Елизавета Терещенко
Мария Равлик
НИУ-ВШЭ, СПб, 2011

Инструменты социологического анализа Интернет-текстов и Интернет-сетей: обзор современного программного обеспечения

В этом обзоре описываются результаты анализа различного рода программного обеспечения (ПО), финальной целью применения которого является анализ текста с целью выделения присутствующей в нем тематики. Помимо различных прикладных характеристик ПО (таких, как его доступность, возможность работать с русским языком и обрабатывать большие массивы данных, вычислительная сложность, быстродействие) особое внимание уделяется оценке релевантности ПО решению стоящих перед проектом задач.

Материалы данного обзора являются результатами совместной работы членов проекта НИУ-ВШЭ по теме «Разработка методологии сетевого и семантического анализа блогов для социологических задач». В проекте изучается, каким образом в блогах отражается различная тематика (например, тематика Ислама) и как на основании этой тематики блоги можно классифицировать. В таком случае объектом анализа выступают массивы текстов, размер которых значителен. Размеры выборки определяют особенности работы с ней: исследователю необходимо собрать целый набор программ, которые позволят автоматизировать процессы сбора и анализа текстов.

Для решения поставленных в проекте задач потребовался поиск различного рода программного обеспечения, которое можно классифицировать следующим образом:

1. технические программы (куда мы относим программы закачки и хранения и программы по предварительной подготовке текстов к анализу);
2. программы для семантического анализа;
3. программы для сетевого анализа.

Понятно, что реально существует гораздо большее количество групп ПО, предназначенных для работы с текстом. К ним могут быть отнесены программы для поиска текстов и data retrieval, программы для извлечения структуры текста (типа семантических решеток и сетей) и смысла текста (автоматическое аннотирование) и пр. Рассмотрим подробнее, какие программы из изученных нами наиболее полно отвечают поставленным задачам.

1. Программы закачки и хранения

Первым этапом работы над проектом является составление базы текстов. В связи с тем, что требовалось собрать значительный объем выборки, причем со строгой структурой (текст, имя пользователя, ссылки и т.п.), возникла необходимость автоматизировать этап сбора данных и найти программу закачки текстов. В соответствии поставленной задачей, к данному типу программного обеспечения предъявлялись такие требования, как работа с большими массивами текстов, возможность скачивать тексты по определенной структуре и хранение полученных результатов в упорядоченном виде. Как показал анализ ПО, ни одна из опробованных программ по закачиванию текстов не соответствует предъявляемым требованиям. Исключение составила программа **The VOSON System**[1]. Она работает на ОС Linux, базе данных MySQL, имеет PHP/javascript веб-интерфейс, веб-краулер, написанный на языке Perl, набор встроенных инструментов для анализа текста, сетевого анализа (каковая функция будет рассмотрена ниже) и последующей визуализации результатов.

Однако в силу непрозрачности работы алгоритмов The VOSON System по сбору и обработке текстовой информации пришлось отказаться от ее использования и принять решение о разработке собственного программного обеспечения. Программа получила название **«Koltran BlogMiner»**. Несомненным достоинством программы является то, что она напрямую предназначена для работы в подобного типа исследованиях. Программа позволяет работать со значительными объемами данных и структурировать их по таким параметрам, как:

– имена (ники) блоггеров (авторов постов), сопоставленные уникальным идентификационным номерам (ID), формируемым при закачке;

– тексты постов в txt формате, сопоставленные ID автора;

– дата и время каждого поста, сопоставленная посту;

– URL адрес каждого поста, сопоставленный посту;

– тексты комменариев, сопоставленные постам;

– дата и время каждого комментария, сопоставленные комментарию;

– имя (ник) автора комментария, сопоставленное комментарию.

Koltran BlogMiner является уникальным, полностью самостоятельным продуктом, реализованным на Delphi 7 и синхронизированным со стандартной оболочкой для создания баз данных и работы с ними, называемой SQL server (в данном случае – Microsoft SQL server).

[1] voson.anu.edu.au

В базе данных реализован полнотекстовый поиск, который, в отличие от поисковой системы типа Yandex, выдает все данные (а не только первую тысячу), осуществляет поиск по очень длинному списку как полных слов, так и лемм (корней слов. База также позволяет делать случайные и шаговые выборки, выборки по дате и др., т.е. приспособлена для социологических задач.

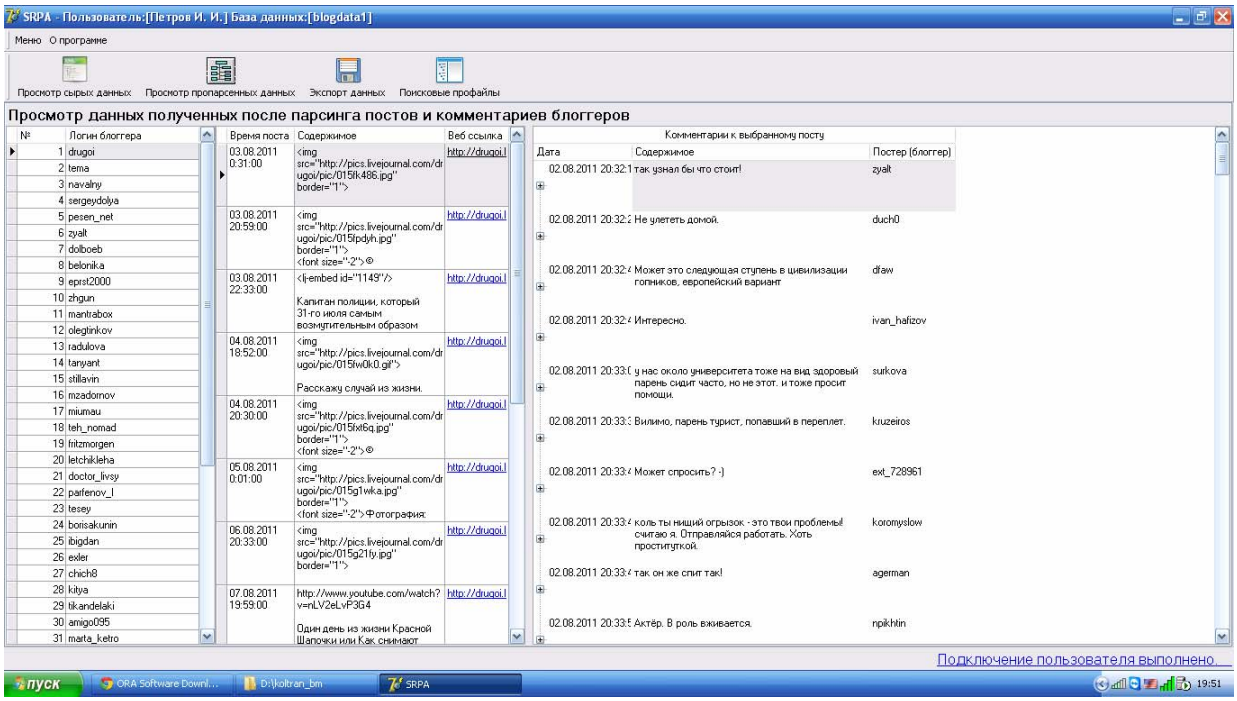

Рисунок 1 – Интерфейс программы Koltran BlogMiner.

На рисунке 1 представлен интерфейс программы Koltran BlogMiner. В левой части ники блоггеров; в средней части – список постов выделенного блоггера (drugoi); в правой части – список комментариев к выделенному посту.

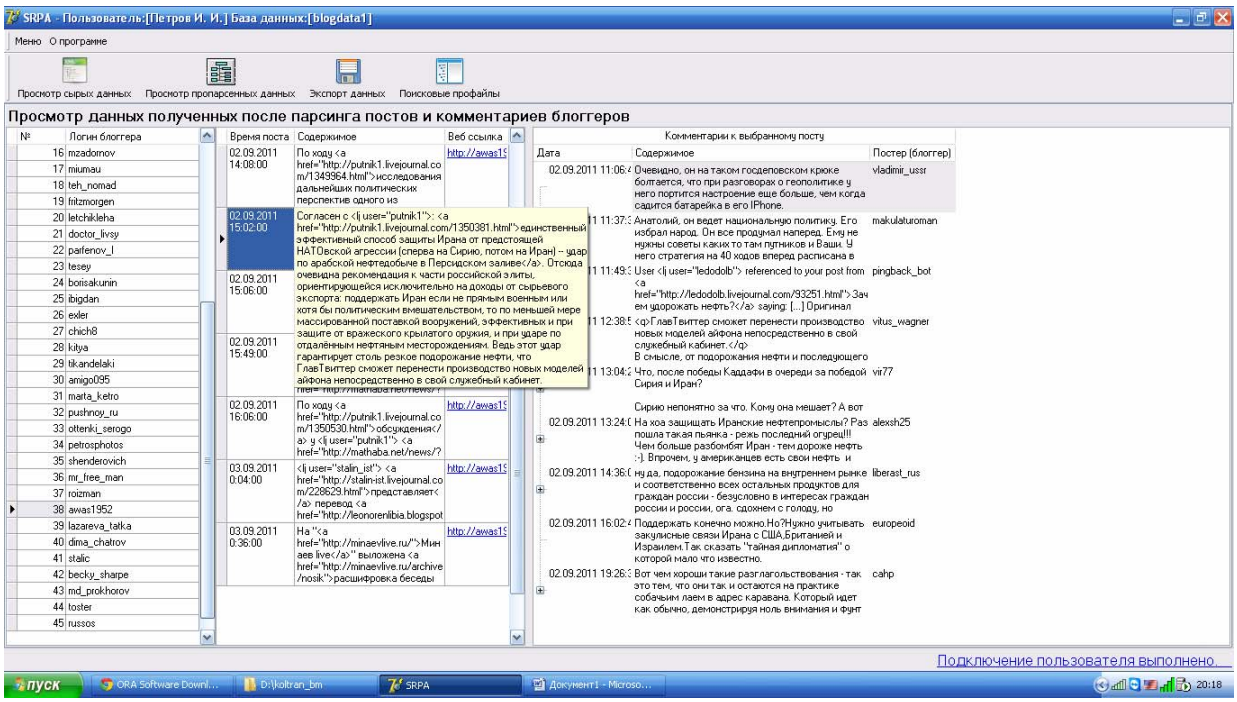

Рисунок 2 – Вид программы в режиме просмотра текста поста

В левой части рисунка 2 – список ников блоггеров; в средней части – всплывающее окно с текстов выделенного поста выделенного блоггера (awas1952); в правой части – тексты комментариев к данному посту.

Следующим этапом в работе по анализу текстов является **преподготовка текстов**. Она включает в себя лемматизацию текстов, удаление несодержательных слов, и соответственно, составление списка подобных слов (стоп-слов). Также для кластерного анализа текстов их необходимо преобразовать в матрицу, т.е. в векторную форму.

Для лемматизации текста была выбрана программа **mystem**[2]. Ее достоинством является то, что она производит морфологический анализ текста на русском языке. Программа разработана компанией Яндекс и, что важно, распространяется в свободном доступе. Функционально mystem представляет собой разновидность стеммера, отсекает от слова окончание и суффиксы, чтобы оставшаяся часть, называемая *stem*, была одинаковой для всех грамматических форм слова. Программа позволяет исследователю видеть несколько вариантов форм слова (если они возможны) и выбирать наиболее точную.

Отметим, что для работы с программой mystem Л. Пивоваровой была написана дополнительная программа **Stem**: использован язык программирования Perl, базовыми функциями программы является запуск mystem, а затем – чистка полученных результатов.

[2] http://company.yandex.ru/technology/mystem/

Для преобразования документа в векторную форму была использована программа **doc2mat**[3], которая разработана George Karypis специально для его программы gCluto. Эта программа конвертирует документы в векторный формат, который используется в программе кластеризации gCluto. Именно для этой программы необходимо, чтобы каждый документ в файле был расположен в одну строку без переносов, тогда в выходном файле число документов будет соответствовать числу строк в исходном файле. Построенная матрица в формате .mat является исходным файлом для программы gCluto.

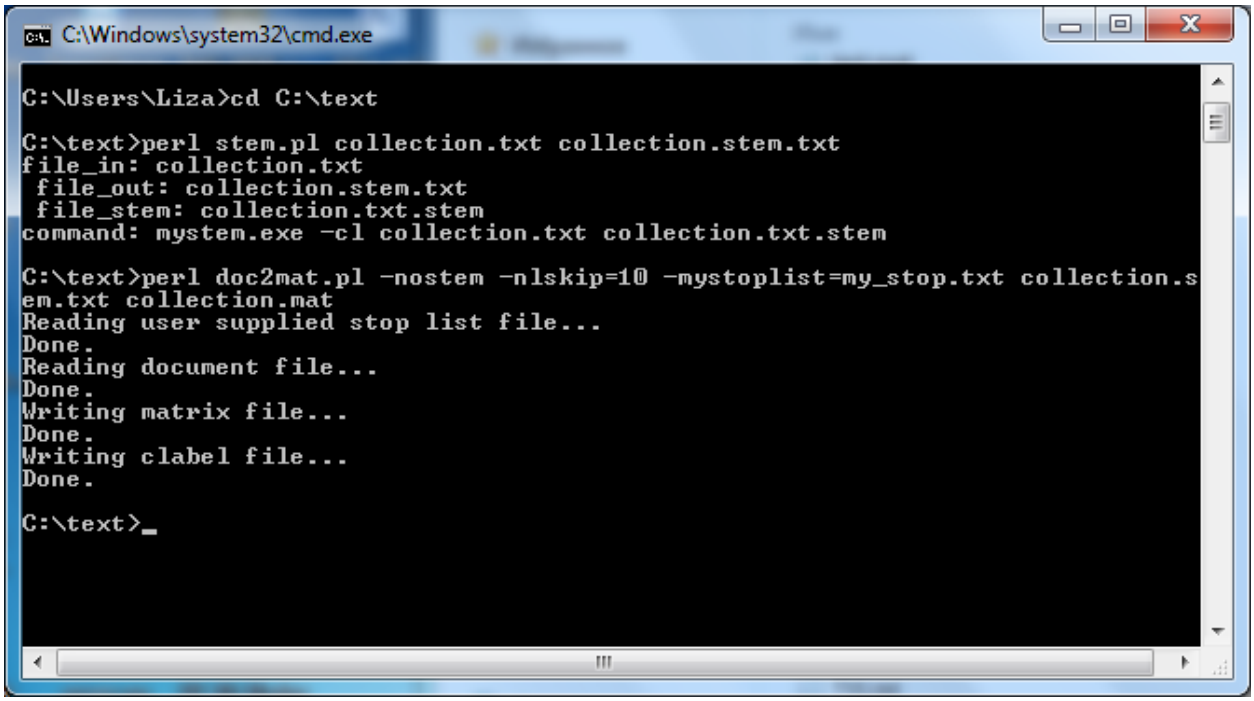

Рисунок 3 – Работа с программами mystem, stem и doc2mat.

Все указанные программы написании на языке программирования Perl, поэтому работа с ними осуществляется через командную строку, внешний вид которой и заданные команды представлены на рисунке 3. Особенностью данной работы является то, что последовательно запускаются три программы – mystem, Stem и doc2mat. Использование такой длинной цепочки для пре-процессинга, безусловно, является сложным и затратным по времени, но более эффективного решения на данный момент не существует. По факту в этом случае пре-процессинг включает такие этапы:

1.запуск программы stem, написанной Л. Пивоваровой. Эта программа работает в 2 этапа: запускает лемматизатор mystem, а затем удаляет из полученных результатов все лишнее. На выходе мы получаем текстовый файл, где все слова приведены к лемме (на рисунке исходный файл collection.txt преобразован в collection.stem.txt)

[3] http://glaros.dtc.umn.edu/gkhome/files/fs/sw/cluto/doc2mat.html

2.запуск программы doc2mat, которая работает уже с файлом collection.stem.txt Для нее прописаны следующие параметры:

– nostem, что означает, что лемматизацию проводить не нужно. Это обусловлено тем, что ее мы провели при помощи программы mystem

– nlskip=10. Этот параметр применяется дли игнорирования возможных идентификаторов документов, которые могут быть в начала строк.

– mystoplist. Для пре-процессинга мы составили собственный список стоп-слов, которые исключаются из текста. Список стоп-слов был составлен на основе списка, предлагаемого Википедией[4] – он включает в себя все существующие в русском языке союзы, местоимения и предлоги.

Doc2mat создает матрицу текста, которая является основой для кластеризации в таких программах как, например, gCluto. Важно отметить, что предварительно пришлось вносить изменения в исходный код этой программы для того, чтобы она могла работать с кириллическим текстом.

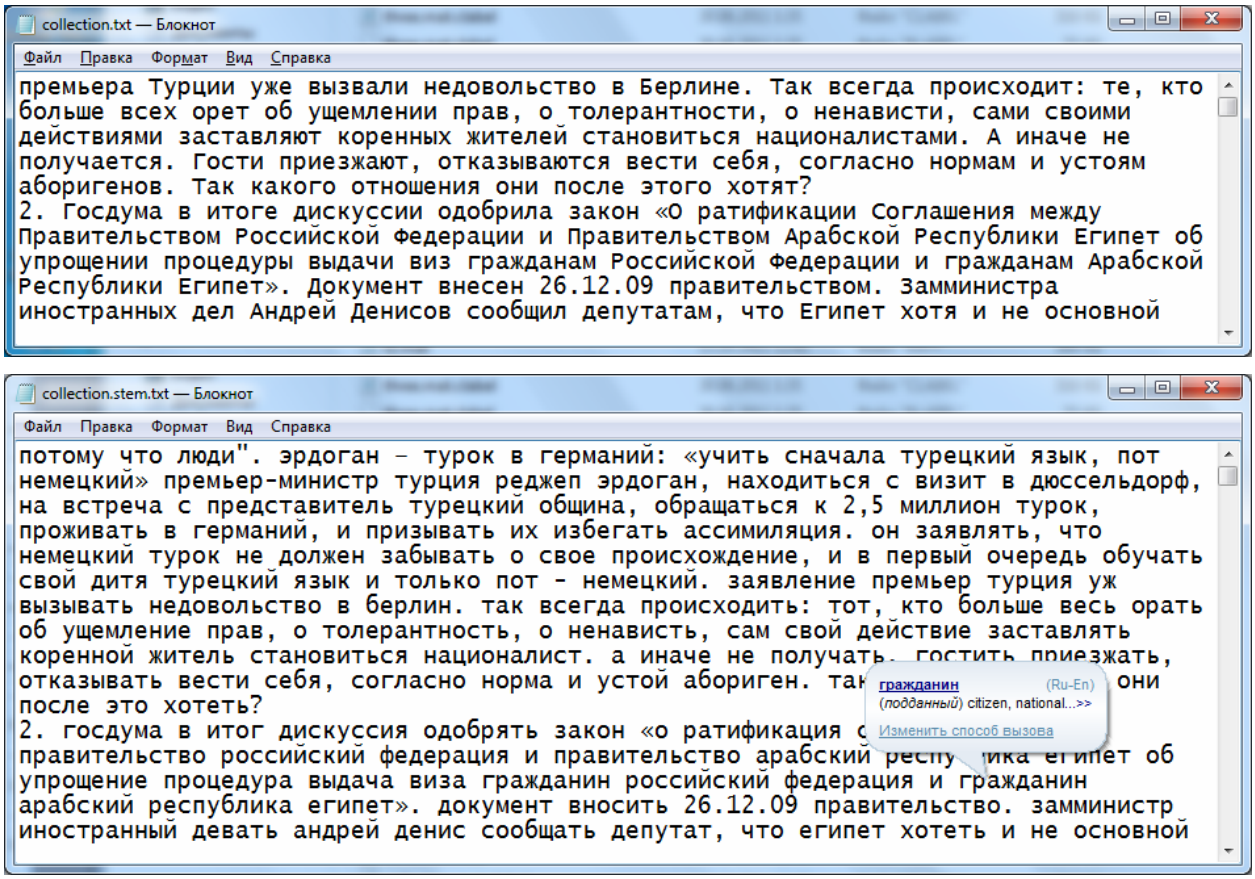

Рисунок 4 – Текст до и после лемматизации в программе mystem

[4] http://ru.wiktionary.org/wiki/Категория:Русский_язык

2. Программы для семантического анализа

Для проведения семантического анализа мы рассматривали программы кластеризации и программы моделирования тем. В соответствии с особенностями выборки для выбора программного обеспечения были составлены критерии, которые представлены в таблице 2. Изучение ПО происходило в два этапа. На первом этапе основным источником информации о ПО служили официальные мануалы к программам и отзывы пользователей. На втором этапе проводилось предварительное тестирование программ на коллекции текстов.

Таблица 2 – Критерии оценки программного обеспечения

Характеристики на 1 этапе (предварительный выбор)	**Характеристики на 2 этапе (апробация программного обеспечения)**
название программного обеспечения, версия программы	быстродействие
язык, с которым работает программа	требования к входным данным
разработчики	легкость установки, настройки и освоения
Сайт разработчиков или ссылка на программное обеспечение	Способность работать с большими коллекциями
Границы доступа к программному обеспечению	Способность самостоятельно определять число кластеров или тем
стоимость программного обеспечения	Вычислительная сложность
требования к системе	способность работать с multidimentionality
Используемый алгоритм	Устойчивость к шумам
Функции программного обеспечения	способность к ко-кластеризации текстов и слов
Примеры использования программного обеспечения	Способность работать с меняющейся коллекцией

Всего было рассмотрено около 40 программ для кластерного анализа. Основной проблемой в большинстве случае является коммерческий характер большинства программ, и, как следствие, закрытые алгоритмы их работы. Это не позволяет предварительно оценить, могут ли они быть использованы для целей нашего исследования и проверить качество проводимой ими кластеризации. Помимо этого, программы кластеризации по принципу работы делятся на работающие в он-лайн и офф-лайн режимах. Соответственно, для наших целей подходят только работающие в оффлайн-режиме, так как они могут обрабатывать значительные объемы информации.

После предварительного анализа функций и отзывов на программное обеспечение было отобрано шесть потенциально пригодных программ.

Таблица 3 – Характеристика программного обеспечения для кластерного анализа

ПО	**Формат**	**Русский**	**Алгоритм КА**	**Год**	**Свободный**	**Особенности**

	входных данных	язык			доступ	
gCLUTO[5] - Graphical Clustering Toolkit	Векторная форма	+\-	17 видов, включая плоскую и нечеткую кластеризации		+	Дополнительные инструменты для визуализации результатов
Carrot2[6]	Интернет-данные	+	Иерархическая и k-means		+	Поисковая программа, работает в он-лайн режиме
HAMLET[7]	Векторная форма	+	иерархический и неиерархический КА		+	Широкий функционал
TextAnalyst[8]	.txt (ANSI, DOS), .rtf	+	не известно	2006	_	Объем выборки не ограничен
PolyAnalyst[9]	не известно	+	не известно	2007	_	

Практически все эти программы обладают широким функционалом, включающим не только кластеризация, но и классификацию, реферирование и прочее.

1. gCLUTO – Graphical Clustering Toolkit – кроссплатформенное графическое приложение. Преимуществом этой программы является открытость алгоритмов (в программе используются такие методы кластеризации, как иерархический агломеративный, и метод кластеризации, основанный на делении). Также достоинством этой программы является ее направленность на визуализацию результатов кластеризации и отсутствие ограничений на объемы входящих данных. Основной проблемой в использовании этой программы является сложность в освоении, т.к. она работает с файлами в векторной форме, и для их преобразования нужно использовать дополнительную программу, которая работает на языке программирования Perl.

2. Carrot2 – представляет собой поисковую программу, которая выдает найденные результаты в виде кластеров. Эта программа использует метод иерархической кластеризации и k-means. Отметим, что по факту любые поисковые программы не соответствуют нашим требованиям, т.к. работают только он-лайн непосредственно с информацией из всей сети Интернет. Данная программа была включена в список рассматриваемых из-за того, что ее разработчики утверждают, что ее она пригодна и для кластеризации любых текстовых массивов (т.е. уже готовых выборок текстов, преобразованных и подготовленных для кластеризации). Также есть платные, расширенные версии программы, которые могут быть использованы в наших целях, например, Lingo3G.

[5] George Karypis, http://glaros.dtc.umn.edu/gkhome
[6] Osiński S., Weiss D., http://project.carrot2.org
[7] ESRC (National Centre for Research Methods) и Bruno Hopp GESIS (Leibniz-Institut für Sozialwissenschaften)
[8] Microsystems, Ltd, http://www.analyst.ru
[9] Megaputer Intelligence, http://megaputer.ru

3. HAMLET – эту программу отличает широкий спектр решаемых задач. Она подходит для решения задач, включающих семантический и символический анализ текстов. Выполняет контент-анализ, кластерный (включающий иерархический и неиерархический кластерный анализ), анализ соответствий и многомерное шкалирование. Для ознакомления с программой есть бесплатный триал.

4. TextAnalyst – эта программа разработана в России и, соответственно, предназначена для работы с русскими текстами. Ее отличает широкий функционал, включая и кластеризацию текстов. Что очень важно, разработчики утверждают, что максимальный объем анализируемой подборки не ограничен и зависит от объема ресурсов компьютера и настройки TextAnalyst. Для ознакомления с программой есть бесплатный триал.

5. PolyAnalyst – программа работает с русским языком, выполняет следующие функции: категоризация, кластеризация, прогнозирование, анализ связей, нахождение ключевых слов и поиск смысла, выявление закономерностей, нахождение аномалий. Недостатком программы является отсутствие пробной версии, что повышает риски при решении о ее покупке.

Из всех вышеперечисленных программ для последующей работы была выбрана программа gCluto. Основная версия программы – CLUTO, запускается из командной строки и работает с языком программирования perl. Это пакет программ для кластеризации низко- и высоко- размерных наборов данных, предназначенный для анализа характеристик различных кластеров. CLUTO написан на языке C++, поддерживает работу на платформах Microsoft и Linux. Для удобства пользования была выбрана версия gCLUTO, т.к. она имеет пользовательский интерфейс и в нее встроены дополнительные инструменты для визуализации полученных кластеров. Также к достоинствам gCLUTO относятся:

– программное обеспечение находится в свободном доступе;

– алгоритмы кластеризации являются открытыми, имеется значительное количество публикаций с их описанием;

– программа опирается на библиотеку CLUTO;

– офф-лайн кластеризацию относительно больших объемов (до 10^4) текстов

– используется 17 алгоритмов, включая плоскую и иерархическую кластеризацию и graph-based алгоритмы;

– программа позволяет проводить кластеризацию текстов, формирует дерево кластеров, отчет по созданным кластерам и строит карту взаиморасположения кластеров.

Проблемой программы является отсутствие способов оценки качества получившихся результатов, а также нестабильная работа с кириллицей.

Таблица 4 – Используемые в gCLUTO алгоритмы кластеризации

Алгоритм	**Описание**
Repeated bisection	Желаемое количество кластеров k строится путем k-1 повторений. Матрица сначала кластеризуется на две группы, потом выбирается 1 группа и делится дальше. Этот процесс не останавливается, пока не будет найдено заданное количество кластеров. На каждом шаге кластер делится пополам. Авторы отмечают, что в таком случае целевая функция кластеризации будет оптимизирована в каждой из групп, но в целом оптимизирована не будет. Выбранные кластеры контролируются параметром –cstype.
Direct	Все кластеры определяются одновременно. Этот метод работает медленнее, чем Repetead bisection, но, с точки зрения качества полученных кластеров, дает лучший результат. Его лучше использовать на небольших выборках.
Agglomerative	Используется метод агломерации, который предполагает оптимизацию целевой функции деления (которая выбирается с использованием параметра the -crfun).
Graph	Используются графы ближайших соседей (каждый объект становится вершиной и каждый объект соединяется с наиболее похожими объектами), затем граф разделяется на кластеры.

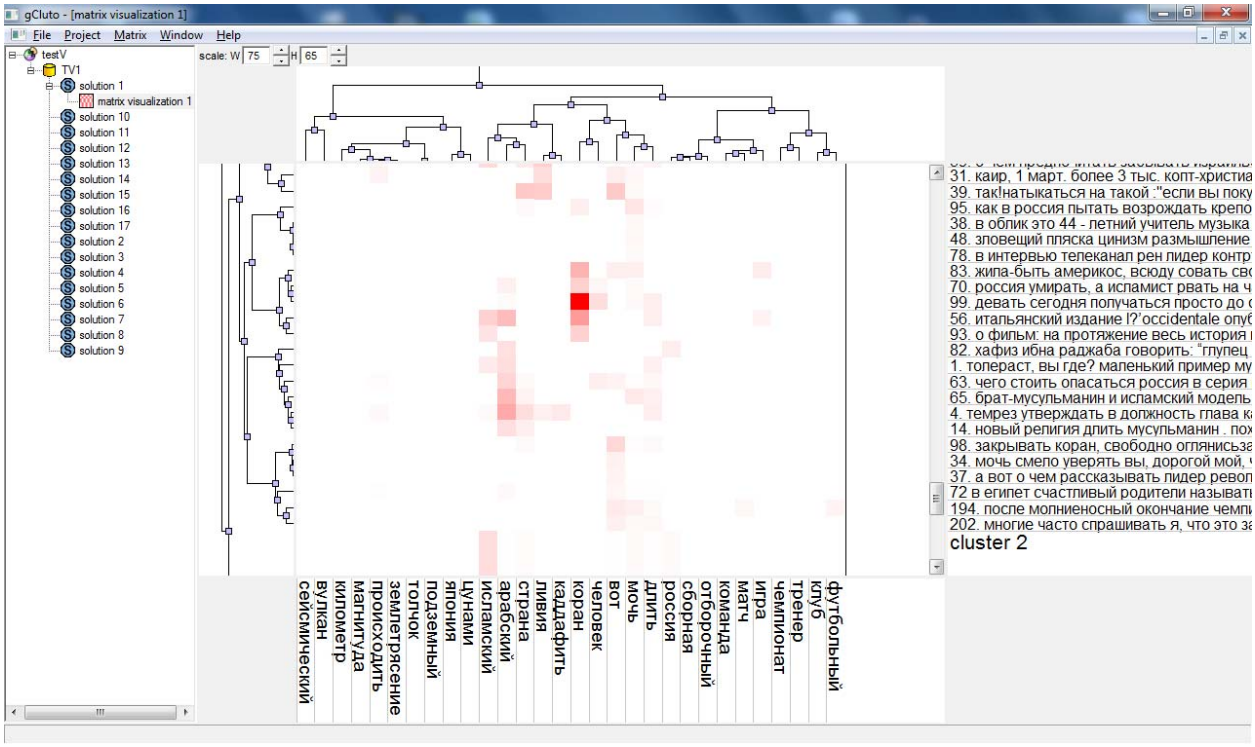

Рисунок 4 - Результат анализа в gCLUTO (иерархическая кластеризация).

В качестве альтернативы кластерному анализу были изучены программы для моделирования тем (topic modeling). Наиболее распространенным алгоритмом, используемым в такого рода программах, является **Latent Dirichlet Allocation (LDA).** Этот алгоритм позволяет извлекать скрытые, латентные темы из текстовых данных. Он полезен для решения задач классификации и совместного фильтрования (collaborative filtering). Каждый документ воспринимается как смешение разных тем. В целом, данный алгоритм похож на вероятностный латентный семантический анализ (pLSA), но дает более осмысленное разделение объектов на группы[10].

Наиболее удачной рограммой, предназначенной для семантического анализа и работающей на данном алгоритме, является **Stanford Topic Modeling Toolbox[11].** Разработчиками являются Daniel Ramage and Evan Rosen (Stanford NLP group). Язык, на котором написана программа – Scala, используемая технология: Topic Modeling.

Программа позволяет импортировать и обрабатывать результаты в Excel, делать саммари текста, выбирать число тем при анализе текста.

Плюсы ПО:

- бесплатное использование;
- разработано специально для социологов и других исследователей, которые не имеют навыков в написании скриптов;
- открытый исходный код;
- возможность обрабатывать результаты в Excel;
- работает с любым набором неструктурированных текстов.

Таким образом, Stanford topic modeling toolbox является ПО с простым интерфейсом, эффективность работы с которым подтверждена многочисленныыми публикациями об используемом алгоритме. Тестирование показало способность ПО работать с данными 10^4 текстов. Недостатком ПО является сложность определения ярлыков для тем.

Также для текстового анализа и выделения присутствующих в нем тем может быть использован программный пакет R, конкретно его приложение Text Mining. Приложение позволяет значительно сократить временные затраты, связанные с пре-процессингом, т.к. лемматизация и удаление стоп-слов происходят одновременно. Однако тестирование R показало его существенные ограничения по приему текстовых данных.

Также к группе ПО для семантического анализа можно отнести **программы sentiment analysis**, которые позволяют оценить эмоциональную окраску текстов. В нашем

[10] http://en.wikipedia.org/wiki/Latent_Dirichlet_allocation
[11] nlp.**stanford**.edu/software/tmt/

проекте в качестве основного для решения данного типа задач рассматривается программное обеспечение Sentistrength.[12]

Разработчик ПО – Mike Thelwall. Назначение программы – анализ настроений, измерение силы положительных и отрицательных эмоций в коротких текстах. Базовым языком для SentiStrength является английский, но могут быть произведены дополнительные настройки для работы с другими языками. Для академических целей программа является бесплатной.

Особенностью работы программы является то, что она предназначена для анализа коротких текстов (и как отмечает разработчик – низкого качества). При этом самой программой длина текста игнорируется, но это приводит к тому, что в результате получаются высокие значения эмоций с обоими полюсами, поэтому подбор текстов для анализа и выбор корректных настроек предопределяют результаты и зависят от мнения исследователя. При этом в качестве подготовки к анализу необходимо составить корпус текстов из не менее тысячи вручную аннотированных текстов и дополнить словарь, которым пользуется программа (EmotionLookupTable.txt) новыми эмоционально-окрашенными терминами, релевантными исследуемому тексту.

3. программы для сетевого анализа

В ходе обзора программ, осуществляющих семантический анализ, было рассмотрено более 80 ПО. Основными требованиями к нему были возможность закачивать данные из Интернета, работать с большими объемами данных и визуализировать результаты. Все программы по сетевому анализу можно отнести к следующим категориям:

- программы-визуализаторы;
- стандартные универсальные пакеты;
- программы, по факту представляющие собой отдельные куски кодов.

Таблица 4 – Обзор программ для сетевого анализа

ПО	Формат входных данных	Год	Свободный доступ	Особенности
Программы-визуализаторы				
SocNetV[13]	.xml (GraphML), .net (pajek), .dot (GraphViz), .sm/.net (Sociomatrix), .net (UCINET)	2010	+	предлагает разные способы графического представления данных
Универсальные пакеты				
Pajek[14]	.net, .paj,	2011	+	принимает данные в почти 10^{10} узлов, однако

[12] http://sentistrength.wlv.ac.uk/
[13] ***socnetv***.*sourceforge.net*

	.dat(UCINET), .ged, .bs, .mac, .mol			он не содержит алгоритмов выявления сообществ
ORA[15]	DyNetML, .csv	2011	+	анализ организационных сетей, в т.ч. динамических, с большим количеством сопровождающих данных, поэтому в целом не рассчитана на большие объемы данных
UCINET[16]	Excel, DL, text, Pajek .net, Krackplot, Negopy, proprietary (##.d & ##.h)	2011	+	содержит алгоритмов выявления сообществ. Используемый алгоритм кластеризации предназначен для группировки вершин, сходных по количеству и составу связей, что не обязательно приводит к выявлению сообществ
Программы, непосредственно работающие с Интернет-данными				
NodeXL[17]	graph data from a variety of file formats, including GraphML, UCINet, Pajek, and matrix a также интернет-данные	2011	+	ограничен количеством строк в Excel
IssueCrawler[18]	Интернет-данные	2002	+\-	ПО расположенно на сервере разработчика. Формирует сети по он-лайн запросу пользователя, исходя из набора первоначальных ссылок. Нет никаких данных о том, что является вершиной сети, как формируются границы сети, каковы алгоритмы визуализации и т.д.
Voson[19]	Интернет-данные	2010	_	Осуществляет сетевой анализа на базе R, синхронизирован с внешними программами текстового анализа и визуализаторами, а также осуществляет выгрузку данных в NodeXL. Интерпретирует в качестве ребер гиперссылки и самостоятельно формирует узлы на основе объединения страниц в сайты
Выполняющие функцию community detection				
igraph[20]	.txt (edge list), .graphml, .gml, .ncol, .lgl, .net	2009	+	Библиотека алгоритмов анализа графов. Используется наибольшее разнообразие продвинутых алгоритмов выявления сообществ. От пользователя требуется знание языков программирования

Программы-визуализаторы специализируются на разных способах графического представления данных. Для целей нашего исследования мы не рассматривали их подробно, т.к. визуализация в отдельном модуле требует дополнительной перекодировки данных на каждом шаге. К стандартным универсальным пакетам сетевого анализа, широко известным социологам, относятся такие программы, как Pajek, ORA, UCINET, а также сходный с ними NodeXL. Эти программы относительно просты в освоении, а NodeXL, являясь надстройкой над Excel, вообще почти не требует освоения. Для них характерны не оптимальные сочетания возможностей по объему данных и применяемых

[14] *vlado.fmf.uni-lj.si/pub/networks/**pajek**/*
[15] *www.casos.cs.cmu.edu/projects/**ora**/*
[16] *www.analytictech.com/**ucinet***
[17] ***nodexl**.codeplex.com*
[18] *www.**issuecrawler**.net*
[19] ***voson**.anu.edu.au*
[20] ***igraph**.sourceforge.net*

алгоритмов. Третьей категорией ПО для сетевого анализа являются программы, непосредственно работающие с Интернет-данными, такие как IssueCrawler, Voson и NodeXL. Общей проблемой является то, что они загружают не те данные и не в такой реляционной структуре, которые требуются для целей данного исследования.

Отдельно стоит выделить igraph – библиотеку алгоритмов анализа графов, реализованную для дальнейшего использования при программировании в языках (средах) C, Питоне или в R. В igraph используется наибольшее разнообразие продвинутых алгоритмов выявления сообществ. Плюсом также является то, что библиотека для C ограничена в приеме данных большого объема только возможностями самих алгоритмов и техническими параметрами компьютера; для R и Питона – тем же + возможностями скриптовых языков программирования по сравнению с компилируемыми. Однако существенным минусом является то, что пользователь фактически должен быть программистом на одном из этих языков.

.

Сводная таблица по рассмотренному программному обеспечению

Решаемая задача	**Наиболее пригодное ПО**	**Достоинства. Проблемы**
программы закачки и хранения	Voson	Собирает сетевые и текстовые Интернет-данные. В связи с тем, что он также закачивает и интерпретирует в качестве ребер гиперссылки (а не комменты) и самостоятельно формирует узлы на основе некоего объединения страниц в сайты, использование его для решения данной задачи не уместно.
	Koltran BlogMiner	Самостоятельно разработанная программа. Позволяет закачивать и хранить текстовые данные, представленные в определенной структуре.
программы для пре-процессинга	mystem	Программа, разработанная Яндексом. Производит морфологический анализ слов. Для запуска требуется дополнительная программа
	stem	Программа, разработанная участником проекта. Предназначена для запуска mystem и лемматизации текстов
	Doc2mat	Строит векторные матрицы текстов
программы для извлечения структуры текста (типа семантических решеток и сетей) и смысла текста (автоматическое аннотирование и пр.)	Hamlet II	Все программы являются платными. Позволяют проводить контент-анализ, реферирование и аннотирование текстов.
	Concordance	
	Content Analyzer	
программы извлечения тем и деления текстов на группы	gCluto	Требует дополнительных изменений в исходном коде для работы с русским языком. Возможность работать с большими массивами тестируется.
	R	Не работает на больших массивах
	Stanford Topic Modeling Toolbox	Тестирование показало способность ПО работать с данными 10^4 текстов. Недостатком ПО является сложность определения ярлыков для тем.
программы sentiment analysis	Sentistrength	Позволяет определять общее настроение коротких текстов. Для работы с русским требуется дополнительная настройка.

программы для сетевого анализа		
программы-визуализаторы	SocNetV	Основная функция - разные способы графического представления данных
Универсальные пакеты	Pajek	принимает данные в почти 10^{10} узлов, однако он не содержит алгоритмов выявления сообществ
	ORA	предназначена, для анализа организационных сетей, в т.ч. динамических, с большим количеством сопровождающих данных, поэтому в целом не рассчитана на большие объемы данных
	UCINET	содержит алгоритмов выявления сообществ. Используемый в ней алгоритм кластеризации предназначен для группировки вершин, сходных по количеству и составу связей, что не обязательно приводит к выявлению сообществ
	NodeXL	ограничен количеством строк в Экселе
программы, непосредственно работающие с Интернет-данными	IssueCrawler	ПО, расположенное на сервере разработчика, которое формирует сети по он-лайн запросу пользователя исходя из набора первоначальных ссылок. Однако совершенно нет никаких данных даже о том, что именно является вершиной сети, как формируются границы сети, каковы алгоритмы визуализации и т.д.
	Voson	осуществляет сетевой анализа на базе R, синхронизирован с внешними программами текстового анализа и визуализаторами, а также осуществляет выгрузку данных в NodeXL. Из-за того, что интерпретирует в качестве ребер гиперссылки (а не комменты) и самостоятельно формирует узлы на основе некоего объединения страниц в сайты, изучение было остановлено
	igraph –	библиотека алгоритмов анализа графов. Используется наибольшее разнообразие продвинутых алгоритмов выявления сообществ. Однако от пользователя требуется знание языков программирования