УДК 004.912

ОПРЕДЕЛЕНИЕ АКТУАЛЬНЫХ НАПРАВЛЕНИЙ НАУЧНЫХ ИССЛЕДОВАНИЙ В ОБЛАСТИ ЭКОНОМИКИ С ПОМОЩЬЮ ТЕХНОЛОГИИ TEXT MINING

***Дзюба Дмитрий Владимирович**, ассистент кафедры статистики и эконометрики, ФГБОУ ВО РГАУ-МСХА имении К. А. Тимирязева, flydmitry68@gmail. com*

***Аннотация:** **В** статье рассмотрены **и** реализованы методы интеллектуального анализа текстовой информации (Text Mining) для выделения актуальных направлений зарубежных научных исследований в области экономики с помощью программной среды R*

***Ключевые слова:** научные исследования, экономика, Text Mining, программная среда R, облако тегов.*

В современном мире экономика является одной из важнейших сфер деятельности человека. Экономические отношения охватывают многие стороны нашей жизни, начиная от выбора потребителя в пользу той или иной продукции в магазине и заканчивая принятием решения крупными предприятиями относительно инвестиционных проектов. Экономика страны довольно тесно связана с государственной политикой, поскольку правительство постоянно придерживается определённого экономического плана, от которого напрямую зависит благополучие любого государства. Для России данная проблема имеет ключевое значение в условиях роста напряжённости её отношений с Европейским Союзом и США. Поэтому перед многими исследователями стоит крайне сложная и ответственная задача выбора направления научного исследования, от решения которой в значительной степени зависит правильный курс экономической политики страны.

Одним из важнейших критериев целесообразности данного выбора является актуальность научно-исследовательской работы. Если ранее учёным для этого приходилось монотонно изучать огромное количество отечественных и зарубежных источников, то сейчас современные информационные технологии существенно упрощают подобный процесс. Так, определение

тенденций научных исследований в ведущих развитых и быстро развивающихся экономиках мира может быть выполнено с использованием методов интеллектуального анализа данных (Data Mining) в различных пакетах прикладных программ. Data Mining позволяет обнаружить практически полезные и доступные интерпретации знания и закономерности среди необработанных данных.

Однако исследователь в этом случае будет работать не с количественными данными, а именно с текстами, в связи с чем возникает потребность в применении технологии Text Mining, зародившейся ещё в конце 90-х годов. Данная технология представляет собой одну из разновидностей методов Data Mining и подразумевает процессы извлечения знаний и высококачественной информации из текстовых массивов. Это обычно происходит посредством выявления шаблонов и тенденций с помощью средств статистического изучения шаблонов.

Такая технология глубинного анализа текстов способна обрабатывать большие объемы неструктурированной информации и выявлять из них только самое значимое, чтобы исследователю не приходилось самому тратить время на добычу ценных знаний «вручную». [4]

Одним из средств анализа текстовой информации, находящимся в открытом доступе, является R - программная среда с открытым исходным кодом. R представляет собой открытое программное обеспечение, получившее широкую популярность среди специалистов, которые занимаются анализом и визуализацией данных. Язык R активно применяется ведущими зарубежными компаниями, такими как Google, Bank of America и др., а также ведущими университетами мира. [5]

Для отражения возможностей среды R в качестве исходных данных была сформирована совокупность, включающая 15 авторских статей из ведущих американских и британских экономических жу рналов (Journal of International Economics, Journal of Development Economics и др.). Каждая единица совокупности имеет такие значения переменных, как название статьи, её ключевые слова и аннотация (см. таблицу).

Одним из пакетов, расширяющих возможности среды R в области обработки текстовой информации, является пакет ***tm.*** Он позволяет исследователям применять многочисленные методики к текстовым структурам данных.

Обладая мощным графическим интерфейсом, R предоставляет возможность визуального представления результатов в виде облака тегов с помощью пакета ***wordcloud.*** Облако достаточно легко интерпретируется, так как наиболее часто используемые слова выделяются в нём крупным планом.

Таблица

Исходные данные для обработки текстовой информации зарубежных статей, опубликованных в 2018 году

Num-ber	Article Title	Key Words	Abstract
1	*2*	*3*	*4*
1	Risk versus ambiguity and international security design	ambiguity aversion, risk aversion, debt/equity choice, international capital flows, international insurance, home bias	We study portfolio allocation and characterize contracts issued by firms in the international financial market when investors exhibit ambiguity aversion and perceive ambiguity in assets issued in foreign locations ....
2	The income elasticity of import demand: Micro evidence and an application	income elasticity, income distribution, import demand, Great Trade Collapse	We construct a synthetic panel of household expenditures from the Consumer Expenditure Survey (CEX) and use the Quadratic Almost Ideal Demand System to estimate expenditure shares and income elasticities of demand that vary by good-time-income....
15	What drives bank efficiency? The interaction of bank income diversification and ownership	income diversification, ownership structure, efficiency, banking	This paper examines the relation between income diversification and bank efficiency across 83 countries over the period 2003-2012...

Но прежде, чем построить данный вид графика, текст должен пройти предварительную обработку. Она заключается в том, чтобы удалить из текстовых данных как можно больше так называемых «шумовых слов», т.е. слов, не несущих смысловой нагрузки (например, «it», «this», «with», «ту» и т.д.). Кроме того, R позволяет удалить все числа и пунктуацию. [2]

После предварительной обработки было получено облако тегов, представленное на рисунке 1.

Рис. 1 **Облако тегов**

Исходя из полученных результатов, мы можем отметить, что наиболее частыми тегами в научных исследованиях авторов являются «иностранный», «банк», «страны» и «доходы» (в переводе с английского). Их конкретные частоты встречаемости можно увидеть с помощью построения в среде R гистограммы.

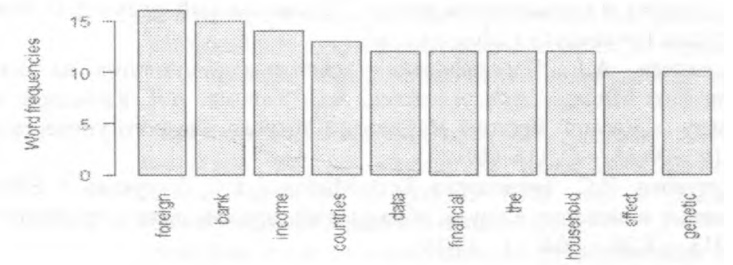

Рис. 2 **Гистограмма часто встречаемых слов**

Каждый отдельный результат не представляет собой ценной информации. Гораздо важнее рассмотреть часто встречаемые словосочетания. В программной среде R существует возможность расчёта корреляции между двумя словами. Для этого достаточно лишь установить минимально возможное значение коэффициента, и консоль отобразит все результаты, которые превысят данный порог.

В частности, тег «доход» имеет высокий коэффициент корреляции (более 0,7) со словами «потребитель», «расходы», «падение», «распределение», «идеальный», «прогнозируемый» и др., тег «банк» — со словами «эффективность», «регулирование», «последствия», «структура», «улучшать» и др., а тег «иностранный» — со словами «корпорации», «эмиграция», «транснациональный», «компании», «выгода» и др. С тегом «страны» какой-либо значимой взаимосвязи не обнаружено.

Таким образом, мы можем предположить, что в настоящее время актуальными являются такие направления экономических исследований, как теория распределения доходов, оценка эффективности регулирования банковского сектора, а также деятельность транснациональных корпораций.

Подводя итоги, отметим, что перед применением технологии Text Mining, текстовая информация всегда должна подвергаться тщательной предварительной обработке. Помимо удаления «шумовых слов», исследователь не должен забывать и о другом важном этапе - стэмминг, где происходит нормализация слов, т.е. их запись в единственном числе, именительном падеже, без особенностей устной речи [3]. Однако это может привести к нарушению семантики, поэтому важно учитывать язык текста. Также для получения более точных и достоверных результатов совокупность необходимо сформировать достаточно большой. Применение технологии Text Mining будет особенно

полезным для магистрантов, аспирантов и независимых учёных при выборе направления научного исследования.

Библиографический список

1. Elsevier [Электронный ресурс] // Режим доступа: https://www.elsevier.com/

2. R: Анализ и визуализация данных [Электронный ресурс] // Режим доступа: https://r-analytics.blogspot.ru/

3. Алексеев, А.А. Классификация текстовых документов на основе технологии Text Mining / А.А. Алексеев, А.С. Катасёв, А.Е. Кириллов, А.П. Кирпичников. - Казань: Вестник Казанского технологического университета, 2016,- Т.19.-№ 18.-С. 116-119

4. Кутукова, Е.С. Технология Text Mining / Е.С. Кутукова // SWorld: Перспективные инновации в науке, образовании, производстве и транспорте. - Одесса, 2013. - Т.30. - №4. - С. 33-36

5. Пиотровская, К.Р. Текст-майнинг: перспективы развития / К.Р. Пиотровская. - СПб.: Российский государственный педагогический университет им. А.И. Герцена, 2014. -№168. - С. 128-134