Министерство образования и науки Российской Федерации
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ (НИ ТГУ)
АООП «Цифровые технологии в социогуманитарных практиках»

ДОПУСТИТЬ К ЗАЩИТЕ В ГЭК
Руководитель ООП,
канд. ист. наук, доцент
_____ Г. В. Можаева
«14» 06 2018 г.

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

РАЗРАБОТКА МОДЕЛИ ПРОГНОЗИРОВАНИЯ ПРИЗНАКОВ ОДАРЕННОСТИ У СТАРШЕКЛАССНИКОВ ПО ИХ ПОДПИСКАМ НА СООБЩЕСТВА В СОЦИАЛЬНОЙ СЕТИ «ВКОНТАКТЕ»

по основной образовательной программе подготовки магистров
направление подготовки 09.04.03 – Прикладная информатика

Корепанов Константин Викторович

Научный руководитель ВКР
канд. ист. наук, доцент
_____ Г. В. Можаева
Научный консультант ВКР
с.преподаватель каф. ГПИ
_____ А. В. Фещенко
подпись
«04» 06 2018 г.

Автор работы
студент группы № 25201
_____ К. В. Корепанов
подпись

Томск-2018

ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
АООП «Цифровые технологии в социогуманитарных практиках»

УТВЕРЖДАЮ

Руководитель ООП,
канд. ист. наук, доцент
_____ Г.В. Можаева
«05» 04 2018 г.

ЗАДАНИЕ
по подготовке магистерской диссертации

магистранту Корепанову Константину Викторовичу группы 25201

1. Тема диссертации Разработка модели прогнозирования признаков одаренности у старшеклассников по их подпискам на сообщества в социальной сети «ВКонтакте»
2. Цель и содержание диссертации

Цель работы: на основе имеющихся данных о старшеклассниках (результаты психологического тестирования и сведения из профиля в социальной сети «Вконтакте») разработать прогностическую модель, определяющую признаки одаренности по открытом пользовательским данным «Вконтакте» для потенциальных абитуриентов ТГУ .

Содержание работы должно включать анализ исследовательских работ и практик по теме исследования, подготовку и обработку исходных данных о старшеклассниках, описание и апробацию методов анализа данных для разработки прогностической модели выявления одарённых абитуриентов, разработку модели прогнозирования признаков одарённости старшеклассников и оценку её точности и др.

3. Перечень вопросов, решаемых по заданию заинтересованных организаций и их наименование:
   1. Состояние российских и зарубежных исследований и практик по анализу данных в социальных сетях;
   2. Подготовка для анализа исходных данных о старшеклассниках: результаты психологического тестирования и сведения из профиля социальной сети «Вконтакте»;
   3. Выбор и апробация методов анализа данных для разработки прогностической модели выявления одарённых абитуриентов;
   4. Поиск и обоснование инструментов для анализа больших данных из социальной сети;
   5. Разработка модели прогнозирования признаков одарённости старшеклассников и оценка её точности;
   6. Автоматизация алгоритма прогнозирования признаков одарённости.
4. Сроки представления завершенной диссертации:
   - в Учебный офис ООП 13.06.2018
   - в ГЭК 15.06.2018
5. Предзащита на кафедре гуманитарных проблем информатики 04.05.2018
6. Консультанты по разделам диссертации: Фещенко А.В, старший преподаватель кафедры гуманитарных проблем информатики Национального исследовательского Томского государственного университета. (введение, основная часть, практическая работа по разработке стратегии продвижения, заключение) _____ А.В. Фещенко
7. Научный руководитель диссертации _____ Г.В. Можаева, к.и.н., зав.

кафедрой гуманитарных проблем информатики Национального исследовательского Томского государственного университета.

Дата «09» 04 2018 г.

Задание принял к исполнению «09» 04 2018 г.

магистрант Корепанов Константин Викторович

Утверждено на заседании совета ООП «09» апреля 2018 г.
Протокол № 6 (12)

АННОТАЦИЯ

Предметом исследования является феномен одарённости у старшеклассников. Объект исследования – прогнозирование признаков одарённости у старшеклассников по открытым пользовательским данным в социальной сети «Вконтакте».

Цель исследования – на основе имеющихся данных о старшеклассниках (результаты психологического тестирования и сведения из профиля в социальной сети «Вконтакте») разработать прогностическую модель, определяющую признаки одаренности по открытом пользовательским данным «Вконтакте» для потенциальных абитуриентов ТГУ.

При выполнении ВКР использованы Python 3.6, пакет библиотек Anaconda и MS Excel.

Результатом работы является прогностическая модель, позволяющая выявлять одарённых абитуриентов по их профилю в социальной сети. Внедрение модели в рекрутинговые компании региональных вузов позволит повысить качество образования за счет привлечения одаренных абитуриентов.

Апробация модели показала высокую точность прогнозирования. Программная реализация модели выполнена на языке программирования Python 3.6., с использованием пакета библиотек Anaconda.

Рассмотренные в работе методики, подходы и проведенное исследование с разработанным алгоритмом обработки больших данных показали, что технологии машинного обучения имеют большой потенциал для анализа информации в социальных сетях.

Магистерская диссертация содержит: 60 с., 25 рис., 4 табл., 2 приложения и 63 литературных источника.

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

В связи со стремительным ростом пользователей в социальных сетях возникает вопрос о методах анализа и способах обработки постоянно возрастающей информации в сети для решения различных задач: от бытовых проблем и заканчивая проблемами мирового масштаба. Ранее было доказано [1], что принадлежность пользователей к сообществам в социальных сетях Facebook и Twitter может быть использована для прогнозирования психологических качеств подписчиков этих сообществ. Facebook – самая крупная по количеству пользователей социальная сеть [2]. В ней зарегистрировано более 1 млрд. человек. В профилях пользователей можно встретить различные демографические атрибуты: пол, возраст, семейное положение, политические и религиозные взгляды и т.д.

Классификаторы, которые извлекаются из авторских постов и профилей пользователей, могут с высокой долей вероятности прогнозировать значения атрибутов, скрытых или вовсе не указанных в профиле у пользователя.

Одной из самых популярных социальных сетей, в качестве ресурса для выявления демографических атрибутов, исследователи называют Twitter. Сервис позволяет публиковать сообщения (блоги) объемом до 140 символов.

Сложность сбора информации в Twitter для обучающей выборки заключается в отсутствии демографических атрибутов в профилях у пользователей. Но и в этом случае тексты небольшой длины имеют свои преимущества, которые были исследованы в 2013 году [3].

Актуальность данного исследования связана с заинтересованностью региона поддерживать высокий уровень подготовки будущих студентов, а также искать и привлекать одаренных старшеклассников в региональные вузы. В связи с этим возникает **проблема**, заключающаяся в отсутствии инструментов по выявлению одаренных старшеклассников. В настоящий момент в социальных сетях генерируются значительные потоки информации, которые

характеризуются высоким уровнем динамичности и масштабности, но они не используются вузами для решения задач по поиску и отбору талантливых абитуриентов.

На основе использования современных технологий анализа больших данных получаемая из социальных сетей информация, которая на первый взгляд представляется разрозненной, может быть распределена по огромному количеству критериев – как общих для отдельных групп пользователей, так и персонально-ориентированных. Эти критерии формулируются в соответствии с социальными и экономическими задачами, решаемыми как на национальном, так и на региональном уровне. Так, регионы, стремясь повысить качество системы высшего образования и расширить рынок предлагаемых образовательных услуг, нацелены на обеспечение высокого уровня подготовки будущих студентов и достижение значительных показателей поступления абитуриентов в региональные вузы.

Для достижения поставленных перед регионом целей в извлекаемых из социальных сетей данных необходимо находить закономерности, на основе которых можно выявить основные психологические и поведенческие характеристики целевой аудитории, ее интересы и профессиональные увлечения, а также определить особенности формирования единого портрета интересующего регион пользователя, т.е. будущего студента вуза.

Была выдвинута **гипотеза**, что по данным профиля социальной сети можно с некоторой вероятностью определить признаки одаренности старшеклассников и автоматизировать алгоритм прогнозирования признаков одарённости при обработке больших объёмов данных. Основаниями для исследования послужили современные научные представления об одаренности, изложенные в работах отечественных и зарубежных ученых: концепции одаренности Ю. Д. Бабаевой, Д. Б. Богоявленской [4], А. М. Матюшкина [5], Н. С. Лейтеса [6], многомерные модели одаренности К. Хеллера [7-8], Дж. Рензулли [9-10] и т.д.

В отечественных и зарубежных работах представлены результаты анализа данных социальных сетей для исследования и прогнозирования пола, возраста,

национальности, психологического типа пользователей, их политических предпочтений, но не одаренности [1, 3, 16, 41, 45 и др.].

Работа над диссертационным исследованием проводилась в рамках проекта Российского фонда фундаментальных исследований (номер проекта: 17-16-70004 РГНФ) «Исследование потенциала социальных сетей для выявления, привлечения и закрепления талантливой молодежи в региональных вузах на основе анализа больших данных». Основная цель проекта заключается в осуществлении высокотехнологичного анализа синтезированных данных социальных сетей и последующем использовании полученных результатов при моделировании механизмов привлечения абитуриентов в региональные вузы. В рамках данного проекта предполагается комплексный подход к решению проблемы, основанный на сочетании методов анализа больших данных, математического моделирования, сетевой визуализации, когнитивной лингвистики, педагогики и цифровой гуманитаристики, позволяющих синтезировать данные различных типов и провести научную интерпретацию полученных результатов [11].

Предметом исследования является феномен одарённости у старшеклассников. **Объект исследования** – прогнозирование признаков одарённости у старшеклассников по открытым пользовательским данным в социальной сети «Вконтакте».

Цель исследования – на основе имеющихся данных о старшеклассниках (результаты психологического тестирования и сведения из профиля в социальной сети «Вконтакте») разработать прогностическую модель, определяющую признаки одаренности по открытым пользовательским данным в социальной сети для потенциальных абитуриентов ТГУ.

Для решения поставленной цели необходимо было выполнить следующий перечень **задач**.

1. Изучить состояние российских и зарубежных исследований и практик по анализу данных в социальных сетях.

2. Подготовить для анализа исходные данные о старшеклассниках:

результаты психологического тестирования и сведения из профиля социальной сети «Вконтакте».

3. Выбрать и апробировать методы анализа данных для разработки прогностической модели выявления одарённых абитуриентов.

4. Выбрать и обосновать инструменты для анализа больших данных.

5. Разработать модель прогнозирования признаков одарённости старшеклассников и оценить её точность.

6. Автоматизировать алгоритм прогнозирования признаков одарённости.

Выпускная квалификационная работа имеет классическую структуру. В первой главе рассматриваются современные методы и инструменты сбора и анализа данных в социальных сетях, анализируется отечественный и зарубежный опыт использования алгоритмов машинного обучения для моделирования когнитивных и психологических характеристик личности, определяется феномен одарённости на основании концепций российских и зарубежных ученых, рассматривается проблема сочетания разных признаков одарённости и трудности их прогнозирования. Содержание первой главы обосновывает возможность прогнозирования признаков одарённости у старшеклассников по их профилю в социальной сети, определяет наиболее соответствующие цели диссертационного исследования методы и инструменты анализа.

Во второй главе представлены основные этапы разработки модели прогнозирования признаков одаренности у старшеклассников по данным из профиля в социальной сети «Вконакте»: описана структура исходных данных, алгоритм их подготовки для анализа, проведен выбор и обоснование программных инструментов, разработана прогностическая модель, проведена её апробация и оценка точности.

В приложениях представлены результаты комплексного профориентационного тестирования старшеклассников в ТГУ и исходные данные о подписках старшеклассников на тематические сообщества «ВКонтакте».

Глава 1. Выбор и обоснование методов и инструментов анализа данных

1.1 Data Mining. Актуальность и характерные особенности

С ростом информационных технологий многие организации начали применять различные методы сбора, обработки и хранения данных с целью их анализа для достижения коммерческих целей, а также для расширения влияния на рынке. Цифровой информации стало настолько много, что возможностей экспертов и ручной обработки уже недостаточно, чтобы проанализировать огромные массивы данных.

В настоящее время все больше ресурсов вкладывается в развитие направления по методам обработки и анализа данных. Разрабатываемые системы в этой области необходимы для минимизации усилий лица, принимающего решения, во время обработки информации, а также в настройке алгоритмов анализа. Большинство интеллектуальных систем не только решают задачи принятия решений в своем каноническом виде, но и находят скрытые закономерности, а также определяют причинно-следственные связи.

Термин Data Mining переводится как «извлечение информации» или «добыча данных» [12]. Анализ данных – это процесс анализа скрытых шаблонов данных в соответствии с различными аспектами классификации информации, которая собирается в хранилища данных, а для эффективной обработки используются алгоритмы интеллектуального анализа данных. Такой подход способствует более точному принятию бизнес-решений, что, в конечном итоге, позволит сократить издержки и увеличить доходы.

Когнитивные возможности человека ограничены к восприятию разнородной информации большого объема. Если не брать во внимание частные случаи, то способности индивида выявлять от трех взаимосвязей в небольших выборках не могут этого сделать.

Существует много общего между интеллектуальным анализом данных и статистикой. Большинство методов, используемых в Data Mining, могут быть

реализованы и в статистике в классическом ее представлении. Однако разница между этими методами все же присутствует.

Традиционные статистические методы требуют значительного взаимодействия с пользователем для корректировки модели. В результате статистические методы могут сложнее автоматизировать. Более того, эти методы часто неприменимы для очень больших наборов данных. Статистические методы основаны на проверке гипотез или нахождении корреляций, основанных на меньших репрезентативных выборках большей популяции. В то время как методы интеллектуального анализа данных подходят для больших наборов данных и их легче автоматизировать.

Повсеместное использование анализа данных ничем не ограничено. Data Mining актуален везде, где присутствуют хоть какие-то данные. На основе многих предприятий статистика показывает, что эффективность от использования Data Mining, в некоторых случаях, достигает 1000%. Например, известны сообщения об экономическом эффекте, в 10-70 раз превысившем первоначальные затраты от 350 до 750 тыс. долларов. Приводятся сведения о проекте в 20 млн. долларов, который окупился за 4 месяца. Другой пример – годовая экономия 700 тыс. долларов за счет внедрения Data Mining в сети универсамов в Великобритании [12].

Основой в области Data Mining является сбор релевантных данных, важных для бизнеса. Данные компании являются транзакционными, или метаданными. Транзакционные данные касаются повседневных операций, таких как продажи, инвентарь, стоимость и т.д. Эксплуатационные данные обычно не прогнозируются, а метаданные связаны с проектированием логической базы данных. Шаблоны и отношения между элементами данных предоставляют соответствующую информацию, что может увеличить доходы организации. Например, розничный гигант Wal-Mart передает всю свою информацию в хранилище данных. Эти данные могут быть легко доступны поставщикам, позволяющим им идентифицировать покупки покупателей. Они также могут генерировать шаблоны покупок, наиболее популярные заказы и другую

информацию, где используются методы интеллектуального анализа данных.

Повседневная деятельность аналитиков и руководителей также не обходится без использования технологии Data Mining. С помощью методов анализа данных различные компании могут получить существенные преимущества в конкурентной борьбе.

Data mining способствует извлечению информации из больших объемов данных. Например, планировщик города может использовать модель, которая предсказывает доход, основанный на демографии, чтобы разработать план постройки жилья для людей с низким доходом. Агентство по лизингу автомобилей может использовать модель, которая идентифицирует сегменты рынка для разработки рекламных кампаний, ориентированных на клиентов с высоким доходом.

1.1.1 Классификация Data Mining

Методы на основе Data Mining решают множество задач, с которыми сталкивается аналитик. Из них основными являются: кластеризация, поиск ассоциативных правил, классификация и регрессия. Их описание представлено следующим образом [13]:

1) Задача **классификации** сводится к построению модели для определения класса объекта на основе его атрибутов. Собирается коллекция записей, каждая запись с набором атрибутов. Одним из атрибутов будет атрибут класса, а цель задачи классификации – как можно точнее назначить атрибут класса новому набору записей. Классификация может использоваться в прямом маркетинге, для снижения маркетинговых затрат путем ориентации на конкретную группу клиентов, которые могут купить новый продукт. Используя имеющиеся данные, можно узнать, какие клиенты приобрели аналогичные продукты, а кто вообще его не покупал. Следовательно, покупка/не покупка принимает форму класса. После присвоения класса клиенту, может быть собрана демографическая

информация и информация об их жизни, с целью дальнейшей рассылки предложений и акций по электронной почте.

2) Как и в задаче классификации, **регрессия** позволяет определить по известным характеристикам объекта значение некоторого его параметра. Разница между классификацией и регрессией заключается в том, что, в данном случае, множество действительных чисел – это значение параметра, а не конечное множество классов как в задаче классификации. Например, модель может прогнозировать доход сотрудника на основе образования, опыта и других демографических факторов, таких как место пребывания, пол и т.д.

3) Задача **ассоциации**. Целью данной задачи является поиск ассоциативных правил (зависимостей или ассоциаций), которые часто возникают между событиями или объектами. Анализ ассоциации используется для управления товарами, рекламы, дизайна каталогов, прямого маркетинга и т. Д. Розничный торговец может идентифицировать продукты, которые обычно покупают покупатели, или даже найти клиентов, которые отвечают за продвижение таких же продуктов.

4) Задача **кластеризации** предполагает поиск отличающихся друг от друга групп (кластеров), а также их особенностей на всей совокупности анализируемых данных. Идентификация объектов в выборке осуществляется по принципу схожести друг с другом. Сходство можно решить на основе ряда факторов, таких как поведение при покупке, склонность к определенным действиям, географическое местоположение и т.д. Например, страховая компания может объединять своих клиентов в зависимости от возраста, места жительства, дохода. Подобная информация будет полезна для лучшего понимания клиентов и, следовательно, предоставления индивидуальных услуг.

5) **Последовательные шаблоны** – предполагают выявление закономерностей между связанными во времени событиями. Правило последовательности говорит, что через определенное время после события X наступит событие Y.

Перечисленные задачи по назначению делятся на описательные и

предсказательные.

Основная особенность в описательных задачах заключается в легкости и прозрачности результатов для восприятия человеком. Есть вероятность, что закономерности, которые были обнаружены в рамках одной исследуемой выборки, не проявят себя в другой, но и в таком случае результаты анализа будут полезны в частном случае. К такому виду задач относятся поиск ассоциативных правил и кластеризация.

Предсказательные задачи делятся на два этапа. В первую очередь, на основе выборки с заранее известными результатами строится модель. А далее прогнозируется результат уже на примере нового набора данных. В предсказательных задачах большую роль играет точность моделей. К такому виду задач относят задачи регрессии и классификации. А также, если полученные результаты используются в качестве прогноза, то сюда можно отнести поиск ассоциативных правил.

Задачи анализа данных также классифицируют по способам их решения на обучение с учителем и обучение без учителя.

В первом случае, для решения задачи необходимо выполнить несколько этапов. В первую очередь, используя любой алгоритм анализа данных строится модель, т.е. классификатор. Этот классификатор проходит процесс обучения. Другими словами, осуществляется качественная проверка работы модели и в случае, если результат неудовлетворительный, классификатору следует провести дополнительное обучение. Это продолжается до тех пор, пока приемлемый уровень качества не будет достигнут, либо станет понятно, что данный алгоритм некорректно работает с данными, а также есть вероятность, что данные не формализованы или не имеют четкой структуры. К такому типу задач относят задачи классификации и регрессии.

К задачам обучения без учителя относят построение описательных моделей, в которых необходимо выявлять закономерности из анализируемой выборки. Примером могут служить закономерности в покупках в большом магазине. Такие задачи решаются без первоначальных знаний об имеющихся

данных. К этой группе задач относятся поиск ассоциативных правил и кластеризация.

В задаче кластеризации исследуемая совокупность объектов делится на группы схожих объектов, которые называются кластерами. Как правило, решение задачи соотнесения множества элементов по кластерам называют кластерным анализом. Подобного рода задачи решаются в различных сферах деятельности, где необходим анализ статистических и экспериментальных данных.

Разница между классификацией и кластеризацией заключается в отсутствии выделенной зависимой переменной для решения последней задачи. В этом случае кластеризация относится к классу задач обучения без учителя. Задача актуальна в том случае, когда имеется мало информации о данных, т.е. в начале проведения исследования. Решение задачи кластеризации способствует лучшему понимаю анализируемых данных, и с этой точки зрения она является описательной задачей.

В ближайшем будущем технология Data Mining ориентирована на развитие в коммерческой деятельности. Программное обеспечение на основе анализа данных может стать таким же необходимым и повседневным, как sms-сообщения, электронная почта, и использоваться пользователями с целью поиска наиболее релевантной продукции по низкой цене.

Что касается будущего технологии Data Mining, то ее развитие прогнозируется в актуальной сфере на любом этапе развития человечества – медицине. Где, например, может быть разработан интеллектуальный анализ пациентов с помощью специального алгоритма для выявления и, возможно, лечения различных заболеваний. В этом случае нивелируется человеческий фактор, что способствует более точной постановке диагноза и последующего лечения.

Но кроме положительных особенностей в анализе данных присутствует и потенциальная опасность. Она заключается в том, что все больше информации в сети Интернет предоставляется в открытом доступе. А это, в свою очередь,

может способствовать ее утечке как по невнимательности, так и с помощью людей, которые обладают достаточными компетенциями для подобной деятельности.

Примером негативного опыта использования технологии Data Mining может служить самый крупный онлайн-магазин «Amazon», который оказался в центре скандала по патенту «Методы и системы помощи пользователям при покупке товаров». Данный патент подразумевает сбор персональной информации о пользователях ресурса. С помощью алгоритма анализа данных прогнозируются дальнейшие запросы пользователя на основании его текущих покупок. Основная цель этого патента заключается в получении как можно большего количества информации о клиентах, в том числе и частного характера.

Таким образом, сбор информации о частной жизни осуществляется не только взрослых, но и их детей, а это уже запрещено законодательством многих стран – для получения информации о несовершеннолетних необходимо разрешение родителей.

1.1.2 Анализ популярных сервисов по аналитике данных в социальной сети «ВКонтакте»

YouScan – первая и лидирующая система для профессионального мониторинга русскоязычных социальных медиа на основе искусственного интеллекта [14].

Сервис отслеживает упоминания ваших брендов, продуктов, конкурентов в блогах, форумах, социальных сетях (Facebook, «ВКонтакте»), Twitter и YouTube, и представляет результаты мониторинга в удобном аналитическом интерфейсе с функциями командной работы.

Платформа позволяет производить мониторинг социальных сетей, который помогает владельцам брендов анализировать отзывы потребителей, размещенные в сети Интернет, о своих продуктах и конкурентах.

Функция сервиса YouScan позволяет автоматически находить такие

вопросы по заданной тематике и показывать их в режиме реального времени в личном кабинете пользователя. Также появляется возможность выгрузить список авторов сообщений и запустить для них автоматические персонализированные рекламные сообщения в социальных сетях. Новая функция базируется на сложных алгоритмах компьютерного анализа текстов и содержит долю «искусственного интеллекта» на основе машинного обучения. Поиск ведется в социальных сетях «Одноклассники», «ВКонтакте» и Facebook.

В сводном отчете сервис может выводить основную статистическую информацию по выбранной теме с данными по числу зафиксированных сообщений, авторам, источникам, динамике и степени увлеченности (частота упоминания слов пользователем). Доступны фильтры по тем же параметрам и сообществам. Имеется экспорт аналитических отчетов в формате xls.

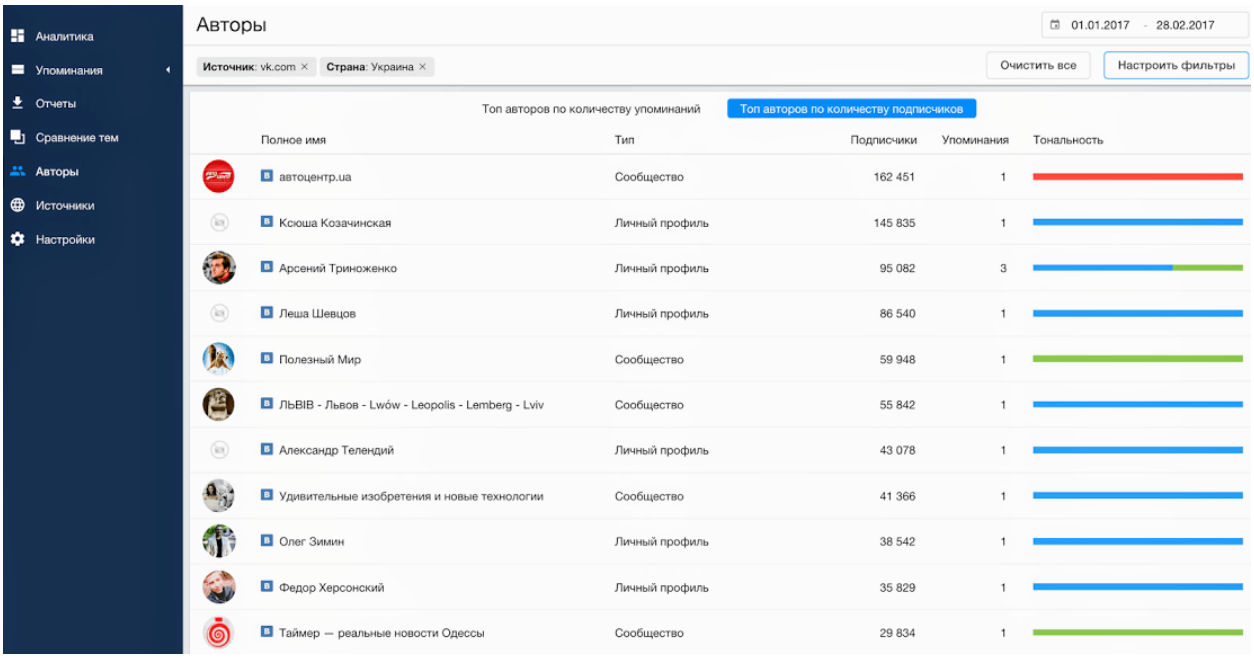


Рисунок 1.1 – Настройка поиска по данным «ВКонтакте»

IQBuzz – сканер социальных сетей. Обрабатывает информацию из Facebook, Twitter, «ВКонтакте», LiveJournal, LiveInternet, Youtube и т.д. Имеет функции для коллективной работы, автоматически определяет позитивные и негативные сообщения, контролирует дубликаты сообщений (часто встречаются

повторные репосты и ретвиты), предоставляет многофункциональный поиск по истории сообщений.

IQBuzz сочетает в себе как автоматизированный сбор данных, так и ручной анализ, предоставленный экспертами. К основным продуктам относятся: маркетинг, PR, бренд-менеджмент репутации, маркетинговые исследования, конкурентный анализ, исследования и разработка продуктов.

Ключевые особенности включают [15]:

– мониторинг в режиме реального времени;

– фильтрация спама и дубликатов;

– фильтрация результатов по типу медиа, языку, полу, возрасту, географии;

– демонстрация графиков и визуализации;

– анализ темы: анализ частоты упоминаний вокруг интересующей темы;

– профилирование и анализ влияния: сегментация рынка (например, распределение авторов по полу, возрасту, местоположению);

– мониторинг и анализ конкурентов;

– составление отчетов.

Бесплатно можно протестировать сервис в течение 7 дней.

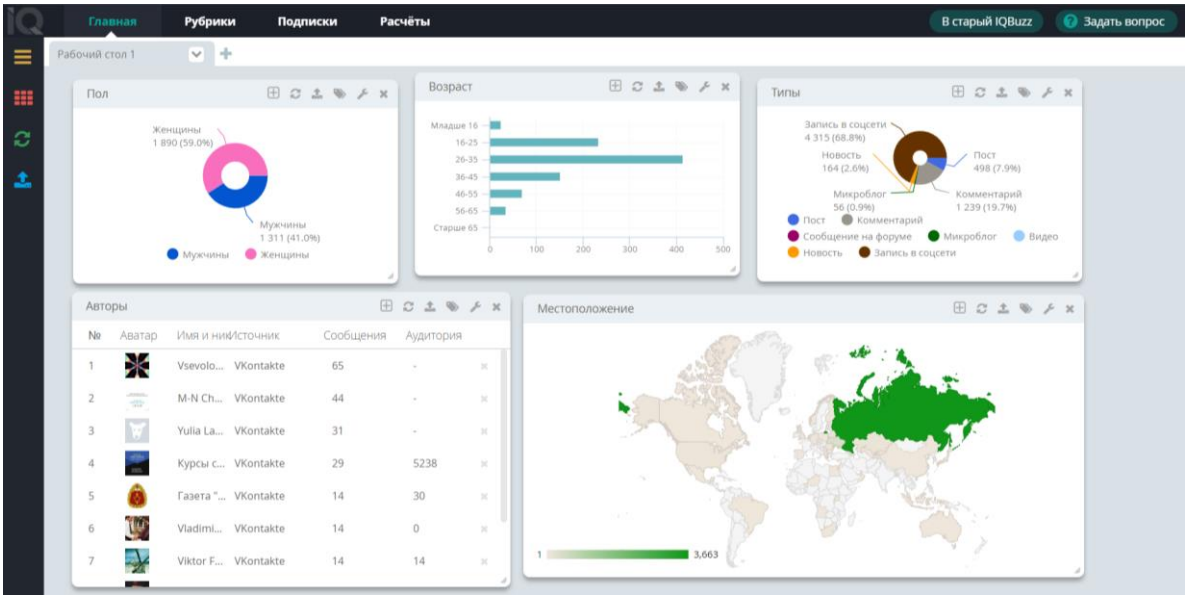


Рисунок 1.2 – Виджеты по ключевому словосочетанию «анализ данных»

Popsters – это онлайн-сервис для аналитики контента и интересов аудитории в конкретных сообществах социальных сетей. При этом один аккаунт в сервисе относится к одной социальной сети. С помощью этого инструмент можно анализировать данные по собственным страницам и страницам конкурентов, черпать информацию об активности, наиболее выгодных временных диапазонах публикаций, вовлеченности и оценке контента.

При этом сервис позволяет сортировать посты по лайкам и репостам. Popsters поддерживает «ВКонтакте», Facebook, Одноклассники, Instagram, Twitter, YouTube и т.д.

К возможностям сервиса относятся [16]:

– сортировка и фильтрация записей сообществ;

– одновременный анализ разных сообществ из разных социальных сетей;

– выгрузка отчетов в xlxs, pptx, pdf, анализ эффективности постов разного формата;

– анализ постов по содержанию хештегов и текста;

– статистика активности;

– графики и их выгрузка.

Сервис предоставляет 7 дней на бесплатное тестирование.

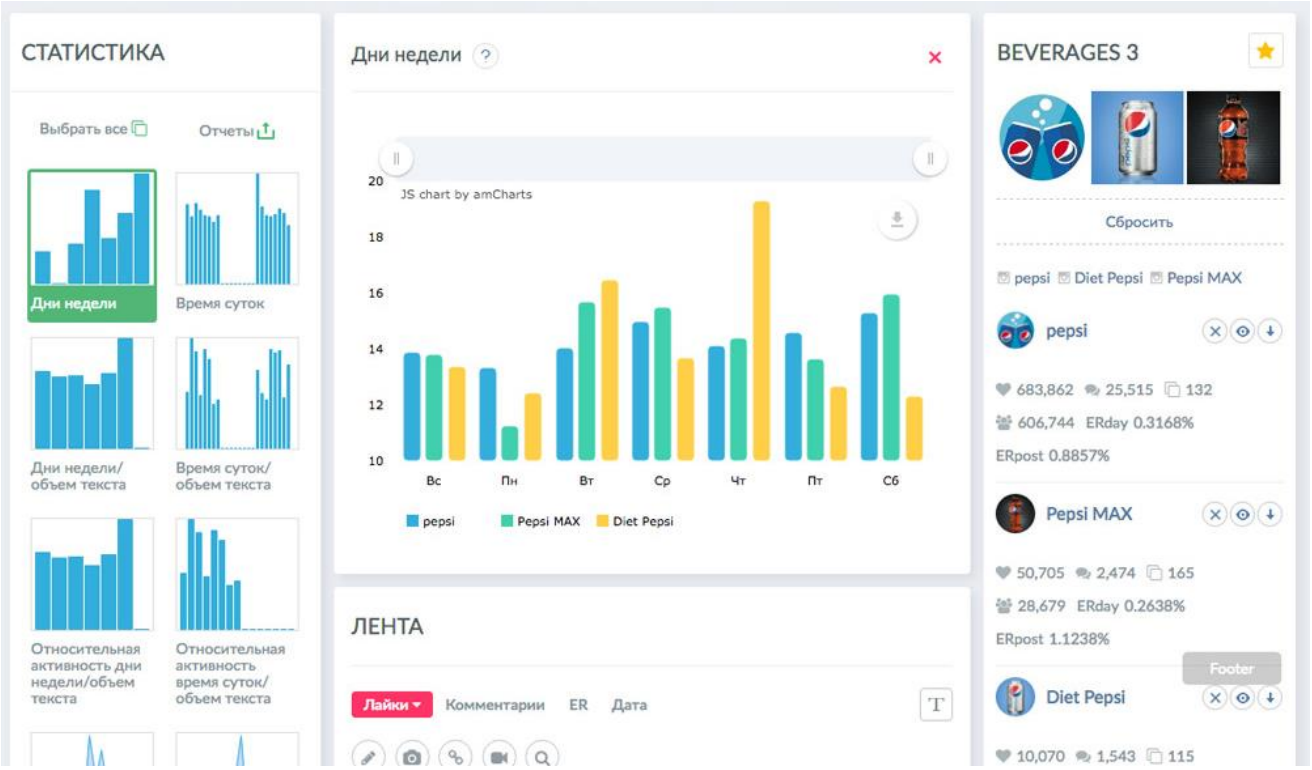


Рисунок 1.3 – Анализ сообществ на примере напитков Pepsi

1.2 Основные концепции одаренности отечественных и зарубежных ученых

В соответствии с представлениями концепции одаренности отечественных и зарубежных ученых, ее принято рассматривать как взаимодействие интеллекта, креативности и мотивационно-личностных особенностей. Одаренными считаются люди, обладающие высоким уровнем развития этих качеств или способные их развить и использовать в любой потенциально ценной деятельности.

В рамках когнитивного подхода, который является доминирующим и наиболее разработанным в изучении одаренности [11], под одаренностью понимается высокий уровень когнитивных способностей и/или интеллекта в целом.

Когнитивный компонент одаренности состоит из конвергентного когнитивного стиля – мышление, сознательно контролируемое и

ориентированное на поиск единственно правильного решения и дивергентного когнитивного стиля – мышление, направленное на создание множества разнообразных новых, нетривиальных решений «открытых», т. е. не имеющих единственного решения проблем. Дж. Гилфорд связывает креативность именно со способностями к дивергентному мышлению. Такие испытуемые обычно отличаются широтой интересов.

В соответствии с концепцией Э. П. Торренса, основными параметрами креативности являются беглость, гибкость, оригинальность и разработанность [17-18]. Беглость отражает способность к генерированию большого числа идей за единицу времени. Гибкость связана со способностью к выдвижению разных идей, переходу от одного аспекта проблемы к другому, использованию разных стратегий ее решения.

Оригинальность – ключевая характеристика креативности, способность к выдвижению необычных, неординарных идей, отличающихся от очевидных, общепринятых или твердо установленных.

В рамках личностного подхода к изучению одаренности главное внимание уделяется мотивационным, эмоциональным и другим личностным характеристикам одаренных людей, рассматриваемым в качестве основных определяющих факторов и условий развития и самореализации. Познавательная мотивация рассматривается как общая характеристика и структурный компонент личности одаренного человека [9, 19-27]. Такие люди отличаются стремлением к познанию, постижению сущности деятельности и получению интеллектуального удовлетворения от деятельности. Полная самореализация человека, с точки зрения К. Г. Юнга, является конечной целью его жизни, однако этой цели достигают только самые способные, высокообразованные и имеющие достаточный досуг люди [28]. А. Маслоу считает, что только одаренные люди, составляющие менее 1% населения, достигают реализации своего потенциала [21]. Мотивационно-личностные качества считаются прогностически благоприятными для существенных достижений в какой-либо деятельности.

Важной особенностью современного понимания одаренности является то,

что она рассматривается не как статическая, а как динамическая характеристика (Ю.Д.Бабаева, А.И.Савенков и другие) [29]. Одаренность реально существует лишь в движении, в развитии. Такое понимание привело к созданию теоретических моделей одаренности, в которые наряду с факторами, характеризующими потенциал личности, включены факторы среды. К таким, например, может быть отнесена модель Ф. Монкса – «мультифакторная модель одаренности» [30]. Ф. Монкс дополняет три уже традиционных пересекающихся «круга Дж. Рензулли» треугольником, обозначающим основные факторы микросреды: «семья», «школа, «сверстники».

1.3 Основные позиции к проблеме соотношения интеллекта и креативности

В сравнении с западно-европейской и американской науками, проблема в отечественной психологии, связанная с интеграцией креативности и интеллектуальных способностей не была так заметна.

Основная причина, вероятно, заключалась в отказе от теории и практики тестирования по системе IQ и в изучении интеллектуальных способностей в основном методом проблемных ситуаций, который прямо ориентировал исследователя на трактование интеллекта (в первую очередь мышления) как комплексной характеристики, рассматривающей креативность как необходимую составляющую.

За последние десять лет произошли изменения, связанные с возникновением проблемы дифференциации творческих и интеллектуальных способностей. Она стала предметом ряда специальных исследований отечественных специалистов (В. Н. Дружинин [31], А. М. Матюшкин [32], И. П. Ищенко [33], М. А. Холодная [34] и другие).

Структура творческой одаренности А.М. Матюшкина: в единой интегративной структуре одаренности выделяются следующие компоненты:

– доминирующая роль познавательной мотивации;

– исследовательская творческая активность, выражающаяся в обнаружении нового, в постановке и решении проблем;

– возможность достижения оригинальных решений;

– возможность прогнозирования и предвосхищения; способность к созданию идеальных эталонов, обеспечивающих высокие эстетические, нравственные, интеллектуальные оценки.

Российский психолог В. Н. Дружинин, анализировал подходы большинства отечественных и зарубежных авторов. В проблеме соотношения интеллекта и креативности он выделил три основных позиции.

1) Отказ от какого бы то ни было разделения этих функций (большинство отечественных ученых; из известных зарубежных исследователей Г. Айзенк);

2) Строится на утверждении, что между интеллектом и креативностью существуют пороговые отношения: для проявления креативности нужен интеллект не ниже среднего;

3) Интеллект и креативность – независимые способности. При максимальном снятии регламентации деятельности в ходе тестирования креативности результаты ее измерения у учащихся не зависят от уровня их интеллекта.

Исходя из результатов, можно предположить, что у специалистов нет единства. Эта ситуация весьма характерна для современной психологической науки в целом.

1.4 Алгоритмы для решения задач обработки данных

В сфере информационных технологий, чтобы описать метод решения задачи, который можно будет реализовать в выбранной среде программирования, используется понятие «алгоритм обработки данных».

В процессе решения задачи, вне зависимости от сферы применения, разработка алгоритма играет важную роль. Прежде чем разрабатывать алгоритм для задачи в реальных условиях, следует оценить степень его сложности,

определить ограничения входных параметров, а также осуществить декомпозицию на подзадачи.

Должна отсутствовать привязка алгоритма к конкретной реализации. В связи с тем, что существует большое разнообразие средств программирования, совместимостей по платформе, требований к ресурсам компьютера, разница в эффективности результатов алгоритмов будет существенна.

В некоторых средах программирования уже встроены большинство библиотек, которые реализуют основные алгоритмы обработки данных (например, в программный пакет Anaconda входят такие библиотеки как: numpy, scipy и т.д.).

Написанные программы должны быть мобильными и актуальными, соответственно, не следует их строго описывать с привязкой к процедурной реализации среды. Поэтому, в первую очередь, необходимо определиться с выбором метода решения на основании специфики задачи.

Метод обработки данных определяется не только за счет сложности имеющейся задачи. Также важно учитывать частоту использования разработанного кода. К примеру, если обращение к реализации редкое, то простые алгоритмы будут предпочтительнее, несмотря на возможное увеличение времени обработки данных.

Результатом разработок и исследовании в течение десятков лет были созданы основные алгоритмы обработки данных. И в настоящее время они также часто применяются в решении различных задач.

К базовым алгоритмам процедурного программирования можно отнести [35]:

– алгоритмы работы со структурами данных. Являются фундаментом для определения базовых принципов и методологии, которые используются для сравнения, анализа и реализации алгоритмов. К таким структурам относятся: абстрактные типы данных, очереди и стеки, связные списки и строки, а также деревья;

– алгоритмы сортировки используются для упорядочения файлов или

массивов. К ним относятся: задачи слияния и выбора, очереди по приоритету;

– алгоритмы поиска осуществляют поиск объектов в большой совокупности элементов. Среди алгоритмов рассматриваемого типа выделяют методы поиска с использованием деревьев – это сбалансированные деревья, хеширование, а также методы, подходящие для обработки файлов с большим объемом информации;

– алгоритмы на графах актуальны для решения важных и сложно структурированных задач. Суть поиска разрабатывается и применяется к задачам связности: задаче о паросочетаниях, о потоках в сетях, поиска кратчайшего пути. Функция, которая лежит в основе унифицированного подхода к этим алгоритмам, основывается на абстрактном типе данных очереди по приоритету;

– алгоритмы обработки строк представляет собой совокупность методов обработки строк (последовательностей символов). В данном случае осуществляется синтаксический анализ посредством поиска по строке и сравнения с эталоном. К этому виду алгоритмов относят и технологию сжатия файлов.

Все эти алгоритмы используются во многих задачах и очень часто формируют связку для достижения результата.

1.5 Исследования на основе методов машинного обучения

Под машинным обучением подразумевается область знаний, в которой алгоритмы обладают способностью обучаться [36]. Методы машинного обучения используют в том случае, когда имеются сложные задачи (например, распознавание речи, текста), для которых алгоритм решения неявный, т.е. присутствует сложность в его разработке. В этом случае можно предпринять обучение компьютера для решения подобного рода задач.

А. Л. Самуэль в 1959 г. [37], создатель самообучающейся компьютерной игры в шашки, один из первых, кто упомянул термин «машинное обучение». Суть обучения заключалась в демонстрации поведения компьютером, которое не

было заложено в него «явно». Однако эта теория была опровергнута, так как не понятно, что подразумевается под наречием «явно».

Более точное определение дал намного позже Т. М. Митчелл: «компьютерная программа обучается на основе опыта E по отношению к некоторому классу задач T и меры качества P, если качество решения задач из T, измеренное на основе P, улучшается с приобретением опыта E» [38].

Фаза обучения может предшествовать фазе работы алгоритма (например, детектирование лиц на фотокамере), но может иметь место обратная ситуация: обучение может проходить в процессе функционирования самого алгоритма (например, определение спама).

В настоящее время машинное обучение имеет многочисленные сферы деятельности, такие, как компьютерное зрение, распознавание речи, компьютерная лингвистика и обработка естественных языков, медицинская диагностика, биоинформатика, техническая диагностика, финансовые приложения, поиск и рубрикация текстов, интеллектуальные игры, экспертные системы и другие.

Анализ социальных данных стремительно набирает популярность во всем мире благодаря появлению в 1990-х годах онлайновых сервисов социальных сетей (LiveJournal, Facebook, Twitter, YouTube) [39-40].

Специалисты из исследовательских центров и компаний по всему миру используют данные социальных сетей для моделирования социальных, экономических, политических и других процессов от персонального до государственного уровня с целью разработки механизмов воздействия на эти процессы, а также создания инновационных аналитических и бизнес-приложений и сервисов.

Так, например, в ИСП РАН (Институт системного программирования им. В.П. Иванникова РАН) разработали стек технологий для анализа пользовательских данных из социальных сетей. Особое внимание уделяется задачам, методам и приложениям анализа сетевых (социальные связи между пользователями) и текстовых (сообщения и профили пользователей) данных:

определение демографических атрибутов пользователей, поиск описаний событий в корпусах сообщений, идентификация пользователей различных сетей, поиск сообществ пользователей и измерение информационного влияния между пользователями [41]. В рамках машинного обучения производится построение модели классификации с использованием онлайнового пассивно-агрессивного алгоритма [42].

Для решения различных задач при анализе социальных сетей используют также стохастические модели. Основная идея вероятностных (стохастических) моделей состоит в том, что каждая социальная сеть может быть рассмотрена как реализация случайного двумерного бинарного массива. Так как элементы этого массива являются зависимыми случайными величинами, то можно анализировать структуру зависимостей между соответствующими участниками социальной сети, находить вероятности существования определенных связей и получать оценки различных параметров сети.

Применение статистических моделей в анализе социальных сетей приведено в исследовании Дженсена Д. из Массачусетского университета, где использовались модели коллективной классификации для выявления корреляции между заголовком веб-страницы и названиями гиперссылок, содержащихся в ней. В данной работе также предлагается применять методы машинного обучения и анализа данных для вычисления относительной автокорреляции, плотности связей и некоторых других характеристик сети [43].

Наивный байесовский классификатор представляет собой одну из самых простых реализаций алгоритмов классификации. Его особенность в том, что все рассматриваемые признаки независимы друг от друга.

В его основе лежит теорема Байеса. Данный алгоритм используется в работах [44-46] для определения пола. Его главным преимуществом является поддержка онлайн-обучения, т.е. модель классификатора не производит дополнительный расчет при добавлении новых объектов в обучающую выборку, а происходит обновление при наличии новых данных.

В случае разделения выборки на два класса используют линейный

классификатор. Метод опорных векторов (SVM – Support Vector Machine) является один из самых популярных алгоритмов для обучения линейного классификатора.

Суть метода заключается в поиске разделяющей гиперплоскости с максимальным зазором до объектов классов [45, 47-49].

В работе [50] решается задача прогнозирования пола пользователей на сайте Youtube на основе двух источников информации: комментариев под видео и социальная среда, описываемая графом пользователи-видео. В данном исследовании пол пользователя предсказывается с точностью более 90%. Также в работе анализируется изменение точности прогноза по разным возрастным группам.

В исследовании, посвященном обзору методов построения социально-демографических профилей пользователей сети Интернет [51], анализируются демографические атрибуты, такие как: возраст, пол, район проживания, религиозные и политические взгляды, отношения с людьми и т.д. Описываются различные методы обработки этих атрибутов на примере машинного обучения с учителем, а для снижения признакового пространства используются алгоритмы выбора признаков. Также в статье исследуются алгоритмы классификации и регрессии, показатели оценки и функции выбора.

Предсказания индивидуальных атрибутов и предпочтений пользователей могут быть использованы для улучшения множества продуктов и услуг. Например, цифровые системы и устройства (интернет-магазины) могут быть спроектированы таким образом, чтобы корректировать поведение клиентов, чтобы наилучшим образом соответствовать профилю каждого пользователя. В работе [52] демонстрируется, что широкий спектр личностных качеств людей, начиная от сексуальной ориентации и заканчивая интеллектом, можно автоматически и точно определить, используя поставленные ими лайки в социальной сети Facebook. Сходство между лайками на Facebook и другими широко распространенными типами цифровых записей, такими как: истории просмотров, поисковые запросы или истории покупок, говорят о том, что

вероятность выявления атрибутов пользователей вряд ли будет ограничена только лайками. Вывод, основанный на наблюдениях за поведением пользователей в социальных сетях, может способствовать формированию новых исследований в области человеческой психологии.

С другой стороны, предсказуемость отдельных атрибутов из цифровых записей пользователей может иметь и отрицательные последствия, поскольку эта информация может быть применена к большому числу людей без получения их согласия. Коммерческие компании, правительственные учреждения или друзья из Facebook могут использовать программное обеспечение для определения таких атрибутов как интеллект, сексуальная ориентация или политические взгляды, которыми человек, возможно, и не собирался делиться. Учитывая тенденцию постоянного роста цифровых следов, которые пользователи оставляют за собой в сети, становится трудно контролировать, какие из их атрибутов предоставляются в открытом доступе.

В связи с этим существует риск того, что растущее осознание цифрового воздействия может негативно сказаться на восприятие людьми цифровых технологий, а также уменьшить их доверие к онлайн-сервисам или даже полностью отказаться от использования современных технологий [52].

Во многих исследованиях по определению автора анализируются количественные параметры его текстов и доказывается существование у каждого из них (не только у профессиональных писателей) индивидуального «языкового отпечатка», формирующего его идиостиль. Этот «языковой отпечаток» получил название «stylome». А люди, имеющие какой-либо общий признак (принадлежность к одному полу, сходные психологические характеристики и т.д.), имеют схожие отпечатки. В связи с чем становится возможной по тексту разработка методик диагностирования личностных характеристик [53].

На основе методов машинного обучения в работе [54] был проведен анализ текстов пользователей социальных сетей. Полученные результаты доказывают наличие у людей индивидуального «языкового отпечатка» – «stylome». Причем особенности языка, словарного запаса обладают более ценной информацией по

сравнению с синтаксическими особенностями автора. И даже при анализе только синтаксической составляющей точность прогноза достаточно высока – 88,7%.

На основе выявления эмоциональной окраски сообщений в исследовании [55] был проведен анализ методов по определению тональности текстовых сообщений в социальной сети «ВКонтакте», а также разработан прототип веб-сервиса, позволяющий анализировать вводимые пользователем сообщения.

Тональность текста позволяет выявить эмоционально окрашенную лексику и проанализировать оценку автора по отношению к объектам, речь о которых идет в тексте. Тональность сообщения определялась по трем категориям: позитивная, нейтральная и негативная. В работе использовались такие методы автоматического извлечения аспектов, как: метод обучения без учителя – метод распространения (bootstrapping) и статистический метод C-value. Данный метод определяет значимость всех n-грамм, входящих в некоторое множество текстов.

Проведенное тестирование системы с применением технологии стемминга и без нее, с различными по длине и эмоциональной окраске сообщениями выявило, что точность работы алгоритма пропорциональна длине сообщения и составляет до 89% при анализе сообщений длиной до 50 символов.

Помимо того, что социальные сети являются посредниками в коммуникациях меду людьми, они еще стали и объектом множественных исследований, обладая ценной информацией для изучения и моделирования поведения пользователя.

В исследовании [56] определяются скрытые показатели личности в профилях в Facebook. Исследуется возможность моделирования личности пользователя на основе предлагаемого набора функций, извлеченных из данных социальной сети Facebook. В работе были протестированы популярные алгоритмы машинного обучения: метод опорных векторов и деревья решений. Их точность составила до 78%.

Исходя из проведенного анализа отечественных и зарубежных практик, сделан вывод, что применение методов машинного обучения позволит с некоторой степенью точности (варьируется, в зависимости от конкретного

психологического качества) идентифицировать пользователей с высоким уровнем развития психологических качеств, значимых для определения одаренности. Применение методов машинного обучения позволяет учитывать большое количество испытуемых и признаков, устанавливать связи между уровнем развития учащегося и его принадлежностью к определенным сообществам социальной сети.

Таким образом, в рамках магистерской диссертации для анализа профилей старшеклассников в социальной сети «ВКонтакте» было принято решение разработать собственную модель по обработке подписок у пользователей, базирующуюся на формулах из статистического анализа. Это обусловлено низкой эксплуатацией мощностей компьютера, а также открытый исходный код позволяет дополнять и улучшать модель.

Глава 2. Разработка модели прогнозирования наличия признаков одаренности по данным из профиля в социальной сети «Вконакте»

2.1 Описание и построение модели

Для проведения исследования была предоставлена информация по комплексному тестированию учащихся старших классов средних общеобразовательных учебных заведений г. Томска. Использовался метод психологического тестирования испытуемых, который является классическим для диагностики одаренности. Была применена компьютеризированная методика «Профориентация», предназначенная для учащихся старших классов образовательных учреждений разного вида [57].

Совокупность этих субтестов выявляют личностные особенности и профессиональную склонность, позволяют определиться с профессиональной направленностью и выбрать подходящий вид деятельности в будущем. В данном исследовании эта информация была необходима с целью построения модели весовых коэффициентов по сообществам, на основании которых и проводился дальнейший расчет абитуриентов 2018 года.

Из результатов более 100 субтестов квалифицированным психологом были отобраны 15 (рис. 2.1), наиболее значимых для определения одаренности учеников в каждом из признаков: интеллект, креативность, мотивация, личность.

Рисунок 2.1 – Классификация субтестов

Исходные данные о старшеклассниках (результаты психологического тестирования и данные о подписках на тематические сообщества) «ВКонтакте» представлены в приложении А и приложении Б.

На первом этапе анализа для каждого учащегося (выборка в 2225 человек, по которым проводилось тестирование) был рассчитан вес по всем 4 признакам:

$$W_j = \frac{\sum_{i=1}^{n} rankdata(mas_i)}{n}, \quad (1)$$

где W_j – коэффициент, характеризующий вес j-го ученика по i-му признаку;

mas_i – результаты по тестированию i-го ученика;

$rankdata$ – функция вычисления рангового индекса списка;

n – количество абитуриентов.

Полученные результаты были распределены на классы от 1 до 3 по каждому признаку (1 класс –высокий уровень: 75 перцентиль и выше, 2 – средний: от 26 до 74 перцентиля, 3 – низкий: 25 перцентиль и ниже) в зависимости от полученных баллов по многопрофильному тестированию (таблица 2.1). На основании этой выборки осуществлялось обучение модели и расчет коэффициентов по сообществам для дальнейшего поиска одаренных абитуриентов на 2018-2019 год обучения.

Таблица 2.1 – Расчет признаков на выборке, по которой проводилось тестирование

ID пользователя	IM	CM	MM	PM	Интеллект	Креативность	Мотивация	Личность
xxx918847	0,52	0,79	0,49	0,46	2	1	2	2
xxx713102	0,12	0,56	0,68	0,44	3	2	1	3
xxx569929	0,54	0,72	0,14	0,4	2	1	3	3
xxx462247	0,1	0,41	0,5	0,4	3	2	2	3
xxx111863	0,27	0,48	0,68	0,44	3	2	1	3
xxx204444	0,39	0,52	0,5	0,46	2	2	2	2
xxx204444	0,58	0,45	0,68	0,47	2	2	1	2
xxx453333	0,41	0,51	0,33	0,39	2	2	3	3
xxx734539	0,38	0,59	0,68	0,46	2	2	1	2
xxx024405	0,18	0,78	0,14	0,47	3	1	3	2
xxx673167	0,66	0,45	0,68	0,46	1	2	1	2
xxx758913	0,22	0,47	0,5	0,38	3	2	2	3
xxx727527	0,52	0,19	0,86	0,34	2	3	1	3
xxx887920	0,04	0,75	0,14	0,39	3	1	3	3
xxx752397	0,74	0,58	0,68	0,48	1	2	1	2
xxx869842	0,56	0,73	0,14	0,43	2	1	3	3
xxx083853	0,85	0,66	0,14	0,42	1	1	3	3
xxx264463	0,78	0,22	0,69	0,26	1	3	1	3
xxx292641	0,48	0,61	0,49	0,35	2	2	2	3

Что касается информации по сообществам, то рассчитанные классы в таблице 2.1 были использованы для построения матрицы, где в строке указывается id сообщества, а в столбцах, на пересечении, количество учащихся, подписавшихся на это сообщество и проранжированных по результатам тестирования (по классам) (таблица 2.2).

Таблица 2.2 – Фрагмент матрицы по анализу сообществ

Group_id	I1	I2	I3	C1	C2	C3	M1	M2	M3	P1	P2	P3
26762265	187	322	187	183	330	183	244	233	219	184	344	168
73375377	166	307	153	157	289	180	235	195	196	181	329	116
60130670	168	315	182	174	318	173	247	212	206	197	318	150
60442626	122	239	138	131	264	104	163	172	164	133	256	110
33064682	167	263	164	169	273	152	206	198	190	181	297	116
57846937	108	174	116	111	195	92	146	135	117	114	181	103
60981357	95	182	109	110	180	96	139	110	137	109	200	77
45745333	103	162	100	88	183	94	144	116	105	110	176	79
93250065	107	184	104	100	181	114	145	131	119	113	185	97

30602036	73	121	83	69	131	77	103	97	77	85	126	66

Для каждого класса, в зависимости от признака, найдена его доля:

$$p_{ij} = \frac{Class_{ij}}{\sum_{j=1}^{n} Class_{ij}}, \quad (2)$$

где p_{ij} – доля учеников в i-ом признаке j-го уровня, n=3.

Далее был рассчитан условный коэффициент, показывающий преобладание в сообществе одаренных учеников: отрицательное значение – в сообществе преобладают учащиеся с низким результатом тестирования, положительное значение – в сообществе преобладают учащиеся с высоким результатом тестирования (таблица 2.3):

$$pm_i = p_{i1} - p_{i3}, \quad (3)$$

где pm_i – разность между высоким и низким классом i-го признака.

Таблица 2.3 – Сообщества с весовыми коэффициентами по каждому признаку

group_id	**IM**	**CM**	**MM**	**PM**
26762265	0,00	0,00	0,04	0,02
73375377	0,02	-0,04	0,06	0,10
60130670	-0,02	0,00	0,06	0,07
60442626	-0,03	0,05	0,00	0,05
33064682	0,01	0,03	0,03	0,11
57846937	-0,02	0,05	0,07	0,03
60981357	-0,04	0,04	0,01	0,08
45745333	0,01	-0,02	0,11	0,08
93250065	0,01	-0,04	0,07	0,04
30602036	-0,04	-0,03	0,09	0,07
40567146	0,02	0,02	0,02	0,08
56106344	-0,01	0,01	-0,02	-0,02
44781847	-0,11	0,00	0,03	0,02
101826369	0,03	0,06	0,01	-0,02
31976785	0,11	0,05	-0,08	0,13
75149440	-0,07	0,06	0,04	0,05
135209264	0,07	0,00	0,01	0,03
30637940	0,09	-0,11	0,15	0,05
111119875	-0,10	0,04	0,13	0,13
25980040	-0,02	-0,03	0,05	0,11

На основании этих результатов можно определить преобладание у ученика одного из четырех признаков (интеллект, креативность, мотивация, личность) и

дающий вес каждому сообществу:

Просуммировав и нормализовав веса всех сообществ, на которые подписан ученик, был рассчитан коэффициент для каждого учащегося. Ранжируя результаты, получается список потенциальных абитуриентов с высоким показателем по какому-либо из признаков.

$$SM_j = \frac{\sum_{i=1}^{n} pm_i}{n}, \quad (4)$$

где SM_j – коэффициент, характеризующий j-го ученика, подписанного на i-ые сообщества;

n – количество маркерных подписок у абитуриента.

Результирующая информация по построенной модели представлена в таблице 2.4.

Таблица 2.4 – Фрагмент итоговых результатов по анализу

ID	**IM**	**Классификация**	**Количество маркерных сообществ**	**Общее количество сообществ**
xxx882243	0,01192	2	44	67
xxx522244	-0,01508	3	202	345
xxx833735	-0,01236	3	13	28
xxx082638	0,00547	2	17	18
xxx768591	0,01694	1	2	13
xxx864080	-0,00127	2	4	25
xxx004946	0,01415	2	10	22
xxx031787	0,03201	1	31	38
xxx682835	0,0204	1	53	82

Таким образом, была построена модель, с помощью которой удалось определить взаимосвязь между способностями старшеклассников и их подписками на сообщества в социальной сети «ВКонтакте».

2.2 Выбор средств разработки

Существует огромное количество прикладных программ, которые позволяют осуществлять расчеты по анализу данных. Каждая из них имеет свои преимущества и недостатки, а используются они в зависимости от поставленной

цели.

Каждый год портал KDnuggets [58] проводит опрос среди своих пользователей, пытаясь выяснить, какие инструменты программирования используют специалисты в той или иной области. Наиболее популярные ресурсы представлены на рис. 2.2. В данном опросе присутствуют как языки программирования, так и библиотеки, программные пакеты, приложения.

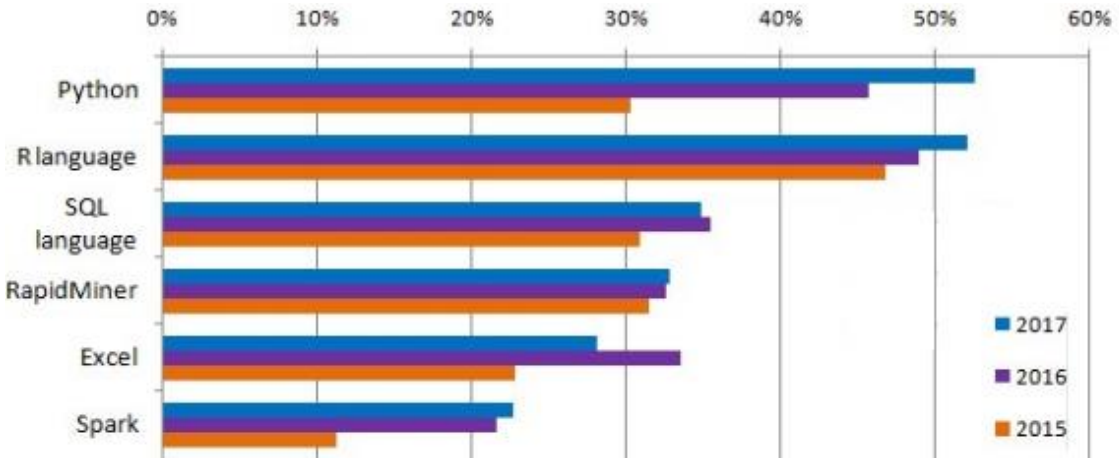


Рисунок 2.2 – Востребованные ресурсы для анализа данных за 2015-2017 гг., портал KDnuggets

Лидирующие позиции делят между собой 2 языка программирования: R и Python. Оба языка появились в конце XX века и уже успели приобрести популярность среди специалистов по анализу данных и разработчиков программного обеспечения. Если в самом начале R использовался только в академической среде, то с ростом интереса к Data Science (наука о данных) он стал популярен и в коммерческих областях применения.

Но за последние годы для языка программирования Python было создано множество инструментов для анализа данных, которые позволили обеспечить существенную конкуренцию R.

Плюсы и минусы языков программирования R и Python представлены следующим образом [59]:

Преимущества R:

– язык создавался специально для анализа данных: запись конструкций языка понятна многим специалистам в области;

– многие функции, необходимые для анализа данных, являются встроенными функциями языка. Проверка статистических гипотез зачастую занимает лишь несколько строк кода;

– установка IDE (RStudio) и необходимых пакетов обработки данных предельно упрощена;

– удобный репозиторий пакетов и обилие готовых тестов практически под все методы Data Science и машинного обучения;

– эффективная работа с векторами и матрицами.

Недостатки R:

– низкая производительность. Однако в системе присутствуют пакеты, позволяющие повысить скорость работы (pqR, FastR и т. д.);

– специфичность в сравнении со стандартными языками программирования, так как язык узкоспециализированный (например, индексация векторов начинается вместо нуля с единицы);

– большая часть кода на R написана людьми, не знакомыми с программированием. Кроме того, не все пользователи следуют рекомендациям по оформлению программного кода;

– R инструмент для статистики и соответствующих независимых приложений, но вызывает проблемы в использовании в тех областях, где традиционно применяются языки общего назначения;

– имеется возможность выполнить один и тот же функционал разными способами. Синтаксис для решения некоторых задач не совсем очевиден;

– в силу большого количества библиотек, документация некоторых менее популярных из них нельзя считать полной.

Преимущества Python:

– универсальный многоцелевой язык: можно осуществить не только обработку данных, но также их поиск и использование результата обработки в веб-приложении;

– Python является одним из тех языков, которые подходит на роль первого языка программирования. В случае потери интереса к Data Science приобретенные навыки пригодятся в других прикладных областях – как при использовании самого Python, так и при освоении родственных языков.

Недостатки Python:

– отсутствие общего репозитория и нехватка альтернатив для многих библиотек R. Однако ситуация значительно улучшилась за последние годы: аналитики, использовавшие ранее несколько различных языков для разных задач, отмечают, что набор инструментов сместился за последние два года в сторону библиотек на Python. Кроме того, вхождение новичков облегчают такие сборки как Anaconda;

– Python – язык с динамической типизацией. Это существенно ускоряет разработку программ, но и усложняет поиск некоторых трудно отслеживаемых ошибок, связанных с неправильным присваиванием различных данных одним и тем же переменным.

Во многих образовательных учреждениях Python занимает позиции первого языка для обучения программированию, вход многих новичков в Data Science со стороны этого языка становится проще, чем дополнительное изучение основ R. Поэтому, в качестве инструмента для решения поставленной задачи, был выбран перспективный в настоящее время язык программирования Python, а именно – программный пакет Anaconda.

2.2.1 Описание программного продукта Anaconda

Anaconda – это открытый пакетный менеджер и дистрибутив языков программирования Python и R. Он широко используется для обработки объемных данных, научных вычислений и прогностического анализа. Anaconda

предлагает набор пакетов с открытым исходным кодом, который на данный момент включает в себя более 720 экземпляров [60].

В основу пакета входит интерпретатор последней версии Python (по умолчанию), графический интерфейс Anaconda Navigator (рис. 2.3) и менеджер пакетов conda. Остальные приложения можно установить при необходимости.

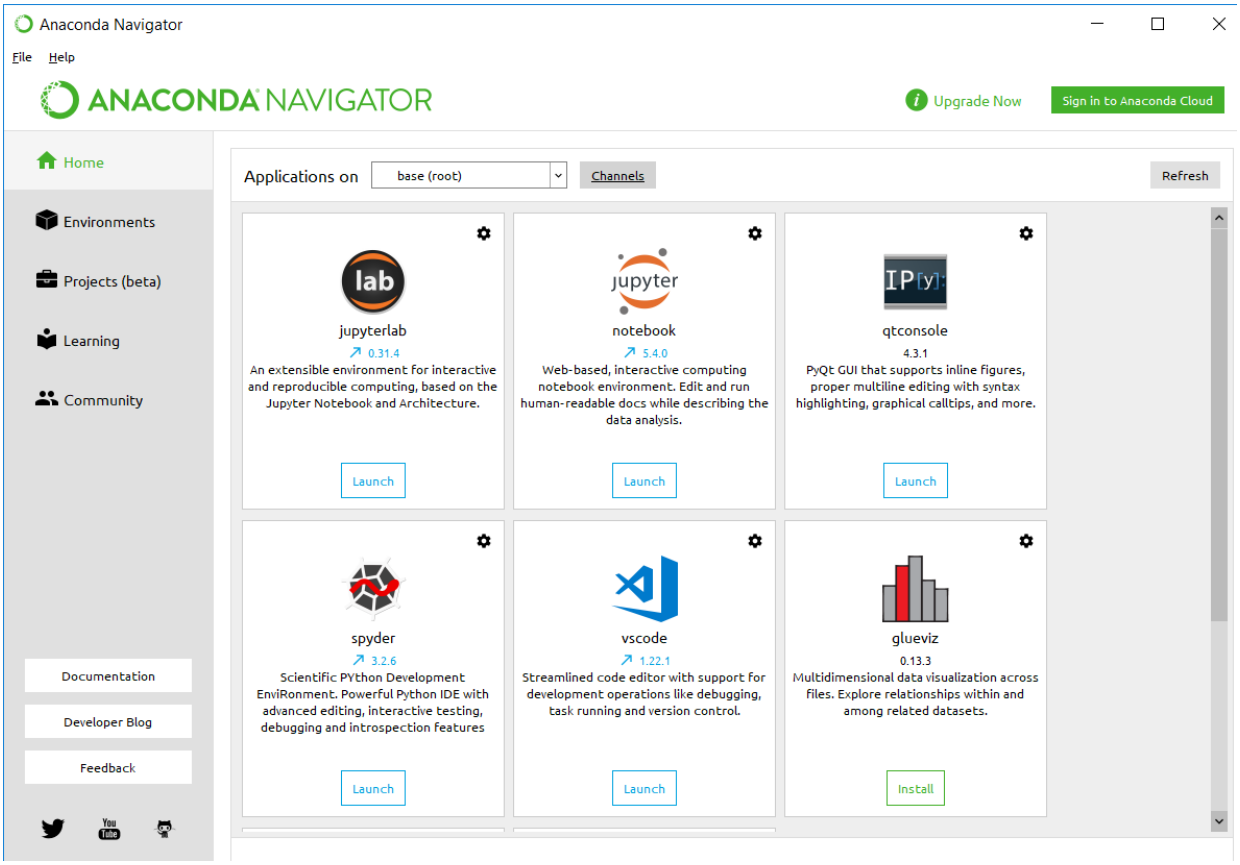


Рисунок 2.3 – Главное окно Anaconda Navigator

Исходя из представленного исследования на портале KDnuggets за основу анализа был взят язык программирования Python 3.6, входящий в состав программного пакета Anaconda и включающий в себя более 250 библиотек, что и явилось причиной выбора именно этого приложения.

Также в программный пакет Anaconda входит большое количество других программ, позволяющих работать с данными. В частности, для данного исследования необходимо было установить интегрированную среду разработки Spyder 3 (рис. 2.4). Spyder (ранее Pydee) – свободная среда разработки для Python.

Она кроссплатформенна и доступна для Windows, Linux и MacOS. Название Spyder расшифровывается как Scientific Python Development Environment, то есть научная среда разработки для Python [61].

Выбор именно этой среды разработки связан с низкими требованиями к ресурсам компьютера, а также с возможностью импорта и экспорта переменных в любой момент времени, что ускоряет процесс обработки данных.

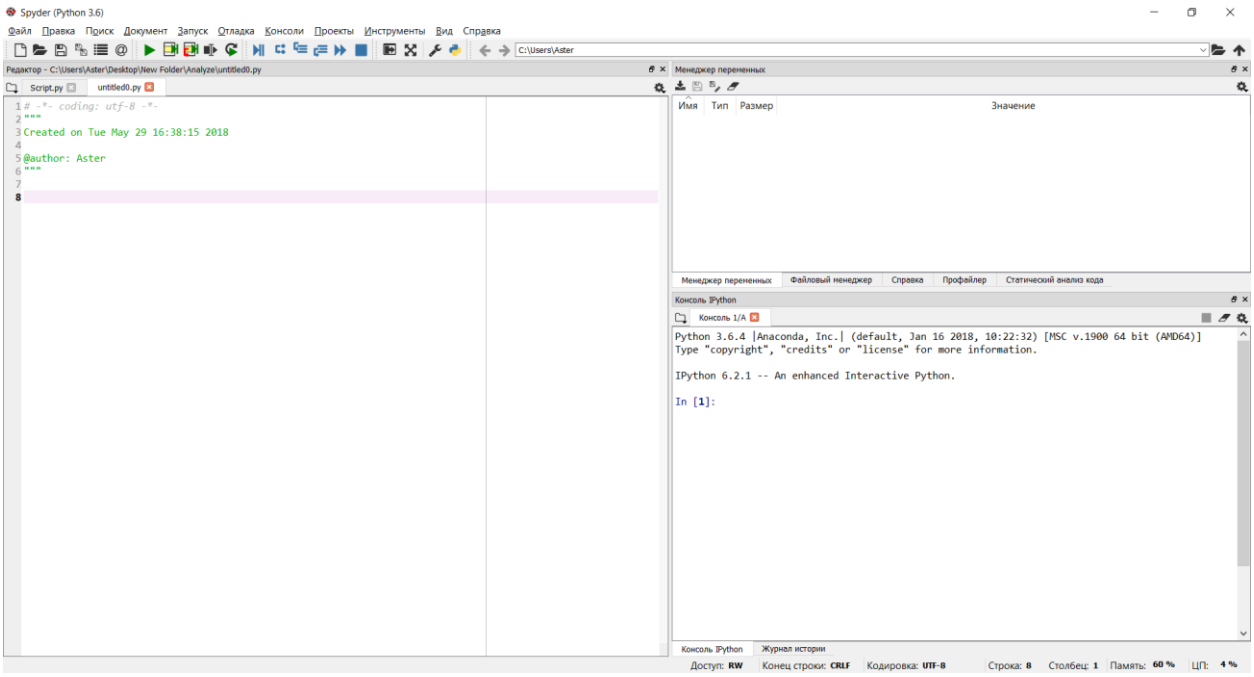


Рисунок 2.4 – Интерфейс Spyder 3

Из основных преимуществ Spyder можно выделить [62]:

– просмотр значений переменных, их импорт/экспорт (рис. 2.5). В процессе выполнения программы они выводятся на панели в виде списка с возможностью просмотра их значений. Также, по данным массива, можно построить график;

– встроенная консоль. Возможность создавать требуемое количество консолей и взаимодействовать с ними как с отдельными процессами, что существенно ускоряет процесс обработки;

– интроспекция кода: автодополнение, переход к выбранному объекту или функции, просмотр кода используемых модулей;

– работа с документацией в режиме онлайн.

Менеджер переменных

Имя	Тип	Размер	Значение
C	float	1	5.62
EvM	list	1098	[[26762265, -0.009999999999999953, -0.03999999999999998, 0.01000000000 ...
I	float	1	1.51
L	float	1	2.1399999999999997
M	float	1	0.94
a	float	1	6931.628989943281
b	int	1	1097
c	int	1	12

Рисунок 2.5 – Список переменных

2.2.2 Описание программы MS Excel

Microsoft Excel является распространенной компьютерной программой, с помощью которой производятся расчеты, составляются таблицы и диаграммы, вычисляются функции различной сложности [63].

Так как программа входит в пакет Microsoft Office, то она установлена практически на всех компьютерах. Возможность составления таблиц, диаграмм и отчетов, произведения самых сложных вычислений делает эту программу популярной среди многих офисных сотрудников. При этом MS Excel отличается понятным интерфейсом и удобством использования с помощью панели инструментов, представленной на рис. 2.6.

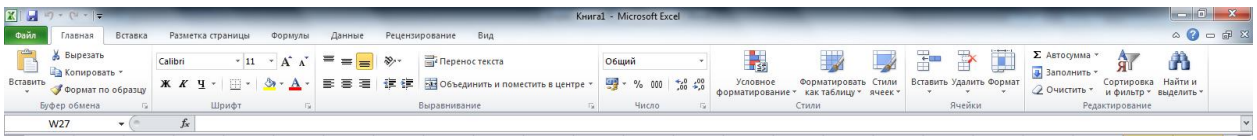


Рисунок 2.6 – Панель инструментов программы MS Excel 2010

Все данные по абитуриентам были предоставлены в форматах xls и csv.

Соответственно, анализ и запись полученных результатов осуществлялись в Microsoft Excel 2010 без использования дополнительных программных средств.

2.3 Апробация модели и расчет точности полученных результатов на примере тестовой выборки

На основе выборки, состоящей из 2225 старшеклассников, построена модель, с помощью которой был рассчитан перечень сообществ с весовым коэффициентом для каждого признака. Результаты представлены по первым 470 учащимся. Расчет осуществлялся на основе следующей дифференциации: 1 класс – высокий уровень способностей, 2 – средний, 3 – низкий.

По признаку креативность наибольшая точность по первому классу достигает 63%, для 2 – 76%, а для третьего класса – 70%. Наблюдается вариация в пределах 60-80% на всем диапазоне анализируемой выборки для всех категорий (рис. 2.7).

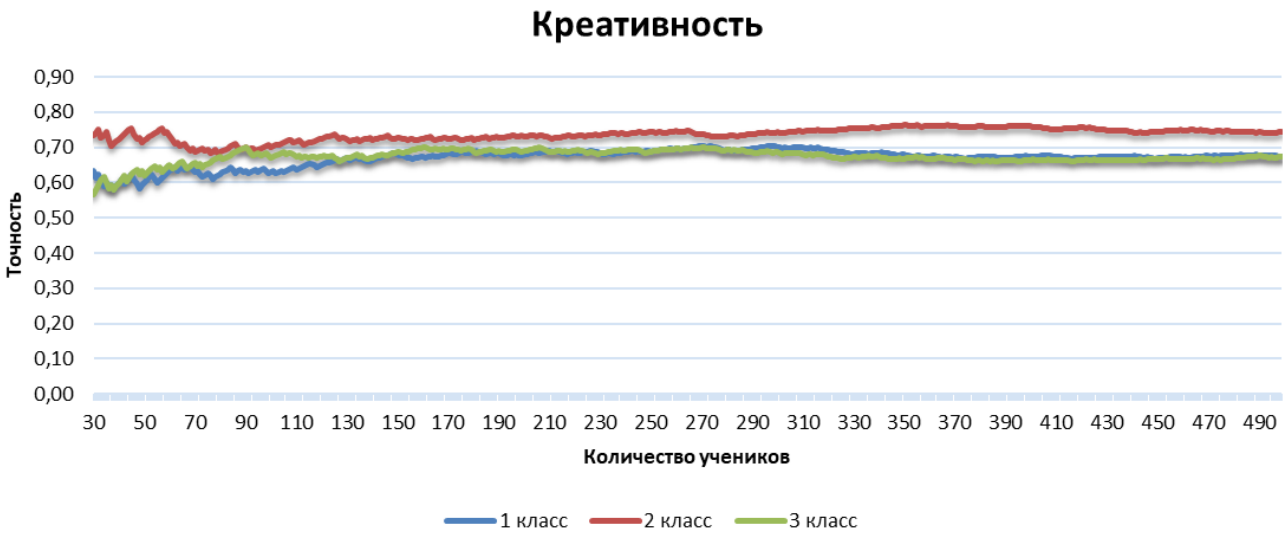


Рисунок 2.7 – Точность модели по признаку креативность

По признаку мотивация наибольшая точность по первому классу достигает 64%, для 2 – 80%, а для третьего класса – 70%. Наблюдается вариация в пределах 55-80% на всем диапазоне анализируемой выборки для всех категорий (рис. 2.8).

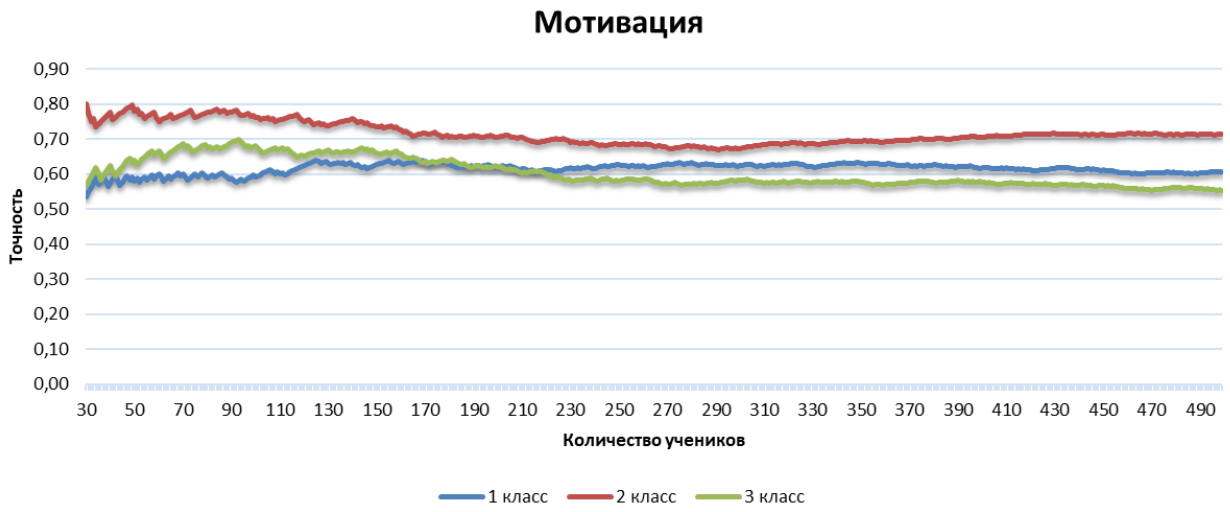


Рисунок 2.8 – Точность модели по признаку мотивация

По признаку интеллект наибольшая точность по первому классу достигает 73%, для 2 – 71%, а для третьего класса – 59%. Наблюдается вариация в пределах 47-73% на всем диапазоне анализируемой выборки для всех категорий (рис. 2.9).

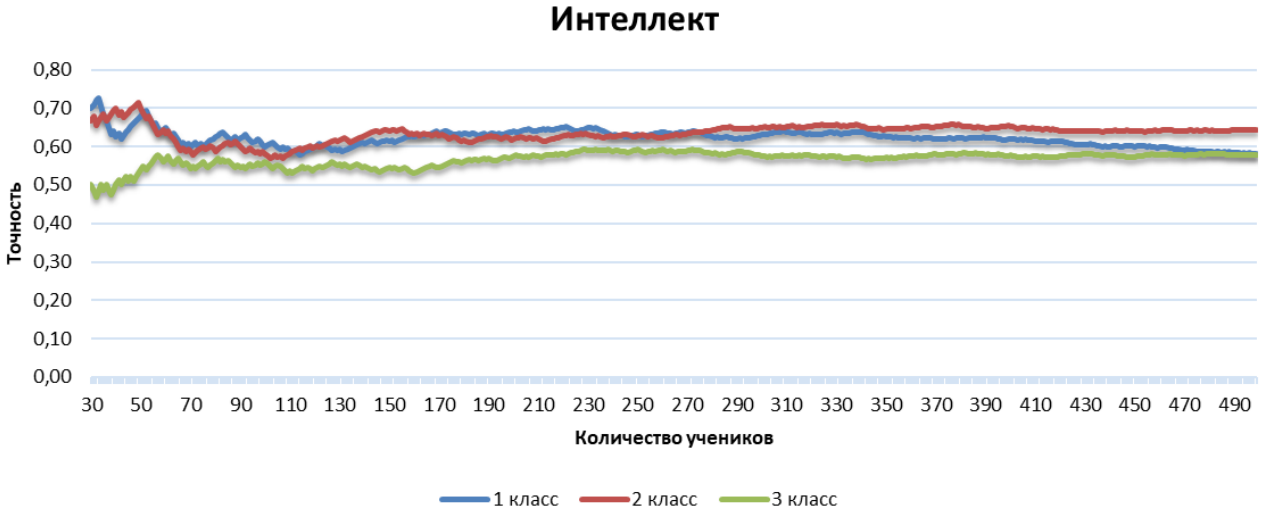


Рисунок 2.9 – Точность модели по признаку интеллект

По признаку личность наибольшая точность по первому классу достигает 73%, для 2 – 74%, а для третьего класса – 80%. Наблюдается вариация в пределах 59-80% на всем диапазоне анализируемой выборки для всех категорий (рис. 2.10).

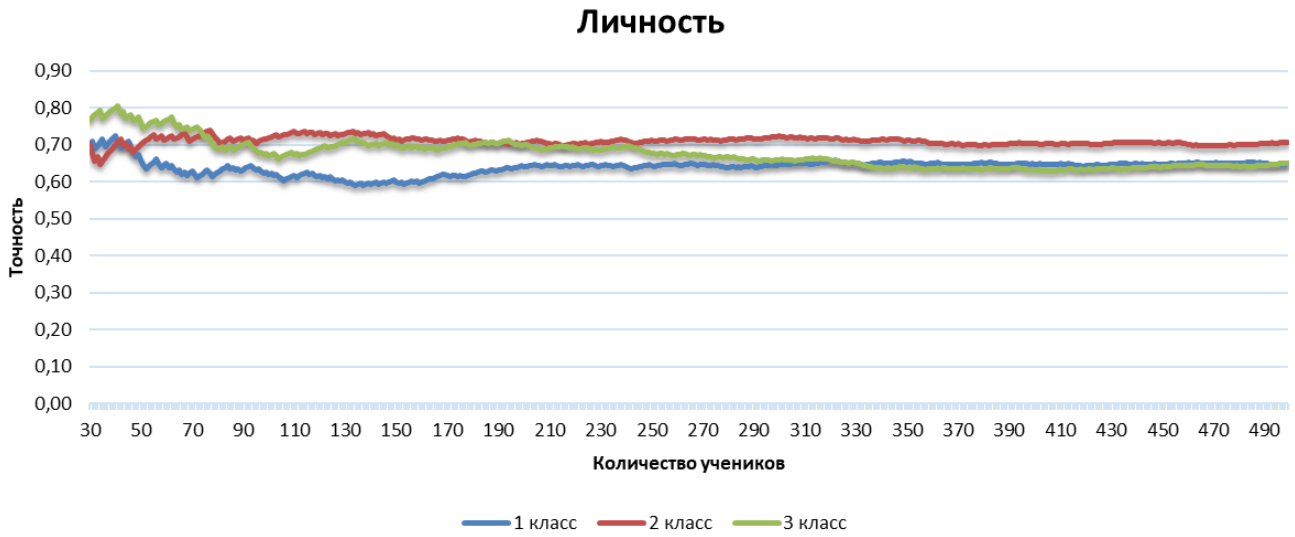


Рисунок 2.10 – Точность модели по признаку личность

Помимо модели для всей выборки была построена и посчитана точность модели с разделением по полу. Совокупность учеников составила 1692 человека, из которых 969 девушек и 723 юноши.

Точность модели определялась также по результатам тестирования и на основании подписок у этих же учащихся. В подсчете участвовали только старшеклассники с высокими показателями по каждому признаку, т.е. относящиеся к 1 классу (рис. 2.11-2.13).

По признаку креативность наибольшая точность по мужскому полу достигает 83%, по женскому – 78%. Наблюдается вариация в пределах 60-82% на всем диапазоне анализируемой выборки для обоих полов (рис. 2.11).

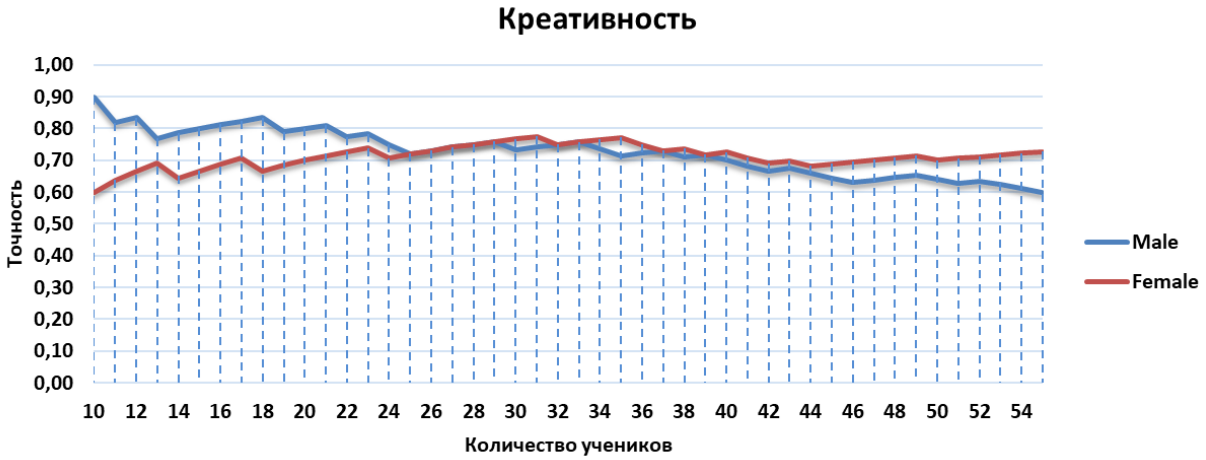


Рисунок 2.11 – Точность модели с разделением по полу, креативность

По признаку мотивация наибольшая точность по мужскому полу достигает 90%, по женскому – 82%. Наблюдается вариация в пределах 62-90% на всем диапазоне анализируемой выборки для обоих полов (рис. 2.12).

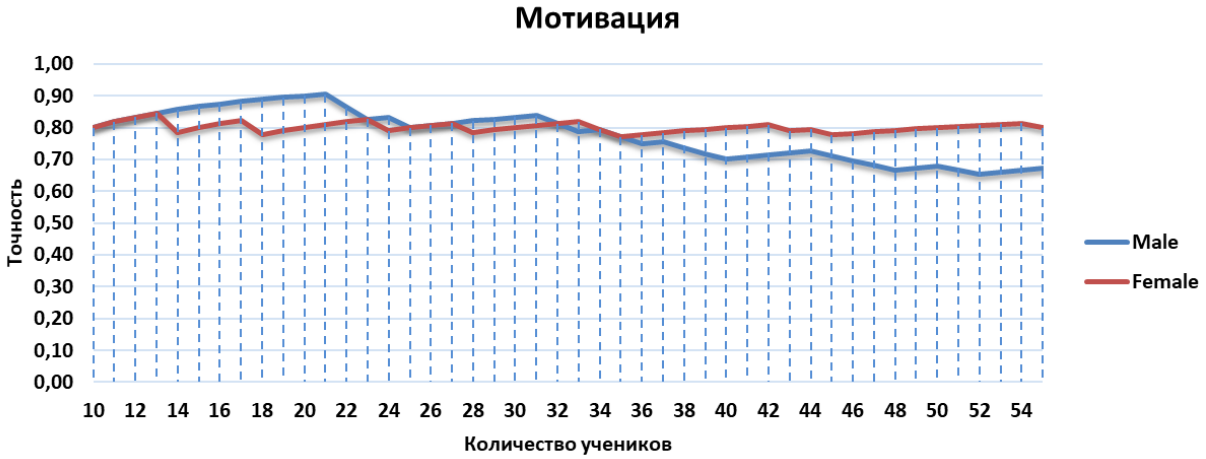


Рисунок 2.12 – Точность модели с разделением по полу, мотивация

Только по признаку интеллект результаты получились достаточно низкими. Наибольшая точность по мужскому полу составила 69%, по женскому – 57%. Наблюдается вариация в пределах 40-69% на всем диапазоне анализируемой выборки для обоих полов (рис. 2.13). В случае с определением интеллектуальных способностей следует подходить к дифференциации комплексно и рассматривать не только расчет модели, но и учитывать

результаты ЕГЭ.

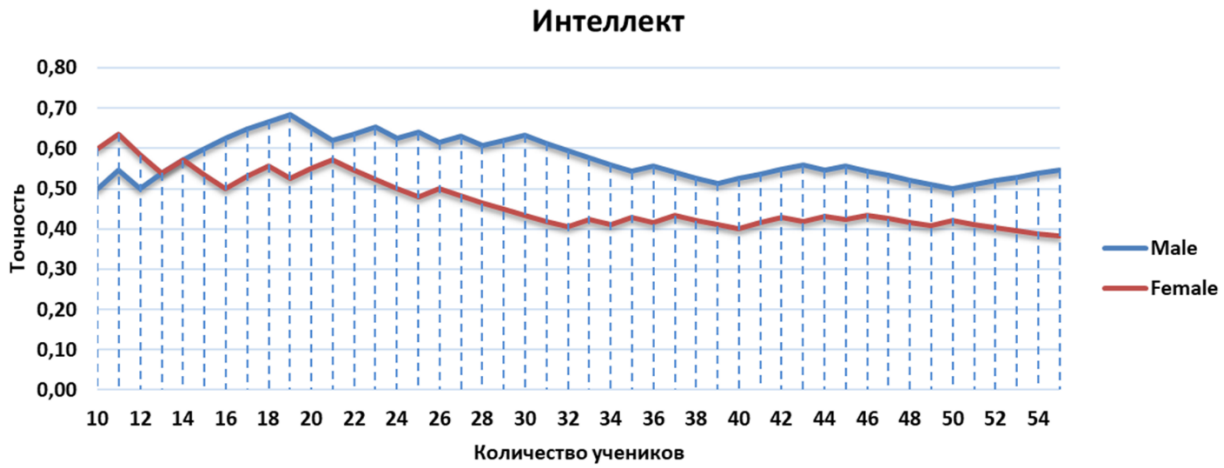


Рисунок 2.13 – Точность модели с разделением по полу, интеллект

Что касается прикладного применения результатов работы, то с помощью модели было выявлено 20 тыс. из 280 тыс. абитуриентов СФО (Сибирский Федеральный Округ) с прогнозируемыми высокими уровнями признаков одаренности. На данный момент проводится индивидуальная работа с каждым из них по приглашению в ТГУ.

2.4 Автоматизация модели

Для реализации поставленной цели использовались следующие библиотеки:

1) Numpy – одна из основных библиотек Python, осуществляющая поддержку больших многомерных массивов и матриц, вместе с библиотекой высокоуровневых математических функций для операций с этими массивами.

2) Openpyxl – это библиотека, используемая для чтения и записи файлов в формате xlsx, xls и т.д.

3) Os – модуль предоставляет множество функций для работы с

операционной системой. В данном исследовании использовалась функция walk для считывания подписок, которые были переданы в виде файлов в формате txt. Для каждого каталога функция возвращает кортеж (путь к каталогу, список каталогов, список файлов).

4) Scipy – содержит модули для оптимизации, интегрирования, специальных функций, обработки сигналов, обработки изображений, генетических алгоритмов, решения обыкновенных дифференциальных уравнений и других задач, обычно решаемых в науке и при инженерной разработке. В работе использовалась функция rankdata для разделения учащихся на три класса. Суть ее работы заключается в ранжировании элементов, присваивая им соответствующий ранг, на основе анализируемого списка.

5) Timeit – библиотека для контроля за выполняемым расчетом. Также ее можно использовать в качестве определения производительности отдельного алгоритма.

Загрузка массива для обработки и отслеживание прогресса было реализовано с помощью фрагмента кода, представленного на рис. 2.14.

```
timer = timeit.default_timer()
mass1=[]
q=0
wb = openpyxl.load_workbook(filename='C:/folder/Spisok.xlsx', data_only=True)
sheet = wb['Лист1']
print('Время открытия файла:',round(timeit.default_timer()-timer,2))
timer = timeit.default_timer()
for i in range(1,sheet.max_row+1):
    mass1.append([sheet.cell(row = i, column=1).value])
    for j in range(2,sheet.max_column):
        if sheet.cell(row = i, column=j).value != None:
            mass1[q].append(sheet.cell(row = i, column=j).value)
        else:
            break;
    q+=1
    if q % 500==0:
        print('Обработано ',q,' записей')
        print('Текущие затраты по времени составляют:',round((timeit.default_timer()-timer)/60,2),'минут')
```

Рисунок 2.14 – Загрузка подписок на сообщества в массив

Импорт списка сообществ для анализа в количестве 44101 представлен на рис. 2.15.

```
com = []
wb = openpyxl.load_workbook(filename='C:/folder/Svodka_po_soobschestvam_absolyutnye_znachenia.xlsx', data_only=True)
sheet = wb['Лист1']
for i in range(2, sheet.max_row+1):
    com.append([sheet.cell(row=i, column=1).value])
    for c in range(13):
        com[i-2].append(0)
```

Рисунок 2.15 – Загрузка перечня сообществ в массив

Информация по подпискам для построения модели получена в виде файлов в формате txt. Для их корректной записи применен метод обработки строк (рис. 2.16).

```
a = []
k = -1
for d, dirs, files in os.walk('C:/podpiski/group7'):
    for f in files:
        a.append([f.replace('.txt', '')])
        b = open(d + '/' + f)
        k += 1
        for line in b.readlines():
            a[k].append(line.replace('\n', ''))
```

Рисунок 2.16 – Импорт подписок для построения модели

```
students=[]
wb = openpyxl.load_workbook(filename='C:/folder/MainModel.xlsx', data_only=True)
sheet = wb['Main']
for i in range(2, sheet.max_row+1):
    students.append([(sheet.cell(row=i, column=1).value).replace('https://vk.com/id', ''), sheet.cell(row=i, column=2).value,
```

Рисунок 2.17 – Выгрузка списка абитуриентов с результатами тестов

Для редактирования массива и расчета перцентилей была создана функция classification (рис. 2.18).

```
proc=[]
def classification(massive):
    massive = list(zip(*massive))
    masss=[]
    masss.append([np.percentile(massive[i],75) for i in range(1,len(massive))])
    masss.append([np.percentile(massive[i],24) for i in range(1,len(massive))])
    return masss

proc=classification(coef)  #вызов функции
```

Рисунок 2.18 – Функция classification

Разделение абитуриентов по классам на основе результатов тестирования (рис. 2.19).

```
for i in range(0,len(coef)):
    for j in range(1,len(coef[i])):
        if coef[i][j]>=proc[0][j-1]:
            if j==1:
                coef[i][j]='1'
            elif j==2:
                coef[i][j]='1'
            elif j==3:
                coef[i][j]='1'
            elif j==4:
                coef[i][j]='1'
            elif j==5:
                coef[i][j]='1'
        elif coef[i][j]<=proc[1][j-1]:
            if j==1:
                coef[i][j]='3'
            elif j==2:
                coef[i][j]='3'
            elif j==3:
                coef[i][j]='3'
            elif j==4:
                coef[i][j]='3'
            elif j==5:
                coef[i][j]='3'
        else:
            if j==1:
                coef[i][j]='2'
            elif j==2:
                coef[i][j]='2'
            elif j==3:
                coef[i][j]='2'
            elif j==4:
                coef[i][j]='2'
            elif j==5:
                coef[i][j]='2'
```

Рисунок 2.19 – Классификация абитуриентов

Характеристика сообществ по количеству в них учащихся, разделенных на 3 класса по каждому признаку (рис. 2.20).

```
k=0
while k!=len(students):
    for i in range(0, len(a)):
        if a[i][0]==students[k][0]:
            for j in range(1, len(a[i])):
                for p in range(0, len(com)):
                    if a[i][j] == str(com[p][0]):
                        if coef[k][1]=='1':
                            com[p][2]+=1
                        elif coef[k][1]=='2':
                            com[p][3]+=1
                        elif coef[k][1]=='3':
                            com[p][4]+=1
                        if coef[k][2]=='1':
                            com[p][5]+=1
                        elif coef[k][2]=='2':
                            com[p][6]+=1
                        elif coef[k][2]=='3':
                            com[p][7]+=1
                        if coef[k][3]=='1':
                            com[p][8]+=1
                        elif coef[k][3]=='2':
                            com[p][9]+=1
                        elif coef[k][3]=='3':
                            com[p][10]+=1
                        if coef[k][4]=='1':
                            com[p][11]+=1
                        elif coef[k][4]=='2':
                            com[p][12]+=1
                        elif coef[k][4]=='3':
                            com[p][13]+=1
    k+=1
```

Рисунок 2.20 – Подсчет подписок на сообщества

Подсчет коэффициента для каждого учащегося с информацией по количеству подписок, содержанию маркерных сообществ, точности, которая рассчитывается исходя из соотношения маркерных сообществ к общему количеству подписок, по классификации и контролю за выполнением процесса расчетов представлены на рис. 2.21-2.22.

```
EvM=[]
sheet = wb['Result']
for i in range(2, sheet.max_row+1):
    EvM.append([sheet.cell(row=i, column=1).value,sheet.cell(row=i, column=27).value,sheet.cell(row=i, column=28).value,
                sheet.cell(row=i, column=29).value,sheet.cell(row=i, column=30).value])
mas_ = []
kk=0
q=0
count=0
timer = timeit.default_timer()
while kk!=len(mass1):
    for i in range(0, len(EvM)):
            for j in range(1,len(mass1[kk])):
                if (EvM[i][0] == mass1[kk][j]) and (len(mass1[kk])>1):
                    q += EvM[i][2]
                    count+=1
    if count!=0:
        mas_.append([mass1[kk][0],round(q/(len(mass1[kk]) - 1),5),round(count/len(EvM),5),count,len(mass1[kk])-1])
        q=0
        count=0
    kk+=1
    if kk % 500==0:
        print('Обработано ',kk,' записей')
        print('Текущие затраты по времени составляют:',round((timeit.default_timer()-timer)/60,2),'минут')
```

Рисунок 2.21 – Расчет коэффициента и точности по каждому абитуриенту

```
proc_=[]
students_=[]

for i in range(2, sheet.max_row+1):
    students_.append([sheet.cell(row=i, column=1).value, sheet.cell(row=i, column=2).value,
                      sheet.cell(row=i, column=3).value])
for i in range(len(students_)):
    students_[i].append(round(students_[i][2]*100/(max([row[2] for row in students_])),1))


proc_=classification(students_)
for i in range(0,len(students_)):
    if students_[i][1]>=proc_[0][0]:
        students_[i][1]='1'
    elif students_[i][1]<=proc_[1][0]:
        students_[i][1]='3'
    else:
        students_[i][1]='2'
```

Рисунок 2.22 – Классификация учащихся

ЗАКЛЮЧЕНИЕ

В рамках магистерской работы на основе данных о подписках абитуриентов 2018 года была разработана модель, которая с некоторой долей вероятности по четырем признакам (креативность, интеллект, мотивация, личность) классифицирует учащихся на одну из трех категорий: 1 – высокий уровень, 2 – средний уровень, 3 – низкий уровень. Полученная информация позволила выявить абитуриентов с выраженными лидерскими качествами, креативным мышлением и интеллектуальными способностями.

Изученный материал отечественных и зарубежных ученых по анализу данных в социальных сетях подтверждает возрастающий интерес к этой области знаний. А на основе анализа их работ удалось разработать алгоритм обработки имеющихся данных с минимальными затратами вычислительных мощностей.

Апробация результатов на тестовой выборке показала точность от 62% до 90%, в зависимости от признака, среди учащихся с высоким уровнем представленности признака одарённости. Именно определение этой категории учеников являлось основной задачей данного исследования.

Автоматизация модели реализована на языке программирования Python 3.6. Анализ, хранение и обработка данных осуществлялась в программном продукте MS Excel.

Рассмотренные в исследовании методики, подходы и проведенное исследование с разработанным алгоритмом обработки данных показали, что технологии машинного обучения имеют большой потенциал для анализа информации в социальных сетях. Мировой опыт применения алгоритмов обработки данных к решению многих актуальных задач может быть успешно использован для создания и развития нового аналитического инструментария социальных и гуманитарных наук, в том числе в области социальных медиа.

На основе исследований по магистерской диссертации был присвоен диплом II степени на Международной конференции «Актуальные проблемы социальных наук» в секции «Гуманитарная информатика: исследование

информационного общества и социальных проблем информатизации».

Что касается дальнейшего развития работы, в силу того, что при моделировании скорость обработки большого объема данных существенно снижается, следует произвести интеграцию с базами данных для оптимизации временных затрат. Помимо рассматриваемых подписок на сообщества в социальной сети планируется расширить анализ такими признаками как: репосты, авторские тексты в профиле, графы социальных связей. В перспективе внедрение модели в региональные вузы страны с целью повышения качества образования за счет привлечения одаренных абитуриентов. Так, регионы, стремясь повысить качество системы высшего образования и расширить рынок предлагаемых образовательных услуг, нацелены на обеспечение высокого уровня подготовки будущих студентов и достижение значительных показателей поступления абитуриентов в региональные вузы. Для достижения поставленных перед регионом целей в извлекаемых из социальных сетей данных необходимо находить закономерности, на основе которых можно выявить основные психологические и поведенческие характеристики целевой аудитории, ее интересы и профессиональные увлечения, а также определить особенности формирования единого портрета интересующего регион пользователя, т.е. будущего студента вуза.

В связи с тем, что информация по абитуриентам 2018 года была предоставлена без разделения по половому признаку, в дальнейшем, планируется выгрузить эту информацию и провести анализ уже с учетом новых данных.

Основной тенденцией расширения социальных сетей в качестве социокультурного феномена, можно выделить четкое понимание принципов поведения человека в обществе. Вследствие этого, необходимо акцентировать внимание на разработку средств для самовыражения, а также на обмен информацией и опытом.

В перспективе высока вероятность дальнейшего развития возможностей и влияния социальных сетей. Вследствие этого, доработка пользовательской

модели согласно новым потребностям на информационном рынке станет актуальной задачей в ближайшее время. Расширение функционала социальных сетей способствует возникновению новых типов данных в виде объектов и связей социального графа, что, в свою очередь, приведет к появлению новых задач и различных алгоритмов для эффективного их решения, связанного с обработкой частной информации.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. A Societal Sentiment Analysis: Predicting the Values and Ethics of Individuals by Analysing Social Media Content / T. Maheshwari et al. // Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. – 2017. – Vol. 1. – P. 731-741.

2. Рейтинг популярности социальных сетей. [Электронный ресурс] // URL : https://gs.seo-auditor.com.ru/dir/ (дата обращения: 15.10.2017).

3. Коршунов А. и др. Определение демографических атрибутов пользователей микроблогов //Труды Института системного программирования РАН. – 2013. – Т. 25, стр. 179-194. DOI: 10.15514/ISPRAS-2013-25-10.

4. Рабочая концепция одаренности / Д. Б. Богоявленская [и др.]. – М. : Мин-во образования РФ, 2003. – 95 с.

5. Матюшкин А. М. Мышление, обучение, творчество / А. М. Матюшкин. – М.: Изд-во МПСИ; Воронеж : НПО «МОДЭК», 2003. – 720 с.

6. Психология одаренности детей и подростков / под ред. Н. С. Лейтеса. – М: Издательский центр «Академия», 1996. – 416 с.

7. Хеллер К. А. Диагностика и развитие одаренных детей и подростков // Основные современные концепции творчества и одаренности. – М., 1997. – С. 243-264.

8. Heller K. A. International trends and issues of research into giftedness // Proceedings of the Second Asian Conference on giftedness: growing up gifted and talented. – 1992. – P. 93-110.

9. Рензулли Дж. Модель обогащающего школьного обучения / Дж. Рензулли, С. М. Рис // Основные современные концепции творчества и одаренности. – М., 1997. – С. 214-242.

10. Renzulli J. S. What is the thing called giftedness, and how do we develop it? A twenty five year perspective // Journal for the education of the gifted. – 1999. – Vol. 23. – № 1. – P. 3-54.

11. Щебланова Е. И. Одаренность как психологическая система: структура

и динамика в школьном возрасте: дис. д-ра психол. наук / Е. И. Щебланова. – М., 2006. – 311 с.

12. Что такое Data Mining? [Электронный ресурс] // URL: https://www.intuit.ru/studies/courses/6/6/lecture/158 (дата обращения: 05.10.2017).

13. Интеллектуальный анализ данных [Электронный ресурс] // URL: https://studfiles.net/preview/6172591/page:6/ (дата обращения: 05.10.2017).

14. Описание YouScan Электронный ресурс] // URL: https://startpack.ru/application/youscan-smm (дата обращения: 02.03.2017).

15. Возможности сервиса IQbuzz Электронный ресурс] // URL: http://iqbuzz.pro (дата обращения: 02.03.2017).

16. 5 сервисов для аналитики групп в социальных сетях [Электронный ресурс] // URL: https://edison.bz/blog/top-5-servisov-dlya-analitiki-grupp-v-sotsialnykh-setyakh.html (дата обращения: 02.03.2017).

17. Torrance E. P. The nature of creativity as manifest in its testing // The nature of creativity. Contemporary psychological perspectives. – Cambridge, 1988. – P. 43-75.

18. Torrance E. P. Torrance tests of creative thinking: Streamlined (revised) manual including norms and direction for administering and scoring figural A and B / E. P. Torrance, Ball O. E. – 1984.

19. Леонтьев Д. А. Личностное в личности: личностный потенциал как основа самодетерминации // Ученые записки кафедры общей психологии МГУ им. М. В. Ломоносова. – Вып. 1. – М.: Смысл, 2002. – С. 56-65.

20. Леонтьев Д. А. Личность как определяющий фактор // Воображение и творчество в образовании и профессиональной деятельности: материалы Четвертой международной конференции памяти Л. С. Выготского. – М. : РГГУ , 2004. – С. 214-223.

21. Маслоу А. Мотивация и личность / А. Маслоу. – СПб. : Питер, 2011. – 352 с.

22. Clark B. Growing up gifted / B. Clark. – New York: Macmillan, 1992. – 674 p.

23. Feldhusen J. F. A conception of giftedness. // Conceptions of giftedness. – Cambridge: Cambridge University Press, 1986. – P. 112-127.

24. Freeman J. Cultural influences on gifted gender achievement // High ability studies. The journal of the European Council for High Ability. – 2004. – № 1. – P. 7-23.

25. Freeman J. Gifted children growing up / J. Freeman. – London: Cassell, 1991.

26. Freeman J. Recent development for the high able in Britain // Education of the gifted in Europe: theoretical and research issues. – Amsterdam, 1992. – P. 58-70.

27. Silverman L. K. Counseling the gifted and talented / L. K. Silverman. – Denver: Love publishing company, 1993.

28. Юнг К. Г. Воспоминания, размышления, сновидения / К. Г. Юнг. – Минск: ООО «Харвест», 2003. – 496 с.

29. Бозаджиев В. Ю., Кукушин В. С., Воронцова М. В. Одаренные дети: Теория и практика обучения и развития. – «Scientific magazine» Kontsep, 2014.

30. Монкс Ф., Ипенбург И. Одаренные дети. – Litres, 2018.

31. Дружинин В. Н. Психология общих способностей. 3-е изд. – Издательский дом «Питер», 2013.

32. Матюшкин А. М. Что такое одаренность: выявление и развитие одаренных детей. – ЧеРо, 2006.

33. Ищенко И. П., Кузнецова Е. Н. Особенности интеллектуальных и творческих способностей городских и сельских подростков //Вестник Курганского государственного университета. Серия «Естественные науки». Выпуск 3. Курган: Изд-во Курганского гос. ун-та, 2010.-96 с. – 2010. – С. 35.

34. Холодная М. А. Принципы и методы выявления одаренных детей //Одаренность: рабочая концепция: ежегодник РПО/отв. ред.: ДБ Богоявленская, ВД Шадриков. –М.: Изд-во РПО. – 2000. – Т. 8. – №. 1. – С. 22-29.

35. Алгоритмы обработки данных [Электронный ресурс] // URL: https://www.intuit.ru/studies/courses/648/504/lecture/11461?page=2 (дата обращения: 15.12.2017).

36. Введение в машинное обучение [Электронный ресурс] // URL: https://www.intuit.ru/studies/courses/10621/1105/lecture/17981 (дата обращения: 15.12.2017).

37. Freund Y., Schapire R. A Decision-Theoretic Generalization of Online Learning and an Application to Boosting Computer and System Sciences. Vol.55, 1997. – pp. 119-139.

38. Cortes C., Vapnik V. N. Support-Vector Networks Machine Learning. 1995. V. 20, № 3. P. 273–297.

39. Najork M., Wiener J. L. Breadth-first crawling yields high-quality pages // Proceedings of the 10th international conference on World Wide Web. – ACM, 2001. – С. 114-118.

40. Leskovec J., Faloutsos C. Sampling from large graphs //Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. – ACM, 2006. – С. 631-636.

41. Коршунов А. и др. Анализ социальных сетей: методы и приложения // Труды Института системного программирования РАН. – 2014. – Т. 26. – №. 1.

42. Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, Yoram Singer. Online Passive-Aggressive Algorithms // JMLR, 7(Mar):551–585, 2006.

43. Батура Т. В. Методы анализа компьютерных социальных сетей //Вестник Новосибирского государственного университета. Серия: Информационные технологии. – 2012. – Т. 10. – №. 4.

44. Miller Z., Dickinson B., Hu W. Gender prediction on twitter using stream algorithms with N-gram character features. – 2012.

45. Burger J. D. et al. Discriminating gender on Twitter //Proceedings of the Conference on Empirical Methods in Natural Language Processing. – Association for Computational Linguistics, 2011. – С. 1301-1309.

46. Alowibdi J. S., Buy U. A., Yu P. Empirical evaluation of profile characteristics for gender classification on twitter //Machine Learning and Applications (ICMLA), 2013 12th International Conference on. – IEEE, 2013. – Т. 1. – С. 365-369.

47. Conover M. D. et al. Predicting the political alignment of twitter users

//Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on. – IEEE, 2011. – С. 192-199.

48. Cheng N., Chandramouli R., Subbalakshmi K. P. Author gender identification from text //Digital Investigation. – 2011. – Т. 8. – №. 1. – С. 78-88.

49. Rao D. et al. Classifying latent user attributes in twitter //Proceedings of the 2nd international workshop on Search and mining user-generated contents. – ACM, 2010. – С. 37-44.

50. Filippova K. User demographics and language in an implicit social network //Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. – Association for Computational Linguistics, 2012. – С. 1478-1488.

51. Гомзин А. Г., Кузнецов С. Д. Методы построения социо-демографических профилей пользователей сети Интернет // Труды Института системного программирования РАН. – 2015. – Т. 27. – №. 4.

52. Private traits and attributes are predictable from digital records of human behavior [Электронный ресурс] // URL: http://www.pnas.org/content/110/15/5802 (дата обращения: 15.10.2017).

53. Литвинова Т. А. К проблеме стабильности характеристик идиостиля // Известия Южного федерального университета. Филологические науки. – 2015. – №. 3. – С. 98-106.

54. New machine learning methods demonstrate the existence of a human stylome / H. Van Halteren et al. // Journal of Quantitative Linguistics. – 2005. – Vol. 12. – № 1. – P. 65-77.

55. Воронина И. Е., Гончаров В. А. Анализ эмоциональной окраски сообщений в социальных сетях (на примере сети «ВКонтакте») //Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии. – 2015. – №. 4. – С. 151-158.

56. Mining Facebook data for predictive personality modeling / D. Markovikj et al. // Proceedings of the 7th international AAAI conference on Weblogs and Social

Media (ICW SM 2013). – Boston, MA, USA. – 2013. – P. 23-26.

57. Богдановская И. М. Компьютерная психодиагностика в профориентационной работе со старшеклассниками / И. М. Богдановская, П. Б. Киселев, А. Н. Кошелева, В. А. Рубан // Психологические проблемы образования и воспитания в современной России: материалы IV конференции психологов образования Сибири. – Иркутск, 2016. – С. 296-302.

58. Data Mining, Analytics, and Knowledge Discovery Websites [Электронный ресурс] // URL: https://www.kdnuggets.com/websites/data-mining.html (дата обращения: 22.12.2017).

59. Python и R: что выбрать для Data Science в 2018? [Электронный ресурс] // URL: https://proglib.io/p/python-vs-r/ (дата обращения: 22.12.2017).

60. Anaconda Cloud [Электронный ресурс] // URL: https://anaconda.org/ (дата обращения: 25.12.2017).

61. Spyder – научная среда разработки для Python [Электронный ресурс] // URL: http://obscurityway.blogspot.com/2011/03/spyder.html (дата обращения: 10.02.2018).

62. Microsoft Excel [Электронный ресурс] // URL: http://obscurityway.blogspot.com/2011/03/spyder.html (дата обращения: 10.02.2018).

63. George Pallis, Demetrios Zeinalipour-Yazti, Marios D. Dikaiakos. Online Social Networks: Status and Trends. New Directions in Web Data Management 1, Studies in Computational Intelligence Volume 331, 2011, pp 213-234.

ПРИЛОЖЕНИЕ А

Результаты комплексного профориентационного тестирования старшеклассников в ТГУ

Таблица А.1 – Графическое представление результатов тестирования учащихся по 15 показателям (фрагмент)

ID пользователя	xxx918847	xxx713102	xxx569929	xxx462247	xxx111863	xxx204444	xxx204444	xxx453333	xxx734539	xxx024405	xxx673167	xxx758913	xxx727527	xxx887920	xxx752397	xxx869842
Аналогии	7	1	18	5	7	7	15	10	2	5	10	6	7	2	17	7
Числовые ряды	4	2	1	1	2	3	2	2	4	2	4	2	4	1	3	4
Беглость	7	5	6	5	6	6	5	6	7	6	7	6	3	8	8	7
Образная адаптивная гибкость	7	7	8	6	5	8	8	8	3	8	7	6	4	7	7	7
Семантическая гибкость	9	6	8	6	6	8	5	6	7	9	7	6	5	4	7	9
Семантическая спонтанная гибкость	9	6	8	5	6	6	7	7	8	8	6	6	4	9	9	9
Оригинальность	8	8	8	6	6	6	6	6	7	7	6	6	5	8	8	8
Независимость	9	7	8	6	7	6	6	6	8	9	6	6	5	9	1	9
Креативное поведение	2	3	2	3	3	2	2	2	2	3	1	3	2	3	1	2
Познавательный мотив	1	2	1	2	2	2	2	2	2	1	2	2	3	1	2	1
Мотив самореализации	3	3	1	2	3	2	3	1	3	1	3	2	3	1	3	3
Инициативность	2	2	1	2	2	2	3	2	2	2	2	1	3	2	2	2
Настойчивость	30	26	32	22	28	19	17	22	18	25	24	32	16	19	20	30
Проф.компетентность	10	12	14	12	11	17	15	13	16	15	10	14	7	11	17	10
Автономия	16	14	16	13	14	17	15	11	18	14	17	14	10	16	17	16

ПРИЛОЖЕНИЕ Б

Первоначальные данные после выгрузки из «ВКонтакте»

A1 | 173882243

	A	B	C	D	E	F	G	H	I	J
1	173882243	28905875	71474813	121715486	93289930	125945331	71729358	157945823	23279823	133178456
2	236522244	83549072	85612530	88474545	110225201	132729641	146093422	118960405	47252548	36164349
3	410813061	124604811	151414392							
4	386833735	93289930	56700006	97630666	121199691	160633958	120103784	132237656	116272634	125301333
5	251082638	66678575	57846937	104332051	28905875	32716331	151140262	133814709	93289930	78996568
6	313768591	23091433	53644646	43509872	100274916	154216159	125559973	157806386	58170807	93289930
7	397864080	27112805	23647100	82558431	80066837	89750322	83434560	129112812	98589889	87230508
8	461004946	33073882	59124711	155341746	105829519	27972579	96050194	63303373	139780732	146866838
9	141031787	127610748	62297798	134121211	98647548	48946342	68029781	31547740	132694187	62297929
10	324682835	153161326	63328642	42903168	98666451	2661911	49690338	144635297	40567146	148867254
11	77870557	8722610	43001537	28905875	41813928	71474813	118162965	23693281	15326149	118703121
12	221782042	146917032	93289930	67920625	134382850	35061290	107413693	39410028	117390149	53197574
13	151772893	143804894	93289930	45745333	57846937	32786443	32540140	137331585	132799222	133206054
14	145565027	61284289	144810593	146917032	131923632	57846937	67281773	103933781	135185874	157341954
15	448030695	26419239	152191544	147166906	58109200	36164349	12382740	131301287	107400177	139923997
16	317068265	108170711	85087785	30602036	460389	117661508	45745333	124249069	76628628	66678575
17	158997867	123695926	142914432	71729358	67580761	19328401	31976785	57846937	30602036	67281773
18	437146799	93289930	124260407	75854410	36092580	138191691	97630666	86747483	149500863	151140262
19	421195250	132729641	155293068	132265	159015406	387766	159439444	26421686	143940544	158062080
20	312581235	26062647	23091433	777107	53023033	51949926	32017129	101933932	145172065	113750664
21	427231581	47544652	157369801	63749724	26492580	147987463	135581003	142969933	131497180	79928855
22	268668920	94674474	23308460	60114472	86218441	102587842	112211940	26614831	45960892	148736004
23	283582073	124156921	44786979	23669758	38802692	78650125	88245281	72113408	26713492	47513151

Рисунок А.1 – Исходные данные о подписках старшеклассников на тематические сообщества «ВКонтакте»

АНТИПЛАГИАТ
ТВОРИТЕ СОБСТВЕННЫМ УМОМ

Отчет о проверке на заимствования №1

Автор: aster32167@gmail.com / ID: 3345803
Проверяющий: (aster32167@gmail.com / ID: 3345803)

Отчет предоставлен сервисом «Антиплагиат»- http://www.antiplagiat.ru

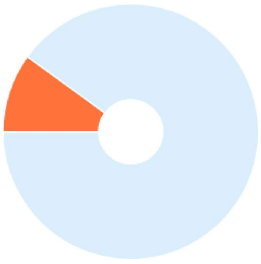
ИНФОРМАЦИЯ О ДОКУМЕНТЕ

№ документа: 78
Начало загрузки: 14.06.2018 00:17:40
Длительность загрузки: 00:00:02
Имя исходного файла: Корепанов_ВКР_2018
Размер текста: 2058 кБ
Символов в тексте: 66570
Слов в тексте: 7846
Число предложений: 431

ИНФОРМАЦИЯ ОБ ОТЧЕТЕ

Последний готовый отчет (ред.)
Начало проверки: 14.06.2018 00:17:42
Длительность проверки: 00:00:02
Комментарии: не указано
Модули поиска:

ЗАИМСТВОВАНИЯ	ЦИТИРОВАНИЯ	ОРИГИНАЛЬНОСТЬ
9,83%	0%	90,17%



Заимствования — доля всех найденных текстовых пересечений, за исключением тех, которые система отнесла к цитированиям, по отношению к общему объему документа.
Цитирования — доля текстовых пересечений, которые не являются авторскими, но система посчитала их использование корректным, по отношению к общему объему документа. Сюда относятся оформленные по ГОСТу цитаты; общеупотребительные выражения; фрагменты текста, найденные в источниках из коллекций нормативно-правовой документации.
Текстовое пересечение — фрагмент текста проверяемого документа, совпадающий или почти совпадающий с фрагментом текста источника.
Источник — документ, проиндексированный в системе и содержащийся в модуле поиска, по которому проводится проверка.
Оригинальность — доля фрагментов текста проверяемого документа, не обнаруженных ни в одном источнике, по которым шла проверка, по отношению к общему объему документа.
Заимствования, цитирования и оригинальность являются отдельными показателями и в сумме дают 100%, что соответствует всему тексту проверяемого документа.
Обращаем Ваше внимание, что система находит текстовые пересечения проверяемого документа с проиндексированными в системе текстовыми источниками. При этом система является вспомогательным инструментом, определение корректности и правомерности заимствований или цитирований, а также авторства текстовых фрагментов проверяемого документа остается в компетенции проверяющего.

№	Доля в отчете	Доля в тексте	Источник	Ссылка	Актуален на	Модуль поиска	Блоков в отчете	Блоков в тексте
[01]	0,96%	2,26%	Основные современные ко...	http://refwin.ru	23 Апр 2016	Модуль поиска Интернет	2	11
[02]	1,07%	1,47%	Человеческое мышление, с...	http://kazzam.ru	06 Апр 2016	Модуль поиска Интернет	8	10
[03]	1,01%	1,01%	Введение в машинное обуч...	http://diss.seluk.ru	07 Фев 2017	Модуль поиска Интернет	5	5

Еще источников: 17
Еще заимствований: 6,78%