МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное учреждение
высшего профессионального образования
"Казанский (Приволжский) федеральный университет"
Институт вычислительной математики и информационных технологий

УТВЕРЖДАЮ

Проректор по образовательной деятельности КФУ

Проф. Таюрский Д.А.

_____ "__" _____ 20___ г.

подписано электронно-цифровой подписью

Программа дисциплины

Анализ интернет-данных Б1.В.ОД.4

Направление подготовки: 38.04.05 - Бизнес-информатика
Профиль подготовки: Математические методы и информационные технологии в бизнесе
Квалификация выпускника: магистр
Форма обучения: очное
Язык обучения: русский

Автор(ы):

Пинягина О.В.

Рецензент(ы):

Лернер Э.Ю.

СОГЛАСОВАНО:

Заведующий(ая) кафедрой: Миссаров М. Д.
Протокол заседания кафедры No ___ от "____" _____ 201__г

Учебно-методическая комиссия Института вычислительной математики и информационных технологий:
Протокол заседания УМК No ____ от "____" _____ 201__г

Регистрационный No 914016

Казань
2016

Содержание

Программу дисциплины разработал(а)(и) доцент, к.н. (доцент) Пинягина О.В. кафедра анализа данных и исследования операций отделение фундаментальной информатики и информационных технологий , Olga.Piniaguina@kpfu.ru

1. Цели освоения дисциплины

Курс охватывает следующие разделы:
- Data mining и Web mining
- Основные Интернет-технологии для Web mining
- Технология R в Web Mining
- Text mining
- Социальные сети. Social mining
- Рекомендательные системы
- Интеллектуальные агенты
- Принципы функционирования поисковых систем

2. Место дисциплины в структуре основной образовательной программы высшего профессионального образования

Данная учебная дисциплина включена в раздел " Б1.В.ОД.4 Дисциплины (модули)" основной образовательной программы 38.04.05 Бизнес-информатика и относится к обязательные дисциплины. Осваивается на 1 курсе, 1 семестр.

Курс "Анализ Интернет-данных" изучается в 1 семестре 1-го года обучения в магистратуре по направлению "Бизнес-информатика".

Для освоения данного курса студенты должны изучить курсы по программе бакалавриата "Бизнес-информатика":
- "Анализ данных"
- "Интернет-технологии"
- "Базы данных",

и иметь навыки работы в среде R.

3. Компетенции обучающегося, формируемые в результате освоения дисциплины /модуля

В результате освоения дисциплины формируются следующие компетенции:

Шифр компетенции	Расшифровка приобретаемой компетенции
ПК-1 (профессиональные компетенции)	готовить аналитические материалы для оценки мероприятий и выработки стратегических решений в области ИКТ (ПК-1);
ПК-11 (профессиональные компетенции)	проводить поиск и анализ инноваций в экономике, управлении и ИКТ (ПК-11);
ПК-16 (профессиональные компетенции)	управлять инновационной и предпринимательской деятельностью в сфере ИКТ (ПК-16);
ПК-2 (профессиональные компетенции)	проводить анализ инновационной деятельности предприятия (ПК-2);
ПК-6 (профессиональные компетенции)	управлять исследовательскими и проектно-внедренческими коллективами (ПК-6);

В результате освоения дисциплины студент:

- знать основные методы Data mining, пригодные для работы с Интернет-данными,
- уметь работать в среде R с пакетами Data mining,
- применять на практике знания, полученные при изучении курса, для извлечения, трансформации и интеллектуального анализа Интернет-данных.

4. Структура и содержание дисциплины/ модуля

Общая трудоемкость дисциплины составляет 4 зачетных(ые) единиц(ы) 144 часа(ов).

Форма промежуточного контроля дисциплины зачет в 1 семестре.

Суммарно по дисциплине можно получить 100 баллов, из них текущая работа оценивается в 50 баллов, итоговая форма контроля - в 50 баллов. Минимальное количество для допуска к зачету 28 баллов.

86 баллов и более - "отлично" (отл.);

71-85 баллов - "хорошо" (хор.);

55-70 баллов - "удовлетворительно" (удов.);

54 балла и менее - "неудовлетворительно" (неуд.).

4.1 Структура и содержание аудиторной работы по дисциплине/ модулю

Тематический план дисциплины/модуля

N	Раздел Дисциплины/ Модуля	Семестр	Неделя семестра	Виды и часы аудиторной работы, их трудоемкость (в часах)			Текущие формы контроля
				Лекции	Практические занятия	Лабораторные работы	
1.	Тема 1. Data mining и Web mining	1	1-2	2	0	2	домашнее задание
2.	Тема 2. Основные Интернет-технологии для Web mining	1	3-5	3	0	3	домашнее задание
3.	Тема 3. Технология R в Web Mining	1	6-8	3	0	3	домашнее задание
4.	Тема 4. Text mining	1	9-10	2	0	2	домашнее задание
5.	Тема 5. Социальные сети. Social mining	1	11-12	2	0	2	домашнее задание
6.	Тема 6. Рекомендательные системы	1	13-14	2	0	2	домашнее задание
7.	Тема 7. Интеллектуальные агенты	1	15-16	2	0	2	домашнее задание

N	Раздел Дисциплины/ Модуля	Семестр	Неделя семестра	Виды и часы аудиторной работы, их трудоемкость (в часах)			Текущие формы контроля
				Лекции	Практические занятия	Лабораторные работы	
8.	Тема 8. Принципы функционирования поисковых систем	1	17-18	2	0	2	домашнее задание
.	Тема . Итоговая форма контроля	1		0	0	0	зачет
	Итого			18	0	18	

4.2 Содержание дисциплины

Тема 1. Data mining и Web mining

лекционное занятие (2 часа(ов)):

Интеллектуальный анализ данных и интеллектуальный анализ Интернет-данных. Web structure mining, Web usage mining, Web content mining.

лабораторная работа (2 часа(ов)):

Установка среды R и RStudio. Форматы данных в R. Чтение и запись данных в файлы. Формат протокола Web-сервера. Работа с базами данных (SQL server, mySQL, Access).

Тема 2. Основные Интернет-технологии для Web mining

лекционное занятие (3 часа(ов)):

Основные Интернет-технологии для Web mining. Клиент-серверная архитектура. Протокол HTTP. Структура запроса клиента. Структура ответа сервера. Формат XML. Технология Xpath. Формат JSON.

лабораторная работа (3 часа(ов)):

Работа с регулярными выражениями в R. Изучение пакета stringr.

Тема 3. Технология R в Web Mining

лекционное занятие (3 часа(ов)):

Технология R в Web Mining. Пакет XML. Пакет Jsonlite. Пакет RCurl. Примеры приложений.

лабораторная работа (3 часа(ов)):

Задача анализа потребительских корзин. Изучение пакета arules.

Тема 4. Text mining

лекционное занятие (2 часа(ов)):

Text mining. Задачи категоризации и аннотирования документов.

лабораторная работа (2 часа(ов)):

Изучение online систем для Web-скрепинга.

Тема 5. Социальные сети. Social mining

лекционное занятие (2 часа(ов)):

Задачи интеллектуального анализа данных на основе информации из социальных сетей. Social mining

лабораторная работа (2 часа(ов)):

Линейная регрессия для задачи прогнозирования.

Тема 6. Рекомендательные системы

лекционное занятие (2 часа(ов)):

Рекомендательные системы. Системы на основе сходства товаров. Системы на основе сходства поведения потребителей.

лабораторная работа (2 часа(ов)):

Применение пакета RCurl для загрузки информации из Интернет.

Тема 7. Интеллектуальные агенты

лекционное занятие (2 часа(ов)):

Интеллектуальные агенты. Агенты с простым поведением. Агенты с поведением, основанным на модели. Целенаправленные агенты. Практичные агенты. Обучающиеся агенты.

лабораторная работа (2 часа(ов)):

Применение пакета http для загрузки информации из Интернет.

Тема 8. Принципы функционирования поисковых систем

лекционное занятие (2 часа(ов)):

Принципы функционирования поисковых систем. Поисковые роботы. Ранжирование документов.

лабораторная работа (2 часа(ов)):

Нейронные сети для задачи прогнозирования.

4.3 Структура и содержание самостоятельной работы дисциплины (модуля)

N	Раздел Дисциплины	Семестр	Неделя семестра	Виды самостоятельной работы студентов	Трудоемкость (в часах)	Формы контроля самостоятельной работы
1.	Тема 1. Data mining и Web mining	1	1-2	подготовка домашнего задания	12	домашнее задание
2.	Тема 2. Основные Интернет-технологии для Web mining	1	3-5	подготовка домашнего задания	18	домашнее задание
3.	Тема 3. Технология R в Web Mining	1	6-8	подготовка домашнего задания	18	домашнее задание
4.	Тема 4. Text mining	1	9-10	подготовка домашнего задания	12	домашнее задание
5.	Тема 5. Социальные сети. Social mining	1	11-12	подготовка домашнего задания	12	домашнее задание
6.	Тема 6. Рекомендательные системы	1	13-14	подготовка домашнего задания	12	домашнее задание
7.	Тема 7. Интеллектуальные агенты	1	15-16	подготовка домашнего задания	12	домашнее задание
8.	Тема 8. Принципы функционирования поисковых систем	1	17-18	подготовка домашнего задания	12	домашнее задание
	Итого				108	

5. Образовательные технологии, включая интерактивные формы обучения

В соответствии с требованиями ФГОС удельный вес занятий, проводимых в активных и интерактивных формах, составляет не менее 40% аудиторных занятий. Так, в процессе изучения дисциплины "Анализ Интернет-данных" студенты выполняют лабораторные работы по извлечению и интеллектуальному анализу Интернет-данных. До 50% лекционных и практических занятий проходят с использованием презентаций MS PowerPoint.

6. Оценочные средства для текущего контроля успеваемости, промежуточной аттестации по итогам освоения дисциплины и учебно-методическое обеспечение самостоятельной работы студентов

Тема 1. Data mining и Web mining

домашнее задание , примерные вопросы:

Задание 1. Анализ Web-логов средствами R. Загрузить индивидуальный фай протокола Apache со страницы курса. В среде R прочитать данные из файла протокола Apache в data.frame. Изучить на практике возможности языка регулярных выражений. С помощью регулярных выражений (пакет stringr) выделить нужные данные и преобразовать их к нужному виду.

Тема 2. Основные Интернет-технологии для Web mining

домашнее задание , примерные вопросы:

Продолжение задания 1. Анализ Web-логов средствами R. Загрузить данные в СУБД, например, в SQL server (пакет RODBC). Ответить на вопросы: -Сколько посетителей было на сайте за месяц? -Сколько в среднем посетителей бывает за час? -Сколько посетителей сделало заказы? -Сколько страниц просмотрел посетитель в среднем, максимум, минимум? -Сколько времени прошло с момента входа на сайт до оформления заказа, в среднем, максимум, минимум? -Сколько в среднем заказов оформляется за день?

Тема 3. Технология R в Web Mining

домашнее задание , примерные вопросы:

Продолжение задания 1. Анализ Web-логов средствами R. Ответить на вопрос: Имеются ли какие-то зависимые товары, которые покупатель кладет в корзину? (Задача анализа покупательских корзин - поиск ассоциативных правил). Для ответа на этот вопрос изучить пакет Arules, прочитать данные из базы, преобразовать в формат транзакций, затем применить метод поиска ассоциативных правил (подсказка: в каждом наборе данных "спрятано" 5 ассоциативных правил) Описать результаты работы в виде Word-документа.

Тема 4. Text mining

домашнее задание , примерные вопросы:

Задание2. Загрузка данных из Интернет (Web-scraping) средствами специализированных программ и анализ данных в R. На сайте avito.ru придумать запрос и задать критерии поиска, чтобы в выборку попало более 50 объектов. Определить список входных параметров для будущей модели прогноза (выходной параметр - цена). Выбрать любую бесплатную систему для web-скрепинга из списка Software for Web Scraping, изучить ее функционал и загрузить данные с сайта avito.ru в файл.

Тема 5. Социальные сети. Social mining

домашнее задание , примерные вопросы:

Продолжение задания 2. Загрузка данных из Интернет (Web-scraping) средствами специализированных программ и анализ данных в R. Если данные были загружены в файл типа json или xml, подключить и изучить необходимые библиотеки R для работы с данным форматом. Прочитать данные в R и преобразовать их к типу data.frame.

Тема 6. Рекомендательные системы

домашнее задание , примерные вопросы:

Продолжение задания 2. Загрузка данных из Интернет (Web-scraping) средствами специализированных программ и анализ данных в R. Изучить и построить модель линейной регрессии для прогнозирования цен на товары. Проанализировать полученные результаты и сделать выводы. Описать результаты работы в виде Word-документа.

Тема 7. Интеллектуальные агенты

домашнее задание , примерные вопросы:

Задание3. Загрузка данных из Интернет (Web-scraping) средствами R и анализ данных в R. На сайте avito.ru или подобном ему торговом сайте сформулировать критерий отбора данных, чтобы получить не менее 100 строк. Можно взять запрос из 2 задания или придумать новый. Дополнить данные информацией из подробного описания товара. Написать сценарий на языке R для загрузки и преобразования данных к формату data.frame.

Тема 8. Принципы функционирования поисковых систем

домашнее задание , примерные вопросы:

Продолжение задания 3. Загрузка данных из Интернет (Web-scraping) средствами R и анализ данных в R. Построить модель прогноза. Проанализировать строки с максимальными отклонениями от прогноза. Убрать или, если возможно, откорректировать их. Сравнить возможности специализированных программ и языка R для загрузки Интернет-данных. Описать результаты работы в виде Word-документа.

Тема . Итоговая форма контроля

Примерные вопросы к зачету:

Темы к зачету:

- Data mining и Web mining
- Основные Интернет-технологии для Web mining
- Технология R в Web Mining
- Text mining
- Социальные сети. Social mining
- Рекомендательные системы
- Интеллектуальные агенты
- Принципы функционирования поисковых систем

7.1. Основная литература:

1.Аверченков, В. И. Мониторинг и системный анализ информации в сети Интернет [электронный ресурс] : монография / В. И. Аверченков, С. М. Рощин. - 2-е изд., стереотип. - М. : ФЛИНТА, 2011. - 160 с. - ISBN 978-5-9765-1270-2

http://znanium.com/bookread2.php?book=453853

2.Кашина О.А.,Миссаров М.Д. Электронный образовательный ресурс "Анализ данных в среде R", 2013

http://zilant.kpfu.ru/course/view.php?id=17341

3.Ярушкина Н. Г. Интеллектуальный анализ временных рядов: Учебное пособие / Н.Г. Ярушкина, Т.В. Афанасьева, И.Г. Перфильева. - М.: ИД ФОРУМ: ИНФРА-М, 2012. - 160 с.:

http://znanium.com/bookread.php?book=249314

7.2. Дополнительная литература:

1. Степанов, Роман Григорьевич. Технология Data Mining: Интеллектуальный анализ данных: учебное пособие / Р. Г. Степанов; Казан. гос. ун-т. - Казань: Казанский государственный университет, 2009.- 110 с.

2. Барсегян, А. А. Анализ данных и процессов: учеб. пособие / А. А. Барсегян, М. С. Куприянов, И. И. Холод, М. Д. Тесс, С. И. Елизаров. - 3-е изд., перераб. и доп. - СПб.: БХВ-Петербург, 2009. - 512 с.: ил. + CD-ROM ? (Учебная литература для вузов).- Режим доступа:

http://www.znanium.com/bookread.php?book=350638

3.Компьютерные технологии анализа данных в эконометрике / Д.М. Дайитбегов. - 2-е изд., испр. и доп. - М.: Вузовский учебник: ИНФРА-М, 2010. - 578 с.: 70x100 1/16. - (Научная книга). (переплет) ISBN 978-5-9558-0191-9

http://www.znanium.com/bookread.php?book=251791

4.Аверченков, В. И. Система формирования знаний в среде Интернет [электронный ресурс] : монография / В. И. Аверченков, А. В. Заболеева-Зотова, Ю. М. Казаков, Е. А. Леонов, С. М. Рощин. - 2-е изд., стереотип. - М. : ФЛИНТА, 2011. - 181 с. - ISBN 978-5-9765-1266-5

http://znanium.com/bookread2.php?book=453908

7.3. Интернет-ресурсы:

The Comprehensive R Archive Network - https://cran.gis-lab.info/index.html

The R Project for Statistical Computing - https://www.r-project.org/

Наглядная статистика. Используем R! (электронный ресурс) - http://ashipunov.info/shipunov/school/books/rbook.pdf

Пользовательский интерфейс для R - https://www.rstudio.com/products/RStudio/

Страница курса на сайте КЭК - http://kek.ksu.ru/EOS/DM/index.html

8. Материально-техническое обеспечение дисциплины(модуля)

Освоение дисциплины "Анализ интернет-данных" предполагает использование следующего материально-технического обеспечения:

Мультимедийная аудитория, вместимостью более 60 человек. Мультимедийная аудитория состоит из интегрированных инженерных систем с единой системой управления, оснащенная современными средствами воспроизведения и визуализации любой видео и аудио информации, получения и передачи электронных документов. Типовая комплектация мультимедийной аудитории состоит из: мультимедийного проектора, автоматизированного проекционного экрана, акустической системы, а также интерактивной трибуны преподавателя, включающей тач-скрин монитор с диагональю не менее 22 дюймов, персональный компьютер (с техническими характеристиками не ниже Intel Core i3-2100, DDR3 4096Mb, 500Gb), конференц-микрофон, беспроводной микрофон, блок управления оборудованием, интерфейсы подключения: USB,audio, HDMI. Интерактивная трибуна преподавателя является ключевым элементом управления, объединяющим все устройства в единую систему, и служит полноценным рабочим местом преподавателя. Преподаватель имеет возможность легко управлять всей системой, не отходя от трибуны, что позволяет проводить лекции, практические занятия, презентации, вебинары, конференции и другие виды аудиторной нагрузки обучающихся в удобной и доступной для них форме с применением современных интерактивных средств обучения, в том числе с использованием в процессе обучения всех корпоративных ресурсов. Мультимедийная аудитория также оснащена широкополосным доступом в сеть интернет. Компьютерное оборудованием имеет соответствующее лицензионное программное обеспечение.

Компьютерный класс, представляющий собой рабочее место преподавателя и не менее 15 рабочих мест студентов, включающих компьютерный стол, стул, персональный компьютер, лицензионное программное обеспечение. Каждый компьютер имеет широкополосный доступ в сеть Интернет. Все компьютеры подключены к корпоративной компьютерной сети КФУ и находятся в едином домене.

Учебно-методическая литература для данной дисциплины имеется в наличии в электронно-библиотечной системе "ZNANIUM.COM", доступ к которой предоставлен студентам. ЭБС "ZNANIUM.COM" содержит произведения крупнейших российских учёных, руководителей государственных органов, преподавателей ведущих вузов страны, высококвалифицированных специалистов в различных сферах бизнеса. Фонд библиотеки сформирован с учетом всех изменений образовательных стандартов и включает учебники, учебные пособия, УМК, монографии, авторефераты, диссертации, энциклопедии, словари и справочники, законодательно-нормативные документы, специальные периодические издания и издания, выпускаемые издательствами вузов. В настоящее время ЭБС ZNANIUM.COM соответствует всем требованиям федеральных государственных образовательных стандартов высшего профессионального образования (ФГОС ВПО) нового поколения.

Компьютерный класс, представляющий собой рабочее место преподавателя и не менее 15 рабочих мест студентов, включающих компьютерный стол, стул, персональный компьютер, лицензионное программное обеспечение. Каждый компьютер имеет широкополосный доступ в сеть Интернет. Все компьютеры подключены к корпоративной компьютерной сети КФУ и находятся в едином домене.

Лабораторные занятия проводятся в компьютерном классе.

Программа составлена в соответствии с требованиями ФГОС ВПО и учебным планом по направлению 38.04.05 "Бизнес-информатика" и магистерской программе Математические методы и информационные технологии в бизнесе .

Автор(ы):
Пинягина О.В. _____
"__" _____ 201 __ г.

Рецензент(ы):
Лернер Э.Ю. _____
"__" _____ 201 __ г.