Оценочные материалы для промежуточной аттестации по дисциплине

«Основы научных исследований в области анализа данных»

Код, направление подготовки	09.04.02 ИНФОРМАЦИОННЫЕ СИСТЕМЫ И ТЕХНОЛОГИИ
Направленность (профиль)	УПРАВЛЕНИЕ ДАННЫМИ
Форма обучения	Очная
Кафедра-разработчик	Информатики и вычислительной техники
Выпускающая кафедра	Информатики и вычислительной техники

Типовые задания для контрольной работы (1 семестр):

Контрольная работа.

ВЫБОР ПРЕДМЕТНОЙ ОБЛАСТИ.
Поставленная задача не привязана к какой-либо конкретной предметной области. Предполагается отойти от принципа выполнения заранее поставленных и четко сформулированных задач, чтобы предоставить гибкость и возможность творческого подхода выполнения. Таким образом, предоставляется возможность самостоятельного выбора интересующей прикладной области, над которой будет проводиться работа. Если же нет своих собственных предпочтений, то предлагаются на выбор предметные области, перечисленные ниже:
«Анализ данных социальных сетей». Например, электронные ресурсы Vkontakte1, Twitter2, Facebook3, LinkedIn4 и др.;
«Анализ рынка вакансий». Например, электронный ресурс HeadHunter5;
«Анализ фильмов». Например: интернет-проект «Кинопоиск»6;
«Анализ журнала запросов к сайту Wikipedia7»;
«Технический радар». Анализ информации с ресурса StackOverFlow8;
«Использование существующих решений и наборов данных». Например, информация с ресурса Kaggle9 (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.

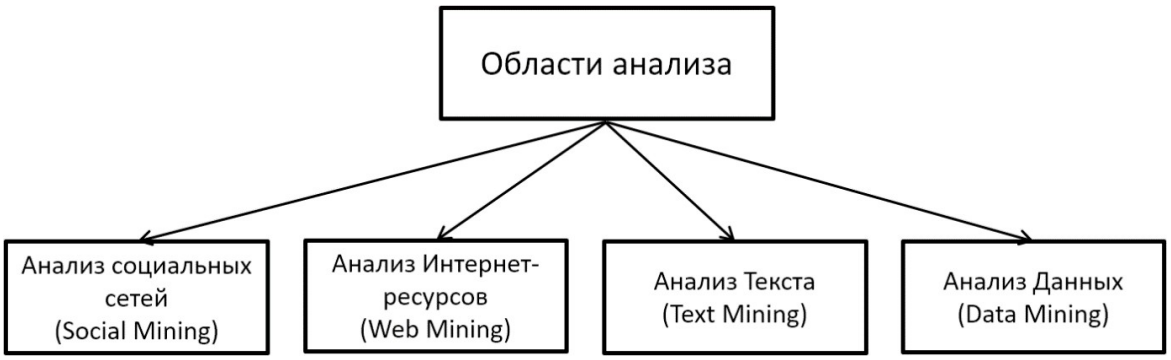

Рис.1 Классификация задач анализа по областям

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

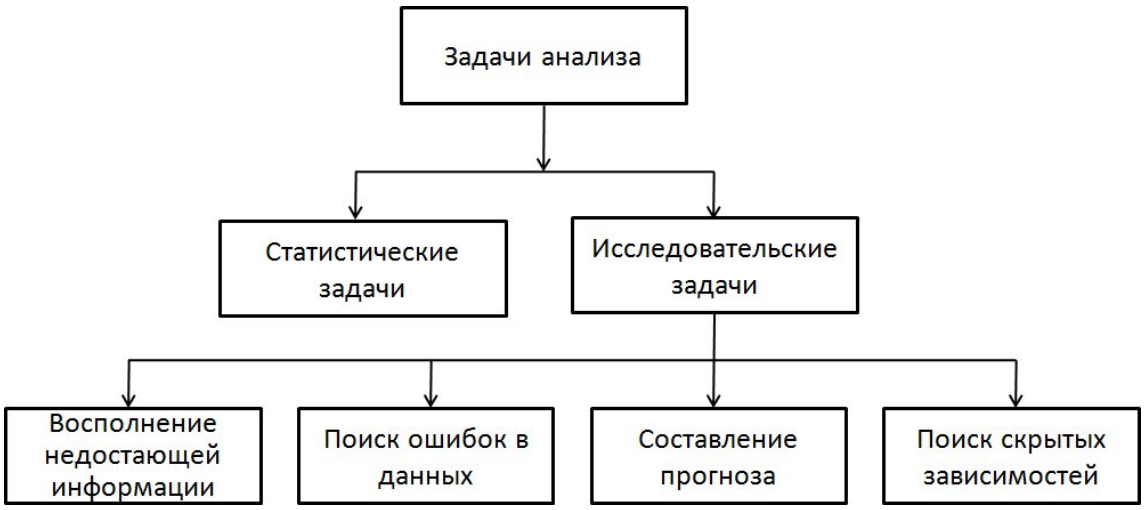

Рис. 2. Классификация задач анализа по типу

---

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

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 (по возможности). 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.

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.

Можно предложить свой вариант темы
(согласовать с преподавателем обязательно)!

При оформлении необходимо соблюдать следующую структуру:

1. Титульный лист
2. Оглавление (сформированное автоматически)
3. Введение
4. Содержательная часть
5. **Теория, решение задачи обязательно проработать и разбираться в алгоритме решения**
6. Заключение
7. Список литературы

Типовые вопросы к зачету (1 семестр):

1. Инструменты анализа данных.
2. Основные методы анализа данных.
3. Прогнозирование при анализе данных.
4. Последовательные модели для анализа данных.
5. Деревья решений.
6. Обработка с запоминанием.
7. Задача классификации.
8. Задача регрессии.
9. Задача прогнозирования.
10. Задача кластеризации.
11. Задача определения взаимосвязей.
12. Задача поиска ассоциативных связей.
13. Анализ текстовой информации.
14. Анализ отклонений.