УДК 004

Исследование систем для Text Mining

Ленкин Алексей Викторович
Приамурский государственный университет им.Шолом-Алейхема
студент

Баженов Руслан Иванович
Приамурский государственный университет им.Шолом-Алейхема
к.п.н., доцент, зав.кафедрой информационных систем, математики и методики обучения

Аннотация
В данной статье рассмотрена проблема наличия в мире большого числа текстовой неструктуризированной информации и способы её структуризации и упрощения с помощью прикладных программных средств.
Ключевые слова: Text Mining, KNIME, GATE, RapidMiner.

The study of systems for Text Mining

Lenkin Aleksei Viktorovich
Sholom-Aleichem Priamursky State University
Student

Bazhenov Ruslan Ivanovich
Sholom-Aleichem Priamursky State University
Candidate of pedagogical sciences, associate professor, Head of the Department of Information Systems, Mathematics and teaching methods

Abstract
This article considers the problem of the availability of the world a large number of textual unstructured information and ways of structuring and simplifying with the help of application software.
Keywords: Text Mining, KNIME, GATE, RapidMiner.

Тема «Исследование систем для Text Mining» считается актуальной, так объемы информации будут удваиваться каждые два года в течение следующих восьми лет. Одним из основных факторов этого роста является увеличение доли автоматически генерируемых данных: с 11% от общего объема в 2005 г. до более 40% в 2020 г. [1]. Больше всего создаётся текстовой информации, и, как правило, это довольно крупные по объёму неструктурированные тексты. Поэтому у некоторых людей, в особенности у работающих в IT сфере, возникают проблемы с их обработкой. Для решения этой проблемы были созданы специальные программы для Text Mining. Text

mining или интеллектуальный анализ текста – процесс автоматического анализа обычных неструктурированных текстовых документов компьютером с целью извлечения высококачественной структурированной информации [2].

Задача исследования: установить несколько программ для Text Mining и проверить их работу.

Цель работы: провести анализ этих программ и показать работу одной из них.

Анализ текстовых спам-сообщений с помощью методов Text Mining провели Р.М. Алгулиев и С.А. Назирова [3]. К.Р. Пиотровская в своей статье «Текст-Майнинг: перспективы развития» [4] рассмотрела наиболее популярные программы этого типа и проанализировала их развитие в будущем. О. Абакумов написал об использовании кластеризации в Text Mining [5]. Е.С. Кутукова говорила о самой технологии Text Mining [6]. Технологии Text Mining как элемент информационной культуры студентов-социологов был описан Е.А. Бердником [7]. Г.П. Кожевникова и А.В. Голикова провели исследование по интеллектуализации поисковых систем на основе технологии Text Mining [8]. Н.С. Мясников сделал обзор методов Text Mining [9]. А.А.Алексеев и др.провели классификацию текстовых документов на основе технологии Text Mining [10]. В англоязычных работах B.S.Kumar и V.Ravi сделали обзор приложений Text Mining для финансовой сферы [11]. M.Perovšek и др. описали TextFlows: платформу визуального программирования для Text Mining [12].

Для исследования были выбраны следующие Text Mining инструменты [13, 14]:

1. GATE (рис.1) [15]

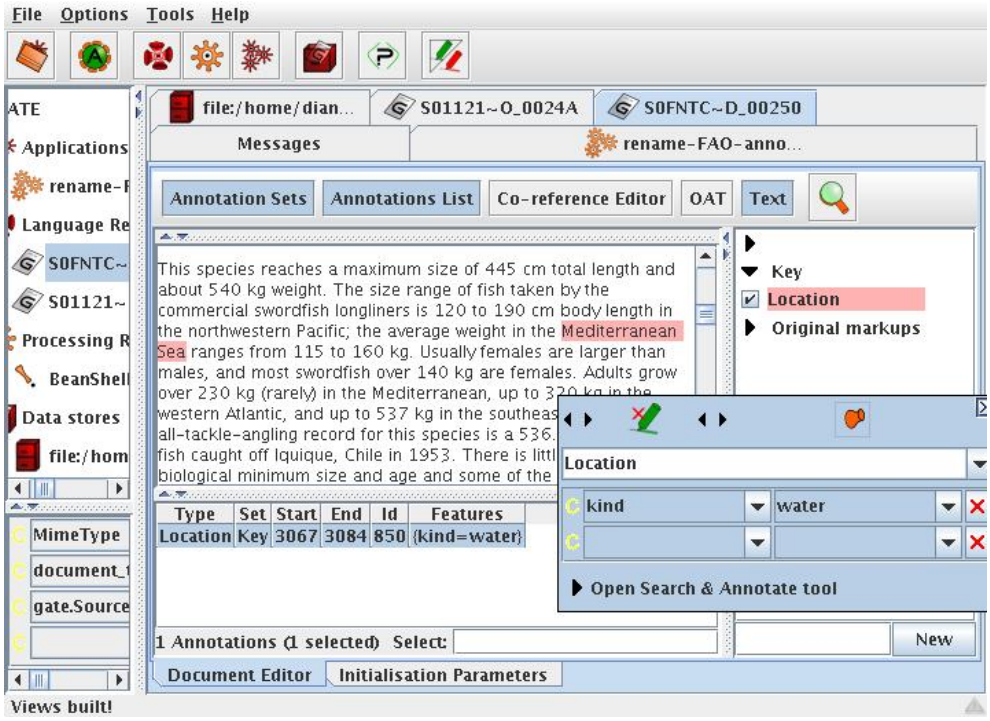

Рисунок 1. Окно программы GATE

General Architecture for Text Engineering (GATE, программа) — система обработки естественного языка с открытым исходным кодом, использующая наборы компонентов на языке Java. Система изначально была разработана в Университете Шеффилда и развивается с 1995 г. [16].

GATE позволяет формализовать и узнать смысловое содержание неструктурированного текста, происходит это при помощи создания аннотаций к частям текста или к отдельным фразам.

2. KNIME (рис. 2) [17]

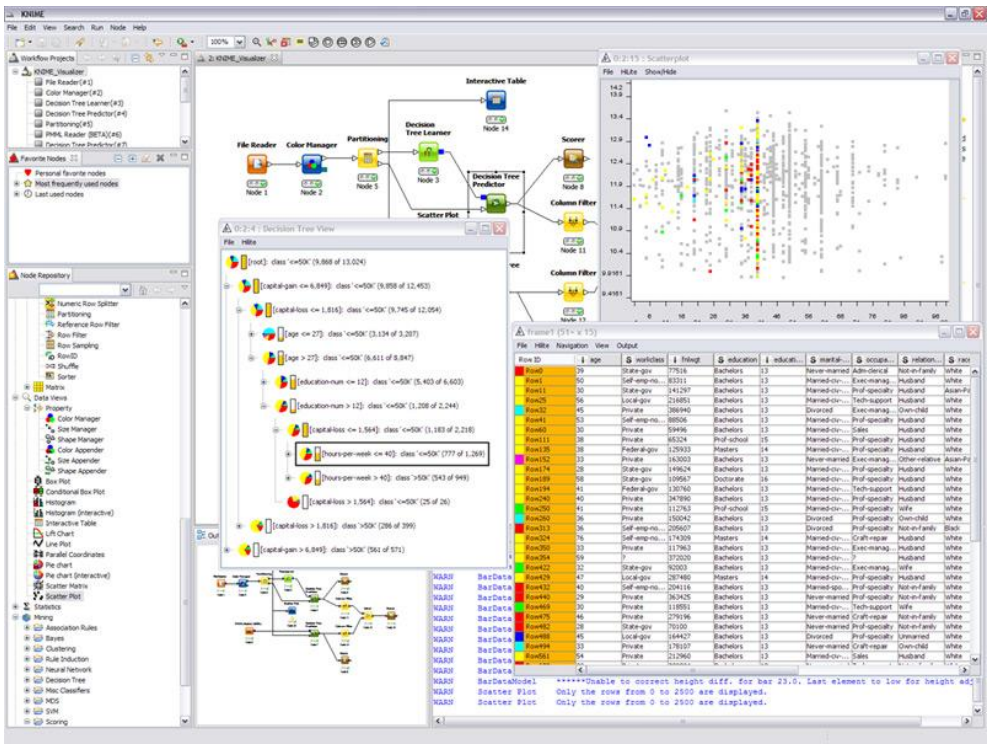

Рисунок 2. Окно программы KNIME

KNIME — это среда на базе Eclipse, первоначально разработанная в 2006 году для анализа данных в фармацевтической отрасли. С тех пор она превратилась в платформу общего назначения для анализа данных, формирования отчетов и интеграции плагинов. Программа работает на основе использования потоков работ, которые отображаются графически как набор узлов, связанных вместе стрелками, указывающими направление потоков данных. Соединяя эти узлы вместе, можно выполнять задачу анализа данных.

Вся логика анализа размещается в узлах, которые служат для приема данных из внешних источников (узлы чтения), преобразования принятых данных и их соединения с данными из других источников. В каждом узле текущее состояние используемых данных можно изучать на экране в виде представления, похожего на электронную таблицу, поэтому можно проверять правильность преобразования данных для конкретной аналитической задачи[18].

3. RapidMiner (рис.3) [19]

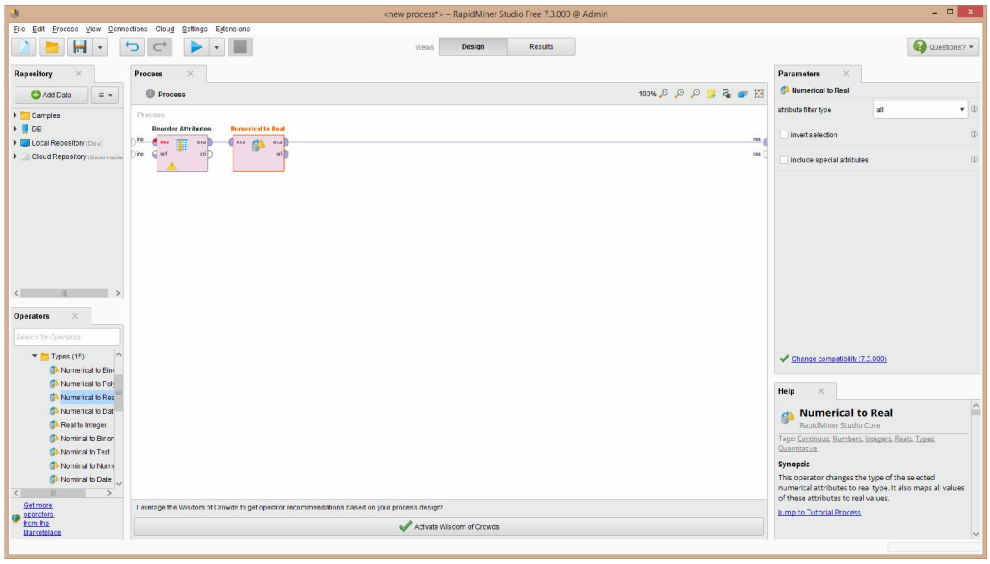

Рисунок 3. Окно программы RapidMiner

RapidMiner (прежнее название YALE) — среда для проведения экспериментов и решения задач машинного обучения и интеллектуального анализа данных. Эксперименты описываются в виде суперпозиций произвольного числа произвольным образом вложенных операторов, и легко строятся средствами визуального графического интерфейса RapidMine*r*-а.

Приложениями RapidMine*r*-а могут быть как исследовательские (модельные), так и прикладные (реальные) задачи интеллектуального анализа данных, включая анализ текста (text mining), анализ мультимедиа (multimedia mining), анализ потоков данных (data stream mining)[20].

Работа RapidMiner схожа с KNIME, так как в нём вся работа сводится также к соединению обрабатывающих узлов. Отличием является то, что RapidMiner может работать с файлами MS Excel.

4. Sisense (рис 4.) [21]

Sisense позволяют не подготовленному пользователю возможность получить доступ к данным и создать визуальное представление информации, отчеты бизнес-аналитики. Sisense включает в себя множество инструментов, чтобы точно определить лучшую визуализацию для данных, таких как: географические карты, линейные диаграммы, чтобы определить тенденции развития, диаграммы рассеяния, чтобы видеть корреляции и круговые диаграммы для четких сравнений.

Рисунок 4. Окно программы Sisense

5. Carrot2 (рис. 5) [22]

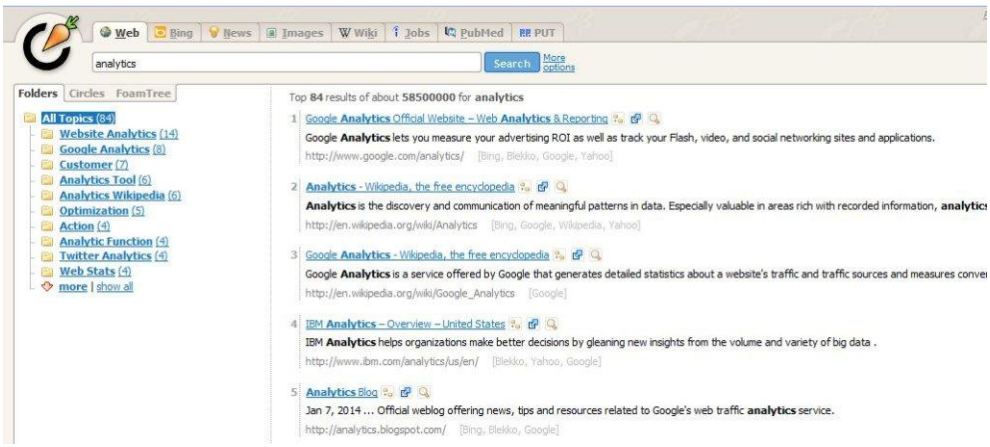

Рисунок 5. Окно программы Carrot2

Carrot2 сортирует текст и результаты поиска по категориям. Он может автоматически сгруппировать небольшие коллекции документов, результаты поиска или краткое описание документов в соответствующие тематические категории. Программа обладает открытым исходным кодом и позволяет изменять группирующий механизм. Кроме поиска внутри программы, Carrot2 также предлагает использовать готовые компоненты для поиска из различных источников такие как GoogleAPI, API Bing, eTools Meta Search, Lucene, SOLR, и другие.

6. KH Coder (рис. 6) [23].

KH Coder - приложение для количественного контент-анализа, Text Mining или корпусной лингвистики. Он может обработать текст на японском, английском, французском, немецком, итальянском, португальском и испанском языках.

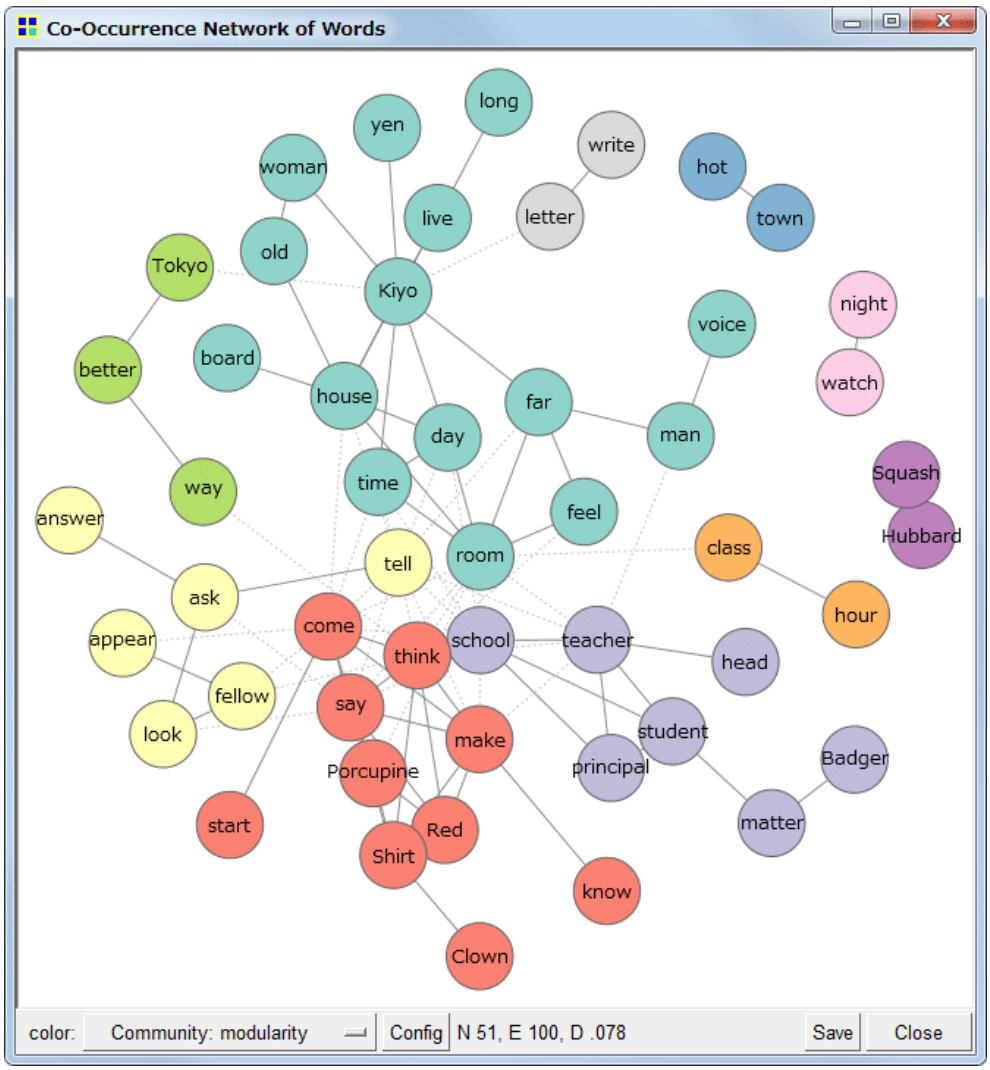

Рисунок 6. Окно программы KH Coder

Для вводимых необработанных текстов можно провести поиск и статистический анализ KWIC, статистику словосочетаний, создать сеть встречаемости, карта связи слов, многомерное масштабирование, кластерный анализ и анализ корреспонденции.. KH Coder обеспечивает различные виды функций поискового и статистического анализа, используя инструменты бэкэнда, такие как Stanford POS Tagger, Snowball stemmer, MySQL и R.

7. tm (Text Mining Infrastructure in R) (рис. 7) [24].

Пакет TM служит основой для приложений Text Mining в R. Функциональность предложений TM для управления текстовыми документами, абстрагирует процесс манипулирования документом и упрощает использование неоднородных текстовых форматов в R. Пакет предоставляет собственную поддержку для чтения в нескольких классических форматах файлов, таких как простой текст, PDFs или XML-файлы. Есть также механизм, чтобы обработать дополнительные форматы файлов. Структуры данных и алгоритмы могут быть расширены, чтобы соответствовать пользовательским требованиям.

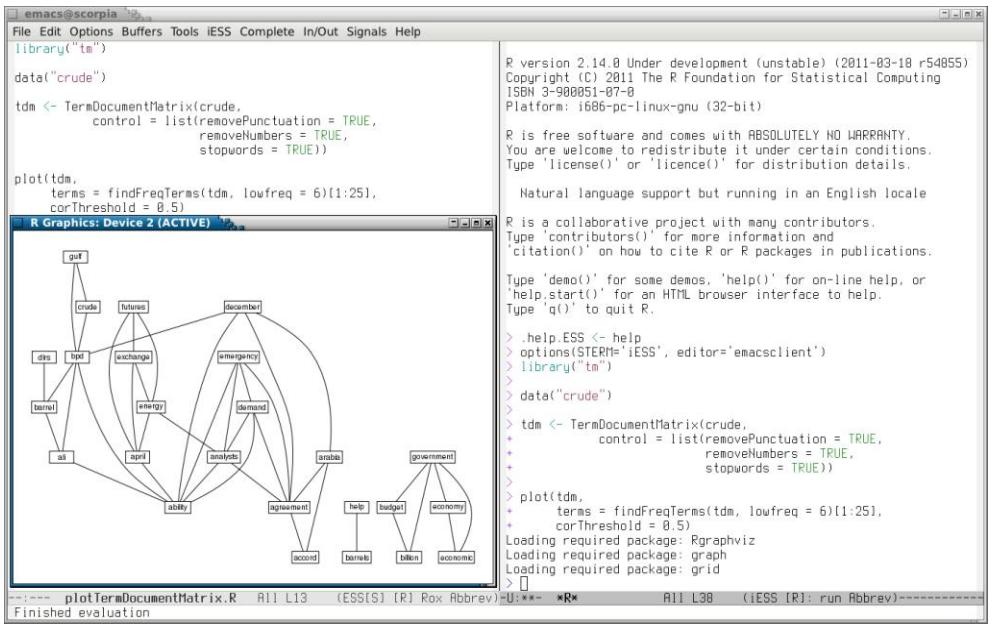

Рисунок 7. Окно программы tm (Text Mining Infrastructure in R)

8. TAMS Analyzer (рис. 8) [25].

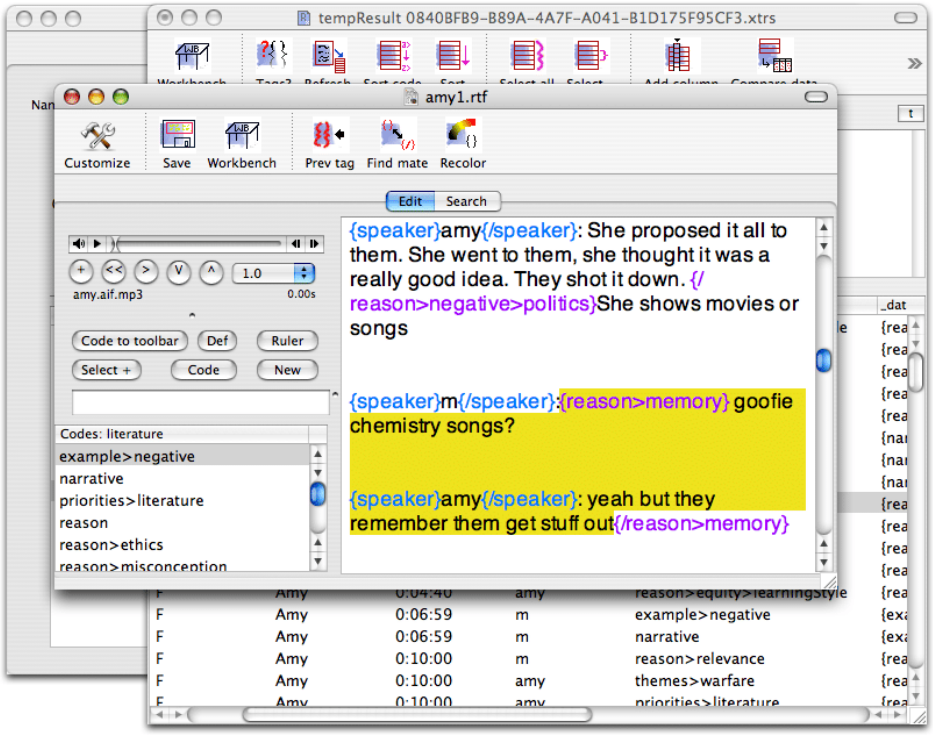

Рисунок 8. Окно программы TAMS Analyzer

TAMS Analyzer для OS X Macintosh является решением для распознания тем в текстах, таких как веб-страницы, интервью, заметки. Он был разработан для разговорного и этнографического поиска. TAMS Analyzer - программа, которая работает с TAMS, чтобы присвоить этнографические коды частям текста, выбирая соответствующий фрагмент и дважды щелкая по имени кода списка. TAMS определяет в тексте участников

разговора, а также главные темы разговора. Он позволяет извлекать, анализировать и сохранять обработанную информацию.

9. STATISTICA Text Miner (рис. 9) [26]

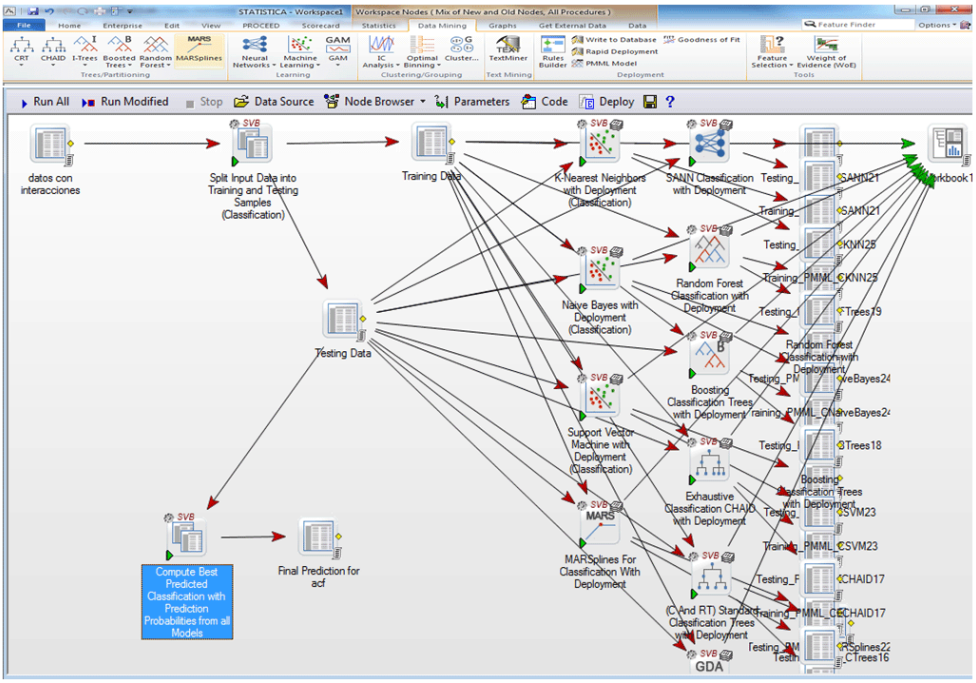

Рисунок 9. Окно программы STATISTICA Text Miner

STATISTICA Text Miner – это дополнительная возможность STATISTICA Data Miner, идеально подходящая для того чтобы переводить неструктурированный текст в легко-читаемую, ценную информацию, пригодную для принятия «золотых» решений. Большинство пользователей, знакомых с системами Text Mining, хорошо знают о том, что, как правило, реальные «необработанные» данные являются не всегда пригодными для восприятия и последующего анализа.

STATISTICA Text Miner позволяет выбрать из потока информации необходимые данные и структурировать их. STATISTICA Text Miner интегрирована в приложение STATISTICA Data Miner.

Настоящее приложение использует много-потоковые компьютерные технологии для достижения максимальной производительности передовых многопроцессорных серверных систем.

Продемонстрируем работу программ Text Mining на примере KH Coder. Создадим карту связей текста.

Откроем программу и первым делом зайдём в Setting, в разделе Word Extraction переключим на Lemmatization with “Stanford POS Tagger”, здесь нажмём на config и переместим в открывшееся окно файл с ключевыми словами (khcoder/tutorial_en/stopwords_sample_en.txt), применим все изменения (рис.10).

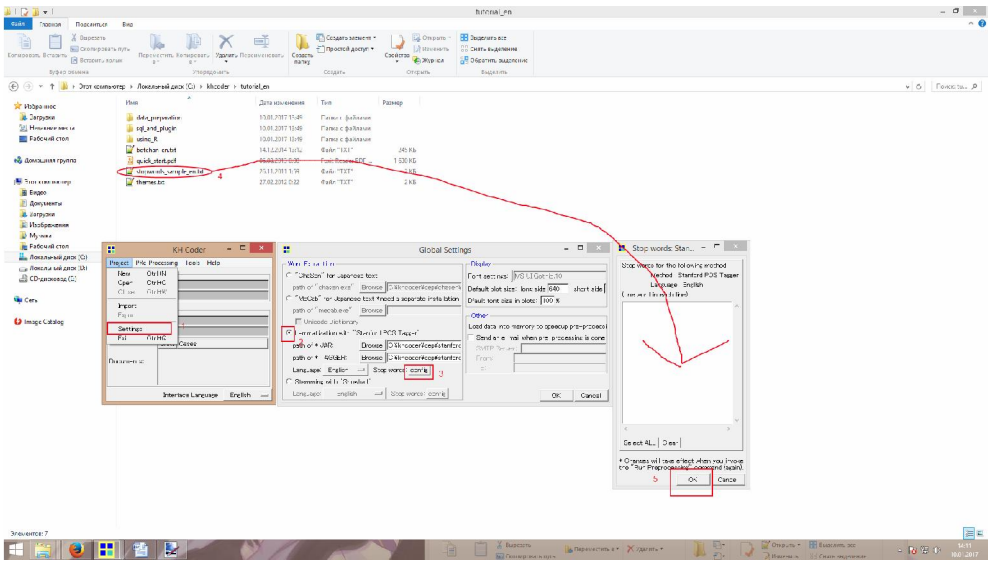

Рисунок 10. Добавление ключевых слов в KH Coder

Создадим новый проект, нажав Project-New project и выберем файл botchan_en.txt (khcoder/tutorial_en/) (рис. 11).

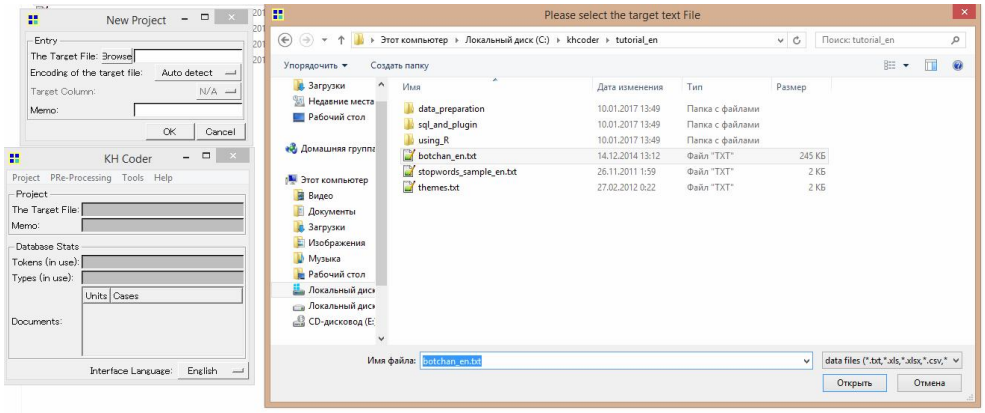

Рисунок 11. Создание проекта и добавление файла в KH Coder

Запустим преобработку текста, нажав Pre-Processing-Run Preprocessing, и дождёмся окончания процесса (рис. 12).

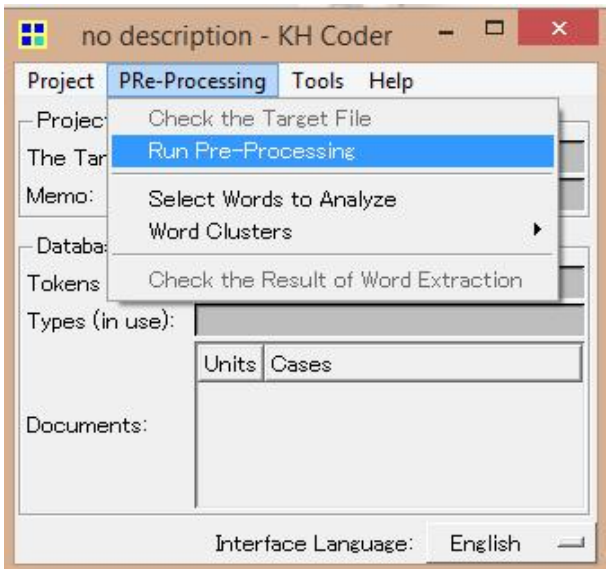

Рисунок 12. Предобработка текста в KH Coder

Осталось только получить карту связей, для этого нажимаем Tools-Words-Co-Occurrence-Network и в открывшемся окне нажимаем Ok (рис. 13).

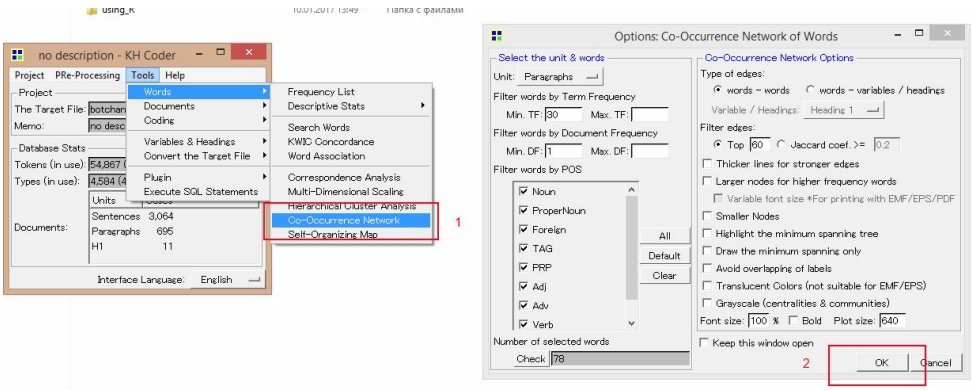

Рисунок 13. Создание карты связей текста в KH Coder

По окончанию процесса была создана карта связей нашего текста (рис. 14).

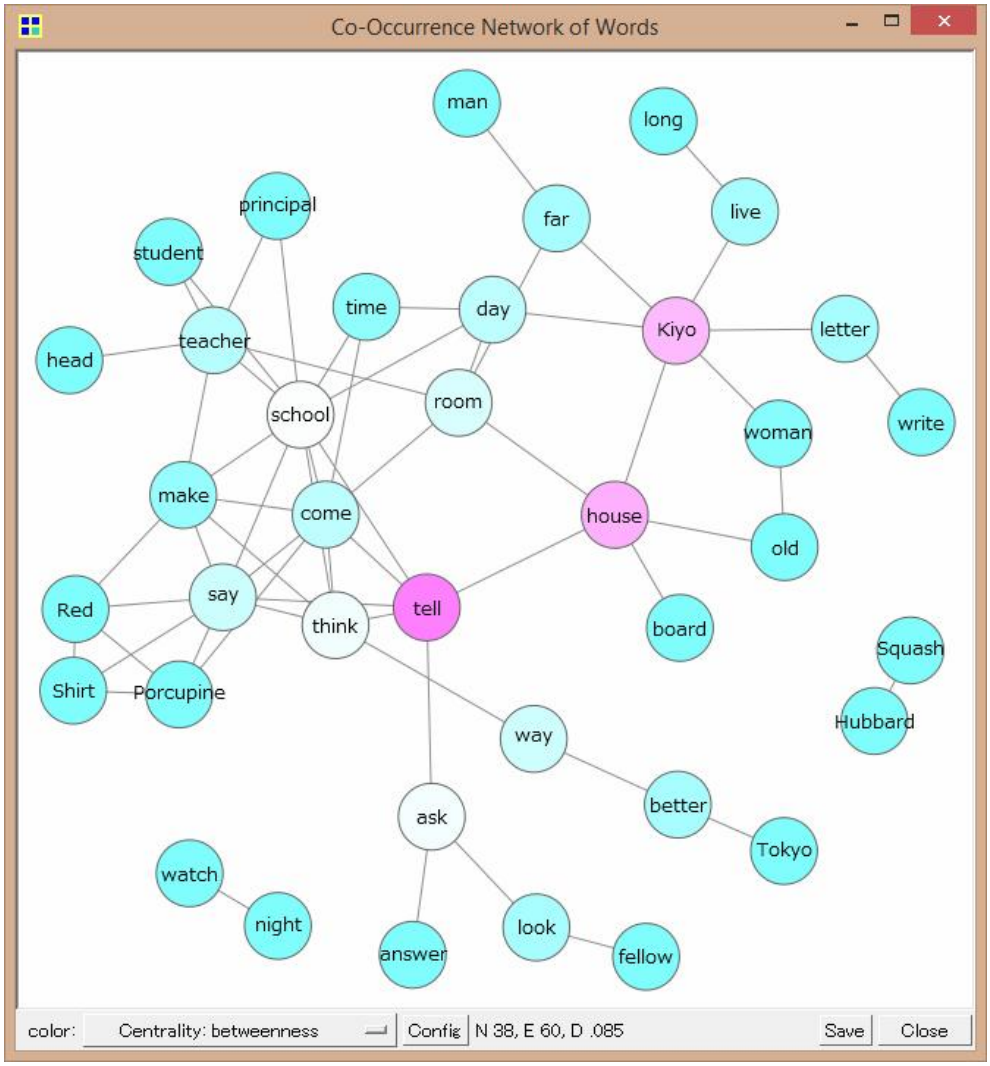

Рисунок 14. Карта связей текста в KH Coder

Таким образом, при проведении обзора было выяснено, что существуют множество специальных программы для обработки большого объема текстовой информации – программы Text Mining. Они отличаются способами работы с данными и сложностью механизмов их обработки, а также способом представления данных.

Данная статья будет полезна для работников IT-сферы, в частности для работающих с большими объёмами данных, так как эти программы смогут существенно упростить им работу.

Библиографический список

1. Рост объема информации - реалии цифровой вселенной // Технологии и средства связи. 2013. №1. С. 24.
2. Пескова О. В. Алгоритмы классификации полнотекстовых документов // Автоматическая обработка текстов на естественном языке и компьютерная лингвистика. М.: МИЭМ (Московский государственный институт электроники и математики), 2011. С. 170-212.
3. Алгулиев Р.М., Назирова С.А. Анализ текстовых спам-сообщений с помощью методов Text Mining // Информационные технологии. 2011. №S9. С. 2-8.
4. Пиотровская К. Р. Текст-майнинг: перспективы развития // Известия РГПУ им. А.И. Герцена. 2014. №168 С.128-134.
5. Абакумов О. Использование кластеризации в Text Mining. // Новые информационные технологии в автоматизированных системах. 2010. № 13. С. 128-129.
6. Кутукова Е.С. Технология Text Mining. // Научные труды SWorld. 2013. Т.30. № 4. С. 33-36.
7. Бердник Е.А. Технологии Text Mining как элемент информационной культуры студентов социологов // В сборнике: Образование и общество Всероссийская социологическая конференция к 20-летию Российского общества социологов. 2009. С. 318.
8. Кожевникова Г.П., Голикова А.В. Интеллектуализация поисковых систем на основе технологии Text Mining. // В сборнике: Инновационный путь развития РФ как важнейшее условие преодоления мирового финансово-экономического кризиса Материалы международной научно-практической конференции. Заседания секций в 2-х томах. 2009. С. 331-334.
9. Мясников Н.С. Обзор методов Text Mining. // В сборнике: Информационные технологии и системы Труды Четвертой Международной научной конференции. Ответственные редакторы: Ю.С. Попков, А.В. Мельников. 2015. С. 92-93.
10. Алексеев А.А., Катасёв А.С., Кириллов А.Е., Кирпичников А.П. Классификация текстовых документов на основе технологии Text Mining // Вестник Казанского технологического университета. 2016. Т. 19. № 18. С. 116-119.

11. Kumar B.S., Ravi V.. A survey of the applications of Text Mining in financial domain // Knowledge-Based Systems. 2016. T.114. C. 128-147.
12. Perovšek M., Kranjc J., Erjavec T., Cestnik B., Lavrač N. TextFlows: A visual programming platform for text mining and natural language processing // ArticleScience of Computer Programming. 2016. T. 121. C. 128-152.
13. Топ-5 инструментов для Text Mining [Электронный ресурс] URL: http://datareview.info/article/top-5-instrumentov-dlya-text-mining/ (дата обращения: 07.01.2017).
14. Top 27 Free Software for Text Analysis, Text Mining, Text Analytics [Электронный ресурс] URL: http://www.predictiveanalyticstoday.com/top-free-software-for-text-analysis-text-mining-text-analytics/ (дата обращения: 07.01.2017).
15. Официальный сайт GATE [Электронный ресурс] URL: https://gate.ac.uk/ (дата обращения: 07.01.2017).
16. Hamish C., Maynard D., Bontcheva K., et al. Developing Language Processing Components with GATE Version 7 (a User Guide) (англ.). The University of Sheeld, 2013.
17. Официальный сайт KNIME [Электронный ресурс] URL: https://www.knime.org/ (дата обращения: 07.01.2017).
18. Снайдер С. Гибкий анализ данных // developerWorks Россия. [Электронный ресурс] URL: https://www.ibm.com/ developerworks/ru/library/d-agile-data-analysis/ (дата обращения: 07.01.2017).
19. Официальный сайт RapidMiner [Электронный ресурс] URL: https://rapidminer.com/ (дата обращения: 07.01.2017).
20. Ingo Mierswa, Michael Wurst, Ralf Klinkenberg, Martin Scholz, Timm Euler. YALE: Rapid Prototyping for Complex Data Mining Tasks // Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-06), 2006.
21. Официальный сайт Sisense [Электронный ресурс] URL: https://www.sisense.com/ (дата обращения: 07.01.2017).
22. Официальный сайт Carrot2 [Электронный ресурс] URL: http://project.carrot2.org/ (дата обращения: 07.01.2017).
23. Официальный сайт KH Coder [Электронный ресурс] URL: http://www.predictiveanalyticstoday.com/kh-coder/ (дата обращения: 07.01.2017).
24. Официальный сайт tm (Text Mining Infrastructure in R) [Электронный ресурс] URL: http://tm.r-forge.r-project.org/ (дата обращения: 07.01.2017).
25. Официальный сайт TAMS Analyzer [Электронный ресурс] URL: http://www.predictiveanalyticstoday.com/tams/ (дата обращения: 07.01.2017).
26. Официальный сайт STATISTICA Text Miner [Электронный ресурс] URL: https://software.dell.com/products/statistica/ (дата обращения: 07.01.2017).