К.Р. Пиотровская (С.-Петербург)

ИНФОРМАЦИОННО- МАТЕМАТИЧЕСКАЯ КУЛЬТУРА СТУДЕНТОВ-ФИЛОЛОГОВ И ОТКРЫТЫЕ РЕСУРСЫ ТЕКСТ-МАЙНИНГА

Совершенно неожиданные результаты были получены вследствии реформирования российской школьной и вузовской системы образования в области филологии, когда в 2000-е годы произошел объективный перенос исследовательских работ с чисто лингвистических площадок на математические и информационные рельсы, что было связано с коммерциализацией и стремительным развитием аппарата информатики, математики и квантитативной лингвистики. Обострению создавшейся ситуации способствовало 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.

Эти события изолировали многих молодых исследователей-филологов от возможности грамотно, на доказательном, современном уровне проводить профессиональную исследовательскую работу. 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 превращаются в библиографический раритет [1, 2, 3].

Хотя значительная интенсификация информационного пространства, увеличение его объемов, рассеяние, дублирование и быстрое старение информации и усиливают значение воспитания 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 лишь 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 ин-

формации, 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 практически без предобработки естественно-языковые тексты.

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 кластеризированной текстовой информации.

Довольно большой набор программных продуктов, предоставляет как пробные бесплатные, так и свободно распространяемые версии для проведения исследований (см. Таблицу 1), также необходимо упомянуть свободные текст-майнинговые библиотеки для языков программирования, таких как C++, Java, Python и R.

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

Активизация применения перечисленных выше технологий в учебной и исследовательской деятельности позволит студентам-филологам научиться 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.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Пиотровский, Р.Г., Бектаев, К.Б., Пиотровская, А.А. Математическая лингвистика. – М.: Высшая школа, 1977.

2. Лесохин М.М. Лукьяненков К.Ф., Пиотровский Р.Г. Введение в математическую лингвистику. – Минск: Наука и техника, 1982.

3. Miner G et all Practical Text Mining and Statistical Analysis for Non-structured Text Data N.Y.: Elsver, 2012.

Таблица 1

Поддерживаемые операции	Названия систем												
	Juxta, Университет штата Вирджиния, США, http://www.juxtasoftware.org/	**RapidMiner,** Дортмундский технический университет, Германия, http://rapidminer.com/	**Carrot 2,** Познаньский технический университет, Польша, http://project.carrot2.org/	**ODAMining**, Provalis Research Group, Канада, http://provalisresearch.com/	**Gate**,Университет г. Шеффелд, Англия, https://gate.ac.uk/	**SEASR**, Университет штата Иллинойс, США, http://www.seasr.org/	**TAPoR Tools**, Университет Альберта, Канада, http://www.tapor.ca/	**Textometrica,** Университет Умео, Швеция,	**AntConc**, Университет г. Васеда, Япония, http://www.laurenceanthony.net/	**Textpresso**, Технологический университет Калифорнии, США, http://www.textpresso.org/	**WordSmith**, Оксфордский университет, Англия, http://www.lexically.net/wordsmith/index.html	**VisualText**, Калифорнийская инкорпорация анализа текста, США,	**Visisimo/Clusty**, Университет Карнеги-Меллона, США
Фильтрация стоп-слов			+			+					+	+	
Выделение ключевых слов		+	+						+	+	+		
Лемматизация			+			+					+		
Токенизация			+		+	+					+		
Тэггизация				+	+						+		
Визуализация результатов	+		+				+						+
Кластеризация		+	+	+					+				+
Категоризация		+		+						+		+	
Суммаризация		+					+					+	
Ассоциации		+											
Коллокация	+	+					+		+		+		
Построение конкордансов							+		+		+		
Построение частотного словаря			+	+			+		+		+		
Статистический анализ текста		+	+	+			+		+	+	+		+