|  |  |
| --- | --- |
| Gerb-BMSTU_01 | **Министерство образования и науки Российской Федерации**  **Федеральное государственное бюджетное образовательное учреждение**  **высшего образования**  **«Московский государственный технический университет**  **имени Н.Э. Баумана**  **(национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |

ФАКУЛЬТЕТ \_Информатика и системы управления\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

КАФЕДРА Программное обеспечение ЭВМ и информационные технологии\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

***К ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЕ***

***НА ТЕМУ:***

***Метод определения признаков авторского стиля***

***для текстов на русском языке \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

***\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***

Студент \_\_ИУ7-85Б\_\_\_\_\_\_ **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ О.П. Кондрашова\_\_\_**

(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель ВКР **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_А.П. Ковтушенко\_\_\_**

(Подпись, дата) (И.О.Фамилия)

Консультант **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Л.Л. Волкова\_\_\_\_\_\_\_\_**

(Подпись, дата) (И.О.Фамилия)

Консультант **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

(Подпись, дата) (И.О.Фамилия)

Нормоконтролер **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_Ю.В. Строганов\_\_\_**

(Подпись, дата) (И.О.Фамилия)

*2021 г.*

# **РЕФЕРАТ**

Расчётно-пояснительная записка, ?? с., ?? рисунков, ?? табл., ?? источников.

СОДЕРЖАНИЕ

[РЕФЕРАТ 2](#_Toc72687540)

[ВВЕДЕНИЕ 4](#_Toc72687541)

[1 Аналитический раздел 6](#_Toc72687542)

[1.1 Описание предметной области 6](#_Toc72687543)

[1.2 Этапы обработки текста 7](#_Toc72687544)

[1.2.1 Токенизация 7](#_Toc72687545)

[1.2.2 Нормализация 9](#_Toc72687546)

[1.3 Этапы анализа текста 9](#_Toc72687547)

[1.4 Методы и решения определения авторского инварианта 11](#_Toc72687548)

[1.4.1 Метод полного синтаксического анализа 12](#_Toc72687549)

[1.4.2 Метод энтропийной классификации 13](#_Toc72687550)

[1.4.3 Метод на основе использования нейронных сетей 14](#_Toc72687551)

[1.4.4 Метод анализа длин слов 15](#_Toc72687552)

[1.4.5 Применения методов из теории вероятности и математической статистики 15](#_Toc72687553)

[1.4.6 Метод выделения N-грамм 16](#_Toc72687554)

[1.5 Классификация текстов 17](#_Toc72687555)

[1.6 Классификация текстов 17](#_Toc72687556)

[1.7 Последовательность решения задачи 17](#_Toc72687557)

[1.8 Выводы 18](#_Toc72687558)

[2 Конструкторский раздел 19](#_Toc72687559)

[3 Технологический раздел 20](#_Toc72687560)

[4 Экспериментальный раздел 21](#_Toc72687561)

[ЗАКЛЮЧЕНИЕ 21](#_Toc72687562)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 22](#_Toc72687563)

# **ВВЕДЕНИЕ**

В последние годы очень быстрыми темпами развивается область обработки естественных языков. Во многом это связано с тем, что с каждым годом объём текстовой информации, используемой человечеством, увеличивается, и растёт потребность в более эффективных алгоритмах обработки и анализа документов, написанных на естественных языках. Особо важную роль играет возможность определить автора текста, основываясь на его стилистических признаках.

Выявление признаков авторского стиля позволяет установить принадлежность текста определенному человеку. Актуальность данного направления в компьютерной лингвистике обусловлена необходимостью в выявлении плагиата или в создании рекомендательной системы для нахождения похожжих текстов. Для определения стиля автора необходимо выделить характерные признаки из принадлежащих ему текстов.

Целью данной работы является разработка метода определения признаков авторского стиля для текстов на русском языке.

Для достижения поставленной цели ставятся следующие задачи:

* изучение предметной области;
* анализ существующих решений определения авторства текста;
* анализ алгоритмов классификации;
* разработка метода определения признаков авторского стиля;
* разработка метода классификации текстов, использующего предложенный метод классификации текстов;
* проектирование структуры и реализация программного обеспечения, реализующего разработанный метод;
* подготовка входных данных и обучение классификатора;
* апробация предложенного метода.

Разрабатываемый метод позволит проводить классификацию текстов на русском языке по авторскому стилю на основании результатов морфологического анализа, не задействуя более трудоемкий и менее доступный синтаксический анализ.

# **Аналитический раздел**

В данном разделе рассматриваются понятия авторского стиля и стилеметрических характеристик, исследуются алгоритмы классификации и существующие методы, и решения определения авторского стиля, анализируются их достоинства и недостатки. Также в этом разделе приводится описание последовательного решения задачи.

## Описание предметной области

Под авторским стилем понимается совокупность характеристик, позволяющих установить авторство текста или выдвинуть предположение, кем может являться автор или к какой группе авторов он может принадлежать. К особенностям авторского стиля можно отнести грамматические конструкции, стилистические приемы, способы построения фраз и абзацев или любой другой набор признаков, который отличает конкретного автора от всех других.

Раздел лингвистики, зани­мающийся измерением стилевых характеристик с целью систематиза­ции и упорядочения текстов, называется стилеметрией [1]. Объектом стилеметрии является текст, созданный конкретным ав­тором. Предметом ис­следования являются элементы стиля, которые понимаются как осо­бенности стиля конкретного автора. Стиль текста описывается как набор некоторых выделенных характеристик. В большинстве случаев в качестве характеризующих параметров текста выбираются статистические характеристики: количество использования определенных частей речи, знаков препинания, количество и длина предложений (измеренная в словах, слогах, знаках), количество полнозначных и служебных слов, средняя длина предложения в тексте и т.д.

Под авторским инвариантом понимаются стилеметрические характеристики, которые однозначно характеризует одного автора или небольшое число «близких авторов» [2]. Анализируемые характеристики должны принимать существенно разные значения для произведений разных групп авторов. Для успешного установления авторства нужно, чтобы число «разных групп» было достаточно велико, и при этом каждая группа должна включать в себя небольшое количество близких по стилю произведений, для которых точно определен автор.

Методы анализа стиля можно разделить их на две большие группы – экспертные и формальные. Экспертные методы предполагают исследование текста профессиональным лингвистом-экспертом. К формальным относятся приемы из теории вероятностей и математической статистики, алгоритмы классификации, кластерного анализа и нейронных сетей.

## Этапы обработки текста

В данном разделе будет выполнен обзор этапов обработки текстов на естественном языке.

### Токенизация

Токенизация — это базовый этап в автоматической обработке текстов, суть которого заключается в разбиении непрерывной строки на отдельные «слова» (токены) [3].

На этапе токенизации также решается задача разбиения текста на предложения. При разделении текста на отдельные предложения недостаточно учитывать только наличие точки и большой буквы после нее. Например, под данное правило не попадают сокращения и инициалы («И.И. Иванов» — три разных предложения).

Токенизация текста состоит из нескольких этапов. В первую очередь текст полностью приводится к нижнему регистру. На данном этапе часть информации может быть утеряна. Например, «ООО» представляет собой аббревиатуру (общество с ограниченной ответственностью), в то время как «ооо» — способ выражения эмоций.

Следующий этап — это замена всех знаков препинания и прочих символов на пробелы. Текст на естественном языке состоит не только из букв, но и из символов: скобок, кавычек, тире. Если все символы, не являющиеся буквами или строками, заменить на пробелы, то слова, разделённые пробелами, можно объявить отдельными токенами. Однако при наличии сложных составных слов замена дефиса на пробел может привести к потере исходного смысла предложения. Например, слово «каком-либо» было бы правильно считать за единый токен, но при замене дефиса на пробел, полчаются два отдельных слова: «каком» и «либо».

Далее каждое слово, отделённое пробелом, объявляется отдельным токеном. На данном этапе стоит учитывать, что некоторые наборы слов должны рассматриваться как одно. Например, названия городов («Нижний Новгород»), или сокращения («к.т.н.», кандидат технических наук). Если рассматривать слова «Нижний» и «Новгород» независимо друг от друг, важная информация, определяющая конкретный текст, может быть утеряна, а разделенные на буквы сокращения и вовсе теряют какой-либо смысл [4].

### Нормализация

Следующий этап после токенизации — это нормализация слов в тексте, то есть приведение каждого слова к его начальной форме.

Существует два основных подхода к нормализации: стемминг и лемматизация [4].

Стемминг — это подход, суть которого заключается в следующем: по некоторым правилам от каждого слова отрезается его окончание. Данный подход может работать некорректно в том случае, если при изменении формы слово меняется целиком (например, «был», «есть», «будет»).

Суть подхода лемматизации заключается в использовании словаря, в который предварительно записано большое количество слов и их форм. В первую очередь слово проверяется по словарю. Если слово найдено в словаре, то с его помощью можно найти известные формы данного слова. В противном случае по определённому алгоритму выводится способ изменения слова, и на его основании делаются выводы о первоначальной форме слова.

Лемматизация дает более точный результат при работе с незнакомыми словами, но в связи с поиском в словаре и работой алгоритма приведения к нормальной форме, является более медленным подходом по сравнению со стеммингом.

## Этапы анализа текста

Полный анализ текста включает в себя несколько этапов [3]. Сложность полного анализа заключается в том, что естественный язык неполон и не всегда формализован, также возникающие сложности связаны с тем, что на практике до сих пор не реализованы все теоретические положения, разработанные на данный момент. Выбор того, какой анализ текста необходимо провести, зависит от поставленной задачи и выбранного пути ее решения.

Графематический анализ – начальный анализ текста на естественном языке, обеспечивающий выделение синтаксических или структурных единиц. Входные данные могут представлять собой линейную структуру, содержащую единый фрагмент текста, но в более общем случае подаваемый на вход текст состоит из многих структурных единиц: основного текста, заголовков, вставок, комментариев и т.д. При машинном переводе ставится задача сохранить структуру исходного текста. Первый вариант – линейная структура текста без вставок – обычно встречается в диалоговых системах. Но и в этом случае графематический анализ должен выделять синтаксические единицы: абзацы, предложения, отдельные слова и знаки препинания.

Морфологический анализ обеспечивает для каждой словоформы в тексте определение ее нормальной формы, от которой данная словоформа образована, а также набора характеризующих ее параметров. Морфологический анализ является основой для нормализации текста и проводится для того, чтобы в дальнейшем ориентироваться только на нормальную форму. Нормализация требует обязательного морфологического анализа текста, распознающего части речи с учетом контекста и многочисленных правил согласования (без него нормализация будет давать значительное количество ошибочных результатов).

Предсинтаксический анализ отвечает за две противоположные задачи: объединение отдельных лексических единиц в одну синтаксическую или, наоборот, ее разделение на несколько. В единую синтаксическую единицу объединяются изменяемые неразрывные словосочетания, т.к. в данном случае при использовании слов по отдельности теряется первоначальный смысл, заложенный в конкретное словосочетание. К задачам предсинтаксического анализа также относится проведение синтаксической сегментации, которая заключается в разметке линейного текста на фрагменты, привязанные правилам следующего этапа синтаксического анализа. Синтаксический анализ является задачей с экспоненциальным ростом сложности, поэтому для ускорения его работы часть задач переносится на этап предсинтакисческиго анализа.

Синтаксический анализ – самая трудоемкая часть анализа текста. На данном этапе необходимо определить роли слов в предложении и их связи между собой. Необходимо сопоставить линейную последовательность лексем (слов, токенов) языка с его формальной грамматикой. Результатом этого этапа является набор синтаксических деревьев, которые наглядно показывают выделенные в тексте связи и являются материалом для дальнейшем обработки.

На этапе постсинтаксического анализа требуется уточнить смысл, заложенный в слова и выраженный при помощи различных средств языка (предлоги, префиксы, аффиксы и т.д.). Необходимо учесть, что одна и та же мысль может быть выражена различными конструкциями языка. В связи с этим полученное на предыдущем этапе синтаксическое дерево необходимо нормализовать, т.е. конструкция, выражающая некоторое действие различным образом для различных языков или ситуаций, должна быть сведена к одному и тому же нормализованному дереву.

Семантический анализ проводится для анализа текста «по смыслу». Семантический анализ уточняет связи, которые не смог уточнить постсинтаксический анализ, так как многие роли выражаются не только при помощи средств языка, но и с учетом значения слова. Словарь для поддержки семантического анализа должен оперировать смыслами и, следовательно, описывать свойства и отношения понятий, а не слов.

## Методы и решения определения авторского инварианта

В данном разделе проанализированы существующие методы и решения для определения признаков авторского стиля.

### Метод полного синтаксического анализа

Под синтаксическим анализом в области обработки текстов на естественном языке понимается построение такой структуры, которая позволяет приблизиться к некоторому формализованному представлению смысла текста и установить связи между словами в предложении.

Большинство моделей синтаксической структуры предложения опираются либо на грамматику составляющих, предложенной в работах Ноама Хомского [5], либо на грамматику зависимостей, для которой основополагающими считаются работы Люсьена Теньера [6] и Игоря Мельчука [7].

Грамматика составляющих предполагает, что предложение на естественном языке может быть представлено в виде иерархии составляющих. Данная иерархия предполагает выделение синтаксических групп, которые не должны пересекаться между собой, но которые в свою очередь состоят из более мелких групп, вплоть до атомарных единиц – слов предложения. Такую иерархическую структуру называют деревом составляющих. Дерево составляющих удовлетворяет свойству проективности, согласно которому при представлении дерева графически связи между структурными единицами не должны пересекаться.

Грамматика зависимостей предполагает, что предложения текста представляют собой деревья зависимостей, в которых слова связаны ориентированными дугами, обозначающими синтаксическое подчинение. Считается, что подход с деревом зависимостей более точно отражает специфику языков с произвольным порядком слов. Грамматика зависимостей допускает наличие непроективных связей, нарушающих свойство проективности дерева

Синтаксический анализ разделяют на глубокий (полный) анализ и поверхностный [8, 9]. Задачей глубокого синтаксического анализа является построение полного синтаксического дерева предложения с максимальной связанностью с учетом дальних связей, а также определение грамматических функций слов предложения (подлежащее, сказуемое, обстоятельства места, времени и т.д.). Понятие поверхностного синтаксического анализа объединяет в себе различные подходы, направленные на построение неполной (частично связанной) синтаксической структуры текста разной сложности. Поверхностный синтаксический анализ охватывает такие задачи как разделение предложения на рекурсивно невложенные синтаксические группы, сегментацию (выделение в предложении различных оборотов и простых предложений в составе сложного), построение поверхностного синтаксического дерева.

Полный синтаксический разбор гарантирует высокое качество анализа, поскольку устанавливает все существующие зависимости при разборе предложения. Но т.к. в русском языке в предложении зачастую встречается свободный порядок слов (в отличие, например, от английского языка), данный метод зачастую требует корректирующего вмешательства эксперта вдобавок к машинному синтаксическому разбору. В силу этих особенностей языка метод синтаксического разбора является очень затратным по времени. К тому же большая часть программного обеспечения, выполняющего синтаксический анализ, является коммерческими проектами и закрыта для использования.

### Метод энтропийной классификации

Метод классификации текстов для задачи определения авторства, предложенный в работе [10], основан на применении алгоритмов сжатия данных.

Суть рассматриваемого метода заключается в том, чтобы добавить текст, автор которого неизвестен, к корпусу текстов, характеризующему конкретного автора, проанализировать, насколько хорошо сжимается полученная после добавления выборка текстов, и сравнить полученные результаты для различных авторов. В рамках энтропийного подхода правильным исходным классом текста является тот класс, на котором получены наилучшие результаты сжатия.

Рассмотрим на примере задачу классификации текста относительно текстов , характеризующих классов. Выбор источника текста должен осуществляться согласно оценке, рассчитываемой по формуле:

где ­– характеристика энтропии текста по отношению к тексту

Существенным преимуществом метода энтропийной классификации является отсутствие необходимости в предварительной обработке текста.

### Рекуррентные нейронные сети

В работе [11] рассматривается способ, основанный на использовании нейронных сетей и методов иерархической кластеризации. Предлагается представлять тексты в виде матриц частот появления признаков, где под частотным признаком понимается любой признак, допускающий возможность нахождения частоты его появления в тексте (например, количество определенных слов). В качестве меры сравнения полученных матриц предлагается использовать меру Кульбака и меру -квадрат. На основе проведенных исследований разработан программный комплекс «Стилеанализатор».

Было предложено использовать нейронные сети, обучающиеся без учителя и предназначенные для обработки больших массивов многомерной информации, – самоорганизующиеся карты Кохонена (Self-organizing map – SOM) [12]. Недостатком SOM-сетей является недостаточное теоретическое обоснование, в основном они опираются на эмпирические результаты.

В итоге был получен вывод о том, что в случае удачного нахождения универсального набора характеристик можно обрабатывать любое число авторов и текстов (большие массивы информации). Достаточно постоянно модифицировать карту, добавляя новые произведения, и оценивать, как они взаимодействуют с уже имеющимися в базе.

Одним из серьезных недостатков метода является невозможность прогнозирования успешного результата. Поиск на заданном наборе текстов может никогда не найти хороший вариант для разделения характеристик.

Также проблемой метода является его трудоемкость. Качество поиска напрямую зависит от количества загруженных текстов, следовательно, по мере увеличения числа текстов, требуется задействовать всё больше ресурсов вычислительной системы (необходим большой объем памяти и мощный процессор).

### Метод анализа длин слов

Для поставленной задачи определения авторства текста существует также решение анализа стилистики текста, основанное на длине слов. Данное решение легло в основу программы «Худломер» [13].

«Худломер» — это метод автоматической классификации функционального стиля текста на основе спектров длин слов. Программа позволяет определять следующие стили: разговорный стиль, стиль художественной литературы, газетно-информационный стиль, научно-деловой стиль.

### Применения методов из теории вероятности и математической статистики

Предлагаемый в работе [14] метод основывается на учете статистики употребления пар элементов любой природы, идущих друг за другом в тексте (букв, морфем, словоформ и т. п.), т. е. на формальной математической модели последовательности букв (и любых других элементов) текста как реализации цепи Маркова. По произведениям, для которых достоверно определен автор, вычисляется матрица переходных частот употребления пар элементов (букв, грамматических классов слов и т. п.). Для каждого автора строится матрица переходных частот и оценивается вероятность того, что именно он написал исследуемый текст. Автором неизвестного текста считается тот, для кого вычисленная оценка вероятности больше.

Примеры того, в виде каких пар можно представить текст:

а) пары букв в словах (в той форме, в которой они употреблены в тексте) и пробелах между ними;

б) пары букв в словах, приведенных к начальной форме;

в) пары наиболее обобщенных грамматических классов слов в их последовательностях в предложениях текста. К таким классам слов относят части речи и некоторые условные категории вроде «конец предложения», «сокращение» и др.;

г) пары менее обобщенных грамматических классов слов. К ним относятся такие семантико-грамматические разряды, как одушевленные и неодушевленные существительные, прилагательные качественные, относительные, притяжательные и т. п.

### Метод выделения N-грамм

N-грамма — это последовательность из N идущих подряд слов в тексте. Использование N-грамм характерно тем, что позволяет учитывать порядок слов.

Для предложения «Наборы подряд идущих токенов» существуют следующие N-граммы:

* униграммы (N = 1): наборы, подряд, идущих, токенов;
* биграммы (N = 2): наборы подряд, подряд идущих, идущих токенов;
* триграммы (N = 3): наборы подряд идущих, подряд идущих токенов.

N-грамма имеет в основе своего использования математическую модель и определяется следующим образом: «N-граммой на алфавите V называют произвольную цепочку длиной N, например последовательность из N букв алфавита V одного слова, одной фразы, одного текста или, в более интересном случае, последовательность из грамматически допустимых описаний N подряд стоящих слов» [15].

Чем больше N, до которого будут найдены N-граммы, тем больше будет получено признаков. При этом увеличение параметра N может привести к переобучению. Например, если N — это длина текста, то у каждого документа будет уникальный признак, и алгоритм сможет переучиться под обучающую выборку.

N-граммы можно использовать не только на словах. В качестве токенов можно рассматривать отдельные символы в предложении. После нахождения таких токенов для них можно также вычислять N-граммы (буквенные N-граммы). Данный подход позволяет учитывать известные слова в незнакомых формах. Часто буквенные N-граммы используют вместе с N-граммами по словам.

Skip-граммы — это чуть более расширенный подход к использованию N-грамм. Более точно они называются k-skip-n-граммы, это наборы из N токенов, причём расстояние между соседними должно составлять не более K токенов [16]. Данный метод позволяет учитывать больше различных N-грамм в предложении, и, в связи с этим, как правило, используется в языковом моделировании в сочетании с другими подходами.

## Извлечение признаков из текста

### Счетчики слов

### TF-IDF

## Классификация текстов

## Последовательность решения задачи

Основываясь на проведённом анализе предметной области, цель данной работы можно уточнить следующим образом: необходимо разработать метод определения признаков авторского стиля для текстов на русском языке.

В первую очередь необходимо провести предварительную обработку текстов из размеченного корпуса. Каждый текст разбивается на отдельные предложения, далее проводится токенизация, лемматизация и поределения части речи для каждого слова. В пределах каждого отдельного предложения выделяются частеречные N-граммы, представляющие собой связку из частей речи, то есть кортеж из словоформ будет представлен как кортеж из частей речи этих словоформ (например, существительное-глагол-прилагательное-существительное).

После анализа каждого текста из имеющейся базы будет получен набор данных – статистическая характеристика распределения N-грамм с определенными частями речи, характерная для конкретного автора. Для оценки важности выделенной N-граммы для текста заданного автора предлагается использовать статистическую меру TF-IDF.

Каждый текст необходимо представить в виде вектора признаков, тогда корпус текстов будет представлять собой матрицу признаков.

На основе полученной матрицы признаков проводится классификация текстов на обучающей выборке из размеченного корпуса текстов.

Следующий этап метода – проведение морфологического и N-граммного анализа для нового текста, не входящего в обучающую выборку. Завершается метод классификацией текста на основе полученного распределения частеречных N-грамм. В результате можно установить автора текста или выдвинуть предположение, кто может быть автором (в таком случае принятие решения предоставляется эксперту).

## Выводы

В данном разделе были рассмотрены этапы обработки текстов на естественном языке. Проведен обзор существующих методов и решений определения авторского стиля. Были проанализированы алгоритмы классификации текстов.

# **2 Конструкторский раздел**

# **Технологический раздел**

# **4 Экспериментальный раздел**

# **ЗАКЛЮЧЕНИЕ**

# 

# **СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ**

1. Мартыненко Г. Я. Стилеметрия: Возникновение и становление в контексте междисциплинарного взаимодействия. Часть 2. Первая половина XX в.: Расширение междисциплинарных контактов стилеметрии. СТРУКТУРНАЯ И ПРИКЛАДНАЯ ЛИНГВИСТИКА, № 11, 2015, С. 9-28.
2. ﻿Фоменко В.П., Фоменко Т.Г. Авторский инвариант русских литературных текстов. Предисловие А.Т. Фоменко // Фоменко А.Т. Новая хронология Греции: Античность в средневековье. Т. 2. М.: Изд-во МГУ, 1996. С. 768-820.
3. ﻿Ермакович М.В. Автоматическое определение границ слова в русском тексте с помощью комплекса лингвистических правил. Компьютерная лингвистика и интеллектуальные технологии: по материалам международной конференции «Диалог 2017» Москва, 31 мая — 3 июня 2017.
4. ﻿Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В.
5. ﻿Chomsky N. Three models for the description of language // IRE Transactions on Information Theory. — 1956. — Vol. 2, no. 3. — P. 113–124.
6. ﻿Tesnière L. Elements de syntaxe structurale. — Editions Klincksieck, 1959.
7. ﻿Mel’cuk I. A. Dependency syntax: theory and practice. — ŠUNY Press, 1988. — P. 428.
8. Abney S. P. Parsing by chunks // Principle-Based Parsing. — Kluwer Academic Publishers, 1991. — P. 257–278.
9. Federici S., Montemagni S., Pirrelli V . Shallow parsing and text chunking: a view on underspecification in syn- tax // Cognitive science research paper-university of Sus- sex CSRP. — 1996. — P. 35–44.
10. ﻿Хмелев Д.В. Классификация и разметка текстов с использованием методов сжатия данных. Краткое введение [электронный ресурс] - режим доступа: URL: <http://compression.ru/download/articles/classif/intro.html> (дата обращения: 11.12.2020).
11. Шевелев О. Г. Методы автоматической классификации текстов на естественном языке: учебное пособие / О. Г. Шевелев ; науч. ред. В. В. Поддубный ; Том. гос. ун-т. - Томск : ТМЛ-Пресс, 2007.
12. T. Kohonen, Self-Organizing Maps (Third Extended Edition), New York, 2001, 501 pages.
13. ﻿Программы анализа и лингвистической обработки текстов. Русская виртуальная библиотека [электронный ресурс] - режим доступа: URL: https://rvb.ru/soft/catalogue/c01.html (дата обращения: 11.12.2020).
14. Кукушкина О. В., Поликарпов А. А., Хмелев Д. В. Определение авторства текста с использованием буквенной и грамматической информации // Проблемы передачи информации. М.: Наука, 2001. Т. 37, № 2. С. 96–108.
15. ﻿Гудков В. Ю., Гудкова Е. Ф. N-граммы в лингвистике // Вестник Челябинского государственного университета. 2011. № 24 (239). Филология. Искусствоведение. Вып. 57. С. 69–71.
16. ﻿Guthrie David, Allison Ben, Liu Wei, Guthrie Louise, Wilks Yorick. (2006). A Closer Look at Skip-gram Modelling. Proc. of the Fifth International Conference on Language Resources and Evaluation.
17. Shahzad Qaiser, Ramsha Ali. Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. International Journal of Computer Applications (0975 – 8887) Volume 181 – No.1, July 2018