УДК 811.161.1’371:004.9

В.А. Яцко

Создание списка стоп-слов русского языка*

Описываются признаки, необходимые для идентификации стоп-слов: статистический, семантический и морфологический. На их основе формулируются новые принципы создания списков стоп-слов. Показывается применение этих принципов для создания списка стоп-слов русского языка. На основе анализа существующих списков и распределения терминов в Национальном корпусе русского языка создан список универсального типа, включающий 535 стоп-слов.

Ключевые слова: *списки стоп-слов, фильтрация, признаки и методы идентификации, принципы создания, русский язык*

DOI: 10.36535/0548-0027-2022-05-4

Одна из распространённых процедур предварительной обработки текстовых документов – фильтрация стоп-слов, под которой понимается исключение некоторых лингвистических единиц из результатов статистического анализа (при том, что их распределение учитывается в общем количестве токенов и уникальных слов). Фильтрацию следует отличать от элиминации, предусматривающей удаление соответствующих лингвистических единиц (терминов) из текста и полное игнорирование данных об их распределении, а также искажающей реальное статистическое распределение терминов текстового документа и поэтому применяющейся достаточно редко.

К типичным стоп-словам относятся артикли, предлоги, частицы, местоимения, союзы. Приведенные в работе [1] подсчеты показали, что десять наиболее частотных английских слов составляют 20 – 30% всех токенов текстов. Их фильтрация позволяет существенно уменьшить размеры обрабатываемых текстовых данных и повысить быстродействие лингвистического программного обеспечения. Это имеет критическое значение для функционирования универсальных информационно-поисковых систем в силу больших объёмов их баз данных. Фильтрация стоп-слов важна и для систем распознавания плагиата, в основе которых лежит выявление совпадающих терминов разных текстов. Очевидно, что предлоги, союзы и другие указанные морфологические классы слов встречаются во всех текстах и их совпадение не может служить опорой при решении этой задачи. Соответственно, фильтрация стоп-слов повышает эффективность решения задачи распознавания плагиата. Предварительная фильтрация применяется и в других видах лингвистических технологий, предусматривающих опору на ключевые термины текста: автоматическом реферировании, анализе мнений пользователей, автоматической классификации, включая тематическую категоризацию, жанровую классификацию, авторскую атрибуцию.

Несмотря на очевидную актуальность и важность фильтрации стоп-слов для автоматической обработки текста, до сих пор не установлены общепринятые критерии и принципы их идентификации. В предлагаемой статье мы рассмотрим основные признаки стоп-слов, подходы, необходимые для их идентификации, а также впервые сформулируем критерии идентификации стоп-слов, которые будут применены для составления списка стоп-слов русского языка.

ПРИЗНАКИ И КРИТЕРИИ ИДЕНТИФИКАЦИИ СТОП-СЛОВ

В настоящее время сложилось два подхода к идентификации стоп-слов: словарный и алгоритмический. Словарный подход предусматривает создание и использование списков стоп-слов, в то время как алгоритмический основан на математических методах анализа распределения и взвешивания терминов. Критерием идентификации стоп-слов в этом случае выступают нулевые или близкие к ним весовые коэффициенты терминов. Наиболее распространённый алгоритмический метод – это фильтрация на основе формулы TF*IDF (англ.: term frequency*inverse document frequency):

$$w_{ij} = tf_{ij} * \log_2 \frac{N}{n}, \qquad (1)$$

где: w_{ij} – вес термина t_i в текстовом документе d_j;
tf_{ij} – частотность термина t_i в документе d_j;
N – общее количество документов в корпусе;
n – количество документов в корпусе, в которых термин t_i встречается хотя бы один раз.

Существуют различные модификации представленной формулы, рассмотренные в [2]. Согласно формуле (1) наибольшие коэффициенты получают термины с высокой частотностью в тексте и наименьшей частотностью в корпусе. Уменьшение частотности в

* Исследование поддержано грантом РФФИ 20-07-00124

конкретном документе и/или увеличение встречаемости в документах корпуса приводит к снижению весового коэффициента. Термины, которые встречаются во всех документах корпуса, получают нулевые коэффициенты (при $N=n$), что позволяет фильтровать стоп-слова автоматически, без применения специальных списков, поскольку именно стоп-слова встречаются во всех текстах.

Применение формулы (1) вызывает трудности, к которым относятся неопределённость размера корпуса, т. е. количество текстов в нём (величина N), а также жанрово-стилистический состав корпуса. Непонятно, какие тексты включать в корпус: относящиеся к тому же жанру, что и анализируемый текст (комплементарное сопоставление), или относящиеся к другому жанру (контрастивное сопоставление). Проведённое в [3] тестирование показало, что состав корпуса существенно влияет на результаты взвешивания. Неправильное составление корпуса может привести к неадекватным результатам взвешивания, начислению стоп-словам высоких числовых коэффициентов. Это относится и к другим алгоритмическим методам, которые также предусматривают сопоставление входного текста с другим текстом/текстами (например, прирост информации, отношение шансов, логарифмическое подобие).

Списки стоп-слов, т. е. словарные методы, позволяют избежать недостатков алгоритмических методов, так как имеют следующие преимущества:

1) простота вычислений – не нужны сложные методы взвешивания и определения пороговых уровней;

2) в отличие от алгоритмических методов, требующих для каждой предметной области составления отдельного корпуса, список стоп-слов является постоянной величиной, независимой от особенностей определённого типа, жанра текстов или предметной области, с которой они соотносятся;

3) количество стоп-слов в языке ограничено, что позволяет решить проблему сокращения размерности текстов. Существующие списки стоп-слов для разных языков включают 200 – 600 терминов. Однако и это количество может быть сокращено в случае применения итеративного порогового уровня, включения в список только терминов с уникальными значениями частотностей;

4) отсутствие производных форм – позволяет обойтись без применения таких алгоритмов предварительной обработки как стемминг и лемматизация, что повышает быстродействие классификатора. Если в списки стоп-слов, на основе которых проводится фильтрация, включаются формы, принимающие суффиксы и окончания, то приводятся все возможные производные формы таких слов;

5) распределение стоп-слов по частотностям специфично для каждого текста и имеет высокую дискриминативную силу (способность идентифицировать класс текста) в случае применения адекватного нормализирующего фактора, в качестве которого в работе [4] было предложено использовать отклонения частотностей стоп-слов от коэффициента Ципфа, что в итоге позволило эффективно решать задачи автоматической классификации текстов.

Применение словарного подхода сталкивается с проблемой выделения критериев идентификации стоп-слов. Полагаем, что в качестве таких критериев могут выступать признаки, отличающие стоп-слова от знаменательных слов языка. Можно выделить три признака стоп-слов: статистический, семантический, морфологический.

Статистический признак указывает на высокую частотностью стоп-слов в разных текстах, независимо от их жанра или функционального стиля. Т. е. этот признак сочетается со свойством жанрово-стилистической независимости распределения терминов. Обычно именно статистический признак считается ведущим, на его основе и происходит выделение стоп-слов из состава остальной лексики. К. Фокс создавал универсальный список стоп-слов английского языка в 1989 г., руководствуясь анализом частотности слов в Брауновском корпусе[1] [5]. Список составлялся по следующей методике. Из Брауновского корпуса было выбрано 278 слов, частотность которых превышала 300; далее из этого списка на основе ручного анализа исключили 32 слова и добавили 149 слов, отобранных в полуавтоматическом режиме на основе частотности и сходства со словами, уже включёнными в список (например, добавлены слова с частотностью более 100 и отличающиеся от слов, включенных в список, на одну букву). В результате получился список из 421 слова. Дополненный список, состоящий из 426 слов (https://iatskota.wixsite.com/yatsko/tf-idf-ranker), мы успешно использовали в целях автоматической классификации текстовых документов. В дополненный список были включены недостающие глагольные словоформы, а также формы, которые используются после апострофа и распознаются современными токенайзерами как отдельные токены ("ve", "ll", "re").

Работа, проведённая К. Фоксом, продемонстрировала проблемы, связанные с составлением списков и словарей стоп-слов: сложность определения пороговых уровней, необходимость учёта словоформ, недостаточная адекватность жанрово-независимого списка. Не вполне понятно, почему в качестве порогового уровня была выбрана частотность 300, а не 299 или 287, также неясно, почему в словарь ввели словоформы *go* и *went*, но не добавили *gone*. Включение в список глагольных форм, таких как *go* допустимо в том случае, если они используются в качестве глагола-связки. По нашим подсчётам в *Corpus of Contemporary American English* (https://www.english-corpora.org/coca/) частотность *go* как глагола-связки составляет 3447586, а в качестве личного глагола – 125320, т. е. в 27,5 раза меньше, что указывает на правомерность включения этой глагольной формы в универсальный список стоп-слов. Вместе с тем следует иметь в виду, что в некоторых текстах, например, художественных повествовательного типа или текстах-инструкциях *go* может преимущественно и достаточно часто использоваться как глагол движения, ср. *John went to his office*. При анализе таких текстов может потребовать-

[1] К. Фокс использовал термин «general text» (текст общего типа).

ся модификация универсального списка, исключение из него соответствующих глагольных форм.

Статистический критерий повсеместно используется и в других проектах, предусматривающих фильтрацию стоп-слов. Списки стоп-слов включены в пакеты библиотек для лингвистического программирования на языках *Python* (127 стоп-слов – https://gist.github.com/sebleier/554280) и *R* (174 стоп-слова – https://campus.datacamp.com/courses/text-mining-with-bag-of-words-in-r/jumping-into-text-mining-with-bag-of-words?ex=12). Источником стоп-слов в этих пакетах выступают списки, составленные М. Портером, выкладываемые на ресурс для поддержки программирования *GitHub*. Все эти списки модифицируются пользователями в зависимости от задач конкретных проектов без учета каких-либо принципов их составления.

Полагаем, что для идентификации стоп-слов следует учитывать также семантический и морфологический признаки, что особенно важно для морфологически развитых языков, таких как русский язык.

Семантический признак указывает на то, что стоп-слова, взятые вне контекста, не выражают значения; по этому признаку стоп-слова отличаются от знаменательных слов. Например, слово "стол" будет обозначать соответствующий объект в ранжированном списке терминов некоторого текста вне контекста его использования, в то время как смысл местоимения "оно" может быть понят только в контексте его применения. По семантическому признаку в списки стоп-слов включается ряд глагольных форм, в первую очередь, глаголы-связки и вспомогательные глаголы, также не выражающие значения в ранжированных списках терминов.

Морфологический признак проявляется в том, что большинство стоп-слов не принимают суффиксов и окончаний и не имеют производных форм. Это особенно характерно для языков со слабо развитой морфологией, таких как английский. Морфологический признак является вспомогательным, так как требование морфологической неизменяемости не всегда может быть соблюдено. Включение в списки стоп-слов глагольных форм предполагает также включение соответствующих деривативных форм, которые есть и в языках морфологически слабо развитых. Например, наряду с глаголом *go* следует учитывать формы *goes, went, gone, going*.

ПРИНЦИПЫ СОЗДАНИЯ СПИСКА СТОП-СЛОВ

На основе проанализированных нами признаков можно выделить следующие принципы идентификации стоп-слов.

1. Принцип равномерности жанрово-стилистического распределения. Соотносится со статистическим признаком стоп-слов и предполагает анализ статистического распределения терминов по основным жанрам эталонного корпуса. Цель анализа – определить, насколько равномерно распределено стоп-слово по жанрам. Под равномерностью распределения понимается небольшой разрыв между частотностями данного стоп-слова в разных жанрах. Пороговым уровнем будет считаться разрыв в 48%, определяемый по простой вероятностной величине $P(i)$, которая находится делением частотности термина на количество его словоформ (токенов) в корпусе. Если разрыв в распределении рассматриваемого термина по жанрам превышает пороговую величину, этот термин не включается в список стоп-слов. Указанный пороговый уровень был получен в результате анализа распределения наиболее типичных стоп-слов[2] в *Национальном корпусе русского языка* (НКРЯ). В таблице приводятся данные о распределении десяти наиболее частотных лемм русского языка по художественным и нехудожественным текстам *Основного корпуса*[2] (https://ruscorpora.ru/new/search-main.html) НКРЯ, а также указывается разрыв в их распределении, вычисленный на основе нормализованных величин (округлялись до трёх десятичных знаков). *Основной корпус* НКРЯ включает 337025184 токенов; размер подкорпусов художественных и нехудожественных текстов составляет, соответственно, 144957347 и 172993559 токенов. Видно, что разрыв в распределении по двум подкорпусам изменяется в диапазоне примерно от 3% (союз *и*) до 48% (местоимение *я*).

2. Принцип выбора оптимального объёма. Объём каждого списка определяется в зависимости от особенностей конкретного языка. Для морфологически развитого языка (например, русского) список стоп-слов должен включать большее количество терминов, чем для морфологически бедного языка, например, английского. Для последнего эталонным считается список Фокса, который был рассмотрен ранее. Мы много раз успешно использовали этот список для решения классификационных проблем, в частности жанровой классификации и авторской атрибуции, поэтому можно принять размер этого списка (426 терминов) в качестве исходного. Однако в русском языке намного больше словоформ, и поэтому список стоп-слов должен быть больше. Полагаем, что определить размер списка стоп-слов русского языка можно, исходя из константы Ципфа A, которая находится по формуле:

$$A = R(w_i) \times \frac{F(w_i)}{N_{(w)}}, \qquad (2)$$

где: $R(w_i)$ – ранг термина w_i;

F – его частотность в некотором репрезентативном для данного языка корпусе;

N – количество токенов в корпусе.

Эмпирическим путём в результате анализа большого массива данных было установлено, что константа A = 0,1 для английского языка [7] и 0,08 – для русского (https://studbooks.net/2056134/informatika/pervyy_zakon_tsipfa_rang_chastota). Соотношение между этими величинами составляет 1,25. Применяя обратную экстраполяцию, получаем приблизительный размер списка стоп-слов: 532 термина (426×1,25).

3. Принцип учёта многоплановости текстов на естественном языке. Исходя из нашей концепции интегративного подхода к анализу текста [8], считаем возможным выделить и отнести к классу стоп-слов термины с модальными, реляционными и прагматическими значениями.

[2] Ранжированный по частотности список приводится в [6].

Таблица

Распределение стоп-слов на основе данных в НКРЯ

№	Стоп-слово	Художественные		Нехудожественные		Разрыв
		частотность	$P(i)$	частотность	$P(i)$	
1	*и*	5703809	0.039	6595468	0.038	3.107%
2	*в*	3564317	0.025	5664104	0.033	24.901%
3	*не*	3004101	0.021	2629688	0.015	26.650%
4	*на*	2371107	0.016	2491069	0.014	11.967%
5	*я*	1634959	0.011	1016649	0.006	47.896%
6	*быть*	1959144	0.014	2196304	0.013	6.063%
7	*он*	3195031	0.022	2299189	0.013	39.701%
8	*с*	1668524	0.012	1854219	0.011	6.881%
9	*что*	1852358	0.013	1825036	0.011	17.442%
10	*а*	1382184	0.010	1154931	0.007	29.983

К модальным относятся модусные слова, выражающие отношение говорящего к истинности вводимой пропозиции, например, *считать*, *полагать*, *знать*, *известно*. Термины с реляционным значением указывают на логико-семантические отношения между компонентами текста, например, *кроме того* (отношение дополнительности), *наоборот* (отношение контрастивности), *потому что*, *так как* (причинно-следственное отношение). Прагматические значения выражаются перформативными компонентами, такими как *обещать поздравлять, требовать, приказывать*. Взятые вне контекста эти термины частично утрачивают значение. Вместе с тем они могут использоваться в текстах, относящимся к разным функционально-стилевым группам и жанрам.

4. Принцип учета морфологической вариативности. С одной стороны, предполагает включение в список стоп-слов парадигмы всех словоформ одной леммы тогда, когда в него вносятся термины, имеющие производные словоформы. Например, если в список добавляется глагол *стать*, то в него должны войти *стал*, *стала*, *стало*, *стали*, *стану*, *станешь*, *станет*, *станете*, *станут*, *стань*, *станьте*, *став*, *ставший*. С другой стороны, в ряде случаев следует учитывать не все, а некоторые словоформы лексемы. Это, прежде всего, относится к терминам, отражающим модальный план текста. В случае с глаголом *считать* имеет смысл ввести в список личные формы глагола, но не инфинитив, который, во-первых, часто используется в омонимичной форме в значении "подсчитывать" (*считать деньги*), во-вторых, неравномерно распределён по художественным и нехудожественным текстам. Разрыв в распределении составляет 56%, что превышает установленный ранее пороговый уровень. То же самое относится к возвратной форме *считается*, которую не следует включать в список, поскольку разрыв в распределении составляет около 80%. Таким образом, словоформы, относящиеся к разным планам текста, необходимо проверять на соответствие принципу равномерного распределения, который является ведущим.

МЕТОДИКА СОЗДАНИЯ СПИСКА СТОП-СЛОВ

Работа над созданием списка русских стоп-слов состояла из нескольких этапов и процедур.

На первом этапе был взят за основу и проанализирован список стоп-слов, составленный М. Портером [9], известным специалистом в области морфологического анализа и автором языка программирования *Snowball*, специально предназначенного для разработки стеммеров. Стеммер Портера и его список стоп-слов включаются в пакеты лингвистического обеспечения для языков программирования *Python* и *R*. Список стоп-слов доступен на сайте автора (http://snowball.tartarus.org/algorithms/russian/stop.txt). Из него мы удалили некоторые слова, а также дубликаты. Например, существительные *жизнь* и *человек*, которые не соответствуют выделенным выше критериям идентификации стоп-слов: статистическому критерию, поскольку не распределены с равномерной частотностью по текстам разных видов и жанров; семантическому критерию, так как выражают значение, будучи взяты вне контекста; морфологическому критерию, так как имеют производные формы. По этим же причинам было удалено наречие *хорошо* и числительное *три*. Включение в списки стоп-слов числительных представляется сомнительным, поскольку они имеют чётко выраженную семантику и не все из них равномерно распределены по типам текстов. В списке Портера мы также исправили термины с орфографическими ошибками. В качестве производных местоимения *она* в списке указаны формы *эи* и *нэи*, под которыми, по-видимому, понимаются *ей* и *ней*. В качестве формы местоимения *этот* указано слово *эты*. Поскольку форма *эти* в списке приводится, вместо *эты* мы ввели форму *эту*, отсутствующую в списке, а также удалили дубликаты, например, слово *все*, которое повторялось три раза. Вместе с тем, добавлена форма *всё*. Буква *ё* использована там, где необходимо, однако были оставлены и формы с *е* вместо *ё,* например, *ее и её*, поскольку в современных текстах *ё* часто заменяется на *е*. В список Портера включены формы перформативов *говорил* (121813

вхождений в НКРЯ), *сказал* (372668 вхождений), *сказала* (109709 вхождений), *говорят* (72845 вхождений), что представляется нами вполне правомерным, так как рассматриваемые формы достаточно частотны и равномерно распределены. К ним мы добавили и другие формы этих глаголов в прошедшем времени. В наш список были включены формы глагола *быть*, которые в списке Портера представлены основой *буд*, по которой можно идентифицировать словоформы *буду, будешь* и остальные. Были введены все формы глаголов *иметь*, *делать*, модальных глаголов и слов *мочь*, *должен*, *надо*, *нужно*, *возможно*, *вероятно*, *видимо*, а также вводных слов, таких как *вообще*, *однако*. В результате, в модифицированном списке оказалось 312 терминов. Отметим, что некоторые вводные и модальные слова в список не включались в случае омонимии. Например, *чай*, *часом* более частотны в качестве существительных, поэтому добавлять их в список стоп-слов нецелесообразно. Список стоп-слов, составленный нами на основе списка Портера, приводится в *Приложении А*.

На следующем этапе получившийся список сопоставили с двумя другими списками: списком, составленным профессором Невшательского университета Ж. Савой (http://members.unine.ch/jacques.savoy/clef/russianST.txt), и списком, доступным на платформе *GitHub* – среды для разработки программного обеспечения (https://raw.githubusercontent.com/ stopwords-iso/stopwords-ru/master/stopwords-ru.txt). Список *Савой* содержит 421 слово, список *GitHub* – 559 слов. Полагаем, что в настоящее время это наиболее авторитетные источники. Ж. Савой – известный специалист в области информационного поиска, на его сайте [10] доступны списки стоп-слов, составленные им и для других языков. Платформой *GitHub* пользуются более 73 млн разработчиков и 4 тыс. организаций. Там также можно найти списки стоп-слов для разных языков. Методом, с помощью которого проводилось сопоставление, было выполнение пересечения и разницы списков, в результате чего находились слова, встречающиеся в обоих списках и только в одном из них. Пересечение и разница выявлялись автоматически с помощью нашего приложения *Y-sets* (https://iatskota.wixsite.com/yatsko/y-sets-application). Из списка *GitHub* были добавлены притяжательные и относительные местоимения *мой*, *твой*, *наш*, *который* и их производные, которых не оказалось в списке Портера, а также ряд наречий, не имеющих производных форм. Нами не были включены такие существительные, как *лицо*, *люди*, *вода*, *нога, ночь*, а также *Россия*, *русский*. Эти слова достаточно частотны, однако неравномерно распределены по жанрам и выражают устойчивые значения. Их фильтрация может привести к ошибкам в интерпретации текста и искажению его смысла. В списке *GitHub*, как и в списке Савой, и некоторых других, были числительные, причем только в форме именительного падежа единственного числа. Они также нами проигнорированы, так как не соответствуют семантическому признаку стоп-слов и имеют производные формы. Далее, дополненный список мы сопоставили со списком Савой, из которого взяты 16 слов, а также формы глаголов *стать* и *становиться*, используемые, в основном, в качестве глаголов-связок. К получившемуся списку добавлены некоторые вводные слова и выражения, взятые из списка на сайте gramota.ru (http://new.gramota.ru/spravka/punctum/punctum-alphabet).

Современные токенайзеры разбивают сложные слова с написанием через дефис на соответствующие компоненты, поэтому в список мы включили компоненты сложных стоп-слов в качестве отдельных терминов, т. е. вместо *кто-нибудь* и *по-моему* в списке присутствуют как отдельные термины *кто*, *нибудь*, *по*, *моему*. Получившийся список стоп-слов, включающий 535 словоформ, приводится в *Приложении Б*.

ВЫВОДЫ

В нашей работе была предпринята попытка сформулировать принципы создания списка стоп-слов, а также применить эти принципы в процессе разработки списка стоп-слов русского языка. Полагаем, что их можно использовать и при создании списков стоп-слов для других языков. При этом следует учитывать, что ведущим принципом является принцип равномерности жанрово-стилистического распределения, соотносящийся со статистическими данными о частотностях слов в репрезентативном корпусе. Принципы учёта многоплановости и морфологической вариативности предполагают анализ равномерности распределения словоформ по разным жанрам.

Очевидно, что качество создаваемого списка непосредственно зависит от репрезентативности корпуса текстов, на основе которого получаются статистические данные. Нами использовался *Национальный корпус русского языка* – ценный, однако, как необходимо с сожалением отметить, недостаточно репрезентативный ресурс. Основной раздел этого корпуса включает 337 025 184 словоформы, в то время как *Corpus of Contemporary American English* – более миллиарда словоформ, и такой объём является стандартным требованием к современным национальным корпусам. Что касается русского языка, то в силу морфологической развитости объем корпуса должен быть ещё больше: на основе предложенной в настоящей работе методики – 1 млрд 250 млн словоформ. Создание эталонного репрезентативного корпуса русского языка крайне важно для развития предметной области, связанной с автоматической обработкой текстов на естественном языке. Сейчас это особенно актуально, поскольку в связи с санкциями закрыт доступ к некоторым ресурсам, в том числе и по корпусной лингвистике. На момент написания нашей статьи недоступен портал https://www.sketchengine.eu/, с которого предоставляется доступ к различным корпусам текстов. При попытке открытия портала выводится сообщение *Until further notice any access to Sketh Engine services is not permitted from Russia and Russian occupied territories*.

Нами был создан универсальный список стоп-слов, который может применяться в различных направлениях компьютерной лингвистики. Вместе с тем, следует учитывать, что при реализации конкретных проектов потребуется его модификация. Современные информационно-поисковые системы используют достаточно небольшие списки стоп-слов, поэтому в целях информационного поиска этот список предлагается сократить, оставив только наиболее типичные стоп-

слова (предлоги, союзы, частицы, местоимения), либо использовать модифицированный список Портера. Вероятно, что для проектов, связанных с анализом узких предметных областей, потребуется удаление некоторых глагольных форм или наречий. В любом случае, предложенный список, как мы надеемся, можно использовать в качестве основы исследований и разработок, посвящённых автоматическому анализу текстовых документов.

СПИСОК ЛИТЕРАТУРЫ

1. Francis W.N., Kucera H. Computational analysis of present day American English. – Providence: Brown University Press, 1967. – 424 p.
2. Roelleke T. Wang U. TF-IDF uncovered: A study of theories and probabilities // ACM SIGIR'08. – Singapore, 2008. – P. 435-442. – URL: http://www.eecs.qmul.ac.uk/~thor/2008/TF-IDF-Uncovered-SIGIR-Talk.pdf.
3. Yatsko V.A. TF*IDF revisited // International journal of computational linguistics and natural language processing. – 2013. – Vol.2, Iss. 6. – P. 385-387. – URL: https://docs.google.com/file/d/0B306nMx7wiLyZ0tFelo4MzY5SWc/edit.
4. Яцко В.А. Новый метод автоматической классификации текстовых документов // Научно-техническая информация. Сер 2. – 2021. – № 6. – С. 27-38; Yatsko V.A. A New Method of Automatic Text Document Classification. – Automatic Documentation and Mathematical Linguistics. – 2021. – Vol. 55, № 3. – P. 122-133.
5. Fox C. A stop list for general text // ACM SIGIR forum. – 1989. – Vol. 24, Iss. 1-2. – P. 19-21. – URL: https://dl.acm.org/doi/pdf/10.1145/378881.378888.
6. Ляшевская О.Н, Шаров С.А. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). – Москва: Азбуковник, 2009. – URL: http://dict.ruslang.ru/freq.php?act=show&dic=freq_freq&title=%D7%E0%F1%F2%EE%F2%ED%FB%E9%20%F1%EF%E8%F1%EE%EA%20%EB%E5%EC%EC
7. Zipf's and Heap's law/Northeastern university. Khoury college of computer sciences. – 2009. – URL: https://www.ccs.neu.edu/home/ekanou/ISU535.09X2/Handouts/Review_Material/ zipfslaw.pdf
8. Яцко В.А. Рассуждение как тип научной речи. – Абакан: Изд-во Хагасского гос. ун-та, 1998. – 182 с.
9. Porter M. The Porter stemming algorithm. – 2006. – URL: https://tartarus.org/martin/PorterStemmer/
10. Savoy J. IR multilingual resources at UniNE. – 2005. – URL: http://members.unine.ch/jacques.savoy/ clef/

Приложение А.

Модифицированный список Портера, 312 терминов

а, без, бесспорно, более, больше, будем, будет, будете, будешь, будто, буду, будут, будучи, будь, бы, бывало, был, была, были, было, быть, вам, вами, вас, вдруг, ведь, вероятно, весь, видимо, видимому, во, возможно, вон, вообще, вот, впрочем, вряд, все, всё, всегда, всего, всей, всем, всеми, всему, всех, всею, всю, вся, вы, где, говорил, говорила, говорили, говорят, да, даже, два, для, до, должен, должна, должны, другой, его, едва, ее, её, ей, ему, если, есть, еще, ею, ж, же, за, зачем, здесь, и, из, или, им, имевший, имеем, имеемый, имеет, имеете, имеешь, имей, имейте, имел, имела, имели, имело, именно, иметь, имею, имеют, имеющий, имея, ими, иногда, их, к, кажется, кажись, казалось, как, какая, какой, когда, конечно, короче, который, кроме, кстати, кто, куда, ли, лишь, лучше, между, меня, мне, много, мной, мною, мог, моги, могите, могла, могли, могло, могу, могут, моему, можем, может, можете, можешь, можно, мой, мочь, моя, мы, на, над, надо, наконец, нам, нами, наоборот, например, напротив, нас, нашему, не, него, нее, неё, ней, ней, нельзя, нему, несомненно, нет, нею, ни, нибудь, никогда, ним, ними, них, ничего, но, ну, нужно, о, об, обычно, один, однако, он, она, они, оно, опять, от, откуда, отчего, очевидно, перед, по, под, поди, пожалуй, позволь, позвольте, помилуй, помилуйте, помимо, после, потом, потому, похоже, почему, почти, при, примерно, про, прочим, проще, раз, разве, разумеется, с, сам, сама, сами, самим, самими, самих, само, самого, самой, самом, самому, самою, саму, свою, себе, себя, сегодня, сейчас, сказал, сказала, сказали, сказать, сколько, скорее, следовательно, случайно, смог, смоги, смогите, смогла, смогли, смогут, сможем, сможете, сможешь, со, собой, собою, собственно, совсем, судя, та, так, такой, там, твоему, те, тебе, тебя, тем, теми, теперь, тех, то, тобой, тобою, тогда, того, тоже, той, только, том, тому, тот, точнее, тою, ту, тут, ты, у, уж, уже, хоть, чего, чей, чем, через, что, чтоб, чтобы, чуть, эта, эти, этим, этими, этих, это, этого, этой, этом, этому, этот, этою, эту, я.

Приложение Б.

Полный список стоп-слов русского языка, 535 терминов

а, алло, без, безусловно, бесспорно, близко, более, больше, будем, будет, будете, будешь, будто, буду, будут, будучи, будь, бы, бывает, бывало, был, была, были, было, быть, в, вам, вами, вас, ваш, ваша, ваше, вашему, ваши, вверх, вверху, ввиду, вдруг, ведь, везде, вернее, верно, вероятнее, вероятно, весь, видимо, видимому, видишь, видно, вне, вниз, внизу, внутри, во, возможно, вокруг, вон, вообще, вот, вперёд, вперед, впрочем, вряд, все, всё, всегда, всего, всей, всем, всеми, всему, всех, всею, всю, всюду, вся, вы, г, где, говорил, говорила, говорили, говорит, говоря, говорят, да, давно, даже, далеко, дальше, данным, действительно, дело, для, до, довольно, должен, должна, должно, должны, допустим, другая, другие, других, другого, другое, другой, е, ё, его, едва, ее, её, ей, ему, если, есть, еще, ещё, ею, ж, же, жаль, за, затем, зато, зачем, здесь, значит, знаете, знаешь, значит, и, из, известно, или, им, имевший, имеем, имеемый, имеет, имеете, имеешь, имей, имейте, имел, имела, имели, имело, именно, иметь, имею, имеют, имеющий, имея, ими, иначе, иногда, их, к, каждая, каждое, каждые, каждый, кажется, кажись, казалось, как, какая, какой,

кем, когда, кого, ком, кому, конечно, короче, которая, которого, которой, которые, который, которых, кроме, кругом, кстати, кто, куда, ли, лишь, лучше, м, мало, между, меня, мимо, мне, мнению, много, мной, мною, мог, моги, могите, могла, могли, могло, могу, могут, мое, моему, моё, можем, может, можете, можешь, можно, мои, мой, мочь, моя, мы, на, наверно, наверное, наверху, над, надо, назад, наконец, нам, нами, наоборот, например, напротив, нарочно, нас, наш, наша, наше, нашему, наши, не, него, недалеко, недавно, нее, неё, ней, некоторая, некоторого, некоторой, некотором, некоторому, некоторую, некоторые, некоторый, некоторым, некоторых, некоторых, нельзя, нем, нём, нему, несомненно, нет, нею, ни, нибудь, низко, никакой, никогда, никого, никому, никто, никуда, ним, ними, них, ничего, ничему, ничто, но, ну, нужно, нх, о, об, оба, обе, обеим, обеими, обеих, обоим, обоими, обоих, обычно, общем, однако, однажды, оказывается, около, он, она, они, оно, опять, особенно, от, отовсюду, отсюда, отчего, очевидно, очень, перед, по, под, поди, пожалуй, пожалуйста, позади, позволь, позвольте, позже, пока, помилуй, помилуйте, помимо, понятно, попросту, пор, пора, поскольку, после, посреди, потом, потому, похоже, почему, почти, поэтому, при, примерно, примеру, про, просто, против, прочим, прочь, проще, раз, разве, разумеется, раньше, с, сам, сама, сами, самим, самими, самих, само, самого, самой, самом, самому, самою, саму, самый, свое, своё, своего, своей, свои, своих, свой, свою, своя, себе, себя, сегодня, сейчас, сзади, сих, скажем, сказал, сказала, сказали, сказать, сколько, скольку, скорее, следовательно, слишком, случайно, слушай, слушайте, смог, смоги, смогите, смогла, смогли, смогло, смогут, сможем, сможете, сможешь, смотри, смотрите, сначала, снова, со, собой, собою, собственно, совсем, соответственно, спасибо, сразу, став, ставший, стал, стала, стали, стало, становимся, становитесь, становится, становиться, становишься, становлюсь, становятся, судя, суть, сути, сущности, считаю, считал, считаешь, считали, считаем, считаете, считает, считают, сюда, т, та, так, такая, также, такие, таким, такими, таких, такого, такое, такой, такому, там, твое, твоё, твоей, твоему, твои, твоим, твоих, твой, твоя, те, тебе, тебя, тем, теми, теперь, тех, то, тобой, тобою, тогда, того, тоже, той, только, том, тому, тот, точнее, тою, ту, туда, тут, ты, у, уж, уже, хоть, хотя, хочу, хочешь, хотите, хочет, хотим, хотят, частности, часто, чего, чей, чем, чему, через, что, чтоб, чтобы, чуть, эта, эти, этим, этими, этих, это, этого, этой, этом, этому, этот, этою, эту, я, являемся, являетесь, является, являешься, являюсь, являются, якобы.

Материал поступил в редакцию 04.04.22.

Сведения об авторе

ЯЦКО Вячеслав Александрович – доктор филологических наук, профессор Хакасского государственного университета им. Н.Ф. Катанова, г. Абакан
e-mail: viatcheslav-yatsko@rambler.ru