Белорусский государственный университет

УТВЕРЖДАЮ
Проректор по учебной работе

_____ А.Л. Толстик

29.

Регистрационный № УД- 3140/ уч.

ВВЕДЕНИЕ В ИНФОРМАЦИОННЫЙ ПОИСК

Учебная программа учреждения высшего образования по учебной дисциплине для специальности второй ступени высшего образования (магистратуры) с углубленной подготовкой специалиста:

1-31 81 09 «Алгоритмы и системы обработки больших объемов информации»

2016 г.

Учебная программа составлена на основе образовательного стандарта высшего образования ОСВО 1-31 81 09-2014 и учебного плана G31-219/уч. от 30.05.2016.

Составители:

Т.А. Хаткевич, инженер-программист ООО «ЯндексБел».

Рекомендована к утверждению:

Кафедрой дискретной математики и алгоритмики Белорусского государственного университета (протокол № 14 от 19 мая 2016 г.);

Методической комиссией факультета прикладной математики и информатики Белорусского государственного университета (протокол № 6 от 24 мая 2016 г.).

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Учебная программа по учебной дисциплине «Введение в информационный поиск» разработана в соответствии с учебным планом и образовательным стандартом второй ступени высшего образования (магистратуры) с углубленной подготовкой специалиста по специальности 1-31 81 09 «Алгоритмы и системы обработки больших объемов информации».

Учебная дисциплина «Введение в информационный поиск» знакомит магистрантов с задачами информационного поиска – автоматической обработкой текстов, индексацией и ранжированием.

Основой для изучения учебной дисциплины являются следующие дисциплины первой ступени высшего образования: «Программирование», «Теория вероятностей и математическая статистика».

Цель преподавания учебной дисциплины «Введение в информационный поиск»: создание базы для понимания и реализации магистрантами составляющих частей информационно-поисковой системы и формирование у магистрантов умения реализовывать различные компоненты информационно-поисковых систем и анализировать данные, получающиеся в результате взаимодействия информационно-поисковых систем с пользователем.

При изложении материала учебной дисциплины важно показать спектр применения информационно-поисковых систем при решении прикладных задач обработки и анализа больших объемов информации, возникающих в различных областях науки, техники, экономики и др.

Основные задачи, решаемые при изучении учебной дисциплины «Введение в информационный поиск»:

– изучение подходов к решению задач информационного поиска;
– изучение особенностей реализации соответствующих алгоритмов;
– изучение методов использования кластерных систем обработки больших объёмов данных для задач информационно-поисковых систем.

В результате изучения дисциплины магистрант должен

знать:

– базовые принципы построения информационно-поисковых систем;
– современные методы анализа Интернета;
– основные методы обработки неструктурированных текстовых данных;

уметь:

– реализовывать различные компоненты информационно-поисковых систем;
– анализировать данные, получающиеся в результате взаимодействия информационно-поисковых систем с пользователем;

владеть:

– методами реализации различных компонент информационно-поисковых систем;
– методами анализа неструктурированных текстовых данных больших объемов.

Освоение образовательной программы магистратуры должно обеспечить формирование следующих групп компетенций:

академических компетенций – углубленных научно-теоретических, методологических знаний и исследовательских умений, обеспечивающих разработку научно-исследовательских, инновационной деятельности, непрерывного самообразования (АК-1. Способность к самостоятельной профессиональной деятельности (анализ, сопоставление, систематизация, абстрагирование, моделирование, проверка достоверности данных, принятие решений и др.), готовность генерировать и использовать новые идеи. АК-2. Методологические знания и исследовательские умения, обеспечивающие решение прикладных задач и инновационной деятельности. АК-3. Способность к постоянному самообразованию);

социально-личностных компетенций – личностных качеств и умений следовать социально-культурным и нравственным ценностям; способностей к социальному, межкультурному взаимодействию, критическому мышлению; социальной ответственности, позволяющих решать социально-профессиональные, организационно-управленческие, воспитательные задачи (Магистр должен: СЛК-1. Учитывать социальные и нравственно-этические нормы в социально-профессиональной деятельности. СЛК-2. Быть способным к сотрудничеству и работе в команде. СЛК-3. Владеть коммуникативными способностями для работы в междисциплинарной и международной среде. СЛК-4. Совершенствовать и развивать свой интеллектуальный и общекультурный уровень, добиваться нравственного и физического совершенствования своей личности. СЛК-5. Формировать и аргументировать собственные суждения и профессиональную позицию. СЛК-6. Логично, аргументированно и ясно строить устную и письменную речь, использовать навыки публичной речи, ведения дискуссии и полемики. СЛК-7. Проявлять инициативу и креативность, в том числе в нестандартных ситуациях);

профессиональных компетенций – углубленных знаний по специальным дисциплинам и способностей решать сложные профессиональные задачи, задачи научно-исследовательской и научно-педагогической деятельности, разрабатывать и внедрять инновационные проекты, осуществлять непрерывное профессиональное самосовершенствование (Магистр должен быть способен: ПК-1. Квалифицированно использовать современные достижения по разработке и анализу алгоритмов и современные информационные технологии. ПК-2. Строить математические модели для прикладных задач и проводить теоретическое и экспериментальное их исследование. ПК-3. Разрабатывать эффективные численные алгоритмы и интегрировать их в компьютерные системы. ПК-4. Обосновывать выбор методов и инструментов для решения прикладных задач. ПК-5. Обосновывать достоверность полученных результатов. ПК-6. Осваивать и реализовывать управленческие инновации в профессиональной деятельности. ПК-7. Формулировать выводы и рекомендации по применению современных достижений науки в инновационной деятельности).

Учебная программа рассчитана на 112 часов, из них 54 аудиторных часа, в том числе 18 лекционных часов и 36 часов практических занятий.

Рекомендуемая форма текущей аттестации – зачет.

СОДЕРЖАНИЕ УЧЕБНОГО МАТЕРИАЛА

Тема 1. Введение. Задачи информационного поиска. Области применения информационного поиска. Архитектура информационно-поисковых систем. Классификация ИПС. Обзор поисковых систем.

Тема 2. Модели информационного поиска. Булев поиск. Понятие инвертированного индекса. Векторная модель документа. Вероятностные модели в информационном поиске. TF-IDF, BM25. Языковые модели.

Тема 3. Компьютерная лингвистика в информационном поиске. Кодировки и языки. Unicode. Нормализация лексем. Стемминг и лемматизация. N-граммные модели. Исправление опечаток. Синонимия. LSA и Word2Vec.

Тема 4. Поисковый робот. Архитектура поискового робота. Robots.txt. Sitemaps. Поддержка актуальности информации в хранилище робота. Точные и нечеткие дубликаты. Алгоритмы обнаружения нечетких дубликатов.

Тема 5. Ссылочный граф интернета. Структура и размер Веб-графа. Ссылочный спам. PageRank. Авторитетность источников и алгоритм HITS. Модель вычислений MapReduce.

Тема 6. Индексация документов. Компоненты индекса. Поисковые структуры для словарей. Списки словопозиций и указатели пропусков. Статический индекс. Блочное индексирование, основанное на сортировке. Индексирование в оперативной памяти. Слияние индексов. Динамический индекс. Распределенное индексирование. Шардирование индекса. Обработка запроса. Static Index Pruning. Методы сжатия словаря. Методы сжатия инвертированного файла. Оценка размера индекса поисковой системы.

Тема 7. Оценка качества поисковых систем. Аспекты качества поисковых систем. Оценка результатов поиска. Оценка релевантности. Статистическая значимость изменений. Тестовые коллекции. Краудсорсинг. A/B-тестирование.

Тема 8. Ранжирование. Статическое ранжирование. Эмпирические методы. Задача обучения ранжированию (learning to rank) и способы ее решения. Градиентный бустинг на решающих деревьях. Алгоритм LambdaMART.

Тема 9. Обратная связь по релевантности. Методы обратной связи по релевантности. Алгоритм Роккио. Обратная связь по псевдорелевантности. Расширение и переформулировка запроса. Автоматическое составление тезауруса.

Тема 10. Социальный поиск. Web 2.0. Поиск по тегам. Поиск сообществ. Социальный граф. Проблема deep web.

УЧЕБНО-МЕТОДИЧЕСКАЯ КАРТА УЧЕБНОЙ ДИСЦИПЛИНЫ

№п/п	Название раздела, темы	Количество часов				Количество часов УСР	Форма контроля знаний
		Аудиторные					
		Лекции	Практ. и сем. занятия	Лаб. занятия	Иное		
1	Введение. Модели информационного поиска	2					
	Изучение систем инфопоиска с открытым исходным кодом		2				
	Токенизация документов, построение простого обратного индекса		2				Отчет по домашнему заданию
2	Компьютерная лингвистика в информационном поиске	2					Устный опрос
	Лемматизация и нормализация текста. Простое текстовое ранжирование		4				Отчет по домашнему заданию
3	Поисковый робот	2					Устный опрос
	Реализация обхода страниц Интернета		4				Отчет по домашнему заданию, выступление на семинаре
4	Ссылочный граф интернета	2					Устный опрос
	Алгоритм PageRank и его модификации		4				Отчет по домашнему заданию, выступление на семинаре
5	Индексация документов	2					Устный опрос
	Индексирование скачанных документов. Сжатие индекса		4				Отчет по домашнему заданию, выступление на семинаре
6	Оценка качества поисковых систем	2					Устный опрос

	Работа с тестовыми коллекциями документов		4				Выступление на семинаре
7	Ранжирование	2					Устный опрос
	Извлечение признаков из пар запрос-документ, Применение открытого алгоритма ранжирования		4				Отчет по домашнему заданию, выступление на семинаре
8	Обратная связь по релевантности	2					Устный опрос
	Взаимодействие компонент поисковой системы		4				Выступление на семинаре
9	Социальный поиск	2					
	Поиск сообществ		4				
ИТОГО		**18**	**36**				

ИНФОРМАЦИОННО-МЕТОДИЧЕСКАЯ ЧАСТЬ

Рекомендуемая литература

Основная

1. Маннинг К. Введение в информационный поиск. / К. Маннинг, П. Рагхаван, Х. Шютце. – М.: Вильямс, 2014. – 528 с.
2. Buettcher S. Information Retrieval: Implementing and Evaluating Search Engines / S. Buettcher, C. Clarke, G. Cormack. – Massachusets Institute of Technology, 2010. – 632 p.
3. Olston, C. Web Crawling / C. Olston, M. Najork. // Foundations and Trends in Information Retrieval. – Vol. 4, No. 3. – 2010. – pp. 175–246.
4. Croft, B. Search Engines: Information Retrieval in Practice. / B. Croft, D. Metzler, T. Strohman – Addison Wesley, 2009. – 552 p.

Дополнительная

1. Berkhin, P. A Survey on PageRank Computing / P. Berkhin // Internet Mathematics. – 2005. – Vol. 2. – No. 1. – pp. 73-120.
2. Burges, C. From RankNet to LambdaRank to LambdaMART: An Overview / C. Burges // Microsoft Research Technical Report – 2010. – MSR-TR-2010-82. – 19 p.
3. Hastie, T. The Elements of Statistical Learning / T. Hastie, R. Tibshirani, J. Friedman. – Springer, 2009. – 745 p.

Рекомендации по контролю качества усвоения знаний и проведению аттестации

На лекционных занятиях по учебной дисциплине «Введение в информационный поиск» рекомендуется используются элементы проблемного обучения: проблемное изложение некоторых аспектов, использование частично-поискового метода. На лекционных занятиях следует акцентировать внимание слушателей на изученных фактах. На лабораторных занятиях рекомендуется реализовывать прототипы изложенных на лекциях алгоритмов и методов.

Перечни рекомендуемых форм диагностики компетенций

Для аттестации обучающихся на соответствие их персональных достижений поэтапным и конечным требованиям образовательной программы создаются фонды оценочных средств, включающие типовые задания, контрольные работы и тесты. Оценочными средствами предусматривается оценка способности обучающихся к творческой деятельности, их готовность вести

поиск решения новых задач, связанных с недостаточностью конкретных специальных знаний и отсутствием общепринятых алгоритмов.

Для диагностики компетенций в рамках учебной дисциплины рекомендуется использовать следующие формы:

1. Устная форма: опросы, выступления на семинарских занятиях.
2. Письменная форма: отчеты по домашним практическим упражнениям.

Контрольные мероприятия проводятся в соответствии с учебно-методической картой дисциплины. В случае неявки на контрольное мероприятие по уважительной причине студент вправе по согласованию с преподавателем выполнить его в дополнительное время. Для студентов, получивших неудовлетворительные оценки за контрольные мероприятия, либо не явившихся по неуважительной причине, по согласованию с преподавателем и с разрешения заведующего кафедрой мероприятие может быть проведено повторно.

Оценка текущей успеваемости рассчитывается как среднее оценок за отчеты по домашним практическим упражнениям и оценок за участие в семинарских занятиях.

Итоговая аттестация предусматривает проведение зачета. При этом рекомендуется использовать оценивание успеваемости на основе модульно-рейтинговой системы.

ПРОТОКОЛ СОГЛАСОВАНИЯ УЧЕБНОЙ ПРОГРАММЫ

Название учебной дисциплины, с которой требуется согласование	Название Кафедры	Предложения об изменениях в содержании учебной программы учреждения высшего образования по учебной дисциплине	Решение, принятое кафедрой, разработавшей учебную программу (с указанием даты и номера протокола)
Программирование	Технологий программирования	Нет	Оставить содержание учебной дисциплины без изменения, протокол № 14 от 19.05.2016 г
Теория вероятностей и математическая статистика	Теории вероятностей и математической статистики	Нет	Оставить содержание учебной дисциплины без изменения, протокол № 14 от 19.05.2016 г

ДОПОЛНЕНИЯ И ИЗМЕНЕНИЯ К УЧЕБНОЙ ПРОГРАММЕ

на _____/_____ учебный год

№№ Пп	Дополнения и изменения	Основание

Учебная программа пересмотрена и одобрена на заседании кафедры дискретной математики и алгоритмики (протокол № ____ от _____ 201_ г.)

Заведующий кафедрой

_____ _____ _____

(ученая степень, звание) (подпись) (И.О. Фамилия)

УТВЕРЖДАЮ
Декан факультета

_____ _____ _____

(ученая степень, звание) (подпись) (И.О.Фамилия)