К. Р. Пиотровская

ТЕХНОЛОГИИ ТЕКСТ-МАЙНИНГА КАК ЭЛЕМЕНТ ИНФОРМАЦИОННОЙ И МАТЕМАТИЧЕСКОЙ КУЛЬТУРЫ СТУДЕНТОВ-ФИЛОЛОГОВ

Неуклонное реформирование российской школьной и вузовской системы привело к неожиданным результатам. С одной стороны, повсеместное закрытие кафедр прикладной лингвистики, сокращение числа ученых советов по специальности **прикладная** и математическая **лингвистика**, отсутствие должной подготовки в области информационных технологий и математических основ обработки информации, а также базовых знаний в области высшей математики, а также изолировали многих филологов-исследователей от возможности грамотно, на доказательном, современном уровне проводить профессиональную исследовательскую работу. Обострению создавшейся ситуации способствовал объективный перенос исследовательских работ с чисто лингвистических площадок на математические и информационные, что было связано с коммерциализацией и стремительным развитием аппарата информатики, математики и квантитативной лингвистики. А высоты, достигнутые в области популяризации применения аппарата математики и теоретической информатики в филологических исследованиях советскими серьезными исследовательскими группами, такими как Всесоюзная группа «Статистика Речи», сегодня во

многом утрачены: хорошие учебники по математике для лингвистов давно не переиздавались в России и постепенно превратились в библиографический раритет [1; 2; 3].

Хотя значительная интенсификация информационного пространства, увеличение его объемов, рассеяние, дублирование и быстрое старение информации усиливают роль информационной и математической культуры для современного студенчества, особенно для студентов-филологов, чья деятельность неразрывно связана с поиском и осмысленной переработкой информации, в то же время, опыт преподавателей свидетельствует, что обращение студентов-филологов к электронным информационным системам и Интернет-ресурсам с целью извлечения знаний путем поиска, синтеза и семантической обработки текстовой информации зачастую имеет негативные результаты. Это связано, в первую очередь, с отсутствием должной подготовки в области оснований математики и математической статистики, а также навыков использования средств, которые предоставляет современная наука для решения этих проблем. К одним из востребованных современной исследовательской деятельностью средств относят технологии дата-майнинг (Data Mining) и текст-майнинг (Text Mining). Эти технологии, основаны на аппарате математической статистики, искусственного интеллекта, экспертных систем, нейронных сетях и др. и предназначены для выявления в текстах скрытой от непосредственного наблюдения информации и ранее неисследованных закономерностей. Оформившись в конце XX века, как направление анализа неструктурированной текстовой информации, технология текст-майнинг стала логическим продолжением дата-майнинга и объединила в себе как классические методы извлечения данных (например, кластеризация) так и методы контент-анализа, статистического анализа и др. [4]. Принципиальное отличие технологии текст-майнинг от дата-майнинга заключается в том, что последняя работает с базами данных, в то время как текст-майнинг позволяет исследователю анализировать обычные тексты, представленные на естественном языке.

На практике использование технологий глубинного анализа текстов открывает для студентов-филологов следующие возможности:

- мониторинг ресурсов Интернет (контент-мониторинг), семантический поиск информации в Интернет и существенное сужение границ поиска за счет включения методов текст-майнинга в современные поисковые системы;
- создание семантических сетей текстов больших объемов, реферирование, классификация и кластеризация текстов, поиск по тексту, интегрирование неструктурированной текстовой информации с существующими структурированными данными, наглядная визуализация кластерезированной текстовой информации.

Довольно большой набор программных продуктов, предоставляет как пробные бесплатные, так и свободно распространяемые версии для проведения исследований:

- QDA Maining, входящий в известный пакет Word Stat, разработанный канадской исследовательской группой Provalis Research, http://provalisresearch.com/;
- RapidMaining, разработка Дортмундского технического университета, Германия, http://rapidminer.com/;

- Carrot2, разработка Познаньского технического университета, Польша, http://project.carrot2.org/;
- Gate, разработка Шеффелдского университета, Англия, https://gate.ac.uk/;
- семейство программ открытого доступа AntConc, разработка Японского университета г. Васеда, http://www.laurenceanthony.net/-software/antconc/;
- Textometrica, разработка университета Умео, Швеция, http://www. simonlindgren. com/textometrica;
- Juxta, разработка университета штата Вирджиния, США, http://www. juxtasoftware.org/;
- SEASR, разработка университета в Иллинойсе, США, http://www. seasr.org/;
- TAPoR Tools, разработка университета Альберта, Канада, http://www. tapor.ca/;
- Textpresso, разработка технологического университета Калифорнии, США, http://www.textpresso.org/;
- Vivisimo/Clusty, разработка университета Карнеги Меллона, США, http://yippyinc.com/about/;
- VisualText, разработка калифорнийской инкорпорации анализа текста, США, http://www.textanalysis.com/;
- Word Hoard, разработка Северо-западный университета, США, http://wordhoard.northwestern.edu/userman/index.html;
- WordSmith, разработка Оксфордского университета, Англия, http://www.lexically.net/wordsmith/index.html;
- свободные текст-майнинговые библиотеки для языков программирования Java, Python, R и C++.

Разработка методики применения этих программных продуктов может послужить основой для составления новых современных учебных пособий по практическому применению студентами-филологами аппарата квантитативной лингвистики, текст-майнинга и автоматической переработки текста.

Активизация их применения в учебной и исследовательской деятельности позволит студентам-филологам ориентироваться в больших информационных потоках, осуществлять более плодотворную деятельность, связанную с поиском, анализом, синтезом электронной текстовой информации, оценивать ее полезность с меньшими временными и энергетическими затратами, формировать систему информационных понятий. Данные обстоятельства делают технологии текст-майнинга необходимым элементом информационной культуры современного студенчества, что обусловливает необходимость преподавания основ этого направления на филологических факультетах.

ЛИТЕРАТУРА

1. *Пиотровский, Р. Г.* Математическая лингвистика / Р. Г. Пиотровский, К. Б Бектаев, А. А. Пиотровская. – М. : Высш. шк., 1977. – 383 с.
2. *Лесохин, М. М.* Введение в математическую лингвистику / М. М. Лесохин, К. Ф. Лукьяненков, Р. Г. Пиотровский. – Минск : Наука и техника, 1982.
3. *Пиотровский, Р. Г.* Методы автоматического анализа и синтеза текста / Р. Г. Пиотровский, В. Н. Билан, А. К. Боркун. – Минск : Выш. шк., 1985.
4. Miner G et all Practical Text Mining and Statistical Analysis for Non-structured Text Data. – N.Y. : Elsver, 2012.