**Санкт-Петербургский государственный университет**

**Р А Б О Ч А Я П Р О Г Р А М М А**

**УЧЕБНОЙ ДИСЦИПЛИНЫ**

Поиск и анализ информации

Information Retrieval and Analysis

**Язык(и) обучения**

русский

Трудоемкость в зачетных единицах: 3

Регистрационный номер рабочей программы: 025194

Санкт-Петербург

2020

**Раздел 1. Характеристики учебных занятий**

**1.1. Цели и задачи учебных занятий**

Дисциплина «Поиск и анализ информации» является одной из важных дисциплин образовательной программы «Технологии программирования», предназначенной для подготовки специалиста в области информационных технологий. Она представляет собой комплекс знаний, умений и навыков, позволяющих овладеть основами поиска, обработки и анализа данных. Отдельные параметры курса могут варьироваться по степени сложности в зависимости от уровня подготовки обучающихся.

Цель изучения дисциплины: ознакомление обучающихся с фундаментальными принципами анализа данных.

Задачи курса: изучение основных инструментариев и приемов поиска и анализа скрытых закономерностей в данных.

**1.2. Требования подготовленности обучающегося к освоению содержания учебных занятий (пререквизиты)**

Программа курса предназначена для обучающихся 3 курса бакалавриата.

Максимальная эффективность программы будет обеспечена при условии, что обучающийся:

* владеет основами информатики;
* умеет программировать.

**1.3. Перечень результатов обучения (learning outcomes)**

* Знание основных алгоритмов, используемых при анализе разнообразных данных;
* умение пользоваться специализированными пакетами анализа данных;
* практические навыки применения алгоритмов анализа данных в разнообразных сферах человеческой деятельности.

Компетенции, формируемые дисциплиной:

* ПКП-4 – способен применять современные информационные технологии при проектировании, реализации, оценке качества и анализа эффективности программного обеспечения для решения задач в различных предметных областях;
* ПКП-5 – способен использовать основные методы и средства автоматизации проектирования, реализации, испытаний и оценки качества при создании конкурентоспособного программного продукта и программных комплексов, а также способен использовать методы и средства автоматизации, связанные с сопровождением, администрированием и модернизацией программных продуктов и программных комплексов;
* ПКП-6 – способен использовать знания направлений развития компьютеров с традиционной (нетрадиционной) архитектурой; современных системных программных средств: операционных систем, операционных и сетевых оболочек, сервисных программ; тенденции развития функций и архитектур проблемно-ориентированных программных систем и комплексов в профессиональной деятельности.

**1.4. Перечень и объём активных и интерактивных форм учебных занятий**

Занятия в активной форме — лекции, предполагающие активную дискуссию с преподавателем (4 ак. часа).

**Раздел 2. Организация, структура и содержание учебных занятий**

**2.1. Организация учебных занятий**

**2.1.1 Основной курс**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Трудоёмкость, объёмы учебной работы и наполняемость групп обучающихся | | | | | | | | | | | | | | | | | | |
| Код модуля в составе дисциплины,  практики и т.п. | Контактная работа обучающихся с преподавателем | | | | | | | | | | | | Самостоятельная работа | | | | Объём активных и интерактивных  форм учебных занятий | Трудоёмкость |
| лекции | семинары | консультации | практические  занятия | лабораторныеработы | контрольныеработы | коллоквиумы | текущий контроль | промежуточная  аттестация | итоговая аттестация | под руководством преподавателя | в присутствии  преподавателя | сам. раб. с использованием  методических материалов | текущий контроль (сам.раб.) | промежуточная аттестация (сам.раб.) | итоговая аттестация  (сам.раб.) |
| ОСНОВНАЯ ТРАЕКТОРИЯ | | | | | | | | | | | | | | | | | | |
| Форма обучения: очная | | | | | | | | | | | | | | | | | | |
| Семестр 5 | 48 |  | 2 |  |  |  |  |  | 2 |  |  |  | 28 |  | 28 |  | 4 | 3 |
|  | 2-100 |  | 2-100 |  |  |  |  |  | 2-100 |  |  |  | 1-1 |  | 1-1 |  |  |  |
| ИТОГО | 48 |  | 2 |  |  |  |  |  | 2 |  |  |  | 28 |  | 28 |  |  | 3 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Виды, формы и сроки текущего контроля успеваемости и промежуточной аттестации | | | | | | |
| Код модуля в составе дисциплины, практики и т.п. | Формы текущего контроля успеваемости | | Виды промежуточной аттестации | | Виды итоговой аттестации  (только для программ итоговой аттестации и дополнительных образовательных программ) | |
| Формы | Сроки | Виды | Сроки | Виды | Сроки |
| ОСНОВНАЯ ТРАЕКТОРИЯ | | | | | | |
| Форма обучения: очная | | | | | | |
| Семестр 5 |  |  | экзамен, устно, традиционная форма | по графику промежуточной аттестации |  |  |

**2.2. Структура и содержание учебных занятий**

**Модуль 1.** Общие сведения о методологии анализа данных.

1. Вводные понятия. Основные задачи DataMining и методы их решения.
2. Основные этапы анализа данных. Выдвижение гипотез. Сбор и систематизация данных. Построение моделей, объясняющих имеющиеся факты. Тестирование моделей и интерпретация результатов. Применение полученных моделей.
3. Данные. Объекты исследования. Переменные (признаки объектов). Генеральная совокупность и выборка. Параметры и статистики. Измерения признаков. Дискретные и непрерывные данные. Шкалы. Номинальная шкала. Порядковая шкала. Интервальная шкала. Относительная шкала. Дихотомическая шкала.
4. Виды данных. Фактоиды. Ряды. Сырые данные (транзакции).
5. Основные источники данных. Социологические опросы. Наблюдения. Документы. Результаты прямых измерений. Социальные сети. Внешние источники.
6. Основные этапы подготовки данных. Загрузка данных. Разделение данных. Приведение к одинаковым единицам измерения. Заполнение отсутствующих значений.
7. Очистка данных. Контроль диапазонов (визуальный и с применением методов математической статистики). Проверка орфографии. Сравнение с регулярными выражениями. Сочетание полей. Устранение дубликатов. Объединение данных из разных источников

**Модуль 2.** Анализ данных.

1. Описательные статистики. Мода. Медиана. Среднее. Средневзвешенное.
2. Визуализация данных. Графики. Диаграммы. Лица Чернова. Иные способы визуализации.
3. Нормировка переменных. Нормировка униполярных и биполярных показателей. Особенности нормировки для бальных шкал.
4. Анализ временных рядов. Выделение трендов, сезонных компонент и шумов. Прогнозирование поведения временных рядов. Аддитивная и мультипликативная модели.
5. Экспоненциальные методы анализа временных рядов.
6. Метрики для оценки качества прогнозирования (MAE, RMSE, MPE, MAPE, AD, MAD, коэффициент несоответствия Тейла, коэффициент детерминации).
7. Задачи классификации. Деревья решений. Метод ближайших соседей. Алгоритм KMeans.
8. Задачи кластеризации. Алгоритм DBScan, методы иерархической кластеризации (восходящие и нисходящие), FTCA- алгоритм (Fast Threshold Clustering Algorithm).
9. Поиск ассоциативных правил. Алгоритмы APRIORI, RELIM.

**Раздел 3. Обеспечение учебных занятий**

**3.1. Методическое обеспечение**

**3.1.1 Методические указания по освоению дисциплины**

Методические материалы включают в себя следующие типы материалов — учебники, учебные пособия, методические указания для обучающихся, Интернет-ресурсы, электронные учебные пособия.

**3.1.2 Методическое обеспечение самостоятельной работы**

В течение семестра обучающимся предлагаются для выполнения практические задания по работе с базой данных. При самостоятельном изучении теоретического материала, выполнении практических целесообразно использовать рекомендованную основную и дополнительную литературу. По согласованию с преподавателем обучающийся может осваивать теоретическую часть курса по одному из онлайн-курсов, представленных на образовательных платформах в п. 3.4.3. При этом преподаватель обязан сообщить обучающимся, какие разделы выбранного им онлайн-курса недостаточно полно раскрывают ту или иную тему курса и порекомендовать дополнительные источники по данной теме.

**3.1.3 Методика проведения текущего контроля успеваемости и промежуточной аттестации и критерии оценивания**

Итоговая оценка по курсу ставится на основании 10 реализованных алгоритмов и сданного экзамена. Каждый реализованный алгоритм оценивается по шкале в диапазоне от 0 до 8 баллов. Высшая отметка возможна, если задание сдается в течение 2 недель с момента выдачи. Отметка за задание снижается при более поздней сдаче задания (1 балл за каждые 2 недели опоздания). Экзамен оценивается по 20 – бальной шкале. Допускается выставление высшей отметки по курсу (5 – оценка СПбГУ, A – оценка ECTS) без экзамена (при условии сдачи всех заданий в отведенный срок с максимальным баллом, т.е. если обучающийся набрал 80 баллов). Остальные отметки выставляются по следующему принципу:

|  |  |  |
| --- | --- | --- |
| Итоговый процент  выполнения, % | Оценка СПбГУ при  проведении экзамена | Оценка  ECTS |
| 90-100 | отлично | A |
| 86-89 | хорошо | B |
| 80-85 | хорошо | C |
| 76-79 | удовлетворительно | D |
| 70-75 | удовлетворительно | E |
| менее 70 | неудовлетворительно | F |

Экзамен проводится в устной форме и состоит из двух вопросов. Время подготовки ответа на вопросы билета составляет 45 минут. Использование конспектов и учебников, а также электронных устройств хранения, обработки или передачи информации при подготовке и ответе на вопросы экзамена категорически запрещено. В случае обнаружения факта использования недозволенных материалов (устройств) составляется акт и обучающийся удаляется с экзамена. После ответа на вопросы билета преподаватель задает несколько дополнительных вопросов, на основании оценки ответов, на которые итоговая оценка по предмету может быть повышена или понижена.

*Критерии выставления баллов за экзамен:*

1. 20 баллов ставятся за полностью раскрытый теоретический материал и правильные ответы на дополнительные вопросы преподавателя.
2. 15-19 баллов с учетом малозначительных погрешностей изложения (-1 балл за каждую погрешность, но не более 5).
3. 10-14 баллов — за изложенный теоретический материал билета с помощью наводящих подсказок преподавателя (-1 балл за каждую подсказку, но не более 5).
4. 5-9 баллов — за знание основных вопросов по выборочным темам (-1 балл за незнание отдельных тем, но не более 5).
5. Меньше 4 баллов — при невыполнении приведенных выше условий.

**3.1.4 Методические материалы для проведения текущего контроля успеваемости и промежуточной аттестации (контрольно-измерительные материалы, оценочные средства)**

*Перечень вопросов к экзамену:*

1. Основные задачи анализа данных.
2. Основные этапы анализа данных.
3. Генеральная совокупность и выборка. Параметры и статистики. Измерения признаков. Дискретные и непрерывные данные. Шкалы.
4. Виды данных. Фактоиды. Временные ряды. Сырые данные (транзакции).
5. Основные источники данных.
6. Основные этапы подготовки данных.
7. Очистка данных.
8. Описательные статистики.
9. Визуализация данных. Графики. Диаграммы. Лица Чернова. Иные способы визуализации.
10. Нормировка переменных. Особенности нормировки для различных шкал.
11. Анализ временных рядов (линейные методы).
12. Экспоненциальные методы анализа временных рядов.
13. Метрики для оценки качества прогнозирования (MAE, RMSE, MPE, MAPE, AD, MAD, коэффициент несоответствия Тейла, коэффициент детерминации).
14. Задачи классификации. Деревья решений.
15. Метод ближайших соседей.
16. Алгоритм KMeans.
17. Задачи кластеризации. Алгоритм DBScan.
18. Методы иерархической кластеризации (восходящие и нисходящие).
19. FTCA-алгоритм (Fast Threshold Clustering Algorithm).
20. Поиск ассоциативных правил. Алгоритм APRIORI.
21. Алгоритм RELIM.

**3.1.5 Методические материалы для оценки обучающимися содержания и качества учебного процесса**

Для оценки обучающимися содержания и качества учебного процесса применяется анкетирование в соответствии с методикой и графиком, утвержденными в установленном порядке.

**3.2. Кадровое обеспечение**

**3.2.1 Образование и (или) квалификация штатных преподавателей и иных лиц, допущенных к проведению учебных занятий**

К преподаванию дисциплины могут быть допущены преподаватели, имеющие диплом о высшем образовании по соответствующему направлению.

**3.2.2 Обеспечение учебно-вспомогательным и (или) иным персоналом**

Специальных требований нет.

**3.3. Материально-техническое обеспечение**

**3.3.1 Характеристики аудиторий (помещений, мест) для проведения занятий**

В аудиториях, где проводятся занятия, необходимо наличие досок и средств письма на них. Аудитории должны соответствовать требованиям СанПиН 2.2.2/2.4.1340-03 в редакции от 21 июня 2016 года.

**3.3.2 Характеристики аудиторного оборудования, в том числе неспециализированного компьютерного оборудования и программного обеспечения общего пользования**

В аудиториях, где проводятся лекционные и практические занятия, необходимо наличие досок и средств письма на них. Для показа слайдов необходим компьютер с установленным программным обеспечением для работы со слайдами в форматах PDF, PPT, PPTXи подключенный к нему мультимедиа проектор с экраном.

Аудитории для проведения практических занятий дополнительно должны быть оснащены компьютером для каждого студента с возможностью загрузки на них студентами файлов с внешних USB-носителей.

**3.3.3 Характеристики специализированного оборудования**

Специальных требований нет.

**3.3.4 Характеристики специализированного программного обеспечения**

Для проведения практических занятий желательно установить в компьютерных классах интерпретатор языка Python.

**3.3.5 Перечень и объёмы требуемых расходных материалов**

Для аудиторий с маркерными досками необходимы стирающиеся маркеры в объеме, достаточном для проведения курса. Для аудиторий с меловыми досками необходим мел в объеме, достаточном для проведения курса.

**3.4. Информационное обеспечение**

**3.4.1 Список обязательной литературы**

1. Иванов О.В. Статистика, Часть 1. Описательная статистика. Теоретико-вероятностные основания статистического вывода. М. 2005.
2. Reena Shaw, Top 10 Machine Learning Algorithms for Beginners, KDnuggets, 2017.

**3.4.2 Список дополнительной литературы**

1. Ian H. Witten, Frank Eibe, Mark A. Hall, Data Mining. Practical Data Learning Tools and Techniques (3rd edition), Morgan Kaufmann publications, 2011.
   * 1. **Перечень иных информационных источников**
2. <https://compscicenter.ru/courses/data-mining-python/2018-spring/classes/> — курс CSC «Анализ данных на Python в примерах и задачах».
3. Ресурсы сети Интернет.

**Раздел 4. Разработчики программы**

Графеева Наталья Генриховна, к.ф.-м. н., доцент кафедры информационно-аналитических систем. N.Grafeeva@spbu.ru