Министерство образования и науки Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра дискретной математики и
информационных технологий

Разработка информационной технологии аналитики текста средствами RapidMiner

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 421 группы
направления 09.03.01 «Информатика и вычислительная техника»
факультета компьютерных наук и информационных технологий
Ковалевского Константина Андреевича

Научный руководитель
д.э.н., профессор _____ Л.В. Кальянов
подпись, дата

Зав. кафедрой
к. ф.-м.н., доцент _____ Л.Б. Тяпаев
подпись, дата

Саратов 2018

ВВЕДЕНИЕ

В современном мире люди постоянно сталкиваются с огромным количеством информации. В связи с необходимостью аналитической обработки больших объемов данных большую популярность обрела технология интеллектуального анализа данных. Интеллектуальный анализ данных включает в себя методы математической статистики и машинного обучения для решения задач анализа данных. Данные технологии открыли новые возможности перед аналитиками, менеджерами, а также руководителями компаний. Системы интеллектуального анализа данных применяются в научных исследованиях, образовании, производстве и многих других отраслях. В связи с большим ростом актуальности анализа данных на рынке IT технологий стремительно начало развиваться программное обеспечение, предназначенное упростить процесс интеллектуального анализа данных. Разработка информационных технологий текстовой аналитики одна из наиболее актуальных и перспективных задач в сфере интеллектуального анализа данных.

Целью выпускной квалификационной работы является разработка информационной технологии текстовой аналитики средствами программы RapidMiner. В ходе работы были поставлены следующие задачи:

- изучение методов интеллектуальной аналитики текста;
- изучение методов анализа тональности текста;
- изучение классификации и математических методов кластеризации текста;
- анализ и сравнение программного обеспечения используемого для аналитики текста;
- обзор программы RapidMiner;
- анализ операторов программы RapidMiner, предназначенных для аналитики текста.

В работе будут описаны основные задачи, методы реализации и области применения технологии интеллектуального анализа текста. Актуальность данной работы обусловлена тем, что с помощью технологии интеллектуальной аналитики

текста можно анализировать большие объемы данных с минимальными ресурсными затратами, что облегчает процесс анализа данных.

Выпускная квалификационная работа состоит из введения, трех глав, заключения и списка используемых источников.

В первой главе приведены определения основных понятий, рассмотрены методы, классификация и анализ основных задач, связанных с интеллектуальной аналитикой текста.

Во второй главе проведен анализ и сравнение программного обеспечения, предназначенного для текстовой аналитики. На основе анализа была выбрана программа RapidMiner, функциональность которой подробно рассмотренная в данной главе.

В третьей главе описана разработка информационных технологий интеллектуальной аналитики текста средствами RapidMiner, решающих задачи кластеризации текста и анализа тональности текста.

1 Методы интеллектуальной аналитики текста. Текстовая аналитика (от англ. text maining), также известная как интеллектуальный анализ текста, представляет собой процесс изучения больших объемов текстовых ресурсов для генерации новой информации и преобразования неструктурированного текста в структурированные данные для использования в дальнейшем анализе [1].

Одними из ключевых задач, решаемых с помощью интеллектуальной аналитики текста являются:

- *Категоризация текстов.* Задача категоризации, то есть отнесение текста к одной или нескольким определенным темам. Решение данной задачи позволяет пользователю значительно упростить процесс структуризации текста.
- *Извлечение информации*. Целью данной задачи является автоматическое извлечение структурированной информации из определенного потока неструктурированных или слабоструктурированных данных. Основным преимуществом такого преобразования является возможность быстрого анализа изначально «хаотичной» информации.
- *Информационный поиск.* Информационный поиск - процедура поиска неструктурированной текстовой информации, удовлетворяющей информационные потребности пользователя.
- *Кластеризация текстов.* Кластеризация текста одна из задач информационного поиска. Целью кластеризации является автоматическое выявление групп семантически похожих документов среди некоторых фиксированных данных.
- *Анализ тональности текста.* Анализ тональности текста - методы текстовой аналитики, предназначенные для выявления в тексте эмоционально окрашенной лексики [8].

Поскольку интеллектуальный анализ текста включает в себя широкий спектр задач, следует остановится на более узкой спецификации аналитики

текста. В данной работе рассмотрены задача анализа тональности текста и задача кластеризации текста.

Тональность - это эмоциональное отношение пользователя к какому-то событию, высказыванию, действию или продукту. Общим примером использования данной технологии является выяснение того, как люди относятся к определенной теме или событию [3]. Анализ тональности позволяет узнать, почему люди думают, что продукт хороший или плохой, путем извлечения точных слов, которые указывают, почему людям нравится или не нравится данный продукт.

Существует пять основных методов решения задачи автоматического определения тональности текста:

- *Статистический метод.* Для данного метода необходимы заранее размеченные по тональности коллекции текстов, с помощью которых происходит обучение модели, с использованием которой происходит определение тональности текста. При статистическом методе решения анализа тональности текста широко используется метод опорных векторов (SVM).
- *Метод, основанный на словарях.* Данный метод используется на основе составленных словарей позитивных и негативных выражений. Этот метод может использовать как списки шаблонов, так и правила соединения тональной лексики внутри предложения.
- *Смешанный метод.* Данный метод основан на комбинации статистического метода и метода, основанного на словарях.
- *Машинное обучение без учителя* (от англ. unsupervised learning). Данный подход основан на идее, что наибольший вес в тексте имеют термины, которые чаще встречаются в этом тексте и в то же время присутствуют в небольшом количестве текстов всей коллекции. Выделив данные термины и определив их тональность, можно сделать вывод о тональности всего текста.
- *Метод, основанный на теоретико-графовых моделях*. В основе этого метода используется предположение о том, что не все слова в

текстовом корпусе документа равнозначны. Какие-то слова имеют больший вес и сильнее влияют на тональность текста [4].

Из всех вышеперечисленных методов самыми распространенными методами являются статистический метод и метод, основанный на словарях, так как они являются наиболее простыми в реализации и использование.

Кластеризация (или кластерный анализ) — это задача разбиения множества объектов на группы, называемые кластерами. В кластеризации, цель заключается не в прогнозировании переменной целевого класса, а в простом захвате возможных «природных» группировок в данных [5].

В кластерном анализе можно выделить следующие методы:

- *N кластеризация на основе прототипа*. При кластеризации на основе прототипов каждый кластер представлен центральным объектом данных, также называемым прототипом. Прототипом каждого кластера обычно является центр кластера, поэтому эта кластеризация также называется центрированной кластеризацией.
- *Плоскостная пространственная кластеризация*. Тип кластеризации на основе плотности, заданного набора точек в каком-либо пространстве. Данный тип кластеризации группирует вместе точки, тесно связанные друг с другом (точки со многими соседями), отмечая их как точки выброса, которые лежат в областях с низкой плотностью.
- *Иерархическая кластеризация*. Иерархическая кластеризация - это процесс, при котором кластерная иерархия создается на основе расстояния между точками данных. Результатом иерархической кластеризации является дендрограмма: древовидная диаграмма, которая показывает различные кластеры в любой точке точности, указанной пользователем.
- *Кластеризация на основе моделей*. Этот метод, также называемый кластеризацией на основе распределения. Кластер можно рассматривать как группу, имеющую точки данных, принадлежащие к той же вероятности распределение.

Следовательно, каждый кластер может быть представлен в виде распределение модели.

Одним из самых распространённых методов кластеризации является метод *k-средних* (от англ. k-means) [6]. Это метод кластеризации на основе прототипа, где набор данных разделен на k кластеров. Метод k-средних стремится минимизировать суммарное квадратичное отклонение точек кластеров от центров этих кластеров. в качестве функции расстояния используется Евклидово расстояние.

Так же, среди методов кластеризации можно выделить алгоритм *пространственной кластеризация приложений с шумом на основе плотности* (от англ. Density-based spatial clustering of applications with noise или DBSCAN). Алгоритм кластеризации DBSCAN идентифицирует кластеры в данных основанных на измерении распределения плотности в n-мерном пространстве [7].

2 Программные инструменты текстовой аналитики. На сегодняшний день на рынке инструментов текстовой аналитики существует несколько основных программных решений. Для сравнения функциональности были выбраны следующие программы:

- GATE (General Architecture for Text Engineering).
- Knime Analytics Platform.
- Orange software.
- LPU (Learning from Positive and Unlabeled Examples).
- RapidMiner.

В результате анализа была построена таблица 1. На основе проведенного анализа был сделан вывод, что что RapidMiner удовлетворяет всем современным требованиям в сфере интеллектуального анализа текста, он хорошо подходит для академических целей так как обладает удобным интерфейсом и обширной справочной системой. RapidMiner - это программа, разработанная компанией RapidMiner, используемая для разработки коммерческих, некоммерческих приложений, а также для научных исследований и образования [8].

Таблица 1 – Сравнение программного обеспечения.

Предметы сравнения	Gate	KNIME	Orange software	LPU	RapidMiner
Обработка больших объемов данных	+	+	+	-	+
Удобный пользовательский интерфейс	+	+	+	-	+
Работа с локальными файлами	-	+	+	-	+
Визуализация процесса	+	+	+	+	+
Справочная система	-	+	-	-	+
Интеграция с приложениями	Java	Java, Python,R	Python	-	Java, Python,R
Поддержка русского языка	+	-	-	-	-

Для решения задач с помощью RapidMiner был проведен подробный обзор функциональности, рассмотрена основная терминология построения процесса. Для более подробного изучения программы был проведен обзор операторов и основных расширений для текстовой аналитики предоставляемых программой RapidMiner.

3 Разработка информационной технологии текстовой аналитики. В ходе работы были разработаны следующие технологии текстовой аналитики:

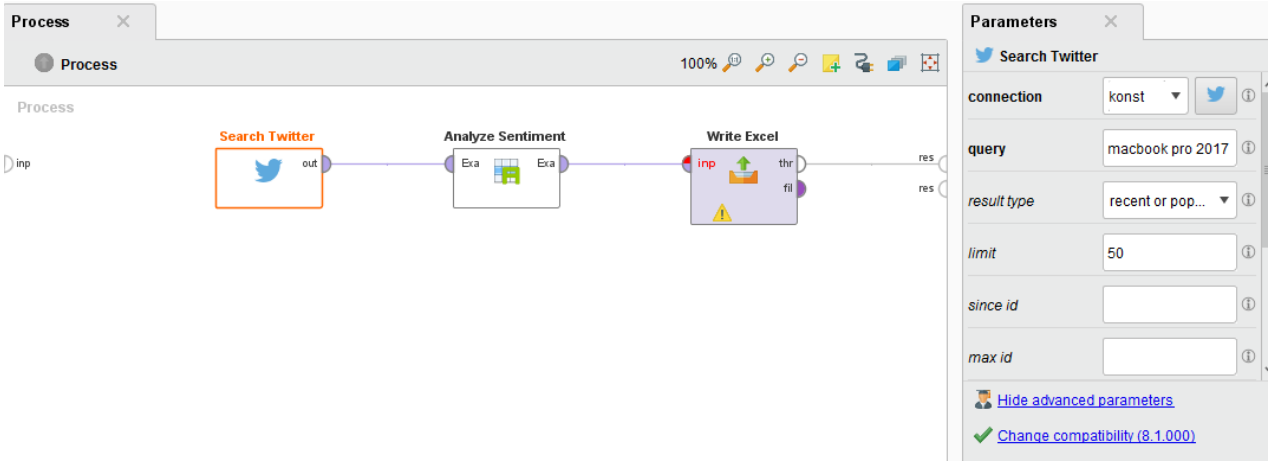

Рисунок 1– Процесс анализа данных Twitter.

Процесс, изображенный на рисунке 1, позволяет производить анализ тональности мнений пользователей социальной сети Twitter.

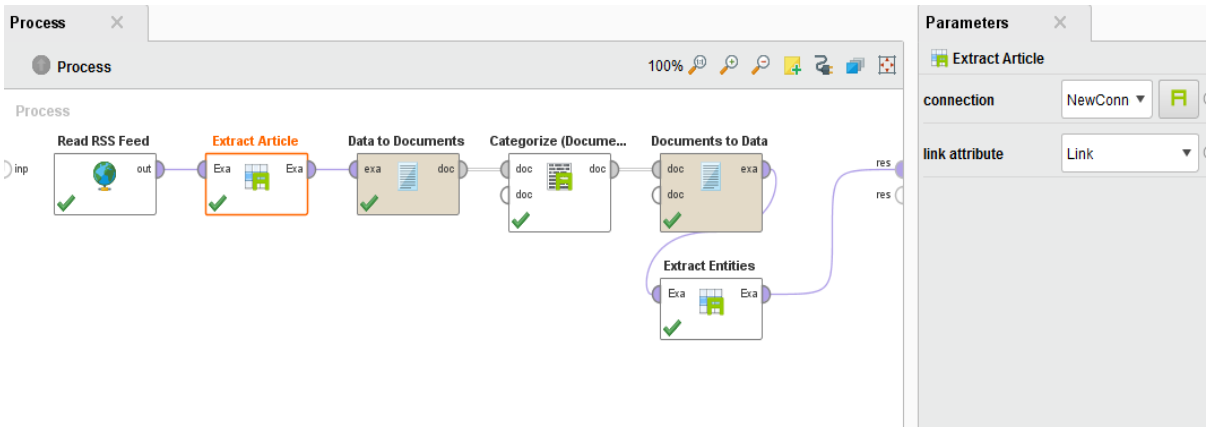

Рисунок 2 – Процесс анализа данных новостных RSS каналов.

Процесс, изображенный на рисунке 2, позволяет анализировать и структурировать новости из различных RSS каналов.

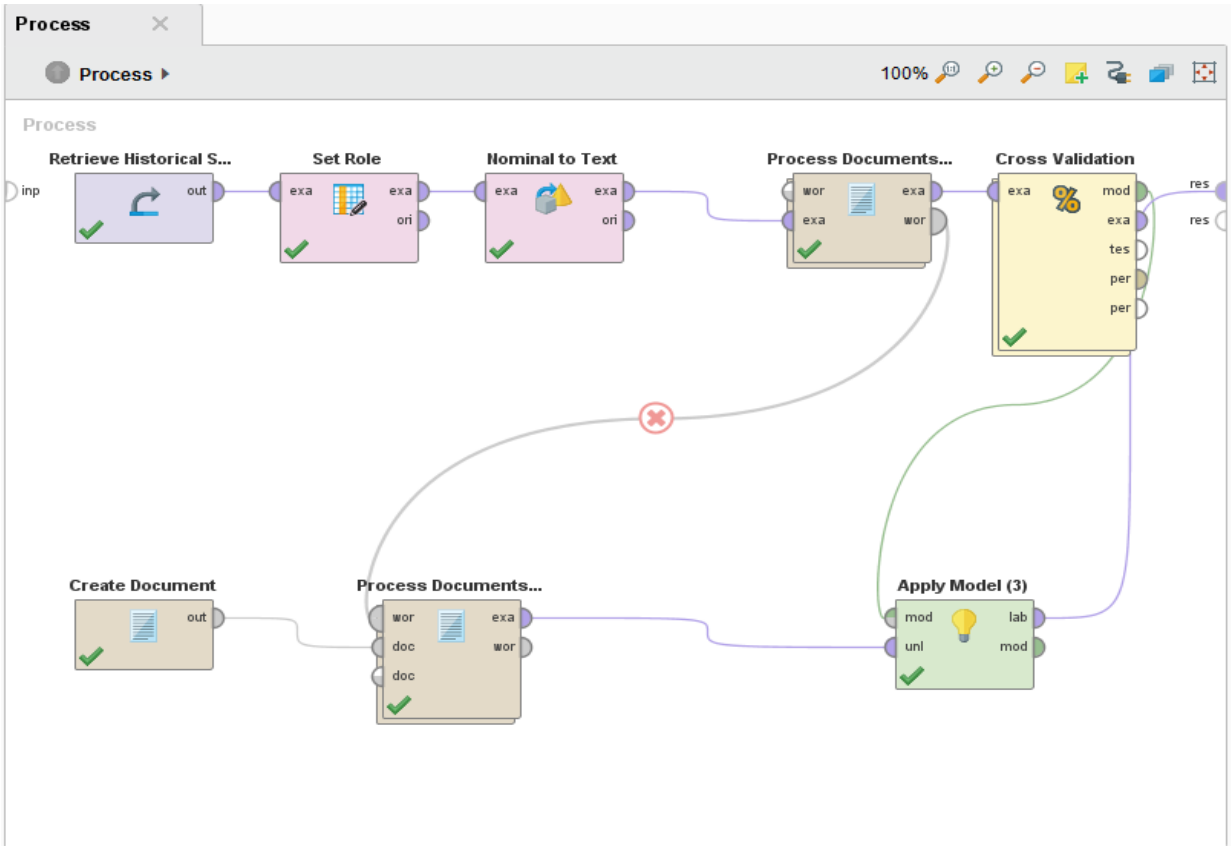

Рисунок 3 – Процесс анализа тональности текста.

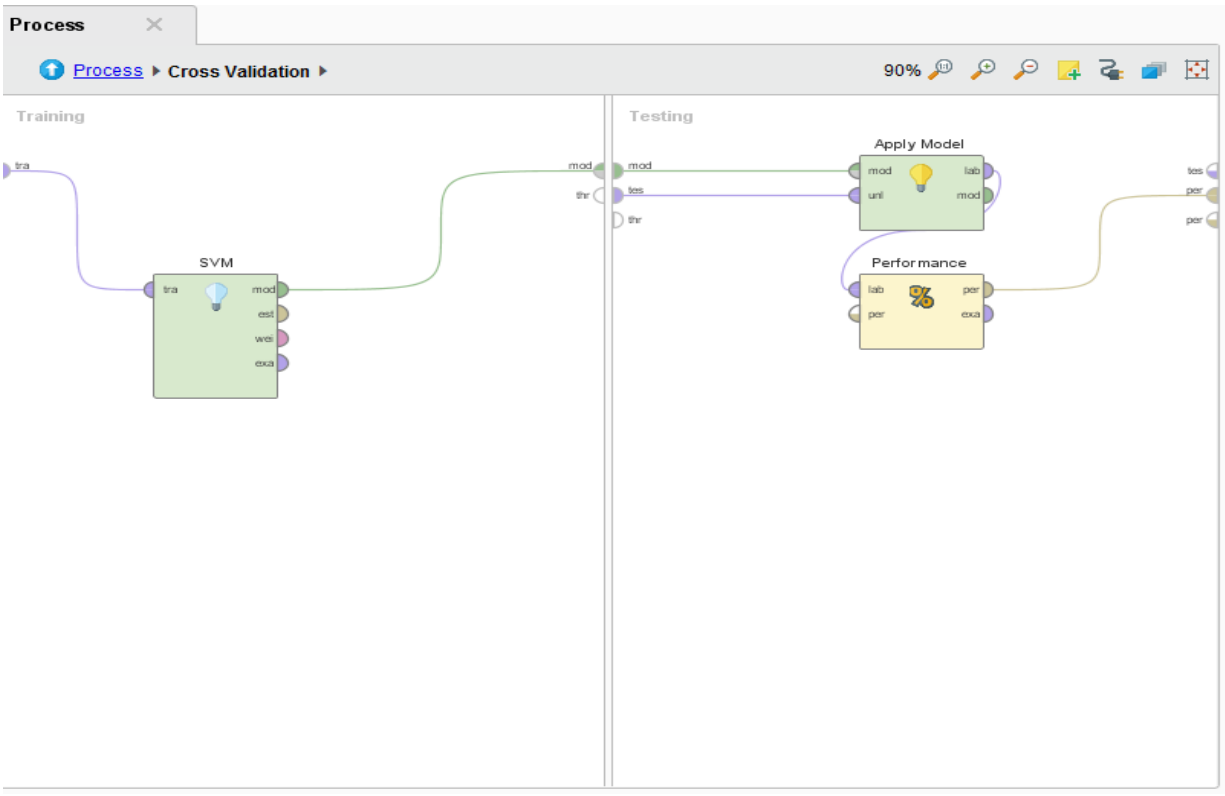

Рисунок 4 – Процесс анализа тональности текста.

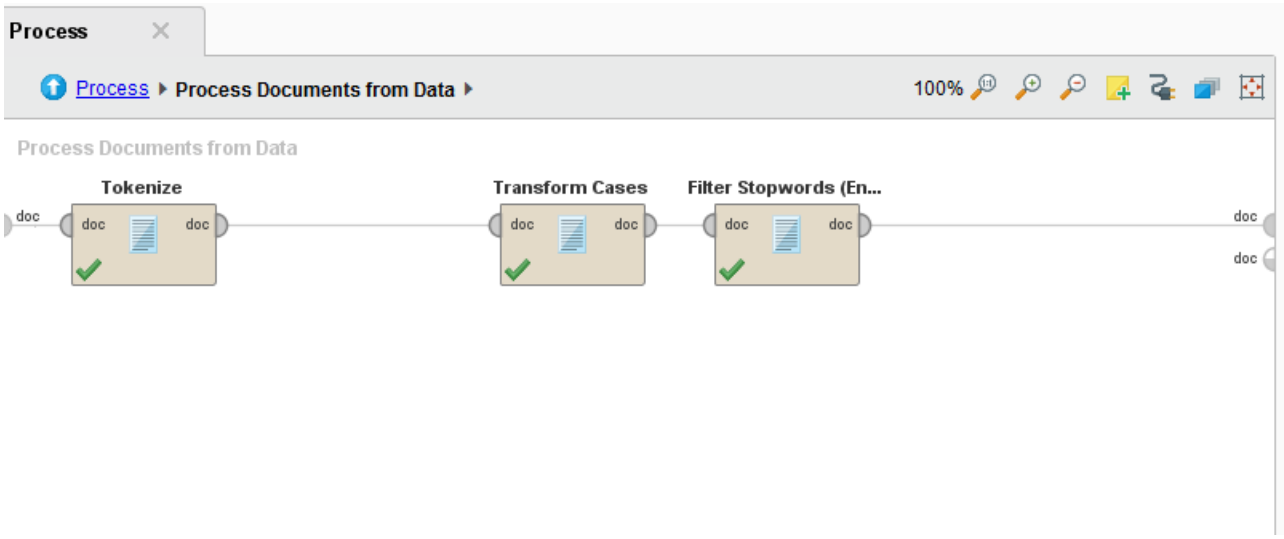

Рисунок 5 – Процесс анализа тональности текста.

Процесс, изображенный на рисунках 3-5, анализирует тональность текста используя метод, основанный на словарях.

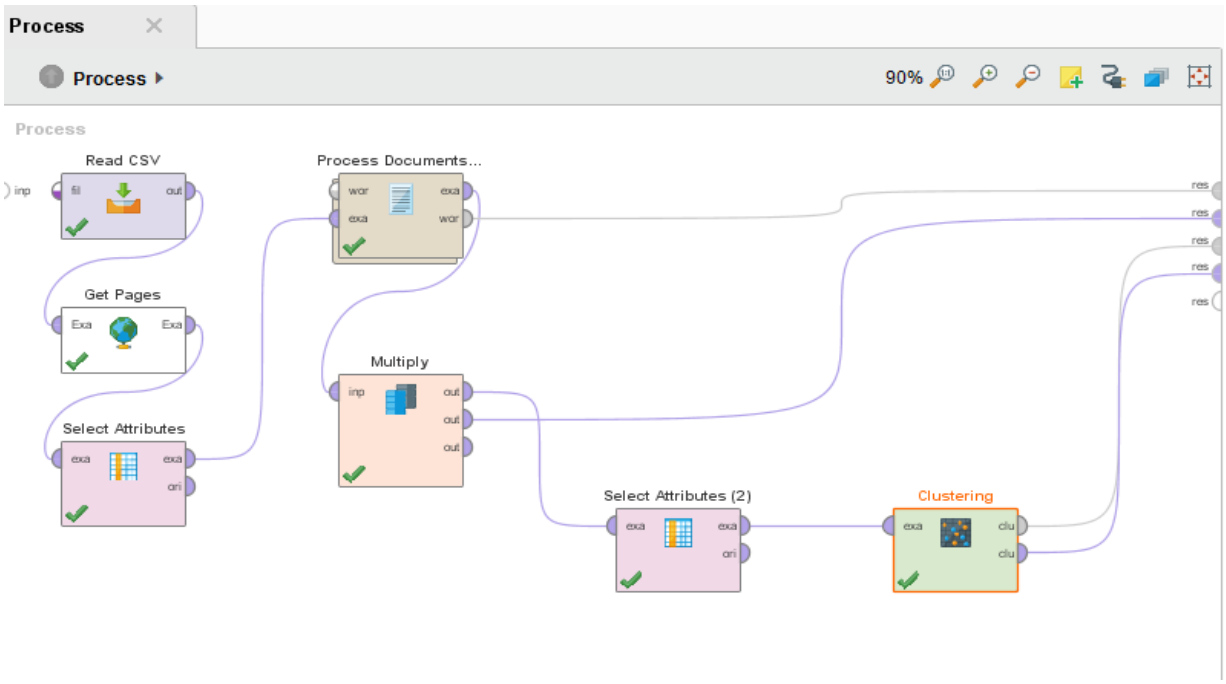

Рисунок 6 – Процесс кластеризации текста.

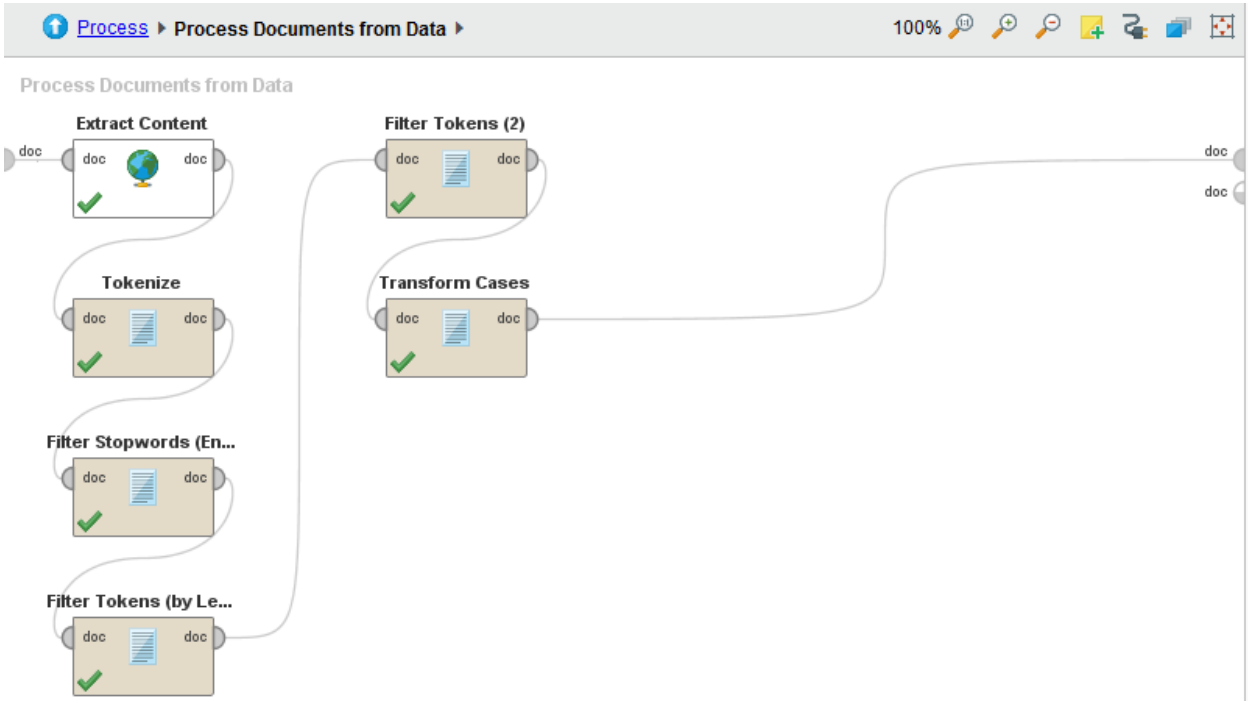

Рисунок 7 – Процесс кластеризации текста.

Процесс, изображенный на рисунках 6-7, кластеризует информацию, содержащуюся на указанном интернет ресурсе. Кластеризация осуществляется на основе ключевых слов с использованием метода k-средних.

ЗАКЛЮЧЕНИЕ

В ходе выполнения выпускной квалификационной работы были получены следующие результаты:

1. Приведены основные определения и понятия, связанные с задачей интеллектуального анализа текста.
2. Рассмотрена классификация задач интеллектуальной аналитики текста.
3. Описаны основные методы анализа тональности текста и методы кластеризации текста.
4. Проведен анализ и сравнение программного обеспечения, предназначенного для интеллектуального анализа текста.
5. Проведен подробный обзор функциональности программы RapidMiner.
6. Разработаны информационные технологии интеллектуальной аналитики текста, с помощью которых были решены задачи анализа тональности и кластеризации текста.

Из данной работы можно сделать вывод, что средства программы RapidMiner позволяют удобно и быстро решать задачи, связанные с текстовым анализом. Благодаря понятному интерфейсу, широкой функциональности и оптимизации RapidMiner является одной из самых удобных и производительных программ для текстовой аналитики.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Data Mining. [Электронный ресурс]: URL: https://ru.wikipedia.org/wiki/Data_mining (дата обращения 02.05.2018)
2. Ronen Feldman, James Sanger, The text mining handbook - Cambridge University 2007
3. Анализ тональности текста. [Электронный ресурс]: URL: https://ru.wikipedia.org/wiki/анализ_тональности_текста (дата обращения 15.04.2018)
4. Сентимент анализ текста. [Электронный ресурс]: URL: https://habr.com/company/palitrumlab/blog/262595 (дата обращения 02.05.2018)
5. Roger Bilisoly, Practical Text Mining - Central Connecticut State University 2008
6. Метод k-средних. [Электронный ресурс]: URL: https://ru.wikipedia.org/wiki/метод_k-средних (дата обращения 10.04.2018)
7. Кластерный анализ. [Электронный ресурс]: URL: https://ru.wikipedia.org/wiki/Кластерный_анализ (дата обращения 18.05.2018)
8. Официальный сайт AYLIEN. [Электронный ресурс]: URL: https://aylien.com (дата обращения 03.04.2018)