ИПМ им.М.В.Келдыша РАН

Абрау-2019 • Труды конференции

А.А. Алексеев, Д.С. Зуев, А.С. Катасёв,
А.Е. Кириллов, А.Ф. Хасьянов

Прототип классификатора электронных документов для информационной системы поддержки принятия судебных решений

Рекомендуемая форма библиографической ссылки

Алексеев А.А., Зуев Д.С., Катасёв А.С., Кириллов А.Е., Хасьянов А.Ф. Прототип классификатора электронных документов для информационной системы поддержки принятия судебных решений // Научный сервис в сети Интернет: труды XXI Всероссийской научной конференции (23-28 сентября 2019 г., г. Новороссийск). — М.: ИПМ им. М.В.Келдыша, 2019. — С. 40-51. — URL: http://keldysh.ru/abrau/2019/theses/98.pdf doi:10.20948/abrau-2019-98

Размещена также презентация к докладу

Прототип классификатора электронных документов для информационной системы поддержки принятия судебных решений

А.А. Алексеев[1], Д.С. Зуев[2], А.С. Катасёв[1], А.Е. Кириллов[3], А.Ф. Хасьянов[1]

[1]Казанский национальный исследовательский технический университет им. А.Н. Туполева
[2]Высшая школа информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета
[3]Арбитражный суд Республики Татарстан

Аннотация. Представлен прототип классификатора электронных документов для создаваемой информационной системы поддержки принятия решений в сфере экономического правосудия. В основу этой системы положены известные алгоритмы текстовой аналитики, а также предложенный алгоритм на основе искусственной нейронной сети. Разработана модель интеллектуального анализа текста для классификации арбитражных документов с целью определения категории (класса) судебного спора. Проведены предварительный анализ судебных документов и отбор значимых признаков, а для классификации электронных документов применены алгоритмы байесовской классификации, *k* ближайшего соседа и деревьев решений. Для повышения точности классификации предложена модель, основанная на искусственной нейронной сети и показавшая на тестовой выборке безошибочное определение типа документа для ряда классов судебных споров в арбитражном судопроизводстве.

Ключевые слова: задача классификации, искусственные нейронные сети, текстовый анализ, алгоритмы классификации, text mining, система поддержки принятия решений

Prototype of classifier for the decision support system of legal documents

A.A. Alekseev[1], D.S. Zuev[2], A.S. Katasev[1], A.E. Kirillov[3], A.F. Khassianov[2]

[1] Kazan National Research Technical University,
[2]Higher School of Information Technologies and Intelligent Systems, Kazan (Volga Region) Federal University
[3]The Arbitration Court of the Republic of Tatarstan

Abstract. We propose a prototype of the classifier of electronic documents for the decision support system in the field of economic justice. The system uses both well-known text analytics algorithms and an original algorithm based on an artificial neural network. A text mining model has been developed to classify court documents to determine the category (class) of a statement of claim. A preliminary analysis of court documents and the selection of significant features were carried out. To choose the best way of solving problem of document classification we implemented Bayesian classification algorithm, k nearest neighbor algorithm and decision trees algorithm. All used algorithms show results with errors on the same sample corpus of texts. To improve the accuracy of classification, an original model based on an artificial neural network was developed, which shows an unmistakable determination of the type of document on a test sample for a number of classes of lawsuits in arbitration proceedings.

Keywords: classification, text mining, artificial neural network, classification algorithms, decision support system.

1. Введение

Судебная система – это сфера деятельности, где объем работы с текстовыми документами весьма значителен, а процесс принятия решения всегда должен быть понятным и прозрачным. Поэтому, особенно в условиях роста нагрузки на сотрудников, работающих в названной сфере, требуются инструменты, позволяющие осуществлять интеллектуальный анализ поступающего информационного массива. В настоящее время в судах Российской Федерации планируется переход к электронному документообороту – рассмотрению арбитражных дел на основании документов в электронной форме [1]. Автоматизированный текстовый анализ позволит выделить важные признаки документов (подсудность, характер спора, участвующие стороны и т. д.), осуществить поиск в судебной базе данных и представить похожие документы, по которым уже приняты решения. Именно на этот аспект работы судебной системы направлено наше исследование: для снижения нагрузки на судей и уменьшения времени рассмотрения экономических споров предложена модель классификации судебных документов на основе интеллектуального анализа текстовых данных (Text Mining) [2], решающая задачу определения типовой категории арбитражного спора. Для определения типового класса (категории) судебного спора поставлены и решены следующие задачи:

1) разработка модели интеллектуального анализа текста для классификации арбитражных документов;

2) моделирование процессов обработки и классификации таких документов в аналитической среде RapidMiner Studio;

3) программная реализация модулей обработки и классификации на языке R;

4) выбор наиболее эффективного алгоритма классификации для дальнейшей апробации в арбитражном судопроизводстве.

Существующие программные решения, используемые сегодня в юридической области, как правило, направлены на автоматизацию документооборота в целом или же представляют собой базы данных тематических документов со скудным набором поисковых инструментов, при этом весь спектр современных семантических технологий и инструментария текстовой аналитики практически не используется. В такой ситуации качественные изменения эффективности работы судей невозможны без использования инструментов автоматизации или существенного увеличения числа сотрудников судебных органов.

На сегодняшний день специалистами Стэндфордского центра юридической информатики, Чикагского юридического колледжа в Кенте и юридического колледжа Южного Техаса создана интеллектуальная система (ИС) [3] на основе машинного обучения и анализа данных, которая прогнозирует судебные вердикты с точностью более 70%. Эта ИС в качестве входных данных использует записи базы данных (БД) Верховного суда США с 1816 по 2015 гг.

Другим примером ИС, близкой по теме нашего исследования, служит система «Case Cruncher Alpha» [4], разрабатываемая в Sidney Sussex College, Cambridge, и ориентированная на прогнозирование решения юридических задач в банках, страховых компаниях и юридических консультациях. Основной ее недостаток (как и многих других иностранных систем) – отсутствие поддержки русского языка и кириллической транскрипции.

2. Анализ текстовых документов методами Text Mining

Предлагаемый нами классификатор является частью разрабатываемой интеллектуальной ИС «Робот-юрист» поддержки принятия судебных решений. Эта информационная система позволит участникам юридического процесса эффективно проводить подготовку дел и осуществлять планирование судебной деятельности. Система ориентирована на арбитражные суды, занимающиеся рассмотрением споров в сфере предпринимательства. Цели и задачи, общая архитектура системы «Робот-юрист», предлагаемые модули и подходы к ее разработке представлены в [5, 6]. В общем виде эта ИС, разрабатываемая на основе интеллектуального анализа текста и искусственных нейронных сетей, схематично изображена на рисунке 1, где модуль 1 – модуль предварительной обработки текстовых документов, модуль 2 – модуль определения основного класса судебного документа, модуль 3 – модуль определения подкласса судебного документа, а модуль 4 – модуль принятия судебного решения. Ниже описана работа первых двух из названных модулей.

Известно, что анализ текстовых документов выполняется в пять шагов [7] (см. рисунок 2):

1) *Поиск информации* – происходит определение документов для дальнейшей обработки и анализа; зачастую пользователи самостоятельно формируют корпусы анализируемых документов, но при огромном числе документов их ручной отбор становится трудоёмким, поэтому возникает необходимость в использовании алгоритмов автоматического выбора.

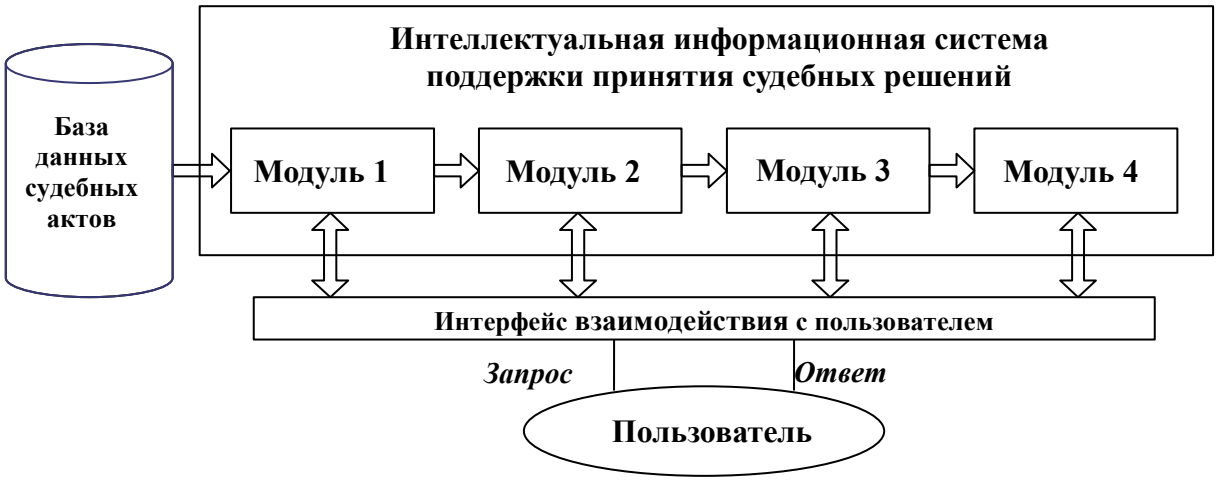

Рис. 1. Схема разрабатываемой системы поддержки принятия судебных решений

2) *Предварительная обработка (предобработка) документов* – текст выбранных документов преобразуются в форму, удобную для применения методов машинного обучения; предобработка проводится с целью удаления терминов, не несущих существенного информационного значения, знаков препинания, а также представления текста в нормализованной форме. Более подробно методы, использованные на данном шаге, рассмотрены далее.

3) *Извлечение значимой информации* – выполняется отбор информативных признаков, которые характеризуют определённые корпусы текстовых документов и являются в них ключевыми.

4) *Применение методов машинного обучения* – основной шаг анализа, на котором формируются новые знания и выявляются скрытые в тексте закономерности.

5) *Интерпретация результатов* – представление результатов анализа на естественном языке в удобной для пользователя форме, в виде графиков и таблиц.

Рис. 2. Этапы анализа текстовых документов методами Text Mining. Предварительная обработка текста

Здесь, как правило, в совокупности применяется несколько методов. Один из них – токенизация текста, то есть операция разбиения текстового документа на отдельные слова. В результате для дальнейшей обработки формируется массив токенов или лексем [8]. Следующим шагом является приведение всех символов к верхнему или нижнему регистру. Например, все слова «text», «Text», «TEXT» приводятся к нижнему регистру «text» [9]. Далее необходимо провести фильтрацию стоп-слов, которые не несут в себе значимого информационного смысла: союзы, предлоги, артикли, междометия, частицы и т. п. Список стоп-слов составляется заранее в зависимости от языка обрабатываемого текста.

Следующий шаг – стэмминг или лемматизация [10], в рамках которого происходит процесс нормализации слов: определение однокоренных слов, отсечение суффиксов и окончаний, приведение терминов к единственному числу, именительному падежу существительного, прилагательного или неопределенной форме глагола. Основной недостаток при проведении таких преобразований – возможное нарушение семантики предложений, словосочетаний, поэтому необходимо учитывать язык оригинала. На текущий момент существует ряд известных реализаций алгоритмов стэмминга и лемматизации для русского языка в виде подключаемых программных библиотек – Snowball, Porter или MyStem [11]. Нами в качестве библиотеки стэмминга при моделировании в среде RapidMiner Studio использован алгоритм Snowball, а при написании программного модуля на языке R – алгоритм лемматизации MyStem.

Для проведения дальнейшего анализа необходимо представить текст в форме, удобной для анализа. Нами использовалась матрица документ–термин (Document-Term Matrix – DTM), представляющая собой таблицу, где каждая строка соответствует документу, а столбец – терминам, встречающимся в корпусе документов [12]. На пересечении строк и столбцов хранятся значения весов терминов в документе.

В рамках создания прототипа классификатора нами проводился анализ судебных документов по четырем категориям: оспаривание решений антимонопольных органов («АНТИМОНОПОЛЬНЫЕ»), оспаривание действий судебных приставов («ПРИСТАВЫ»), привлечение к ответственности за нарушение условий лицензирования («ЛИЦЕНЗИИ»), споры о неисполнении или ненадлежащем исполнении обязательств по договорам поставки («ПОСТАВКИ»).

После этапа предобработки документов была сформирована матрица DTM размерностью 167*5419, где 167 – общее число документов (38 по классу «ПРИСТАВЫ», 32 по классу «АНТИМОНОПОЛЬНЫЕ», 61 по классу «ПОСТАВКИ», 36 по классу «ЛИЦЕНЗИИ»), а 5419 – число терминов, содержащихся во всех документах. Значениями матрицы была выбрана мера *TF*. После первоначального формирования матрицы DTM было замечено, что из 5419 терминов не все являются значимыми для определения класса документа, поэтому перед дальнейшими действиями необходимо произвести отбор информативных признаков.

Отбор информативных признаков. Как правило, методы классификации текстовой информации основаны на предположении, что документы, относящиеся к одной категории, содержат одинаковые признаки (слова или словосочетания), и наличие или отсутствие таких признаков в документе говорит о его принадлежности или непринадлежности к тому или иному классу (см., например, [7]). Для определения группы признаков (терминов), которые характеризуют категории обрабатываемых документов, нами применялись следующие методы: энтропийный метод прироста информации (Information Gain), хи-квадрат (Chi Square) и индекс Джини (Gini Index) [14, 16]. В таблице 1 представлены фрагменты результатов отбора признаков этими методами. Из приведенной таблицы видно, что все три использованных метода выявили практически одинаковые термины, незначительно отличающиеся по своим коэффициентам.

Таблица 1. Результаты отбора значимых признаков

Information Gain		Gini Index		Chi Square	
Термин	Коэффициент	Термин	Коэффициент	Термин	Коэффициент
истц	1.0	Истц	1.0	судебн	1.000
истец	0.899	Истец	0.925	заявител	0.935
исполнител	0.898	обязательств	0.923	ответствен	0.921
пристав	0.898	заявител	0.878	Рф	0.890
обязательств	0.888	Ответчик	0.874	ответчик	0.861
заявител	0.855	Накладн	0.873	исполнител	0.806
договор	0.846	Договор	0.849	пристав	0.806
незакон	0.840	исполнител	0.813	Са	0.801
накладн	0.835	Пристав	0.813	Коап	0.783
ответчик	0.831	Приста	0.758	исполнительн	0.766
взыскател	0.799	Взыскател	0.754	Привлека	0.759
исполнительн	0.799	исполнительн	0.736	производств	0.749
приста	0.793	Коап	0.731	Истц	0.744
взыскан	0.790	Иск	0.729	предусмотрен	0.725
антимонопольн	0.784	производств	0.721	Взыскан	0.724

Для формирования обучающей и тестовой выборок взято 40 терминов, полученных на этапе отбора информативных признаков, – итоговая матрица документ-термин имеет размерность 167*40. В перспективе планируется увеличение числа терминов, однако на этапе разработки прототипа системы выбранного ограниченного набора терминов вполне достаточно. Очевидно, что количество рассматриваемых терминов влияет на размерность матрицы и, следовательно, на время, необходимое на обучение классификатора. Для формирования обучающей и тестовой выборок полученная матрица документ-термин была разбита на

матрицы 117*40 и 50*40 (соотношение 70/30). Полученные значения матрицы использовались в качестве входных данных для алгоритмов классификации.

3. Классификация электронных документов

Напомним, что задача классификации [15] формулируется следующим образом. Даны набор текстовых документов $D=\{X_1, ..., X_n\}$ и набор k различных дискретных значений $\{1, ..., k\}$, каждое из которых соответствует индексу (метке) класса (категории). Для каждого документа X_i. необходимо определить его категорию (соответствующую значению индекса).

Данная задача, как правило, решается с помощью алгоритмов обучения с учителем, где обучающая выборка документов (т. е. документов с известными метками категорий) используется для построения модели классификации, которая определяет связь признаков в определённом документе с одной из меток класса. Для элементов тестовой выборки, где у документов класс заранее не известен, разработанная обученная модель должна определить метку класса. Для уточнения алгоритмов работы классификатора модель должна периодически проходить переобучение.

Для определения метода решения, наиболее оптимального на имеющейся выборке данных, было апробировано несколько алгоритмов классификации. Первоначально проверялись следующие методы классификации: наивный байесовский классификатор [7], метод k-ближайших соседей [8] и деревья решений [15]. Результаты работы данных классификаторов на тестовой выборке представлены ниже в таблицах 2–4 (матрицах сопряженности или матрицах классификации).

Таблица 2. Результаты работы байесовского классификатора

	true ПРИСТАВЫ	true АНТИМОНОПОЛЬНЫЕ	true ПОСТАВКИ	true ЛИЦЕНЗИИ	class precision
pred. ПРИСТАВЫ	11	0	0	0	100.00%
pred. АНТИМОНОПОЛЬНЫЕ	0	10	0	2	83.33%
pred. ПОСТАВКИ	0	0	18	0	100.00%
pred. ЛИЦЕНЗИИ	0	0	0	9	100.00%
class recall	100.00%	100.00%	100.00%	81.82%	

Таблица 3. Результаты работы алгоритма k ближайших соседей

	true ПРИСТАВЫ	true АНТИМОНОПОЛЬНЫЕ	true ПОСТАВКИ	true ЛИЦЕНЗИИ	class precision
pred. ПРИСТАВЫ	11	0	0	0	100.00%
pred. АНТИМОНОПОЛЬНЫЕ	0	10	0	1	90.91%
pred. ПОСТАВКИ	0	0	18	0	100.00%
pred. ЛИЦЕНЗИИ	0	0	0	10	100.00%
class recall	100.00%	100.00%	100.00%	90.91%	

Таблица 4. Результаты работы деревьев решений

	true ПРИСТАВЫ	true АНТИМОНО-ПОЛЬНЫЕ	true ПОСТАВКИ	true ЛИЦЕН-ЗИИ	class precision
pred. ПРИСТАВЫ	11	0	0	0	100.00%
pred. АНТИМОНО-ПОЛЬНЫЕ	0	10	0	0	100.00%
pred. ПОСТАВКИ	0	0	18	1	94.74%
pred. ЛИЦЕНЗИИ	0	0	0	10	100.00%
class recall	100.00%	100.00%	100.00%	90.91%	

Отметим, что ни один из примененных алгоритмов не показал стопроцентной точности классификации. Так, на тестовой выборке были допущены следующие ошибки: классификатор Байеса – 4%, точность классификации 96%; kNN – 2%, точность классификации 98%; деревьев решений – 2%, точность классификации 98%. Учитывая, что выборка данных весьма ограничена, полученные результаты нельзя считать удовлетворительными.

Для более точной классификации судебных документов использована модель на основе искусственных нейронных сетей (ИНС) [16]. Разработанная нейронная сеть имеет следующие параметры: 40 нейронов во входном слое, 1 скрытый слой с 4 нейронами, 4 выходных нейрона; активационная функция: сигмоида. Результат решения задачи классификации по методу ИНС представлен в таблице 5.

Результат работы классификатора на основе нейронной сети показал 100%-ю точность классификации на тестовой выборке для классификации по четырем признакам.

Таблица 5. Результаты работы ИНС

	true ПРИСТАВЫ	true АНТИМОНО-ПОЛЬНЫЕ	true ПОСТАВКИ	true ЛИЦЕНЗИИ	class precision
pred. ПРИСТАВЫ	11	0	0	0	100.00%
pred. АНТИМОНО-ПОЛЬНЫЕ	0	10	0	0	100.00%
pred. ПОСТАВКИ	0	0	18	0	100.00%
pred. ЛИЦЕНЗИИ	0	0	0	11	100.00%
class recall	100.00%	100.00%	100.00%	100.00%	

Заключение

В отдельно взятой предметной области – арбитражном судопроизводстве – практически применены существующие методы интеллектуального анализа текста Предложена нейросетевая модель классификации текстовых документов по типовым категориям. В рамках решения задачи классификации проведены предварительный анализ судебных документов и отбор значимых признаков для определенных категорий судебного спора, применены алгоритмы байесовской

классификации, *k* ближайшего соседа, деревьев решений. Для повышения абсолютной точности классификации предложена модель, основанная на искусственной нейронной сети и показавшая безошибочное определение типа документа на тестовой выборке. Разработан программный комплекс на языке R, выполняющий предобработку текстовых документов. Апробация предложенных модели и программного комплекса проведена в Арбитражном суде РТ.

На следующем этапе развития системы планируется увеличить выборку исковых арбитражных заявлений и рассмотреть большее число типов возможных судебных споров, а также разработать программные модули, выполняющие задачи отбора информативных признаков и классификации. После тестовой эксплуатации модуль будет включен в качестве отдельного сервиса в разрабатываемую систему «Робот-Юрист» [5, 6].

Работа выполнена за счет средств субсидии, выделенной Казанскому федеральному университету для выполнения государственного задания в сфере научной деятельности, проект 1.2368.2017/ПЧ, и при частичной финансовой поддержке РФФИ и Правительства Республики Татарстан в рамках научного проекта № 18-47-160012.

Литература

1. Постановление Пленума Высшего Арбитражного Суда РФ от 25 декабря 2013 г. № 100 «Об утверждении Инструкции по делопроизводству в арбитражных судах Российской Федерации (первой, апелляционной и кассационной инстанций)».
2. Fayyad U., Piatetsky-Shapiro G., Smyth P. From Data Mining to Knowledge Discovery: an Overview, Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996. — P. 1–34.
3. Katz D.M., Bommarito M.J. II, Blackman J. A general approach for predicting the behavior of the Supreme Court of the United States. PLoS ONE 12(4): e0174698., 2017. — https://doi.org/10.1371/journal.pone.0174698ю
4. Case Crunch Alfa [Электронный ресурс]. — http://www.case-crunch.com, свободный/
5. Зуев Д.С., Марченко А.А., Хасьянов А.Ф. Применение инструментов интеллектуального анализа текстов в юриспруденции // CEUR Workshop Proceedings. — 2017. — V. 2022. — P. 214–218. — http://ceur-ws.org/Vol-2022/paper35.pdf.
6. Алексеев А.А., Зуев Д.С., Катасёв А.С., Тутубалина Е.В., Хасьянов А.Ф. Интеллектуальная информационная система поддержки принятия судебных решений в сфере экономического правосудия// Научный сервис в сети Интернет: труды XIX Всероссийской научной конференции (17-22 сентября 2018 г., г. Новороссийск). — М.: ИПМ им. М.В. Келдыша, 2018.

7. Барсегян А.А., Елизаров С.И., Куприянов М.С., Холод И.И., Тесс М.Д. Анализ данных и процессов: учеб. пособие / 3-е изд., перераб. и доп. — СПб.: БХВ-Петербург, 2009. — 512 с.: ил. + CD-ROM.
8. Aggarwal C.C. Machine Learning for Text. Springer International Publishing AG, part of Springer Nature, 2018. — 493 p.
9. Hofmann M., Klinkenberg R. RapidMiner: Data Mining Use Cases and Business Analytics Applications. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series — CRC Press, 2016. —525 p. — ISBN: 9781482205503.
10. Маннинг К.Д., Рагхаван П., Шютце Х. Введение в информационный поиск: Пер. с англ. — М.: ООО «И. Д. Вильямс», 2011. — 528 с.: ил. — Парал. тит. англ.
11. Segalovich I. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine // Proceedings of the International Conference on Machine Learning, 2003, Las Vegas, Nevada, USA.
12. Williams G. Hands-On Data Science with R, Text Mining, 5th November 2014.
13. Feinerer I., Hornik K., Meyer D. Text Mining Infrastructure in R// Journal of Statistical Software. — V. 25, Issue 5, March 2008. — 54 p.
14. Kotu V., Deshpande B. Predictive Analytics and Data Mining. Concepts and Practice with RapidMiner. Morgan Kaufmann, 225 Wyman Street, Waltham, MA 02451, USA, 425 p.
15. Aggarwal C.C. Data Classification: Algorithms and Applications. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series — CRC Press, 2015. — 707 p. — ISBN: 9781498760584.
16. Глова В.И., Аникин И.В., Катасёв А.С., Кривилёв М.А., Насыров Р.И. Мягкие вычисления: Учебное пособие. Казань: Изд-во Казан. гос. техн. ун-та, 2010. — 206 с.

References

1. Postanovleniye Plenuma Vysshego Arbitrazhnogo Suda RF ot 25 dekabrya 2013 g. № 100 "Ob utverzhdenii Instruktsii po deloproizvodstvu v arbitrazhnykh sudakh Rossiyskoy Federatsii (pervoy, apellyatsionnoy i kassatsionnoy instantsiy)".
2. Fayyad U., Piatetsky-Shapiro G., Smyth P. From Data Mining to Knowledge Discovery: an Overview, Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996. — P. 1–34.
3. Katz D.M., Bommarito M.J. II, Blackman J. A general approach for predicting the behavior of the Supreme Court of the United States}, PLoS ONE 12(4): e0174698., 20. DOI: 10.1371/journal.pone.0174698.
4. Case Crunch Alfa. — http://www.case-crunch.com.

5. Zuev D.S., Marchenko A.A., Khassianov A.F. Text mining tools in legal documents // CEUR Workshop Proceedings. — 2017. — V. 2022. — P. 214–218. — http://ceur-ws.org/Vol-2022/paper35.pdf (In Russian).
6. Alekseev A.A., Zuev D.S., Katasev A.S., Tutubalina E.V., Khassianov A.F. Intellectual information decision support system in the field of economic justice, Nauchnyy servis v seti Internet: trudy XIX Vserossiyskoy nauchnoy konferentsii (17–22 sentyabrya 2018 g., g. Novorossiysk), Moscow, Keldysh Institute of Applied Mathematics, 2018 (in Russian).
7. Barsegyan A.A., Yelizarov S.I., Kupriyanov M.S., Kholod I.I., Tess M.D. Analiz dannykh i protsessov, 3 ed. SPb. BKHV-Peterburg, 2009. — 512 p.
8. Aggarwal C.C. Machine Learning for Text. Springer International Publishing AG, part of Springer Nature 2018. — 493 p.
9. Hofmann M., Klinkenberg R. RapidMiner: Data Mining Use Cases and Business Analytics Applications. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, CRC Press, 2016 — 525 p. — ISBN: 9781482205503.
10. Manning C. D., Raghavan P., Schutze H. S. Introduction to Information Retrieval, Cambridge University Press, Cambridge, England, 2008. — ISBN: 978-0521865715.
11. Segalovich I. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine // Proceedings of the International Conference on Machine Learning, 2003, Las Vegas, Nevada, USA.
12. Williams G. Hands-On Data Science with R, Text Mining, 5th November 2014.
13. Feinerer I., Hornik K., Meyer D. Text Mining Infrastructure in R. Journal of Statistical Software. — V. 25, Issue 5, March 2008. — 54 p.
14. Kotu V., Deshpande B. Predictive Analytics and Data Mining. Concepts and Practice with RapidMiner. Morgan Kaufmann, 225 Wyman Street, Waltham, MA 02451, USA. — 425 p.
15. Aggarwal C.C. Data Classification: Algorithms and Applications. Chapman& Hall/CRC Data Mining and Knowledge Discovery Series — CRC Press, 2015 — 707 p. — ISBN: 9781498760584.
16. Glova V.I., Anikin I.V., Katasov A.S., Krivilov M.A., Nasyrov R.I. Myagkiye vychisleniya, Uchebnoye posobiye. Kazan': Izd-vo Kazan. gos. tekhn. un-ta, 2010. — 206 s.