НЕДЕЛЯ1. ВЫБОР ПРЕДМЕТНОЙ ОБЛАСТИ.

Поставленная перед слушателями задача не привязана к какой-либо конкретной предметной области. Предполагается отойти от принципа выполнения заранее поставленных и четко сформулированных задач, чтобы предоставить исполнителю гибкость и возможность творческого подхода выполнения. Таким образом, исполнителю предоставляется возможность самостоятельного выбора интересующей его прикладной области, над которой в рамках курса будет проводиться работа. Если же исполнитель не имеет своих собственных предпочтений, то ему предлагаются на выбор предметные области, перечисленные ниже:

— «Анализ данных социальных сетей». Например, электронные ресурсы Vkontakte[1], Twitter[2], Facebook[3], LinkedIn[4] и др.;

— «Анализ рынка вакансий». Например, электронный ресурс HeadHunter[5];

— «Анализ фильмов». Например: интернет-проект «Кинопоиск»[6];

— «Анализ журнала запросов к сайту Wikipedia[7]»;

— «Технический радар». Анализ информации с ресурса StackOverFlow[8];

— «Использование существующих решений и наборов данных». Например, информация с ресурса Kaggle[9] (см. условия выставления итоговой оценки). Например, «задача Титаника»[10].

---

[1] Vkontakte. [Электронный ресурс]. Режим доступа: // http://www.vk.com.

[2] Twitter. [Электронный ресурс]. Режим доступа: // http://www.twitter.com.

[3] Facebook. [Электронный ресурс]. Режим доступа: // http://www.facebook.com.

[4] LinkedIn. [Электронный ресурс]. Режим доступа: // http://www.linkedin.com.

[5] HeadHunter — *качественная база резюме* и *вакансий* и *современные сервисы* для поиска работы и персонала. [Электронный ресурс]. Режим доступа: // http://www.hh.ru.

[6] Кинопоиск — русскоязычный интернет-проект, посвящённый кинематографу, [Электронный ресурс]. Режим доступа: // http://www.kinopoisk.ru.

[7] Wikipedia — свободная общедоступная мультиязычная универсальная интернет-энциклопедия, [Электронный ресурс]. Режим доступа: // http://www.wikipedia.org.

[8] StackOverFlow — популярная система вопросов и ответов о программировании, [Электронный ресурс]. Режим доступа: // http://www.stackoverflow.com.

Приветствуются темы из следующих областей: «Образование», «Наука», «Здравоохранение», «Информационные технологии» (ИТ) и др.

Для выбранной предметной области требуется сформулировать от 5 до 20 задач для проведения анализа. Задачи могут быть отнесены к следующим областям анализа: анализ социальных сетей (Social Mining), анализ Интернет-ресурсов (Web Mining), анализ текста (Text Mining), анализ данных (Data Mining). Классификация задач анализа по областям приведена на рис.1.

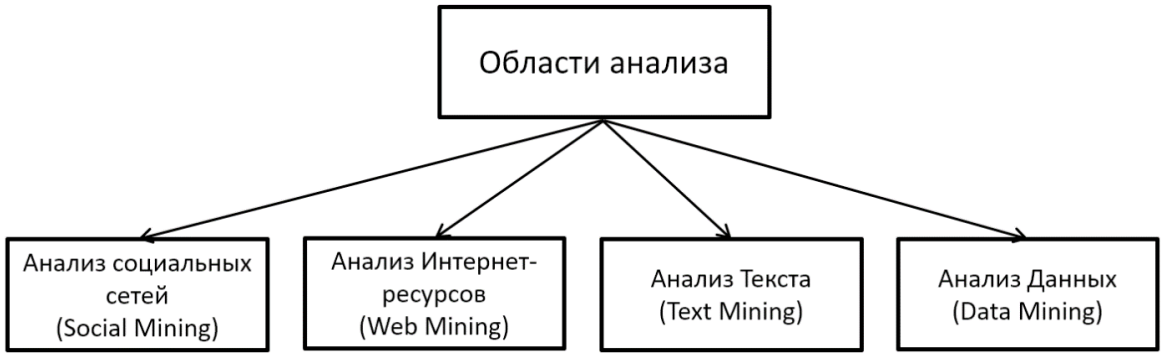

Рис.1 Классификация задач анализа по областям

В тоже время задачи анализа можно классифицировать по типу: задачи статистического типа и задачи исследовательского типа. Классификация приведена на рис.2.

---

[9] Kaggle - англоязычный ресурс, посвященный задачам анализа и науке о данных, [Электронный ресурс]. Режим доступа: // http://www.kagle.com.

[10] «Тита́ник» (англ. *Titanic*) — британский трансатлантический пароход. «Задача Титаника» - создание модели для предсказания выживших пассажиров парохода в зависимости от характеристик пассажира: его пол, возраст, номер каюты и т. д..

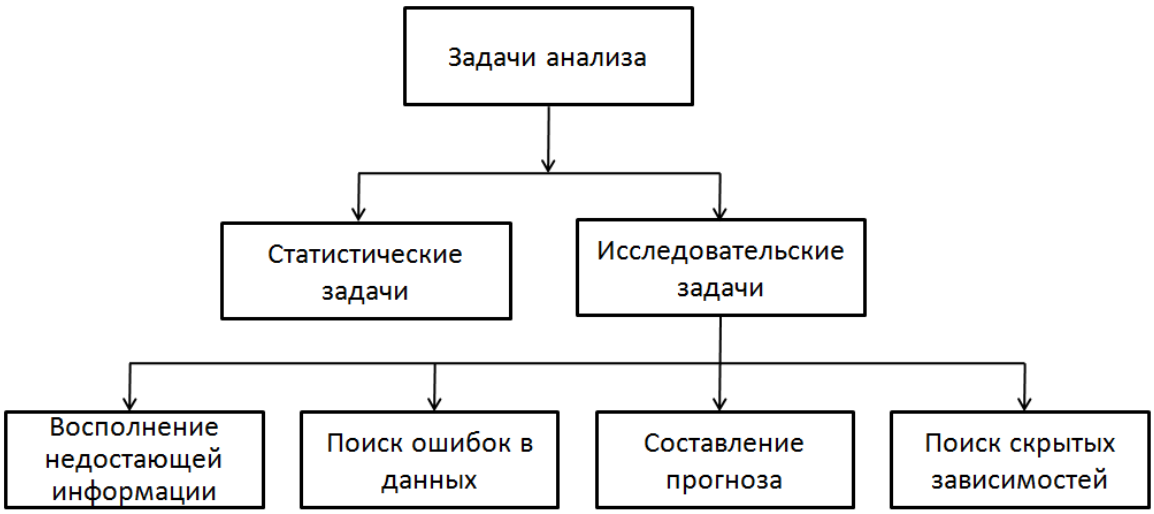

Рис. 2. Классификация задач анализа по типу

Статистические задачи относятся к традиционной обработке известного набора данных, объектов и их атрибутов для получения численных характеристик. Традиционно принято считать, что статистические задачи относятся к категории бизнес-аналитики (Business Intelligence). Они призваны помочь ответить на вопросы: «Какие численные показатели получила отрасль за прошлое время?», «Как правильно настроить рабочие процессы на основе прошлых, исторических данных?». Иными словами, результаты решения статистических задач помогают понять, что же произошло в прошлом и как на основе этих данных оптимизировать бизнес или производственные процессы и получить выгоду, зачастую экономическую. Особенностью реализации этого типа задачи являются: большое количество записей, большой объем информации и реализация алгоритмов обработки средствами и фреймворками для высокопроизводительных и распределенных вычислений.

Исследовательские задачи (Data Science), в отличие от статистических, подразумевают поиск скрытых зависимостей и паттернов в данных, восстановление недостающей информации, поиск ошибок в данных, а также составление некоторых прогнозов на будущее. Особенностью этого типа задач является использование инновационных, современных и прогрессивных методов анализа, которые в том числе позволяют построить своего рода экспертную систему.

При формулировании задач анализа необходимо, чтобы были представлены на утверждение задачи из каждой категории. Проработка каждой задачи анализа требует проявления фантазии и собственной заинтересованности в получении ответа на поставленный вопрос, потому что именно личностная заинтересованность может привести к высокому качеству выполнения проекта.

Стоит принять во внимание, что данные, подвергаемые анализу, могут обладать рядом неприятных свойств: неполнота, противоречивость, некорректность и разнородность. Если не учитывать возможность наличия таких свойств в данных, то результаты решения задач анализа могут находиться в другой плоскости относительно истинного решения. Для того, чтобы результаты решения задач были корректными, необходимо осуществлять валидацию и верификацию подвергаемой анализу информации. Зачастую применяют следующие подходы для проверки данных на корректность: методы машинного обучения, поиск нечетких связей и соответствий, и выявление обратной связи между атрибутами объектов, результатами решения задачи и входных данных.

Если рассматривать предметную область «Вакансии» с web-ресурса «HeadHunter», то в роли задач анализа могут выступать следующие приведенные статистические и исследовательские задачи.

Статистические задачи:

— анализ наиболее востребованных на рынке информационных технологий языков программирования в заданные интервалы времени (начиная с 2002 по 2016 гг.);

— определение распределения вакансий в области информационных технологий по регионам в зависимости от года;

— поиск наиболее популярных профессий в Российской Федерации;

— нахождение зависимости зарплаты от специализации;

Исследовательские задачи:

— поиск скрытых зависимостей между характеристиками работодателя и представленных вакансий;

— прогнозирование заработной платы в области IT на 2030 год.

Для предметной области «Социальные сети» в роли статистических задач анализа могут выступать:

— определение перечня городов, из которых в вузы Санкт-Петербурга приезжают для поступления абитуриенты, в том числе и зарубежные;

— нахождения перечня стран и городов, в которых работают выпускники вузов Санкт-Петербурга;

— установление параметров корреляции популярных тем обсуждений в социальных сетях с событиями в новостях.

Исследовательскими задачами для социальных сетей могут быть:

— прогнозирование количества приезжих абитуриентов в вузы Санкт-Петербурга;

— поиск скрытых зависимостей между родным городом абитуриента и Санкт-Петербургом.

Перечисленные выше примеры задач анализа могут показаться достаточно простыми и требующими создания одного или нескольких запросов к базам данных (БД). Исполнителю нужно сформулировать задачи анализа разной сложности, чтобы каждая из задач решалась с использованием разных подходов и методов обработки информации.