи, однако общим недостатком большинства из них является тот факт, что темы представляются исключительно униграммами. Это заметно ухудшает точность выделения тем и усложняет их интерпретацию исследователем. Несмотря на некоторые успешные реализации идеи включения *n*-грамм в тематические модели, на сегодняшний день нет универсального метода, позволяющего однозначно решить данную проблему. Одной из целей нашего исследования было создание подобного алгоритма.

Для достижения цели исследования были изучены различные вероятностные тематические модели и разработан собственный алгоритм для добавления в темы биграмм, основывающийся на их выделении в текстовой коллекции вначале на этапе предобработки текста, а затем – на

выявленных темах. Алгоритм был реализован на языке программирования *Python* и проверен на двух русскоязычных корпусах: на корпусе специальных текстов по радиоэлектронике, ракетостроению и технике и на корпусе текстов по лингвистике.

Полученные результаты можно считать удовлетворительными, поскольку более 70% выделенных в темах биграмм действительно таковыми являются. Таким образом, поставленные в начале данной работы задачи были решены.

В дальнейшем планируется усовершенствовать выделение биграмм с использованием частеречной разметки текста, обеспечить приведение биграмм к согласованной форме, а также проверить работу алгоритма на корпусах текстов других стилей.

Список литературы

1. Большая советская энциклопедия: в 30 т. / Гл. ред. А. М. Прохоров. — 3-е изд. — М. : Сов. энцикл., 1969 – 1978.
2. Бузикашвили Н.Е., Самойлов Д.В., Крылова Г.А. N-граммы в лингвистике // Сборник: Методы и средства работы с документами. М.: Диториал УРРС. 2000. 376 с. С. 91-130.
3. Воронцов К.В. Вероятностное тематическое моделирование // www.machinelearning.ru : web. — 2013.
4. Дубовик А.Р. *Автоматическое определение стилистической принадлежности текстов по их статистическим параметрам* // Международный научный симпозиум «Интернет и современное общество». СПб., 2017 [в печати].
5. Захаров В.П. *Корпусная лингвистика: Учебно-метод. пособие.* – СПб., 2005. – 48 с.
6. Кольцов С.Н., Кольцова О.Ю., Митрофанова О.А., Шиморина А.С. *Интерпретация семантических связей в текстах русскоязычного сегмента Живого Журнала на основе тематической модели LDA* // Технологии информационного общества в науке, образовании и культуре: сборник научных статей. Материалы XVII Всероссийской объединенной конференции «Интернет и современное общество» IMS–2014, Санкт- Петербург, 19 - 20 ноября 2014 г. СПб., 2014. С. 135–142.
7. Математическая энциклопедия / Ред. коллегия: И.М. Виноградов (глав. ред.) [и др.]. - Т. 1. А-Г. - М., 1977. - 1152 стб. (576 с.)
8. Митрофанова О.А. *Моделирование тематики специальных текстов на основе алгоритма LDA.* // Санкт-Петербург, 11—16 марта 2013 г.: Избранные труды. СПб.: Филологический факультет СПбГУ, а. 2014.-С. 220–233.

9. Нокель М.А. *Методы улучшения вероятностных тематических моделей текстовых коллекций на основе лексико-терминологической информации:*): дис. … канд.физ-мат.наук. –. М., 2016.– 159 с.
10. Нокель М.А., Лукашевич Н.В. *Тематические модели: добавление биграмм и учет сходства между кнграммами и биграммами.* // Вычислительные методы и программирование. -2015.- Т.6 - С. 215 – 234.
11. Aletras N. *Interpreting Document Collections with Topic Models.* PhD dissertation. University of Sheffield, Sheffield, UK. 2014.
12. Baker, J. K. *Stochastic modeling for automatic speech understanding.* // Readings in Speech Recognition, 1990. -P. 297–307.
13. Bharucha-Reid A. T. *Elements of the Theory of Markov Processes and Their Applications*. New York: McGraw-Hill, 1960.
14. Blei D.M, McAuliffe J.D. *Supervised topic models.* // In: Advances in Neural Information Processing Systems (NIPS) . Cambridge, MA, MIT Press, 2007.-P.121-128.
15. Blei D.M, Ng A., Jordan M.. *Latent Dirichlet Allocation* // Journal of Machine Learning Research. 2003. Т. 3. -P. 993–1022.
16. Blei D.M., Lafferty J.D. *Dynamic topic models* // In Proceedings of the 23rd international conference on Machine learning (ICML 2006). New York: ACM Press, 2006. С. 113–120.
17. Boyd-Graber J.L., Blei D.M., Zhu X. *A Topic Model for Word Sense Disambiguation*. // Proceedings of the Joint Meeting of the Conference on Empirical Methods in Natural Language Processing and The Conference on Computational Natural Language Learning. Czech Republic: Prague; 2007.
18. Cano Basave A.E., He Y., Xu R. *Automatic Labelling of Topic Models Learned from Twitter by Summarisation* // Proceedings of the 52nd

Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Stroudsburg, PA, USA: Association for Computational Linguistics, 2014. C. 618–624.

19. Choueka Y. *Looking for Needles in A Haystack, or Locating Interesting Collocational Expressions in Large Textual Databases.* // In Proceedings of Recherche d'Informations Assistee par Ordinateur 1988 (RIAO'88). Cambridge, USA, 1988. C.609–623.
20. Clark A., Fox C., Lappin S. *The Handbook of Computational Linguistics and Natural Language Processing*. Hoboken, NJ: Wiley-Blackwell, 2013.
21. Cohen J. *A Coefficient of Agreement for Nominal Scales*. Educational and Psychological Measurement, 1960:37-46.
22. Cohn D., Hofmann T. *The missing link- a probabilistic model of document content and hypertext connectivity. //* In: Advances in Neural Information Processing Systems (NIPS) 13. Cambridge, MA, MIT Press, 2001.-7 p.
23. Darling W.M. *A theoretical and practical impleentation tutorial on topic modeling and Gibbs sampling*. School of Computer Science, University of Guelph, 2011.-10 p.
24. Daud A., Li J., Zhou L., Muhammad F. *Knowledge discovery through directed probabilistic topic models: a survey* // Frontiers of Computer Science in China. 2010. T. 4. № 2. -P. 280–301.
25. Fleiss J.L. *Statistical Methods for Rates and Proportions (2nd ed.).* New York: John Wiley, 1981.
26. Gaussier, E., Renders, J.-M., Matveeva, I., Goutte, C., Déjean, H.. *A geometric view on bilingual lexicon extraction from comparable corpora.* // In Proceeding of the 42th Annual Meeting of the Association for Computational Linguistics. -2004.-P.526-533.

27. Gernot A. Fink *Markov Models for Pattern Recognition: From Theory to Applications.* Springer-Verlag New York, NJ, USA 2007.
28. Greene D., O'Callaghan D., Cunningham P. *How many topics? stability analysis for topic models. //* Joint European Conference on Machine Learning and Knowledge Discovery in Databases - Springer Berlin Heidelberg -2014.- P. 498 – 513.
29. Griffiths T., Steyvers M., Tenenbaum J. *Topics in semantic representation. Psychological Review. //* American Psychological Association - Vol. 114, № 2. –2007.- P. 211-244.
30. Griffiths T.L, Steyvers M. *Finding scientific topics. //* In: Proceedings of the National Academy of Sciences. USA. 2004.-P. 5228–5235.
31. Griffiths T.L. , Steyvers M., Blei D. M, Tenenbaum J.B. *Integrating topics and syntax. //* In: Advances in Neural Information Processing Systems (NIPS) 17. Cambridge, MA, MIT Press. 2005. -18 p.
32. Gwet L. K. *Testing the Difference of Correlated Agreement Coefficients for Statistical Significance.* Educational and Psychological Measurement 2016, Vol. 76(4) 609–637
33. Harris Z. *Distributional Structure*. // In Word 10 (23). 1954. C. 146-162.
34. Hastie T., Tibshirani R., Friedman J. *The Elements of Statistical Learning: Data mining, Inference, and Prediction.* New York: Springer, 2009. C. 485–586. ISBN 978-0-387-84857-0.
35. Hindle A., Ernst N.A., Godfrey M.W., Mylopoulos J. *Automated Topic Naming: Supporting Cross-Project Analysis of Software Maintenance Activities* // Empirical Software Engineering. 2013. T. 18. № 6. C. 1125–1155.
36. Hofmann T. *Probabilistic latent semantic analysis. //* In: Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence (UAI), Stockholm, Sweden, 1999.- P. 289-296.

37. Hu, W., Shimizu, N., Nakagawa, H., And Sheng, H. *Modeling Chinese Documents with Topical Word-Character Models.* // In Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008). Manchester, UK, 2008. C. 345–352.
38. Jelinek, F. *Self-organized language modeling for speech recognition.* // In Readings in Speech Recognition, 1990.-P. 450-506.
39. Jurafsky D., M. H. James. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition.* Dorling Kindersley Pvt, Ltd., 2014.
40. Lau J. H., Baldwin T., Newman D. . *On Collocations and Topic Models.* // ACM 131 Transactions on Speech and Language Processing. – ACM Press. -– Vol. 10, № 3. – 2013.-P. 1-14.
41. Lau J.H., Newman D., Karimi S., Baldwin T. *Best Topic Word Selection for Topic Labelling* // COLING'10 In Proceedings of the 23rd International Conference on Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2010.- P. 605–613.
42. Magatti D., Calegari S., Ciucci D., Stella F. *Automatic labeling of topics* // ISDA 2009 - 9th International Conference on Intelligent Systems Design and Applications. Pisa: IEEE, 2009. C. 1227–1232.
43. Manning C., Schütze H. *Foundations of Statistical Natural Language Processing*. MA, USA: MIT Press Cambridge, 1999.
44. McCallum A., Corrada-Emmanuel A., Wang X. *The author-recipient-topic Model for Topic and Role Discovery in Social Networks: Experiments with Enron and Academic Email.* 2004. -16 p.
45. Mei Q., Shen X., Zhai C. *Automatic labeling of multinomial topic models* // In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '07. New York, New York, USA: ACM Press, 2007. C. 490.

46. Mirzagitova A., Mitrofanova O. *Automatic assignment of labels in Topic Modelling for Russian Corpora* // In Proceedings of 7th Tutorial and Research Workshop on Experimental Linguistics, ExLing 2016 / ed. A. Botinis. – Saint Petersburg: International Speech Communication Association, 2016. P. 115-118.
47. Nolasco D., Oliveira J. *Detecting Knowledge Innovation through Automatic Topic Labeling on Scholar Data* // 49th Hawaii International Conference on System Sciences (HICSS). Koloa, HI: IEEE Computer Society, 2016. C. 358–367.
48. Pedregosa et al. *Scikit-learn: Machine Learning in Python.* // Journal of Machine Learning Research 12.-2011.-P. 2825-2830.
49. Powers D. *The Problem with Kappa* // Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, 2012. C. 345–355
50. Segalovich I. *A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine.* // In Proceedings of the International Conference on Machine Learning; Models, Technologies and Applications. MLMTA'03, June 23 - 26, 2003, Las Vegas, Nevada, USA, 2003.
51. Segalovich I., Maslov M. *Russian Morphological Analysis and Synthesis With Automatic Generation of Inflection Models For Unknown Words. //* Dialog'98 (in Russian) , 1998.
52. Shannon, C. E. *A mathematical theory of communication.* // Bell System Technical Journal, 27(3), 1948.-P.379–423.
53. Steyvers M., Smyth P., Rosen-Zvi M., Griffiths T. *Probabilistic author-topic models for information discovery. //* In Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, Washington, 2004.- P. 306-315.

54. Vulic I., De Smet W., M-F. Moens. *Identifying Word Translations from Comparable Corpora Using Latent Topic Models* // In Proceeding of the 49th Annual Meeting of the Association for Computational Linguistics. Portland, Oregon, 2011. -P. 479-484.

55. Wallach H. *Topic Modeling: Beyond Bag-Of-Words* // In Proceedings of the 23rd International Conference on Machine Learning. -2006.-P. 977-984.

56. Wang X., McCallum A., Wei X. *Topical N-Grams: Phrase and Topic Discovery, with an Application to Information Retrieval* // Seventh IEEE International Conference on Data Mining (ICDM 2007). NY: IEEE, 2007.- P. 697–702.

Электронные ресурсы

1. MyStem

   URL: https://tech.yandex.ru/mystem/

   (дата последнего обращения 27.04.2017)
2. Scikit-Learn

   URL: http://scikit-learn.org/stable

   (дата последнего обращения 27.04.2017)
3. NLTK

   URL: http://www.nltk.org/

   (дата последнего обращения 27.04.2017)
4. GenSim

   URL: https://radimrehurek.com/gensim/

   (дата последнего обращения 27.04.2017)
5. ReCal

   URL: http://dfreelon.org/

   (дата последнего обращения 21.05.2017)

Приложение 1. Список стоп-слов на основе словарей служебных слов и оборотов НКРЯ

а, аа, а-а, ааа, а-а-а, а-а-а-а, абы, авось, ага, аж, аз, ай, ай-ай-ай, айда, ай-яй-яй, аки, але, али, алле, алло, аль, а-ля, аминь, ан, апчхи, атас, ау, аф, ах, ахти, аще, б, ба, бабах, ба-бах, баста, бах, бац, без, безо, бен, бис, бишь, благо, благодаря, близ, блин, бляха-муха, бо, боже, более, больше, бом, браво, брр, бррр, брысь, бу-бу-бу, буде, будто, буль-буль, бум, бы, было, быть, в, вай, ван, вау, ваш, вблизи, ввиду, вглубь, вдоль, ведь, ведь, везде, весь, взамен, виват, вишь, включая, вкруг, вместо, вне, внизу, внутри, внутрь, во, во-во, возле, вокруг, вон, вон, вона, вообще-то, во-он, во-от, вопреки, восемнадцатый, восемнадцать, восемь, восемьдесят, восемьсот, вослед, восьмеро, восьмидесятый, восьмой, вот, вперед, впереди, впрямь, вроде, все, всегда, всего, все-таки, вслед, вследствие, всюду, всякий, всяко, всякое, второй, вы, выше, где, где-либо, где-нибудь, где-то, геть, глядь, гм, гоп, гы, да, да-а, да-а-а, дабы, давай, давайте, да-да, да-да-да, даже, дай, дайте, дак, данный, два, двадцатый, двадцать, двенадцатый, двенадцать, двести, двое, д-да, де, девяносто, девяностый, девятнадцатый, девятнадцать, девятый, девять, девятьсот, дель, ден, дер, десятеро, десятый, десять, ди, для, до, добро, доколе, дон, доселе, дотоле, другие, другое, другой, дудки, ды, дык, его, едва, ее, ежели, ежли, ей-богу, ей-ей, ейный, елки-палки, е-мое, если, ет, ето, ето, етот, еще, ж, же, за, заместо, зато, зачем, зачем-то, здесь, здеся, здорово, здравствуй, здравствуйте, здрасте, значит, зы, и, ибн, ибо, , идти, иже, иже, из, из-за, изнутри, изо, из-под, и-и, или, иль, именно, имхо, ин, иначе, иначе, иной, исключая, исключительно, итак, ить, их, ихний, ишь, ишь, к, ка, ка-ак, кабы, каждый, кажный, как, как-либо, как-нибудь, как-никак, како, каков, каково, каковой, какой, какой-либо, какой-нибудь, какой-никакой, какой-то, как-то, касательно, кис-кис, ко, когда, когда, когда-либо, когда-нибудь, когда-то, кое, кое-где, кое-как,

кое-какой, кое-кто, кое-что, кой, кой-какой, кой-кто, кой-то, кой-что, коли, коль, конечно, который, кроме, кругом, кто, кто-кто, кто-либо, кто-нибудь, кто-то, ку, куда, куда-либо, куда-нибудь, куда-то, куды, ку-ку, кыш, ла, ладно, ле, ли, ли, либо, лишь, лучше, ль, любой, м, мало, марш, мда, м-да, мдя, меж, между, меньше, мерси, мимо, мля, мм, м-м, ммм, м-м-м, , мочь, может, многие, многий, много, многое, много-много, мой, мол, мы, мяу, на, навроде, навсегда, над, надо, на-ка, накануне, наперекор, наподобие, напротив, насчет, нате, наш, н-да, не, неа, не-а, небось, невесть, не-е, не-е-ет, не-ет, нежели, неизвестно, некий, некогда, некого, некоторые, некоторый, некто, немало, немногие, немногий, немного, немногое, немножко, несколько, нет, нет-нет, нет-нет-нет, неужели, неужто, нехай, нечего, нечто, нешто, ни, нибудь, нигде, ниже, никак, никакой, никогда, никой, никто, никуда, ниоткуда, нипочем, нисколечко, нисколько, ниче, ниче, ничего, ничей, ничо, ничто, ничуть, ништяк, н-не, н-нет, н-ну, но, но-но, ну, ну-ка, ну-ко, ну-ну, ну-с, ну-у, нэ, о, об, оба, обо, обоего, оглы, ого, ого-го, о-го-го, один, одиннадцатый, одиннадцать, однако, одно, ой, ой-ой-ой, ок, о'кей, около, окрест, окромя, о'кэй, он, она, они, оно, оный, о-о, о-о-о, оп, остальное, остальной, остальные, от, откуда, откуда-нибудь, откуда-то, относительно, ото, отовсюду, отсюда, отсюдова, оттого, оттого-то, оттуда, оттудова, отчего, отчего-то, офф, ох, ох-хо-хо, очень, пам, пардон, первый, перед, передо, пи, пиф-паф, пли*, по, по-вашему, поверх, повсюду, по-всякому, под, поди, подле, подо, подобно, по-другому, поелику, пожалуйста, по-за, позади, по-иному, пока, покамест, покуда, полноте, полста, полтора, полтораста, полундра, помимо, по-моему, по-над, по-нашему, понеже, поперек, по-своему, посему, посередине, посередь, поскольку, после, посреди, посредине, посредством, постольку, по-твоему, потом, потому, потому-то, почем, почему, почему-либо, почему-то, почто, пошто, поэтому, поэтому-то, правда, превыше, пред, предо, прежде, при, притом, причем, про, промеж, просто, против, противу, прочая, прочее,

прочий, прям, прямо, пу, пускай, пусть, путем, пущай, пшел, пятеро, пятидесятый, пятнадцатый, пятнадцать, пятый, пять, пятьдесят, пятьсот, равно, равняйсь, ради, раз, разве, ровно, р-раз, с, сам, самый, самый-самый, сверх, свое, свой, свыше, се, себе, себя, седьмой, сей, сейчас, сем, семеро, семидесятый, семнадцатый, семнадцать, семь, семьдесят, семьсот, середь, сзади, сие, сиречь, сичас, сквозь, сколь, сколько, сколько-нибудь, сколько-то, сколь-нибудь, словно, со, собственно, согласно, сообразно, соответственно, сорок, сороковой, сорри, сотый, спасибо, спасибочки, спустя, среди, средь, сродни, становиться, сто, столь, столько, столько-то, стоп, стук, супер, супротив, сю, сюда, сюды, сяк, сякой, сям, та, та-ак, так, также, таки, тако, таков, таковой, таковский, такой, такой-сякой, такой-то, так-так, так-таки, так-так-так, так-то, там, тама, там-то, та-та, та-та-та, твой, те, тем, теперь, тик-так, типа, то, тогда, тогда-то, тож, той, тока, токмо, токо, только, только-то, топ, то-се, тот, то-то, то-то, тот-то, точно, тра-та-та, трах, третий, три, тридевять, тридцатый, тридцать, тринадцатый, тринадцать, трис, триста, трое, тсс, тс-с, ттт, туда, туда-сюда, туда-то, туды, тук, тук-тук, тук-тук-тук, тут, тута, тут-то, ту-ту, ты, тьфу, тьфу-тьфу, тьфу-тьфу-тьфу, тю, у, уа, увы, угодно, угу, уж, ужели, ужель, уй, ура, усе, у-у, ууу, у-у-у, уф, ух, фи, фон, фра, фу, ха, ха-ха, ха-ха-ха, хвать, хе, хех, хе-хе, хе-хе-хе, хи, хи-хи, хи-хи-хи, хлоп, хм, хны, хо, хорошо, хоть, хотя, хо-хо, хо-хо-хо, хошь, хр, хрясь, хто, цоб, цыц, чаво, чай, чао, че, чево, чего, чегой-то, чего-то, чей, чей-либо, чей-нибудь, чей-то, чем, через, черт-т, четверо, четвертый, че-то, четыре, четыреста, четырнадцатый, четырнадцать, чи, чик-чик, чмок, чо, чого, чрез, что, чтоб, чтобы, чтой-то, что-либо, что-нибудь, что-нить, что-о, что-о-о, что-то, что-что, чу, чур, чуть, ч-черт, ша, шалом, шестеро, шестидесятый, шестисотый, шестнадцатый, шестнадцать, шестой, шесть, шестьдесят, шестьсот, шо, шоб, што, штоб, шу, ща, щелк, э, эвон, эврика, эге, эдак, эдакий, эй, эк, эка, экий, эль, энный, эт, этак, этакий, это, этот, эт-то, эх, ээ, э-э, э-эх, эээ, э-э-э, я, яко, якобы, 1, 2, 3, 4, 5, 6, 7, 8, 9,

0, ,, ., !, №, @, #, $, ^, ;, %, :, &, ?, *, (,), _, +, =, -, {, }, [,], ", ', |, \, /, ~, a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z, A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, W, X, Y, Z.

Приложение 2. Список стоп-слов, дополняющий список стоп-слов на основе словарей служебных слов и оборотов НКРЯ

использовать, позволять, определять, осуществлять, следовать, иметь, предполагать, являться, рассматривать, рассматриваться, показывать, сформировывать, замечать, описываться, использоваться, располагать, подчеркивать, оказываться, описывать, возникать, допускать, удовлетворять, использоваться, определяться, находить, означать, приводить, составлять, называть, происходить, принимать, называться, получать, выбираться, заключаться, учитывать, вычислять, иметь, иметься, описывать, полагать, повторять, содержаться, сравниваться, находиться, обозначать, основываться, соответствовать, представлять, давать, появление, применять, применяться, требовать, интерпретировать, фиксировать, производиться, характеризовать, разрабатывать, видеть, входить, образовываться, можно, должно, подставлять, даваться, содержать, принадлежать, знать, выражаться, наличие, отсутствие, обнаружение, соответствующий, соотношение, использование, прохождение, следовательно, помощь, вычисление, действительный, например, действительно, определенный, рассмотрение, выход, ошибка, рис, схема.