МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования
«Новосибирский национальный исследовательский государственный университет»

Факультет информационных технологий

**УТВЕРЖДАЮ**
Декан факультета
_____Федотов А.М.

« ___» _____ 2014 г.

# РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ
## Аналитика больших массивов данных

(наименование дисциплины)

Кафедра общей информатики факультета информационных технологий
(*наименование кафедры, обеспечивающей преподавание дисциплины)*

Образовательная программа
09.04.01 ИНФОРМАТИКА И ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА.

(наименование профиля, для дисциплин базовой части не указывается)

Уровень высшего образования
Магистратура

Форма обучения очная

Статус дисциплины: _**вариативная**_____
(*базовая, вариативная, вариативная по выбору, факультативная)*

Объем дисциплины _3_зачетных единицы, в том числе в академических часах по видам учебных занятий:

| Семестр | учебные занятия | | | | | | | | форма промежуточной аттестации (зачет, дифференцированный зачет, экзамен) |
|---|---|---|---|---|---|---|---|---|---|
| | Общий объем | в том числе | | | | | | СРС | |
| | | контактная работа обучающихся с преподавателем | | | | | | | |
| | | Всего | из них | | | | | | |
| | | | Лекции | Лабораторные занятия | Практические занятия | КСР | Консультации | | |
| 1 | 108 | 33 | 16 | | 16 | 1 | | 39 | Экзамен (36) |

Рабочая программа дисциплины составлена в соответствии с требованиями ФГОС ВО к структуре и результатам освоения образовательных программ магистратуры по направлению подготовки 09.04.01 «Информатика и вычислительная техника».

Разработчики:

1) к.ф.-м.н. Павловский Евгений Николаевич, кафедра общей информатики;

2) Зырянов Александр Олегович, кафедра общей информатики.

Рабочая программа дисциплины одобрена на заседании Методической комиссии факультета информационных технологий
от _____ года, протокол № _____.

**Аннотация рабочей программы дисциплины**

Дисциплина «Аналитика больших массивов данных» входит в вариативную часть образовательной программы магистратуры 09.04.01 Информатика и вычислительная техника, магистерская программа «Технология разработки программных системы».

Дисциплина реализуется на факультете информационных технологий НГУ кафедрой общей информатики.

Содержание дисциплины охватывает круг вопросов, связанных с интеллектуальным анализом данных, а именно, его приложением к большим массивам данных.

Дисциплина нацелена на формирование общекультурных ОК-2, ОК-4, ОК-8, общепрофессиональных компетенций ОПК-2, ОПК-5, ОПК-6, и профессиональных компетенций ПК-4, ПК-7, ПК-15 выпускника.

Преподавание дисциплины предусматривает проведение следующих видов учебных занятий: лекций и семинарских занятий.

Рабочая программа дисциплины предусматривает проведение следующих видов контроля: текущий контроль успеваемости в форме приёма заданий и промежуточный контроль в форме экзамена.

Объем дисциплины 3 зачетных единицы, в том числе в академических часах по видам учебных занятий:

| Семестр | учебные занятия | | | | | | | | форма промежуточной аттестации (зачет, дифференцированный зачет, экзамен) |
|---|---|---|---|---|---|---|---|---|---|
| | Общий объем | в том числе | | | | | | | |
| | | контактная работа обучающихся с преподавателем | | | | | | СРС | |
| | | Всего | из них | | | | | | |
| | | | Лекции | Лабораторные занятия | Практические занятия | КСР | Консультации | | |
| 1 | 108 | 33 | 16 | | 16 | 1 | | 39 | Экзамен (36) |

## 1. Цели освоения дисциплины

Курс «Аналитика больших массивов данных» имеет своей целью: формирование у студентов профессиональной компетенции в области разработки и использования систем обработки и анализа больших массивов данных. Данная цель соотносится с целью образовательной программы в части с технологий разработки специализированных программных систем, отвечающих за обработку больших данных. Изучение данной дисциплины готовит выпускника к выполнению следующих профессиональных задач:

- Постановка задачи анализа данных.
- Предварительная обработка данных.
- Визуализация данных.
- Разработка, реализация и применение методов интеллектуального анализа данных к большим массивам данных.
- Представление результатов работы.

## 2. Место дисциплины в структуре образовательной программы

Курс входит в вариативную часть профессионального цикла основной образовательной программы «Технология разработки программных систем» по направлению подготовки 090401 «Информатика и вычислительная техника».

Требования к «входным» знаниям, умениям и готовностям обучающегося в объёме компетенций бакалавра:

1. Материал курса ««Объектно-ориентированный анализ и дизайн»» необходим в части знания основных принципов объектно-ориентированного проектирования программных систем;
2. Материал курса «Программирование» необходим в части знания основных алгоритмов обхода дерева, поиска и сортировки;
3. Материал курса «Логические основы программирования» необходим в части знания основ лямбда-исчисления, основ логического программирования;
4. Материал курса «Математическая логика и теория алгоритмов» необходим в части знания принципов вычислимости и разрешимости, понимания признаков алгоритмически неразрешимых проблем, знакомства с нечёткими логиками.
5. Материал курса «Теория вероятности и математическая статистика» необходим в части знания формул комбинаторики, функций распределения случайной величины, проверки гипотез, критериев согласия, исследования статистической зависимости.

6. Материал курса «Методы оптимизации» необходим в части знания задач выпуклого программирования, градиентных методов, методов отсечения.
7. Материал курса «Интеллектуальный анализ данных» необходим в части понимания основных задач анализа данных, умения классифицировать практические задачи с их помощью, владения методами интеллектуального анализа данных.

Курс «Аналитика больших массивов данных» необходим для тех магистрантов, чья работа по диссертации связана с обработкой и анализом больших массивов данных, а также с созданием инструментов для такой обработки и анализа.

## 3. Компетенции обучающегося, формируемые в результате освоения дисциплины (перечень планируемых результатов обучения)

В результате освоения данной дисциплины обучающийся демонстрирует следующие общекультурные, общепрофессиональные и (или) профессиональные:

| Код компетенции | Формулировка компетенции из ФГОС | Планируемые результаты обучения (показатели достижения заданного уровня освоения компетенций) |
|---|---|---|
| ОК-2 | Способность понимать роль науки в развитии цивилизации, соотношение науки и техники, иметь представление о связанных с ними современных социальных и этических проблемах, понимать ценность научной рациональности и ее исторических типов. | Понимать важность феномена больших данных для развития общества и науки. Знать причины возникновения тренда больших данных. Знать проблемы и возможности, связанные с появлением больших данных.<br>Понимать важность применения научных методов для извлечения пользы из больших массивов данных |

| | | |
|---|---|---|
| ОК-4 | Способность заниматься научными исследованиями | Понимать возможности технологий анализа больших данных при проведении научных исследований.<br>Уметь применять научные методы, в т.ч. методы интеллектуального анализа данных, к большим данным. |
| ПК-4 | Владение существующими методами и алгоритмами решения задач распознавания и обработки данных. | Уметь формулировать алгоритмы в парадигме Map Reduce.<br>Владеть методами интеллектуального анализа данных, в т.ч. методами оценки качества моделей, алгоритмов, методами экспериментальной проверки гипотез, методами обоснования гипотез. |
| ОК-8 | Способность к профессиональной эксплуатации современного оборудования и приборов (в соответствии с целями магистерской программы). | Уметь: выбрать подходящую технологию хранения и обработки больших данных, использовать современные высоконагруженные системы хранения и обработки больших данных. |
| ОПК-5 | Владение методами и средствами получения, хранения, переработки и трансляции информации посредством современных компьютерных технологий, в том числе, в глобальных компьютерных сетях | Знать: существующие современные технологии высоконагруженных систем хранения и обработки данных, принципы работы высоконагруженных систем.<br>Владеть: технологией Map Reduce и ее реализацией Hadoop. |
| ОПК | Культурой мышления, спо- | Знать: существующие в современном |

| | | |
|---|---|---|
| -2 | собностью выстраивать логику рассуждений и высказываний, основанных на интерпретации данных, интегрированных их разных областей науки и техники, выносить суждения на основании неполных. | мире источники данных.<br>Уметь: интегрировать данные из разных источников, интерпретировать их в контексте поставленной задачи, делать выводы, основанные на анализе полученных данных.<br>Владеть: методами получения данных из различных доступных источников. |
| ОПК-6 | Способность анализировать профессиональную информацию, выделять в ней главное, структурировать, оформлять и представлять в виде аналитических обзоров с обоснованными выводами и рекомендациями. | Знать: существующие в современном мире источники и типы информации.<br>Уметь: визуализировать имеющиеся данные, отбрасывать несущественную информацию, структурировать информацию в рамках поставленной задачи.<br>Владеть: современными средствами визуализации, методами предварительной подготовки данных. |
| ПК-7 | Применение перспективных методов исследования и решения профессиональных задач на основе знания мировых тенденций развития вычислительной техники и информационных технологий. | Знать: тенденции больших данных.<br>Уметь: формулировать бизнес-задачи в терминах анализа данных. |
| ПК-15 | Способность к созданию программного обеспечения для анализа, распознавания и обработки информации, систем цифровой обработки сигналов | Знать: основные элементы процесса анализа больших данных, основные подходы к обработке больших массивов данных. |

## 4. Объем, структура и содержание дисциплины

*4.1. Общая трудоемкость курса составляет 3 зачетных единицы, 108 часов.*

*4.2. Структура дисциплины*

| № п/п | Раздел (тема) дисциплины | Семестр (из учебного плана) | Неделя семестра (из учебного плана) | Контактная работа обучающихся с преподавателем по видам учебных занятий *(из учебного плана, в часах)* | | | | | Самостоятельная работа обучающихся *(из учебного плана, в часах)* | Формы текущего контроля успеваемости *(по неделям семестра)*<br><br>Форма промежуточной аттестации *(по семестрам, из учебного плана)* |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | лекции | Лабораторные занятия | Практические занятия | Контроль самостоятельной работы (КСР) | Консультации | | |
| 1 | **Введение в большие данные.** | 1 | 1 | 2 | | | | | 2 | Тест (пятиминутка) |
| 2 | **Жизненный цикл анализа больших данных.** | | 3 | 2 | | | | | 2 | Тест (пятиминутка) |
| 3 | **Корреляция и регрессия. Их роль в аналитике больших данных.** | | 2,4 | 2 | | 2 | | | 4 | Контрольное задание |
| 4 | **Задачи классификации и кластеризации. Ассоциативные правила.** | | 4,6 | 2 | | 2 | | | 8 | Контрольное задание |
| 5 | **Языки Python и R, стек библиотек анализа данных. Готовые решения анализа данных (Weka и т.д.).** | | 6,8 | 2 | | 4 | | | 8 | Контрольное задание |
| 6 | **Преподготовка данных. Визуализация данных. Понима-** | | 8,10 | 2 | | 2 | | | 4 | Контрольное задание |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **ние данных.** | | | | | | | | | |
| 7 | **Парадигма Map Reduce. Ее реализация Hadoop.** | | 10, 12 | 2 | | 2 | | | 5 | Контрольное задание |
| 8 | **Проблема переобучения. Регуляризация. Нейронные сети. Машина опорных векторов.** | | 12, 14 | | | 4 | | | 4 | Контрольное задание |
| 9 | **Научные проблемы в области больших данных** | | 15-16 | 2 | | | | | 2 | Тест (пятиминутка) |
| 10 | **Приём заданий** | | 16 | | | | 1 | | | |
| 11 | **Выходной контроль** | | 17 | | | | | | 36 | Экзамен |
| | **Итого за семестр** | | | 16 | | 16 | 1 | | 75 | 108 |

## *4.3. Содержание дисциплины, структурированное по темам (разделам)*

Лекции – *16 часов (из учебного плана)*

| Раздел (тема), Код компетенции | № занятия | Содержание занятий и ссылки на рекомендуемую литературу | Кол-во часов | |
|---|---|---|---|---|
| | | | всего | В интерактивной форме |
| Тема1. (ОК-2) | 1 | Введение в большие данные. Роль аналитика данных (Data Scientist). Ключевые компетенции аналитика. Отличия BI от Data Science.<br>Литература:<br>1. C. B. B. D. Manyika, “Big Data: The Next Frontier for Innovation, Competition, and Productivity,” McKinsey Global Institute, 2011. URL: http://www.mckinsey.com/~/media/McKinsey/dotcom/Insights%20and%20pubs/MGI/Research/Technology%20and%20Innovation/Big%20Data/MGI_big_data_full_report.ashx<br>2. Виктор Маер-Шенбергер, Кеннет Кукьер. Большие данные: Революция, которая изменит то, как мы живем, работаем и мыслим. — М.: «Манн, Иванов и Фербер», 2013, 240 с. ISBN 978-5-91657-936-9 (http://www.mann-ivanov-ferber.ru/books/paperbook/big_data/) | 2 | |

| | | | | |
|---|---|---|---|---|
| Тема2 (ОПК-2, ОПК-6) | 2 | Жизненный цикл анализа больших данных. «Песочница».<br>Литература:<br>1. Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data // EMC Education Services. 2015. — 432p. — ISBN: 978-1-118-87613-8.<br>2. DJ Patil. Building Data Science Teams. O'Reilly. 2011. ISBN: 978-1-449-31623-5 (http://cdn.oreilly.com/radar/2011/09/Building-Data-Science-Teams.pdf) | 2 | |
| Тема3 (ПК-4) | 3 | Корреляция и регрессионный анализ. Коэффициент корреляции. Графическое представление. Постановка задачи регрессионного анализа. Линейная регрессия. Метод наименьших квадратов. Их роль в аналитике больших данных.<br>Литература:<br>1. Лбов, Геннадий Сергеевич. Анализ данных и знаний : учебное пособие / Г.С. Лбов ; Федер. агентство по образованию, Новосиб. гос. ун-т, Мех.-мат. фак .— Новосибирск : Новосибирский государственный университет, 2010 .— 107 с. | 2 | |
| Тема4 (ОК-4) | 4 | Задачи классификации и кластеризации. Ассоциативные правила.<br>Литература:<br>1. Trevor Hastie, Elements of statistical learning, Springer, 2009. (URL:http://statweb.stanford.edu/~tibs/ElemStatLearn/ ) | 2 | |
| Тема5 (ПК-15) | 5 | Языки Python и R. Синтаксис языка R, основные типы данных.<br>Литература:<br>1. J. Adler. R in a Nutshell. Second Edition. O'Reilly Media Inc. 2012. ISBN: 978-1-449-31208-4 | 2 | |
| Тема6 (ОК-8, ОПК-2, ОПК-5, ОПК-6) | 6 | Подготовка данных. Визуализация данных. Понимание данных.<br>Литература:<br>1. Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data // EMC Education Services. 2015. — 432p. — ISBN: 978-1-118-87613-8. | 2 | |
| Тема7 (ОПК-5, ПК-4) | 7 | Парадигма Map Reduce. Ее реализация Hadoop.<br>Литература:<br>1. Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data // EMC Education Services. 2015. — 432p. — ISBN: 978-1-118-87613-8. | 2 | |

| | | | | |
|---|---|---|---|---|
| Тема8 (ОК-2, ОК-4) | 8 | Научные проблемы в области больших данных<br>Литература:<br>1. Frontiers in Massive Data Analysis, National Research Council, 2013 http://www.nap.edu/catalog/18374/frontiers-in-massive-data-analysis<br>2. J. Hopcroft, R. Kannan. Foundations of Data Science. 2013. — 412 p. (https://www.dropbox.com/s/j2s5dn5w5g7ics5/Data%20Science%20Foundations%20book-dec-30-2013.pdf) | 2 | |
| | | Итого: | 16 | |

Практические (семинарские) занятия – *16 часов (из учебного плана)*

| Раздел (тема), Код компетенции | № занятия | Содержание занятий и ссылки на рекомендуемую литературу | Кол-во часов | |
|---|---|---|---|---|
| | | | всего | В интерактивной форме |
| Тема1 Корреляция и регрессия. Их роль в аналитике больших данных. (ОК-2, ОК-4 ПК-4) | 1 | Корреляция. Регрессионный анализ. Задачи в области больших данных, решаемые методом регрессионного анализа.<br>Литература:<br>Лбов, Геннадий Сергеевич. Анализ данных и знаний : учебное пособие / Г.С. Лбов ; Федер. агентство по образованию, Новосиб. гос. ун-т, Мех.-мат. фак .— Новосибирск : Новосибирский государственный университет, 2010 .— 107 с. | 2 | 2 |
| Тема 2 Задачи классификации и кластеризации. Ассоциативные правила (ОК-2, ОК-4 ПК-4, ПК-15) | 3 | Постановка задачи классификации. Постановка задачи кластеризации. Задача построения ассоциативных правил.<br>Литература:<br>Trevor Hastie, Elements of statistical learning, Springer, 2009.<br>(URL:http://statweb.stanford.edu/~tibs/ElemStatLearn/ ) | 4 | 4 |
| Тема 3 Языки Python и R, стек библиотек анализа данных. Готовые решения анализа данных (Weka и т.д.) (ОПК-5, ПК-15) | 6 | Роль языков программирования Python и R в аналитике больших данных. Необходимый набор библиотек. Готовые решения анализа данных и их роль в области больших данных.<br>Литература:<br>J. Adler. R in a Nutshell. Second Edition. O'Reilly Media Inc. 2012. ISBN: 978-1-449-31208-4 | 2 | 2 |

| | | | | |
|---|---|---|---|---|
| Тема 4 Подготовка данных. Визуализация данных. Понимание данных. (ОПК-5, ОПК-6) | 7 | Методы предварительной подготовки данных. Инструменты и методы визуализации данных. Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data // EMC Education Services. 2015. — 432p. — ISBN: 978-1-118-87613-8. | 2 | 2 |
| Тема 5 Парадигма Map Reduce. Ее реализация Hadoop. *(ОПК-5, ОПК-6, ПК-15)* | 8 | Парадигма Map Reduce. Роль Map Reduce в аналитике больших данных. | 2 | 2 |
| Тема 6 Проблема переобучения. Регуляризация. Нейронные сети. Машина опорных векторов. (ОК-2, ОК-4 ПК-4, ПК-15) | 9 | Проблема переобучения и регуляризация. Разбор алгоритма нейронных сетей. Разбор алгоритма SVM<br>Литература:<br>Trevor Hastie, Elements of statistical learning, Springer, 2009.<br>(URL:http://statweb.stanford.edu/~tibs/ElemStatLearn/ ) | 4 | 4 |
| | | Итого: | **16** | **16** |

*Самостоятельная работа – 76 часов (из учебного плана)*

*Самостоятельная работа студентов предусматривает выполнение следующих заданий:*

| Раздел (тема), Код компетенции | *№ темы* | *Содержание темы для самостоятельного изучения и ссылки на литературу* | *Кол-во часов* | *Форма контроля* |
|---|---|---|---|---|
| *Тема1* Корреляция и регрессия. Их роль в аналитике больших данных. (ОК-2, ОК-4 ПК-4) | *1* | *Линейная регрессия*<br>Литература:<br>Trevor Hastie, Elements of statistical learning, Springer, 2009.<br>(URL:http://statweb.stanford.edu/~tibs/ElemStatLearn/ ) | *2* | *Прием заданий* |
| | *2* | *Метод наименьших квадратов*<br>Литература:<br>Trevor Hastie, Elements of statistical learning, Springer, 2009.<br>(URL:http://statweb.stanford.edu/~tibs/ElemStatLearn/ ) | *3* | *Прием заданий* |
| | *3* | *Примеры использования корреляции и регрессионного анализа в области больших данных*<br>Литература:<br>Trevor Hastie, Elements of statistical learning, Springer, 2009.<br>(URL:http://statweb.stanford.edu/~tibs/ElemStatLearn/ ) | *2* | *Прием заданий* |

| | | | | |
|---|---|---|---|---|
| *Тема 2* Задачи классификации и кластеризации. Ассоциативные правила (ОК-2, ОК-4 ПК-4, ПК-15) | *4* | *Логистическая регрессия*<br>Литература:<br>Trevor Hastie, Elements of statistical learning, Springer, 2009.<br>(URL:http://statweb.stanford.edu/~tibs/ElemStatLearn/ ) | *5* | *Прием заданий* |
| | *5* | *Наивный классификатор Байеса*<br>Литература:<br>Воронцов К.В. Математические методы обучения по прецедентам<br>http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf | *5* | *Прием заданий* |
| | *6* | *Алгоритм априори* | *5* | *Прием заданий* |
| | *7* | *Алгоритм k-means. Матрица парных расстояний. Виды метрик.*<br>Литература:<br>Trevor Hastie, Elements of statistical learning, Springer, 2009.<br>(URL:http://statweb.stanford.edu/~tibs/ElemStatLearn/ ) | *5* | *Прием заданий* |
| *Тема 3* Языки Python и R, стек библиотек анализа данных. Готовые решения анализа данных (Weka и т.д.) (ОПК-5) | *8* | *Анализ стандартных наборов данных (iris, mtcars и т.д.) при помощи Weka или Orange.* | *4* | *Прием заданий* |
| | *9* | *Пройти онлайн курс R Programming [5]* | *28* | *Прием заданий* |
| *Тема 4* Преподготовка данных. Визуализация данных. Понимание данных. (ОПК-5) | *10* | *Визуализация стандартных наборов данных при помощи Tableau* | *3* | *Прием заданий* |
| | *11* | *Изучение части курса Introduction to Data Science [6], посвященной визуализации.* | *1* | *Прием заданий* |
| *Тема 5* Парадигма Map | *12* | *Подсчет кол-ва слов, вычисление индекса TF-IDF, реализация алгоритма k-means в рамках парадигмы Map Reduce с использованием Hadoop.* | *5* | *Прием заданий* |

| | | | | |
|---|---|---|---|---|
| Reduce. Ее реализация Hadoop. *(ОПК-5, ОПК-6, ПК-15)* | | | | |
| Тема 6 Проблема переобучения. Регуляризация. Нейронные сети. Машина опорных векторов. (ОК-2, ОК-4 ПК-4, ПК-15) | *13* | *Регуляризация для метода наименьших квадратов.*<br>Литература:<br>Trevor Hastie, Elements of statistical learning, Springer, 2009.<br>(URL:http://statweb.stanford.edu/~tibs/ElemStatLearn/ ) | *3* | *Прием заданий* |
| | *14* | *Реализовать нейронную сеть или машину опорных векторов.*<br>Литература:<br>Trevor Hastie, Elements of statistical learning, Springer, 2009.<br>(URL:http://statweb.stanford.edu/~tibs/ElemStatLearn/ )<br>Воронцов К.В. Математические методы обучения по прецедентам<br>http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf | *5* | *Прием заданий* |
| | | | *итого* | *76* |

## 5. Образовательные технологии

До начала каждой новой лекции студенты имеют возможность ознакомиться с её заголовками. Это ознакомление служит элементом ориентирования во время изучения нового материала и позволяет студенту внимательно следить за материалом новой лекции.

Знание теоретического материала закрепляется тестами, проводимыми с помощью системы опроса всех слушателей в виде небольшого теста. Тесты проводятся в первые 15 минут лекции. Тесты проводятся с целью выявления непонимания среди слушателей и служат элементом обратной связи от слушателей к лектору для своевременной корректировки изложения и дополнения пройденного.

На практических занятиях студенты выполняют контрольные задания, в которых закрепляют теоретический материал, имеют возможность наработать умение использовать программные продукты для проведения анализа больших данных. За каждое выполненное задание начисляются баллы, сумма которых определяет возможность допуска к экзамену.

## 6. Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине

Методические рекомендации по самостоятельной работе обучающихся приводятся в приложении к рабочей программе дисциплины (Приложение А)

## 7. Фонд оценочных средств для проведения текущей и промежуточной аттестации обучающихся по дисциплине

7.1. Перечень компетенций с указанием этапов их формирования приведен в описании образовательной программы

| Компетенция | Знания, умения, навыки | Процедура оценивания |
|---|---|---|
| Способность понимать роль науки в развитии цивилизации, соотношение науки и техники, иметь представление о связанных с ними современных социальных и этических проблемах, понимать ценность научной рациональности и ее исторических типов (ОК-2) | Понимать важность феномена больших данных для развития общества и науки. | Дискуссия |
| | Знать причины возникновения тренда больших данных. | Тест-пятиминутка |
| | Знать проблемы и возможности, связанные с появлением больших данных. | Тест-пятиминутка |
| | Понимать важность применения научных методов для извлечения пользы из больших массивов данных | Дискуссия |
| Способность заниматься научными исследованиями (ОК-4) | Понимать возможности технологий анализа больших данных при проведении научных исследований. | Тест-пятиминутка |
| | Уметь применять научные методы, в т.ч. методы интеллектуального анализа данных, к большим данным. | Домашнее задание |
| Владение существующими методами и алгоритмами решения задач | Уметь формулировать алгоритмы в парадигме Map Reduce. | Домашнее задание |

| | | |
|---|---|---|
| распознавания и обработки данных (ПК-4) | Владеть методами интеллектуального анализа данных, в т.ч. методами оценки качества моделей, алгоритмов, методами экспериментальной проверки гипотез, методами обоснования гипотез. | Домашнее задание |
| Способность к профессиональной эксплуатации современного оборудования и приборов (в соответствии с целями магистерской программы) (ОК-8) | Уметь: выбрать подходящую технологию хранения и обработки больших данных, использовать современные высоконагруженные системы хранения и обработки больших данных. | Домашнее задание |
| Владение методами и средствами получения, хранения, переработки и трансляции информации посредством современных компьютерных технологий, в том числе, в глобальных компьютерных сетях (ОПК-5) | Знать: существующие современные технологии высоконагруженных систем хранения и обработки данных, принципы работы высоконагруженных систем. | Тест-пятиминутка |
| | Владеть: технологией Map Reduce и ее реализацией Hadoop. | Домашнее задание |
| Культурой мышления, способностью выстраивать логику рассуждений и высказываний, основанных на интерпретации данных, интегрированных их разных областей науки и техники, выносить суждения на основании неполных (ОПК-2) | Знать: существующие в современном мире источники данных. | Тест-пятиминутка |
| | Уметь: интегрировать данные из разных источников, интерпретировать их в контексте поставленной задачи, делать выводы, основанные на анализе полученных данных. | Домашнее задание |
| | Владеть: методами получения данных из различных доступных источников. | Домашнее задание |

| | | |
|---|---|---|
| Способность анализировать профессиональную информацию, выделять в ней главное, структурировать, оформлять и представлять в виде аналитических обзоров с обоснованными выводами и рекомендациями (ОПК-6) | Знать: существующие в современном мире источники и типы информации. | Тест-пятиминутка |
| | Уметь: визуализировать имеющиеся данные, отбрасывать несущественную информацию, структурировать информацию в рамках поставленной задачи. | Домашнее задание |
| | Владеть: современными средствами визуализации, методами предварительной подготовки данных. | Домашнее задание |
| Применение перспективных методов исследования и решения профессиональных задач на основе знания мировых тенденций развития вычислительной техники и информационных технологий (ПК-7) | Знать: тенденции больших данных. | Тест-пятиминутка |
| | Уметь: формулировать бизнес-задачи в терминах анализа данных. | Дискуссия |
| Способность к созданию программного обеспечения для анализа, распознавания и обработки информации, систем цифровой обработки сигналов (ПК-15) | Знать: основные элементы процесса анализа больших данных, основные подходы к обработке больших массивов данных. | Тест-пятиминутка |

7.2. Описание показателей и критериев оценивания компетенций, описание шкал оценивания

Выделяются три показателя уровня сформированности компетенции:

| Уровень | Показатели (что обучающийся должен продемонстрировать) | Оценочная шкала | | |
|---|---|---|---|---|
| | | удовлетворительно | хорошо | отлично |

| | | | | |
|---|---|---|---|---|
| Пороговый | Понимать важность феномена больших данных для развития общества и науки. | Имеет фрагментарное представление о феномене больших данных и его влиянии на общество и науку. | Может сформулировать несколько факторов влияния больших данных на науку | Демонстрирует целостное представление о феномене больших данных в контексте развития общества. |
| Базовый | Знать причины возникновения тренда больших данных. | Называет причины правильно, но не может объяснить. | Объясняет как минимум одну причину и ее генезис. | Имеет целостное понимание причин, а также что не является причиной. Может фильтровать маркетинговый шум. |
| Базовый | Знать проблемы и возможности, связанные с появлением больших данных. | Имеет представление о проблемах больших данных, но не может объяснить их происхождение. | Имеет представление о проблемах и их происхождении, способен назвать возможности, открываемые анализом больших данных. | Имеет целостное представление о проблемах и возможностях больших данных. |
| Базовый | Понимать важность применения научных методов для извлечения пользы из больших массивов данных | Называет один научный метод, и приводит пример извлечения пользы из больших массивов данных, с помощью данного метода. | Отличает применение научных от ненаучных методов в анализе больших данных. | Может обосновать применение того или иного метода анализа больших данных. |
| Продвинутый | Понимать возможности технологий анализа больших данных при проведении научных исследований. | Называет несколько конкретных случаев, где технологий анализа больших данных принесли пользу при проведении научных исследований. | Объясняет, каким образом польза была получена. | Имеет целостное понимание о областях применимости той или иной технологии анализа больших данных. |
| Базовый | Уметь применять научные методы, в т.ч. методы интеллектуального анализа данных, к большим данным. | Хорошо применяет один метод, с пониманием. | Может применить более одного метода, с пониманием. | Может применять все основные методы интеллектуального анализа данных. Может обосновать выбор того или иного метода. |
| Базовый | Уметь формулировать алгоритмы в парадигме Map Reduce. | Может сформулировать один алгоритм, но затрудняется объяснить его работу. | Объясняет работу алгоритма, но допускает неточности. | Безошибочно может сформулировать алгоритм в парадигме Map Reduce |

| | | | | |
|---|---|---|---|---|
| Базовый | Владеть методами интеллектуального анализа данных, в т.ч. методами оценки качества моделей, алгоритмов, методами экспериментальной проверки гипотез, методами обоснования гипотез. | Хорошо применяет один метод, с пониманием. | Может применить более одного метода, с пониманием. | Может применять все основные методы интеллектуального анализа данных. Может обосновать выбор того или иного метода. Может оценить качество работы метода. |
| Базовый | Уметь: выбрать подходящую технологию хранения и обработки больших данных, использовать современные высоконагруженные системы хранения и обработки больших данных. | Правильно выбирает технологию под поставленную задачу, но не может обосновать свой выбор. | Правильно выбирает технологию под поставленную задачу. В некоторых случаях способен обосновать свой выбор. | Может обоснованно выбрать конкретную технологию под поставленную задачу из множества представленных технологий. |
| Пороговый | Знать: существующие современные технологии высоконагруженных систем хранения и обработки данных, принципы работы высоконагруженных систем. | Знает несколько, но не все основные технологии хранения и обработки больших данных. | Знает наиболее распространенные технологии хранения и обработки больших данных | Знает наиболее распространенные технологии хранения и обработки больших данных. Знает их отличия, возможности и недостатки. |
| Базовый | Владеть: технологией Map Reduce и ее реализацией Hadoop. | Владеет базовым пониманием принципов действия парадигмы Map Reduce.. | Может объяснить, как на практике применить парадигму Map Reduce. Может привести пример подсчета кол-ва слов. | Демонстрирует целостное знание как парадигмы Map Reduce, так и ее реализации Hadoop. Знает сильные и слабые стороны, а так же области применимости данной парадигмы и ее реализации. |
| Пороговый | Знать: существующие в современном мире источники данных. | Имеет поверхностное знание об источниках данных. | Способен детально рассказать об нескольких источниках данных. | Способен взаимодействовать с несколькими источниками данных. |

| | | | | |
|---|---|---|---|---|
| Базовый | Уметь: интегрировать данные из разных источников, интерпретировать их в контексте поставленной задачи, делать выводы, основанные на анализе полученных данных. | Может проводить интеграцию стандартных данных с помощью коллег, товарищей или преподавателя. | Самостоятельно интегрирует данные. | Интегрирует данные из нестандартных источников, понимает и объясняет методы, которые использует. |
| Базовый | Владеть: методами получения данных из различных доступных источников. | Владеет одним методов получения данных из стандартных источников. | Владеет несколькими методами получения данных из стандартных источников. | Владеет несколькими методами получения данных из нестандартных источников. |
| Базовый | Уметь: визуализировать имеющиеся данные, отбрасывать несущественную информацию, структурировать информацию в рамках поставленной задачи. | Использует стандартные графики. | Правильно определяет способ визуализации выбранных данных. | Свободно ориентируется в способах визуализации, предлагает и обосновывает их для конкретной задачи. |
| Продвинутый | Владеть: современными средствами визуализации, методами предварительной подготовки данных. | Знает основные средства визуализации и может ими воспользоваться. | Способен осмысленно использовать методы визуализации и преподготовки данных. | Способен сформировать понимание данных и предложить метод их обработки на основе визуализации данных. |
| Базовый | Знать: тенденции больших данных. | Называет основные тенденции больших данных. | Может называть несколько причин развития больших данных. | Знает и понимает тенденции больших данных, может прогнозировать их развитие. |
| Базовый | Уметь: формулировать бизнес-задачи в терминах анализа данных. | Может привести два примера бизнес-задач в терминах анализа данных. | Может перевести стандартную бизнес-задачу на язык анализа данных. | Безошибочно определяет класс задачи анализа данных по бизнес-задаче, может определять класс нетривиальных бизнес-задач. |

| | | | | |
|---|---|---|---|---|
| Базовый | Знать: основные элементы процесса анализа больших данных, основные подходы к обработке больших массивов данных. | Называет все элементы процесса анализа данных. | Может объяснить сущность некоторых элементов. | Называет все элементы в правильной последовательности и может объяснить сущность каждого. |

Если хотя бы одна из компетенций не сформирована, то положительной оценки по дисциплине быть не может.

7.3. Типовые контрольные задания

1) Текущий контроль. В течении семестра студенты выполняют 11 контрольных заданий. Выполнение контрольных заданий является обязательным для всех студентов, а результаты текущего контроля служат основанием для выставления оценок в ведомость контрольной недели на факультете. В случае невыполнения более 3х контрольных заданий студент не допускается до экзамена. Решение о допуске принимает лектор вместе с ассистентом, ведущим практические занятия в данной группе. Отсутствие более одного контрольного задания являться поводом для снижения оценки за экзамен.

Тесты (пятиминутки) (выбрать правильный ответ, или указать число):

1. О соотношении аналоговой и цифровой информации:
    1. Большинство данных в мире в 2011 году содержалось:
        i. В цифровом виде
        ii. В аналоговом виде
    2. В каком веке произошёл перевес объёмов накопленных человечеством данных в сторону цифровых?
    3. Объём накопленных человечеством цифровых данных на 2012 год измеряется:
        i. Петабайтами
        ii. Зеттабайтами
        iii. Экзабайтами
        iv. Йоттабайтами
    4. Сколько Петабайт в Зеттабайте?
2. История больших данных
    1. Укажите фактор, способствовавший появлению тренда больших данных
        i. Маркетинговые кампании крупных корпораций
        ii. Снижение издержек на хранение данных
        iii. Появление новых технологий обработки потоковых данных
        iv. Выпуск баз данных с обработкой данных в памяти
    2. Какие вероятные разочарования тренда больших данных?

i. Из-за угрозы безопасности личной жизни (privacy) граждан будут усложнены процедуры сбора данных, что приведёт к падению ценности больших данных.

3. Отметьте значимые события, повлиявшие на формирование тренда больших данных:
   i. Разработка Hadoop
   ii. Изобретение принципа MapReduce
   iii. Разработка языка Pyhton
   iv. Победа Deepblue в матче с Г.Каспаровым.

3. Определение больших данных:
   1. Выберите верный ответ
      i. Большие данные – это обработка или хранение более 1 Тб информации.
      ii. Проблема больших данных – это такая проблема, когда при существующих технологиях хранения и обработки сущностная обработка данных затруднена или невозможна.
      iii. Большие данные – это огромная PR-акция крупных вендоров и не более того.
      iv. Большие данные – это явление, когда цифровые данные наиболее полно представляют изучаемый объект.
   2. Выберите неверный ответ:
      i. Большие данные – это данные объёма свыше 1 Тб
      ii. Проблема больших данных – это проблема, когда при существующих технологиях хранения и обработки сущностная обработка данных затруднена или невозможна.
      iii. Большие данные – это тренд в области ИТ, подогреваемый маркетинговыми кампаниями крупных вендоров.
      iv. Большие данные как правило не структурированы.
   3. Отметьте те из вариантов, в которых данные структурированы:
      i. Данные о продажах компании, представленные в виде помесячных отчётов в формате MS Word.
      ii. Таблица с ежедневными показаниями температуры помещения за год в файле формата csv.
      iii. Текст педагогической поэмы А.С. Макаренко, представленный в формате PDF.
      iv. Библиотека фильмов, представленных в формате mpeg4 на одном жестком диске.

4. Характеристики Big Data:
   1. Перечислите четыре основных характеристики Big Data:
      i. Virtualization, Volume, Variability, Vehicle
      ii. Variety, Velocity, Volume, Value
      iii. Verification, Volume, Velocity, Visualization
      iv. Video, Value, Variety, Volume
   2. Выберите неверное высказывание:

i. Большие объёмы данных приводят к слабой их структуризации, поэтому появляется такое разнообразие данных.
ii. Увеличившаяся производительность телекоммуникационных каналов привела к росту объёмов передаваемой информации.
iii. Удешевление систем хранения на единицу информации привело к росту рынка больших данных.
iv. Большое разнообразие источников данных

3. Отметьте неверное понимание Variety в контексте характеристик Big Data:
    i. Высокая скорость генерирования данных.
    ii. Разные типы данных в колонках таблиц реляционных СУБД.
    iii. Разнообразие отраслей, являющихся источниками данных.
    iv. Разнообразие типов данных, включающих в себя структурированные, полуструктурированные и неструктурированные.

5. Принцип MapReduce
    1. Принцип MapReduce состоит в том, чтобы
        i. Производить вычисления на узлах, где информация изначально была сохранена
        ii. Использовать вычислительные мощности систем хранения
        iii. Использовать функциональное программирование для решения задач массивно-параллельной обработки
    2. Выберите одно неверное высказывание про MapReduce:
        i. Интерфейс для массово-параллельной обработки данных, где вычисления производятся на узлах, где информация изначально была сохранена
        ii. MapReduce – это две операции: распределения и сборки данных
        iii. MapReduce был придуман разработчиками Hadoop
        iv. MapReduce был анонсирован разработчиками Google
    3. Каков теоретический прирост производительности при подсчёте числа слов в тексте при работе MapReduce при переходе от одного узла к двум?
6. Технологии хранения
    1. Какие из следующих технологий СУБД не используют принцип MapReduce
        i. Hadoop
        ii. Cassandra
        iii. HDInsight
        iv. Redis
    2. Какие СУБД полностью полагаются на оперативную память при хранении информации:
        i. Oracle Exalytics
        ii. SAP HANA
        iii. BigTable
        iv. HBase

3. В чём преимущество колоночно-ориентированных СУБД?
    i. Они позволяют выполнять более сложные SQL-запросы по сравнению с реляционными СУБД
    ii. Они позволяют динамически дополнять содержание записей новыми полями
    iii. Они имеют более гибкие возможности аналитики.
    iv. Они позволяют эффективно делать межколоночные сравнения.

7. «Песочница» в аналитическом процессе
    1. Для чего аналитику необходима «песочница»?
        i. Для высокопроизводительной аналитики за счёт использования оперативной памяти и inDB операций.
        ii. Для хранения всех полученных от заказчика данных.
        iii. Для построения отчётов о результатах анализа
        iv. Для снижения затрат, связанных с репликацией данных
    2. Какие из следующих средств разумно использовать для анализа данных, представленных единственным csv-файлом размера более 100Гб:
        i. Hadoop
        ii. Data Warehouse
        iii. «Песочница»
        iv. Python
    3. Выберите верное утверждение:
        i. Data Warehouse создаются для проверки гипотез при анализе больших данных.
        ii. «Песочница» используется для снижения нагрузки на основной Data Warehouse.
        iii. Каждый Data Warehouse должен содержать «песочницу».
        iv. «Песочница» необходима для любого процесса аналитики.

8. CRISP-DM
    1. Расставьте последовательность этапов проекта аналитики в соответствии с CRISP-DM.
        i. Понимание бизнеса (Business understanding)
        ii. Понимание данных (Data Understanding)
        iii. Подготовка данных (Data Preparation)
        iv. Моделирование (Modeling)
        v. Оценка (Evaluation)
        vi. Внедрение (Deployment)
    2. На каком из этапов процесса CRISP-DM происходит проверка гипотез?
        i. Понимание бизнеса (Business understanding)
        ii. Понимание данных (Data Understanding)
        iii. Моделирование (Modeling)
        iv. Оценка (Evaluation)

3. Вы являетесь владельцем и аналитиком в компании из 10 человек, в которой требуется проанализировать продажи за 1 год (1 млн. продаж). Какие из этапов CRISP-DM можно опустить:
    i. Понимание бизнеса (Business understanding)
    ii. Подготовка данных (Data Preparation)
    iii. Моделирование (Modeling)
    iv. Оценка (Evaluation)

9. Hadoop
    1. Пример благоразумного использования Hadoop
        i. Анализ 10 Гб данных.
        ii. Ежедневное сохранение данных температуры, поступающих со всех городов России (по одному показанию на город, всего городов 1100 шт).
        iii. Посекундное сохранение данных температуры, поступающих со всех городов России (по одному показанию на город, всего городов 1100 шт).
        iv. Построение графика пульса пациента в реальном времени.
    2. Начиная с каких размеров данных обоснованно применение кластера Hadoop для хранения данных?
        i. 100Гб
        ii. 1Тб
        iii. 100Тб
        iv. 1Пб
    3. Hadoop – это:
        i. Набор утилит, и программный каркас для выполнения распределённых программ, работающих на кластерах.
        ii. Распределённая СУБД, позволяющая обрабатывать большие данные.
        iii. Язык выполнения заданий в парадигме MapReduce.
        iv. Распределённая файловая система, предназначенная для хранения файлов большого объёма.

2) Промежуточная аттестация – экзамен. Студент случайным образом выбирает билет, содержащий вопрос по теории, читаемой в лекционной части курса, и вопрос, касающийся практической части курса. Если студент не сдал какие-либо контрольные задания, то он может попросить дополнительный вопрос по данным темам, чтобы избежать снижения оценки за экзамен.
3) Темы контрольных заданий.

1. Метод наименьших квадратов применительно к задаче линейной регрессии.
2. Логистическая регрессия.
3. Наивный классификатор Байеса.
4. Алгоритм k-means.
5. Алгоритм Априори.

6. Использование готовых решений анализа данных (Weka, Orange, и т.д.).
7. Визуализация данных с помощью Tableau.
8. Алгоритм k-means, реализация в рамках парадигмы Map Reduce.
9. Регуляризация метода наименьших квадратов.
10. Нейронная сеть.
11. Алгоритм SVM.

4) Вопросы к экзамену.

1. Определение больших данных, ключевые характеристики. Примеры задач больших данных. Основные виды данных.
2. Роль аналитика по данным (Data Scientist). Ключевые компетенции аналитика. Отличия BI от Data Science.
3. Корреляция и регрессионный анализ. Коэффициент корреляции. Графическое представление. Постановка задачи регрессионного анализа. Линейная регрессия. Метод наименьших квадратов. Привести примеры использования регрессионного анализа.
4. Классификация. Признаковое описание объекта и таблица объект-свойства. Постановка задачи. Отличия задачи классификации от задачи регрессии. Определение модели и алгоритма. Процесс обучения. Проблема переобучения. Регуляризация. Cross validation. Привести примеры использования алгоритмов классификации.
   Дополнительный вопрос: привести модель в линейной регрессии.
5. Кластеризация. Метрики. Матрица парных расстояний. Постановка задачи кластеризации. Отличие от задачи классификации. Привести примеры использования алгоритмов кластеризации.
6. Ассоциативные правила. Определение. Достоверность и поддержка. Отличия построения ассоциативного правила от решающего правила задачи классификации. Привести примеры использования ассоциативных правил.
7. Парадигма Map Reduce. Описать принцип работы. Нарисовать диаграмму. Перечислить слабые и сильные стороны. Обозначить области применимости. Привести примеры использования.
8. Визуализация. Дать определение визуализации. Показать важность визуализации в аналитике больших данных. Привести примеры использования визуализации.
9. «Жизненный цикл» проекта по аналитике больших данных. Типовая архитектура проекта в области больших данных. Перечислить используемые технологии, указать степень вовлеченности каждой из технологий на каждом этапе работы над проектом. Перечислить основные роли исполнителей проекта.
10. Научные проблемы больших данных. Показать значимость проблем, актуальность, связь с областями математики и инженерии.

5) Критерии выставления оценки. Оценка складывается из двух показателей: 30% * «знание теоретического материала» + 70% * «выполнение контрольных заданий».

Теоретический материал:
«отлично» –
«хорошо» – в
«удовлетворительно» –
«не удовлетворительно» –

Контрольные задания:
«отлично» – средняя оценка за контрольные задания превышает 4,5 б.
«хорошо» – средняя оценка за контрольные задания превышает 3,5 б.
«удовлетворительно» – средняя оценка за контрольные задания превышает 2,5 б.
«не удовлетворительно» – средняя оценка за контрольные задания менее 2,5 балла.

Каждое контрольное задание оценивается по сформированности соответствующих компетенций.

7.4. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности, характеризующих этапы формирования компетенций

**8. Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины**

а) основная литература:

1. Trevor Hastie, Elements of statistical learning, Springer, 2009. (URL:http://statweb.stanford.edu/~tibs/ElemStatLearn/ )
2. Лбов, Геннадий Сергеевич. Анализ данных и знаний : учебное пособие / Г.С. Лбов ; Федер. агентство по образованию, Новосиб. гос. ун-т, Мех.-мат. фак .— Новосибирск : Новосибирский государственный университет, 2010 .— 107 с. (Абонемент учебной и научной литературы: НК-7840, 35 экз).
3. Мерков, Александр Борисович. Распознавание образов : введение в методы статистического обучения / А.Б. Мерков ; Рос. акад. наук, Ин-т систем. анализа .— Москва : УРСС = URSS, 2010 .— 254 с. : ил. ; 22 см. — На обороте тит. л. в макете: 2011 .— Библиогр.: с.243-249 (87 назв.) .— Предм. указ.: с.250-254 .— ISBN 978-5-354-01337-1. (Книгохранение: 1 экз, 3-8 М523). — (URL:http://www.recognition.mccme.ru/pub/RecognitionLab.html/slbook.pdf)

б) дополнительная литература:

4. Christopher M. Bishop, Pattern recognition and machine learning, Springer, 2006 http://www.springer.com/computer/image+processing/book/978-0-387-31073-2
5. Committee on the Analysis of Massive Data; Committee on Applied and Theoretical Statistics; Board on Mathematical Sciences and Their Applications; Division on Engineering and Physical Sciences; Frontiers in Massive Data Analysis, National Research Council, 2013 http://www.nap.edu/catalog/18374/frontiers-in-massive-data-analysis
6. Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data // EMC Education Services. 2015. — 432p. — ISBN: 978-1-118-87613-8.
7. C. B. B. D. Manyika, “Big Data: The Next Frontier for Innovation, Competition, and Productivity,” McKinsey Global Institute, 2011. URL: http://www.mckinsey.com/~/media/McKinsey/dotcom/Insights%20and%20pubs/MGI/Research/Technology%20and%20Innovation/Big%20Data/MGI_big_data_full_report.ashx
8. *J. Hopcroft, R. Kannan*. Foundations of Data Science. 2013. — 412 p. (https://www.dropbox.com/s/j2s5dn5w5g7ics5/Data%20Science%20Foundations%20book-dec-30-2013.pdf)
9. *Martin Hilbert.* Big Data for Development: From Information- to Knowledge Societies", – 2013. – SSRN Scholarly Paper No. ID 2205145). Rochester, NY: Social Science Research Network;
10. *Hortonworks.* 7 Key Drivers for the Big Data Market. – 2012
11. Big Data analytics: Future architectures, Skills and roadmaps for the CIO – 2011. – IDC/SAS
12. *Hasso Plattner, Alexander Zeier.* In-Memory Data Management: Technology and Applications.
13. *Gaurav Vaish.* Getting Started with NoSQL - 2013. - Packt Publishing. ISBN-13: 978-1849694988. (URL: http://it-ebooks.info/book/2748/)
14. *Jim Webber, Emil Eifrem.* Graph Databases by Ian Robinson. - 2013 - O'Reilly Media. ISBN-13: 978-1449356262 (URL: http://it-ebooks.info/book/2571/)
15. *Виктор Маер-Шенбергер, Кеннет Кукьер.* Большие данные: Революция, которая изменит то, как мы живем, работаем и мыслим. — М.: «Манн, Иванов и Фербер», 2013, 240 с. ISBN 978-5-91657-936-9 (http://www.mann-ivanov-ferber.ru/books/paperbook/big_data/)
16. *DJ Patil.* Building Data Science Teams. O’Reilly. 2011. ISBN: 978-1-449-31623-5 (http://cdn.oreilly.com/radar/2011/09/Building-Data-Science-Teams.pdf)
17. *J. Adler*. R in a Nutshell. Second Edition. O’Reilly Media Inc. 2012. ISBN: 978-1-449-31208-4 (URL: http://it-ebooks.info/book/1014/)

в) программное обеспечение и Интернет-ресурсы:

1. Apache Hadoop, HBase.
2. Apache Spark, Mahout.
3. Weka, RapidMiner, Orange

4. Excel PowerPivot
5. IRIS Dataset. URL: http://aima.cs.berkeley.edu/data/iris.csv
6. Tableau public

## 9. Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины

1. Воронцов К.В. Математические методы обучения по прецедентам http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf
2. Математические методы распознавания образов Автор: Л.М. Местецкий (Интернет университет высоких технологий) http://www.intuit.ru/department/graphics/imageproc/4/1.html
3. Онлайн курс Machine learning https://www.coursera.org/course/ml
4. Онлайн курс Big Data Overview https://education.emc.com/academicalliance/elearning/Big_Data_Overview/index.htm
5. Онлайн курс R progamming https://www.coursera.org/course/rprog
6. Онлайн курс Introduction to Data Science https://www.coursera.org/course/datasci
7. Онлайн курс «Введение в аналитику больших массивов данных» http://bit.ly/IntuitBDA.
8. Учебник по статистическому обучению http://statweb.stanford.edu/~tibs/ElemStatLearn/

## 10. Методические указания для обучающихся по освоению дисциплины

Присутствие на лекции не должно сводиться лишь к автоматической записи изложения предмета преподавателем. Каждый студент должен разработать для себя систему ускоренного фиксирования на бумаге материала лекции. Поэтому рекомендуется формализация записи посредством использования общепринятых логико-математических символов, сокращений, алгебраических (формулы) и геометрических (графики), системных (схемы, таблицы) фиксаций изучаемого материала. Овладение такой методикой, позволяет каждому студенту не только ускорить процесс изучения, но и повысить его качество, поскольку успешное владение указанными приемами требует переработки, осмысления и структуризации материала.

Вузовская подготовка студентов должна обеспечивать приобретение ими не только знаний, но и умений использовать полученные знания на практике. Это требование и положено в основу целей и методов проведения практических занятий. Практические задания предлагаются в соответствии с рабочей программой в рамках каждой темы.

## 11. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине, включая перечень программного обеспечения и информационных справочных систем (при необходимости)

1. Доступ к кластеру Hadoop. Студенты самостоятельно устанавливают Hadoop (открытая лицензия).

2. Доступ к программно-аппаратному комплексу SAP HANA (или Oracle Exalytics), например, через облака Amazon.
3. Для обучения каждому студенту необходимы:
    - персональный компьютер, с установленным следующим ПО:
        - Weka, RapidMiner
        - Excel PowerPivot

**12. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине**

1. Компьютерный класс компьютерами, объединенными в локальную сеть и доступом к Интернет.
2. Во время лекционных занятий также необходим проектор, подключенный к ПК с установленным Microsoft Office, для наглядной демонстрации изучаемого материала и проведения лекционных занятий.

Приложение А
***(обязательное)***

**Методические рекомендации по самостоятельной работе обучающихся по дисциплине «Аналитика больших массивов данных»**

1. План-график выполнения СРС по дисциплине

В процессе изучения дисциплины предусмотрено выполнение следующих видов самостоятельной работы:

| Вид самостоятельной работы | Номер недели семестра | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 |
| Чтение литературы | 2 | | 2 | | | | | | 2 | | | | | 2 | | | |
| Выполнение заданий | | 2 | 2 | 4 | 4 | 2 | | | | | 4 | 2 | | | | | |
| Прохождение онлайн-курса | | | | | | | 4 | 4 | 2 | | | | | | | | |
| Подготовка к экзамену | | | | | | | | | | | | | | | | | 36 |
| Итого в неделю часов | 2 | 2 | 4 | 4 | 4 | 2 | 4 | 4 | 4 | 0 | 4 | 2 | 0 | 2 | 0 | 0 | 36 |

2. Характеристика и описание заданий на СРС

В процессе изучения дисциплины предусмотрены следующие контрольные точки:

| Контрольная точка | Срок сдачи (номер недели семестра) |
|---|---|
| Контрольное задание №1 | 2 |
| Контрольное задание №2 | 4 |
| Контрольное задание №3 | 6 |
| Контрольное задание №4 | 8 |
| Контрольное задание №5 | 10 |
| Контрольное задание №6 | 12 |
| Контрольное задание №7 | 14 |
| Контрольное задание №8 | 16 |

3. Примерные нормы времени на выполнение заданий контрольных точек

| Контрольная точка | Норма времени на выполнение (в часах) |
|---|---|
| Контрольное задание №1 | 2 |
| Контрольное задание №2 | 4 |

| | |
|---|---|
| Контрольное задание №3 | 4 |
| Контрольное задание №4 | 4 |
| Контрольное задание №5 | 2 |
| Контрольное задание №6 | 2 |
| Контрольное задание №7 | 6 |
| Контрольное задание №8 | 2 |

4. Рекомендуемая литература (основная и дополнительная)

а) основная литература:


1. Trevor Hastie, Elements of statistical learning, Springer, 2009. (URL:http://statweb.stanford.edu/~tibs/ElemStatLearn/ )
2. Лбов, Геннадий Сергеевич. Анализ данных и знаний : учебное пособие / Г.С. Лбов ; Федер. агентство по образованию, Новосиб. гос. ун-т, Мех.-мат. фак .— Новосибирск : Новосибирский государственный университет, 2010 .— 107 с. (Абонемент учебной и научной литературы: НК-7840, 35 экз).
3. Мерков, Александр Борисович. Распознавание образов : введение в методы статистического обучения / А.Б. Мерков ; Рос. акад. наук, Ин-т систем. анализа .— Москва : УРСС = URSS, 2010 .— 254 с. : ил. ; 22 см. — На обороте тит. л. в макете: 2011 .— Библиогр.: с.243-249 (87 назв.) .— Предм. указ.: с.250-254 .— ISBN 978-5-354-01337-1. (Книгохранение: 1 экз, З-8 М523). — (URL:http://www.recognition.mccme.ru/pub/RecognitionLab.html/slbook.pdf)


б) дополнительная литература:


4. Christopher M. Bishop, Pattern recognition and machine learning, Springer, 2006 http://www.springer.com/computer/image+processing/book/978-0-387-31073-2
5. Committee on the Analysis of Massive Data; Committee on Applied and Theoretical Statistics; Board on Mathematical Sciences and Their Applications; Division on Engineering and Physical Sciences; Frontiers in Massive Data Analysis, National Research Council, 2013 http://www.nap.edu/catalog/18374/frontiers-in-massive-data-analysis
6. Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data // EMC Education Services. 2015. — 432p. — ISBN: 978-1-118-87613-8.
7. C. B. B. D. Manyika, “Big Data: The Next Frontier for Innovation, Competition, and Productivity,” McKinsey Global Institute, 2011. URL: http://www.mckinsey.com/~/media/McKinsey/dotcom/Insights%20and%20pubs/MGI/Research/Technology%20and%20Innovation/Big%20Data/MGI_big_data_full_report.ashx
8. *J. Hopcroft, R. Kannan*. Foundations of Data Science. 2013. — 412 p. (https://www.dropbox.com/s/j2s5dn5w5g7ics5/Data%20Science%20Foundations%20book-dec-30-2013.pdf)

9. *Martin Hilbert.* Big Data for Development: From Information- to Knowledge Societies", – 2013. – SSRN Scholarly Paper No. ID 2205145). Rochester, NY: Social Science Research Network;
10. *Hortonworks.* 7 Key Drivers for the Big Data Market. – 2012
11. Big Data analytics: Future architectures, Skills and roadmaps for the CIO – 2011. – IDC/SAS
12. *Hasso Plattner, Alexander Zeier.* In-Memory Data Management: Technology and Applications.
13. *Gaurav Vaish.* Getting Started with NoSQL - 2013. - Packt Publishing. ISBN-13: 978-1849694988. (URL: http://it-ebooks.info/book/2748/)
14. *Jim Webber, Emil Eifrem.* Graph Databases by Ian Robinson. - 2013 - O'Reilly Media. ISBN-13: 978-1449356262 (URL: http://it-ebooks.info/book/2571/)
15. *Виктор Маер-Шенбергер, Кеннет Кукьер.* Большие данные: Революция, которая изменит то, как мы живем, работаем и мыслим. — М.: «Манн, Иванов и Фербер», 2013, 240 с. ISBN 978-5-91657-936-9 (http://www.mann-ivanov-ferber.ru/books/paperbook/big_data/)
16. *DJ Patil.* Building Data Science Teams. O'Reilly. 2011. ISBN: 978-1-449-31623-5 (http://cdn.oreilly.com/radar/2011/09/Building-Data-Science-Teams.pdf)
17. *J. Adler*. R in a Nutshell. Second Edition. O'Reilly Media Inc. 2012. ISBN: 978-1-449-31208-4 (URL: http://it-ebooks.info/book/1014/)

## 5. Требования к представлению и оформлению результатов СРС

Студенты выполняют контрольные задания самостоятельно и результат показывают преподавателю. Преподаватель выставляет оценку в соответствии с критериями, изложенными в программе курса.

## 6. Оценка выполнения СРС

Результаты тестов оцениваются по правильности ответа на поставленный вопрос (правильно/неправильно).

По результатам выполнения заданий преподаватель оценивает уровень сформированности каждой компетенции.

Данные тестов и выполнения заданий служат первичной информацией для оценки компетенций по таблице в п.7.2. По результатам оценки всех компетенций путём простого среднего выставляется общая оценка.

Приложение Б
*(обязательное)*

**Фонд оценочных средств по дисциплине «Аналитика больших массивов данных»**

1) Текущий контроль. В течении семестра студенты выполняют 11 контрольных заданий. Выполнение контрольных заданий является обязательным для всех студентов, а результаты текущего контроля служат основанием для выставления оценок в ведомость контрольной недели на факультете. В случае невыполнения более 3х контрольных заданий студент не допускается до экзамена. Решение о допуске принимает лектор вместе с ассистентом, ведущим практические занятия в данной группе. Отсутствие более одного контрольного задания являться поводом для снижения оценки за экзамен.
Тесты (пятиминутки) (выбрать правильный ответ, или указать число):

1. О соотношении аналоговой и цифровой информации:
    1. Большинство данных в мире в 2011 году содержалось:
        i. В цифровом виде
        ii. В аналоговом виде
    2. В каком веке произошёл перевес объёмов накопленных человечеством данных в сторону цифровых?
    3. Объём накопленных человечеством цифровых данных на 2012 год измеряется:
        i. Петабайтами
        ii. Зеттабайтами
        iii. Экзабайтами
        iv. Йоттабайтами
    4. Сколько Петабайт в Зеттабайте?
2. История больших данных
    1. Укажите фактор, способствовавший появлению тренда больших данных
        i. Маркетинговые кампании крупных корпораций
        ii. Снижение издержек на хранение данных
        iii. Появление новых технологий обработки потоковых данных
        iv. Выпуск баз данных с обработкой данных в памяти
    2. Какие вероятные разочарования тренда больших данных?
        i. Из-за угрозы безопасности личной жизни (privacy) граждан будут усложнены процедуры сбора данных, что приведёт к падению ценности больших данных.
    3. Отметьте значимые события, повлиявшие на формирование тренда больших данных:
        i. Разработка Hadoop
        ii. Изобретение принципа MapReduce
        iii. Разработка языка Pyhton
        iv. Победа Deepblue в матче с Г.Каспаровым.
3. Определение больших данных:

1. Выберите верный ответ
    i. Большие данные – это обработка или хранение более 1 Тб информации.
    ii. Проблема больших данных – это такая проблема, когда при существующих технологиях хранения и обработки сущностная обработка данных затруднена или невозможна.
    iii. Большие данные – это огромная PR-акция крупных вендоров и не более того.
    iv. Большие данные – это явление, когда цифровые данные наиболее полно представляют изучаемый объект.
2. Выберите неверный ответ:
    i. Большие данные – это данные объёма свыше 1 Тб
    ii. Проблема больших данных – это проблема, когда при существующих технологиях хранения и обработки сущностная обработка данных затруднена или невозможна.
    iii. Большие данные – это тренд в области ИТ, подогреваемый маркетинговыми кампаниями крупных вендоров.
    iv. Большие данные как правило не структурированы.
3. Отметьте те из вариантов, в которых данные структурированы:
    i. Данные о продажах компании, представленные в виде помесячных отчётов в формате MS Word.
    ii. Таблица с ежедневными показаниями температуры помещения за год в файле формата csv.
    iii. Текст педагогической поэмы А.С. Макаренко, представленный в формате PDF.
    iv. Библиотека фильмов, представленных в формате mpeg4 на одном жестком диске.

4. Характеристики Big Data:
    1. Перечислите четыре основных характеристики Big Data:
        i. Virtualization, Volume, Variability, Vehicle
        ii. Variety, Velocity, Volume, Value
        iii. Verification, Volume, Velocity, Visualization
        iv. Video, Value, Variety, Volume
    2. Выберите неверное высказывание:
        i. Большие объёмы данных приводят к слабой их структуризации, поэтому появляется такое разнообразие данных.
        ii. Увеличившаяся производительность телекоммуникационных каналов привела к росту объёмов передаваемой информации.
        iii. Удешевление систем хранения на единицу информации привело к росту рынка больших данных.
        iv. Большое разнообразие источников данных
    3. Отметьте неверное понимание Variety в контексте характеристик Big Data:
        i. Высокая скорость генерирования данных.
        ii. Разные типы данных в колонках таблиц реляционных СУБД.

iii. Разнообразие отраслей, являющихся источниками данных.
iv. Разнообразие типов данных, включающих в себя структурированные, полуструктурированные и неструктурированные.

5. Принцип MapReduce
    1. Принцип MapReduce состоит в том, чтобы
        i. Производить вычисления на узлах, где информация изначально была сохранена
        ii. Использовать вычислительные мощности систем хранения
        iii. Использовать функциональное программирование для решения задач массивно-параллельной обработки
    2. Выберите одно неверное высказывание про MapReduce:
        i. Интерфейс для массово-параллельной обработки данных, где вычисления производятся на узлах, где информация изначально была сохранена
        ii. MapReduce – это две операции: распределения и сборки данных
        iii. MapReduce был придуман разработчиками Hadoop
        iv. MapReduce был анонсирован разработчиками Google
    3. Каков теоретический прирост производительности при подсчёте числа слов в тексте при работе MapReduce при переходе от одного узла к двум?
6. Технологии хранения
    1. Какие из следующих технологий СУБД не используют принцип MapReduce
        i. Hadoop
        ii. Cassandra
        iii. HDInsight
        iv. Redis
    2. Какие СУБД полностью полагаются на оперативную память при хранении информации:
        i. Oracle Exalytics
        ii. SAP HANA
        iii. BigTable
        iv. HBase
    3. В чём преимущество колоночно-ориентированных СУБД?
        i. Они позволяют выполнять более сложные SQL-запросы по сравнению с реляционными СУБД
        ii. Они позволяют динамически дополнять содержание записей новыми полями
        iii. Они имеют более гибкие возможности аналитики.
        iv. Они позволяют эффективно делать межколоночные сравнения.
7. «Песочница» в аналитическом процессе
    1. Для чего аналитику необходима «песочница»?

i. Для высокопроизводительной аналитики за счёт использования оперативной памяти и inDB операций.
ii. Для хранения всех полученных от заказчика данных.
iii. Для построения отчётов о результатах анализа
iv. Для снижения затрат, связанных с репликацией данных

2. Какие из следующих средств разумно использовать для анализа данных, представленных единственным csv-файлом размера более 100Гб:
   i. Hadoop
   ii. Data Warehouse
   iii. «Песочница»
   iv. Python
3. Выберите верное утверждение:
   i. Data Warehouse создаются для проверки гипотез при анализе больших данных.
   ii. «Песочница» используется для снижения нагрузки на основной Data Warehouse.
   iii. Каждый Data Warehouse должен содержать «песочницу».
   iv. «Песочница» необходима для любого процесса аналитики.

8. CRISP-DM
   1. Расставьте последовательность этапов проекта аналитики в соответствии с CRISP-DM.
      i. Понимание бизнеса (Business understanding)
      ii. Понимание данных (Data Understanding)
      iii. Подготовка данных (Data Preparation)
      iv. Моделирование (Modeling)
      v. Оценка (Evaluation)
      vi. Внедрение (Deployment)
   2. На каком из этапов процесса CRISP-DM происходит проверка гипотез?
      i. Понимание бизнеса (Business understanding)
      ii. Понимание данных (Data Understanding)
      iii. Моделирование (Modeling)
      iv. Оценка (Evaluation)
   3. Вы являетесь владельцем и аналитиком в компании из 10 человек, в которой требуется проанализировать продажи за 1 год (1 млн. продаж). Какие из этапов CRISP-DM можно опустить:
      i. Понимание бизнеса (Business understanding)
      ii. Подготовка данных (Data Preparation)
      iii. Моделирование (Modeling)
      iv. Оценка (Evaluation)

9. Hadoop
   1. Пример благоразумного использования Hadoop
      i. Анализ 10 Гб данных.

ii. Ежедневное сохранение данных температуры, поступающих со всех городов России (по одному показанию на город, всего городов 1100 шт).
iii. Посекундное сохранение данных температуры, поступающих со всех городов России (по одному показанию на город, всего городов 1100 шт).
iv. Построение графика пульса пациента в реальном времени.

2. Начиная с каких размеров данных обоснованно применение кластера Hadoop для хранения данных?
   i. 100Гб
   ii. 1Тб
   iii. 100Тб
   iv. 1Пб
3. Hadoop – это:
   i. Набор утилит, и программный каркас для выполнения распределённых программ, работающих на кластерах.
   ii. Распределённая СУБД, позволяющая обрабатывать большие данные.
   iii. Язык выполнения заданий в парадигме MapReduce.
   iv. Распределённая файловая система, предназначенная для хранения файлов большого объёма.

2) Промежуточная аттестация – экзамен. Студент случайным образом выбирает билет, содержащий вопрос по теории, читаемой в лекционной части курса, и вопрос, касающийся практической части курса. Если студент не сдал какие-либо контрольные задания, то он может попросить дополнительный вопрос по данным темам, чтобы избежать снижения оценки за экзамен.

3) Темы контрольных заданий.

12. Метод наименьших квадратов применительно к задаче линейной регрессии.
13. Логистическая регрессия.
14. Наивный классификатор Байеса.
15. Алгоритм k-means.
16. Алгоритм Априори.
17. Использование готовых решений анализа данных (Weka, Orange, и т.д.).
18. Визуализация данных с помощью Tableau.
19. Алгоритм k-means, реализация в рамках парадигмы Map Reduce.
20. Регуляризация метода наименьших квадратов.
21. Нейронная сеть.
22. Алгоритм SVM.

4) Вопросы к экзамену.

11. Определение больших данных, ключевые характеристики. Примеры задач больших данных. Основные виды данных.

12. Роль аналитика по данным (Data Scientist). Ключевые компетенции аналитика. Отличия BI от Data Science.
13. Корреляция и регрессионный анализ. Коэффициент корреляции. Графическое представление. Постановка задачи регрессионного анализа. Линейная регрессия. Метод наименьших квадратов. Привести примеры использования регрессионного анализа.
14. Классификация. Признаковое описание объекта и таблица объект-свойства. Постановка задачи. Отличия задачи классификации от задачи регрессии. Определение модели и алгоритма. Процесс обучения. Проблема переобучения. Регуляризация. Cross validation. Привести примеры использования алгоритмов классификации.
    Дополнительный вопрос: привести модель в линейной регрессии.
15. Кластеризация. Метрики. Матрица парных расстояний. Постановка задачи кластеризации. Отличие от задачи классификации. Привести примеры использования алгоритмов кластеризации.
16. Ассоциативные правила. Определение. Достоверность и поддержка. Отличия построения ассоциативного правила от решающего правила задачи классификации. Привести примеры использования ассоциативных правил.
17. Парадигма Map Reduce. Описать принцип работы. Нарисовать диаграмму. Перечислить слабые и сильные стороны. Обозначить области применимости. Привести примеры использования.
18. Визуализация. Дать определение визуализации. Показать важность визуализации в аналитике больших данных. Привести примеры использования визуализации.
19. «Жизненный цикл» проекта по аналитике больших данных. Типовая архитектура проекта в области больших данных. Перечислить используемые технологии, указать степень вовлеченности каждой из технологий на каждом этапе работы над проектом. Перечислить основные роли исполнителей проекта.
20. Научные проблемы больших данных. Показать значимость проблем, актуальность, связь с областями математики и инженерии.