СПИСОК ЛИТЕРАТУРЫ

1. Реализация искусственных нейронных сетей в Линукс: научный журнал Novainfo : [сайт]. — URL: https://novainfo.ru/article/2010. — Текст : электронный.
2. Scaled-YOLOv4 : [сайт]. — URL: https://github.com/WongKinYiu/ScaledYOLOv4 (дата обращения: 16.05.2022). — Текст : электронный.
3. YOLOv5 : [сайт]. — URL: https://github.com/ultralytics/yolov5 (дата обращения: 16.05.2022). — Текст : электронный.
4. D. Cochard. YOLOv5: The Latest Model for Object Detection : [сайт]. — URL: https://medium.com/axincai/yolov5-the-latest-model-for-object-detection-b13320ec516b. — Текст : электронный.
5. Documentation Mapbox : [сайт]. — URL: https://docs.mapbox.com/ (дата обращения: 16.05.2022). — Текст : электронный.

А. И. Стрельников, М. С. Воробьева

Тюменский государственный университет, г. Тюмень

УДК 004.912

ИССЛЕДОВАНИЕ МЕТОДОВ АНАЛИЗА ИНФОРМАЦИОННОЙ И ЛЕКСИЧЕСКОЙ НАСЫЩЕННОСТИ НАУЧНЫХ ТЕКСТОВ

Аннотация. *В статье представлен обзор существующих методов анализа информационной и лексической насыщенности. Выдвинута гипотеза корреляции критериев лексической и информационной насыщенности. Гипотеза проверена на основе результатов анализа 752 ВКР студентов ИМиКН за 6 лет обучения.*

Ключевые слова: *информационная насыщенность, информативность, абстрактность, водность, плотность ключевых слов, лексическая насыщенность, лексическое многообразие, лексическая сложность*.

Введение. При написании текстов научного стиля авторы заинтересованы в повышении информационной насыщенности своих работ при сохранении понятности для читателя. Необходимо провести исследование методов определения информационной насыщенности и факторов, влияющих на насыщенность. Авторы Ю. В. Морозова и И. А. Уртамова в статье [1] и Н. К. Криони и др. в статье [2]

используют для анализа научных и учебных текстов количественные критерии информационной насыщенности (информативность, абстрактность, водность, плотность ключевых слов). В работах [3, 4] анализируется лексическая насыщенность текстов. Исследования информационной насыщенности применяют количественные критерии, однако в них не учитываются критерии лексической насыщенности.

Проблема исследования. В рамках данного исследования была выдвинута гипотеза о зависимости критериев информационной насыщенности от метрик лексической насыщенности и сформулирована цель — выявить корреляцию критериев лексической и информационной насыщенности. Для достижения данной цели были сформулированы следующие задачи:

- изучить различные критерии анализа информационной насыщенности текста и определить критерии анализа научных текстов;
- разработать и исследовать методы анализа информационной насыщенности;
- проанализировать информационную насыщенность текстов ВКР студентов различных направлений Института математики и компьютерных наук ТюмГУ;
- провести ранжирование работ по выделенным критериям для проверки корреляции лексической и информационной насыщенности текста.

Формальная постановка задачи

Пусть даны тексты работ $T_i, \mathrm{i} = 1..\mathrm{I}$

Для каждого текста T_i необходимо извлечь количественные критерии информационной насыщенности (1) и критерии лексической насыщенности (2):

$$K_m, m = 1..M \tag{1}$$

$$L_n, n = 1..N \tag{2}$$

Для текста T_i провести корреляционный анализ каждой пары критериев (3):

$$\{K_m ; L_n\}, m = 1..M, n = 1..N \tag{3}$$

Материалы и методы

Количественные критерии информационной насыщенности

Для проведения анализа введем следующие понятия: информативность, абстрактность, водность, плотность ключевых слов.

Для вычисления критерия информативности в анализируемом тексте считается доля введенных определений. Перед анализом проводится токенизация текста и приведение слов к начальной форме. Определение узнается по характерным диагностическим признакам, словам в начальной форме (например: «называть», «означать», «обозначать», «называться», «обозначаться», «определяться», «пониматься») [1, 2]. Также в анализируемом тексте проводится поиск с помощью регулярного выражения (4), чтобы обнаружить признак определения «*слово — слово*».

<[А-Яа-яеA-Za-z]+ – [А-Яа-яеA-Za-z]{2,}> (4)

Для вычисления абстрактности в анализируемом тексте считается доля абстрактных слов, которыми считаются существительные со заданными суффиксами («ние», «тие», «ств», «аци», «есть», «ость», «изм», «изн», «ота», «ина», «ика», «тив»).

Для вычисления критерия водности в ходе исследования был составлен список стоп-слов в алфавитном порядке. Для каждого слова в анализируемом тексте проведен бинарный поиск в списке стоп-слов.

Для вычисления плотности ключевых слов из анализируемого текста извлекаются N ключевых слов (по умолчанию N = 10) и считается их средняя доля в тексте.

Критерии лексической насыщенности

К. Кайл в своей работе [4] выделяет три группы критериев лексической насыщенности: критерии плотности, многообразия и сложности.

Критерии лексического многообразия оперируют такими понятиями, как токен и тип. Токенами являются отдельные слова в тексте, а типами — уникальные слова. Все токены приведены к начальной форме с помощью программы mystem от компании Яндекс [5]. Самый базовый способ определения лексического многообразия —

это вычисление метрики ttr (5) (Type-Token Ratio) или ее вариаций root_ttr (6) и log_ttr (7). Более сложной вариацией TTR является индекс Мааса (8).

$$\frac{NType}{NToken} \tag{5}$$

$$\frac{NType}{\sqrt{NToken}} \tag{6}$$

$$\frac{lg(NType)}{lg(NToken)} \tag{7}$$

$$\frac{lg(Ntoken) - lg(NType)}{(lg(Ntoken))^2} \tag{8}$$

Метрики ttr зависимы от длины текста, поэтому Ф. Зенкер и К. Кайл [6] предлагают использовать вариации ttr, не зависящие от длины текста:

- msttr (mean segment ttr): текст разбивается на сегменты по 50 слов и считается среднее сегментное ttr. Если последний сегмент содержит меньше 50 слов, то он отсекается;
- mattr (moving average ttr): считается среднее ttr для слов 1-51, 2-52 и т. д.

Еще одной метрикой, не зависящей от длины текста, является mtld (measure of textual lexical diversity) — среднее число слов, которое требуется, чтобы TTR достигла критического значения 0.72. Для вычисления критериев данной группы используется библиотека для Python lexical_diversity.

Средняя частотность и средний диапазон документов относятся к критериям лексической сложности. Для вычисления данных критериев требуется корпус текстов. В корпусе для каждого слова вычисляется его средняя частотность на миллион слов и число документов, в которых это слово встречается [7]. Используются значения частотности и диапазона из нового частотного словаря русской лек-

сики [8]. Данный частотный словарь составлялся на основе Национального корпуса русского языка объемом 100 млн. словоупотреблений, содержащего тексты различных стилей. Словарь содержит около 52000 наиболее частотных слов.

Доля слов из Russian Academic Vocabulary List в тексте является критерием лексической сложности [9]. Список русских академических слов содержит 640 слов, которые часто используются в академических текстах и нечасто в других. Предварительно список отсортирован в алфавитном порядке, для каждого слова в оригинальном тексте проведен бинарный поиск в списке академических слов.

Результаты анализа. В рамках исследования была проанализирована информационная насыщенность 752 ВКР студентов Института математики и компьютерных наук ТюмГУ по 10 направлениям за 6 лет обучения (2016-2021 гг.). Использование ВКР различных направлений обеспечивает разнообразие корпуса, на основе которого был проведен корреляционный анализ категорий лексической и информационной насыщенности (рис. 1).

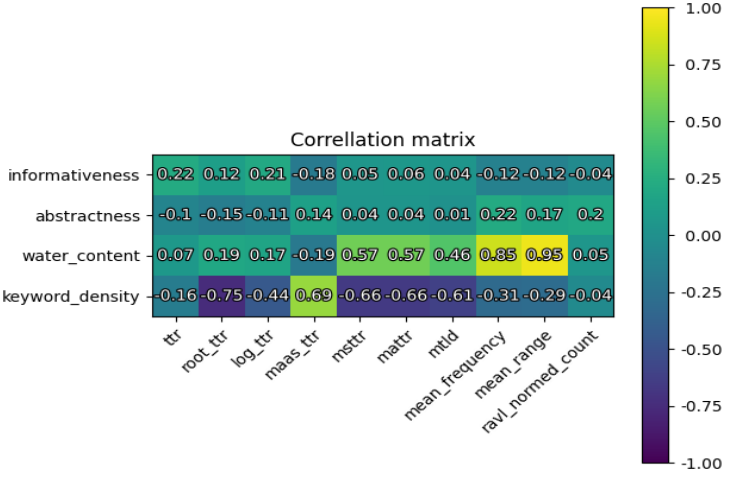

Рис. 1. Корреляционная матрица

Результаты говорят о том, что критерий информативности слабо коррелирует ($r \leq 0.3$) с лексической насыщенностью. Следовательно,

для повышения информативности текста автору не требуется тщательно следить за используемыми словами, а только вводить больше новых понятий. Критерий абстрактности также слабо коррелирует с лексической насыщенность: число абстрактных слов набирается естественным образом в ходе написания научного текста.

Водность текста сильно коррелирует ($r \geq 0.5$) с теми метриками лексического многообразия, которые не зависят от длины текста, а также со средней частотностью слов в тексте (рис. 2), поскольку водным текст делают местоимения, предлоги, союзы и другие служебные слова, которые являются наиболее частотными словами русского языка.

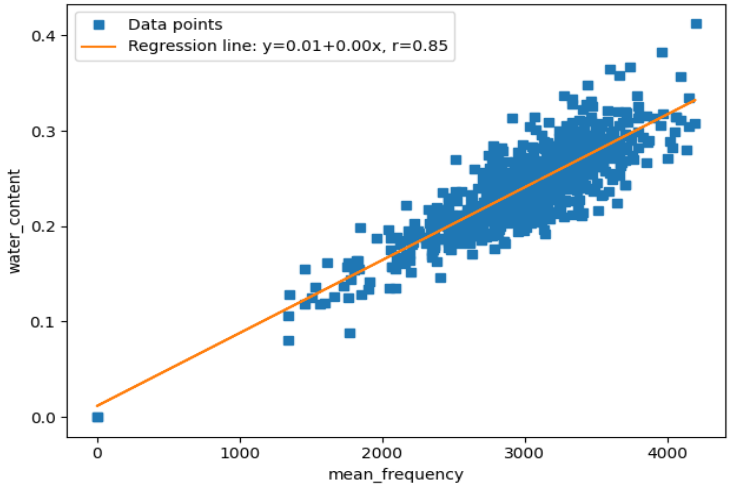

Рис. 2. График зависимости водности от средней частотности слов

Плотность ключевых слов имеет сильную отрицательную корреляцию с метриками лексического многообразия (рис. 3). Из этого следует, что чем более разнообразная лексика используется в тексте, тем меньше внимания уделено ключевым понятиям для данного текста.

Кроме того, плотность ключевых слов имеет среднюю отрицательную ($r \approx -0.3$) корреляцию со средней частотностью слов в тексте (рис. 4). Это объясняется тем, что ключевые слова зачастую

являются специфичными для тематики конкретного текста и редко встречаются в общих корпусах текстов.

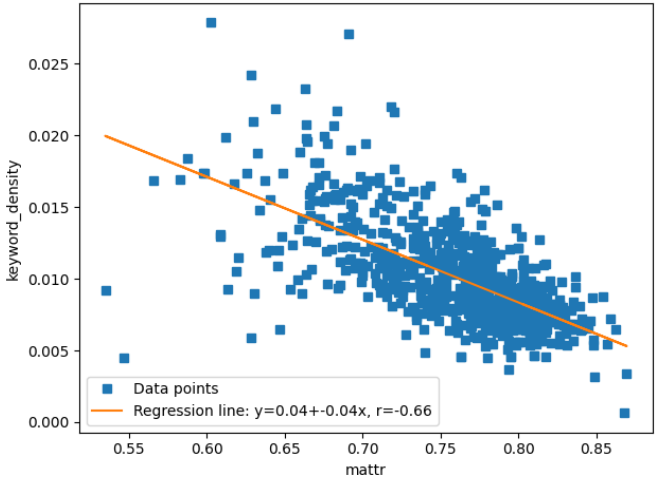

Рис. 3. График зависимости плотности ключевых слов от метрики mattr

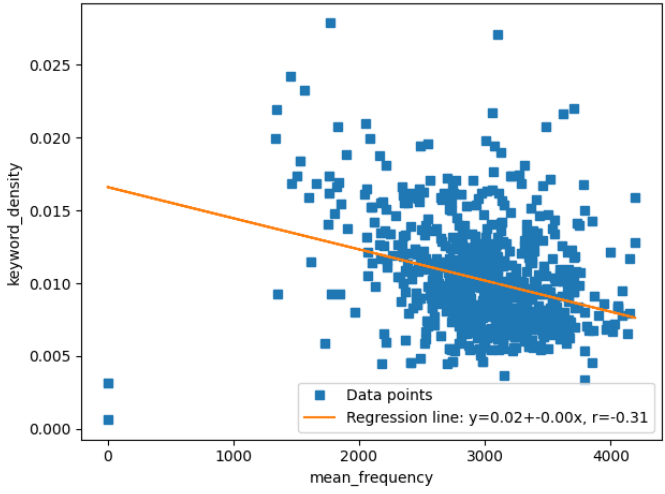

Рис. 4. График зависимости плотности ключевых слов от средней частотности слов

Метрики mattr, msttr и mtld показывают схожие результаты, поэтому целесообразно использование только метрики mattr, которую рекомендует К. Кайл в работе [4].

Заключение. В ходе исследования был проведен обзор существующих критериев информационной и лексической насыщенности. С помощью критериев был проведен анализ информационной насыщенности ВКР студентов Института математики и компьютерных наук.

В результате корреляционного анализа было решено использовать при анализе текстов все количественные критерии информационной насыщенности и все критерии лексической сложности. Однако было решено отказаться от метрик лексического многообразия, которые сильно зависят от длины анализируемого текста.

СПИСОК ЛИТЕРАТУРЫ

1. Морозова Ю. В. Методика анализа электронного учебного контента / Ю. В. Морозова, И. А. Уртамова. — Текст : непосредственный // Открытое и дистанционное образование. — 2017. — № 4. — С. 68.
2. Криони Н. К. Автоматизированная система анализа сложности учебных текстов / Н. К. Криони, А. Д. Никин, А. В. Филиппова. — Текст : непосредственный // Вестник Уфимского государственного авиационного технического университета. — 2008. — Т. 11, № 1. — С. 101-107.
3. Torruella J. Lexical statistics and tipological structures: a measure of lexical richness / J. Torruella, R. Capsada. — Text : direct // Procedia-Social and Behavioral Sciences. — 2013. — Т. 95. — С. 447-454.
4. Kyle K. Measuring lexical richness / K. Kyle. — Text : direct // The Routledge handbook of vocabulary studies. — 2019. — С. 454-475.
5. Segalovich I. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine / I. Segalovich. — Text : direct // MLMTA. — 2003. — Vol. 2003. — P. 273.
6. Zenker F. Investigating minimum text lengths for lexical diversity indices / F. Zenker, K. Kyle. — Text : direct // Assessing Writing. — 2021. — Vol. 47. — P. 100505.
7. Kyle K. Automatically assessing lexical sophistication: Indices, tools, findings, and application / K. Kyle, S. A. Crossley. — Text : direct // Tesol Quarterly. — 2015. — Vol. 49, № 4. — P. 757-786.
8. Ляшевская О. Н. Новый частотный словарь русской лексики / О. Н. Ляшевская, С. А. Шаров. — Текст : электронный // Словари на основе Национального корпуса русского языка : [сайт]. — URL: http://dict.ruslang.ru/freq.php (дата обращения: 24.05.2022).

9. Talalakina E. Developing and Validating an Academic Vocabulary List in Russian: A Computational Approach / E. Talalakina, D. Stukal, M. Kamrotov. — Text : direct // The Modern Language Journal. — 2020. — Vol. 104, № 3. — P. 618-646.

Д. К. Зитцер, А. Г. Ивашко
Тюменский государственный университет, г. Тюмень
УДК 004.912

ПОДСЧЕТ ПОСЕТИТЕЛЕЙ В РИТЕЙЛЕ НА ОСНОВЕ МОДЕЛИ ДЕТЕКЦИИ И АЛГОРИТМА ОТСЛЕЖИВАНИЯ

Аннотация. *В работе представлен процесс разработки программного обеспечения для подсчета посетителей в ритейле. Подробно описаны процессы подготовки данных и обучения модели детекции.*
Ключевые слова: *подсчет посетителей, компьютерное зрение, машинное зрение, нейронные сети, детекция объектов, трекинг объектов.*

Введение. Высокий уровень конкуренции в российском ритейле ставит управленцев перед выбором: как улучшать обслуживание клинетов, отслеживать максимум показателей магазина — при это не разориться на дорогостоящих IT-продуктах. В данной работе будет представлен простой, с точки зрения разработки, способ подсчета посетителей торговой точки.

Зачем считать посетителей?

Подсчет посетителей в торговой точке необходим для того, чтобы:

1) корректировать график работы персонала по времени и дням недели;

2) выстраивать систему мотивации персонала на основе коэффициента конверсии для улучшения качества обслуживания;

3) планировать и проводить маркетинговые акции, оценивать их результаты и корректировать бюджет.

Существующие способы подсчета

Инфракрасные счетчики (табл. 1):

- технология 2000-х гг.;
- наиболее распространенная система;
- установка по бокам от входной группы.