А.А. Ключникова [1], Ю.В. Боганюк[1,2]

[1] Тюменский государственный университет, г. Тюмень

[2] Научно-технический университет «Сириус», г.Сочи

УДК 004.912

АНАЛИЗ УРОВНЯ ПРОФЕССИОНАЛИЗМА СТУДЕНТОВ НА ОСНОВЕ ТЕКСТОВ ИХ КУРСОВЫХ, ПРОЕКТНЫХ И ВЫПУСКНЫХ РАБОТ

Аннотация. В статье представлен модуль для анализа уровня профессионализма студентов, а также возможные пути повышения эффективности такого анализа.

Ключевые слова: качество обучения, анализ данных, текст, ширина словаря.

Введение

Повышение качества образования всегда была и остается актуальной задачей для любого университета. Основной проблемой является выявление критичных мест в программах обучения, подходах и технологиях образования, в восприятии материала студентами.

Задача управления образовательным процессом невозможна без анализа имеющейся информации о процессе обучения, ведь его успех в первую очередь зависит от участников самого процесса. Информация об образовательном процессе включает в себя не только данные об успеваемости студентов, об учебных программах, но и текстовые данные обучающихся: курсовые работы, отчеты о различных практических работах, рефераты, эссе и так далее. Текстовые данные содержат много информации об уровне подготовки студентов, но, в отличие от анализа числовых данных, анализ текстовой информации гораздо сложнее и более

затратен по времени.

Использование технологий машинного обучения могут помочь в решении данной проблемы, не только уменьшая затраты по времени на анализ данных, но и повышая уровень анализа. Основная идея решения основана на следующей гипотезе: в процессе обучения используемая литература становится более специализированной, количество терминов в работах увеличивается (многие материалы издаются на английском языке и в русском языке еще нет устоявшихся терминов, поэтому термины могут встречаться и на английском языке), при таких предположениях уровень профессионализма студента на каждом году обучения можно выявить, основываясь на количестве терминов, используемых в его работах. Таким образом, чем глубже и детальнее студент разбирается в какой-то теме, тем больше терминов, приведенных как на русском, так и на английском языке, появляется в его работах.

Целью исследования является анализ имеющихся текстовых данных студентов для проверки справедливости сформулированной гипотезы.

Материалы и методы

Для анализа использовались тексты студентов института математики и компьютерных наук тюменского государственного университета. Были проанализированы выпускные квалификационные работы 16 студентов за 2015-2016 учебный год и тексты магистерских работ этих же 16 студентов за 2017-2018 учебный год. Студенты обучались на разных направлениях института математики и компьютерных наук.

Для анализа данных был реализован модуль на языке Python. Входными данными модуля являлись следующие данные: идентификационный номер студента, год написания работы и непосредственно текст работы.

Был разработан и реализован следующий алгоритм:

1. Работы каждого студента сортировались по дате написания.
2. Для каждой работы студента проводились следующие операции:
    a. Обработка исходного текста
    b. Подсчет количества слов (поиск ширины словаря)
    c. Подсчет количества терминов
3. После анализа работ всех студентов подводилась итоговая информация:
    a. Подсчет числа студентов, в работах которых увеличилось число слов за время обучения
    b. Подсчет числа студентов, в работах которых увеличилось число терминов за время обучения
    c. Подсчет числа студентов, в работах которых увеличилось соотношение числа слов к числу терминов за время обучения

Обработка исходных текстов заключалась в следующем. Исходные данные текстов были обработаны путем извлечения слов без знаков препинания. Затем слова были лемматизированы с помощью pymorphy2.MorphAnalyzer [1] и применен стемминг [2] для русских и английских слов с помощью nltk.stem.SnowballStemmer [3]. Применение данных методов было сделано с целью поиска слов, которые были использованы в различных формах, а также однокоренных слов, что позволило более точно распознать различные слова.

Подсчет различных слов в тексте, то есть поиск ширины словаря студента, был осуществлен с помощью инструмента sklearn.feature_extraction.text.TfidfVectorizer [4].

Для поиска терминов в тексте были извлечены слова из словаря «Англо-русский и русско-английский словарь технических терминов» 2001г., В.Д., Табунщиков Ю.А., Бродач М.М. [5] Количество слов в пересечении слов текстов работ и слов из словаря является количеством

терминов в каждой работе.

Обсуждение результатов

Проведенный анализ привел к следующим результатам:

1. Число студентов, в работах которых увеличилось число слов составляет 8 из 16 исследуемых студентов. Максимальная разница в количестве слов составляет 1120 слов, то есть 71.13% от максимального количества (начальные значения 451 и 1571 слов). Минимальная разница составляет 78 слов, то есть 0.08 % от максимального количества (начальные значения 840 и 918 слов).

2. Число студентов, в работах которых увеличилось число терминов составляет 8 из 16 исследуемых студентов. Максимальная разница в количестве терминов составляет 227 терминов, то есть 74.18% от максимального количества (начальные значения 79 и 306 терминов). Минимальная разница в количестве терминов составляет 3 термина, что составляет 0.01% от максимального количества (начальные значения 253 и 256 терминов)

3. Число студентов, в работах которых увеличилось соотношение числа терминов к ширине словаря составляет 7 из 16 исследуемых студентов. Максимальная разница составляет 0.056, то есть 25% от максимального количества (начальные значения 0.168 и 0.224 терминов к количеству слов). Минимальная разница в количестве терминов составляет 0.0003, что составляет 0.17% от максимального количества (начальные значения 0.1776 и 0.1779 терминов к количеству слов)

Такие результаты предварительного анализа данных подтверждают существование проблемы разного уровня усвоения материала студентами, а соответственно и разного уровня профессионализма после обучения в университете.

Так как известно, что исследуемые студенты обучались на разных

направлениях, то следующим шагом является разделение студентов на группы по направлениям обучения и проведение такого анализа по этим группам для более детального анализа. Также планируется увеличение общего числа исследуемых студентов и их работ. Такой подход позволит определить отличия в подготовке для различных направлений.

Также планируется улучшить анализ текстов за счет использования следующих категорий поиска и предположений:

- Поиск цитат и ссылок на литературу. Ссылкой считаем предложение или несколько предложений, после которых указан источник в [ ]
- Поиск терминов на английском и русском языках отдельно
- Объем цитат более заданного процента от общего объема работы свидетельствует о желании студента увеличить объем работы.

Благодарности

Статья подготовлена в рамках разработки образовательного кейса для НТУ Сириус при финансовой поддержке РФФИ в рамках научного проекта № 19-37-51028.

СПИСОК ЛИТЕРАТУРЫ

1. Морфологический анализатор pymorphy2 [Электронный ресурс]. – Режим доступа: https://pymorphy2.readthedocs.io/en/latest/user/guide.html, свободный (Дата обращения: 01.05.2020)

2. Стемминг [Электронный ресурс]. – Режим доступа: https://textis.ru/stemming/, свободный (Дата обращения: 01.05.2020)

3. Stemmers [Электронный ресурс]. – Режим доступа https://www.nltk.org/howto/stem.html, свободный (Дата обращения: 01.05.2020)

4. Документация библиотеки scikit-learn [Электронный ресурс] – Режим доступа: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html, свободный (Дата обращения: 01.05.2020)

5. Коркин В.Д., Табунщиков Ю.А., Бродач М.М. Англо-русский и русско-английский словарь технических терминов, Москва Авок Пресс, 2001г. [Электронный ресурс]. – Режим доступа: http://padabum.com/d.php?id=18822 (Дата обращения: 17.05.2020)