Санкт-Петербургский государственный университет

Р А Б О Ч А Я П Р О Г Р А М М А

учебной дисциплины

Автоматизированные системы сбора и обработки данных

Automated Systems of Data Collection and Processing

Язык(и) обучения

русский

Трудоёмкость (границы трудоёмкости) в зачетных единицах: 3

Регистрационный номер рабочей программы: _001129

Санкт-Петербург

2018

Раздел 1. Характеристики учебных занятий

1.1. Цели и задачи учебных занятий

Цель дисциплины – получение базовых знаний об основных проблемах и базовых методах автоматизированного сбора и обработки информации в больших коллекциях веб-документов, а также приобретение практических навыков разработки и использования специализированных инструментов, применяемых в областях информационного поиска и анализа данных для сбора и обработки информации.

Задачами дисциплины являются изучение применяемых в информационном поиске и анализе данных основных технологий сбора данных на примере в сети Веб; изучение классических и специализированных методов сбора информации на примере в сети Веб; изучение методов обработки и хранения информации с использованием индексных структур; получение практических навыков разработки и программирования инструментария для автоматизированного сбора данных в больших коллекциях веб-документов для решения прикладных задач.

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

Освоение дисциплин: дискретная математика, основы программирования.

Знание одного из высокоуровневых языков программирования (C++, Java, C#, Python).

1.3. Перечень результатов обучения (learning outcomes)

Дисциплина способствует формированию следующих компетенций:

УК-2 Способен управлять проектом на всех этапах его жизненного цикла

ОПК-1 Способен находить, формулировать и решать актуальные проблемы прикладной математики, фундаментальной информатики и информационных технологий

ОПК-5 Способен инсталлировать и сопровождать программное обеспечение информационных систем, осуществлять эффективное управление разработкой программных средств и проектов

ПКП-2 Способен проводить анализ качества, эффективности применения и соблюдение информационной безопасности при разработке программных продуктов и программных комплексов;

ПКП-6 Способен применять основные концептуальные положения функционального, логического, объектно-ориентированного и визуального направлений программирования, методы, способы и средства разработки программ в рамках этих направлений;

ПКП-8 Способен принимать участие в управлении проектами создания информационных систем на стадиях жизненного цикла.

В результате изучения дисциплины студент должен

знать классические и специализированные методы автоматизированного сбора, обработки и хранения информации для больших коллекций документов, классические индексные структуры хранения тексто-ориентированных данных.

уметь разрабатывать автоматизированные системы сбора и обработки информации, используя изложенные в курсе методы и техники сбора и обработки; уметь реализовывать методы сбора, обработки информации и используемые индексные структуры хранения собранной информации средствами языков программирования высокого уровня; уметь экспериментально (с помощью компьютера) исследовать эффективность методов и программных инструментов.

владеть навыками автоматизированного сбора информации в больших коллекциях веб-документов и ее обработки с использованием индексных структур; навыками командной работы над реализацией проекта; навыками работы с существующими программными средствами для решения задач автоматизированного сбора и обработки информации; практическими навыками самостоятельного проектирования, кодирования, отладки, тестирования и документирования программ с применением инструментальных средств современных интегрированных сред разработки.

1.4. Перечень активных и интерактивных форм учебных занятий

Все практические занятия необходимо проводить с привлечением интерактивных методов: работа в малых группах, групповое обсуждение материалов лекций, метод проектов, представление самостоятельно выполненных индивидуальных заданий и коллективное обсуждение полученных результатов – 16 часов.

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																		
Период обучения (модуль)	Контактная работа обучающихся с преподавателем												Самостоятельная работа				Объём активных и интерактивных форм учебных занятий	Трудоёмкость
	лекции	семинары	консультации	практические занятия	лабораторные работы	контрольные работы	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																		
очная форма обучения																		
Семестр 3	12	12	2	34				6	2				33		7		34	3
ИТОГО	**12**	**12**	**2**	**34**				**6**	**2**				**33**		**7**		**34**	**3**

Формы текущего контроля успеваемости, виды промежуточной и итоговой аттестации			
Период обучения (модуль)	Формы текущего контроля успеваемости	Виды промежуточной аттестации	Виды итоговой аттестации (только для программ итоговой аттестации и дополнительных образовательных программ)
ОСНОВНАЯ ТРАЕКТОРИЯ			
очная форма обучения			
Семестр 3		экзамен	

2.2. Структура и содержание учебных занятий

Период обучения (модуль): **Семестр 3**

№ п/п	Наименование темы (раздела, части)	Вид учебных занятий	Количество часов
1	Введение	лекции	1
2	Веб-пространство и его особенности	лекции	3
		практические занятия	11
		по методическим материалам	3
3	Поисковый робот как автоматизированная система сбора и обработки данных в сети Веб.	лекции	4
		практические занятия	11
		по методическим материалам	3
4	Задача индексирования. Индексные структуры	лекции	4
		практические занятия	12
		по методическим материалам	4

Содержание учебных занятий

Введение.

Веб-пространство и его особенности. Общая информация об Интернете, гипертексте. Веб – как коллекция неструктурированных документов большой размерности. Исследования Netcraft. Веб-пространство, его особенности и свойства. Модель Веб-пространства Брёдера «Bow-tie». Открытый/ закрытый Веб. Модель Бартона-Кеблера. Информационно-поисковые системы в Вебе.

Поисковый робот как автоматизированная система сбора и обработки данных в сети Веб. Веб-краулеры (поисковые роботы). Основные задачи, решаемые поисковыми роботами. Архитектурные особенности поисковых роботов. Классические и специализированные Веб-краулеры. Поиск и обновление значимых Веб-страниц. Основные метрики значимости Веб-страниц.

Задача индексирования. Индексные структуры. Построение индекса, основные проблемы и задачи. Индексные структуры: инвертированный файл (блочное индексирование), сигнатурные файлы, суффиксные деревья.

Раздел 3. Обеспечение учебных занятий

3.1. Методическое обеспечение

3.1.1. Методические указания по освоению дисциплины

Самостоятельная работа студентов включает в себя решение задач, изучение лекционного материала, учебников, учебных пособий и иных материалов. Время и место самостоятельной работы выбираются студентами по своему усмотрению с учетом рекомендаций преподавателя.

Самостоятельную работу над дисциплиной следует начинать с изучения учебно-методического комплекса, который содержит основные требования к знаниям, умениям, навыкам. Необходимо также вспомнить рекомендации преподавателя, данные в ходе лекционных занятий или консультаций, затем приступать к изучению отдельных разделов и тем.

Подготовка к лекции заключается в следующем:

- внимательно изучить материал предыдущей лекции;
- целесообразно составить краткий конспект или схему, отображающую смысл и связи основных понятий данного раздела, включенных в него тем, а затем, полезно изучить выдержки из литературы;
- узнать тему предстоящей лекции;
- ознакомиться с учебным материалом по учебнику и учебным пособиям;
- записать возможные вопросы, которые вы зададите лектору на лекции.

Подготовка к практическим занятиям:

- выполнить практические задания домашней работы;
- внимательно изучить материал лекций, относящихся к данному семинарскому занятию, ознакомиться с учебным материалом по учебнику, учебным пособиям или рекомендованным электронным ресурсам;
- выписать основные термины;
- уяснить, какие учебные элементы остались неясными, и сформулировать вопросы, которые необходимо задать преподавателю на занятии или консультации;
- готовиться можно индивидуально, парами или в составе малой группы, последние являются эффективными формами работы.

Изучение дисциплины заканчивается экзаменом. При непосредственной подготовке к экзамену рекомендуется тщательно изучить формулировку каждого вопроса, понять его сущность. В соответствии со смыслом составить план ответа. План ответа желательно развернуть, приложив к нему ссылки на конкретные источники. Отметить пробелы в знаниях, которые следует ликвидировать в ходе консультации.

3.1.2. Методическое обеспечение самостоятельной работы

Комплект слайдов презентаций материалов.

Конспекты лекций по темам, предусматривающим самостоятельную работу.

Комплект заданий с указаниями для самостоятельного выполнения студентами индивидуальных заданий.

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

Промежуточная аттестация проходит в форме презентаций командных разработок, описаний роли каждого обучающегося и научно-практического вклада в командную разработку проекта, и подготовленных докладов по темам семинаров.

Для прохождения промежуточной аттестации необходимо выполнить все практические задания в командном проекте, подготовить доклад, представить результаты проведенных экспериментов и сдать экзамен по курсу.

Экзамен проводится в устной форме по билетам. Список экзаменационных вопросов предоставляется обучающимся не позднее, чем за две недели до экзамена. Билет содержит два теоретических вопроса. По окончании подготовки к ответу обучающийся устно излагает содержание экзаменационных вопросов экзаменатору и отвечает на вопросы экзаменатора. После устного ответа по вопросам экзаменационного билета экзаменатор вправе задать обучающемуся любые вопросы из списка экзаменационных вопросов (дополнительные вопросы).

Итоговая оценка «отлично» ставится при условии успешного выполнения всех практических заданий в течении семестра (с учетом активной работы в команде над проектом), наличия подготовленного доклада и полного ответа на два теоретических вопроса по курсу, знания основных определений и методов.

Итоговая оценка «хорошо» ставится при условии успешного выполнения всех практических заданий в течении семестра (с учетом активной работы в команде над проектом), наличия подготовленного доклада и полного ответа хотя бы на один теоретический вопрос по курсу.

Итоговая оценка «удовлетворительно» ставится при условии успешного выполнения всех практических заданий в течении семестра (с учетом активной работы в команде над проектом), неполного ответа на теоретический вопрос по курсу.

Итоговая оценка «неудовлетворительно» ставится при условии отсутствия хотя бы одного практического задания, а также при пассивной работе в команде над проектом, отсутствия знаний основных определений и методов.

Преподаватель имеет право предоставить информацию о задолженностях студента в аттестационную комиссию.

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

Примерные вопросы к экзамену по дисциплине:

1) Общая информация об Интернете, гипертексте, Веб-пространстве. Исследования Netcraft. Статистические и динамические части Веб-пространства, уравнение Бартона-Кеблера. Проблемы хранимой информации в Веб-пространстве.
2) Модель Веб-пространства. Модель Веб-пространства Брёдера.
3) Основные задачи информационного поиска. Понятие релевантности. Информационно-поисковые системы (ИПС) в Веб-пространстве. Классификация информационно-поисковых систем в Веб-пространстве (каталог, поисковые машины, мета-поисковые системы).
4) Методы ранжирования результатов информационного поиска в Вебе.
5) Интеллектуальный анализ логов запросов пользователей в современных поисковых машинах (Mining query logs to improve web search engines' operations).
6) Методы обработки текстовой информации: Bag of words, Word2Vec, WordNet, лемматизация, стемминг.
7) Веб-краулеры (поисковые роботы). Основные задачи, решаемые поисковыми роботами. Архитектурные особенности поисковых роботов.
8) Тематические Веб-краулеры. Поиск и обновление значимых Веб-страниц. Основные метрики значимости Веб-страниц.
9) Методы обновления данных в индексе с помощью поисковых роботов. Стратегии равномерного и пропорционального обновления.
10) Алгоритм PageRank для вычисления весов Веб-странц сайтов.
11) Меры центральности графов, используемые при анализе данных социальных сетей (Indegree, Outdegree, closeness, eigenvector-centrality, cliques, betweenness и т.п.).
12) Задача индексирования. Индексная структура инвертированный файл/индекс, их принципы структуризации информации и архитектурные особенности.
13) Задача индексирования. Индексная структура сигнатурные файлы, их принципы структуризации информации и архитектурные особенности.
14) Задача индексирования. Индексная структура суффиксные деревья, их принципы структуризации информации и архитектурные особенности.
15) Распределенное (MapReduce) и динамическое индексирование.
16) Алгоритмы сжатия: непараметрические алгоритмы дельта- и гамма-кодирования Элиаса.
17) Алгоритмы сжатия: параметрический алгоритм кодирования Голомба.

Примерный перечень тем докладов:

1) Модуль I:
   1.1. Методы ранжирования результатов информационного поиска в Вебе.
   1.2. Mining query logs to improve web search engines' operations.
   1.3. Методы обработки текстовой информации: Bag of words, Word2Vec, WordNet, лемматизация, стемминг.
2) Модуль II:
   2.1. Методы обновления данных в индексе с помощью поисковых роботов. Стратегии равномерного и пропорционального обновления.
   2.2. Алгоритм PageRank для вычисления весов Веб-странц сайтов.

2.3. Меры центральности графов, используемые при анализе данных социальных сетей.

3) Модуль III:

3.1. Распределенное (MapReduce) и динамическое индексирование.

3.2. Алгоритмы сжатия: непараметрические алгоритмы дельта- и гамма-кодирования Элиаса.

3.3. Алгоритмы сжатия: параметрический алгоритм кодирования Голомба.

Примеры практических заданий:

1) Модуль I:

1.1. Разработка и тестирование синтаксического анализатора для обработки html-страниц.

1.2. Разработка и тестирование синтаксического анализатора для автоматизированной обработки документов форматов .pdf и .doc

2) Модуль II:

2.1. Разработка простейшей модели поискового робота с классическим алгоритмом сбора и обработки данных в сети Веб.

2.2. Автоматизированный сбор данных с помощью простейшей модели поискового робота и специализированного алгоритма обхода на примере сайта СПбГУ и МГУ.

2.3. Сбор статистики обработанных страниц: объем страниц и всех ссылок, количество неработающих страниц, количество внутренних поддоменов, количество страниц на внешние ресурсы.

3) Модуль III:

3.1. Реализация и тестирование индексной структуры на основе инвертированного файла.

3.2. Применение метода сжатия инвертированного файла с использованием дельта- и гамма-кодирования Элиаса.

3.3. Тестирование процесса индексирования на собранных на 2м этапе веб-страниц сайта СПбГУ или МГУ (скорость процесса индексирования по количеству текстовых документов около 40 тыс., проверить на сколько эффективно индексирование с использованием алгоритма сжатия уменьшает объемы хранимой информации по сравнению с классической ситуацией, не предусматривающей использование алгоритма сжатия).

Практические задания выполняются в команде, состоящей программиста, инженера по контролю качества (QA) и менеджера проекта.

Проджект менеджер - подготавливает презентацию проекта (докладывает тему семинара), распределяет задачи по проекту для программиста и QA (в виде таблицы, кто и что делал, время выполнения каждой задачи), совместно с программистом продумывает и формализует архитектуру программного комплекса. Готовит полную отчетность по проекту.

Программист - имплементирует функциональность (таким образом, чтобы QA было удобно писать юнит-тесты, которые покрывают функциональность) и помогает QA.

QA - проектирует тест-кейсы и разрабатывает юнит-тесты. Собирает статистику о кол-ве тестов, их специфики прохождения/не прохождения.

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

Для оценки обучающимися содержания и качества учебного процесса используется анкета-отзыв установленная локальными актами СПбГУ.

3.2. Кадровое обеспечение

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

К преподаванию привлекаются преподаватели, имеющие ученую степень, а также главные и ведущие специалисты в этой области. Допускается проведение занятий обучающимся в аспирантуре (под руководством научного руководителя) для прохождения педагогической практики.

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

Не требуется.

3.3. Материально-техническое обеспечение

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

Компьютерный класс с количеством рабочих мест соответствующим количеству обучающихся с учетом рабочего места преподавателя, мультимедийный проектор, доска.

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

Универсальные компьютеры, объединенные в локальную сеть, мультимедийное оборудование (проектор, экран). Системное программное обеспечение общего назначения (MS Windows любой современной версии).

3.3.3. Характеристики специализированного оборудования

Отсутствуют.

3.3.4. Характеристики специализированного программного обеспечения

Microsoft Visual Studio 2015 (или старше).

3.3.5. Перечень и объёмы требуемых расходных материалов

Не требуются.

3.4. Информационное обеспечение

3.4.1. Список обязательной литературы

1. К. Маннинг, П. Рагхаван, Х. Шютце. Введение в информационный поиск // М.: ИД «Вильямс». – 2011. – С. 528.

2. И.С. Блеканов, Е.С. Романенко, Г.С. Шиманская. Введение в информационный веб-поиск. Часть 1: Обобщенная структура сети Веб. Поисковые роботы. Индексные структуры // Учебно-методическое пособие. – СПб. 2019. – 46с.

3.4.2. Список дополнительной литературы

1. David Easley and Jon Kleinberg. Networks, Crowds, and Markets: Reasoning About a Highly Connected World // Cambridge University Press. – 2010. – P. 820. (http://www.cs.cornell.edu/home/kleinber/networks-book/)

2. Thelwall, M. Big Data and Social Web Research Methods // University of Wolverhampton Press. – 2014. – P. 142. (http://www.scit.wlv.ac.uk/~cm1993/papers/IntroductionToWebometricsAndSocialWebAnalysis.pdf)

3.4.3. Перечень иных информационных источников

1. Официальный сайт электронной библиотеки The ACM Digital Library: http://dl.acm.org/

2. Официальный сайт электронной библиотеки Google Book: http://books.google.com

4. Официальный сайт электронной библиотеки IEEE Xplore Digital Library: http://ieeexplore.ieee.org

Раздел 4. Разработчики программы

Фамилия, имя, отчество	Учёная степень	Учёное звание	Должность	Контактная информация (служебный адрес электронной почты, служебный телефон)
Блеканов Иван Станиславович	к.т.н.		доцент	i.blekanov@spbu.ru