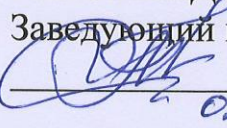


МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение
высшего образования
«ТЮМЕНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

ИНСТИТУТ МАТЕМАТИКИ И КОМПЬЮТЕРНЫХ НАУК
Кафедра программного обеспечения

РЕКОМЕНДОВАНО К ЗАЩИТЕ В ГЭК
Заведующий кафедрой, к.т.н, доцент

М. С. Воробьева
02.07. 2021 г.

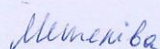
ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
магистерская диссертация

**РАЗРАБОТКА СИСТЕМЫ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ ПО
ИЗМЕНЕНИЮ СТРУКТУРЫ И СОДЕРЖАНИЯ ОБРАЗОВАТЕЛЬНОЙ
ПРОГРАММЫ**

02.04.03 Математическое обеспечение и администрирование информационных систем

Магистерская программа «Разработка технологий Интернета вещей и больших данных»

Выполнила работу
студентка 2 курса
очной формы обучения



Метелёва Елена Сергеевна

Научный руководитель
д.п.н., профессор



Захарова Ирина Гелиевна

Рецензент
руководитель ИТ-проектов
ООО "Инкомтехнологии
Групп"



Боганюк Юлия Викторовна

Тюмень
2021

Оглавление

ВВЕДЕНИЕ.....	4
ГЛАВА 1. ОБРАБОТКА ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ.....	6
1.1. ОСОБЕННОСТИ ОБРАБОТКИ ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ.....	6
1.2. ВЕКТОРНАЯ МОДЕЛЬ	7
1.3. МЕРЫ БЛИЗОСТИ ВЕКТОРОВ	8
1.4. КЛАСТЕРИЗАЦИЯ	9
1.5. ТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ	10
1.6. ВЕРОЯТНОСТНЫЕ ТЕМАТИЧЕСКИЕ МОДЕЛИ.....	11
1.6.1. ГРАФИЧЕСКИЕ МОДЕЛИ	12
1.6.2. ВЕРОЯТНОСТНОЕ ЛАТЕНТНО-СЕМАНТИЧЕСКОЕ ИНДЕКСИРОВАНИЕ.....	13
1.6.3. СКРЫТОЕ РАЗМЕЩЕНИЕ ДИРИХЛЕ	14
1.6.4. ИЕРАРХИЧЕСКОЕ СКРЫТОЕ РАЗМЕЩЕНИЕ ДИРИХЛЕ..	16
1.7. НЕПАРАМЕТРИЧЕСКИЕ МОДЕЛИ.....	18
1.8. МЕТОДЫ ОЦЕНИВАНИЯ КАЧЕСТВА РЕЗУЛЬТАТОВ ПОСТРОЕНИЯ ТЕМАТИЧЕСКОЙ МОДЕЛИ.....	19
1.8.1. ОБОБЩАЮЩАЯ СПОСОБНОСТЬ МОДЕЛИ.....	19
1.8.2. ИНТЕРПРЕТИРУЕМОСТЬ	20
ГЛАВА 2. ПОСТАНОВКА ЗАДАЧИ И ЕЕ РЕШЕНИЕ	21
2.1. ПОСТАНОВКА ЗАДАЧИ.....	21
2.2. ИСПОЛЬЗУЕМЫЕ ИНСТРУМЕНТЫ.....	21
2.3. ИСХОДНЫЕ ДАННЫЕ	21
2.4. ИЗВЛЕЧЕНИЕ ДАННЫХ.....	22

2.4. ПРЕДОБРАБОТКА ДАННЫХ	24
2.5. ВЫБОР ДАННЫХ	24
2.6. ПРОГРАММНАЯ РЕАЛИЗАЦИЯ	29
2.6.1. УНИВЕРСАЛЬНЫЕ МЕТОДЫ.....	29
2.6.2. КЛАСТЕРИЗАЦИЯ	32
2.6.3. ТЕМАТИЧЕСКИЕ МОДЕЛИ.....	33
2.6.4. АЛГОРИТМ ПОИСКА КУРСА	44
2.6.5. ОПРЕДЕЛЕНИЕ НЕОБХОДИМЫХ ЗНАНИЙ ДЛЯ ИЗУЧЕНИЯ КУРСА.....	48
2.6.6. АЛГОРИТМ ОПРЕДЕЛЕНИЯ НЕОБХОДИМЫХ ЗНАНИЙ ДЛЯ ИЗУЧЕНИЯ КУРСА	50
ЗАКЛЮЧЕНИЕ	54
СПИСОК ЛИТЕРАТУРЫ	55
ПРИЛОЖЕНИЕ 1. ПАРСИНГ РПД.....	58
ПРИЛОЖЕНИЕ 2. ОПРЕДЕЛЕНИЕ РАССТОЯНИЯ МЕЖДУ ЗАПРОСОМ ПОЛЬЗОВАТЕЛЯ И КОРПУСОМ ТЕКСТОВ	60
ПРИЛОЖЕНИЕ 3. ПОСТРОЕНИЕ ТЕМАТИЧЕСКОЙ МОДЕЛИ.....	61
ПРИЛОЖЕНИЕ 4. ПОИСК КУРСА ПО КЛЮЧЕВЫМ СЛОВАМ.....	62
ПРИЛОЖЕНИЕ 5. ПАРСИНГ ТАБЛИЦЫ ОБЕСПЕЧИВАЮЩИХ ДИСЦИПЛИН	63

ВВЕДЕНИЕ

В настоящее время в высшем образовании России сформировался определенный перечень образовательных услуг. Большая часть высших учебных заведений вынуждена конкурировать друг с другом из-за недостатка абитуриентов и дефицита рабочих мест на рынке труда [18]. Признаком высокой конкурентоспособности вузов является сохранение плановых показателей набора. Способность вузов конкурировать друг с другом зависит от наличия в учебных планах дисциплин, которые способствуют формированию компетенций студентов, наиболее востребованных у работодателей. Высокие темпы научно-технического прогресса способствовали тому, что преподавание многих актуальных дисциплин составляет менее пяти лет.

Динамичность рынка труда требует постоянной актуализации образовательных программ. В каждом вузе за каждой образовательной программой стоит руководитель, который вносит в нее изменения, сравнивает дисциплины, анализируя другие программы. В силу ограниченности человеческих ресурсов отслеживание изменений является трудновыполнимой задачей. К тому же, отсутствует инструмент для оценки возможности той или иной модификации образовательной программы.

Данная проблема была сформулирована в рамках работы над проектом «Цифровой след студента», целью которого является применение методов и технологий машинного обучения для сопровождения индивидуальных образовательных траекторий на основе анализа цифрового следа обучающихся. Цифровой след — это постоянно пополняемый набор данных, включающий как значения традиционных показателей (различные аттестации, посещение занятий и т. д.), так и тексты, созданные самими студентами. В их числе рефераты и обзоры литературы, курсовые, отчеты по практикам, описания проектов, выпускные квалификационные работы, эссе и мотивационные письма на конкурсы. Именно из этих текстов современные методы анализа данных позволяют извлечь объективную информацию для диагностики

профессиональной компетентности выпускника и выявить факторы, которые повлияли на ее формирование. Исследование выполнено при поддержке РФФИ и НТУ «Сириус» в рамках научного проекта № 19-37-51028.

Цель работы: разработать модуль, предназначенный для актуализации образовательной программы. При этом, изменение образовательных программ предполагается в разрезе отдельных дисциплин.

Для достижения данной цели требуется выполнить следующие задачи:

1. Изучить архитектуру системы и хранилища проекта «Цифровой след студента»
2. Сформировать корпус текстов для исследования
3. Изучить методы анализа текстов на естественном языке
4. Разработать алгоритм работы модуля по актуализации образовательной программы

Для подготовки и защиты выпускной квалификационной работы использовались поиск, анализ информации, системный подход для решения поставленных задач; приемы критического анализа проблемных ситуаций, а также средства и методы саморазвития и самореализации; методики межкультурного взаимодействия; умение расставлять приоритеты собственной деятельности при работе в общем проекте в соответствии с командной стратегией для достижения поставленной цели.

Формулирование выводов по итогам проведенной работы осуществлялись с учетом применения современных коммуникативных технологий (в том числе на иностранном языке) для представления результатов на академических, профессиональных, экспертных ИТ-мероприятиях.

СПИСОК ЛИТЕРАТУРЫ

1. Blei D., Ng A., Jordan M. Latent Dirichlet allocation // Journal of Machine Learning Research. 2003. №3. P. 993–1022.
2. DiGraph — Directed graphs with self loops — NetworkX 2.5 documentation [сайт]. URL: <https://networkx.org/documentation/stable/reference/classes/digraph.html> (дата обращения 28.05.2021)
3. Distance computations (scipy.spatial.distance) — SciPy v1.6.3 Reference Guide [сайт]. URL: <https://docs.scipy.org/doc/scipy/reference/spatial.distance.html> (дата обращения 25.03.2021)
4. Gensim: Topic modelling for humans [сайт]. URL: <https://radimrehurek.com/gensim/> (дата обращения 10.04.2021)
5. Hierarchical topic models and the nested Chinese restaurant process / D. Blei, T. Griffiths, T. Jordan, J. Tenenbaum // Neural Information Processing Systems. 2003. №16. P. 17–24.
6. Hofmann T. Probabilistic Latent Semantic Indexing // SIGIR. 1999. P. 50–57.
7. Knowledge discovery through directed probabilistic topic models: a survey. / D. Ali, L. Juanzi, Z. Lizhu, M. Faqir // In Proceedings of Frontiers of Computer Science in China. 2010, P. 280–301. — перевод на русский К. В. Воронцов, А. В. Темлянец и др.
8. Matplotlib: Python plotting — Matplotlib 3.4.2 documentation [сайт]. URL: <https://matplotlib.org/> (дата обращения 25.03.2021)
9. Mikolov T., Sutskever I., Chen K., Corrado G., Dean J. Distributed Representations of Words and Phrases and their Compositionality (2013). URL: <https://papers.nips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf> (дата обращения 15.05.2021)
10. Newman, Lau, Grieser, Baldwin. Automatic Evaluation of Topic Coherence // NAACL HLT. 2010. P.100–108.
11. Nltk.stem package — NLTK 3.6.2 documentation [сайт]. URL: <https://www.nltk.org/api/nltk.stem.html> (дата обращения 1.04.2021)

- 12.Re — Regular expression operations — Python 3.9.5 documentation [сайт]. URL: <https://docs.python.org/3/library/re.html> (дата обращения 29.03.2021)
- 13.Single Word — wordcloud 1.8.1 documentation [сайт]. URL: https://amueller.github.io/word_cloud/auto_examples/single_word.html (дата обращения 25.03.2021)
- 14.Sklearn.feature_extraction.text.TfidfVectorizer — scikit-learn 0.24.2 documentation [сайт]. URL: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html (дата обращения 29.03.2021)
- 15.Автоматическая обработка текстов на естественном языке и анализ данных: учеб. пособие / Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. М.: Изд-во НИУ ВШЭ, 2017. 269 с.
- 16.Велихов П.Е. Меры семантической близости статей Википедии и их применение к обработке текстов // Информационные технологии и вычислительные системы. 2009. №1. С. 21 – 37.
- 17.Воронцов К.В. Вероятностное тематическое моделирование, 16 октября 2013 г.. URL: http://www.machinelearning.ru/wiki/images/9/9a/Sem1_knn.pdf (дата обращения 24.04.2021)
- 18.Емельянов А.А., Власова Е.А. Актуализация образовательных программ и планирование подготовки преподавателей // Высшее образование в России. 2009. №1. С.100 – 111.
- 19.Кластеризация документов [сайт]. URL: <https://amp.ru.google-info.cn/234548/1/klasterizatsiya-dokumentov.html> (дата обращения 12.04.2021)
- 20.Кластеризация. [сайт]. URL: <http://neerc.ifmo.ru/wiki/index.php?title=Кластеризация> (дата обращения 13.04.2021)

21. Коршунов А., Гомзин А. Тематическое моделирование текстов на естественном языке // Труды Института системного программирования РАН (электронный журнал). 2021. №23. С. 215 – 242.
22. Половикова О.Н., Фокина В.В. Использование евклидова и манхэттенского расстояний в качестве меры близости для решения задачи классификации // Известия АГУ. 2010. №1(65). С. 101 – 102.
23. Тематическое моделирование с помощью Gensim (Python) - Еще один блог веб-разработчика [сайт]. URL: <https://webdevblog.ru/tematicheskoe-modelirovanie-s-pomoshhju-gensim-python/> (дата обращения 10.04.2021)