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

Текстовая аналитика средствами Rapidminer

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 421 группы
направления 09.03.01 «Информатика и вычислительная техника»
факультета компьютерных наук и информационных технологий
Сулейманова Расима Равиковича

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

Компьютеризация общества значительно расширила наши возможности для генерации и сбора данных из разных источников. Громадный объем данных заполнил почти каждый аспект нашей жизни: наука, медицина, финансы, торговля, маркетинг, информационные технологии и т.д.

Из-за большого роста хранящихся и текущих данных появилась необходимость в новых методах и автоматизированных инструментах, которые могут эффективно и результативно помочь в преобразовании огромного объема данных в полезную информацию и знания.

Это привело к появлению такого направления в области компьютерных наук, как data mining, или интеллектуальный анализ данных, и его различных приложений. Knowledge discovery from data (KDD), так же широко используемый термин для обозначения data mining, представляет собой автоматическое или удобное извлечение шаблонов, представляющих знания, неявно хранящихся или скрытых в больших базах данных, хранилищах данных, в Интернете, других массивных информационных хранилищах или потоках данных.

Целью данной дипломной работы является выявление актуальных тем интернет публикаций средствами текстовой аналитики.

Для выполнения этой цели были поставлены следующие задачи: изучение интернет ресурса habr.com, моделирование процесса в приложении rapidminer для загрузки статей публикаций, анализа частоты появления статей определенных тем на данном ресурсе, путем классификации данных статей по темам.

Для решения данной задачи использовались средства data mining, text mining, web mining.

Для решения данной задачи были изучены и применены следующие методы, модели и алгоритмы: TF-IDF, SVM, CV, K-NN[1][2][3].

1 Классификация текстов в rapidminer

1.1 Среда разработки

Для решения данной задачи в качестве среды разработки выбор пал на rapidminer studio версии 8.2.000.

RapidMiner - инструмент анализа данных с достаточно хорошим набором операторов решающих большой спектр задач получения и обработки информации из разнообразных источников (базы данных, файлы и т.п.).

Возможности RapidMiner могут быть расширены с помощью дополнений. Система поддерживает все этапы глубинного анализа данных, включая результирующую визуализацию, проверку и оптимизацию[15][16][17][18].

1.2 Выбор объекта исследований

Для решения данной задачи объектом исследования был выбран популярный интернет ресурс для IT-специалистов habr.com.

Средствами web mining было загружено более 8000 статей с целью их дальнейшей классификации на темы:

- Разработка
- Администрирование
- Гиктаймс
- Управление
- Маркетинг
- Дизайн
- Разное

На основании данной классификации был сделан вывод о наиболее популярных и менее популярных темах[19].

1.3 Описание процесса

Для решения данной задачи в Rapidminer Studio был создан процесс. Внутри данного процесса реализована загрузка статей с habr.com, добыча данных и полученных текстов, нормализация текста, обработка данных, обучение модели и ее реализация для классификации текстов.

Классификация - это процесс нахождения модели (или функции), которая описывает и отличает классы данных или общее представление. Модель выводится на основе анализа набора данных обучения (Объектов данных, для которых известны метки классов). Модель используется для прогнозирования метки класса объектов, для которых метка класса неизвестна.

1.4 Добыча данных

С сайта habr.com было загружено более 8000 статей для дальнейшего анализа.

Добыча данных производилась с помощью оператора web crawl, расширения Web Mining.

Для загрузки исключительных страниц с помощью регулярных выражений были установлены условия следования и загрузки страниц.

Данные были сохранены на жестком диске.

Также для обучения модели классификации было подобрано по 200 статей для каждой категории классификации[20].

1.5 Нормализация текста. Создание векторов

Данные для анализа и данные для обучения были загружены в обработчик с помощью оператора Process Documents from Files, расширения Text Processing.

С помощью вложенных операторов Tokenize, Transform Cases, Stem и Filter Tokens, данные были обработаны для более успешного анализа.

Оператор Tokenize: делит текст на последовательность токенов (отдельных слов). На данном этапе удаляются знаки препинания, числа, и остальные не буквенные символы.

С помощью оператора Transform Cases все заглавные буквы были заменены строчными.

Был создан словарь стоп слов, которые не несут никакого значения в контексте данного исследования (такие, как города, месяцы, предлоги, союзы, местоимения и т.д.) С помощью оператора Filter Stopwords данные слова были удалены из списка.

Оператор Stem отсекает от слова окончания и суффиксы, чтобы оставшаяся часть, называемая stem, была одинаковой для всех грамматических форм слова.

Оператор Filter Tokens удалил все токены длинной менее 3 и более 25 букв.

Далее, для каждого получившегося токена был создан вектор TF-IDF и высчитано его значение[8].

1.6 Работа с данными

Filter Empty Text - фильтрация атрибутов по назначению.

Replace Missing Values - пустые значения атрибутов заменяются на ноль.

Set Role - атрибутам под именем label назначается тип label.

Filter unknown - данные для обучения и обработки фильтруются и направляются на обучение и обработку соответственно.

Weight by SVM - вычисление значимости данных для каждой категории, путем подсчета веса атрибутов для данной категории. Подсчеты производятся с применением алгоритма SVM.

Select by Weights - выбор наиболее значимых атрибутов для каждой категории.

1.7 Cross Validation. Обучение модели

Следующий оператор проводит кросс валидацию данных.

Cross-validation или скользящий контроль - метод оценки аналитической модели и её поведения на независимых данных. При оценке модели имеющиеся в наличии данные разбиваются на частей. Затем на частях данных производится обучение модели, а оставшаяся часть данных используется для тестирования. Процедура повторяется раз; в итоге каждая из частей данных используется для тестирования. В результате получается оценка эффективности выбранной модели с наиболее равномерным использованием имеющихся данных.

Модель обучается по алгоритму k-NN. Метод k - ближайших соседей - метрический алгоритм для автоматической классификации объектов и регрессии.

Классификатор ближайших соседей основан на обучении по аналогии, то есть путем сравнения обучающей выборки и анализируемых данных. Каждый объект выборки представлен точкой в n-мерном пространстве.

При определении класса объекта классификатор ищет ближайшие объекты к искомому объекту.

Затем модель применяется на обучающей выборке для проверки с помощью оператора Performance.

1.8 Применение модели. Результат работы

Обученная модель классификации применяется на ранее загруженных текстах и классифицирует данные статьи по категориям.

Результат классификации представлен на диаграмме, которую можно видеть на рисунке 1.

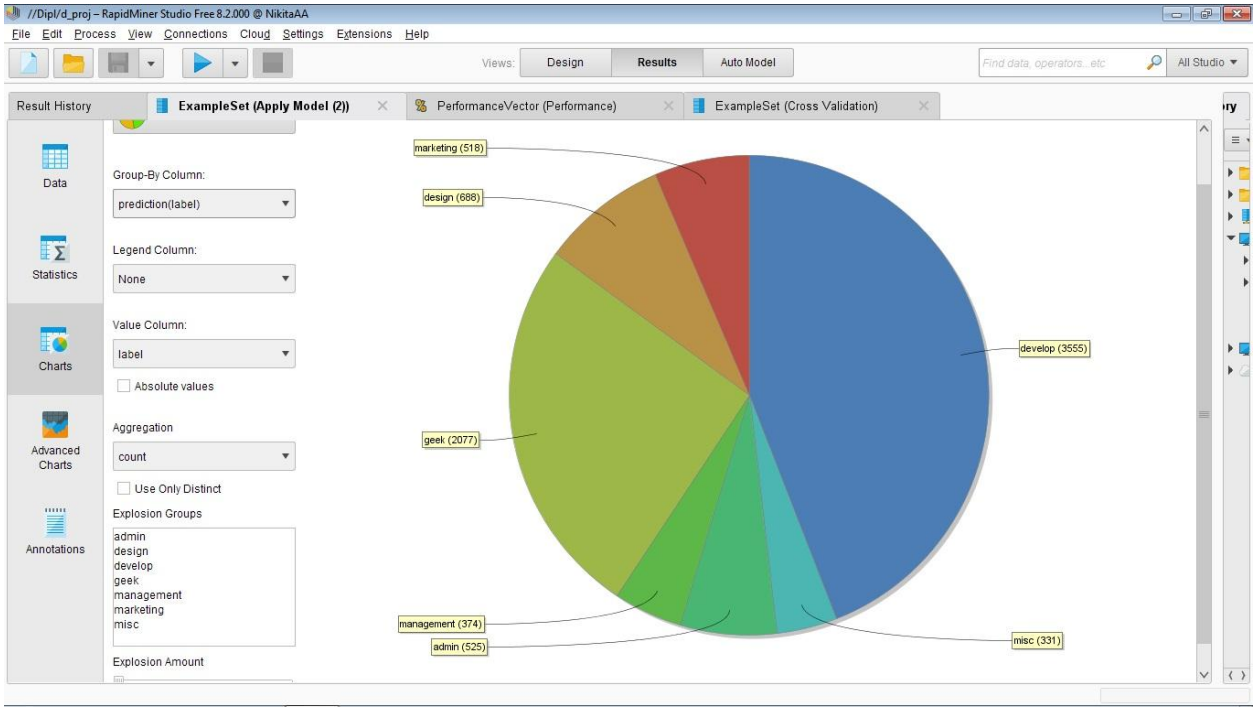

Рисунок 1 – Результат работы.

На данной диаграмме можно увидеть сколько статей попало в каждую категорию:

Разработка (develop): 3555 статей (44%);

Разное (misc): 331 статей (4.3%);

Администрирование (admin): 525 статей (6.5%);

Управление (management): 374 статей (4.6%);

Гиктаймс (geek): 2077 статей (25.7%);

Дизайн (design): 688 статей (8.5%);

Маркетинг (marketing): 518 статей (6.4%).

ЗАКЛЮЧЕНИЕ

В ходе работы был разработан процесс в приложении rapidminer, проанализированы и классифицированы статьи интернет ресурса habr.com и были выявлены наиболее и наименее актуальные темы. Для решения данной задачи использовались средства data mining, text mining, web mining.

Средствами текстовой аналитики были выявлены актуальные темы интернет публикаций.

Для выполнения этой цели были решены следующие задачи: изучение интернет ресурса habr.com, моделирование процесса в приложении rapidminer для загрузки статей публикаций, анализа частоты появления статей определенных тем на данном ресурсе, путем классификации данных статей по темам.

Для решения данной задачи использовались средства data mining, text mining, web mining.

Для решения данной задачи были изучены и применены следующие методы, модели и алгоритмы: TF-IDF, SVM, CV, K-NN.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 Ian H. Witten, Frank Eibe, Mark A. Hall. Data mining : practical machine learning tools and techniques.—3rd ed. 2011 ISBN 978-0-12-374856-0.

2 Интеллектуальный анализ данных: базовые понятия [Электронный ресурс] сайт. https://www.intuit.ru/studies/courses/2312/612/lecture/13260 Дата обращения 11.05.2018.

3 Интеллектуальный анализ данных. [Электронный ресурс] сайт. URL: http://www.seun.ru/content/learning/4/science/2/doc/Интеллектуальный%20анализ%20данных.pdf. Дата обращения 15.05.2018.

4 В. Дюк, А. Самойленко. Data mining. Учебный курс, 2001. ISBN 5-318-0022707.

5 А. А. Барсегян, М. С. Куприянов, И. И. Холод, М. Д. Тесс, С. И. Елизаров. Анализ данных и процессов: учеб. пособие, 3-е изд., перераб., 2009г. ISBN 978-5-9775-0368-6.

6 Jiawei Han and Micheline Kamber. Data Mining: Concepts and Techniques, Second Edition 2006. ISBN 978-1-55860-901-3.

7 Скользящий контроль. [Электронный ресурс] сайт. URL: http://www.machinelearning.ru/wiki/index.php?title=CV Дата обращения 12.05.2018.

8 Text segmentation - Wikipedia. [Электронный ресурс] сайт. URL: https://en.wikipedia.org/wiki/Text_segmentation#Word_segmentation. Дата обращения: 17.05.2018.

9 Web mining: основные понятия. [Электронный ресурс] сайт. URL: https://basegroup.ru/community/articles/basic-conceptions. Дата обращения 20.05.2018.

10 Кутукова Е.С. Технология Text mining. [Электронный ресурс] сайт. URL: https://www.sworld.com.ua/konfer33/121.pdf. Дата обращения 15.05.2018.

11 Текстовая аналитика. [Электронный ресурс] сайт. URL: http://3itech.ru/production/tekstovaya-analitika. Дата обращения: 15.05.2018.

12 Технология web mining. [Электронный ресурс] сайт. URL: https://elibrary.ru/download/elibrary_25777733_62938668.pdf. Дата обращения 20.05.2018.

13 Web mining. [Электронный ресурс] сайт. URL: https://ru.wikipedia.org/wiki/Web_mining. Дата обращения 20.05.2018.

14 Сферы примерения data mining. Web mining. [Электронный ресурс] сайт. URL: https://www.intuit.ru/studies/courses/6/6/lecture/170?page=4. Дата обращения 16.05.2018.

15 Rapidminer Studio. [Электронный ресурс] сайт. URL: https://rapidminer.com Дата обращения 14.05.2018.

16 Наиболее полный список инструментов для анализа данных и машинного обучения. [Электронный ресурс]. URL: http://ru.datasides.com/big-data-analytic-tools/ Дата обращения 16.05.2018.

17 Введение в rapidminer. [Электронный ресурс] сайт. URL: https://habr.com/post/269427/ Дата обращения 16.05.2018.

18 RapidMiner Data Mining Use Cases and Business Analytics Applications 2014 Markus Hofmann, Ralf Klinkenberg ISBN 978-1-4822-0550-3.

19 Крупнейший в Европе ресурс для IT-специалистов. [Электронный ресурс] сайт. URL: https://habr.com Дата обращения 16.05.2018.

20 Шпаргалка по регулярным выражениям. [Электронный ресурс] сайт. URL: http://website-lab.ru/article/regexp/ Дата обращения: 20.05.2018