**Федеральное государственное бюджетное образовательное учреждение высшего образования**
**«РОССИЙСКАЯ АКАДЕМИЯ НАРОДНОГО ХОЗЯЙСТВА И ГОСУДАРСТВЕННОЙ СЛУЖБЫ ПРИ ПРЕЗИДЕНТЕ РОССИЙСКОЙ ФЕДЕРАЦИИ»**

ЭКОНОМИЧЕСКИЙ ФАКУЛЬТЕТ
ОТДЕЛЕНИЕ ЭКОНОМИКИ
Кафедра Системного анализа и информатики

УТВЕРЖДЕНА
на заседании кафедры
Системного анализа и информатики
Протокол от «1» сентября 2017г. № 1

# РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Индекс Б1.В.ОД.5 «Сбор, обработка и хранение распределенных данных»

по направлению подготовки 38.04.01 «Экономика»

направленность «Системы больших данных в экономике»

квалификация магистр

очная форма обучения

Год набора - 2017

Москва, 2017г.

**Автор(ы)–составитель(и):**
к.т.н., доцент кафедры Системного анализа и информатики Стефановский Д.В.

**Заведующий кафедрой**
Системного анализа и информатики, к.т.н., доцент, Маруев С.А.

## СОДЕРЖАНИЕ

## 1. Перечень планируемых результатов обучения по дисциплине, соотнесенных с планируемыми результатами освоения программы

1.1. Дисциплина Б1.В.ОД.5 «Сбор, обработка и хранение распределенных данных» обеспечивает овладение следующими компетенциями:

| Код компетенции | Наименование компетенции | Код этапа освоения компетенции | Наименование этапа освоения компетенции |
|---|---|---|---|
| ПК-И-1 | Способность применять современные информационные технологии для работы с экономическими данными | ПК-И-1.2 | Способен к отбору и формированию состава собственных и приобретаемых данных и информации с указанием источников данных и условий их получения и доставки в соответствии с требованиями технического задания |

1.2. В результате освоения дисциплины у студентов должны быть сформированы:

| Профессиональные действия | Код этапа освоения компетенции | Результаты обучения |
|---|---|---|
| Проведение аналитического исследования в соответствии с согласованными требованиями | ПК-И-1.2 | **на уровне знаний:**<br>• Методы разработки отчетной аналитической документации |
| | | **на уровне умений:**<br>• Адаптировать и развертывать модели в предметной среде |
| Разработка и согласование технического проекта методической и технологической инфраструктуры больших данных | | **на уровне знаний:**<br>• Современный опыт использования анализа больших данных<br>• Методы управления жизненным циклом ИТ-инфраструктуры организации<br>• Методы управления проектами создания ИТ-инфраструктуры организации<br>• Современные методы и инструментальные средства анализа больших данных<br>• Рекомендации и опыт использования методов анализа больших данных<br>• Рекомендации по использованию, опыт использования и интеграции современных инструментальных средств сбора, хранения, обработки и анализа больших данных<br>• Методы интерпретации и визуализации больших данных<br>• Источники информации и условия их использования<br>• Методы извлечения информации и знаний из гетерогенных, мульти структурированных, |

| | | |
|---|---|---|
| | | неструктурированных источников, в том числе при потоковой обработке<br>• Современная технологическая инфраструктура высокопроизводительных и распределенных вычислений<br>• Методы обеспечения и оценки качества информации |
| | | **на уровне умений:**<br>• Проведение сравнительного анализа и выбора методов и методик анализа больших данных и составление рекомендаций по их использованию, оценивать условия их приобретения и использования<br>• Разработка спецификации и описания разрабатываемых методов и методик анализа больших данных, инструментальных средств или их компонент в соответствии с требованиями технического задания |

## 2. Объем и место дисциплины в структуре ОП ВО

**Объем дисциплины**

5 ЗЕ, 64 ак. часа на контактную работу с преподавателем, 80 ак. часов на самостоятельную работу обучающихся;

**Место дисциплины в структуре ОП ВО**

– Б1.В.ОД.5 «Сбор, обработка и хранение распределенных данных», 1 курс, 1 семестр;

– дисциплина реализуется после изучения дисциплин:

экономическая информатика (в объеме бакалавриата);

– форма промежуточной аттестации – экзамен.

## 3. Содержание и структура дисциплины

Таблица 2.

| № п/п | Наименование тем (разделов) | Объем дисциплины, ак. час./час | | | | | | Форма текущего контроля успеваемости*, промежуточной аттестации |
|---|---|---|---|---|---|---|---|---|
| | | Всего | Контактная работа обучающихся с преподавателем по видам учебных занятий | | | | СР | |
| | | | Л | ЛР | ПЗ | КСР | | |
| Очная форма обучения | | | | | | | | |
| Тема 1 | Распределенный анализ данных | 18 | 2 | 6 | | | 10 | Опрос |
| Тема 2 | Data Mining в реальном времени (Real-Time Data Mining) | 18 | 2 | 6 | | | 10 | Опрос |
| Тема 3 | Извлечение знаний из Web — Web Mining | 18 | 2 | 6 | | | 10 | Опрос |
| Тема 4 | Программирование операций с RDD. Работа с парами ключ/значение | 18 | 2 | 6 | | | 10 | Опрос |
| Тема 5 | Выполнение в кластере | 18 | 2 | 6 | | | 10 | Опрос |
| Тема 6 | Spark SQL | 18 | 2 | 6 | | | 10 | Опрос |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Тема 7 | Spark Streaming | 18 | 2 | 6 | | | 10 | Опрос |
| Тема 8 | Машинное обучение с MLlib | 18 | 2 | 6 | | | 10 | Опрос, ДЗ (1-8), КР (1-8) |
| Промежуточная аттестация | | | | - | | | - | Экзамен |
| Всего: | | 180/ 135 | 16/ 12 | 48/ 36 | | | 80/ 60 | 36 |

Примечание – формы текущего контроля успеваемости: контрольная работа (КР), домашнее задание (ДЗ)

## Содержание дисциплины

***Тема 1. Распределенный анализ данных.***

Системы мобильных агентов. Основные понятия. Стандарты многоагентных систем. Системы мобильных агентов. Система мобильных агентов JADE. Использование мобильных агентов для анализа данных. Проблемы распределенного анализа данных. Агенты-аналитики. Варианты анализа распределенных данных.

Система анализа распределенных данных. Общий подход к реализации системы. Агент для сбора информации о базе данных. Агент для сбора статистической информации о данных. Агент для решения одной задачи интеллектуального анализа данных. Агент для решения интегрированной задачи интеллектуального анализа данных.

***Тема 2. Data Mining в реальном времени (Real-Time Data Mining).***

Идея Data Mining в реальном времени. Адаптация системы к общей концепции. Адаптивная добыча данных. Статический Data Mining и Data Mining в реальном времени. Применение Data Mining в реальном времени.
Рекомендательные машины. Классификация рекомендательных машин. Подход на основе содержания. Совместное фильтрование. Анализ рыночной корзины и секвенциальный анализ. Усиление обучения и агенты.

Инструменты Data Mining в реальном времени

***Тема 3. Извлечение знаний из Web — Web Mining***

Web Mining. Проблемы анализа информации из Web. Проблемы анализа информации из Web. Этапы Web Mining. Web Mining и другие интернет-технологии. Категории Web Mining.

Методы извлечения Web-контента. Извлечение Web-контента в процессе информационного поиска. Извлечение Web-контента для формирования баз данных

Извлечение Web-структур. Представление Web-структур. Оценка важности Web-структур. Поиск Web-документов с учетом гиперссылок. Кластеризация Web-структур.

Исследование использования Web-ресурсов. Исследуемая информация. Этап препроцессинга. Этап извлечения шаблонов. Этап анализа шаблонов и их применение.

***Тема 4. Программирование операций с RDD. Работа с парами ключ/значение***

Основы RDD. Создание RDD. Операции с RDD. Преобразования. Действия. Отложенные вычисления. Передача функций в Spark. Применры на: Python, Scala, Java. Простые наборы RDD. Преобразование типов RDD. Сохранение (кэширование).

Создание наборов пар. Преобразования наборов пар. Агрегирование. Группировка данных. Соединения. Сортировка. Действия над наборами пар ключ/значение. Управление распределением данных. Определение объекта управления распределением RDD. Операции, получающие выгоды от наличия информации о распределении. Операции, на которые влияет порядок распределения. Пример: PageRank. Собственные объекты управления распределением.

***Тема 5. Выполнение в кластере***

Архитектура среды Spark времени выполнения. Драйвер. Исполнители. Диспетчер кластера. Запуск программы. Итоги. Развертывание приложений с помощью spark-submit. Упаковка программного кода и зависимостей. Сборка приложения на Java с помощью Maven. Сборка приложения на Scala с помощью sbt. Конфликты зависимостей. Планирование приложений в Spark. Диспетчеры кластеров. Диспетчер кластера Spark Standalone. Hadoop YARN. Apache Mesos. Amazon EC2. Выбор диспетчера кластера.

***Тема 6. Spark SQL***

Включение Spark SQL в приложения. Использование Spark SQL в приложениях. Инициализация Spark SQL. Пример простого запроса. Наборы данных SchemaRDD. Кэширование. Загрузка и сохранение данных. Apache Hive. Parquet. JSON. Из RDD. Сервер JDBC/ODBC. Работа с программой Beeline. Долгоживущие таблицы и запросы. Функции, определяемые пользователем. Spark SQL UDF. HiveUDF. Производительность Spark SQL.

***Тема 7. Spark Streaming***

Архитектура и абстракция. Преобразования без сохранения состояния. Преобразования с сохранением состояния. Операции вывода. Источники исходных данных. Основные источники. Дополнительные источники. Множество источников и размеры кластера. Круглосуточная работа. Копирование в контрольных точках. Повышение отказоустойчивости драйвера. Отказоустойчивость рабочих узлов. Отказоустойчивость приемников. Гарантированная обработка. Веб-интерфейс Spark Streaming. Проблемы производительности. Интервал пакетирования и протяженность окна. Степень параллелизма. Сборка мусора и использование памяти.

***Тема 8. Машинное обучение с MLlib***

Системные требования. Основы машинного обучения. Пример: классификация спама. Типы данных. Векторы. Алгоритмы. Извлечение признаков. Статистики. Классификация и регрессия. Кластеризация. Коллаборативная фильтрация и рекомендации. Понижение размерности. Оценка модели. Советы и вопросы производительности. Выбор признаков. Настройка алгоритмов. Кэширование наборов RDD для повторного использования. Разреженные векторы. Степень параллелизма. Высокоуровневый API машинного обучения.

## 4. Материалы текущего контроля успеваемости обучающихся и фонд оценочных средств промежуточной аттестации по дисциплине

### 4.1. Формы и методы текущего контроля успеваемости.

**4.1.1.** В ходе реализации дисциплины Б1.В.ОД.5 «Сбор, обработка и хранение распределенных данных» используются следующие методы текущего контроля успеваемости обучающихся: опросы, домашнее задание и контрольная работа.

| Тема | Методы текущего контроля успеваемости |
|---|---|
| Тема 1 | Опрос |
| Тема 2 | Опрос |
| Тема 3 | Опрос |
| Тема 4 | Опрос |
| Тема 5 | Опрос |
| Тема 6 | Опрос |
| Тема 7 | Опрос |
| Тема 8 | Опрос, домашнее задание по темам 1-8, контрольная работа по темам 1-8 |

**4.1.2. Экзамен проводится с применением следующих методов (средств)**: в письменной форме в виде контрольной работы.

**4.2. Материалы текущего контроля успеваемости обучающихся**

**Типовые оценочные материалы по теме 1**

Опрос:

- Расскажите о системах мобильных агентов и основных понятиях. Приведите примеры.
- Перечислите основные стандарты многоагентных систем. Приведите примеры использования.
- Опишите основные особенности использование мобильных агентов для анализа данных.
- Перечислите основные проблемы распределенного анализа данных.
- Опишите основные особенности понятия Агенты-аналитики.
- Приведите примеры анализа распределенных данных.
- Опишите основные особенности систем анализа распределенных данных. Приведите примеры.
- Опишите общий подход к реализации системы.
- Опишите основные особенности систем для сбора информации о базе данных. Приведите примеры.
- Опишите основные особенности систем для сбора статистической информации о данных. Приведите примеры.
- Опишите основные особенности систем для решения одной задачи интеллектуального анализа данных. Приведите пример.

**Типовые оценочные материалы по теме 2**

Опрос:

- Опишите основные особенности систем Data Mining в реальном времени.
- Приведите примеры адаптации системы к общей концепции.
- Опишите основные особенности адаптивной добыча данных. Приведите примеры.
- Опишите основные особенности статического Data Mining. Приведите примеры.
- Опишите основные особенности понятия "Рекомендательные машины". Приведите примеры использования.
- Приведите классификацию рекомендательных машин.
- Опишите основные особенности существующих инструментов Data Mining в реальном времени.

**Типовые оценочные материалы по теме 3**

Опрос:

- Web Mining. Перечислите основные проблемы анализа информации из Web.
- Опишите основные особенности и этапы Web Mining. Приведите примеры.
- Опишите основные особенности и методы извлечения Web-контента. Приведите пример.
- Опишите основные особенности и методы извлечение Web-контента для формирования баз данных. Приведите пример.
- Опишите основные особенности и методы поиска Web-документов с учетом гиперссылок.
- Опишите основные особенности и этапы Web-mining.

**Типовые оценочные материалы по теме 4**

Опрос:

- Опишите основные особенности RDD. Приведите примеры.
- Опишите основные особенности концепции "Отложенные вычисления". Опишите основные особенности передачи функций в Spark. Приведите примеры на: Python, Scala, Java.

- Опишите основные особенности: Простые наборы RDD. Преобразование типов RDD. Приведите примеры.
- Опишите основные особенности: создание наборов пар, преобразования наборов пар, агрегирования. Приведите примеры.
- Опишите основные особенности: группировка данных, соединения, сортировка. Приведите примеры.
- Дайте определение объекта управления распределением RDD. Приведите пример операций, получающих выгоды от наличия информации о распределении.
- Опишите основные особенности: операций, на которые влияет порядок распределения. Приведите пример.

**Типовые оценочные материалы по теме 5**

Опрос:

- Опишите основные особенности и дайте необходимые определения: архитектуры среды Spark.
- Опишите основные особенности развертывания приложений с помощью spark-submit. Приведите пример упаковки программного кода и зависимостей.
- Опишите основные особенности сборки приложения на Java с помощью Maven. Приведите примеры необходимых команд.
- Опишите основные особенности сборки приложения на Scala с помощью sbt. Приведите примеры необходимых команд.
- Опишите основные особенности планирования приложений в Spark. Приведите примеры необходимых команд.
- Опишите основные особенности диспетчера кластеров Spark Standalone.
- Опишите основные особенности и правила выбора диспетчера кластера. Поясните на примере.

**Типовые оценочные материалы по теме 6**

Опрос:

- Опишите основные особенности включения Spark SQL в приложения. Приведите примеры.
- Опишите основные особенности использования Spark SQL в приложениях. Опишите основные особенности использования Apache Hive. Проясните на примере.
- Опишите основные особенности использования долгоживущих таблиц и запросов. Поясните на примере.
- Опишите основные особенности использования функций, определяемых пользователем. Поясните на примере.
- Опишите основные особенности использования Spark SQL UDF. Поясните на примере.

**Типовые оценочные материалы по теме 7**

Опрос:

- Опишите архитектуру и абстракцию для потоковой обработки.
- Опишите основные особенности использования для потоковой обработки: операций вывода, источники исходных данных. Поясните на примере.
- Опишите основные особенности концепции "Гарантированная обработка".
- Опишите основные особенности использования проблемы производительности. Поясните на примере.

**Типовые оценочные материалы по теме 8**

Опрос:

- Опишите основные подходы использования библиотеки MLib. Поясните на примере. Опишите основные типы данных библиотеки MLib. Поясните на примере
- Опишите основные особенности MLib, связанные с выбором признаков и настройки алгоритмов. Поясните на примере.

- Опишите основные особенности высокоуровневого API машинного обучения. Поясните на примере.

Домашнее задание:
Встройте предложенный преподавателем метод из библиотеки scikit-learn (http://scikit-learn.org/stable/) в бот Slack-statsbot(https://github.com/backspace/slack-statsbot).

Контрольная работа:
Для предложенного преподавателем набора данных для зависимой переменной определите оптимальный набор признаков и настройте алгоритм машинного обучения для построения классификатора.

**4.3. Оценочные средства для промежуточной аттестации.**

**4.3.1. Перечень компетенций с указанием этапов их формирования в процессе освоения образовательной программы. Показатели и критерии оценивания компетенций с учетом этапа их формирования.**

| Код компетенции | Наименование компетенции | Код этапа освоения компетенции | Наименование этапа освоения компетенции |
|---|---|---|---|
| ПК-И-1 | Способность применять современные информационные технологии для работы с экономическими данными | ПК-И-1.2 | Способен к отбору и формированию состава собственных и приобретаемых данных и информации с указанием источников данных и условий их получения и доставки в соответствии с требованиями технического задания |

| Этап освоения компетенции | Показатель оценивания | Критерий оценивания |
|---|---|---|
| ПК-И-1.2 | Способен к отбору и формированию состава собственных и приобретаемых данных и информации с указанием источников данных и условий их получения и доставки в соответствии с требованиями технического задания | Понимает и может обосновать целесообразность выбранных собственных и приобретаемых данных и информации с указанием источников данных и условий их получения и доставки в соответствии с требованиями технического задания |

**4.3.2. Типовые оценочные средства.**

**Вопросы к экзамену:**

- Расскажите о системах мобильных агентов и основных понятиях. Приведите примеры.
- Перечислите основные стандарты многоагентных систем. Приведите примеры использования.
- Опишите основные особенности использование мобильных агентов для анализа данных.
- Перечислите основные проблемы распределенного анализа данных.
- Опишите основные особенности понятия Агенты-аналитики.

- Приведите примеры анализа распределенных данных.
- Опишите основные особенности систем анализа распределенных данных. Приведите примеры.
- Опишите общий подход к реализации системы.
- Опишите основные особенности систем для сбора информации о базе данных. Приведите примеры.
- Опишите основные особенности систем для сбора статистической информации о данных. Приведите примеры.
- Опишите основные особенности систем для решения одной задачи интеллектуального анализа данных. Приведите пример.
- Опишите основные особенности систем Data Mining в реальном времени.
- Приведите примеры адаптации системы к общей концепции.
- Опишите основные особенности адаптивной добыча данных. Приведите примеры.
- Опишите основные особенности статического Data Mining. Приведите примеры.
- Опишите основные особенности понятия "Рекомендательные машины". Приведите примеры использования.
- Приведите классификацию рекомендательных машин.
- Опишите основные особенности существующих инструментов Data Mining в реальном времени.
- Web Mining. Перечислите основные проблемы анализа информации из Web.
- Опишите основные особенности и этапы Web Mining. Приведите примеры.
- Опишите основные особенности и методы извлечения Web-контента. Приведите пример.
- Опишите основные особенности и методы извлечение Web-контента для формирования баз данных. Приведите пример.
- Опишите основные особенности и методы поиска Web-документов с учетом гиперссылок.
- Опишите основные особенности и этапы Web-mining.
- Опишите основные особенности RDD. Приведите примеры.
- Опишите основные особенности концепции "Отложенные вычисления". Опишите основные особенности передачи функций в Spark. Приведите примеры на: Python, Scala, Java.
- Опишите основные особенности: Простые наборы RDD. Преобразование типов RDD. Приведите примеры.
- Опишите основные особенности: создание наборов пар, преобразования наборов пар, агрегирования. Приведите примеры.
- Опишите основные особенности: группировка данных, соединения, сортировка. Приведите примеры.
- Дайте определение объекта управления распределением RDD. Приведите пример операций, получающих выгоды от наличия информации о распределении.
- Опишите основные особенности: операций, на которые влияет порядок распределения. Приведите пример.
- Опишите основные особенности и дайте необходимые определения: архитектуры среды Spark.
- Опишите основные особенности развертывания приложений с помощью spark-submit. Приведите пример упаковки программного кода и зависимостей.
- Опишите основные особенности сборки приложения на Java с помощью Maven. Приведите примеры необходимых команд.
- Опишите основные особенности сборки приложения на Scala с помощью sbt. Приведите примеры необходимых команд.

- Опишите основные особенности планирования приложений в Spark. Приведите примеры необходимых команд.
- Опишите основные особенности диспетчера кластеров Spark Standalone.
- Опишите основные особенности и правила выбора диспетчера кластера. Поясните на примере.
- Опишите основные особенности включения Spark SQL в приложения. Приведите примеры.
- Опишите основные особенности использования Spark SQL в приложениях. Опишите основные особенности использования Apache Hive. Проясните на примере.
- Опишите основные особенности использования долгоживущих таблиц и запросов. Поясните на примере.
- Опишите основные особенности использования функций, определяемых пользователем. Поясните на примере.
- Опишите основные особенности использования Spark SQL UDF. Поясните на примере.
- Опишите архитектуру и абстракцию для потоковой обработки.
- Опишите основные особенности использования для потоковой обработки: операций вывода, источники исходных данных. Поясните на примере.
- Опишите основные особенности концепции "Гарантированная обработка".
- Опишите основные особенности использования проблемы производительности. Поясните на примере.
- Опишите основные подходы использования библиотеки MLib. Поясните на примере. Опишите основные типы данных библиотеки MLib. Поясните на примере
- Опишите основные особенности MLib, связанные с выбором признаков и настройки алгоритмов. Поясните на примере.
- Опишите основные особенности высокоуровневого API машинного обучения. Поясните на примере.
- Встройте предложенный преподавателем метод из библиотеки scikit-learn (http://scikit-learn.org/stable/) в бот Slack-statsbot(https://github.com/backspace/slack-statsbot).
- Для предложенного преподавателем набора данных для зависимой переменной определите оптимальный набор признаков и настройте алгоритм машинного обучения для построения классификатора.

**Шкала оценивания.**

Оценка определяется по формуле:

¼ опрос + ¼ д.з.+ ¼ к.р.+ ¼ сдача зачета.

| 10-бальная шкала | Традиционная шкала | «Зачтено»/ «Не зачтено» | Определение |
|---|---|---|---|
| 10 | Отлично | Зачтено | Полные, глубокие и систематические знания, полный и правильный ответ на теоретический вопрос, полное и правильное решение задачи. |
| 9 | Отлично | Зачтено | Глубокие и систематические знания, правильный ответ на теоретический вопрос, правильное решение задачи. |
| 8 | Отлично | Зачтено | Систематические знания, правильный ответ на теоретический вопрос, правильное решение задачи. |

| 10-бальная шкала | Традиционная шкала | «Зачтено»/ «Не зачтено» | Определение |
|---|---|---|---|
| 7 | Хорошо | Зачтено | Систематические знания, правильный ответ на теоретический вопрос с незначительными не точностями, правильное решение задачи. |
| 6 | Хорошо | Зачтено | Систематические знания, правильный ответ на теоретический вопрос с незначительными не точностями, правильное решение задачи с незначительными неточностями. |
| 5 | Удовлетворительно | Зачтено | Ответ на теоретический вопрос неполный, правильное решение задачи с незначительными неточностями. |
| 4 | Удовлетворительно | Зачтено | Ответ на теоретический вопрос неполный, решение задачи содержит арифметические ошибки, не влияющие на правильность хода решения задачи. |
| 3 | Неудовлетворительно | Не зачтено | Ответ на теоретический вопрос неполный, решение задачи содержит идеологические ошибки. |
| 2 | Неудовлетворительно | Не зачтено | Ответ на теоретический вопрос неверный и/или решение задачи содержит идеологические ошибки. |
| 1 | Неудовлетворительно | Не зачтено | Ответ на теоретический вопрос неверный и решение задачи отсутствует. |
| 0 | Неудовлетворительно | Не зачтено | Ответ на теоретический вопрос отсутствует и решение задачи отсутствует. |

### 4.4. Методические материалы по проведению промежуточной аттестации

Экзамен проводится в аудитории. Отсчет времени, отведенного на письменную работу, идет по завершении процедуры размещения студентов и раздачи заданий.

Студент обязан являться на письменный контроль в указанное в расписании время. В случае опоздания время, отведенное на письменный контроль знаний, не продлевается.

При себе студенты могут иметь только письменные принадлежности. Необходимую для выполнения работы бумагу выдает преподаватель.

Преподаватель раздает варианты работы, содержащий 2 вопроса. Листы с заданиями должны быть повернуты текстом вниз, чтобы студенты до окончания процедуры раздачи не могли начать выполнение работы. По окончании раздачи вариантов студентам разрешается перевернуть текст задания и одновременно приступить к выполнению работы. По окончании отведенного времени студенты одновременно заканчивают выполнение работы. Если работа завершена существенно раньше срока, то по разрешению преподавателя студент может покинуть аудиторию досрочно.

Мобильные телефоны должны быть выключены и убраны со столов, допускается использование калькуляторов, выполняющих только простые арифметические вычисления.

Во время проведения письменного контроля знаний студентам не разрешается пользоваться учебными программами, справочниками и прочими источниками информации.

Использование материалов, а также попытка общения с другими студентами или иными лицами, в том числе с применением электронных средств связи,

несанкционированные перемещения и т.п. являются основанием для удаления студента из аудитории и последующего проставления в ведомость оценки «неудовлетворительно».

Во время проведения письменного контроля знаний студентам разрешается покинуть аудиторию только при условии сдачи работы в объеме, выполненном к моменту выхода из аудитории. Дальнейшее продолжение работы запрещается.

Ответы в работе без объяснений не засчитываются. Рисунки должны быть четкими, все линии графиков, используемых при ответах на вопросы задач, должны быть подписаны.

Продолжительность экзаменационной письменной работы 120 минут.

## 5. Методические указания для обучающихся по освоению дисциплины

Любой вид занятий, создающий условия для зарождения самостоятельной мысли, познавательной и творческой активности студента связан с самостоятельной работой. В широком смысле под самостоятельной работой понимают совокупность всей самостоятельной деятельности студентов как в учебной аудитории, так и вне ее, в контакте с преподавателем и в его отсутствие. Самостоятельная работа может реализовываться: непосредственно в процессе аудиторных занятий – на лекциях, практических и семинарских занятиях, при выполнении контрольных и лабораторных работ и др.; в контакте с преподавателем вне рамок аудиторных занятий – на консультациях по учебным вопросам, в ходе творческих контактов, при ликвидации задолженностей, при выполнении индивидуальных заданий и т.д.; в библиотеке, дома, в общежитии, на кафедре и других местах при выполнении студентом учебных и творческих заданий.

Лекции

Главное в период подготовки к лекционным занятиям – научиться методам самостоятельного умственного труда, сознательно развивать свои творческие способности и овладевать навыками творческой работы. Для этого необходимо строго соблюдать дисциплину учебы и поведения. Четкое планирование своего рабочего времени и отдыха является необходимым условием для успешной самостоятельной работы. В основу его нужно положить рабочие программы изучаемых в семестре дисциплин. Каждому студенту следует составлять еженедельный и семестровый планы работы, а также план на каждый рабочий день. С вечера всегда надо распределять работу на завтрашний день. В конце каждого дня целесообразно подводить итог работы: тщательно проверить, все ли выполнено по намеченному плану, не было ли каких-либо отступлений, а если были, по какой причине это произошло. Нужно осуществлять самоконтроль, который является необходимым условием успешной учебы. Если что-то осталось невыполненным, необходимо изыскать время для завершения этой части работы, не уменьшая объема недельного плана.

Самостоятельная работа на лекции Слушание и запись лекций – сложный вид вузовской аудиторной работы. Внимательное слушание и конспектирование лекций предполагает интенсивную умственную деятельность студента. Краткие записи лекций, их конспектирование помогает усвоить учебный материал. Конспект является полезным тогда, когда записано самое существенное, основное и сделано это самим студентом. Не надо стремиться записать дословно всю лекцию. Такое
«конспектирование» приносит больше вреда, чем пользы. Запись лекций рекомендуется вести по возможности собственными формулировками. Желательно запись осуществлять на одной странице, а следующую оставлять для проработки учебного материала самостоятельно в домашних условиях. Конспект лекци лучше подразделять на пункты, параграфы, соблюдая красную строку. Этому в большой степени будут способствовать пункты плана лекции, предложенные преподавателям. Принципиальные места,

определения, формулы и другое следует сопровождать замечаниями «важно», «особо важно», «хорошо запомнить» и т.п. Можно делать это и с помощью разноцветных маркеров или ручек. Лучше если они будут собственными, чтобы не приходилось присить их у однокурсников и тем самым не отвлекать их во время лекции. Целесообразно разработать собственную «маркографию» (значки, символы), сокращения слов. Не лишним будет и изучение основ стенографии. Работая над конспектом лекций, всегда необходимо использовать не только учебник, но и ту литературу, которую дополнительно рекомендовал лектор. Именно такая серьезная, кропотливая работа с лекционным материалом позволит глубоко овладеть знаниями.

Семинар и проведение опроса

Каждый студент должен начать с ознакомления с планом занятия, который отражает содержание предложенной темы. Тщательное продумывание и изучение вопросов плана основывается на проработке текущего материала лекции, а затем изучения обязательной и дополнительной литературы, рекомендованную к данной теме. На основе индивидуальных предпочтений студенту необходимо самостоятельно выбрать тему доклада по проблеме семинара и по возможности подготовить по нему презентацию. Если программой дисциплины предусмотрено выполнение практического задания, то его необходимо выполнить с учетом предложенной инструкции (устно или письменно). Все новые понятия по изучаемой теме необходимо выучить наизусть и внести в глоссарий, который целесообразно вести с самого начала изучения курса. Результат такой работы должен проявиться в способности студента свободно ответить на теоретические вопросы семинара, его выступлении и участии в коллективном обсуждении вопросов изучаемой темы, правильном выполнении практических заданий и контрольных работ.

Работа с литературными источниками.

В процессе подготовки к семинарским занятиям, студентам необходимо обратить особое внимание на самостоятельное изучение рекомендованной учебно-методической (а также научной и популярной) литературы. Самостоятельная работа с учебниками, учебными пособиями, научной, справочной и популярной литературой, материалами периодических изданий и Интернета, статистическими данными является наиболее эффективным методом получения знаний, позволяет значительно активизировать процесс овладения информацией, способствует более глубокому усвоению изучаемого материала, формирует у студентов свое отношение к конкретной проблеме. Более глубокому раскрытию вопросов способствует знакомство с дополнительной литературой, рекомендованной преподавателем по каждой теме семинарского или практического занятия, что позволяет студентам проявить свою индивидуальность в рамках выступления на данных занятиях, выявить широкий спектр мнений по изучаемой проблеме.

Методические указания по выполнению самостоятельной работы:

***Тема 1. Распределенный анализ данных.***

Студент изучает предложенные материалы темы, решает задачи по теме и готовится к опросу по следующей тематике:

Системы мобильных агентов. Основные понятия. Стандарты многоагентных систем. Системы мобильных агентов. Система мобильных агентов JADE. Использование мобильных агентов для анализа данных. Проблемы распределенного анализа данных. Агенты-аналитики. Варианты анализа распределенных данных.

Система анализа распределенных данных. Общий подход к реализации системы. Агент для сбора информации о базе данных. Агент для сбора статистической информации о данных. Агент для решения одной задачи интеллектуального анализа данных. Агент для решения интегрированной задачи интеллектуального анализа данных.

***Тема 2. Data Mining в реальном времени (Real-Time Data Mining).***

Студент изучает предложенные материалы темы, решает задачи по теме и готовится к опросу по следующей тематике:

Идея Data Mining в реальном времени. Адаптация системы к общей концепции. Адаптивная добыча данных. Статический Data Mining и Data Mining в реальном времени. Применение Data Mining в реальном времени.
Рекомендательные машины. Классификация рекомендательных машин. Подход на основе содержания. Совместное фильтрование. Анализ рыночной корзины и секвенциальный анализ. Усиление обучения и агенты.

Инструменты Data Mining в реальном времени

***Тема 3. Извлечение знаний из Web — Web Mining***

Студент изучает предложенные материалы темы, решает задачи по теме и готовится к опросу по следующей тематике:

Web Mining. Проблемы анализа информации из Web. Проблемы анализа информации из Web. Этапы Web Mining. Web Mining и другие интернет-технологии. Категории Web Mining.

Методы извлечения Web-контента. Извлечение Web-контента в процессе информационного поиска. Извлечение Web-контента для формирования баз данных

Извлечение Web-структур. Представление Web-структур. Оценка важности Web-структур. Поиск Web-документов с учетом гиперссылок. Кластеризация Web-структур.

Исследование использования Web-ресурсов. Исследуемая информация. Этап препроцессинга. Этап извлечения шаблонов. Этап анализа шаблонов и их применение.

***Тема 4. Программирование операций с RDD. Работа с парами ключ/значение***

Студент изучает предложенные материалы темы, решает задачи по теме и готовится к опросу по следующей тематике:

Основы RDD. Создание RDD. Операции с RDD. Преобразования. Действия. Отложенные вычисления. Передача функций в Spark. Применры на: Python, Scala, Java. Простые наборы RDD. Преобразование типов RDD. Сохранение (кэширование).

Создание наборов пар. Преобразования наборов пар. Агрегирование. Группировка данных. Соединения. Сортировка. Действия над наборами пар ключ/значение. Управление распределением данных. Определение объекта управления распределением RDD. Операции, получающие выгоды от наличия информации о распределении. Операции, на которые влияет порядок распределения. Пример: PageRank. Собственные объекты управления распределением.

***Тема 5. Выполнение в кластере***

Студент изучает предложенные материалы темы, решает задачи по теме и готовится к опросу по следующей тематике:

Архитектура среды Spark времени выполнения. Драйвер. Исполнители. Диспетчер кластера. Запуск программы. Итоги. Развертывание приложений с помощью spark-submit. Упаковка программного кода и зависимостей. Сборка приложения на Java с помощью Maven. Сборка приложения на Scala с помощью sbt. Конфликты зависимостей. Планирование приложений и в приложениях Spark. Диспетчеры кластеров. Диспетчер кластера Spark Standalone. Hadoop YARN. Apache Mesos. Amazon EC2. Выбор диспетчера кластера.

***Тема 6. Spark SQL***

Студент изучает предложенные материалы темы, решает задачи по теме и готовится к опросу по следующей тематике:

Включение Spark SQL в приложения. Использование Spark SQL в приложениях. Инициализация Spark SQL. Пример простого запроса. Наборы данных SchemaRDD. Кэширование. Загрузка и сохранение данных. Apache Hive. Parquet. JSON. Из RDD. Сервер JDBC/ODBC. Работа с программой Beeline. Долгоживущие таблицы и запросы. Функции, определяемые пользователем. Spark SQL UDF. HiveUDF. Производительность Spark SQL.

***Тема 7. Spark Streaming***

Студент изучает предложенные материалы темы, решает задачи по теме и готовится к опросу по следующей тематике:

Архитектура и абстракция. Преобразования без сохранения состояния. Преобразования с сохранением состояния. Операции вывода. Источники исходных данных. Основные источники. Дополнительные источники. Множество источников и размеры кластера. Круглосуточная работа. Копирование в контрольных точках. Повышение отказоустойчивости драйвера. Отказоустойчивость рабочих узлов. Отказоустойчивость приемников. Гарантированная обработка. Веб-интерфейс Spark Streaming. Проблемы производительности. Интервал пакетирования и протяженность окна. Степень параллелизма. Сборка мусора и использование памяти.

***Тема 8. Машинное обучение с MLlib***

Студент изучает предложенные материалы темы, решает задачи по теме и готовится к опросу по следующей тематике:

Системные требования. Основы машинного обучения. Пример: классификация спама. Типы данных. Векторы. Алгоритмы. Извлечение признаков. Статистики. Классификация и регрессия. Кластеризация. Коллаборативная фильтрация и рекомендации. Понижение размерности. Оценка модели. Советы и вопросы производительности. Выбор признаков. Настройка алгоритмов. Кэширование наборов RDD для повторного использования. Разреженные векторы. Степень параллелизма. Высокоуровневый API машинного обучения.

## 6. Учебная литература и ресурсы информационно-телекоммуникационной сети "Интернет", включая перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине

### 6.1. Основная литература.

1. Карау, Холден. Изучаем Spark: молниеносный анализ данных / Холден Карау, Энди Конвински, Патрик Венделл, Матей Захария. - Москва: ДМК Пресс, 2015. - 304 с.: ил.; ISBN 978-5-97060-323-9

### 6.2. Дополнительная литература.

1. Натан Марц, Джеймс Уоррен. Большие данные. Принципы и практика построения масштабируемых систем обработки данных в реальном времени.

### 6.3. Учебно-методическое обеспечение самостоятельной работы.

Отдельное обеспечение не предусмотрено.

### 6.4. Нормативные правовые документы.

Не предусмотрены.

### 6.5. Интернет-ресурсы.

1. http://citeseer.ist.psu.edu/ основной источник знаний по Computer Science, по многим статьям есть полные тексты
2. http://citeseer.cs.msu.su/— русскоязычная электронная библиотека научных статей
3. http://arxiv.org/ — библиотека электронных публикаций, в основном по физике, но доля «Computer Science» в последнее время стремительно увеличивается

4. http://rexa.info/ библиографическая поисковая система по статьям, авторам и грантам
5. http://elibrary.ru/ - российская научная электронная библиотека
6. http://liinwww.ira.uka.de/bibliography/index.html
7. библиографическая база данных для работы с BibTeX
8. http://www.gotai.net/ -- русскоязычный сайт об искусственном интеллекте
9. Math-Net.ru -- общероссийский математический портал

### 6.6. Иные источники.

Не предусмотрены.

## 7. Материально-техническая база, информационные технологии, программное обеспечение и информационные справочные системы

Для лекций:

1. Персональный компьютер
2. Мультимедийный проектор
3. Доска, мел или маркеры

Для лабораторных занятий:

1. Компьютерный класс,
2. Виртуальная машина Ubuntu 15.04 b выше с установленным Postgresql и MongoDB
3. Мультимедийный проектор
4. Доска, маркеры
5. Компилятор R-2.15.1 – GNU - http://www.r-project.org/ либо интегрированная среда разработки RStudio – GNU AGP - http://www.rstudio.com/ide/.
6. Jupyter Notebook - бесплатная интерактивная оболочка для языка программирования Python, позволяющая объединить код, текст и диаграммы.
7. Компилятор Scala – http://www.scala-lang.org/
8. Программный комплекс анализа новостного сайта -”Crawler-Persona”
9. База данных “Централизация государственных закупок в 2014 г”.
10. База данных учебно-методических материалов по дисциплине "Макроэкономика".
11. База данных Бюджетная и социально-экономическая статистика субъектов Российской Федерации.