Построение языковой сети правового пространства для судебных приложений искусственного интеллекта

Крылов Владимир Владимирович
Профессор, д.т.н.,
Национальный исследовательский университет Высшая школа экономики
603155, Н. Новгород, ул. Большая Печорская, 25/12
vkrylov@hse.ru

Крылов Сергей Владимирович
Доцент, к.т.н.,
Национальный исследовательский университет Высшая школа экономики
603155, Н. Новгород, ул. Большая Печорская, 25/12
skrylov@hse.ru

Жигалов Григорий Максимович
Магистрант,
Национальный исследовательский университет Высшая школа экономики
603155, Н. Новгород, ул. Большая Печорская, 25/12
zhigalov8@gmail.com

Аннотация: Показано, что математическое «правовое» пространство является обособленным кластером естественного языка, объективные понятия близости в таком пространстве порождают понятия связанности юридических утверждений и тем самым возможность предсказания правовой классификации фактов, высказанных на естественном языке. Для иллюстрации некоторых возможностей использования модели разработан классификатор оценки успешности гражданского дела. Построен прототип чат-бота помощника юриста для интерпретации произвольного высказывания в терминах нормативных документов.

Ключевые слова: модель правового пространства, эмбеддинги юридических терминов, машинное обучение.

Law Space Language Grid Generation for Judge AI Applications

Vladimir V. Krylov
National Research University Higher School of Economics.
Russia, Nizhny Novgorod, Bolshaya Pecherskaya st, 25/12
vkrylov@hse.ru

Sergei V. Krylov
National Research University Higher School of Economics.
Russia, Nizhny Novgorod, Bolshaya Pecherskaya st, 25/12
skrylov@hse.ru

Gregory M. Zhigalov
National Research University Higher School of Economics.
Russia, Nizhny Novgorod, Bolshaya Pecherskaya st, 25/12
Zhigalov8@gmail.com

Abstract: We propose the model of Judge AI implemented by the approach based on commutative diagram of mappings. We suggest using the constellation for the grid of normative acts and the narrative together with the corresponding divergence. In this paper, the pre-trained models known in NLP as doc2vec and Fast Text are described. The evaluation of the quality of models carried out using open databases of court decisions.

Keywords: Law space model, legal terms embedding, machine learning.

1 Введение

Вопрос разработки роботов-юристов стал часто обсуждаться как в среде юристов, так и среди разработчиков систем искусственного интеллекта. Большинство сегодняшних разработок нацелены на автоматизацию рутинных процессов составления или проверки заполнения готовых форм на основе широкого набора имеющихся документов, относящихся к решаемой задаче. Целью настоящей работы является попытка представить модель, которая описывала бы основные закономерности в правовых документах с точки зрения математических понятий. Такая модель позволит подойти к проблеме установления соответствия действующим нормам права любых жизненных ситуаций на основании формальных соотношений и правил, что может заметно сократить влияние человеческого фактора на принимаемые решения. Робот, выполняющий действия на основании математической модели права, будет действовать с максимально возможной для программного решения эффективностью, поскольку будет использовать не плохо формализованные описания правовых норм, а их согласованные представления в рамках единой модели. Вообще, любая наука, опирающаяся математическую платформу, получает возможность более объективного описания собственного предмета и приобретает возможности получения выводов из утверждений, прогнозирования последствий на основании достаточно жестких формальных предиктивных моделей. Успехи физики и прикладных наук вокруг нее практически во всем обязаны использованию математических моделей. То же происходит сегодня и в биологии, впитывающей математический стиль в экспериментальные исследования и все больше обращающей внимание на построение формальных моделей объектов и процессов. Не только наука, но и искусство получают успешную подпитку от применения моделей машинного обучения как в живописи, так и написании музыки. В настоящей работе мы исходим из того определения права, которое толкует его как некий общественный договор о том, что считать в данном обществе недопустимым в действиях любого члена общества или их группы. Будучи выраженным через систему права и воплотившись через нормативно-правовые акты в вид деятельности государства, это неотъемлемая часть существования человека стала существенно влиять на его поведение в не меньшей степени, чем природная среда обитания или экономические взаимоотношения. Однако, если последние факторы поддаются в какой-то мере моделированию и предикции, то правовое поле остается вне возможностей формализации и построения кибернетических помощников. Это во многом связано с тем, что правовые нормы формулируются в виде текстов на естественном языке, а с помощью государственного механизма их влияние на человека материализуется в экономические, социально-биологические и даже терминальные последствия (смертная казнь или пожизненное лишение свободы). Тексты в праве стали материализоваться с серьезными последствиями для всех без исключения человеческих индивидуумов. Исследователи философии права и философии естественных языков обратили внимание на то, что нормы регулирования существования субъектов отождествляются с некоторыми языковыми конструкциями, что порождает серьезные проблемы, поскольку совершенные правовые нормы могут неправильно регулировать сообщество из-за несовершенства их языковых представлений.

В этой работе мы пытаемся заменить набор текстов, образующих правовое пространство, в некоторую математическую модель, которая может помочь более точному применению действующего права и анализу его внутренней структуры с целью совершенствования. Компьютерная обработка текстов на сегодняшний день сосредоточенная в области технологий NLP/NLU – Natural Language Processing/Natural language Understanding, имеет немало средств отображения текстов в различные модели, имеющие математическую основу. Эти модели представляют широкий спектр представлений одного и того же текста как синтаксической структуры, семантической модели, совокупности именованных сущностей, общего эмоционального выражения, принадлежности определенному ав-

тору, типа текста из заданного набора типов и т.п. Авторы поставили своей задачей получить для нормативно-правовых актов некоторую единую модель, которая может служить как объект для изучения с точки зрения структурного и количественного анализа, так и для выполнения правовой классификации произвольного описания потока событий, выраженного на естественном языке (нарратива). Говоря упрощенно, как может быть решена задача машинного перевода с естественного языка на язык юридический.

В исследовании будет использоваться русский язык и правовые нормы в форме Уголовного и Гражданского Кодекса Российской Федерации. Метод, который используется – математические модели языка и правовых норм. Очевидно, что язык есть эмпирический объект и рассуждать о нем в абстрактных терминах в нашей задаче будет неправомерно. Для представления языка будет использован, подход, известный под названием “корпусная лингвистика”. Инструментом для работы будет аппарат NLU – Natural Language Understanding. В данной работе мы показали, как может быть построена такая модель, и как на ее основе могут решаться задачи прогнозирования исхода гражданских процессов, квалифицироваться поведение и поступки людей в соответствии с уголовным кодексом. Прототип чат-бота для демонстрации возможностей инкрементального обучения модели робота-юриста в области уголовного законодательства, также описывается в этой работе.

2 История вопроса

Понимание текстов на естественном языке охватывает целый спектр задач отображения текста в набор однозначно интерпретируемых сущностей, чем последовательность символов, которой любой текст первично является. Остановимся здесь на типовых задачах, которые обычно решаются.

2.1 Лексико-грамматический парсинг

Важнейшей единицей текста является предложение. Каждое предложение в результате синтаксического парсинга может быть представлено графом связи слов, позволяющим установить объекты и субъекты и действия над ними, описанные в данном предложении. Из предложения могут быть извлечены именованные сущности и проведены замены конструкций-указателей на антецеденты, для построения консистентной текстовой структуры проводится прореживание удалением несодержательных стоп-слов, отдельные слова могут быть лематизированы. Описанная выше обработка текстов используется в настоящее время в большинстве компьютерных систем поддержки процессов в юриспруденции и является важнейшей частью современных роботов-юристов. Так известный робот-юрист, созданный 19-летним студентом-программистом Стэнфордского университета Joshua Browder для помощи автовладельцам, оспаривающим штрафы за парковку, за год выиграл в судах дела на общую сумму в три миллиона долларов.[1]. Искусственный интеллект робота не учитывает факторы, которые учитывает человек-юрист, - например, кто находился за рулем автомобиля в момент паркования, или сопутствующие дорожные условия. Программа лишь задает вопросы о происшествии и формулирует текст апелляции в суд. Сбербанк внедрил робота-юриста, который заменит 3 тыс. сотрудников. который извлекает из бумаг нужные данные и составляет текст искового заявления. [2] Над созданием робота-юриста совместно работают представители юрфака, программисты и математики Казанского университета. Проект назвали «Искусственный интеллект в юриспруденции». Целью проекта является создание системы, которая могла бы на основании имеющихся данных выдавать возможный вариант решения судьи. В случае если данных для принятия решения не хватает, она бы показала, чего именно не достает, Программа также должна будет выявлять нелогичные постановления, что позволит сократить ошибки и возможности для коррупции. [3]. Электронный ассистент ROSS - программа, созданная в компании ROSS Intelligence и работающая на когнитивном компьютере IBM Watson, оснащённом вопросно-ответной системой искусственного интеллекта - использует естественный язык для того, чтобы понять вопросы юристов и сообщить им информацию по интересующим их судебным делам и законодательству с необходимыми ссылками. Машина избавляет юристов от необходимости просматривать множество материалов в поисках наиболее подходящих прецедентов. Также Ross постоянно отслеживает новые судебные решения и может вовремя проинформировать о тех, которые помогут в деле. Весь процесс занимает секунды. На сегодня робота-ассистента уже протестировали более 20 фирм. Пока он успел изучить банкротное право, но уже готовится освоить трудовое право, право интеллектуальной собственности, налоговое право и вопросы причинения ущерба здоровью. Цель, заявленная основателем проекта Эндрю Аррудой — позволить каждому юристу пользоваться помощью искусственного интеллекта. Учитывая, какие компании одобряют начинание, проект будет успешным: один из инвесторов - NextLaw Labs, компания, занимающаяся юридическими технологическими стартапами в Пало Альто (так, ранее компания инвестировала в стартап Apperio, занимающийся разработкой юридических технологий) и поддерживаемая Dentons. Робот уже заслужил признание. В мае юрфирма BakerHostetler дала ему лицензию на работу в банкротной практике. Пару недель спустя то же самое сделали Latham & Watkins LLP и Briesen & Roper SC. Помочь юристам работать быстрее, лучше — и дешевле - такова цель подобных разработок, уверены в компании. [4]. Таких примеров существует уже немало и, несмотря на то, что в существующих публикациях нередко просматривается больше желаемого, чем уже реализованного функционала, общая тенденция на использование NLU как основного инструмента для превращения текста в запросы к базам данных и базам знаний является очевидной и успешной.

2.2 Семантическое моделирование текстов и извлечение тематики

Семантическое моделирование основано, как правило на LSA - латентно-семантическом анализе текста, его вероятностном варианте – pLSA и LDA – латентном размещении Дирихле. LSA - это математический / статистический метод для извлечения и выведения отношений ожидаемого контекстуального использования слов в отрывках дискурса. Метод не использует построенные человеком словари, базы знаний, семантические сети, грамматики, синтаксические синтаксические анализаторы или морфологии и т. п. и принимает в качестве входных данных только необработанный текст, разбитый на слова, определенные как уникальные символьные строки, и разделенные на значимые отрывки или примеры, такие как предложения или абзацы [5]. LSA строит интерпретацию текста темами (topics), извлекаемыми из корпуса текстов в процессе unsupervised learning. Эти темы носят абстрактный характер, то есть определяются группами слов, которые связаны друг с другом. В документе может быть несколько тем. Модель текста как суперпозиции тем помогает исследовать большие объемы текстовых данных, выделять абстрактные темы и находить семантическое сходство между документами. Эта модель используется также в поисковых системах. В LSA используется идея распределенной семантики, состоящей в анализе отношений между набором документов и содержащимися в них терминами путем создания набора понятий, связанных с документами и терминами. LSA предполагает, что слова, близкие по значению, будут встречаться в похожих фрагментах текста (статистическая гипотеза распределения). Матрица, содержащая количество слов в каждом абзаце (строки представляют собой уникальные слова, а столбцы представляют каждый абзац), построена из совокупности текста, а математический метод, называемый разложением по собственным числам матрицы (SVD), используется для уменьшения количества строк при сохранении структуры столбцов. Мы можем контролировать размерность скрытого пространства во время SVD, поэтому количество тем представления текста можно выбирать. Другие упомянутые методы pLSA и LDA [6,7]. представляют собой вероятностные модели разложения текста на самоизвлекаемые темы. Таким образом введение математической модели позволяет решать более сложные задачи представления текстов. Следующим по глубине применяемого уровня математических моделей, является классификация текстов, например, по их эмоциональной окраске - sentiment analysis.

2.3 Модели текстов для перевода и классификации по скрытым признакам

Машинный перевод текстов, также как и классификация по некоторых общим скрытым признакам, может осуществляться путем построения отображения текстовой последовательности в другую последовательность или конечное множество (натуральных чисел) путем построения математической модели в форме некоторой нейронной структуры (одной или нескольких связанных нейронных сетей) и обучения коэффициентов. Именно машинный перевод инициировал исследования в области построения математических моделей естественных языков, связи текстов с их авторами и других задач лингвистики. Самым значительным продвижением математических моделей в лингвистическую проблему при этом можно считать появление различных способов отображения слов и затем целых фрагментов текстов в многомерное числовое пространство. Получаемую модель и результат работы называют эмбеддингом. До этих работ использовался простейший эмбеддинг - модель текста, называемая «мешок слов». В этой модели каждое слово текста представлено в виде вектора фиксированной длины, длина которого равна размеру словаря. Каждое измерение этого вектора - это количество вхождений слова в конкретный текст. Мешок слов приводит к разреженному и многомерному представлению. Структурная информация документа удаляется, и модели должны выяснить, какие векторные размеры семантически похожи. Например, отображение «кошачьих» и «кошек» в разных измерениях менее интуитивно понятно, так как модель вынуждена изучать корреляцию между разными измерениями. Однако такая модель имела много недостатков. Одна из проблем представления «мешком слов» заключается в том, что векторы документа семантически связаны. Существует несколько различных моделей основанных на этом предположении, но все они основаны на гипотезе совместного распределения. Это означает, что «слово характеризуется компанией, которую оно хранит». Целью новых методов эмбеддинга стало выявление семантических и синтаксических закономерностей в языке из больших неконтролируемых наборов документов, таких как, например, Википедия. Слова, которые встречаются в одном и том же контексте, должны быть представлены близкими по расстоянию друг от друга векторами. Эмбеддинги на уровне слов приводят к представлениям документов, которые больше не имеют фиксированной длины. Наиболее эффективными на практике оказались модели, описанные Mikolov et al. в 2013 году [8]. Модель под названием word2vec стала очень популярной моделью эмбеддинга слов. На основе гипотезы статистического распределения существует несколько других моделей эмбеддинга слов, предложений и целых документов. Исторически развитие лингвистических моделей породило такие все более усложняющиеся архитектуры представлений как GloVe[9], FastText [10], BERT[11], ELMO [12] и некоторые другие. Практическое использование таких эмбеддингов к настоящему времени позволяет считать математическую модель лингвистического объекта в виде одного или нескольких векторов многомерного числового пространства корректной формой представления.

3 Модель правового пространства

Предметом настоящего исследования являются модели для текстов, относящихся к юридической области деятельности. Язык юридических документов уже внешне достаточно сильно отличается от бытового языка, которым обычно излагаются факты из реальной жизни. Во многом задачей юристов является перевод с «юридического» на

бытовой язык и обратно. Однако такой перевод имеет сильнейшие отличия от перевода с одного естественного языка на другой. Во-первых, оба эти языка используют слова из одного и того же естественного языка. Поэтому проблема нахождения адекватных языковых конструкций для правильно понимаемого перевода здесь отсутствует. Однако, принципиальное значение здесь имеет «точность» перевода, поскольку она будет определять правовые последствия использования сделанного перевода. Специфика проблемы порождает необходимость изучения структуры пространства математических представлений текстов в юридической области, их базу, заданную нормативно-правовыми актами, основные отношения в модели, соответствующие правовым отношениям в реальном мире.

3.1 Данные для конструирования модели

В качестве текстов, образующих сегмент бытового языка, был выбран корпус русского языка проекта Open Corpora (76882 текстов) [13]. Сегмент нормативно-правовых норм был образован текстами Уголовного кодекса РФ (496 текстов) и Гражданского кодекса РФ (1551 текст). В целом «юридический язык» был представлен текстами судебных решений первой инстанции судов общей юрисдикции РФ [14]:

- Описательная и мотивировочная части судебных решений по уголовным делам, рассматриваемым в первой инстанции (в общей сложности около 20 тыс. текстов);
- Описательные части судебных решений по гражданским делам, рассматриваемых в судах первой инстанции. (около 25 тыс. текстов)

Тексты статей кодексов и судебных актов в большей степени состоят из очень длинных предложений, это происходит вследствие необходимости использования громоздких формулировок законов, частых перечислений и для достижения максимальной однозначности семантик. Для упрощения анализа все тексты в корпусе были обработаны с помощью алгоритма Textrank [15]. Это графовый метод извлечения ключевой информации из текста. Каждый из текстов корпуса был очищен и разделен на токены, каждый из которых приведен к стандартной форме с использованием библиотеки Mystem [16]. Токены фильтровались по частоте использования в текстах, выбирались только часто встречающиеся слова. Кроме того, для объединения токенов в «фразы» или n-граммы использовался алгоритм нахождения коллокаций токенов в тексте [17], имплементированный в библиотеке Gensim, что позволило выделить часто встречающиеся группы токенов и сократить размерность словаря. Поиск, загрузка и обработка необходимых для исследования данных производилась в течение всего хода работы и в некоторых случаях требовала полного пересмотра предыдущих алгоритмов и решений. Данные и программный код, с помощью которого производилась загрузка и предобработка данных выложены в открытом доступе [18].

Описанные здесь тексты были использованы для генерации математических моделей в виде 300 мерных числовых векторов. За минимальную лексическую единицу, отображаемую в вектор, было принято нормализованное на предварительном этапе предложение. Трудность вызывал размер большинства правовых и описательных текстов. Для уменьшения вычислительной нагрузки и ускорения интерпретаций было решено использовать методы извлечения ключевой информации из текстов, как один из этапов подготовки текстов. Эти трансформации позволили строить текстовый объект, основываясь на сравнительной важности информации, содержащейся в последовательностях внутри этого объекта. Модель использовала предложения и части предложения длиной не более 30 слов, как минимальную семантическую единицу текста. Каждый текст разбивался на минимальные семантические единицы (с номером i для текста m) и модель обучалась строить отображение такой единицы в вектор x_i^m в пространстве $\mathbb{R}^n$.

3.2 Модель на основе алгоритма doc2vec

Алгоритм doc2vec возвращает векторное представление для каждого из предложений, которое поступает ему на вход. Модель языка на основе алгоритма doc2vec с внутренней размерностью представления 300, была обучена на всех материалах, как на едином корпусе русского языка. Таким образом были получены вектора в 300-мерном пространстве, представляющем различные языковые конструкции – статьи, параграфы, документы.

Полученное семантическое пространство было 2D визуализировано при помощи метода главных компонент и раскрашено по источнику данных. По результату небольшой выборки и графику плотности кластеров (Рис.1) видно, как тексты корпуса русского языка «Opencorpora» отделяются в пространстве от правовых текстов. Небольшое пересечение вызвано наличием в общем корпусе русского языка статей из уголовного кодекса и текстов правового содержания.

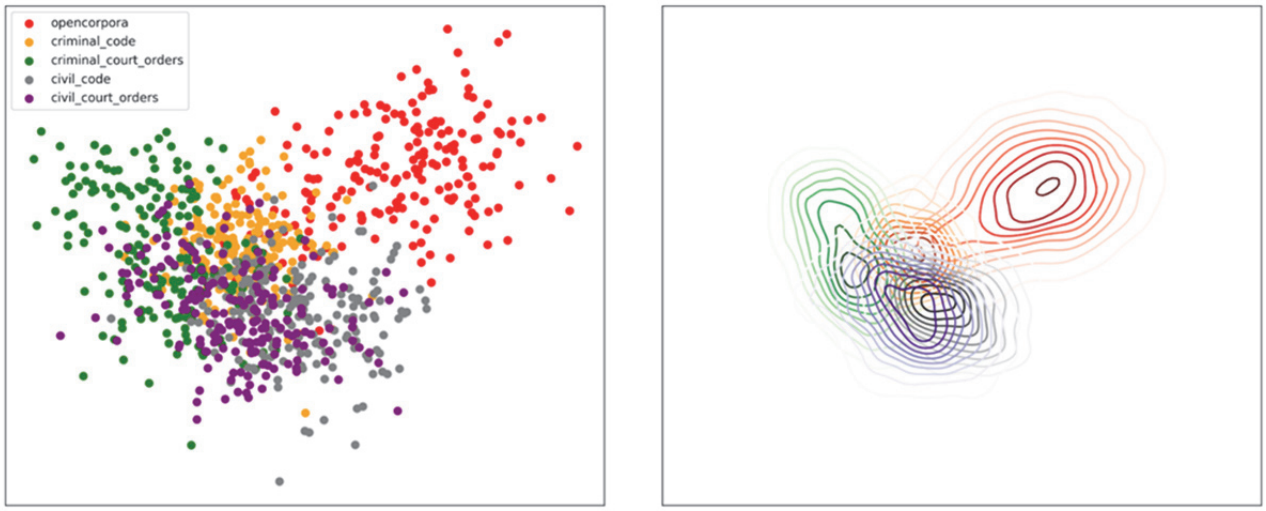

Рисунок 1 – Пересечение правовых пространств УК РФ и ГК РФ с корпусом русского языка.
слева: 500 случайных документов из каждого класса; справа: Оценка плотности кластеров.

Если рассмотреть отдельно Уголовный Кодекс Российской Федерации, тексты судебных решений по уголовным преступлениям и массив текстов общего русского языка, становится заметна тенденция. Тексты художественного и публицистического характера заметно выбиваются из основной массы правовых текстов (**Ошибка! Источник ссылки не найден.**), что свидетельствует о наличии некоторого подпространства юридических терминов в русском языке, которое стоит обособлено от остальных.

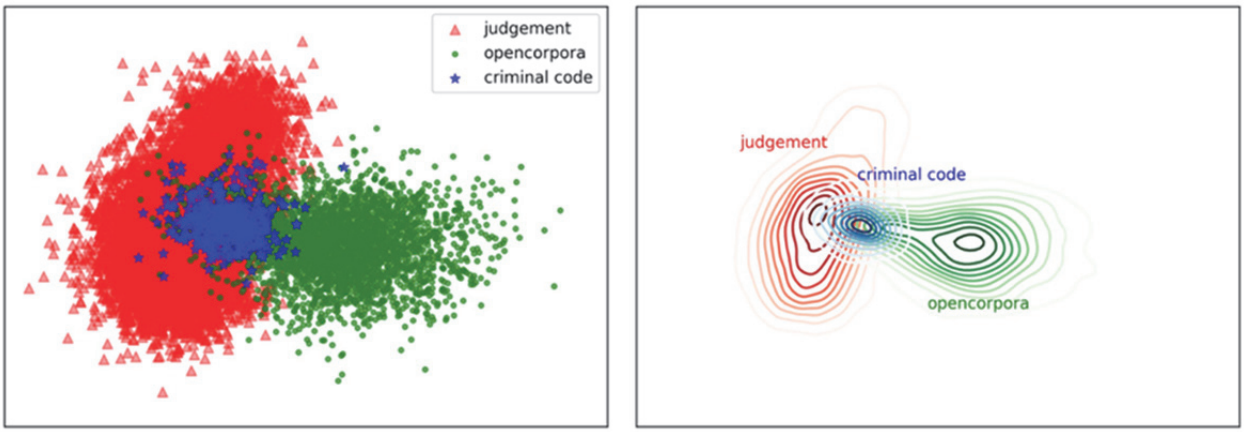

Рисунок 2 – Пересечение правового пространства Уголовного Кодекса РФ и модели естественного языка на примере корпуса «Opencorpora».

Теперь рассмотрим взаимное расположение нормативных актов и общего представления правовых документов. Если рассмотреть тексты судебных решений по уголовным делам и тексты статей УК РФ, а также соответственно судебных решений по гражданским делам и ГК РФ, то их взаимное расположение выглядит следующим образом (Рис. 3)

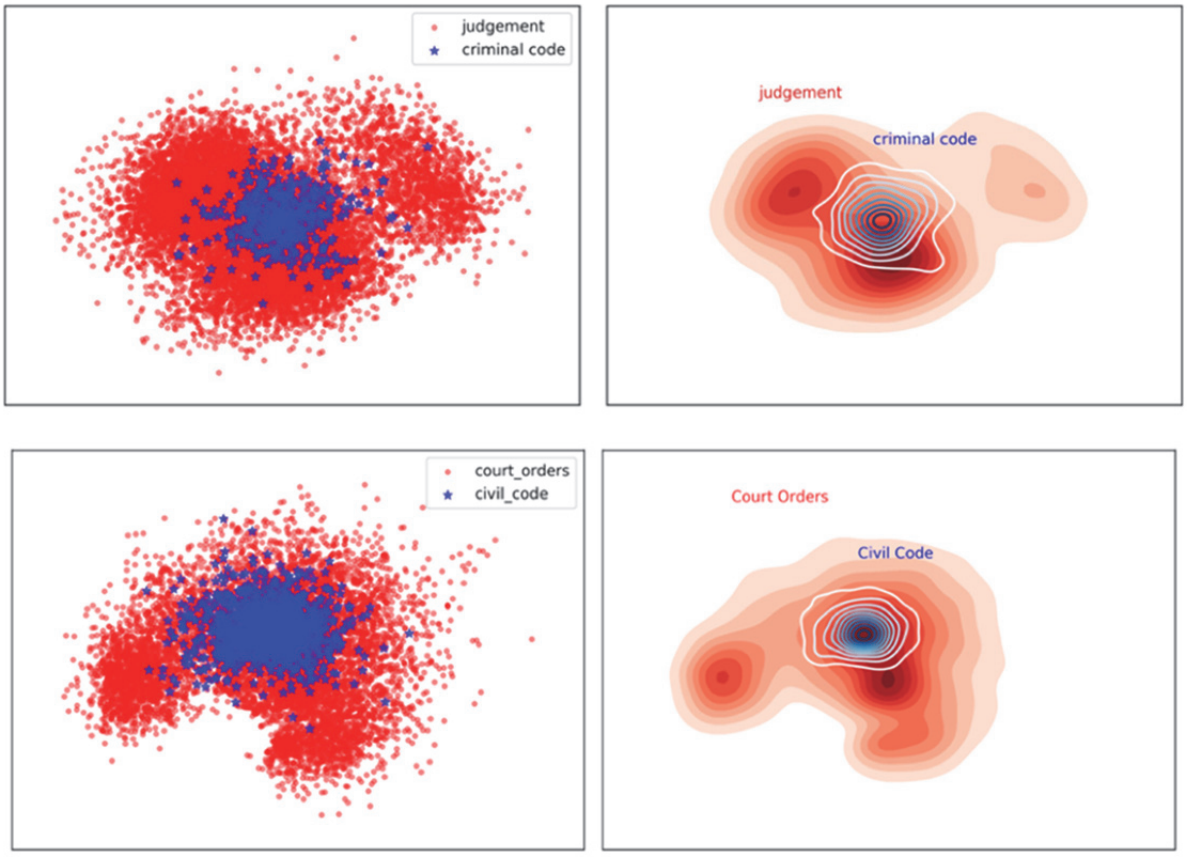

Рисунок 3 – Изображение правового пространства Уголовного Кодекса РФ. (вверху слева). Оценка плотности кластеров. (вверху справа). Правовое пространство Гражданского Кодекса РФ. (внизу слева). Оценка плотности кластеров. (внизу справа)

3.3 Модель на основе алгоритма FastText

Разработанный в Facebook AI Research алгоритм FastText учитывает информацию, которая заключена в отдельных частях слов, что сильно увеличивает качество построения векторных представлений для морфологически богатых языков, к которым относится русский.

Алгоритм FastText не генерирует векторные представления отрывков текста автоматически, для работы с ним в нашей процедуре необходимо было добавить дополнительный шаг. Результатом работы алгоритма являются векторные представления слов. Для того чтобы спроецировать каждый минимально значимый семантический отрывок текста (предложения, либо отдельные части длинного предложения) нами были взяты обученные вектора слов, нормализованы и усреднены. Использование данного подхода позволяет подняться с уровня векторных отображений слов к документам.

Описанный метод позволяет взглянуть на полученное, в результате работы алгоритма векторное пространство (Рис. 4). На представленном ниже изображении можно заметить, что сохранилась общая тенденция, по которой вектора предложений описательных правовых текстов организуются вокруг векторных представлений нормативно-правовых документов. Следовательно, тот факт, что для использования этого алгоритма пришлось опуститься на более детальный уровень рассмотрения текстов – уровень слов, не изменил структуры получаемого правового пространства.

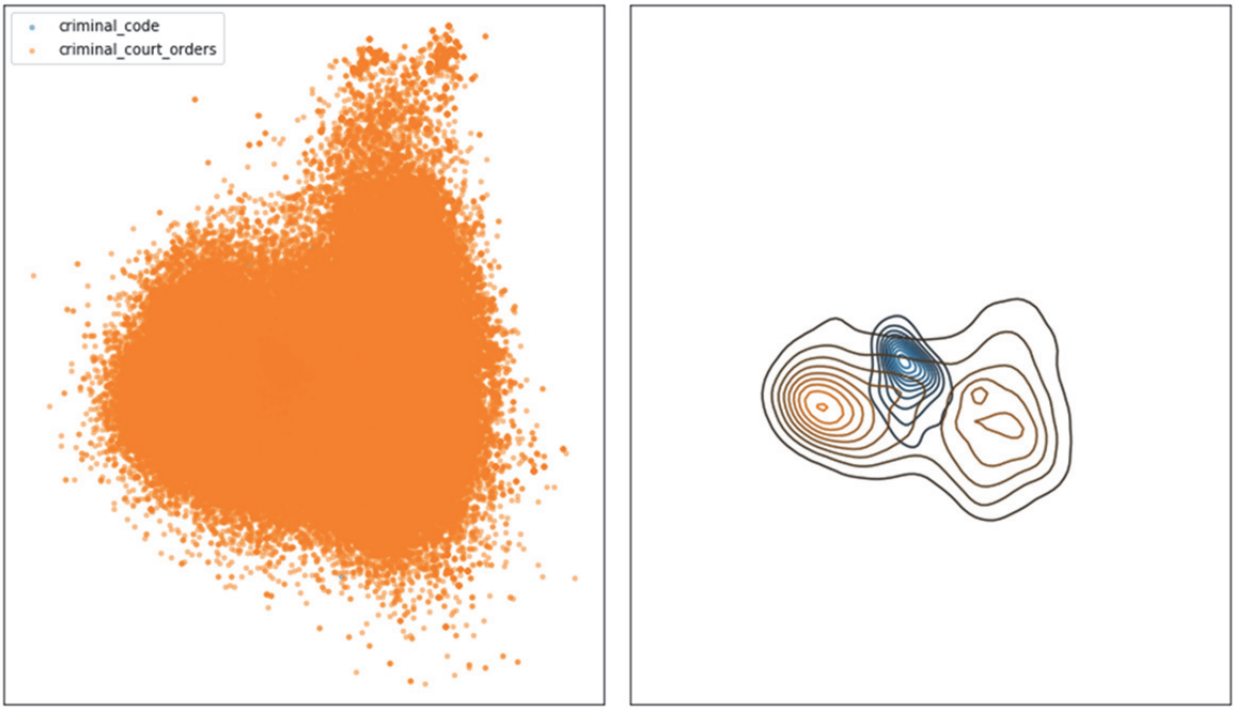

Рисунок 4 – Изображение правового пространства, полученного алгоритмом FastText (слева). Диаграмма плотности кластеров(справа)

3.4 Метрические свойства правового пространства

Использованные выше представления предложений с помощью числовых векторов x_i^m, полученных семантическим эмбеддингом, могут рассматриваться как элементы пространства $\mathbb{R}^n$. Для них существует естественная мера близости, определяемая эвклидовым расстоянием или чаще косинусной мерой похожести, которая обычно используется:

$$sim(x_i, x_j) = \frac{< x_i, x_j >}{\|x_i\|\|x_j\|}$$

Использование этой меры позволяет сравнивать сравнительную близость отдельных предложений и даже реализовать некоторую простейшую «алгебру» предложений. Однако наша задача требует искать связи и соотношения между полными текстами, поскольку именно они формулируют правовые нормы и представляют нарративы для интерпретации, перевода с бытового языка на «юридический».

Пусть текст представляет собой набор из M предложений, тогда его представление в $\mathbb{R}^n$будет множеством

$$X^m = \{x_i^m\}, \qquad i = 1, \dots, M$$

Будем называть далее такое множество созвездием данного текста. В этих терминах все семантическое пространство текстов оказывается множеством таких созвездий. Каждое созвездие уникально, как и текст, однако несколько созвездий могут иметь общие вектора, поскольку порождающие их тексты могут иметь идентичные предложения. Для наделения пространства текстов какими-либо метрическими свойствами, нужно корректно определить близость на множестве их созвездий.

Введем функцию расхождения между созвездиями $\partial(X^a, X^b)$, которая должна обладать следующими свойствами

$$\partial(X^a, X^a) = 0;\ \forall a$$
$$\partial(X^a, X^b) \geq 0;\ \forall a, b$$

При этом будем допускать анизотропию расстояний, то есть считать, что равенство $\partial(X^a, X^b) = \partial(X^b, X^a); \forall a, b$ может нарушаться. Мы определим функцию расхождения созвездия X^a от созвездия X^b как:

$$\partial(X^a, X^b) = \sum_{i=1}^{M} \min_k d\left(x_i^a, x_k^b\right)$$

Здесь используется расстояние между предложениями $d(x_i, x_j)$ как между соответствующими векторами в $\mathbb{R}^n$, согласованное с используемой функции похожести, например, ангулярное:

$$d(x_i, x_j) = \arccos\left(sim(x_i, x_j)\right)/\pi$$

На рис. 5 показано из каких величин расстояния (в работе мы использовали в основном косинусное расстояние) состоит функция расхождения созвездия a, состоящего из двух точек и созвездия b из трёх.

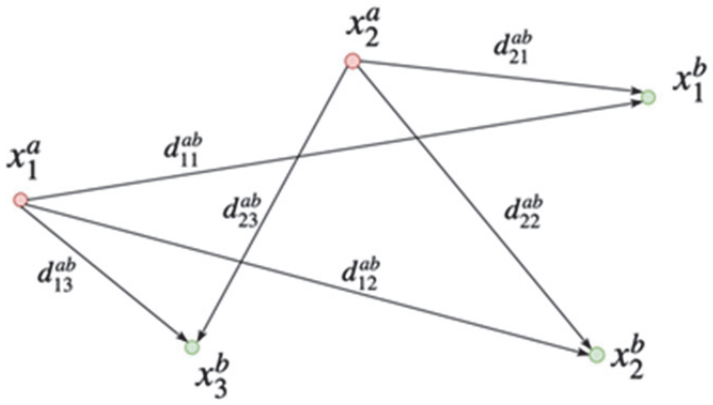

Рисунок 5 – Внутренняя структура функции расхождения созвездий

Описанная функция расхождения созвездий может использоваться для реализации алгоритма поиска k-ближайших соседей. Пусть $k > 0$ – количество текстов, которых необходимо найти как наиболее близкие для текста, представленном в виде созвездия $X^c = \{x_i^c\};\ i = 1, \dots, M$. Мы можем вычислить отличие X^C от остальной коллекции знаков $\{X^1, X^2, \dots, X^S\}$. После этого мы сможем выстроить созвездия в порядке возрастания функции расхождения:

$$\partial(X^c, X^{i1}) \leq \partial(X^c, X^{i2}) \leq \cdots \leq \partial(X^c, X^S)$$

Первые k созвездий в этом порядке будут являться результатом поиска методом k-ближайших соседей.

3.5 Графовая модель пространства нормативных документов

Остановимся на построении графовой модели только для УК РФ как исчерпывающем примере. Сначала посмотрим, как выглядит распределение расхождений созвездий для УК РФ. Как видно, созвездия весьма плотно расположены и заполняют практически весь возможный спектр расстояний

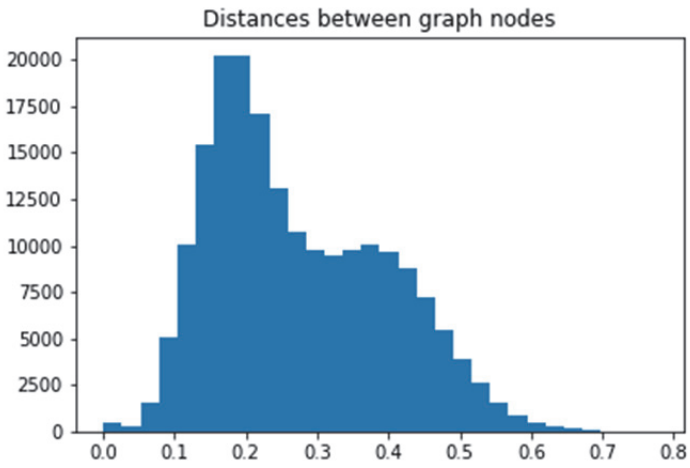

Рисунок 6 – Гистограмма расхождений созвездий для УК РФ

Это говорит о том, что графовая модель, которая может быть построена на созвездиях как вершинах, будет иметь вид полного графа с весами ребер, соответствующим расхождению, распределенными как показывает вышеприведенный на Рис. 5 график. Рассмотрим возможность построения пороговых графов, то есть получаемых из полного путем отсечения ребер с весами меньшими порога. На Рис.7 показаны три графа, полученных при выборе разных порогов отсечения.

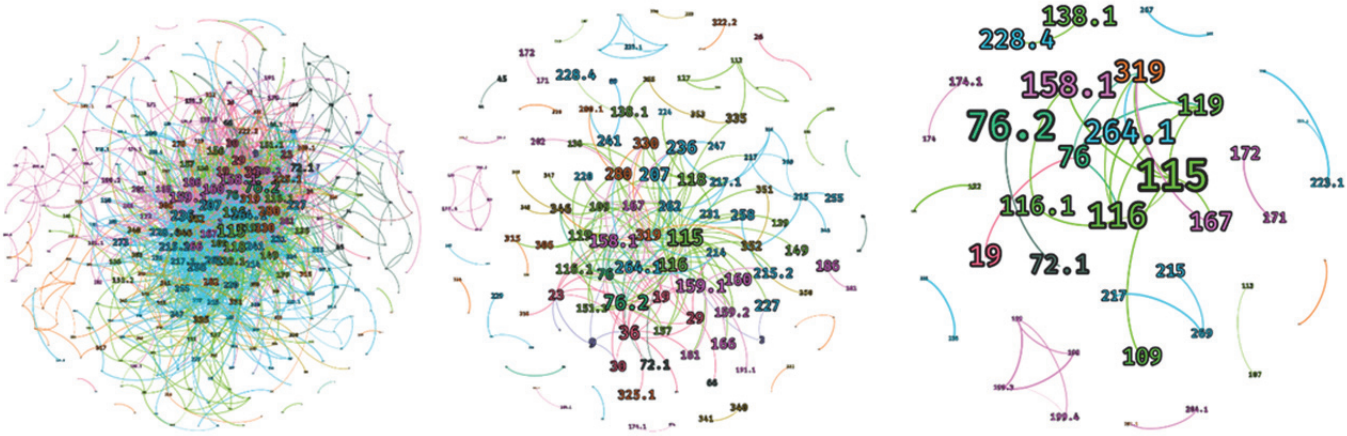

Рисунок 7 – Пороговые графы близости статей УК РФ. Вершины раскрашены цветом по разделам в кодексе, всего 11 разделов. Размер вершины определяется взвешенной степенью. Ребра раскрашены по цветам вершин. Пороговое значение слева направо: 0.4, 0.3, 0.2

На вершинах порогового графа - статьи и части статей УК, а ребра ассоциируются с расхождением, которое стремится к нулю, в том случае, если статьи определяют похожие по описанию на русском языке правонарушения. По полученной матрице расстояний было построено несколько отсеченных графов, с разными пороговыми значениями.

Исходя из информации, представленной на рисунке, можно сказать, что при уменьшении значения порога очевидно выделяются кластеры вершин, относящихся к одному разделу уголовного кодекса, однако также присутствуют нетривиальные связи между статьями, которые нуждаются в изучении и интерпретации, которая выходит за рамки настоящей статьи.

4 Результаты использования модели и возможные приложения

Практическую применимость построенной модели правового пространства мы продемонстрируем на примерах нескольких задач выбора ближайшего соседа из нормативно-правовых документов к некоторому тексту на бытовом языке, в качестве которого будем использовать тексты, полученные извлечением только описательной части различных судебных решений. Такой подход позволяет использовать имеющуюся модель достаточно просто и дает возможность говорить о нем, как основе построения робота-юриста, интерпретирующего нарративный текст описания совокупности фактов статьями УК или ГК РФ. Эту задачу можно рассматривать как простейшую форму перевода текста на бытовом языке в текст нормативного акта, наилучшим образом относящегося к содержанию заданного текста.

Оценим качество подхода на построенной модели. Нами был составлен набор данных, содержащий описательную часть судебного акта и набор статей уголовного кодекса из резолютивной части. Этот набор, по мнению судьи (экспертная разметка профессионального юриста) наиболее точно описывает дело. В такой постановке задачи мы имеем возможность сравнить качество интерпретации алгоритма с экспертной разметкой. Для измерения качества работы нашей процедуры была использована метрика «слабой точности» модели (Weak Accuracy), которая показывает долю текстов, для которых процедура смогла обнаружить хотя бы одну статью УК из размеченных судебным экспертом.

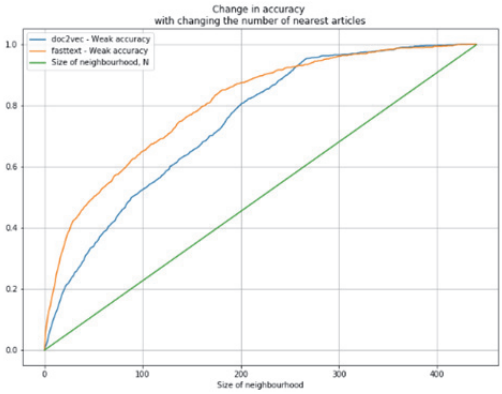

Рисунок 8 – Сравнение метрики "слабой" точности для процедуры, использующей алгоритм doc2vec (синим) и fastText (оранжевым) при увеличении размера окрестности поиска.

Была написана процедура, которая используя модель правового пространства и метод ближайших соседей генерировала на выходе заранее заданное количество статей кодексов, которые алгоритм посчитал наиболее адекватно описывающими представленный на вход отрывок текста. Степень близости измеряется с помощью функции расхождения двух созвездий. Очевидно, что чем больше будет задана область поиска ближайших к тексту соседей, тем больше шанс, что как минимум один из соседей попадет в итоговую выборку, что даст точность 1 при количестве соседей равному числу семиотических знаков. Качество процедуры выше, если в маленькую область поиска попадает наибольшее количество корректно предсказанных статей УК. В работе мы показали, как величина метрики «слабой точности» зависит от размера выбранной окрестности. На Рис. 8 можно увидеть, как изменяются значения метрики для постепенно увеличивающейся области поиска ближайших соседей в сравнении для модулей, использующих алгоритмы doc2vec и fastText соответственно. Из данных результатов, которые далеки до идеальных можно сделать два утверждения: разработанная процедура работает и даёт положительные результаты; применение алгоритма fastText повышает качество работы данной процедуры и дает заметный прирост в точности для более маленьких окрестностей.

Полученные результаты говорят, что пока полученного качества работы модели, видимо, недостаточно для практического применения в ответственных областях права, таких как судебные разбирательства. Однако, авторы посчитали возможным выполнить пробную имплементацию модели в прототипе консультирующего робота-юриста. Для постоянного совершенствования модели в прототип встроен механизм инкрементального обучения [19]. Он построен на основе интерактивного взаимодействия с пользователем, который имеет возможность корректировать перевод, выполненный роботом. Эта коррекция модифицирует созвездие введенного пользователем текста, приближая его к созвездию, соответствующему введенному нормативному акту.

В ходе работы был разработан и протестирован прототип механизма работы с процедурой в режиме автономного робота – чат бота на базе мессенджера «Telegram».

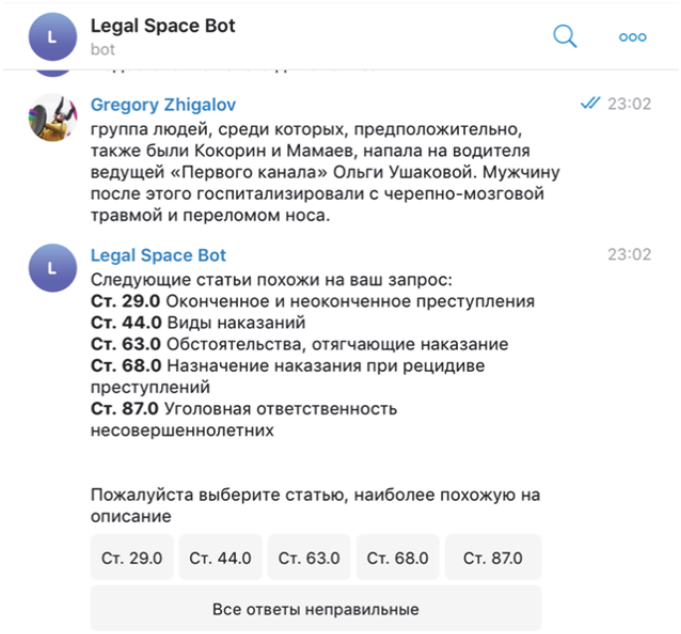

Рисунок 9 – Интерфейс чат-бота помощника юриста в мессенджере Telegram

Разработанный чат-бот предназначен для работы с аудиторией, которая может сформулировать описательный текст в терминах близких к правовым. На данном этапе решение может быть полезно для помощников судей, адвокатов и профессиональных юристов.

В роботе-помощнике предусмотрен механизм сбора пользовательских оценок (Рис. 9.) результатов работы модели, что в дальнейшем может быть использовано для подкрепления и переобучения модели на корпусах пользовательских текстов. Обученные и сериализованные на этапе тренировки модели и вектора для семиотических знаков сохранялись в облачное файловое хранилище. Модели в автоматическом режиме принимались в обработку вспомогательной системой робота и предыдущую версию алгоритмов процедуры, обеспечивая постоянный цикл обновления и актуальности версии процедуры внутри робота.

Разработанное решение выложено в открытом доступе и может быть использовано для реализации автоматических систем с похожим функционалом. Данный опыт показывает, что процедура перевода на основе разработанной модели может успешно использоваться в продуктивной среде, для решения инженерных задач связанных с анализом текстов.

5 Выводы и направление будущих работ

Построение роботов-юристов будет тем успешнее, чем более тонкие математические модели будут применены для реализации основных выполняемых ими процессов. Сравнивая основной процесс выполняемый юристами – трактовку текстов в терминах нормативно-правовых актов, с работой переводчиков, авторы пришли к выводу о необходимости в первую очередь разработки модели правового пространства, как лингвистической, представленной эмбеддингом в высокоразмерное числовое пространство. Применяя современные алгоритмы эмбеддинга doc2vec и FastText, были построены модели, представления как юридических, так и бытовых текстов. Для корректного отражения семантических связей в правовом пространстве был разработан и применен алгоритм нахождения расхождений созвездий текстов. На основе вычисления расхождений созвездий нормативно-правовых актов и нарративных текстов разработаны процедуры нахождения ближайших нормативных актов к текстам на бытовом языке. Была исследована численно плотность распределения эмбеддингов УК РФ как примера нормативно-правовых актов и получено семейство пороговых графов частей статей УК РФ. Было показано, что кроме естественной иерархии и сущности статей, в УК присутствует немало необъясненных сильных связей между внешне отдаленными статьями. Исследование таких связей может быть одним из направлений будущих работ. Наряду с этим предполагается перспективным направлением совершенствование модели правового пространства путем исследования более контрастивных алгоритмов оценки расхождения текстов, которые позволили бы снизить плотность окружения созвездий нормативно-правовых актов.

Благодарности автора

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта №19-011-00320.

Список использованной литературы

[1] Shannon Liao, World’s first robot lawyer’ now available in all 50 states, https://www.theverge.com/2017/7/12/15960080/chatbot-ai-legal-donotpay-us-uk
[2] Сбербанк внедрил робота-юриста, который заменит 3 тыс. сотрудников https://www.business-gazeta.ru/news/334226
[3] О. Луговой, В Казани начали создавать робота-юриста, способного предсказать решение судьи, https://www.nnov.kp.ru/daily/26599/3615075/
[4] Робот, а не человек: как искусственный интеллект перестроит работу юристов, https://pravo.ru/story/view/131655/
[5] Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (2013). *Handbook of latent semantic analysis.* Psychology Press.
[6] T. Hofmann, «Probabilistic Latent Semantic Analysis,» в Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, San Francisco, CA, USA, 1999.
[7] Moody, C. E. (2016). Mixing dirichlet topic models and word embeddings to make lda2vec. arXiv preprint arXiv:1605.02019.
[8] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. CoRR, abs/1310.4546.
[9] Jeffrey Pennington, Richard Socher, Christoper D. Manning, Glove: Global Vectors for Word Representation, Conference: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp.1532-1543.
[10] Joydeep Bhattacharjee, fastText Quick Start Guide, Publisher: Packt Publishing, Release Date: July 2018
[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv preprint arXiv:1810.04805, 24 May 2019.
[12] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer, Deep contextualized word representations, arXiv preprint arXiv:1802.05365, 22 Mar. 2018.
[13] Сайт проекта «Открытый корпус» (OpenCorpora). (б.д.). Получено из http://opencorpora.org
[14] Государственная Автоматизированная система Российской Федерации "Правосудие". (б.д.). Получено 24 Май 2019 г., из https://sudrf.ru
[15] Mihalcea, R., & Tarau, P. (2004). Textrank: Bringing order into text. Proceedings of the 2004 conference on empirical methods in natural language processing.
[16] Морфологический анализатор текста на русском языке mystem [Электронный ресурс]; // Компания Яндекс [сайт] — 2003–2013. — URL: http://company.yandex.ru/technologies/mystem/
[17] Ulla, G., & Virpi, H. (May 2017 г.). Tones and traits - experiments of text-based extractions with cognitive services. Finnish Journal of eHealth and eWelfare, 9, 82-94.
[18] Открытый программный код исследования. (б.д.). Получено из https://github.com/donfaq/legal-space-research
[19] Alexander Gepperth, Barbara Hammer, Incremental learning algorithms and applications, ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ.,