ЦИФРОВЫЕ РЕШЕНИЯ В НАУЧНО-ОБРАЗОВАТЕЛЬНОЙ СФЕРЕ

Научная статья
УДК 371.26:004.912
DOI: 10.17853/2587-6910-2024-12-9-15

ПРИМЕНЕНИЕ TEXT MINING В ТЕСТИРОВАНИИ

Владимир Владимирович Братищенко
Доцент кафедры математических методов и информационных технологий Кандидат физико-математических наук, e-mail: vvb@bgu.ru
ФГБОУ ВО «Байкальский государственный университет», Россия, Иркутск

Аннотация. В работе предлагается использовать методы обработки текстов на естественном языке для оценивания ответов на тестовые задания открытого типа. Для тестирования нужен корпус заданий с вариантами правильных ответов. Для оценки соответствия ответа тестируемого образцу предлагается использовать два метода. Первый метод использует параметризацию текста в виде «мешка слов» и обученный по корпусу заданий классификатор, который оценивает соответствие ответа образцу как вероятность соответствующей классификации. Второй метод использует эмбеддинговую модель корпуса текстов по теме тестирования. Оценка соответствия ответа образцу вычисляется как косинус угла вектора образца и вектора ответа. Численные эксперименты продемонстрировали возможности применения предлагаемых оценок.

Ключевые слова: тестовые задания открытого типа; обработка текстов на естественном языке; классификация текстов; «мешок слов»; эмбеддинги предложений

Для цитирования: Братищенко В. В. Применение text mining в тестировании // Новые информационные технологии в образовании и науке. 2024. № 12. С. 9–15 https://doi.org/10.17853/2587-6910-2024-12-9-15

APPLYING TEXT MINING IN TESTING

Vladimir Vladimirovich Bratischenko

Associate Professor of the Department of Mathematical Methods and Information Technologies, Candidate of Physical and Mathematical Sciences

Baikal State University, Russia, Irkutsk

Abstract. The paper proposes to use natural language text processing methods to estimate answers to open-type test tasks. For testing, you need a corpus of tasks with correct answers. To assess the compliance of the test person's response with the sample, it is proposed to use two methods. The first method uses text parameterization in the form of a “bag of words” and a classifier trained on a corpus of tasks, which evaluates the correspondence of the answer to the sample as the probability of the corresponding classification. The second method uses the embedding model of a corpus of texts on the topic of testing. The assessment of the response's correspondence to the sample is calculated as the cosine of the angle of the sample vector and the response vector. Numerical experiments demonstrated the possibilities of using the proposed estimates.

Keywords: open test tasks; natural language text processing; text classification; "bag of words"; sentence embedding

For citation: Bratischenko V. V. Applying text mining in testing. *New information technoljgies in tducftion and science.* 2024;12: 9-15. (In Russ.). doi:10.17853/2587-6910-2024-12-9-15

Тесты традиционно относят к инструментам объективного измерений уровня знаний [1]. Распространению технологии тестирования способствует активное применение компьютерных технологий. В данной работе описывается применение компьютерных технологий обработки текстов на естественном языке (text mining) для обработки ответов на открытые вопросы текстов.

К недостаткам тестирования относят [2] достаточно большую вероятность угадывания правильных ответов на задания закрытого типа с фиксированным набором

вариантов ответов. Для заданий открытого типа, в которых необходимо ввести правильный ответ, такая вероятность значительно меньше. Все известные типы компьютерных тестовых заданий характеризуются «жесткостью» — достаточно небольшого отклонения от правильного ответа, чтобы результат выполнения был отрицательным.

Предлагаемая в работе [3] методика «мягкого» тестирования требует более детальной проработки процедуры измерения точности ответа. Для этого требуется «Ввести в тестовые задания многозначные логические отношения, создать критериально-ориентированную технику оценки выполнения заданий, включающую не только полюсные («верно» и «не верно») варианты оценки, но и более широкий спектр, в том числе двумерную, матричную шкалу». Предложенная методика, очевидно, требует трудоемкой методической подготовки, а «критериально-ориентированная техника оценки» будет отражать пристрастия автора.

Как альтернативу такой методике, предлагается использовать задания открытого типа, предполагающие ответ в виде предложения на естественном языке. Для оценивания ответов на такие задания можно использовать в достижения в области Text Mining [4]. В качестве оценки предлагается использовать степень близости ответа на задание одному или нескольким образцам правильного ответа. В этом случае тестовые задания могли бы применяться для проверки знания определений или инструкций. Оценка за такое задание может определяться мерой соответствия ответа тестируемого одному из образцов, причем максимальная оценка — 1 балл — должна соответствовать точному ответу.

Один из самых простых методов параметризации текстов [5] — «мешок слов» — характеризует текст вектором частот термов, входящих в текст. Для русского языка требуется предобработка, которая заключается в замене слов словарными формами — леммами. Кроме этого, рассматриваются комбинации слов — n-граммы, исключаются предлоги, междометия, вводные слова и другие аналогичные случаи употребления слов, объединяемые в так называемый стоп-список. Альтернативой стоп-списку является включение в параметрическое описание только терминов из специального словаря.

Полученные параметрические описания текстов из достаточно большого корпуса заданий открытого типа становятся исходными данными для настройки модели

классификации. Корпус заданий должен быть связан с контекстом тестирования. Например, он может включать все определения в рамках некоторой области знания, а также задания, предполагающие ответы в виде предложений. Для классификации текстов применяются различные модели [5] от простых, таких как наивная байесовская модель и логистическая регрессия, до достаточно сложных, например, нейронных сетей и случайного леса. Обучение модели выполняется по всему корпусу заданий: для каждого задания должны быть определены параметрические описания правильных ответов. Такая разметка не требует особых трудозатрат. Разработчику нужно составить множество пар «задание — правильный ответ», причем правильных ответов может быть несколько. Параметризация корпуса заданий, обучение моделей и определение класса ответа тестируемого может быть выполнено с применением свободно распространяемых модулей, таких как scikit-learn (https://scikit-learn.org/stable/). Обученная модель классификации для ответа на задание выдает вероятность его отнесения к правильному ответу на задание.

В таблице 1 приведены результаты предсказания классов для некоторых вариантов ответов. Первые два ответа соответствуют разным определениям базы данных, использованным для обучения классификатора, им соответствует высокая вероятность распознавания. Третий ответ содержит совсем неточное определение, которое классификатор с небольшой вероятностью отнес к СУБД. Четвертое утверждение соответствует определению системы управления базами данных. Приведенный пример демонстрирует возможность применения предложенной технологии оценки тестовых заданий открытого типа.

Хорошие перспективы для определения близости текстов имеет применение эмбеддинговых моделей [6], которые создают векторные описания текстов с учетом контекстов употребления слов. Основным механизмом word embeddings является модель word2vec [7]. Данная модель обучает нейронную сеть по большому корпусу текстов определять числовой вектор слова на основе вариантов его использования: близкие по контексту слова должны иметь близкие векторы. Обучение нейронной сети выполняется по огромному корпусу текстов.

Для определения близости ответов к образцам больше подходит аналогичная модель — doc2vec [8], позволяющая получать вектор для всего сообщение. Мерой

близости двух сообщений может быть косинус угла между соответствующими векторами.

Таблица 1.

Предсказание класса по тексту ответа

№	Леммы определения	Предсказанный класс	Вероятность класса «База данных»	Вероятность класса «СУБД»
1	поименованный целостный единый система данные организовывать по определенный правило который предусматривать общий принцип описание хранение и обработка данные	База данных	0,916	0,028
2	совокупность данные организованный в соответствие с некоторый концептуальный модель данный который описывать характеристика этот данные и взаимоотношение между соответствующий они реалия и который предназначать для информационный обеспечение один или несколько приложение	База данных	0,958	0,014
3	набор информация который храниться упорядоченный в электронный вид	Система управления базами данных	0,092	0,414
4	совокупность программный и лингвистический средство общий или специальный назначение обеспечивать управление создание и использование база данных	Система управления базами данных	0,008	0,952

Для апробации такой технологии использовался курс лекций по дисциплине «Базы данных». Текст был разбит на 3625 предложений, обработан лемматизатором Mystem (https://yandex.ru/dev/mystem/) и использован для настройки модели Doc2Vec модуля gensim (https://radimrehurek.com/gensim/models/doc2vec.html). Для двух определений базы данных и двух определений системы управления базами данных с помощью настроенной модели были определены соответствующие векторы и косинусы углов для всех возможных пар векторов (см. табл. 2). Близкие определения характеризуются значением косинуса близким к единице.

Приведенные примеры свидетельствуют о возможности применения методов обработки текстов на естественном языке для создания и применения тестовых заданий открытого типа. Предложенные меры близости могут быть использованы для формирования оценки в случае незначительных отклонений от образцового ответа. Для надежности можно ввести границу меры близости. При значениях ниже границы тестируемый получает нулевую оценку за задание.

Таблица 2.

Меры близости (косинусы углов) для векторов четырех предложений

i	Леммы i-го определения	Косинус угла между i-м и j-м векторами			
		1	2	3	4
1	поименованный целостный единый система данные организовывать по определенный правило который предусматривать общий принцип описание хранение и обработка данные	1	0,88	-0,09	0,16
2	поименованный система данные организовывать по определенный правило который предусматривать общий принцип хранение и обработка данные	0,88	1	0,15	0,27
3	комплекс программа обеспечивать хранение добавление удаление и выбор данные	-0,09	0,15	1	0,71
4	программный обеспечение для создание бд хранение добавление удаление и выбор необходимый информация	0,16	0,27	0,71	1

Технологии Text Mining могут быть использованы для создания чат-ботов выполняющих различные функции обучения: тестирование, предоставление дополнительных учебных материалов и заданий по результатам тестирования, консультирование обучаемых. Все это поможет сделать процесс обучения более гибким, интерактивным и привлекательным.

Список источников

1. Челышкова М. Б. Теория и практика конструирования педагогических тестов. М.: Логос, 2002. 431 с.

2. Желнин М. Э., Кудинов В. А., Белоус Е. С. Преимущества и недостатки тестирования в сравнении с другими методами контроля знаний // Ученые записки. Электронный научный журнал Курского государственного университета. 2012. № 1 (21). С. 244–248. URL: https://cyberleninka.ru/article/n/preimuschestva-i-nedostatki-

testirovaniya-v-sravnenii-s-drugimi-metodami-kontrolya-znaniy (дата обращения: 06.02.2024).

3. Морев И. А. Образовательные информационные технологии. Ч. 2. Педагогические измерения. Владивосток: Изд-во Дальневост. ун-та, 2004. 174 с.

4. Автоматическая обработка текстов на естественном языке и анализ данных / Е. И. Большакова, К. В. Воронцов, Н. Э. Ефремова, Э. С. Клышинский, Н. В. Лукашевич, А. С. Сапин. М.: Изд-во НИУ ВШЭ, 2017. 269 с.

5. Силен Д., Мейсман А., Али М. Основы Data Science и Big Data. Python и наука о данных. СПб.: Питер, 2017. 336 с.

6. Спивак А. И., Лапшин С. В., Лебедев И. С. Классификация коротких сообщений с использованием векторизации на основе ELMo // Известия ТулГУ. Технические науки. 2019. № 10. С. 410–418. URL: https://cyberleninka.ru/article/n/klassifikatsiya-korotkih-soobscheniy-s-ispolzovaniem-vektorizatsii-na-osnove-elmo (дата обращения: 10.11.2023).

7. Efficient Estimation of Word Representations in Vector Space / T. Mikolov, K. Chen, G. Corredo, J. Dean // arXiv:1301.3781. 2013. https://doi.org/10.48550/arXiv.1301.3781.

8. Le Q., Mikolov T. Distributed representations of sentences and documents // arXiv:1405.4053. 2014. https://doi.org/10.48550/arXiv.1405.4053.