**Федеральное государственное бюджетное образовательное учреждение  
высшего образования  
«Саратовский государственный технический университет имени  
Гагарина Ю.А.»**

*На правах рукописи*

Петров Андрей Сергеевич

**МАТЕМАТИЧЕСКИЕ МОДЕЛИ И МЕТОДЫ ИЗВЛЕЧЕНИЯ ТЕРМИНОВ ИЗ РУССКОЯЗЫЧНЫХ ТЕКСТОВ**

*(наименование темы научно-квалификационной работы)*

|  |
| --- |
| 09.06.01 Информатика и вычислительная техника |
| *(код и наименование направления) по направленности* |
| Математическое моделирование, |
| численные методы и комплексы программ |
| *(наименование направленности)* |

Научно-квалификационная работа (диссертация)

Научный руководитель:

д.ф.-м.н., профессор,

*(ученая степень, ученое звание)*

проф. каф. ИКСП

*(должность)*

Шульга Татьяна Эриковна

*(ФИО полностью)*

Саратов 2019

Содержание

[ВВЕДЕНИЕ 3](#__RefHeading___Toc24382_1389336206)

[1. Анализ моделей и методов, применяемых при извлечении терминов из текста 10](#__RefHeading___Toc1143_773439566)

[1.1. Определение термина 10](#__RefHeading___Toc1145_773439566)

[1.2. Существующие обзоры методов извлечения терминов из текста 14](#__RefHeading___Toc1145_7734395661)

[1.3 Алгоритм работы методов извлечения терминов из текста 16](#__RefHeading___Toc4813_773439566)

[1.4. Методы на основе статистического параметра частоты вхождений 20](#__RefHeading___Toc4815_773439566)

[1.5. Методы на основе контекстов вхождений 25](#__RefHeading___Toc4809_773439566)

[1.6. Методы оценки эффективности 28](#__RefHeading___Toc4817_773439566)

[1.7. Выводы 30](#__RefHeading___Toc4811_773439566)

[2. Математическая модель русскоязычного текстового документа, предназначенная для извлечения терминов из текста 32](#__RefHeading___Toc12842_3435704646)

[2.1. Математическая модель русскоязычного текстового документа 32](#__RefHeading___Toc21931_3435704646)

[2.2. Применение математической модели русскоязычного текстового документа для извлечения терминов из текста 37](#__RefHeading___Toc21931_34357046461)

[2.3. Выводы 39](#__RefHeading___Toc21931_343570464611)

[3. Разработка комплекса проблемно-ориентированных программ для извлечения терминов из текста 40](#__RefHeading___Toc12217_3224744665)

[3.1. Анализ требований и проектирование 40](#__RefHeading___Toc12219_3224744665)

[3.2. Разработка 42](#__RefHeading___Toc11098_2718481933)

[3.3. Выводы 45](#__RefHeading___Toc25264_2718481933)

[ЗАКЛЮЧЕНИЕ 46](#__RefHeading___Toc24434_1389336206)

[СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ 48](#__RefHeading___Toc24436_1389336206)

# ВВЕДЕНИЕ

**Актуальность темы.** Специалисты по терминоведению и компьютерной лингвистике длительное время ведут споры по поводу природы понятия «термин». Под термином обычно понимается слово или словосочетание на естественном языке, описывающее понятие конкретной предметной области. Извлечение терминов из текста — это процесс выделения из текста и распознавания ключевых слов и терминов, описывающих понятия определенной предметной области. Извлечение терминов из текста является важным этапом в решении ряда задач, связанных с обработкой текстов предметной области, среди которых — информационный поиск, машинный перевод, классификация и кластеризация документов, построение онтологий, глоссариев и тезаурусов.

Существует большое количество методов автоматического извлечения терминов из текста на естественном языке. Часть из них создавалась с учетом особенностей ряда таких языков, относящихся к индоевропейским языкам романо-германской группы, как английский и французский, и не учитывает особенности грамматики и морфологии русского языка. Кроме того, эффективность работы данных методов остается низкой и варьируется в зависимости от предметной области и стиля исходного текста. Также большое количество методов извлечения терминов из текстов для обработки используют входные данные, размеченные вручную. Данные особенности и недостатки ограничивают их область применимости.

Большая часть существующих методов относятся к категории статистических методов извлечения терминов из текста. Исследованиям в области применения статистических критериев для выбора ключевых слов и понятий посвящены работы Roberto Navigli, Paola Velardi, Katerina Frantzi, Sophia Ananiadou, Hideki Mima, David A. Evans, Robert G. Lefferts. В исследованиях указанных авторов основным критерием, использующимся для процесса извлечения терминов из текста, является критерий частоты вхождения кандидатов в рассматриваемую коллекцию текстовых документов. На таком принципе построена работа методов Domain Consensus [1], C-Value [2], TF-IDF [3]. Большая часть понятий в предметных областях являются многословными терминами. Однако в текстах, написанных на естественных языках, вне зависимости от предметной области такие понятия встречаются значительно реже по сравнению с однословными понятиями. Существующие методы по-разному компенсируют данный эффект. Так, методы, основывающиеся на подсчете меры ассоциации, учитывают вероятность совместного употребления слов в составе определенного термина. К таким метрикам относятся, в частности, взаимная информация (Mutual Information, MI), логарифмическое правдоподобие (Loglikelihood Ratio). В ряде методов учитывается вложенность понятий предметной области. К таким, например, относится C-Value [2]. Вложенные термины (англ. nested terms) - это понятия, содержащиеся в исходном тексте как по отдельности, так и в составе других понятий [49]. Другим подходом к улучшению результатов работы методов извлечения ключевых слов из текста является учёт контекста вхождений. Одним из методов, использующим данный подход, является, например, NC-Value [2].

Еще одним способом улучшения результатов извлечения многословных терминов является частота вхождений слов и словосочетаний во внешнюю по отношению к рассматриваемому тексту коллекцию документов с точки зрения предметной области. Причем документы подбираются таким образом, чтобы их тематика не ограничивалась определенной предметной областью. То есть в случае обработки исходного текста военно-исторической предметной области подбирается корпус новостей или художественной литературы. Текстовый корпус – это сформированная по определенным правилам, структурированная коллекция текстов. К методам, учитывающим данную метрику, относятся, в частности, Domain Relevance [5], Weirdness [7], Relevance [7], Domain Pertinence [8].

Часть методов извлечения терминов из текста применяют внешние ресурсы: например, поисковые машины [9, 10] или энциклопедию Википедия [11, 12]. Большая часть подобных методов не использует заранее подготовленную коллекцию текстовых документов, применяя исключительно внешние ресурсы для поиска и извлечения терминов. Однако, при таком подходе без предварительного составления текстового корпуса достаточно сложно добиться высокой точности извлечения перечня терминов.

Таким образом, значительная часть существующих методов извлечения терминов из текста, основывающихся на использовании статистических метрик для выбора понятий, работает только с текстами выбранной предметной области. Как правило, тексты, используемые в процессе обработки, обычно не являются размеченными текстовыми корпусами, в связи с чем не содержат в себе необходимого объема, в том числе, лингвистической и морфологической информации для автоматического извлечения терминов. Ряд методов для решения этой проблемы используют внешние ресурсы, такие как поисковые машины или контрастные корпуса текстов других предметных областей. Текстовые документы в большинстве своем не имеют структуры и позволяют использовать только статистическую информацию о частоте встречаемости слов и словосочетаний в тексте без учета особенностей рассматриваемой предметной области.

В качестве модели представления исходных текстовых данных большая часть статистических методов извлечения терминов из текста использует модель «мешок слов» (англ. Bag of words). Данная модель представляет текстовый документ в виде множества составляющих его слов, а также их частот встречаемости. Разновидностью модели Bag of words является частотная модель текста, в которой каждому слову соответствует весовой коэффициент, который определяется в зависимости от выбранной метрики: например, частоты вхождения слова в документе TF (англ. Term frequency), логарифма частоты вхождения слова Log TF или обратной частоты документов IDF (англ. Inverse document frequency). Однако вышеописанные модели не учитывают порядок слов и могут быть использованы для извлечения лишь однословных понятий из текста.

Для решения этой проблемы вводится модель коллокации N-грамма (англ. N-gram). Коллокация — словосочетание, являющееся синтаксически и семантически целостной единицей. Данная модель позволяет представить словосочетания из N слов, при N=2 словосочетание имеет название биграммы, при N=3 — триграммы. Таким образом, для извлечения многословных терминов возможно использование частотной модели текста вместе с моделью коллокаций N-грамма. Однако, данные модели не принимают во внимание грамматические особенности русского языка, с учётом которых возможно повышение качества результатов методов извлечения терминов из текста.

Вышесказанное определило актуальность настоящей работы, а также, следующие из неё цели и задачи.

**Цель работы** – cодержательная постановка задачи автоматического извлечения терминов из текста, разработка математической модели русскоязычного текстового документа, которая может быть использована для решения рассматриваемой задачи, проведение исследования статистических методов выделения понятий на текстах исторической предметной области

**Задачи работы**. Для достижения поставленной цели были поставлены и решены следующие задачи:

1. Исследовать существующие математические модели и методы, применяемые при автоматическом извлечении терминов из текста.
2. Разработать математическую модель русскоязычного текстового документа.
3. Создать комплекс проблемно-ориентированных программ, реализующий разработанную математическую модель и методы автоматического извлечения терминов из русскоязычного текста.

**Объект исследования** – процесс извлечения терминов из русскоязычных текстовых документов.

**Предмет исследования** – математические модели и методы, применяемые при выделении ключевых слов и понятий.

**Методы исследования**: аппарат теории множеств, методы обработки многомерных данных, методы параллельного и объектно-ориентированного программирования, методы статического анализа.

**Научная новизна работы** (соответствует пунктам 1, 4 паспорта специальности 05.13.18) заключается в следующем:

Разработана математическая модель текстового документа на русском языке, учитывающая грамматические особенности русского языка, описанная в форме терминологии с использованием аппарата теории множеств. С помощью разработанного комплекса проблемно-ориентированных программ исследована применимость методов автоматического извлечения терминов из текста на текстах военно-исторической предметной области.

**Практическая значимость работы** обусловлена созданием комплекса проблемно-ориентированных программ для извлечения терминов из коллекции текстовых документов.

**Основные положения и результаты**, выносимые на защиту:

1. Математическая модель предметной области текстового документа на русском языке, учитывающая грамматические особенности русского языка, описанная в форме терминологии с использованием аппарата теории множеств.
2. Разработанный комплекс проблемно-ориентированных программ «Программный комплекс для извлечения терминов из текстов», на основе предложенной модели.

**Достоверность результатов**.

Достоверность и обоснованность результатов работы определяется корректной постановкой задач, применением методов статистического анализа, обработки многомерных данных. Результаты исследования подтверждены вычислительными экспериментами и практической апробацией.

**Апробация работы**. По основным результатам работы сделано 4 доклада на 4 международных, всероссийских и региональных конференциях:

1. Международная научно-практическая конференция «Проблемы управления в социально-экономических и технических системах» (ПУ-2016).
2. Международная научная конференция «Информационно-коммуникационные технологии в науке, образовании и производстве» (ICIT-2016).
3. Международная научная конференция «Информационно-коммуникационные технологии в науке, образовании и производстве» (ICIT-2017).
4. Международная научная конференция «Проблемы управления, обработки и передачи информации» (УОПИ-2017).

**Публикации**

Результаты работы опубликованы в 5 изданиях, 1 из которых является изданием, рекомендованным ВАК, 4 индексируются в базе РИНЦ. Имеется свидетельство о государственной регистрации программы для ЭВМ №2017663283 от 28 ноября 2017 г.

**Соответствие темы диссертации требованиям паспорта специальностей научных работников.** Диссертационная работа выполнена в соответствии с паспортом специальности 05.13.18 – Математическое моделирование, численные методы и комплексы программ, п. 1. Разработка новых математических методов моделирования объектов и явлений; п. 4. Реализация эффективных численных методов и алгоритмов в виде комплексов проблемно-ориентированных программ для проведения вычислительного эксперимента.

**Структура и объем работы.** Научно-квалификационная работа состоит из введения, трех глав, заключения, списка использованной литературы и приложения. Работа содержит 53 страниц, включая 2 рисунка, библиографический список из 58 наименований.

# 1. Анализ моделей и методов, применяемых при извлечении терминов из текста

В первой главе приведено описание рассматриваемой задачи и существующих методов решения, а также описание множества существующих определений понятия «термин».

## 1.1. Определение термина

Исследованиям в области терминоведения посвящено большое количество работ, первые из которых появились более 80 лет назад. За долгие годы развития данной науки появилось большое количество определений понятия «термин», отражающих множество различных аспектов исследований, которые проводились и проводятся в данной области. Так, согласно Татаринову В.А. в настоящий момент изучаются как лингвистические аспекты, так и, в частности, психолингвистические, логические, гносеологические, системные, дидактические, информационные аспекты, переводческие аспекты [13].

Несмотря на большое количество работ, посвященных исследованию понятия «термин», авторами отмечается отсутствие единого, универсального определения. Так, Мякшин К.А. в своей работе отмечает, что неоднократные попытки лингвистов сформулировать удовлетворяющее всех определение понятия «термин» оказались малопродуктивными [15]. Рассмотрим некоторые из существующих определений, используемые в компьютерной лингвистике, а также признаки термина.

А. А. Реформатский дает следующее определение: «термины - это слова, ограниченные своим особым назначением; слова, стремящиеся быть однозначными как точное выражение понятий и называние вещей. Термины существуют не просто в языке, а в составе определенной терминологии ...» [16].

К.А. Мякшин выделяет две точки зрения на понятие «термин»: субстанциональную и функциональную. [15]. Согласно субстанциональной точке зрения, термины — это особые слова и словосочетания, обладающие определенным набором критериев, к которым, в частности, относятся моносемантичность, независимость от контекста и нейтральность. Согласно функциональной точке зрения «в роли термина может выступать любое слово», кроме того, «термины — это не особые слова, а слова в особой функции» [17]. Также стоит отметить, что во многих предметных областях существуют консубстанциональные термины: лексические единицы, присутствующие как в профессиональной, так и в обыденной речи, «которые вызывают ряд трудностей при выделении терминологической лексики из словарного состава языка» [14]. Кроме того, среди немаловажных явлений, не рассматриваемых с субстанциональной точки зрения на понятие «термин», существуют явления терминологизации и детерминологизации. Терминологизацией называется процесс перехода общеупотребительной лексики в категорию терминов, в то время как детерминологизация — это процесс потери термином специального значения [18].

Ряд работ зарубежных исследователей посвящено вопросу взаимоотношений между лексической единицей, представляющей собой термин, и понятием, выражаемым термином.

Так, например, по мнению О. Вюстера предметные области состоят из наборов понятий, или мыслительных конструкций, в то же время как термины служат текстовым представлением этих понятий [19].

Международная Организация по Стандартизации, в 2000 году опубликовала стандарт, описывающий понятия терминологии – ISO 1087-1:2000 «Terminology work -- Vocabulary -- Part 1: Theory and application» [20]. Данный документ дает следующее значение понятию «термин»: «term: verbal designation (3.4.1) of a general concept (3.2.3) in a specific subject field» (Перевод: термин — это словесное обозначение определенного понятия в определенной предметной области). Понятие «designation» (обозначение) в данном определении обозначает «representation of a concept (3.2.1) by a sign which denotes it» (перевод: представление понятия с помощью символьного обозначения). Понятие «concept» (понятие) обозначает «unit of knowledge created by a unique combination of characteristics» (перевод: единица знаний, образованная путем уникального объединения характеристик). А понятие «characteristic» (характеристика) - это «abstraction of a property of an object (3.1.1) or of a set of objects» (перевод: абстракция свойств объекта или множества объектов).

Рассмотрим описание понятия «термин» через описание характерных признаков, отличительных по сравнению с общеупотребительной лексикой. В работе Н.А. Астраханцева [11] была предложена классификация признаков, сформулированная на основе трёх аспектов термина, выделенных А. Д. Хаютином в работе [21]. Данными аспектами являются синтаксический, семантический и прагматический.

**Синтаксические признаки**

К данной группе относятся признаки, обусловленные формой термина.

1. Номинативность. В качестве терминов обычно рассматриваются имена существительные или словосочетания, построенные на их основе [22].
2. Нормативность. Соответствие термина языковым нормам.
3. Терминологическая инвариантность. Отсутствие разнообразия в написании и произношении термина [21, 23].
4. Мотивированность или самообъяснимость термина. «Максимальное соответствие структуры термина содержательной структуре выражаемого им понятия» [23].

**Семантические признаки**

К данной группе относятся признаки, обусловленные содержанием термина.

1. Системность. Принадлежность термина к системе понятий определенной предметной области.
2. Соответствие обозначаемому понятию. Отсутствие противоречий между значением термина в определенной предметной области и лексическим значением слов, из которых состоит термин.
3. Однозначность или моносемантичность термина. Однозначность термина в рамках терминологии определенной предметной области. В разных сферах употребления термин может иметь разные значения.
4. Содержательная точность. Точность и ограниченность значения термина.

**Прагматические признаки**

К данной группе относятся признаки, обусловленные спецификой функционирования термина.

1. Внедрённость или общепринятость.
2. Дефиницированность. Наличие у термина дефиниции, понятия - определения или толкования слова; «установление смысла незнакомого понятия с помощью знакомых и уже осмысленных понятий, или путем включения в контекст знакомых слов» [18].
3. Независимость от контекста. Признак, являющийся следствием моносемантичности термина.
4. Вариационная устойчивость. Воспроизводимость в текстах рассматриваемой предметной области как слов, так и словосочетаний, образующих термин данной терминологии. Другими словами этот признак обозначает высокую встречаемость термина в текстах рассматриваемой предметной области.
5. Благозвучность.

Таким образом, нами были рассмотрены признаки термина, которые отличают термин от слов и словосочетаний, употребляемых в текстах, ими не являющихся. На основе вышеописанных особенностей и построены алгоритмы решения задачи построения перечня ключевых слов из текста (извлечение терминов из текста). Перейдем к описанию существующих обзоров методов автоматического извлечения терминов из текста, рассмотрим критерии оценки работы методов, используемые авторами, а затем перейдем к описанию самих методов.

## 1.2. Существующие обзоры методов извлечения терминов из текста

Проблемой извлечения терминов из текста еще в 1996 года занимались Кагеура К. и Унимо В., которые в своей статье, посвященной задаче извлечения ключевых слов и терминов из текста и обзору существующих на тот момент методов решения данной задачи, подробно анализировали подходы к решению не только задачи получения перечня терминов из текста, но и автоматического индексирования текстовых документов [24]. Большая часть данного исследования посвящена методам на основе мер ассоциаций, а также методам, являющимся развитием TF-IDF. Для оценки результатов данные авторы предлагают два новых аспекта понятия «термин»: «соединенность» (англ. unithood) — параметр, отражающий характеристику связи слов в многословных терминах; и «терминологичность» (англ. termhood) — степень близости термина к рассматриваемой предметной области. Кроме того, в данной работе выделяются два класса методов извлечения терминов из текста: лингвистические и статистические.

В исследовании других авторов отмечается, что современные работы рассматривают лингвистические методы выделения ключевых слов в качестве набора фильтров и явно не проводят разделение на вышеуказанные классы [25]. Также в этом обзоре делается упор на рассмотрении таких ассоциативных мер, как отношение функций правдоподобия, t-тест, χ2 - тест, z-тест, *MI*, *MI2* и некоторых других. Кроме того, в данной работе рассматриваются методы, пытающиеся определить принадлежность термина предметной области такие, как частота вхождений, C-Value, Co-Occurrence.

З. Чжан и др. в статье [26] провели обзор методов, поддерживающих извлечение как однословных, так и многословных терминов, а также их провели экспериментальное сравнение. В перечне указанными авторами методов - C-Value [2], TF-IDF [3], TermExtractor [5], Weirdness [7] и Glossex [32]. Авторы проводили исследования на текстах двух предметных областей: биомедицины и зоологии. В результате исследования они отмечают, что, несмотря на относительную близость двух выбранных предметных областей, результаты работы методов различаются в зависимости от наборов данных. Кроме того, в этом обзоре показывается превосходство алгоритма голосования как метода комбинации отдельных признаков.

П.И. Браславский и Е.А. Соколов в своей работе [27] сравнили четыре метода извлечения терминов, описываемые словосочетаниями из двух слов. Рассматриваемые методы используют следующие ассоциативные метрики: частоту вхождений, t-тест, χ2 - тест и отношение функций правдоподобия. Авторы отмечают, что лучшие результаты показывают методы, использующие параметры частоты вхождений и t-тест, а также выделяют основной тип ошибок, встречающийся в результирующем списке терминов: «выделение устойчивых общеупотребительных словосочетаний, удовлетворяющих шаблонам».

В другой, более поздней, работе П.И. Браславский и Е.А. Соколов [28] сравнивают пять методов выделения терминов произвольной структуры: MaxLen [29], C-Value [2], k-factor [4], Window [30], именные группы, выделенные с помощью синтаксического анализатора АОТ [31]. В качестве оценки эффективности авторы использовают комбинацию экспертной и формальной оценок по предзаданному словарю, выступающему в роли «эталонного списка». Согласно полученным результатам, «сравниваемые методы дают в целом похожие результаты», однако, как отмечают авторы, наибольшую эффективностью показывают методы C-Value и k-factor, в то время как наименьшую - метод на основе синтаксического анализа. По результатам комбинированной оценки (экспертной и формальной) авторы делают вывод, что формальные методы оценки эффективности годятся для сравнения больших списков кандидатов в термины.

Таким образом, существующие в настоящее время исследования в области извлечения ключевых слов и терминов из текста предлагают различные методы решения данной задачи, однако, полученные результаты не отличаются высокой эффективностью, не обладают свойством универсальности, могут быть использованы только в узкой предметной области, и, следовательно, должны быть продолжены.

## 1.3 Алгоритм работы методов извлечения терминов из текста

При решении задачи извлечения из коллекции документов заданного количества терминов любой длины, не различая вхождения одного термина, можно выделить общий алгоритм её решения, который реализуется большей частью существующих статистических методов. Согласно этому алгоритму, работа метода извлечения терминов из текста выполняется в три этапа.

1. Сбор кандидатов - этап, включающий в себя фильтрацию слов и словосочетаний, извлеченных из коллекции текстовых документов, по статистическим и лингвистическим критериям.
2. Подсчет признаков - этап, в ходе которого производится перевод каждого терминологического кандидата в вектор пространства признаков.
3. Вывод на основе признаков - этап, в котором производится оценка вероятности быть термином для каждого кандидата на основе значений признаков. Также на данном этапе осуществляется последующая сортировка всех кандидатов по рассчитанной оценке и отбор заранее определенного числа кандидатов.

Методы, реализующие этап сбора кандидатов в термины, в свою очередь, состоят из следующих этапов [33]. На первом шаге применяются лингвистические фильтры. Лингвистический фильтр — это набор алгоритмов, который используется для ограничения типов слов и словосочетаний. В работе [2] описывается две категории лингвистических фильтров: закрытые и открытые фильтры.

Закрытые фильтры ограничивают тип слов в терминах-кандидатах определенной частью речи. Примером закрытого фильтра является Noun+, ограничение которого состоит в том, что в качестве терминов могут выступать только существительные и словосочетания существительных. Непосредственным результатом применения такого типа фильтров является повышение параметра точности выделения терминов и понижение параметра полноты, так как большое количество терминов представляют собой словосочетания, являющиеся комбинацией из слов разных частей речи (например, прилагательное и существительное). Более подробная информация о таких параметрах оценки эффективности методов автоматического извлечения терминов, как точность выделения терминов и полнота приводится в пункте 1.6 настоящей работы.

Открытые лингвистические фильтры накладывают иные ограничения (например, допускаются сочетания существительных и прилагательных). Благодаря использованию фильтров открытого типа, среди результирующих терминов допускаются именные группы. Однако данный тип фильтров имеет недостаток, выражающийся в снижении точности результирующего списка. В то же время применение открытых фильтров позволяется добиться увеличения параметра полноты.

Таким образом, цель работы лингвистического фильтра заключается в ограничении слов и словосочетаний, извлеченных из исходного текста, только существительными и именными группами, словосочетаниями с существительным в роли главного слова. Такие частеречные ограничения накладываются на терминологические кандидаты в соответствии с таким признаком термина, как номинативность. Для этого применяется или поверхностный синтаксический разбор (англ. shallow parsing, chunking) [34], или фильтрация N-грамм по предопределенным шаблонам частей речи (существительное, прилагательное + существительное) [2].

На последующих этапах работы методов сбора кандидатов производится дополнительная фильтрация с целью снижения шума. Фильтрация производится по следующему ряду критериев:

1. По частоте. Из перечня терминологических кандидатов исключаются слова и словосочетания с числом вхождений меньше 2 или 3. Фильтрация по этому критерию выполняется по причине неприменимости многих статистических признаков к таким кандидатам в термины;
2. По содержанию. Терминологические кандидаты проходят фильтрацию при помощи заранее составленного списка стоп-слов. Список стоп-слов или стоп-список (англ. stop list) — это перечень, содержащий слова, появление которых не является ожидаемым в списке терминов указанной предметной области [2, 36]. Назначение стоп-списков заключается в отсеивании слов и словосочетаний, которые, с большей долей вероятности, не являются терминами заданной предметной области. Многие слова, например, такие как прилагательные «добрый» или «интенсивный», очень редко входят в состав терминов, несмотря на то, что их частота встречаемости в исходном тексте может быть достаточно большой (например, «интенсивная артиллерийская подготовка»). Их применение позволяет добиться повышения точности выделения терминов. Однако стоит отметить, что в силу меняющегося характера ряда предметных областей, существует необходимость в периодическом пересмотре и обновлении стоп-списков из-за возникновения новых терминов [33].
3. По длине слов кандидата или содержанию в них особых символов [37]. Из перечня терминологических кандидатов часто исключаются символы, не относящиеся к алфавитным, а также слова, состоящие из одной буквы.

В пунктах 1.4 и 1.5 данной работы подробно рассматривается этап , вычисления значений признаков для кандидатов в термины.

В качестве модели представления исходных текстовых данных большая часть статистических методов извлечения терминов из текста использует модель Bag of words. Данная модель представляет текстовый документ в виде неупорядоченного множества пар: слово, взятое из документа, и частота встречаемости. Недостатком этой модели является отсутствие возможности учета порядка слов в тексте.

Разновидностью модели Bag of words является частотная модель текста, в которой каждому слову соответствует весовой коэффициент, который определяется в зависимости от выбранной метрики: например, частоты вхождения слова в документе TF (англ. Term frequency), логарифма частоты вхождения слова Log TF или обратной частоты документов IDF (англ. Inverse document frequency). Однако данная модель также, как и модель Bag of Words, не учитывает порядок слов и может быть использована для извлечения лишь однословных понятий из текста.

Для решения этой проблемы вводится модель коллокации N-грамма (англ. N-gram). Коллокация — словосочетание, являющееся синтаксически и семантически целостной единицей. Данная модель позволяет представить словосочетания из N слов, при N=2 словосочетание имеет название биграммы, при N=3 — триграммы. Таким образом, для извлечения многословных терминов возможно использование частотной модели текста вместе с моделью коллокаций N-грамма. Однако, данные модели не принимают во внимание грамматические особенности русского языка, с учётом которых возможно повышение качества результатов методов извлечения терминов из текста.

Как отмечает Н.А. Астраханцев, в исследованиях, посвященных методам извлечения терминов из текста часто взаимозаменяемо применяются два термина: «признак» и «метод» [11]. Существует разница в значениях данных понятий: «признак — это отображение кандидата в некоторое число, а метод — это последовательность действий, позволяющая получить ранжированный список кандидатов для заданной коллекции документов, которая включает в себя вычисление одного или нескольких признаков» [11]. Данные термины взаимозаменяемы, так как любой метод может рассматриваться в качестве признака, а большая часть признаков изначально создавалась в качестве отдельных методов.

Перейдем к рассмотрению методов решения задачи подсчета признаков терминологических кандидатов.

## 1.4. Методы на основе статистического параметра частоты вхождений

Рассмотрим методы, учитывающие частоту вхождений кандидатов в коллекции документов и вложенность кандидатов.

Самым первым методом можно считать частоту вхождений кандидата в термины — TF. Данный метод подсчитывает количество упоминаний кандидата в исходном тексте.

К ранним признакам можно отнести и метод, появившийся изначально в области информационного поиска, — TF-IDF (англ. аббр. Term frequency - Inverse document frequency).

(1.1)

где *t* — терминологический кандидат,

*TFd (t)* — количество текстовых документов, в которых встретился кандидат *t*.

Этот метод присваивает высокие значения признака кандидатам, часто встречающихся лишь в малом числе документов. Одним из первых исследований применимости указанного метода для решения задачи извлечения термина является работа Д. Эванса и Р. Лефферетса [3].

В качестве признака, используемого для методов решения задачи подсчета признаков терминологических кандидатов, также используется Domain Consensus [1]. Данная метрика создавалась для распознавания терминов, равномерно распределенных по всей коллекции документов:

(1.2)

где *d* — документ,

*D* — множество исходных текстовых документов.

В отдельную категорию «меры ассоциации» (англ. word association measures) выделяют признаки, оценивающие параметр синтагматичности. Синтагматичность или [синтагматическая целостность](http://www.multitran.ru/c/m.exe?t=4795089_2_1&s1=unithood) (англ. unithood) — это мера, отражающая степень взаимосвязи между собой слов в термине-словосочетании [32]. Иными словами, данный признак оценивает насколько сильно связаны слова в составе термина. Данная категория признаков может быть применена только к многословным терминам, а в ряде случаев только к двухсловным терминам. К этой группе относятся следующие методы: z-тест [38], t-тест [39], отношение функций правдоподобия, χ2 — тест [40], взаимная информация (англ. Mutual Information, *MI*) [41], MI2 , MI3 [42], Term Cohesion [31], Lexical Cohesion [43].

К методам, созданным на основе статистики вхождений, и учитывающим параметр синтагматичности, относится также признак C-Value [2]. Метрика C-Value учитывает длину и вложенность терминологического кандидата. Вложенные термины (англ. nested terms) - это понятия, содержащиеся в исходном тексте как по отдельности, так и в составе других понятий [44].

(1.3)

где *t* — кандидат в термины,

|*t*| — длина кандидата *t*, выраженная в количестве слов,

*f(t)* — частота встречаемости *t* в коллекции текстов,

*s* — множество кандидатов, включающих *t*, то есть таких кандидатов, что t является их подстрокой.

В данном признаке вес кандидата уменьшается, если он является частью других кандидатов. Данная особенность учитывает тот факт, что в этом случае частота встречаемости терминологического кандидата суммируется с частотой вхождения кандидатов, его включающих. Так, например, словосочетание point arithmetic (арифметические операции с точкой) имеет не меньшую частоту вхождений, чем термин floating point arithmetic (арифметические операции с плавающей точкой). Однако, термином из двух предложенных вариантов является floating point arithmetic.

Необходимо отметить, что признак C-Value предназначен для извлечения только многословных терминов: для однословных терминов значение метрики будет равным нулю.

Развитием данного метода, учитывающего указанный недостаток, является работа Баррона-Кедено и др. [36]. Признак C-Value в данном исследовании обобщается на случай однословных терминов путем добавления константы *i* к логарифму (см. формулу 1.4). В ходе исследований авторы пришли к выводу, что наибольшая эффективность метода C-Value достигается при *i*=1.

(1.4)

где *i* — константа,

Другой модификацией метода C-Value, решающей недостаток невозможности извлечения однословных терминов, является работа Лоссио-Вентура и др. [35]. Признак C-Value в данном исследовании обобщается на случай однословных терминов путем добавления константы к логарифму.

(1.5)

Еще одной модификацией метода C-Value является работа Дж. Бордо и др. [45]. Предложенный метод Basic предназначен для извлечения терминов средней специфичности.

(1.6)

где *et*— количество кандидатов, содержащих кандидата *t*.

Метод Basic применим только для многословных терминов, также как и C-Value. Однако в отличие от C-Value, в котором учитываемая частота кандидата уменьшается, если он является вложенным, в данном методе значение признака увеличивается с ростом количества содержащих его терминологических кандидатов, поскольку среднеспецифичные термины часто служат для образования более специфичных терминов. Авторы работы приводят в качестве примера термин information retrieval (информационный поиск), который может использоваться для создания таких более специфичных терминов, как information retrieval system (система информационного поиска) и information retrieval metric (метрика информационного поиска).

Метод, описанный в работе [4], не имеет собственного названия. В исследовании Браславского и др. [27] указанный метод фигурирует под названием k-factor. Данный метод реализован в рамках информационной системы BootCaT, предназначенной для формирования текстового корпуса из Веба при помощи поисковой системы. Метод k-factor базируется на использовании статистической метрики частоты встречаемости строки в тексте и, также как и метод C-value, учитывает вложенность терминологических кандидатов. Метод функционирует следующим образом: если строка *a* является вложенной по отношению к *b* и выполняется неравенство (см. формулу 1.7), то из двух кандидатов в финальном перечне терминов метод оставляет *b*.

(1.7)

Метод GlossEx был создан с учетом особенностей, присущих документальной информационной базе данных технической предметной области и содержащейся в ней информации. Данный метод учитывает специфику неоднородной коллекции документов. Помимо частоты встречаемости кандидата в исходном текстовом корпусе, данный метод использует метрики частоты и вероятности слов, составляющих определенный терминологический кандидат. Каждой строке *T* присваивается оценка, определяемая по следующей формуле.

(1.8)

где *α, β* - константные значения, определяющие степень влияния *TD* и *TC* на итоговую оценку, *α + β = 1*;

*TD* — параметр доменной специфичности (см. формулу 1.9);

*TC* — параметр терминологической связанности (см. формулу 1.10).

Доменная специфичность или терминологичность — это параметр, отражающий степень отношения кандидата в термины к понятиям указанной предметной области [32]. Синтагматичность или [синтагматическая целостность](http://www.multitran.ru/c/m.exe?t=4795089_2_1&s1=unithood) - это мера, отражающая степень взаимосвязи между собой слов в термине-словосочетании [32].

(1.9)

(1.10)

где *T* - терминологический кандидат, состоящий из слов *Wi*,

|*T*| - длина *T*, выраженная в количестве слов,

*f*(˙) - частота встречаемости кандидата или слова,

- вероятность встречаемости слова *Wi* в документе определенной предметной области*,*

- вероятность встречаемости слова *Wi* во всех документах текстового корпуса.

## 1.5. Методы на основе контекстов вхождений

Рассмотрим методы, учитывающие контекст вхождений кандидатов в коллекции документов. Методы данной группы, в частности NC-Value [2], основаны на предположении, что контексты терминов и обычных слов отличаются. Авторы признака NC-Value подразумевают под контекстом существительные, глаголы или прилагательные, непосредственно предшествующие или следующие за вхождением термина.

Вычисление данного признака состоит из трех этапов. На первом этапе извлекаются 200 терминов с наибольшим значением статистической метрики. В качестве признака авторы использовали C-Value, однако, вместо него возможно использование любого другого метода.

На втором этапе вычисляются веса для слов контекста по формуле:

(1.11)

где *w* - слово контекста (существительное, глагол или прилагательное);

*t(w)* - количество терминов, в контексте которых встретилось *w*;

*n* - общее количество рассматриваемых терминов.

На третьем этапе вычисляется финальное значение по формуле:

(1.12)

где *t* - кандидат в термины;

*Ct -* множество слов, встречающихся в контексте кандидата *t*;

*w* - слово из *Ct*;

*ft(w)* - частота, с которой слово *w* встречается в контексте кандидата *t*.

Модификацией метода NC-Value для случая извлечения среднеспецифичных терминов является метод Domain Coherence, предложенный в работе Дж. Бордо и др. [45]. Авторы вводят следующие ограничения на контекстные слова, называемые «моделью домена»:

1. слова должны присутствовать не менее чем в четверть документов входной коллекции текстовых документов;
2. слова должны быть существительными, глаголами или прилагательными;
3. слова должны быть семантически близки ко многим специфичным терминам предметной области.

В качестве входных данных метод использует список кандидатов, полученный с помощью метода Basic. В отличие от подсчета соотношения терминов, перед которыми или после которых встретилось слово, применяемого в NC-Value, в методе Domain Coherence предлагается использовать метрику *PMI* (см. формулу 1.13).

(1.13)

где *w* - слово, рассматриваемое в качестве кандидата в модель домена;

*T* - множество 200 лучших терминов, извлеченных с помощью метода Basic;

*P(t, w)* - вероятность появления слова w в контексте термина t;

*P(t)* и *P(w)* - вероятности появления термина *t* и слова *w*, соответственно. Указанные вероятности оцениваются на основе частот вхождения во входной коллекции документов; в качестве контекста рассматривается окно в 5 слов.

Для вычисления финального значения признака Domain Coherence также применяется метрика *PMI*, вычисляемая между каждым кандидатом в термины *t* и словом из модели домена *w*.

Перейдем к описанию методов оценки эффективности работы методов автоматического извлечения терминов из текста.

## 1.6. Методы оценки эффективности

Вопрос об оценке эффективности систем извлечения терминов остается нерешенным. Как отмечают Г. Бернье-Колборн и П. Дроин в своем исследовании [46] оценки эффективности регулярно публикуются в работах, посвященных данной тематике, однако методология отличается от работы к работе, затрудняя таким образом какое-либо сравнение.

Возможно выделить два принципиальных подхода к оценке эффективности методов извлечения терминов:

1) оценка результатов работы метода вручную с помощью экспертов предметной области;

2) использование заранее созданного списка терминов-эталонов. В работах П.И Браславского и Е.А. Соколова ([25, 27]) данный список именуется «формальной оценкой».

В то время как первый подход позволяет произвести наиболее точную оценку, второй подход позволяет обеспечить повторяемость результатов, в результате чего появляется возможность многократной настройки параметров без дополнительного привлечения труда эксперта предметной области. При втором подходе возможно также возможно осуществить сравнение разных методов на при одном наборе данных.

Существует несколько способов осуществить оценку эффективности, используя заранее созданный список терминов-эталонов. В зависимости от способа получения списка терминов-эталонов можно выделить следующие способы:

1. Разметка всех документов вручную (например, [33]);
2. Разметка небольшой части документов вручную (например, [47]);
3. Адаптация существующих ресурсов к задаче извлечения терминологии (например, [37]).

Первый способ является наиболее точным, но, в то же время, наиболее трудозатратным и времязатратным.

Второй способ позволяет уменьшить время выполнения процедуры предварительной разметки документов. С его помощью сохраняется возможность вычисления признаков на основе всей исходной коллекции документов. В то же время оценка эффективности работы методов извлечения терминов из текста проводится только для тех терминов, которые будут встречаться в размеченных документах.

Третий способ осуществим при наличии внешних ресурсов, таких как ранее созданные вручную тезаурусы или словари [27].

Перейдем к описанию существующих метрик эффективности результатов выполнения методов решения задачи извлечения заданного количества терминов любой длины из коллекции документов

1. Точность выделения терминов (англ. precision) — это параметр, показывающий долю терминов, относящихся к заданной предметной области, в результирующем списке понятий в процентном отношении. Так, например, если при помощи алгоритма автоматического выделения терминов был получен список из 100 понятий и значение параметра точности составляет 2%, из этого можно сделать вывод, что только два понятия относятся к заданной предметной области [32].

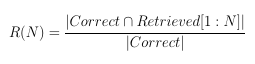
 (1.14)

где *N* — количество учитываемых лучших кандидатов;

*Correct* — множество терминов-эталонов;

*Retrieved*[*1* : *N*] — множество лучших *N* кандидатов в термины в соответствии с весами, назначенными оцениваемым методом.

1. Полнота выделения терминов (англ. recall) — это параметр, показывающий процентное соотношение извлеченных терминов, относящихся к заданной предметной области, к множеству всех терминов заданной предметной области. То есть, показатель полноты выделения терминов в 2% означает, что в результирующем списке понятий содержится 2% от общего множества понятий заданной предметной области [32].

 (1.15)

1. Средняя точность (англ. average precision):

 (1.16)

Стоит отметить, что параметры полноты и точности имеют обратно пропорциональную зависимость: высокий показатель точности выделения терминов может быть достигнут при более низком показателе полноты и наоборот. Астраханцев Н.А. в работе [11] отмечает, что наиболее широко используемой метрикой оценки эффективности работы методов извлечения терминов из текста является средняя точность, так как данный показатель является интегральной оценкой по множеству значений N .

## 1.7. Выводы

В данной главе был проведен обзор существующих определений базовых понятий для задачи извлечения терминов, определена cодержательная постановка задачи автоматического извлечения терминов из текста, проанализированы существующие методы ее решения.

Необходимо выделить несколько основных проблем, связанных с рассматриваемой задачей:

1. Отсутствие общепринятых определений термина. На настоящий момент времени существует множество несогласованных, а иногда и вовсе противоречивых определений. Также стоит отметить неформальность множества определений понятию «термин».
2. Отсутствие формальной постановки задачи извлечения терминов.
3. Как следствие из пункта 2 — сложность оценки эффективности и сравнения разработанных методов в связи с отсутствием общепринятых наборов данных и методологии оценки эффективности.
4. Зависимость многих существующих методов не только от предметной области исходных текстовых данных, но и от выбранного языка. Данная проблема затрудняет или делает невозможным использование одним и тех же методов извлечения терминов из текста на текстовых документах других предметных областей и различных языков.

В следующей главе будет описана модель русскоязычного текстового документа, учитывающая особенности грамматики русского языка, которую можно использовать для извлечения ключевых слов и терминов из текста с помощью статистических методов.

# 2. Математическая модель русскоязычного текстового документа, предназначенная для извлечения терминов из текста

Проведенный анализ методов извлечения ключевых слов и терминов из текста, а также математических моделей, используемых существующими методами, показал следующее. Модель текста Bag of Words, частотная модель текста и модель N-gram позволяют работать не только с текстовыми документами не только на любом языке, но и с многоязычными документами. Однако несмотря на простоту данных моделей, в контексте задачи обработки моноязычного текста на русском языке они не принимают во внимание его грамматические особенности, с учётом которых возможно повышение качества результатов методов извлечения терминов из текста.

Во второй главе приведено разработанное автором данной работы описание математической модели русскоязычного текстового документа, учитывающей особенности грамматики русского языка, а также приведены способы использования данной математической модели для извлечения ключевых слов и терминов из текста с помощью статистических методов.

## 2.1. Математическая модель русскоязычного текстового документа

Рассмотрим математическую модель русскоязычного текстового документа. Пусть дан текстовый документ определенной предметной области *D,* состоящий из множества предложений *si* (см. формулу 2.1).

*D* = {*s1,…,* *si,…,sn* }, (2.1)

где *si* - предложение

Предложение *si* состоит из множества слов *wij*, разделенных между собой символами-разделителями слов *sep'*ib′, а также из символов конца предложения sepik (см. формулу 2.2).

*,* (2.2)

где *i* - номерпредложения,

n - количество предложений в текстовом документе *D*

k′- номер символа конца предложения,

k - количество символов конца предложения si

b′ - номер символа-разделителя слова,

b - количество символов-разделителей слов wij в предложении si

*wij* - слово, *wij* ∈ W

Опишем слово *wij* предложения *si* как множество, состоящее из символов русского алфавита и символа дефиса (см. формулу 2.3).

, (2.3)

где j - номер слова в предложении si,

m - количество слов в предложении si

lija′ - символ в слове wij,

*L* – множество символов кириллического алфавита, а также символ дефиса (‘-’)

a′ - номер буквы в слове wij,

a - количество символов в слове wij

sepik′ - символ окончания предложения si,

sep′ib′ - разделитель слов wij и wij+1 в предложении si,

Sep - множество символов окончания предложений. Значения данного множества представлены в формуле 2.4.

Sep = {’.’, ’!’, ’?’} (2.4)

Sep′ - множество символов-разделителей слов в предложении. Значения данного множества представлены в формуле 2.5.

Sep′ = {’ ’, ’,’, ’;’, ’-’, ’:’} (2.5)

Каждое слово обладает морфологической парадигмой *Wf* — системой словоформ, образующих одну лексему [17]. Словоформа – это слово в некоторой грамматической форме [48]. Лексема — это единица словаря языка, объединяющая разные формы одного слова (например, метод, метода, методу, методы и т.д.), а также разные смысловые варианты слова, зависящие от контекста. Таким образом, *Wf* - это множество словоформ одного слова.

, (2.6)

где *wfe'* – словоформа,

e - количество словоформ одной парадигмы

e′ - номер словоформы, .

Опишем множество частей речи русского языка Pos. Элементами данного множества являются:

Pos = (’Существительное’, ’Местоимение’, ’Прилагательное’, ’Предлог’, ’Частица’, ’Союз’, ’Наречие’, ’Числительное’, ’Причастие’, ’Деепричастие’, ’Глагол’).

Каждая часть речи характеризуется особой системой грамматических категорий [48]. Таким образом, в зависимости от части речи словоформа *wfe'* может быть описана по-разному. Пусть дано множество *Pos'*, Pos′ ∈ *Pos*, элементами которого являются именные части речи, а также причастие.

Pos′ = (’Существительное’, ’Местоимение’, ’Прилагательное’, ’Причастие’, ’Числительное’).

Для именных частей речи, а также для такой отглагольной части речи, как причастие, словоформа может быть описана следующим образом (см. формулу 2.7).

для

, (2.7)

где w - слово, w ∈ W,

case - падеж, case ∈ Case,

n - число, n ∈ N

*g* – род, g ∈ G

Case = (’именительный’, ’родительный’, ’дательный’, ’винительный’, ’творительный’, ’предложный’)

N = (’единственное’, ’множественное’)

G = (‘мужской’, ’средний’, ’женский’)

Глагольная словоформа может быть описана следующим образом (см. формулу 2.8).

для ,

, (2.8)

где w – слово,

*t –* время, t ∈ T

n - число, n ∈ N

*p* – лицо, p ∈ *P*

T = {‘прошедшее’, ‘настоящее’, ‘будущее’}

P = {‘первое’, ‘второе’, ‘третье’}

Глаголы в форме прошедшего времени характеризуются, помимо грамматических категорий времени, числа и лица, также и такой категорией, как род *g*, g ∈ *G* (например, читал, читала). Таким образом, глагольные словоформы прошедшего времени могут быть описаны следующим образом (см. формулу 2.9).

для

(2.9)

В русском языке деепричастие, наречие, предлог, частица и союз являются неизменяемыми частями речи. Поэтому морфологической парадигмой *Wf* слов, принадлежащих к данным частям речи, является само слово *w*. Таким образом, *Wf =* {*w*}.

У каждого слова w имеется лемма wfnf. Лемма - это исходная, базовая или нормальная форма слова, зафиксированная в словаре [50]. Для именных частей речи, а также для причастия, нормальной является форма именительного падежа единственного числа. Лемма является одной из словоформ слова w. Таким образом,

, (2.10)

где case′ = ′именительный′,

n′ = ′единственное′

Нормальной формой глагола является инфинитив. В инфинитиве морфологически не выражены такие категории глагола, как время, лицо и число [48].

Перейдем к описанию использованных функций. Обозначим DPos функцией определения части речи слова (см. формулу 2.10), *DNF* - функцией нормализации слова (см. формулу 2.11). Нормализация слова или лемматизация - это процесс приведения слова к лемме или начальной форме [50].

(2.11)

(2.12)

Перейдем к описанию возможности применения описанной математической модели русскоязычного текстового документа для анализа текста.

## 2.2. Применение математической модели русскоязычного текстового документа для извлечения терминов из текста

Применительно к рассматриваемой задаче данную модель можно использовать для описания текстового корпуса. Текстовый корпус представляет собой сформированную по определенным правилам, структурированную коллекцию текстов [51, 52]. Одно из главных отличий корпуса текстов от коллекции текстов заключается в аннотированности корпуса. Суть разметки или аннотирования состоит в присвоении текстам корпуса и их компонентам дополнительной информации, называемой метаданными или тегами. В частности, описываемая в данной работе математическая модель учитывает лингвистическую информацию о словах. Другой немаловажной особенностью, характеризующей текстовые корпусы, является свойство репрезентативности [53]. В то время как текстовые коллекции могут объединять в себе семантически малосвязанные документы, корпусы текстов создаются на основе документов, релевантных для целей создания корпуса. Например, текстовый корпус статей о термодинамике не должен включать в себе информацию о биографиях авторов данных статей.

Перейдем к описанию понятия терминологического кандидата. Кандидат в термины — это слово или словосочетание, удовлетворяющее заданным критериям и потенциально являющееся термином определенной предметной области. Пусть P - множество терминологических кандидатов. Элементами данного множества являются слова w и словосочетания p.

 (2.13)

где o - количество кандидатов-слов,

r - количество кандидатов-словосочетаний,

 (2.14)

где u - номер словосочетания, 

c - номер слова в словосочетании pu,

z - максимальная длина словосочетания;

эмпирически установлено, что z ≤ 5

Большинство методов решения задачи автоматического извлечения терминов включают два этапа. На первом этапе производится извлечение кандидатов в термины из текстового корпуса. Опишем функцию DP, выполняющую извлечение перечня терминологических кандидатов P из текстового документа D.

 (2.15)

На втором этапе путем фильтрации и ранжирования списка, полученного на предыдущем этапе с помощью функции DP, формируется результирующий список терминов. На втором этапе применяются, в частности, статистические методы и методы машинного обучения. Суть статистического метода состоит в подсчете метрики M с помощью функции нахождения значения статистической метрики Fm.

M - статистическая метрика, M ∈ R,M ≥ 0

 (2.16)

Описав все необходимые понятия и функции, перейдем к формальной постановке задачи извлечения терминов из текста.

Дано:

*TD* - текстовый корпус

 (2.17)

где - количество текстовых документов в текстовом корпусе

Найти:

*T* - множество пар: терминологические кандидаты и их значения метрик

 (2.18)

где v′ - номер кандидата в текстовом документе, v′ = 1,v

v - количество терминов в текстовом документе

Множество T проранжировано в порядке убывания значений m. Таким образом, чем ближе к началу в итоговом множестве T находится кандидат p, тем с большей долей вероятности он является термином заданной предметной области.

## 2.3. Выводы

В данной главе было приведено описание математической модели русскоязычного текстового документа, учитывающей особенности грамматики русского языка. Данную модель можно использовать для описания текстового корпуса, который будет являться входными данными для методов извлечения ключевых слов и терминов из текста. Также модель учитывает лингвистическую информацию, необходимую для работы лингвистического фильтра на этапе сбора терминологических кандидатов.

В следующей главе будет описан процесс разработки комплекса проблемно-ориентированных программ, реализующих предложенную модель русскоязычного текстового документа, а также ряд существующих статистических методов извлечения терминов из текста.

# 3. Разработка комплекса проблемно-ориентированных программ для извлечения терминов из текста

В третьей главе рассматривается процесс разработки комплекса проблемно-ориентированных программ, реализующих функционал извлечения терминов из коллекции текстовых документов.

## 3.1. Анализ требований и проектирование

Для целей исследований применимости статистических методов извлечения терминов из русскоязычных текстов была сформирована коллекция из военно-исторических документов 1941-1945 гг., опубликованных в рамках двенадцати выпусков “Сборника боевых документов Великой Отечественной Войны”, выпущенных в период с 1947 по 1950 гг. [55]. Текстовые документы представлены в виде документов формата HyperText Markup Language (html), Portable Document Format (PDF) и в виде обычного текста (англ. plain text).

Для разрабатываемого программного обеспечения, реализующего предложенную в предыдущей главе математическую модель текстового документа и существующие статистические методы извлечения терминов из текста, были сформулированы следующие базовые функциональные требования для разрабатываемого программного обеспечения:

* чтение текстовых документов в форматах txt, html, pdf;
* выделение предложений в каждом из документов;
* морфологический разбор токенов в предложениях, хранение грамматических характеристик;
* реализация лингвистических фильтров Noun+  и (Adj|Noun)+Noun;
* реализация механизма фильтрации с помощью стоп-списка;
* реализация статистических методов C-Value, GlossEx, k-factor;
* сохранение промежуточной информации после проведения лексического и морфологического анализа в виде файла(ов);
* сохранение промежуточной информации после обработки терминологических кандидатов с помощью лингвистических фильтров в виде файла(ов);
* сохранение результирующего списка терминов в виде файла(ов).

Требуемый функционал рациональнее было разделить на модули, что позволило следовать принципу единственности ответственности (англ. SRP, Single Responsibility Principle) и упростило последующее тестирование. Таким образом, была спроектирована следующая структура ПО:

* Модуль Structures. Содержит классы, реализующие математическую модель русскоязычного текстового документа:
  + Case — падеж,
  + PartOfSpeech — часть речь,
  + TaggedWord — слово с грамматическими свойствама,
  + Separator — символ-разделитель.
  + Collocation — коллокация.
* Модуль TextImporter. Содержит классы для распознавания текста из исходных файлов формата pdf, html и txt. В случае, если в качестве входного файла будет передан неоцифрованный документ pdf, данный модуль должен произвести распознавание символов (англ. optical character recognition, OCR).
* Модуль LinguisticFilter. Содержит классы для реализации общего механизма фильтрации с помощью лингвистических фильтров, структур шаблонов, применяемых лингвистическими фильтрами, а также непосредственные реализации фильтров Noun+ и (Adj|Noun)+Noun.
* Модуль Stoplist. Содержит классы для реализации механизма фильтрации с помощью стоп-списка.
* Модуль Morph. Содержит классы для морфологического анализа слов.
* Модуль SentenceFetcher. Содержит класс для разбиения текста на предложения.
* Модуль StatMethods. Содержит классы, реализующие статистические методы C-Value, GlossEx, k-factor.

## 3.2. Разработка

Для разработки комплекса проблемно-ориентированных программ, реализующих функционал извлечения терминов из коллекции текстовых документов было принято решение использовать IDE PyCharm 2017.1 для Linux. Языком программирования для разработки был выбран Python версии 3.7.

Основная часть процесса разработки велась в соответствии с принципами разработки через тестирование (англ. Test Driven Development, TDD), когда для требуемого функционала и классов сначала пишутся автоматические модульные тесты, а уже потом программный код позволяющий их успешно выполнить. Это позволяет обеспечить простоту и структурированность создаваемого кода, избегая излишнего функционала.

Алгоритм, реализованный в разработанном программном обеспечении, выполняет следующую последовательность этапов:

1. Для каждого слова из исходного текстового корпуса проводится морфологический разбор.
2. Применяется лингвистический фильтр, предназначенный для ограничения типов слов и словосочетаний. Реализованы два лингвистических фильтра: Noun+ и (Adj|Noun)+Noun.
3. Полученный список кандидатов фильтруется с применением списка стоп-слов. Стоп-слова — это слова, появление которых является нежелательным в определенной предметной области.
4. Подсчитывается статистическая метрика. Итоговое множество слов и словосочетаний сортируются в порядке убывания значений метрики.

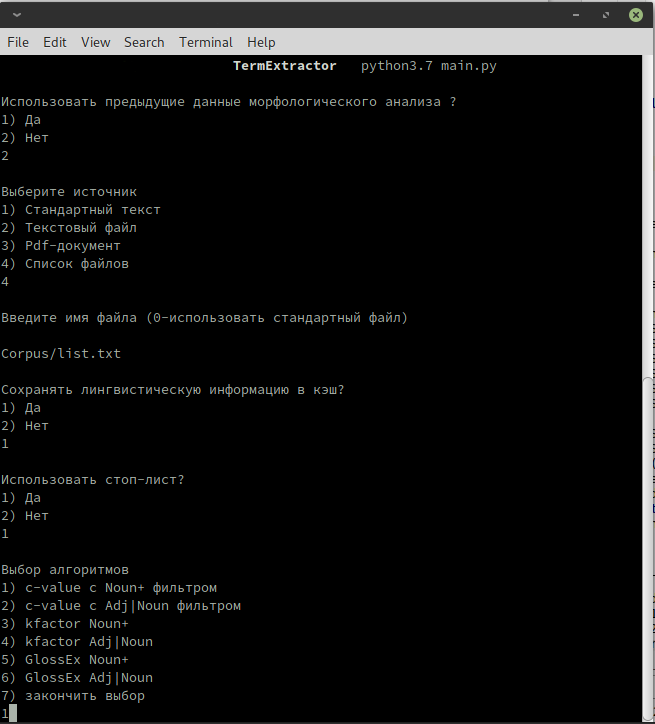


Рисунок 3.2.1 – Стартовый вид программы TermExtractor

Программное обеспечение реализовано с применением консольного интерфейса (см. рисунок 3.2.1). При запуске пользователю дается возможность выбора источника входных данных, возможность включения или отключения стоп-списка, возможность сохранения промежуточных результатов, выбор статистических алгоритмов с целью запуска ПО на обработку входных данных.

На рисунке 3.2.2 отображен фрагмент программного кода модуля LinguisticFilter, принимающего на вход предложение, состоящее из слов с тегами и символами-разделителями, а возвращающего перечень коллокаций, слов или словосочетаний, удовлетворяющих условию, заданному выбранным типом лингвистического фильтра.

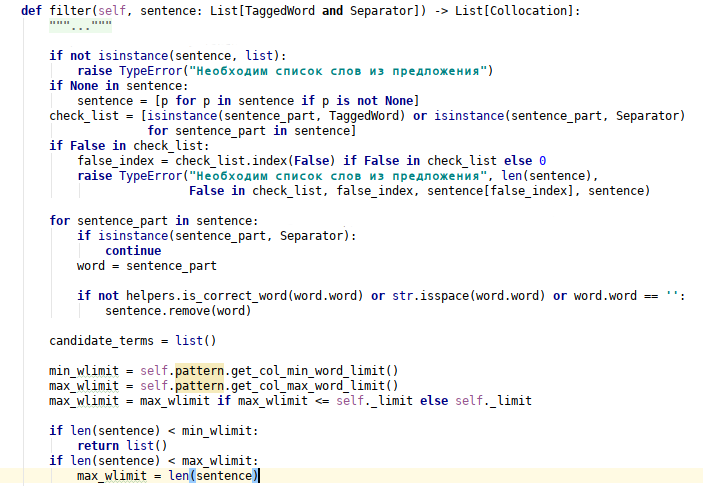


Рисунок 3.2.2 – Фрагмент программного кода реализации лингвистического фильтра

Входным параметром является исходная коллекция текстовых документов, которая может быть извлечена как из текстовых файлов, так и из pdf-документов, html-документов. Для выполнения первого этапа применяется библиотека pymorphy2 [47]. Исходный текст делится на предложения, каждое из которых преобразуется в массив слов с тегами и разделителями слов в предложении. Имеется возможность выбора типа лингвистического фильтра, а также методов подсчета статистической метрики для каждого кандидата. По мере выполнения программы формируется журнал событий, в котором отображается вся информация о ходе текущего выполнения программы. По окончании работы программы формируется ряд текстовых файлов, в каждом из которых записываются результаты выполнения одного выбранного метода.

## 3.3. Выводы

В данной главе был описан процесс разработки комплекса проблемно-ориентированных программ, реализующих функционал извлечения терминов из коллекции текстовых документов. Комплекс проблемно-ориентированных программ реализует предложенную математическую модель русскоязычного текстового документа, учитывающую особенности грамматики русского языка.

# ЗАКЛЮЧЕНИЕ

В процессе научно-практических исследовательских работ были поставлены и успешно решены следующие задачи:

1. Исследование существующих математических моделей и методов, применяемых при автоматическом извлечении терминов из текста.
2. Разработка математической модели русскоязычного текстового документа.
3. Создание комплекса проблемно-ориентированных программ, реализующего разработанную математическую модель и методы автоматического извлечения терминов из русскоязычного текста.

При решении первой задачи были изучены существующие определения, базовые понятия для задачи извлечения терминов, определена cодержательная постановка задачи автоматического извлечения терминов из текста, проанализированы существующие методы ее решения, описан общий алгоритм работы ряда существующих методов, проанализированы математические модели, использующиеся при извлечении терминов из текста.

При решении второй задачи была создана математическая модель русскоязычного текстового документа, учитывающая особенности грамматики русского языка. Данную модель можно использовать для описания текстового корпуса, который будет являться входными данными для методов извлечения ключевых слов и терминов из текста. Разработанная в ходе исследования модель также учитывает лингвистическую информацию, необходимую для работы лингвистического фильтра на этапе сбора терминологических кандидатов.

При решении третьей задачи был разработан комплекс проблемно-ориентированных программ, реализующих функционал извлечения терминов из коллекции текстовых документов. Созданное программное обеспечение реализует предложенную математическую модель русскоязычного текстового документа, учитывающую особенности грамматики русского языка.

Таким образом, поставленные в ходе исследования задачи были решены, цель работы достигнута. По основным результатам работы сделано 4 доклада на 4 международных, всероссийских и региональных конференциях (Саратов). Имеются следующие публикации по исследуемой научной тематике в периодических изданиях и материалах конференций: [33], [44], [49], [57], [58].

Полученные результаты будут использованы в последующих научных исследованиях при разработке методов и алгоритмов извлечения терминов из текста и автоматизированного построения онтологий из текста.

# СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Navigli R., Velardi P. Semantic interpretation of terminological strings // Proc. 6th Int’l Conf. Terminology and Knowledge Eng. 2002. P. 95–100.

2. Frantzi, K. Automatic recognition of multi-word terms: The C-value/NC-value method / K. Frantzi, S. Ananiadou, H. Mima // International Journal on Digital Libraries. — 2000. — Т. 3, No 2. — С. 115—130.

3. Evans D. A. CLARIT-TREC experiments / D. A. Evans, R. G. Lefferts // Information processing & management. 1995. Vol. 31, №. 3. P. 385–395.

4. Baroni, M. BootCaT: Bootstrapping corpora and terms from the web / M. Baroni, S. Bernardini // Proceedings of LREC. — 2004. — Т. 4. — С. 1313—1316.

5. Sclano, F. TermExtractor: a Web Application to Learn the Common Terminology of Interest Groups and Research Communities / F. Sclano, P. Velardi // Proceedings of the 9th Conference on Terminology and Artificial Intelligence (TIA 2007). — 2007.

6. Park, Y. Automatic glossary extraction / Y. Park, R. J. Byrd, B. K. Boguraev // Proceedings of the 19th international conference on Computational linguistics -. — 2002. — Т. 1. — С. 1—7.

7. University of surrey participation in trec8: Weirdness indexing for logical document extrapolation and retrieval (wilder) / K. Ahmad, L. Gillam, L. Tostevin et al. // The Eighth Text REtrieval Conference (TREC-8). 1999.

8. Meijer Kevin, Frasincar Flavius, Hogenboom Frederik. A semantic approach for extracting domain taxonomies from text // Decision Support Systems. 2014. Т. 62. С. 78–93.

9. Браславский П.И., Соколов Е.А. Автоматическое извлечение терминологии с использованием поисковых машин Интернета // Компьютерная лингвистика и интеллектуальные технологии. Тр. Международной конференции «Диалог». 2007. С. 89-94.

10. Голомазов Д. Д. Методы и средства управления научной информацией с использованием онтологий // Диссертация кандидата физико-математических наук. Москва. 2012.

11. Астраханцев, Н. А. Методы и программные средства извлечения терминов из коллекции текстовых документов предметной области // Диссертация кандидата физико-математических наук. Москва. 2014.

12. Vivaldi J., Rodrı́guez H. Using Wikipedia for term extraction in the biomedical domain: first experiences // Procesamiento del Lenguaje Natural. 2010. Vol. 45. P. 251–254.

13. Татаринов В.А. Терминологическая лексика русского языка: Эволюция проблем и аспектов изучения // Русский язык в современном обществе: Функциональные и статусные характеристики / РАН. ИНИОН; Отв. ред. Опарина Е.О., Казак Е.А. Теория и история языкознания. ИНИОН РАН, Москва, 2006. С. 133–164.

14. Гринев-Гриневич, С. В. Терминоведение // М.: Академия. 2008. Т. 304.

15. Мякшин К.А. Разнообразие подходов к определению понятия «термин» // Альманах современной науки и образования, сер. «Языкознание и литературоведение в синхронии и диахронии и методика преподавания языка и литературы». 2007. Т. 3, № 3. С. 175–178.

16. Реформатский, А. А. Что такое термин и терминология [Текст] / М.: Академия наук СССР. Институт языкознания, 1959. 16 с.

17. Винокур Г.О. Грамматические наблюдения в области технической терминологии // Труды МИИФЛИ. 1939. Т. 5.

18. Жеребило Т.В. Словарь лингвистических терминов. // Назрань: Пилигрим, 2010.

19. Wüster E. Einführung in die allgemeine Terminologielehre und terminologische Lexikographie (1979) // København: Handelshøjskolen. 1985.

20. ISO 1087-1:2000. Terminology work — Vocabulary — Part 1: Theory and application, URL: https://www.iso.org/obp/ui/#iso:std:iso:1087:-1:ed-1:v1:en (дата обращения 20.05.2019).

21. Хаютин А. Д. Составные термины - функциональный тип сложных лингвистических единиц (СЛЕ) с позиций лексикографии // Отраслевая терминология и лексикография. Воронеж, 1981.

22. Ахманова О. С. Терминология лингвистическая // Лингвистический энциклопедический словарь. Москва, 1990.

23. Мякшин К.А. К вопросу об основных признаках термина // Альманах современной науки и образования, сер. «Языкознание и литературоведение в синхронии и диахронии и методика преподавания языка и литературы». 2008. Т. 2, № 21. С. 17–22.

24. Kageura K., Umino B. Methods of automatic term recognition: A review // Terminology. 1996. Vol. 3, no. 2. P. 259–289.

25. Pazienza M., Pennacchiotti M., Zanzotto F. Terminology extraction: an analysis of linguistic and statistical approaches // Knowledge Mining. 2005. P. 255–279.

26. Zhang Z. A comparative evaluation of term recognition algorithms / Z. Zhang, J. Iria, C. Brewster, F. Ciravegna // Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008), Marrakech, Morocco. – 2008. – С. 2108–2113.

27. Браславский П. И. Сравнение четырех методов автоматического извлечения двухсловных терминов из текста / П.И. Браславский, Е.А. Соколов // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции «Диалог». М.: Издательство РГГУ, 2006. - С. 88–94.

28. Браславский П. И. Сравнение пяти методов извлечения терминов произвольной длины / П.И. Браславский, Е.А. Соколов // Компьютерная лингвистика и интеллектуальные технологии: Труды Международной конференции Диалог (Бекасово, 4–8 июня 2008). М.: Издательствово РГГУ, 2008. - № 7 – С. 67-74.

29. Bourigault D. Surface grammatical analysis for the extraction of terminological noun phrases // Proceedings of the 14th conference on Computational linguistics-Volume 3 / Association for Computational Linguistics. 1992. P. 977–981.

30. Добров Б.В. Формирование базы терминологических сочетаний по текстам предметной области / Б.В. Добров, Н.В. Лукашевич, С.В. Сыромятников // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды пятой Всероссийской научной конференции. 2003. С. 201–210.

31. Синтаксический анализ. Проект АОТ: Tech. Rep.: URL: http://www.aot.ru/docs/synan.html.

32. Kozakov L. Glossary extraction and utilization in the information search and delivery system for IBM Technical Support / L. Kozakov, Y. Park, T. Fin, Y. Drissi, Y. Doganata, T. Cofino // IBM Systems Journal. 2004. Vol. 43, № 3. P. 546–563.

33. Шульга Т.Э. О задаче автоматического извлечения терминов из текста / Т.Э. Шульга, А.С. Петров // Информационно-коммуникационные технологии в науке, производстве и образовании ICIT-2016: материалы Международной научно-практической конференции, Саратов, 23-28 августа 2016 г. - 2016. - С. 112-117.

34. Judea A. Unsupervised Training Set Generation for Automatic Acquisition of Technical Terminology in Patents / A. Judea , H. Schütze, S. Bruegmann // Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, 2014. August. P. 290–300.

35. Lossio-Ventura J.A. Combining C-value and Keyword Extraction Methods for Biomedical Terms Extraction / J. A. Lossio-Ventura, C. Jonquet, M. Roche, M.Teisseire // LBM’2013: International Symposium on Languages in Biology and Medicine. 2013. P. 45–49.

36. Barrón-Cedeño A. An improved automatic term recognition method for Spanish / A. Barrón-Cedeño, G. Sierra, P. Drouin, S. Ananiadou // Computational Linguistics and Intelligent Text Processing. 10th International Conference, CICLing 2009, Mexico City, Mexico, March 1-7, 2009. Proceedings. Springer, 2009. P. 125–136.

37. Bordea G. Domain adaptive extraction of topical hierarchies for Expertise Mining. Ph.D. thesis. 2013.

38. Dennis S. F. The construction of a thesaurus automatically from a sample of text // Proceedings of the Symposium on Statistical Association Methods For Mechanized Documentation, Washington, DC. 1965. P. 61–148.

39. Church K.W. Using Statistics in Lexical Analysis / K.W. Church, W.A. Gale, P. Hanks, D. Hindle // Lexical acquisition: exploiting on-line resources to build a lexicon. 1991. P. 115.

40. Dunning T. Accurate methods for the statistics of surprise and coincidence / T. Dunning // Computational linguistics. 1993. Vol. 19, no. 1. P. 61–74.

41. Church K. W., Hanks P. Word association norms, mutual information, and lexicography / K.W. Church, P. Hanks // Computational linguistics. 1990. Vol. 16, №. 1. P. 22–29.

42. Daille B. Combined approach for terminology extraction: lexical statistics and linguistic filtering. Ph.D. thesis: Ph. D. thesis, University Paris 7. 1994.

43. Park Y., Byrd R., Boguraev B. Automatic glossary extraction: beyond terminology identification // Proceedings of the 19th international conference on Computational linguistics-Volume 1 / Association for Computational Linguistics. 2002. P. 1–7.

44. Петров А.С. О методах решения задачи автоматического извлечения терминов из русскоязычных текстовых документов / А.С. Петров, Т.Э. Шульга // Проблемы управления, обработки и передачи информации (УОПИ-2017): сборник трудов V Международной юбилейной  научной конференции 2017. - C. 334-340.

45. Bordea G. Domain-independent term extraction through domain modelling / G. Bordea, P. Buitelaar, T. Polajnar // the 10th International Conference on Terminology and Artificial Intelligence (TIA 2013), Paris, France / 10th International Conference on Terminology and Artificial Intelligence. 2013.

46. Bernier-Colborne G., Creating a test corpus for term extractors through term annotation / G. Bernier-Colborne, P. Drouin // Terminology. 2014. Vol. 20, № 1. P. 50–73.

47. Astrakhantsev N.. Automatic Enrichment of Informal Ontology by Analyzing a Domain-Specific Text Collection / N. Astrakhantsev, D. Fedorenko, D. Turdakov // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue”. 2014. Vol. 13. P. 29–42.

48. Лингвистический энциклопедический словарь / Под ред. В. Н. Ярцевой // М.: Советская энциклопедия, 1990.

49. Петров А.С. Математическая модель русскоязычного текстового документа для решения задачи автоматического извлечения терминов из текста / А.С. Петров, Т.Э. Шульга // Вестник Воронежского Государственного Университета. Серия: Системный анализ и информационные технологии. - 2017. - № 3. - С. 195-203.

50. Захаров В.П., Богданова С.Ю. Корпусная лингвистика: Учебник для студентов направления «Лингвистика» // СПб.: СПбГУ. РИО. Филологический факультет, 2013.

51. Николаев И.С., Митренина О.В., Ландо Т.М. Прикладная и компьютерная лингвистика / И.С. Николаев, О.В. Митренина, Т.М. Ландо, М.: URSS, 2016. - 320 с.

52. McEnery T., Xiao R., Tono Y. Corpus-based language studies: An advanced resource book // Taylor & Francis, 2006.

53. Leech G. The state of the art in corpus linguistics // English Corpus Linguistics: Linguistic Studies in Honour of Jan Svartvik. 1991.

54. Nation P., Waring R. Vocabulary size, text coverage, and word lists // Schmitt; McCarthy, Vocabulary: Description, Acquisition and Pedagogy. 1997

55. Cборник боевых документов Великой Отечественной Войны / ред. В.А. Небучинов, М.Н. Шарохин. - М.: Воениздат, 1947-1950. Т. 1-12.

56. Korobov, M. Morphological Analyzer and Generator for Russian and Ukrainian Languages / M. Korobov // Analysis of Images, Social Networks and Texts. Т. 542 / под ред. M. Y. Khachay [и др.]. — Springer International Publishing, 2015. — С. 320—332.

57. Петров А.С. Применение статистических методов для решения задачи автоматического извлечения терминов из русскоязычных текстовых документов / А.С. Петров, Т.Э. Шульга // Информационно-коммуникационные технологии в науке, производстве и образовании ICIT-2017: материалы Международной научной конференции, Саратов, 21-22 сентября 2017 г. - 2017. - С. 239-246.

58. Петров А.С. Обзор статистических методов извлечения терминов из текста / А.С. Петров // Проблемы управления в социально-экономических и технических системах: сборник научных статей - 2016. С. 163-167.