УДК 81′33: 519.76
DOI 10.25205/1818-7900-2020-18-1-74-82

Автоматическая обработка текстов на основе платформы TXM с учетом анализа структурных единиц текста

Ф. Н. Соловьев

Федеральный исследовательский центр «Информатика и управление» РАН
Москва, *Россия*

Аннотация

В настоящей работе мы приводим описание интеграции средств автоматической обработки текста (выделение псевдооснов, именных групп, анализ глагольного управления) для расширения аналитических возможностей платформы корпусного анализа TXM. Представленные средства объединены в единый программный комплекс, что позволяет эффективно осуществлять подготовку специальных корпусов для последующего анализа средствами платформы TXM.

Ключевые слова

автоматический анализ текстов, платформа TXM, псевдоосновы, именные группы, глагольное управление

Благодарности

Работа выполнена при финансовой поддержке РФФИ в рамках научного проекта № 19-07-00806

Для цитирования

Соловьев Ф. Н. Автоматическая обработка текстов на основе платформы TXM с учетом анализа структурных единиц текста // Вестник НГУ. Серия: Информационные технологии. 2020. Т. 18, № 1. С. 74–82. DOI 10.25205/1818-7900-2020-18-1-74-82

Embedding Additional Natural Language Processing Tools into the TXM Platform

F. N. Soloviev

Federal Research Center "Informatics and Management"
Moscow, *Russian Federation*

Abstract

In our work we present a description of integration of natural language processing tools (pseudostem extraction, noun phrase extraction, verb government analysis) in order to extend analytic facilities of the TXM corpora analysis platform. The tools introduced in the paper are combined into a single software package providing TXM platform with an effective specialized corpora preparation tool for further analysis.

Keywords

natural language processing, TXM platform, pseudostems, noun phrases, verb government

Acknowledgements

The research was supported by the RFBR, grant no. 19-07-00806

For citation

Soloviev F. N. Embedding Additional Natural Language Processing Tools into the TXM Platform. *Vestnik NSU. Series: Information Technologies*, 2020, vol. 18, no. 1, p. 74–82. (in Russ.) DOI 10.25205/1818-7900-2020-18-1-74-82

Введение

Задачи автоматической обработки текстов, например классификация и выделение нетривиальных языковых конструкций, в настоящее время представляют серьезный интерес как для научного сообщества, так и для решения практических задач, и были, в частности, исследованы нами в [1; 2]

В настоящей работе в контексте решения указанных задач мы опираемся на программный комплекс – платформу ТХМ (http://textometrie.org). Платформа ТХМ является эффективным средством корпусного анализа, позволяющим проводить комплексный анализ корпусов (анализ соответствий, кластеризация, построение лексических таблиц, поиск сложных лексических конструкций, выделение подкорпусов по различным параметрам). Платформа ТХМ интегрирована с расширением TreeTagger [3], позволяющим проводить лишь морфологический анализ и лемматизацию словоупотреблений. Она использует словоупотребления в качестве структурных единиц анализа.

Для повышения эффективности таких используемых ТХМ методов, как анализ специфичности и анализ соответствий, целесообразно ввести в рассмотрение новые единицы анализа, опирающиеся на процедуры автоматизированной обработки текстов на естественных языках, описанные в [4].

Мы предлагаем ряд расширений, позволяющих дополнить и усложнить анализ корпусов, включающий автоматический морфологический анализ словоформ и приведение их к канонической форме, выделение псевдооснов, выделение именных и глагольных групп и комбинирование результатов работы предлагаемых расширений. Конечной целью дополнений к платформе ТХМ является создание механизмов для исследования применимости различных дифференцирующих признаков при решении задачи классификации текстов и создания тематических корпусов текстов.

В [5–8] мы провели эксперименты по использованию псевдооснов и именных групп для выявления экстремистской направленности текстов. В данной работе к этим характеристикам добавлены возможности учета глагольного управления.

Псевдоосновы

Для определения дифференцирующих признаков коротких текстов сети Интернет, характеризующихся особыми тематическими и психолингвистическими свойствами, содержащих неологизмы и жаргонизмы, большой интерес представляет использование аналитического метода выделения псевдооснов, так как он позволяет обрабатывать отсутствующие в стандартных словарях формы.

Используемый способ выделения псевдооснов представляет собой метод структурных схем, описанный подробно в [9]. Суть метода состоит в получении псевдоосновы словоформы путем рассмотрения ее словоизменительных аффиксов. Словообразовательные аффиксы считаются в рамках этого метода элементом корневой части и не отбрасываются. Далее под аффиксами мы будем понимать исключительно словоизменительные аффиксы. С каждым словом можно сопоставить отвечающую ему последовательность аффиксов. Такие последовательности называются структурами некорневой части слова. Отсюда происходит название метода. Как и в традиционном морфологическом анализе, аффиксы подразделяются на префиксы и суффиксы в соответствии с их позицией относительно корня слова. Псевдоосновой называется часть слова, не содержащая суффиксов и префиксов. Способ автоматического выделения псевдооснов состоит в сопоставлении рассматриваемой словоформы с множеством допустимых в языке структур некорневой части слова. Псевдооснова слова выделяется отбрасыванием всех соответствующих определенной структурной схеме аффиксов (т. е. допустимой в данном языке максимальной комбинации префиксов и суффиксов). У глаголов, в частности, отбрасываются показатели лица, числа, рода, времени, причастной формы. Видовые префиксы не отбрасываются, так как они могут влиять на лексическое значение слова.

Псевдооснова не всегда совпадает с основой слова в традиционном понимании. Например, в словоформе *людьми* единственным аффиксом, который можно отбросить согласно продуктивной структурной схеме, является *-и*, поэтому выделяется псевдооснова *людьм*.

Данный подход позволяет анализировать текстовые конструкции, опираясь не только на точные словоформы, и тем самым повышает полноту и гибкость корпусного анализа.

Морфологические характеристики

Возможность привести словоформу к канонической форме позволяет анализировать различные элементы словоизменительной парадигмы как одну и ту же структурную единицу текста. Это, в свою очередь, позволяет более корректно проводить содержательный статистический анализ текста, например, путем рассмотрения частот лексем вместо частот отдельных словоформ.

При предобработке всех русскоязычных текстов мы осуществляем автоматический морфологический анализ словоформ на основе словарной компьютерной морфологии, описанной в [4]. Используемая стандартная в отечественной компьютерной лингвистике морфологическая модель относит каждое слово к одному из 24 морфологических классов, включающих, помимо частей речи в традиционном понимании, такие разряды, как «неизменяемое слово», «аббревиатура», «топоним». Каждый из этих морфологических классов характеризуется набором грамматических характеристик: род, падеж, число, наклонение и др. В программной реализации словарной морфологии русского языка применяется специализированная структура данных, позволяющая осуществлять поиск словоформ за линейное по числу букв словоформы время. Каждая словоформа содержит свои грамматические характеристики и ее каноническую (начальную) форму.

В настоящей работе мы также использовали интегрированный в TXM программный пакет TreeTagger [3], предоставляющий возможность совместного морфологического анализа слов предложения на основе статистической модели путем сопоставления словоупотреблений, снабженных специальными метками, кодирующими морфологические характеристики. Преимуществом данной процедуры разметки является однозначность морфологического анализа, но при таком анализе существует риск ошибок, который возрастает, если текст содержит большое количество неологизмов и нестандартных написаний слов. Все виды морфологической разметки использовались в дальнейшем для сопоставительного анализа текстов корпуса.

Дополнительную информацию о специфическом содержании текста можно почерпнуть, анализируя не только словоформы, но и именные группы и глагольные группы целиком. В отличие от отдельных слов, выделенные именные и глагольные группы несут информацию о конкретных отдельных аспектах содержания текста.

Выделение именных групп

Именная группа определяется нами как группа слов, у которой главное слово существительное, а другие слова связаны с ним подчинительными синтаксическими связями. Рассмотрение частотных именных групп и их сочетаний в совокупности с анализом отдельных словоупотреблений позволяет получить более полную картину семантических и стилистических характеристик текста, релевантных его содержанию.

Определенную трудность при выделении именных групп представляет разрешение омонимической неопределенности, проистекающей из множественности морфологических разборов отдельных словоупотреблений, которая, как правило, имеет место. Наш метод выделения именных групп предполагает рассмотрение всего множества возможных морфологических разборов каждого слова.

Используемый нами алгоритм подробно описан в [4]. Алгоритм состоит из трех этапов: установление подчинительных синтаксических связей в предложении между парами слов; установление синтаксических связей внутри конструкций с однородными членами; выделение именных групп как цепочки последовательно связанных подчинительными связями слов.

Приведем вкратце основные моменты и определения, относящиеся к алгоритму выделения именных групп. В нашем изложении будем двигаться от простых структурных единиц к сложным: от отдельных словоупотреблений к именным группам.

Словоупотребление *w* можно рассмотреть как список результатов морфологического анализа, объединенных в группы по частям речи. Введем отображение $MA(w)$ из множества словоупотреблений во множество вариантов морфологического разбора:

$$MA: w \to (G_1, ..., G_n),$$

где конкретный морфологический разбор словоупотребления *w*:

$$G_i = (GI_1, ..., GI_k), \quad GI_j = (pos, gc, gender, number),$$

$$pos \in PoS, \quad gc \in GC, \quad gender \in Genders, \quad number \in Numbers,$$

где

PoS – множество частей речи русского языка;

GC – множество падежей русского языка;

$Genders$ – множество родов русского языка;

$Numbers$ – множество показателей числа;

GI_k – вариант морфологического разбора (морфологический разбор);

G_i – группа морфологических разборов.

Групп морфологических разборов G_i, вообще говоря, может оказаться несколько из-за явления омонимии в русском языке. Результаты морфологического анализа словоупотребления *w* могут быть представлены в разгруппированном виде, в качестве плоского списка вариантов морфологического разбора: $MA(w) = (GI_1, ..., GI_p)$.

Введем подчинительную синтаксическую связь как пару $(w, RL(w))$, где *w* – словоупотребление, а $RL(w)$ – список подчиненных ему словоупотреблений. Обозначим за *R* множество подчинительных синтаксических связей.

Входными данными для алгоритма выделения именных групп являются предложения *Sent*, состоящие из словоупотреблений: $Sent = w_1, ..., w_n,$, где $w_i \in \Sigma^+$, при этом множество Σ – алфавит русского языка. Алгоритм состоит из трех этапов.

1. Определение подчинительных синтаксических связей в паре (lw, rw), *lw* – левое слово, *rw* – правое слово в паре слов. Например, если есть предложение *Sent*, состоящее из слов w_1, w_2, w_3, то пара (w_1, w_2) может образовать синтаксическую связь, в которой $lw = w_1$, $rw = w_2$.
2. Установление синтаксических связей внутри конструкций с однородными членами.
3. Выделение именных групп.

На множестве подчинительных синтаксических связей *R* как на ребрах (дугах) может быть построен ненаправленный граф $G = (V, E)$, $E = E_q \cup E_u$. Вершины этого графа – слова предложения, дуги – подчинительные синтаксические связи. Различают два типа ребер: *q*-ребра и *u*-ребра.

Если

$$e_{q_l} = \left(v_i, v_j\right),$$

то

$$\exists r_p \in R : v_j \in RL\left(v_i\right),$$

если

$$e_{u_l} = \left(v_i, v_j\right),$$

то

$$\exists r_p \in R : v_j \in RL\left(v_i\right),\ pos_{v_j} = ADJ.$$

Построенный граф служит выделению именных групп: в таком графе именная группа – это путь $p = \left(v_1, ..., v_k\right)$, удовлетворяющий одному из следующих условий:

1) если $k = 2$, то $e = \left(v_1, v_2\right)$, $e \in E_q$;

2) если $k > 2$, то $\exists e_i \in E_u$ и $\exists e_j \in E_q$.

Множество $NG = \left\{p_1, ..., p_f\right\}$ путей, полученных в результате работы алгоритма, будет множеством именных групп предложения.

Выделение глагольных групп

Выделение глагольных групп (словосочетаний, главным словом которых является глагол), т. е. установление связей выделенных именных групп с глаголами, представляет важную, необходимую составляющую синтаксического анализа предложения. Данная задача решается нами при помощи анализа глагольного управления.

Глагольным управлением называется разновидность синтаксической подчинительной связи типа управления, в которой главным словом является глагол. Как и в связи управления вообще, при глагольном управлении главным словом (глаголом) накладываются ограничения на употребление зависимого словосочетания в виде набора вариантов допустимых комбинаций грамматических характеристик зависимого словосочетания и, возможно, необходимых служебных частей речи. Такие комбинации мы в соответствующем контексте называем для удобства просто *ограничениями.* Указанные ограничения напрямую зависят от семантических свойств главного слова.

Пусть имеется множество глаголов A и a, типичный представитель этого множества – глагол.

С каждым глаголом может быть сопоставлен набор ограничений, накладываемых им на зависимые словосочетания. Такой набор мы называем *парадигмой управления* данного глагола или *парадигмой глагольного управления*, если мы отвлекаемся от самого глагола и рассматриваем набор его ограничений сам по себе. Обозначим как C_a парадигму управления глагола a.

Таким образом, каждую конкретную глагольную группу vg_a можно представить в виде тройки $vg_a = \left(a, p, c_a\right)$ – глагола a, выступающего в качестве главного слова, зависимого словосочетания p и представителя c_a парадигмы управления C_a глагола a, которому удовлетворяет рассматриваемое зависимое словосочетание p.

В нашей работе мы занимаемся выделением только таких глагольных групп, в которых зависимым словосочетанием является именная группа.

В русском языке ограничения на зависимые словосочетания указанного типа имеют простое общее представление в виде требования определенного предлога и падежа следующего за предлогом словосочетания либо требования определенного падежа и отсутствия предлога.

Обозначим множество предлогов как S, а его элементы – s, в которое для удобства включим специальный элемент s_0, обозначающий «пустой» предлог, или отсутствие предлога. Множество падежей мы ввели выше как GC. Тогда рассматриваемые ограничения – это пары (s, gc).

Так, к примеру, словосочетание *сделал в последний момент* может быть представлено как глагол *сделал*, зависимое словосочетание *в последний момент* и ограничение употребления словосочетания, состоящее из служебной части речи – предлога *в*, и ограничения на грамматическую характеристику падежа – винительный падеж.

Другой пример: *написал статью*. Здесь главное слово – *написал*, зависимое слово – *статью*, ограничение требует отсутствия предлога и винительного падежа.

Отдельно заметим, что поскольку природа ограничений носит характер, связанный со значением глагола, то сами ограничения не зависят от морфологических характеристик употребления глагола. Из указанного следует, что решение задачи выделения глагольных групп посредством анализа ограничений глагольного управления требует наличия словарей глагольного управления: особых словарей, приводящих глаголы вместе с их парадигмами глагольного управления.

Формально такой словарь можно представить в виде отображения

$$F_c : a \mapsto C_a.$$

Результаты морфологического анализа и процедуры выделения именных групп позволяют, используя словарь глагольного управления, выявить синтаксические связи для определения глагольных групп. Выделение глагольных групп в предложении осуществляется путем анализа всех возможных пар (глагол, именная группа) предложения на предмет соответствия именной группы парадигме управления соответствующего глагола, а именно поиска в парадигме ограничения, которому удовлетворяет рассматриваемая именная группа. Если такая группа находится, принимается решение о наличии связи управления между глаголом и именной группой. Если же в парадигме нет ни одного ограничения, которому бы удовлетворяла рассматриваемая именная группа, то принимается решение об отсутствии связи управления между глаголом и именной группой.

Более формально алгоритм выделения глагольных групп можно описать следующим образом. Пусть в предложении *Sent* содержатся глаголы $A_{Sent} = \{a \in A | a \in Sent\}$, и именные группы $P_{Sent} = \{p | p \in Sent\}$. Пусть имеется словарь (отображение) глагольного управления F_c. Пусть F_{gc} – отображение, определяющее множество возможных падежей именной группы:

$$F_{gc} : p \mapsto \{gc_1, ..., gc_k\}.$$

Тогда множество VG_{Sent} глагольных групп может быть определено как

$$VG_{Sent} = \left\{ (a, p, c_a) \middle| c_a \in F_c(a), c_a = (s, gc), gc \in F_{gc}(p), s \wedge p \in Sent \right\},$$

где выражение $s \wedge p$ обозначает предлог s с последующей именной группой p.

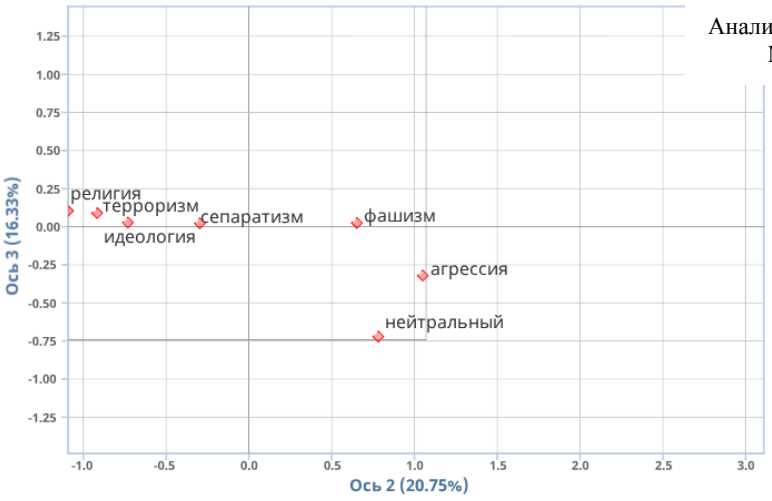
Для анализа глагольного управления был использован электронный словарь глагольного управления, в который вошли первые две тысячи наиболее частотных глаголов русского языка по материалам Национального корпуса русского языка (ruscorpora.ru). Словарь глагольного управления содержит парадигмы глагольного управления, состоящие из ограничений употреблений именных групп вида *предлог + падеж* или *отсутствие предлога + падеж*.

Анализ подкорпусов

Удобным инструментом количественной оценки «необычности» специального подкорпуса относительно всего корпуса является показатель специфичности [10]. Анализ специфичности позволяет составить своего рода «профиль» подкорпуса, выделенного на каких-либо внешних основаниях (например, автор, жанр, тематика или идеологическая направленность текста), путем выявления наиболее характерных или нехарактерных для него словоформ (лексем, псевдооснов, именных и глагольных групп и т. п.). Этот «профиль» может быть использован для диагностики нового текста.

Другим подходом к анализу разделенного на части (подкорпуса) по определенному критерию корпуса является анализ соответствий. Методика анализа соответствий, используемая TXM, была предложена Ж.-П. Бензекри [11] и имплементирована в пакете FactoMineR для платформы R [12]. Анализ соответствий демонстрирует взаимную «близость» или «удаленность» подкорпусов на основе анализа частот совместного появления значений переменных (словоформ, начальных форм, псевдооснов, именных групп, морфологических тегов и т. д.).

Экспериментальный корпус был проанализирован с использованием двух обозначенных выше функций TXM – специфичность и анализ соответствий. Детально были рассмотрены следующие лексические объекты: словоформы; начальные формы слов, полученные по словарной морфологии; начальные формы слов с морфологическими характеристиками, полученные с помощью TreeTagger; псевдоосновы слов; именные группы, составленные из словоформ; именные группы, составленные из начальных форм; именные группы, составленные из псевдооснов вместо отдельных словоупотреблений; глагольные группы.



Анализ соответствий для глагольных групп
Matching analysis for verb groups

На рисунке приведен пример применения анализа соответствий для глагольных групп, в рамках которых слова заменены на их псевдоосновы. Анализируемый корпус состоит из подкорпусов агрессивной, идеологической, националистической, фашистской, религиозной, сепаратистской, террористической и нейтральной направленностей. При делении текстов на подкорпуса есть возможность интерпретировать близость или разделенность значений рассматриваемых характеристик подкорпусов относительно друг друга как оценку, указывающую на сходство или различие маркированных подкорпусов между собой и по отношению к «нейтральному» подкорпусу.

Различные выделяемые дифференцирующие признаки, описанные выше, демонстрируют схожую разделяющую способность.

Заключение

Проведенная работа по интеграции инструментов автоматической обработки текста и платформы корпусного анализа ТХМ показала, что такая интеграция позволяет расширить возможности статистического анализа текстов.

Детально рассмотрены такие лексические объекты, как леммы, псевдоосновы, именные и глагольные группы различной структуры. Упомянутые средства были объединены в набор утилит, позволяющих вычислять для текстовых корпусов ряд характеристик языковых единиц, входящих в их состав. Корпуса с вычисленными характеристиками преобразуются нами в формат для импорта пакетом ТХМ.

В силу выявленных особенностей и противопоставленности нейтрального подкорпуса остальным сформированный корпус может быть использован для машинного обучения в задачах классификации текстов на предмет выявления заданного содержания с целью их углубленного экспертного анализа.

Список литературы

1. **Поляков И. В., Соловьев Ф. Н., Чеповский А. А., Чеповский А. М.** Задача распознавания для текстов на естественных языках. М.: Национальный открытый университет «ИНТУИТ», 2017.
2. **Поляков И. В., Соколова Т. В., Чеповский А. А., Чеповский А. М.** Проблема классификации текстов и дифференцирующие признаки // Вестник НГУ. Серия: Информационные технологии. 2015. Т. 13, № 2. С. 55–63.
3. **Schmid H.** Probabilistic Part-of-Speech Tagging Using Decision Trees. In: Proceedings of International Conference on New Methods in Language Processing. Manchester, UK, 1994.
4. **Чеповский А. М.** Информационные модели в задачах обработки текстов на естественных языках. Второе издание, переработанное. М.: Национальный открытый университет «ИНТУИТ», 2015.
5. **Ананьева М. И., Девяткин Д. А., Кобозева М. В., Смирнов И. В., Соловьев Ф. Н., Чеповский А. М.** Исследование характеристик текстов противоправного содержания // Тр. Ин-та системного анализа РАН. 2017. Т. 67, № 3. С. 86–97.
6. **Лаврентьев А. М., Смирнов И. В., Соловьев Ф. Н., Суворова М. И., Фокина А. И., Чеповский А. М.** Создание специальных корпусов текстов на основе расширенной платформы ТХМ // Системы высокой доступности. 2018. Т. 14, № 3. С. 76–81.
7. **Лаврентьев А. М., Соловьев Ф. Н., Суворова М. И., Фокина А. И., Чеповский А. М.** Новый комплекс инструментов автоматической обработки текста для платформы ТХМ и его апробация на корпусе для анализа экстремистских текстов // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2018. Т. 16, № 3. С. 19–31.
8. **Chepovskiy A., Devyatkin D., Smirnov I., Ananyeva M., Kobozeva M., Solovyev F.** Exploring linguistic features for extremist texts detection (on the material of Russian-speaking illegal texts). In: 2017 IEEE International Conference on Intelligence and Security Informatics: Security and Big Data, ISI 2017. Institute of Electrical and Electronics Engineers Inc., 2017, p. 188–190.
9. **Egorova E., Chepovskiy A., Lavrentiev A.** A structural pattern based method for automated morphological analysis of word forms in a natural language. *Journal of Mathematical Sciences*, 2016, vol. 214, no. 6, p. 802–813.
10. **Lafon P.** Sur la variabilité de la fréquence des formes dans un corpus. *Mots*, 1980, no. 1, p. 127–165.
11. **Benzécri J.-P.** L'analyse des données: l'analyse des Correspondances. 2nd ed. Paris, Dunod, 1979, vol. 2.
12. **Lê S., Josse J., Husson F.** FactoMineR: An R package for multivariate analysis. *Journal of statistical software*, 2008, no. 25 (1), p. 1–18.

References

1. **Polyakov I. V., Soloviev F. N., Chepovskiy A. A., Chepovskiy A. M.** Zadacha raspoznavaniya dlya tekstov na yestestvennyh yazykah. Moscow, Natsional'nyy otkrytyy universitet "INTUIT", 2017. (in Russ.)
2. **Polyakov I. V., Sokolova T. V., Chepovskiy A. A., Chepovskiy A. M.** Problema klassifikatsii tekstov i differentsiruyushchiye priznaki. *Vestnik NSU. Series: Information Technologies*, 2015, vol. 13, no. 2, p. 55–63. (in Russ.)
3. **Schmid H.** Probabilistic Part-of-Speech Tagging Using Decision Trees. In: Proceedings of International Conference on New Methods in Language Processing. Manchester, UK, 1994.
4. **Chepovskiy A. M.** Informatsionnyye modeli v zadachah obrabotki tekstov na yestestvennyh yazykah. 2nd ed. Moscow, Natsional'nyy otkrytyy universitet "INTUIT", 2015. (in Russ.)
5. **Ananieva M. I., Devyatkin D. A., Kobozeva M. V., Smirnov I. V., Soloviev F. N., Chepovskiy A. M.** Issledovaniye harakteristik tekstov protivopravnogo soderzhaniya. *Trudy Instituta sistemnogo analiza Rossiyskoy akademii nauk*, 2017, vol. 67, no. 3, p. 86–97. (in Russ.)
6. **Lavrentiev A. M., Smirnov I. V., Soloviev F. N., Suvorova M. I., Fokina A. I., Chepovskiy A. M.** Sozdaniye spetsial'nyh korpusov tekstov na osnove rasshirennoy platformy TXM. *Sistemy vysokoy dostupnosti*, 2018, vol. 14, no. 3, p. 76–81. (in Russ.)
7. **Lavrentiev A. M., Soloviev F. N., Suvorova M. I., Fokina A. I., Chepovskiy A. M.** Novyi kompleks instrumentov avtomaticheskoy obrabotki teksta dlya platformy TXM i yego aprobatsiya na korpuse dlya analiza ekstremistskih tekstov. *Vestnik NSU. Series: Linguistics and Intercultural Communication*, 2018, vol. 16, no. 3, p. 19–31. (in Russ.)
8. **Chepovskiy A., Devyatkin D., Smirnov I., Ananyeva M., Kobozeva M., Solovyev F.** Exploring linguistic features for extremist texts detection (on the material of Russian-speaking illegal texts). In: 2017 IEEE International Conference on Intelligence and Security Informatics: Security and Big Data, ISI 2017. Institute of Electrical and Electronics Engineers Inc., 2017, p. 188–190.
9. **Egorova E., Chepovskiy A., Lavrentiev A.** A structural pattern based method for automated morphological analysis of word forms in a natural language. *Journal of Mathematical Sciences*, 2016, vol. 214, no. 6, p. 802–813.
10. **Lafon P.** Sur la variabilité de la fréquence des formes dans un corpus. *Mots*, 1980, no. 1, p. 127–165.
11. **Benzécri J.-P.** L'analyse des données: l'analyse des Correspondances. 2nd ed. Paris, Dunod, 1979, vol. 2.
12. **Lê S., Josse J., Husson F.** FactoMineR: An R package for multivariate analysis. *Journal of statistical software*, 2008, no. 25 (1), p. 1–18.

Материал поступил в редколлегию
Received
07.06.2019

Сведения об авторе

Соловьев Федор Николаевич, младший научный сотрудник, Федеральный исследовательский центр «Информатика и управление» РАН (Вавилова, 44/2, Москва, 119333, Россия)
the0@yandex.ru

Information about the Author

Fyodor N. Soloviev, Junior Researcher, Federal Research Center "Informatics and Management" of the Russian Academy of Sciences (44/2 Vavilov Str., Moscow, 119333, Russian Federation)
the0@yandex.ru