5. Математическое и информационное моделирование : материалы Всероссийской конференции молодых ученых. — Тюмень : Издательство Тюменского государственного университета, 2018. — 496 с. — Текст : непосредственный.
6. Математическое и информационное моделирование : материалы Всероссийской конференции молодых ученых. — Тюмень : Издательство Тюменского государственного университета, 2017. — 534 с. — Текст : непосредственный.
7. Математика и междисциплинарные исследования : материалы Всероссийской научно-практической конференции молодых ученых с международным участием. — Пермь : Пермский государственный национальный исследовательский университет, 2020. — 343 с. — Текст : непосредственный.
8. Krapivin M. Large dataset for keyphrases extraction : Technical report / M. Krapivin, A. Autaeu, M. Marchese ; University of Trento. — Povo-Trento, 2008. — 6 p. — Text : direct.
9. Lin C. Y. Rouge: A package for automatic evaluation of summaries / C. Y. Lin. — Text : direct // Text summarization branches out. — Barcelona, 2004. — P. 74-81.

А. А. Кузьминых, А. А. Ступников
Тюменский государственный университет, г. Тюмень

УДК 004.912

РАЗРАБОТКА И АНАЛИЗ МОДЕЛЕЙ КЛАСТЕРИЗАЦИИ И КЛАССИФИКАЦИИ ДЛЯ ДАННЫХ ОБ УЧЕБНЫХ ДИСЦИПЛИНАХ ИМиКН

Аннотация. *В статье рассматривается подход к анализу соответствия содержания дисциплин, представленного в учебно-методических комплексах (УМК), определенному разделу компьютерных наук.*

Ключевые слова: *кластеризация, классификация, машинное обучение, анализ текстовых данных.*

Введение. Образовательные организации должны поддерживать высокий уровень качества профессиональной подготовки будущих квалифицированных специалистов. Одним из основных факторов образовательной деятельности является содержательное наполне-

ние дисциплин, входящих в учебный план по отдельным специальностям. Основной способ заочно оценить содержание любой дисциплины — провести анализ ее учебно-методического комплекса (УМК). В данном исследовании рассматривается подход к анализу тематического содержания дисциплин, связанный с соответствием преподаваемых учебных дисциплин с определенными разделами компьютерных наук.

Подобный анализ предполагает выполнение классификации текстов УМК по разделам знаний. Схожие задачи возникают, в частности, при классификации статей по соответствующим предметным областям. При этом применяются нейронные сети различной архитектуры. Довольно популярной является модель, основанная на полносвязных слоях и LSTM, использование LSTM обусловлено тем, что она имеет преимущество перед стандартной рекуррентной нейронной сетью [2, 6]. Для классификации статей по 109 заранее выделенным предметным областям (по рефератам WoS) применялась глубокая нейросеть архитектуры DANN [3]. Особенностью нашего исследования является автоматическое выделение предметных областей (тематик) для УМК.

Проблема исследования. Так как при формировании УМК составители должны опираться на то, как дисциплина будет выглядеть в общем комплексе изучаемых профессиональных дисциплин было выдвинуто следующее предположение: для классификации содержания УМК дисциплин, преподаваемых в Институте математики и компьютерных наук, можно построить классификатор, определяющий отношение данной дисциплины к тому или иному разделу компьютерных наук с точностью не менее 0.7 (отношения количество правильно классифицированных документов к размеру выборки) на тестовой выборке.

Задача исследования: конструирование и реализация механизма классификации содержания УМК по дисциплинам направлений Института математики и компьютерных наук, позволяющего отнести дисциплину УМК к конкретному разделу знаний и на его основе оценка соответствия содержания УМК дисциплины ее тематике.

Материалы и методы. Исходными данными для проведения исследования являются данные, содержащие тексты УМК, полученные из базы данных гранта о цифровом следе студента, реализуемого Институтом математики и компьютерных наук ТюмГУ. Данные извлекались из двух баз данных PostgreSQL и Hbase.

В комплексе применяемых методов и моделей можно выделить три структурных блока: предобработка данных (A), выделение тематических групп дисциплин (B) и построение и применение классификатора (C). Рассмотрим эти блоки ниже.

A. Этапы предобработки текстовых данных

1. Исключение повторяющихся УМК и УМК общей направленности.

2. Извлечение тематических глав УМК.

Под тематическими главами понимаются следующие главы УМК: «Перечень планируемых результатов обучения по дисциплине», «Содержание дисциплины», «Темы лабораторных работ». Тематические главы извлекаются с помощью регулярных выражений.

3. Токенизация.

Под словами понимаются буквенные последовательности, как русского, так и английского языка без знаков препинания и чисел. Для токенизации использовалась функция “word_tokenize” из библиотеки NLTK для русского языка, слова определяются с помощью регулярного выражения. Далее из списка токенов удаляются стоп-слова, содержащие предлоги и союзы.

4. Лемматизация.

В функцию лемматизации библиотеки PyMystem3 передается строка из токенов, разделенных пробелом и возвращается список лемм. Далее, если имеется список биграмм (из п. 6), производится замена двух идущих подряд лемм на одну общую лемму, если они составляют биграмму.

5. Удаление стоп-слов.

Список слов для удаления был взят из библиотеки NLTK, а также дополнен списком слов, вручную составленным на основе тематики исследования («тема», «балл», «студент»).

6. Замена токенов на биграммы.

Список биграмм был составлен с помощью модели Phrases (библиотека Gensim), данная модель позволяет автоматически определять распространенные фразы — выражения. Полученные биграммы были сохранены в файл.

7. Векторизация полученных лемм.

Векторизация происходила по трем моделям: TF-IDF из библиотеки Sklearn, Word2Vec (min_count = 5, vector_size = 100, window = 3, epochs = 50, sg = 0) и Doc2Vec (min_count = 2, vector_size = 64, window = 3, epochs = 25), взятые из библиотеки Gensim. В итоге было получено 3 датасета, каждый из них имел 326 строк и 843 столбца для модели TF-IDF, 100 столбцов для модели Word2Vec и 64 столбца для модели Doc2Vec соответственно.

В. Получение тематических групп с помощью моделей кластеризации

Кластеризация производилась по трем моделям — Kmeans, SpectralClustering (gamma=0.3, affinity='rbf'), Agglomerative Clustering (affinity='cosine', linkage='complete') для каждого из трех датасетов. В качестве метрики расстояния для Kmeans была выбрана метрика косинусного расстояния между векторами на основе статьи [5]. Оптимальное количество кластеров определялось с помощью метрик «Силуэт» и индекса Дэвиса–Болдина. На рис. 1 представлены графики с оценками эффективности модели Kmeans с косинусовой метрикой для датасета, сформированном с помощью модели Doc2Vec.

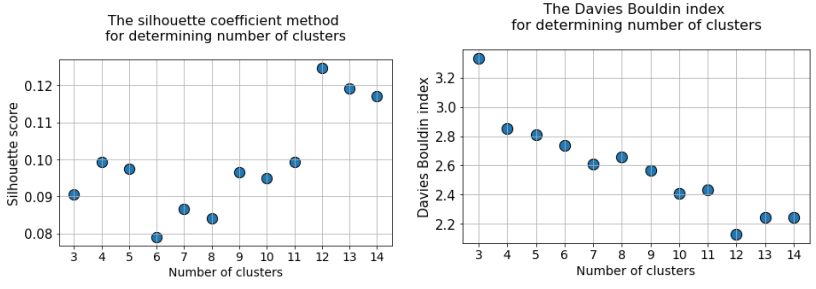

Рис. 1. Оценки эффективности модели кластеризации при различных значениях количества кластеров

Наибольшее значение по метрике «Силуэт» было получено на 12 кластерах, в свою очередь минимальное значение индекса Дэвиса–Болдина соответствует 12 кластерам. Данная модель показала наилучший результат по кластеризации: большинство текстов разбились на группы, объединенные по тематикам, была выделена одна группа, целиком состоящая из предметов по «Операционным системам».

По результатам лучшей модели были получены следующие тематические группы (в скобках указано число дисциплин, образующих данную группу):

1. Общая математика (41): математический анализ, банаховы алгебры и гармонический анализ, действительный анализ и т. п.

2. Логика/алгебра/теория вероятностей (30): алгебра, непрерывные группы, алгебра и математическая логика, теория игр и т. п.

3. Базы данных (19): системы управления базами данных, администрирование и безопасность MS SQL Server и т. п.

4. Информационная безопасность (36): основы информационной безопасности, история криптографии и т. п.

5. Операционные системы (14): операционные системы, администрирование операционных систем и т. п.

6. Фундаментальное программирование (35): компьютерная графика, технологии разработки программного обеспечения т. п.

7. Веб и мобильные приложения (19): Технологии web-программирования, интернет-технологии, разработка мобильных приложений и т. п.

8. Сети и администрирование (12): администрирование информационных систем, корпоративные информационные системы и т. п.

9. Механика / мехатроника (29): механика деформируемого твердого тела, вычислительные методы математической физики и т. п.

10. Информационные системы (63): современные информационные системы, информационные технологии, управление проектами и т. п.

11. Алгоритмы (28): структуры и алгоритмы компьютерной обработки данных, дискретная математика и т. п.

C. Архитектура моделей классификации

Для построения моделей классификации используется высокоуровневый API для разработки моделей — Keras, который позволяет инкапсулировать общие парадигмы машинного обучения в код [4, p. 8]. На основе данной были построены две модели классификации, представленные на рис. 2. Архитектура моделей подбиралась экспериментально: варьировалось число слоев, количество узлов, функции активации, для Droupout доля отбрасываемых входных единиц, были отобраны модели, показавшие самую высокую точность. В качестве функции потерь в моделях сетей использовалась “categorical_crossentropy”, данная функция потерь используется для мультиклассовой классификации.

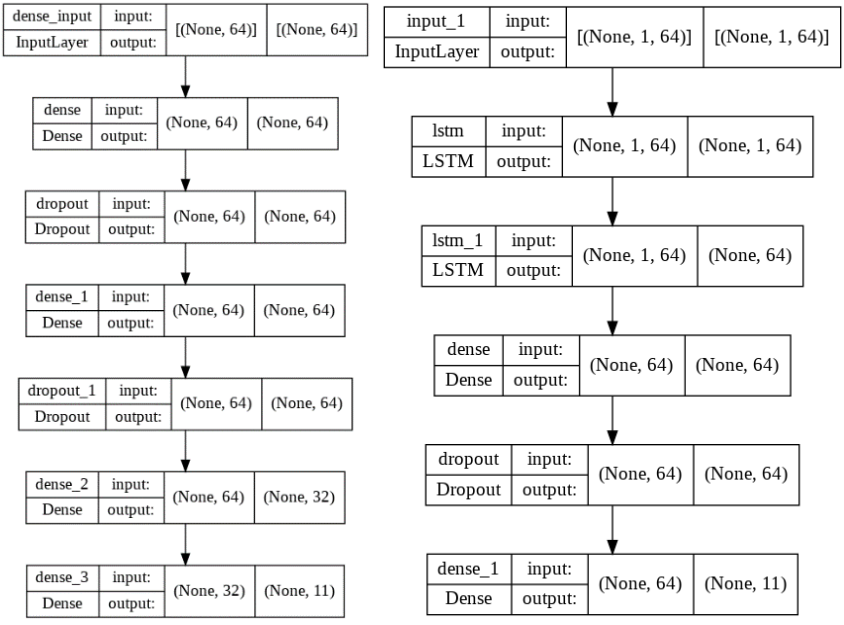

Рис. 2. Архитектура сети “Simple model” и “Model with LSTM”

Результаты. Вычислительный эксперимент происходил на машине на базе x64 с 11th Gen Intel(R) Core(TM) i7-11800H, 2.30GHz, 2304 Мгц, 8 ядер, 16 логических процессоров, RAM 32 Гб, сравнение

моделей происходило на тестовых данных по времени работы, средней точности и метрики ROC-AUC.

Полученные значения, представленные в табл. 1, показывают, что наилучшие результаты обучения были получены у сети “Simple model” при 100 эпохах по средней точности и метрике ROC-AUC [1], разница в средней точности у 50 и 100 эпохах незначительная. Обе модели показали хорошие результаты по средней точности, начиная с 50 эпох. По времени обучения у обоих моделей разница оказалась незначительной.

Таблица 1

Результаты классификации на тестовой выборке

Модель	*Кол-во эпох*	*Средняя точность*	*ROC-AUC*	*Время работы*
Simple model	10	0.70	0.96	4.24
	50	0.87	0.989	4.14
	100	0.88	0.995	6.46
	120	0.85	0.991	7.78
Model with LSTM	10	0.42	0.93	5.23
	50	0.82	0.987	6.18
	100	0.85	0.9876	7.96
	120	0.84	0.99	8.97

Построенный классификатор позволил абсолютно корректно определить классы 89% УМК. При этом наиболее нестабильная классификация имела место для дисциплин класса «информационные системы»: из 21 дисциплины 2 дисциплины попали в класс «веб и мобильные приложения», другие 2 дисциплины попали в класс «информационная безопасность» и 1 в класс «механика/мехатроника», и аналогично 1 дисциплина из класса «веб и мобильные приложения» и 2 из класса «механика/мехатроника» были отнесены к данной группе, вероятно, это связно с тем, что класс «информационные системы» имеет достаточно обширные тематики из различных разделов компьютерных наук и могут быть близки тематически к данным классам. Также имеются ошибки классификации для

класса «общая математика»: из 21 дисциплины 2 определились в группу «логика/алгебра/теория вероятностей», другие 2 определились в группу «механика/мехатроника», так как это довольно близкие тематики, содержащие в себе математическую базу. Ошибка в классификации между классами «сети и администрирование» и «операционные системы» вероятно также вызвана близостью тематик. В целом модель показала хороший результат и подтверждает выдвинутое предположение.

Заключение. В рамках исследования, был проведен сбор программ дисциплин в рамках образовательной программы, в общей сложности было получено 324 файла. Каждый файл проходил предобработку данных: токенизация, удаление стоп-слов, лемматизация, повторное удаление стоп-слов, векторизация по моделям TF-IDF, Word2Vec, Doc2Vec, в итоге было получено 3 датасета, по которым была произведена кластеризация.

Кластеризация проводилась по 3 алгоритмам: Kmeans, Spectral Clustering, Agglomerative Clustering. Наилучшие результаты по кластеризации были достигнуты алгоритмом Kmeans с косинусной метрикой расстояния на основе векторизации данных по модели Doc2Vec. Далее, были реализованы 2 модели классификации на основе нейронный сетей с использованием библиотеки Keras для TensorFlow. По результатам было получено, что модель под названием “Simple model” является наилучшей для классификации учебных дисциплин ИМиКН. Точность на тестовой выборке составила 0.88, что выше предполагаемой точности. Содержание УМК по дисциплинам профильной тематики ИМиКН позволяет организовать достаточно надежное автоматическое средство отнесения их к соответствующим разделам. Кроме выявление общей содержательной картины цикла дисциплин отдельного направления разработанный инструмент позволяет оценить соответствие содержания УМК отдельной дисциплины ее предметной области.

СПИСОК ЛИТЕРАТУРЫ

1. Davis J. The relationship between Precision-Recall and ROC curves / J. Davis, M. Goadrich. — Text : direct // Proceedings of the 23rd international conference on Machine learning. — 2006. — P. 233-240.

2. Hochreiter S. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies / S. Hochreiter, Y. Bengio, P. Frasconi, J. Schmidhuber. — Text : electronic // A field guide to dynamical recurrent neural networks. — Wiley-IEEE Press, 2001. — P. 237-243. — URL: http://www.bioinf.jku.at/publications/older/ch7.pdf (date of the application: 23.05.2022).
3. Kandimalla B. Large scale subject category classification of scholarly papers with Deep Attentive Neural Networks / B. Kandimalla, S. Rohatgi, J. Wu. — Text : electronic // Frontiers in Research Metrics and Analytics. — 2021. — Vol. 5 — P. 1-12. — URL: https://www.frontiersin.org/articles/ 10.3389/frma.2020.600382/pdf (date of the application: 23.05.2022).
4. Moroney L. AI and Machine Learning for Coders / L. Moroney. — Text : direct. — USA, Highway North, Sebastopol : O'Reilly Media, 2020. — 390 p.
5. Sahi L. An improved K-means algorithm using modified cosine distance measure for document clustering using Mahout with Hadoop / L. Sahi, B. Mohan. — Text : direct // 9th International Conference on Industrial and Information Systems (ICIIS). — 2014. — P. 1-5.
6. Semberecki P. Deep learning methods for subject text classification of articles / P. Semberecki, H. Maciejewski. — Text : direct // Federated Conference on Computer Science and Information Systems (FedCSIS). — 2017. — Vol. 11. — P. 357-360.