УДК 004.738.52

МЕТОД АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ ПОЛНОТЕКСТОВЫХ ОПИСАНИЙ КЕРНОВ С ИСПОЛЬЗОВАНИЕМ СЛОВАРЕЙ

А. П. Антонов[1] [0009-0007-3642-7734], С. А. Афонин[2] [0000-0003-3058-9269],
А. С. Козицын[3] [0000-0002-8065-9061], В. М. Староверов[4] [0000-0001-8289-2273]

[1, 4]*Московский государственный университет им. М. В. Ломоносова, г. Москва, Россия*

[2, 3]*Научно-исследовательский институт механики МГУ им. М. В. Ломоносова, г. Москва, Россия*

[1]alexey.p.antonov@gmail.com, [2]serg@msu.ru, [3]alexanderkz@mail.ru, [4]staroverovvl@yandex.ru

Аннотация

Использование методов автоматической обработки текстов, в том числе методов классификации полнотекстовых описаний, позволяет достичь существенного снижения трудозатрат при обработке экспериментальных данных. В настоящей работе рассмотрено применение метода автоматической классификации текстов в области обработки и классификации элементов керна и определения литофаций. Литофациями называют одновозрастные геологические тела (отложения), которые по своему составу или строению отличаются от соседних слоев.

При проведении оценки нефтегазового потенциала месторождений требуется выполнять построение карт и схем распространения литофаций. Для этого необходимо осуществить классификацию большого количества полнотекстовых описаний участков керна, выполненных специалистами. Алгоритм, представленный в статье, позволяет на основе заданных правил и словарей провести классификацию с учетом порядка и значимости ключевых слов в предложениях. Преимуществами такого подхода являются возможность различать близкие литофации, возможность использования архивных данных, простота настройки на новые классы, адаптация к русскоязычным описаниям кернов и возможность локального использования без необходимости передавать описания кернов сторонним приложениям.

***Ключевые слова**: классификация текстов, литофации, словари, информационные системы.*

ВВЕДЕНИЕ

Опыт крупных мировых компаний показывает, что для увеличения эффективности разведки и разработки нефтяных и газовых месторождений необходимо внедрять в производственный процесс методы машинного обучения, в том числе при разведке и оценке месторождений [1]. Одной из важных задач при проведении оценки нефтегазового потенциала месторождений является построение карт и схем распространения литофаций [2]. Литофациями называют одновозрастные геологические тела (отложения), которые по своему составу или строению отличаются от соседних слоев. Классификация литофаций является двухуревневой. На первом уровне определяется название фации («фация морен», «фация аллювиальных конусов выноса», «фация речной поймы» и др.), на втором уровне – компонента лито- («алевро-глинистая», «углисто-алевролитовая», «битуминозно-кремнисто-глинистая» и др.). Одним из часто используемых методов построения карт распространения литофаций является анализ результатов бурения. Полученные в процессе бурения керны исследуются специалистами и описываются в свободном текстовом формате с разделением всей глубины керна на отдельные однородные участки. Результатом такого анализа является набор данных, включающий координаты и параметры скважины, глубину и полнотекстовое описание состава породы, полученной в результате бурения пробы. На основе выполненного текстового описания (литологического описания керна, содержащего, обычно, от одного до десяти предложений) необходимо сопоставить каждому участку керна один из заданных классов литофаций. Следует отметить, что в разных исследовательских группах и организациях существуют различные стандарты на классификаторы литофаций. В зависимости от предъявляемых требований и поставленных задач могут использоваться классификаторы, содержащих от 5–8 классов до сотни классов.

Автоматизация процесса проведения такой классификации позволяет значительно упростить и унифицировать обработку полнотекстовой информации, полученной от различных специалистов по десяткам тысяч кернов за последние

десятилетия, при анализе описаний кернов со скважин исследуемого региона. Возможные ошибки в работе впоследствии могут корректироваться при построении карт за счет сглаживания данных на соседних участках.

ОБЗОР СУЩЕСТВУЮЩИХ ПОДХОДОВ

Использование методов машинного обучения и автоматизация процессов определения литофаций керна в настоящий момент развиваются по трем основным направлениям. Наибольшее количество работ посвящено автоматизации распознавании изображений шлифов. Здесь следует отметить такие инструментальные средства, как ИС ABAI [3], сервис DeepCore копании Digital Petrelium [4], комплекс DHD [5], программный комплекс «Цифровой керн» [6] и «Нейросетевое распознавание текстурных особенностей графических керновых данных» [7].

В составе ИС ABAI (Advanced Base Artificial Intelligence) модуль «Автоматическая интерпретация керна» позволяет проводить автоматическое определение литофации по фотографии керна с использованием методов распознавания изображения. Обучение системы проводилось по 1300 скважинам. Классификатор содержит пять типов фаций и отдельный шестой тип для неопределенных изображений, может обрабатывать как фотографии, так и видео. Разрабатывается и эксплуатируется казахской компанией «КазМунайГаз».

Сервис DeepCore компании ООО «Диджитал Петролеум» производит обработку изображений кернов с использованием сверточных нейронных сетей. Обученная модель производит классификацию изображений кернов по 14 классам с точностью 70%. Сервис не заменяет полностью эксперта-геолога, однако, согласно данным, приведенным в статье [4], использование сервиса позволяет ускорить работу по описанию литофаций керна в 7 раз.

В программном комплексе DHD [5] компании Шлюмберже реализован модуль цифрового анализа керна (ЦАК), который по данным рентгеновских микрофотографий строит трехмерную цифровую модель керна с возможностью оценки ее физических свойств и сегментации.

Программный модуль «Цифровой керн» [6] компании Норникель позволяет на основе изображения керна в видимом диапазоне описывать характеристики керна и проводить классификацию участков керна по процентному содержанию сульфидов в исследуемом образце. Это позволяет в режиме online находить

и анализировать наличие рудного материала в керне по фотографии и с высокой вероятностью определять процент рудной минерализации.

Подобные решения позволяют существенно сократить трудозатраты при анализе данных бурения, поскольку не требуют или значительно уменьшают ручную работу специалистов по составлению описания керна. Однако такая технология неприменима для уже накопленного экспериментального материала. Кроме того, количество выделяемых классов фиксировано для каждой системы и оказывается существенно меньшим, чем при традиционных методах обработки текстовых данных.

В ряде работ предложено проводить классификацию литофаций по числовым характеристикам и измеряемым физическим свойствам пород, например, в работе [8] сравнены результаты классификации пород на 4 класса в соответствии со значением функций давления и внутреннего трения между частицами породы. Подобные методы также неприменимы для распределения по большим классификаторам, содержащим десятки классов.

В работе [9] исследованы методы определения физических характеристик керна на основе изображений, в том числе с использованием компьютерной томографии, что позволяет упростить получение данных о характеристиках кернового материала, а также дополнить результаты лабораторных и натурных исследований свойств пластов.

В ряде работ для распознавания и классификации полнотекстовых описаний предложено использовать современные системы с искусственным интеллектом. Например, в работе [10] описан метод классификации описаний с использованием векторных моделей текстов и нейронных сетей. Для проведения анализа полнотекстовых описаний кернов авторы преобразуют текст в векторное описание в модели GeoVec [11]. Эта модель построена на основе модели GloVe[12], обученной на текстах 280 тыс. англоязычных статей по геологии, доступных для скачивания через Elsevier ScienceDirect APIs, и отобранных вручную страницах Википедии "List_of_rock_types", "List_of_minerals", "List_of_landforms", "Rock_(geology)", "USDA_soil_taxonomy", "FAO_soil_classification" и др. Обученная модель позволяет автоматически определять близость отношений геологических

понятий, например, оказывается близкой векторная разность таких пар, как «гранит» – «магматический», «гнейс» – «метаморфический», «известняк» – «осадочный», «туф» – «вулканический», или пар «песок» – «песчаник», «гравий» – «конгломерат». На основе построенной модели авторы вычисляют для каждого предложения усреднение вектора его слов в модели GeoVec, которые подают на вход обученной нейронной сети. Классификация текстовых описаний кернов производилась по 18-ти классам литофаций.

Основным недостатком подобного подхода является отсутствие учета порядка слов в предложении. Поскольку для обучения и анализа в качестве входных данных используется усредненный вектор слов, элементы описания «глины с вкраплениями песка» и «песок с вкраплением глин» становятся неразличимыми. Соответственно, такой подход неприменим для проведения анализа с целью распределения текстов по детализированным классификаторам литофаций, включающим такие классы, как «глинисто-песчанные» и «песчанно-глинистые». Дополнительным ограничением применения машинного обучения на основе векторных моделей представления текстов является отсутствие предобученных моделей на русском языке, аналогичных GeoVec. Использование пословного перевода обученных моделей не дает удовлетворительного результата ввиду многозначности значительной части слов, используемых в английском и русском языках.

В работе [13] рассмотрена задача классификации полнотекстовых описаний кернов по 9 классам с использованием подходов, основных на сверточных нейронных сетях для классификации текста (TextCNN), сетей двунаправленной длительной-кратковременной памяти (BiLSTM) и сетей представлений двунаправленного кодера (BERT). Процент правильного распознавания существенно зависит от класса. Для трех классов вероятность правильного распознавания составила 99%, для оставшихся шести классов – от 24% до 32%. Для токенизации была использована модель RuBERT, обученная на русскоязычном варианте Википедии и новостных лентах, поскольку специализированные геологические модели для русского языка отсутствуют.

В этой связи для построения систем более точной классификации необходимо использование моделей и алгоритмов, учитывающих порядок слов в предложениях и адаптированных к русскому языку. Для этого можно применить методы, которые используются в классических задачах тематического анализа по ключевым словам [14, 15].

ОПИСАНИЕ АЛГОРИТМА

Разработанный нами алгоритм опирается на использование словарей, составленных геологами, с описанием характеристик различных литофаций. На вход алгоритму поступает полнотекстовое описание участка керна, выполненное специалистом при анализе результатов бурения. Результатом работы алгоритма является ранжированный список возможных литофаций, которые соответствуют заданному текстовому описанию. Полнотекстовое описание керна дается в свободном формате, но, как правило, содержит строгие формулировки. Например, «Переслаивание мощных пачек чередования хорошо сортированных песчаников, алевролитов с преимущественно песчаными прослоями в верхней части разреза и алевролитовыми – в нижней. Текстура пород преимущественно линзовидная, волнистослоистая».

Настройка алгоритма на специфику предметной области производится при помощи формирования словаря описаний фаций на основе возможных признаков фации, которые должны встречаться в описании (характерные для данной фации) или, наоборот, не могут встречаться в ее описании. Например, в описании фации русла рек не могут встречаться морские организмы, а в описании морен не может встречаться текстура воздушной ряби и орагнические включения в виде кораллов и трилобитов. Каждый признак является словом или словосочетанием и относится к определенному типу. В текущей программной реализации были рассмотрены следующие характеристики: название породы (например, «песчаник», «глина», «аргиллит», «песок»), ее цвет (например, «серого», «бежевого», «бурого», «рыжего»), структура (например, «псаммитовая», «алевритовая», «мелкозернистые»), текстура (например, «горизонтальная», «линзовидная», «массивная»), включения флоры и фауны (например, «трилобиты», «кораллы»,

«криноидеи», «радиолярии»), окатность (например, «неокатанные», «угловатые», «плохо окатанные»), сортировка (например, «сортировка плохая», «сортировка средняя»), границы (например, «волнистые», «ровные», «нечеткие»). При необходимости список анализируемых характеристик может пополняться. В качестве словосочетаний рекомендуется использовать пары вида «существительное прилагательное» или «наречие причастие», например, «граница ровная» или «хорошо окатанные». Такой подход позволяет получить достаточно точные формализованные критерии, которыми пользуются геологи при решении аналогичной задачи в ручном режиме.

Для учета специфики формирования полнотекстовых описаний используют вспомогательные словари, которые не несут в себе информации о предметной области, но позволяют правильно расставлять акценты и определять значимость ключевых элементов описания. Словарь ослабляющих слов и выражений с глаголами (например, «изредка встречаются») определяет слова и выражения, после которых значимость всех ключевых терминов уменьшается до конца предложения или глагола. Словарь усиливающих слов и выражений с глаголами (например, «основной», «превалирует», «обильно», «часто», «значительно», «многочисленный») позволяет задавать усиление значимости всех ключевых терминов до конца предложения или глагола. Кроме того, используется словарь глаголов-исключений, которые не прерывают действие ослабляющих и повышающих слов (например, «обладать», «содержать», «содержаться», «слагать»), словари синонимов (например, «аргиллит – глина – глинистые») и гиперонимов (например, «органогенный детрит: шлам, раковинный детрит, обломки раковин»). Следует отметить различие в обработке синонимов и гиперонимов. При сравнениях все синонимы считаются совпадающими терминами с учетом транзитивных зависимостей. Гипонимы и гиперонимы при обработке текстовых описаний считаются совпадающими терминами, но без учета транзитивных зависимостей между терминами. Словарь «лито» содержит возможные определения для каждой составляющей возможных описаний каждой компоненты лито (например, «глинистая, глина, глинисто, аргиллит, аргиллитовый, аргиллитовая»).

Разработанный алгоритм состоит из четырех этапов. Схема алгоритма представлена на рис. 1.

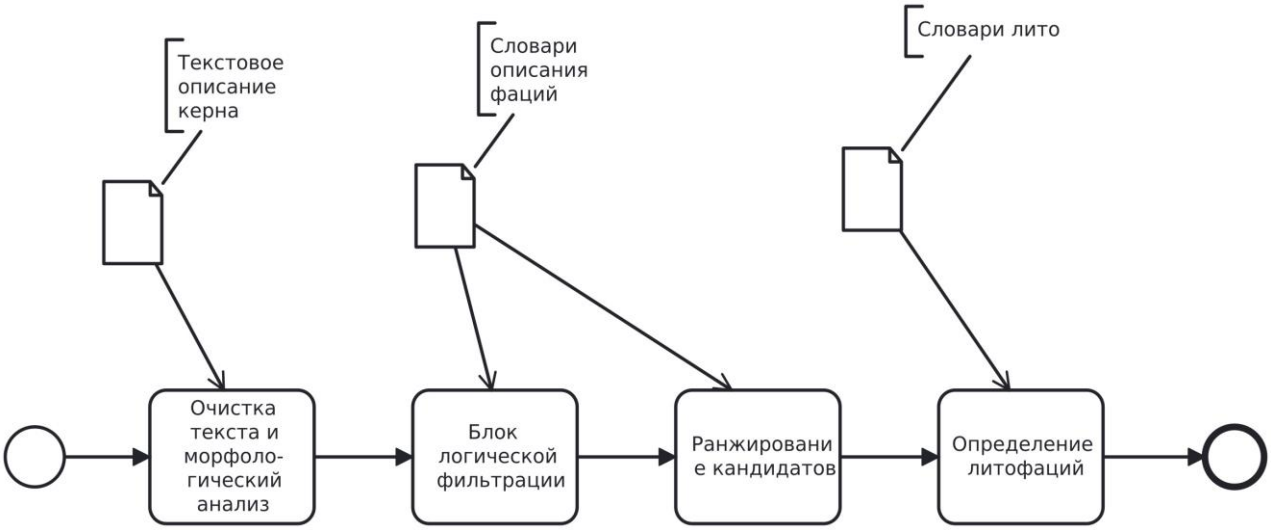


Рис. 1. Схема работы алгоритма классификации

На первом этапе работы алгоритма производятся подготовка текста для анализа, включая очистку текста от лишних символов, разметка предложений, разметка областей, заключенных в скобочки, а также проводится морфологический анализ. Для морфологического анализа используется свободно распространяемая библиотека pymorphy3 для языка Python, которая позволяет получать по заданной словоформе возможные варианты нормальной формы слова. В случае наличия нескольких вариантов алгоритм рассматривает все возможные случаи и для каждого ищет совпадения в используемых словарях. В процессе анализа для каждого найденного термина (слова или словосочетания) указывается его относительная позиция в тексте, принадлежность словарям характеристик фаций, наличие повышения или понижения его значимости в соответствии с наличием повышающих («преобладает», «преимущественно») и понижающих («иногда бывать», «изредка», «иногда», «прослои») слов. После повышающего или ослабляющего выражения до конца предложения, другого повышающего слова или глагола не из списка глаголов исключений все термины помечаются соответственно повышенными или ослабленными. Такой подход позволяет корректно обрабатывать описания вида «ниже залегает толща доломиты серых с прослоями мергелей и ангидритов». Для данного случая будет отмечено, что основной породой являются доломиты, а мергели и ангидриты имеют второстепенное значение, поскольку встречаются только в виде прослоек. Результатом работы первого этапа

алгоритма является структурированное описание анализируемого участка керна с разметкой позиций и значимости терминов.

В блоке логической фильтрации осуществляется фильтрация фаций-кандидатов по запрещающим правилам. Для каждой фации в словарях указывается, каких терминов из различных типов описаний в них не может встречаться. В случае, если хотя бы один из запрещенных для фации терминов встретился в анализируемом описании керна, указанная фация исключается из списка возможных кандидатов. Кроме того, фация удаляется из списка кандидатов, если в описании керна не встретилось ни одного термина из какого-либо значимого (непустого и с коэффициентом значимости «1») описания данной фации. При анализе встречаемости терминов в описании фаций используются словари синонимов («песок, песчанник») и гипонимов («красный: светло-красный, темно-красный, алый»). Результатом работы второго этапа алгоритма является список фаций, характеристики которых не противоречат анализируемому описанию участка керна.

В блоке ранжирования производится анализ списка всех терминов, найденных в описании керна, отдельно для каждого типа характеристики. Итоговый ранг фации вычисляется как сумма баллов, набранных фацией за соответствие каждого типа характеристики рассматриваемому описанию керна по следующей формуле:

$$css \sum_{i \in H} \sum_{w \in W} \begin{cases} dsk_i ss_{p_i} \cdot t_{iw}, & w \in D_i, \\ -k_2 \cdot dsk_i \cdot ssn_{p_i} \cdot t_{iw}, & w \notin D_i, \end{cases}$$

где CSS– нормирующий коэффициент фации, зависящий от количества описанных типов характеристик для рассматриваемой фации, dsk_i– коэффициент значимости типа признака *i* в расчете ранга, ss_j – коэффициент учета в ранге найденных в описании слов из признака со значимостью *j*, ssn_j – коэффициент учета в ранге не найденных в описании слов из признака со значимостью *j*, p_i – заданная экспертом значимость признака *i* для рассматриваемой фации (принимает значение 1, 2, 3 (1 – наиболее важное, 3 – наименее важное)); k_2 – размер штрафа за ненайденное слово, $t_{iw} = clf_i \cdot t_{wi}(w)$ – вес термина w_i; clf_i – способ учета длины описания типа признака *i* в рассматриваемой фации, в зависимости от значения параметра обучения либо 1, либо $1/n_i$; n_i – количество терминов в описании типа

признака *i* для фации, $t_{wi}(w)$ – функция учета idf для слова w, H – множество признаков фации, W – множество слов в описании, D_i – описание признака фации.

Конкретные значения коэффициентов определяются на этапе обучения модели. Результатом работы третьего этапа алгоритма является ранжированный список наиболее вероятных фаций, соответвующих текстовому описанию участка керна.

На последнем этапе для каждой найденной фации выбирается название лито из списка возможных значений для каждой фации, заданного экспертом в словаре. Например, для фации кам возможными лито будут «глинисто-песчаная» и «песчаная», для фации аллювиальных конусов выноса возможными фациями будут «алевро-песчаная», «мелко-грубообломочная» и «грубо-мелкообломочная».

Основным принципом построения названия лито является повышение приоритета слова от начала к концу термина. Например, название лито «углисто-алевро-глинистая» означает, что его основу составляют глины, в меньшей степени встречаются алевролитовые слои и в совсем небольшом количестве в образце присутствуют угольные вкрапления. Для поиска подходящего названия лито из описания керна выделяются все термины, сортируются с учетом позиции текста и значимости (значимые передвигаются вперед, незначимые – назад), после этого все термины с дефисом переворачиваются. Для каждого возможного лито определяется порядок встречаемости его терминов в получившемся списке. Описания лито, термины которых не встретились в списке или встретились не в том порядке, исключаются из рассмотрения. Среди оставшихся названий лито выделяется название, термины которого оказались ближе к началу списка. Результатом работы данного шага алгоритма является наиболее вероятное название лито для каждой рассматриваемой фации. Дополнительно в случае нахождения единственного возможного названия лито или нескольких возможных вариантов лито к рангу фации, вычисленному на предыдущем шаге, добавляется дополнительный коэффициент, поднимающий ее выше в ранжированном списке возможных вариантов фаций.

Результатом работы алгоритма является ранжированный список фаций с указанием лито для каждой фации. Этот список может использоваться для автоматической или автоматизированной классификации. В случае автоматической обработки описанию керна сопоставляется литофация с наибольшим рангом. В случае автоматизированной работы пользователю вместе с описанием предлагается выбор из нескольких фаций, с наиболее высоким рангом, что позволяет значительно сократить время поиска среди большого списка позиций.

ПРОГРАММНАЯ РЕАЛИЗАЦИЯ И РЕЗУЛЬТАТЫ

Программная реализация алгоритма выполнена на языке Python 3.11. Морфологический анализ осуществляется с использованием библиотеки pymorphy3. Для работы также используются библиотеки dataclass, math и numpy. Для описания словарей использовались файлы в текстовом формате. Такой подход позволяет легко редактировать словари в любом текстовом редакторе, например, создавать описания новых фаций или корректировать существующие. Для ускорения работы по исходным словарям в процессе работы в автоматическом режиме строятся дополнительные индексные файлы.

Процесс обучения разработанной программной реализации алгоритма заключался в подборе значений параметров алгоритма, указанных в табл. 1. Обучение проводилось на основе обучающей выборки из 66 примеров, для каждого из которых экспертом был выбран правильный вариант ответа. Подбор параметров производился с использованием метода градиентного спуска. Метрика качества L для проведения обучения рассчитывалась по следующей формуле

$$L = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{k_i},$$

где N – количество тестовых примеров, k_i – порядковый номер правильного ответа в полученном ранжированном списке для примера i.

В случае наличия в ранжированном списке нескольких примеров с одинаковым рангом порядковый номер для них усредняется. Например, для ранжированного (с указанием ранга) списка A(5.2), B(4.7), C(4.7), D(4.7)), E(1.6) при правильном ответе B значение k_i будет равняться 3 (средняя позиция B, C и D). В результате обучения на обучающей выборке было достигнуто значение метрики качества L = 0.681.

Табл. 1. Параметры обучения алгоритма.

Параметр	Возможные значения	Оптимальное значение
Учитывать ли длину описания фации	0 – нет, 1 – да	0
Нормирующий коэффициент фации, зависящий от количества описанных типов характеристик для рассматриваемой фации	0 – не учитывается, 1 – учитывается линейно; иначе с заданным основанием логарифма (например, 10)	10
Коэффициент учета в ранге найденных в описании слов признака со значимостью 1	Число	1
Коэффициент учета в ранге найденных в описании слов признака со значимостью 2	Число	0.9
Коэффициент учета в ранге найденных в описании слов признака со значимостью 3	Число	0.8
Коэффициент учета в ранге не найденных в описании слов со значимостью 1	Число	1
Коэффициент учета в ранге не найденных в описании слов со значимостью 2	Число	2/3
Коэффициент учета в ранге не найденных в описании слов со значимостью 3	Число	4/9
Коэффициент понижения значимости для «незначимых» слов	Число	0

Размер штрафа за ненайденное слово	Число	1
Тип учета idf	Целое число: 0 – не учитывается, 1 – учитывается линейно, иначе учитывается как логарифм с указанным основанием	0
Коэффициент значимости типа признака в расчете ранга	Список значений коэффициентов для каждого типа признака	{"rock":2, "Color":0.6, "structures":1, "texture":1, "inclusion":1, "inclusionorg":1, "roundness":1, "sort":1, "border":1}
Добавочный коэффициент за единственное найденное название лито	Число	3
Добавочный коэффициент за несколько возможных найденное названий лито	Число	0.5

Полученные в результате процесса обучения значения параметров использовались для проведения тестирования результатов работы на тестовой выборке описаний кернов.

Тестирование проводилось на 58 примерах. Результаты тестирования: L = 0.576, на первом месте нужная фация при выборе из 42 фаций оказалась в 25 примерах, на втором – в 7 примерах, на третьем – в 8 примерах. Компонента лито была определена правильно в 46 примерах.

ЗАКЛЮЧЕНИЕ

Представленный алгоритм классификации полнотекстовых описаний кернов может использоваться для автоматизации процесса определения классов литофаций при построении литофационных карт, в том числе в разрабатываемых в настоящее время системах, которые должны заменить Petromod в национальных корпорациях. При обработке специалистом описаний кернов алгоритм подбирает наиболее вероятные классы, сокращая время разметки исходного материала. Преимуществами алгоритма являются возможность обработки архивных данных и данных сторонних исследований, адаптация к русскому языку, возможность локального использования, а также возможность учета порядка слов в описаниях.

СПИСОК ЛИТЕРАТУРЫ

1. Искусственный интеллект в нефтегазовой индустрии Китая. URL: https://nntc.pro/tpost/h2hoet4se1-iskusstvennii-intellekt-v-neftegazovoi-i (дата обращения: 11.12.2025)

2. *Антонов А.П., Афонин С.А., Козицын А.С. и др*. Автоматизированное построение реалистичных литофациальных карт методами комбинаторной оптимизации // Интеллектуальные системы. Теория и приложения. 2024. Т. 28, № 4. С. 5–20.

3. Информационная система ABAI. URL: https://kmge.kz/abai/ (дата обращения: 11.12.2025)

4. *Барабошкин Е.Е., Панченко Е.А., Демидов А.Е. и др.* Система автоматического описания керна в производственном процессе. Опыт применения // Пути реализации нефтегазового потенциала Западной Сибири: Материалы XXV научно-практической конференции, Ханты-Мансийск, 23–26 ноября 2021 года / Под редакцией Э.А. Вторушиной, Е.Е. Оксенойд, С.А. Алёшина, Н.Н. Захарченко, Е.В. Олейник, Т.Н. Печёрина. Ханты-Мансийск: Автономное учреждение Ханты-Мансийского автономного округа – Югры. Научно-аналитический центр рационального недропользования им. В.И. Шпильмана, 2022. С. 293–299.

5. Комплекс DHD. URL: https://magazine.neftegaz.ru/articles/tsifrovizatsiya/682038-tsifrovoy-analiz-

kerna-v-zadachakh-proektirovaniya-razrabotki-neftyanykh-i-gazovykh-mestorozhdeniy-/ (дата обращения: 11.12.2025)

6. Программный комплекс «Цифровой керн».
URL: https://globalcio.ru/projects/10448/ (дата обращения: 11.12.2025)

7. *Аристов А.И., Зеленин А.В., Катанов Ю.Е.* Нейросетевое распознавание текстурных особенностей графических керновых данных. Свидетельство о регистрации программы для ЭВМ RU 2024615647, 11.03.2024. Заявка № 2024614650 от 11.03.2024.

8. *Li H, Wan B, Chu D, Wang R, Ma G, Fu J, Xiao Z.* Progressive Geological Modeling and Uncertainty Analysis Using Machine Learning // ISPRS International Journal of Geo-Information. 2023. Vol. 12 (3). 97.
https://doi.org/10.3390/ijgi12030097

9. *Химуля В.В.* Применение технологии цифрового анализа керна для изучения фильтрационно-емкостных свойств и структуры высокопроницаемых пород подземных хранилищ газа // RJES. 2024. №5. С. 1–15.
URL: https://rjes.ru/temp/fddc89c0f81314f3d14bad3446565446.pdf (дата обращения: 11.12.2025).

10. *Fuentes I., Padarian J., Iwanaga T.,* Vervoort R.W. 3D lithological mapping of borehole descriptions using word embeddings // Computers & Geosciences. 2020. Vol. 141. 104516. https://doi.org/10.1016/j.cageo.2020.104516.
URL: https://www.sciencedirect.com/science/article/pii/S0098300419306533

11. *Padarian J., Fuentes I.* Word embeddings for application in geosciences: development, evaluation, and examples of soil-related concepts // SOIL. 2019. Vol. 5. P. 177–187. https://doi.org/10.5194/soil-5-177-2019.
URL: https://soil.copernicus.org/articles/5/177/2019/

12. *Pennington J., Socher R., Manning C.* Glove: Global vectors for word representation // Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014. P. 1532–1543.

13. *Катанов Ю.Е., Аристов А.И., Ягафаров А.К., Новрузов О.Д.* Цифровой керн: нейросетевое распознавание текстовой геолого-геофизической информации // Известия высших учебных заведений. Нефть и газ. 2023. № 3 (159). С. 35–54.

14. *Денисов Д.В.* Анализ методов машинного обучения для тематической классификации текстов// Международный журнал информационных технологий и энергоэффективности. 2024. Т. 9, № 4 (42). С. 5–11.

15. *Козицын А.С.* Алгоритмы тематического поиска данных в наукометрических системах // Программная инженерия. 2022. Т. 13. № 6. С. 291–300.

METHOD FOR AUTOMATIC CLASSIFICATION OF FULL-TEXT DESCRIPTIONS OF CORES USING DICTIONARIES

A. P. Antonov[1] [0009-0007-3642-7734], S. A. Afonin[2] [0000-0003-3058-9269],
A. S. Kozitsin[3] [0000-0002-8065-9061], V. M. Staroverov[4] [0000-0001-8289-2273]

[1, 4]*Lomonosov Moscow State University, Moscow, Russia*

[2, 3]*Institute of Mechanics, Lomonosov Moscow State University, Moscow, Russia*

[1]alexey.p.antonov@gmail.com, [2]serg@msu.ru, [3]alexanderkz@mail.ru, [4]staroverovvl@yandex.ru

Abstract

The use of automatic text processing methods, including full-text description classification methods, allows achieving a significant reduction in labor costs when processing experimental data. This paper discusses the use of the automatic text classification method in the field of processing and classifying core elements and determining lithofacies. Lithofacies are coeval geological bodies (deposits) that differ in composition or structure from adjacent layers. When assessing the oil and gas potential of fields, it is necessary to construct maps and diagrams of lithofacies distribution. This requires classifying a large number of full-text descriptions of core sections prepared by specialists. The algorithm presented in the article allows, based on specified rules and dictionaries, to conduct classification taking into account the order and significance of keywords in sentences. The advantages of this approach are: the ability to distinguish between close lithofacies, the ability to use archival data, ease of adjustment to new classes, adaptation to Russian-language core descriptions and the possibility of local use without the need to transfer core descriptions to third-party applications.

Keywords*: text classification, lithofacies, dictionaries, information systems.*

REFERENCES

1. Iskusstvennyi intellekt v neftegazovoi industrii Kitaia. URL: https://nntc.pro/tpost/h2hoet4se1-iskusstvennii-intellekt-v-neftegazovoi-i

2. *Antonov A.P., Afonin S.A., Kozitsyn A.S. i dr.* Avtomatizirovannoe postroenie realistichnykh litofatsialnykh kart metodami kombinatornoi optimizatsii // Intellektualnye sistemy. Teoriia i prilozheniia. 2024. Vol. 28, № 4. S. 5–20.

3. Informatsionnaia sistema ABAI. URL: https://kmge.kz/abai/

4. *Baraboshkin E.E., Panchenko E.A., Demidov A.E. i dr.* Sistema avtomatich-eskogo opisaniia kerna v proizvodstvennom protsesse. Opyt primeneniia // Puti real-izatsii neftegazovogo potentsiala Zapadnoi Sibiri: Materialy XXV nauchno-prakticheskoi konferentsii, Khanty-Mansiisk, 23–26 noiabria 2021 goda / Pod redaktsiei E.A. Vtorushinoi, E.E. Oksenoid, S.A. Aleshina, N.N. Zakharchenko, E.V. Oleinik, T.N. Pecher-ina. Khanty-Mansiisk: Avtonomnoe uchrezhdenie Khanty-Mansiiskogo avtonomnogo okruga – Iugry "Nauchno-analiticheskii tsentr ratsionalnogo nedropolzovaniia im.V.I.Shpilmana", 2022. S. 293–299.

5. Kompleks DHD. URL: https://magazine.neftegaz.ru/articles/tsifrovizatsiya/682038-tsifrovoy-analiz-kerna-v-zadachakh-proektirovaniya-razrabotki-neftyanykh-i-gazovykh-mestorozh-deniy-/ (11.12.2025)

6. Programmnyi kompleks "Tsifrovoi kern". URL: https://globalcio.ru/projects/10448/

7. *Aristov A.I., Zelenin A.V., Katanov Iu.E.* Neirosetevoe raspoznavanie teksturnykh osobennostei graficheskikh kernovykh dannykh. Svidetelstvo o registratsii programmy dlia EVM RU 2024615647, 11.03.2024. Zaiavka № 2024614650 11.03.2024.

8. *Li H, Wan B, Chu D, Wang R, Ma G, Fu J, Xiao Z.* Progressive Geological Modeling and Uncertainty Analysis Using Machine Learning // ISPRS International Jour-nal of Geo-Information. 2023. Vol. 12(3). 97. https://doi.org/10.3390/ijgi12030097

9. *Khimulia V.V.* Primenenie tekhnologii tsifrovogo analiza kerna dlia izuche-niia filtratsionno-emkostnykh svoistv i struktury vysokopronitsaemykh porod pod-zemnykh khranilishch gaza // RJES. 2024. №5. S. 1–15. URL: https://rjes.ru/temp/fddc89c0f81314f3d14bad3446565446.pdf

10. *Fuentes I., Padarian J., Iwanaga T., Vervoort R.W.,* 3D Lithological mapping of borehole descriptions using word embeddings // Computers & Geosciences. 2020. Vol. 141. 104516. https://doi.org/10.1016/j.cageo.2020.104516
URL: https://www.sciencedirect.com/science/article/pii/S0098300419306533

11. *Padarian J., Fuentes I.* Word embeddings for application in geosciences: development, evaluation, and examples of soil-related concepts // SOIL. 2019. Vol. 5. P. 177–187. https://doi.org/10.5194/soil-5-177-2019, 2019.
URL: https://soil.copernicus.org/articles/5/177/2019/

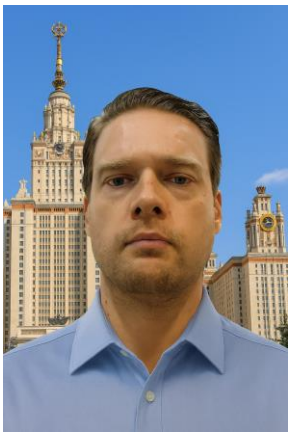
12. *Pennington J., Socher R., Manning C.* Glove: Global vectors for word representation // Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014. P. 1532–1543.

13. *Katanov Iu.E., Aristov A.I., Iagafarov A.K., Novruzov O.D.* Tsifrovoi kern: neirosetevoe raspoznavanie tekstovoi geologo-geofizicheskoi informatsii // Izvestiia vysshikh uchebnykh zavedenii. Neft i gaz. 2023. № 3 (159). S. 35–54.

14. *Denisov D.V.* Analiz metodov mashinnogo obucheniia dlia tematicheskoi klassifikatsii tekstov // Mezhdunarodnyi zhurnal informatsionnykh tekhnologii i energoeffektivnosti. 2024. Vol. 9, № 4(42). S. 5–11.

15. *Kozitsyn A.S.* Algoritmy tematicheskogo poiska dannykh v naukometricheskikh sistemakh // Programmnaia inzheneriia. 2022. Vol. 13, № 6. S. 291–300.

СВЕДЕНИЯ ОБ АВТОРАХ



АНТОНОВ Алексей Петрович – доцент, к. ф.-м. н., окончил мехмат МГУ им. М.В. Ломоносова. Специалист в области гармонического анализа.

Alexey Petrovich ANTONOV – Associate Professor, Ph.D., graduated from Lomonosov Moscow State University. Specialist in the field of harmonic analysis.

email: alexey.p.antonov@gmail.com

ORCID:0009-0007-3642-7734

АФОНИН Сергей Александрович – ведущий научный сотрудник, к. ф.-м. н., окончил мехмат МГУ им. М. В. Ломоносова. Специалист в области регулярных языков и информационных систем.

Sergey Alexandrovich AFONIN – Leading Researcher, Ph.D., graduated from Lomonosov Moscow State University. Specialist in the field of regular languages and information systems.

email: serg@msu.ru

ORCID:0000-0003-3058-9269

КОЗИЦЫН Александр Сергеевич – ведущий научный сотрудник, к. ф.-м. н., окончил мехмат МГУ им. М. В. Ломоносова. Специалист в области информационного поиска и баз данных.

Alexander Sergeevich KOZITSYN – Leading Researcher, Ph.D., graduated from Lomonosov Moscow State University. Specialist in the field of information retrieval and database.

email: alexanderkz@mail.ru

ORCID: 0000-0002-8065-9061

СТАРОВЕРОВ Владимир Михайлович – доцент, к. ф.-м. н., окончил мехмат МГУ им. М. В. Ломоносова. Специалист в области геологического моделирования.

Vladimir Mikhailovich STAROVEROV – Associate Professor, Ph.D., graduated from Lomonosov Moscow State University. Specialist in the field of geological modelling.

email: staroverovvl @yandex.ru

ORCID: 0000-0001-8289-2273

Материал поступил в редакцию 18 декабря 2025 года