Д. В. Степанова, М. А. Жданович

ОНЛАЙН-СЕРВИС ISTIO И ЕГО ВОЗМОЖНОСТИ ДЛЯ СЕМАНТИКО-СТАТИСТИЧЕСКОГО АНАЛИЗА ТЕКСТОВ

В настоящее время существует потребность в оперативной, а значит автоматической, обработке и анализе значительных массивов научно-технических текстов на английском языке. Для решения такой актуальной задачи разработаны различные программные средства, многие из которых находятся в открытом доступе. Такое программное обеспечение, как правило, действует на основе статистических методов анализа данных и нуждается в постоянном обновлении и совершенствовании. Одним из таких программных продуктов для семантико-статистического анализа текстов является онлайн-сервис Istio, который доступен пользователям по адресу www.istio.com.

Данное программное обеспечение было использовано нами в качестве средства для выделения ключевых слов из отобранных научно-технических текстов предметной области «Информационные технологии» с целью моделирования глоссария терминов. Для проведения эксперимента по определению возможностей сервиса Istio для статистического анализа научно-технических текстов нами были выбраны 20 статей из журнала «Future Generation Computer Systems» за 2019 год. В данном журнале рассматривается широкий круг вопросов: обработка и анализ больших данных, системы баз данных, протоколы и стандарты, безопасность, разработка алгоритмов и программ, сети, облачные вычисления, Интернет вещей и т.п. При проведении исследования было принято решение отбирать только названия статей, их аннотации и списки ключевых слов, что позволило повысить концентрацию терминов в создаваемом корпусе, снижая вероятность вхождения в частотные словари общеупотребительных слов. Количественные результаты автоматической обработки отобранного материала представлены на следующем рисунке.

Параметр	Значение
Длина с пробелами ?	32989 символов
Длина без пробелов ?	**28281** символов
Всего слов ?	4699
Водность ?	43%
Тошнота ?	17.23
Топ10 слов ?	datum, algorithm, propose, model, network, cloud, base, ontology, method, detection
Словарь ?	1200 слов
Словарь ядра ?	1002 слов
Язык текста ?	eng
Тематика ?	Неопределена

Без стоп-слов | Со стоп-словами | Словарь | Анализ позиций

Статистика по словам, **за исключением** стоп-слов. ×

#	Слово	Кол-во ?	Рел. ?	% в ядре ?	% в тексте ?
1	datum	64	3.71	2.4%	1.3%
2	algorithm	38	2.2	1.4%	0.8%
3	propose	36	2.08	1.3%	0.7%
4	model	34	1.97	1.2%	0.7%
5	network	32	1.85	1.2%	0.6%
6	cloud	28	1.62	1%	0.5%
7	base	27	1.56	1%	0.5%
8	ontology	20	1.16	0.7%	0.4%
9	method	18	1.04	0.6%	0.3%
10	detection	18	1.04	0.6%	0.3%

Фрагмент окна онлайн-сервиса Istio

Статистические данные разделены на два блока, содержащие семантико-статистический анализ всего обработанного материала и информацию о каждой отдельной лексической единице текста. Первый блок дает представление о таких параметрах текста, как его длина с пробелами и без пробелов, количество слов, водность, тошнота, 10 ключевых слов, список всех слов и словарь ядра, язык и тематика текста. Такие параметры как *длина с пробелами* и *без пробелов* и *всего слов* являются достаточно универсальными и могут быть определены также текстовым редактором Microsoft Word. Тем не менее, сравнительный анализ статистических данных, полученных в обеих программах, показал ряд любопытных отличий. Было установлено, что сервис Istio автоматически выставляет пробел в начале и конце каждого абзаца, что увеличивает общее количество символов с пробелами по сравнению с результатом редактора Microsoft Word.

Погрешность в определении количества слов объясняется тем, что сервис Istio игнорирует числа, в то время как Microsoft Word считает их отдельными словоупотреблениями. Кроме того, были выявлены различия в восприятии программами слов, записанных через косую черту. Например, *fog/edge computing* в Word определяется как двухкомпонентное словосочетание, а в Istio – трехкомпонентное. Других причин расхождения количественных данных в предложенной выборке найдено не было. Следовательно, в рамках проводимого нами исследования, более качественными являются результаты, полученные путем применения сервиса Istio, так как игнорирование чисел и косой черты при статистическом анализе повышает концентрацию ключевых слов в тексте.

Последующие параметры, отраженные на рис. 1, являются специфическими, а для их определения необходимо использование систем автоматической обработки текстов. Параметр *водность* означает процент слов и словесных связок, не несущих в тексте смысловой нагрузки, или стоп-слов. Под стоп-словами разработчики понимают не только предлоги, союзы, вспомогательные глаголы, артикли и т.п., но и слова длиной в 1–2 символа, вводные слова, оценочные эпитеты, клише и ряд других слов без смысловой нагрузки. Как отмечается в документации, имеющейся на сайте, допустимой является водность в 30–60 %. Такой показатель позволяет создать одновременно содержательный и доступный для понимания текст. Таким образом, полученный показатель водности 43 % является удовлетворительным для анализируемого материала.

Следующий параметр, *тошнота*, указывает на отношение самых частотных и значимых слов по специальной формуле, которая не приводится на сайте. Данный параметр не является существенным в проводимом нами исследовании, так как частотность слова указывает на его распространенность, следовательно, значимость для определенной предметной области.

На наш взгляд, недостатком сервиса Istio является отражение в списке *ключевых слов* только 10 лексических единиц, т.е. порогом является вхождение в десятку самых частотных слов, а не фиксированный показатель, рассчитан-

ный по определенной формуле. Фрагментарный анализ предложенной выборки показал, что слова, обладающие одинаковой частотой, располагаются в конечном списке в порядке, соответствующем их позиции в тексте. Сначала фиксируются единицы, которые встречаются в тексте позже всего.

Параметр *словарь* указывает на количество всех слов, содержащихся в тексте. При этом сервис автоматически объединяет словоформы слов, что, несомненно, является достоинством программы. После объединения всех словоформ сервис Istio обнаружил в представленном тексте 1200 слов. Согласно таблице, *словарь ядра* составляет 1002 слова. Следовательно, сервис выявил 198 стоп-слов.

Также программа позволяет автоматически определить *язык* текста и на основе ключевых слов выделить *тематику* текста. В нашем случае тематика не определена, но представленные ключевые слова *datum*, *algorithm*, *network*, *base*, *method* зафиксированы в стандарте ISO/IEC 2382:2015(en) по ИТ в качестве терминов, а слова *detection* и *model* входят в состав терминологических словосочетаний.

Рассматриваемая программа позволяет получить более детальный анализ каждой отдельной единицы, что отражено на рис. 1 во втором блоке, который содержит такие столбцы, как порядковый номер слова, слово, количество, релевантность, % в ядре, % в тексте. В столбце «*Количество*» отражено число словоупотреблений каждой единицы во всем анализируемом тексте. *Релевантность* представляет собой количественные показатели того, насколько определенное слово отражает содержание текста. На сайте не предоставлена информация о способе вычисления данного параметра, что затрудняет его учет в рамках лингвистических исследований.

Следующий параметр, *процент в ядре*, показывает процентное соотношение отдельного слова со всеми значимыми словами текста. Завершающий параметр, *процент в тексте*, отображает процентное соотношение количества отдельного словоупотребления к общему количеству слов в тексте, включая стоп-слова. Данный параметр был определен в качестве первоначального критерия для статистического выделения ключевых слов текста. Следует отметить, что в зависимости от целей исследования можно включить или исключить из списка стоп-слова. Кроме того, во вкладке «*Словарь*» представлены все слова текста в виде списка, организованного по частотному принципу. В данном списке отсутствуют стоп-слова, т.е. он является «обезвоженным».

Таким образом, проведенный анализ возможностей онлайн-сервиса Istio показывает, что данный программный продукт может быть использован не только для семантико-статистического анализа текстов, но и для автоматического определения основных статистических характеристик лексических единиц и для оперативного выделения ключевых слов текстов с учетом их словоформ.