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

УТВЕРЖДАЮ
Проректор по образовательной деятельности КФУ
Проф. Д.А. Таюрский
_____ «___»_____20__г.

подписано электронно-цифровой подписью

Программа дисциплины

Анализ интернет-данных

Направление подготовки: 01.04.02 - Прикладная математика и информатика
Профиль подготовки: Анализ данных и его приложения
Квалификация выпускника: магистр
Форма обучения: очное
Язык обучения: русский
Год начала обучения по образовательной программе: 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. Перечень планируемых результатов обучения по дисциплине (модулю), соотнесенных с планируемыми результатами освоения ОПОП ВО

Обучающийся, освоивший дисциплину (модуль), должен обладать следующими компетенциями:

Шифр компетенции	Расшифровка приобретаемой компетенции
ОК-1	способность к абстрактному мышлению, анализу, синтезу
ОК-2	готовность действовать в нестандартных ситуациях, нести социальную и этическую ответственность за принятые решения
ОК-3	готовность к саморазвитию, самореализации, использованию творческого потенциала
ОПК-4	способностью использовать и применять углубленные знания в области прикладной математики и информатики
ПК-2	способностью разрабатывать концептуальные и теоретические модели решаемых научных проблем и задач
ПК-4	способностью разрабатывать концептуальные и теоретические модели решаемых задач проектной и производственно-технологической деятельности
ПК-5	способностью управлять проектами, планировать научно-исследовательскую деятельность, анализировать риски, управлять командой проекта
ПК-9	способностью к преподаванию математических дисциплин и информатики в образовательных организациях основного общего, среднего общего, среднего профессионального и высшего образования

Обучающийся, освоивший дисциплину (модуль):

Должен демонстрировать способность и готовность:

- 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.

2. Место дисциплины (модуля) в структуре ОПОП ВО

Данная дисциплина (модуль) включена в раздел "Б1.В.ДВ.4 Дисциплины (модули)" основной профессиональной образовательной программы 01.04.02 "Прикладная математика и информатика (Анализ данных и его приложения)" и относится к дисциплинам по выбору.
Осваивается на 2 курсе в 3 семестре.

3. Объем дисциплины (модуля) в зачетных единицах с указанием количества часов, выделенных на контактную работу обучающихся с преподавателем (по видам учебных занятий) и на самостоятельную работу обучающихся

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

Контактная работа - 42 часа(ов), в том числе лекции - 0 часа(ов), практические занятия - 0 часа(ов), лабораторные работы - 42 часа(ов), контроль самостоятельной работы - 0 часа(ов).

Самостоятельная работа - 102 часа(ов).

Контроль (зачёт / экзамен) - 0 часа(ов).

Форма промежуточного контроля дисциплины: зачет в 3 семестре.

4. Содержание дисциплины (модуля), структурированное по темам (разделам) с указанием отведенного на них количества академических часов и видов учебных занятий

4.1 Структура и тематический план контактной и самостоятельной работы по дисциплине (модулю)

N	Разделы дисциплины / модуля	Семестр	Виды и часы контактной работы, их трудоемкость (в часах)			Самостоятельная работа
			Лекции	Практические занятия	Лабораторные работы	
1.	Тема 1. Data mining и Web mining	3	0	0	2	4
2.	Тема 2. Основные Интернет-технологии для Web mining	3	0	0	8	24
3.	Тема 3. Технология R в Web Mining	3	0	0	16	24
4.	Тема 4. Text mining	3	0	0	8	26
5.	Тема 5. Социальные сети. Social mining	3	0	0	2	6
6.	Тема 6. Рекомендательные системы	3	0	0	2	6
7.	Тема 7. Интеллектуальные агенты	3	0	0	2	6
8.	Тема 8. Принципы функционирования поисковых систем	3	0	0	2	6
	Итого		0	0	42	102

4.2 Содержание дисциплины (модуля)

Тема 1. Data mining и Web mining

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

Тема 2. Основные Интернет-технологии для Web mining

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

Тема 3. Технология R в Web Mining

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

Тема 4. Text mining

Text mining. Задачи категоризации и аннотирования документов.

Тема 5. Социальные сети. Social mining

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

Тема 6. Рекомендательные системы

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

Тема 7. Интеллектуальные агенты

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

Тема 8. Принципы функционирования поисковых систем

Принципы функционирования поисковых систем. Поисковые роботы. Ранжирование документов.

5. Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине (модулю)

Самостоятельная работа обучающихся выполняется по заданию и при методическом руководстве преподавателя, но без его непосредственного участия. Самостоятельная работа подразделяется на самостоятельную работу на аудиторных занятиях и на внеаудиторную самостоятельную работу. Самостоятельная работа обучающихся включает как полностью самостоятельное освоение отдельных тем (разделов) дисциплины, так и проработку тем (разделов), осваиваемых во время аудиторной работы. Во время самостоятельной работы обучающиеся читают и конспектируют учебную, научную и справочную литературу, выполняют задания, направленные на закрепление знаний и отработку умений и навыков, готовятся к текущему и промежуточному контролю по дисциплине.

Организация самостоятельной работы обучающихся регламентируется нормативными документами, учебно-методической литературой и электронными образовательными ресурсами, включая:

Порядок организации и осуществления образовательной деятельности по образовательным программам высшего образования - программам бакалавриата, программам специалитета, программам магистратуры (утвержден приказом Министерства образования и науки Российской Федерации от 5 апреля 2017 года №301)

Письмо Министерства образования Российской Федерации №14-55-996ин/15 от 27 ноября 2002 г. "Об активизации самостоятельной работы студентов высших учебных заведений"

Устав федерального государственного автономного образовательного учреждения "Казанский (Приволжский) федеральный университет"

Правила внутреннего распорядка федерального государственного автономного образовательного учреждения высшего профессионального образования "Казанский (Приволжский) федеральный университет"

Локальные нормативные акты Казанского (Приволжского) федерального университета

6. Фонд оценочных средств по дисциплине (модулю)

Фонд оценочных средств по дисциплине (модулю) включает оценочные материалы, направленные на проверку освоения компетенций, в том числе знаний, умений и навыков. Фонд оценочных средств включает оценочные средства текущего контроля и оценочные средства промежуточной аттестации.

В фонде оценочных средств содержится следующая информация:

- соответствие компетенций планируемым результатам обучения по дисциплине (модулю);

- критерии оценивания сформированности компетенций;

- механизм формирования оценки по дисциплине (модулю);

- описание порядка применения и процедуры оценивания для каждого оценочного средства;

- критерии оценивания для каждого оценочного средства;

- содержание оценочных средств, включая требования, предъявляемые к действиям обучающихся, демонстрируемым результатам, задания различных типов.

Фонд оценочных средств по дисциплине находится в Приложении 1 к программе дисциплины (модулю).

7. Перечень литературы, необходимой для освоения дисциплины (модуля)

Освоение дисциплины (модуля) предполагает изучение основной и дополнительной учебной литературы. Литература может быть доступна обучающимся в одном из двух вариантов (либо в обоих из них):

- в электронном виде - через электронные библиотечные системы на основании заключенных КФУ договоров с правообладателями;

- в печатном виде - в Научной библиотеке им. Н.И. Лобачевского. Обучающиеся получают учебную литературу на абонементе по читательским билетам в соответствии с правилами пользования Научной библиотекой.

Электронные издания доступны дистанционно из любой точки при введении обучающимся своего логина и пароля от личного кабинета в системе "Электронный университет". При использовании печатных изданий библиотечный фонд должен быть укомплектован ими из расчета не менее 0,5 экземпляра (для обучающихся по ФГОС 3++ - не менее 0,25 экземпляра) каждого из изданий основной литературы и не менее 0,25 экземпляра дополнительной литературы на каждого обучающегося из числа лиц, одновременно осваивающих данную дисциплину.

Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины (модуля), находится в Приложении 2 к рабочей программе дисциплины. Он подлежит обновлению при изменении условий договоров КФУ с правообладателями электронных изданий и при изменении комплектования фондов Научной библиотеки КФУ.

8. Перечень ресурсов информационно-телекоммуникационной сети "Интернет", необходимых для освоения дисциплины (модуля)

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

9. Методические указания для обучающихся по освоению дисциплины (модуля)

Рекомендации. Для выполнения 1 задания можно воспользоваться следующим алгоритмом.

В среде R прочитать данные из файла в data.frame

С помощью регулярных выражений (пакет stringr, [1, Глава 8]) выделить нужные данные и преобразовать их к нужному виду

Загрузить данные в СУБД, например, в SQL server (пакет RODBC)

Все запросы, кроме последнего, легко написать на SQL

Для ответа на последний 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)

Рекомендации. Для выполнения 2 задания можно воспользоваться следующим алгоритмом.

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.

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.

Изучить и построить модель линейной регресии для прогнозирования цен на товары.

Проанализировать полученные результаты и сделать выводы.

Рекомендации. Примерный алгоритм работы для выполнения 3 задания.

Найти источники данных и сформулировать цель задачи.

Загрузить данные.

Преобразовать данные (очистить, удалить стоп-слова, провести стэмминг).

Применить к данным методы анализа.

Проанализировать полученные результаты и сделать выводы.

При необходимости вернуться и повторить предыдущие этапы работы.

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

Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), включая перечень программного обеспечения и информационных справочных систем, представлен в Приложении 3 к рабочей программе дисциплины (модуля).

11. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине (модулю)

Материально-техническое обеспечение образовательного процесса по дисциплине (модулю) включает в себя следующие компоненты:

Помещения для самостоятельной работы обучающихся, укомплектованные специализированной мебелью (столы и стулья) и оснащенные компьютерной техникой с возможностью подключения к сети "Интернет" и обеспечением доступа в электронную информационно-образовательную среду КФУ.

Учебные аудитории для контактной работы с преподавателем, укомплектованные специализированной мебелью (столы и стулья).

Компьютер и принтер для распечатки раздаточных материалов.

Мультимедийная аудитория.

Компьютерный класс.

12. Средства адаптации преподавания дисциплины к потребностям обучающихся инвалидов и лиц с ограниченными возможностями здоровья

При необходимости в образовательном процессе применяются следующие методы и технологии, облегчающие восприятие информации обучающимися инвалидами и лицами с ограниченными возможностями здоровья:

- создание текстовой версии любого нетекстового контента для его возможного преобразования в альтернативные формы, удобные для различных пользователей;

- создание контента, который можно представить в различных видах без потери данных или структуры, предусмотреть возможность масштабирования текста и изображений без потери качества, предусмотреть доступность управления контентом с клавиатуры;

- создание возможностей для обучающихся воспринимать одну и ту же информацию из разных источников - например, так, чтобы лица с нарушениями слуха получали информацию визуально, с нарушениями зрения - аудиально;

- применение программных средств, обеспечивающих возможность освоения навыков и умений, формируемых дисциплиной, за счёт альтернативных способов, в том числе виртуальных лабораторий и симуляционных технологий;

- применение дистанционных образовательных технологий для передачи информации, организации различных форм интерактивной контактной работы обучающегося с преподавателем, в том числе вебинаров, которые могут быть использованы для проведения виртуальных лекций с возможностью взаимодействия всех участников дистанционного обучения, проведения семинаров, выступления с докладами и защиты выполненных работ, проведения тренингов, организации коллективной работы;

- применение дистанционных образовательных технологий для организации форм текущего и промежуточного контроля;

- увеличение продолжительности сдачи обучающимся инвалидом или лицом с ограниченными возможностями здоровья форм промежуточной аттестации по отношению к установленной продолжительности их сдачи:

- продолжительности сдачи зачёта или экзамена, проводимого в письменной форме, - не более чем на 90 минут;

- продолжительности подготовки обучающегося к ответу на зачёте или экзамене, проводимом в устной форме, - не более чем на 20 минут;

- продолжительности выступления обучающегося при защите курсовой работы - не более чем на 15 минут.

Программа составлена в соответствии с требованиями ФГОС ВО и учебным планом по направлению 01.04.02 "Прикладная математика и информатика" и магистерской программе "Анализ данных и его приложения".

Приложение 2
к рабочей программе дисциплины (модуля)
Б1.В.ДВ.4 Анализ интернет-данных

Перечень литературы, необходимой для освоения дисциплины (модуля)

Направление подготовки: 01.04.02 - Прикладная математика и информатика
Профиль подготовки: Анализ данных и его приложения
Квалификация выпускника: магистр
Форма обучения: очное
Язык обучения: русский
Год начала обучения по образовательной программе: 2016

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

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

Приложение 3
к рабочей программе дисциплины (модуля)
Б1.В.ДВ.4 Анализ интернет-данных

Перечень информационных технологий, используемых для освоения дисциплины (модуля), включая перечень программного обеспечения и информационных справочных систем

Направление подготовки: 01.04.02 - Прикладная математика и информатика
Профиль подготовки: Анализ данных и его приложения
Квалификация выпускника: магистр
Форма обучения: очное
Язык обучения: русский
Год начала обучения по образовательной программе: 2016

Освоение дисциплины (модуля) предполагает использование следующего программного обеспечения и информационно-справочных систем:
Операционная система Microsoft Windows 7 Профессиональная или Windows XP (Volume License)
Пакет офисного программного обеспечения Microsoft Office 365 или Microsoft Office Professional plus 2010
Браузер Mozilla Firefox
Браузер Google Chrome
Adobe Reader XI или Adobe Acrobat Reader DC
Kaspersky Endpoint Security для Windows
Учебно-методическая литература для данной дисциплины имеется в наличии в электронно-библиотечной системе "ZNANIUM.COM", доступ к которой предоставлен обучающимся. 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, учебно-методические комплексы, 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 образования (ФГОС ВО) нового поколения.