Федеральное государственное автономное образовательное учреждение высшего образования
"Национальный исследовательский университет
"Высшая школа экономики"

Факультет Компьютерных Наук
Базовая кафедра Яндекс

Рабочая программа дисциплины «Информационный поиск»

для образовательной программы «Науки о данных»
направления подготовки 01.04.02 "Прикладная математика и информатика"
уровень магистра

Разработчик программы
Федотов С.Н., sfedotov@hse.ru

Одобрена на заседании базовой кафедры Яндекс
«___»_____ 2017 г.

Заведующий Кафедрой
М.А. Бабенко_____

Утверждена Академическим советом образовательной программы
«___»_____ 2017 г., № протокола_____

Академический руководитель образовательной программы
С.О. Кузнецов _____

Москва, 2017

1 Область применения и нормативные ссылки

Настоящая программа учебной дисциплины устанавливает требования к образовательным результатам и результатам обучения студента и определяет содержание и виды учебных занятий и отчетности.

Программа предназначена для преподавателей, ведущих дисциплину Информационный поиск», учебных ассистентов и студентов направления подготовки/специальности 01.04.02
«Прикладная математика и информатика», обучающихся по образовательной программе «Науки о данных».

Программа учебной дисциплины разработана в соответствии с:

- Образовательным стандартом федерального государственного автономного образовательного учреждения высшего профессионального образования «Национального исследовательского университета «Высшая школа экономики»;
- Образовательной программой подготовки магистра по направлению 01.04.02 «Прикладная математика и информатика», специализации «Анализ Интернет-данных».
- Объединенным учебным планом университета по образовательной программе «Науки о данных», утвержденным в 2018 г.

2 Цели освоения дисциплины

Целями освоения дисциплины является получение студента основ информационного поиска, освоение методов построения информационно-поисковых систем. Будут изучены задачи информационного поиска и архитектура поисковых систем, машинное обучение в поиске и компьютерная лингвистика, построение поискового индекса и обнаружение дубликатов, поисковый робот и оценка качества. Решение предлагаемых практических заданий связано со знакомством с широким спектром технологий и алгоритмов, применяемых на практике при построениии компонентов поисковой системы.

3 Компетенции обучающегося, формируемые в результате освоения дисциплины

В результате освоения дисциплины студент осваивает компетенции:

Компетенция	Код по ОС ВШЭ	Уровень формирования компетенции	Дескрипторы – основные признаки освоения (показатели достижения результата)	Формы и методы обучения, способствующие формированию и развитию компетенции	Форма контроля уровня сформированности компетенции
Способен вести исследовательскую деятельность, включая анализ проблем, постановку целей и задач, выделение объекта и	УК-6	РБ	Умеет распознавать базовые закономерности в задачах информационного поиска.	Стандартные (лекционно- семинарские). Самостоятельные внеаудиторные занятия.	Итоговый экзамен

Компетенция	Код по ОС ВШЭ	Уровень формирования компетенции	Дескрипторы – основные признаки освоения (показатели достижения результата)	Формы и методы обучения, способствующие формированию и развитию компетенции	Форма контроля уровня сформированности компетенции
предмета исследования, выбор способа и методов исследования, а также оценку его качества					
Способен описывать проблемы и ситуации профессиональной деятельности, используя язык и аппарат математики	ПК-1	СД	Умеет формулировать математически и проводить анализ задач относящиеся к комбинаторной оптимизации	Стандартные (лекционно- семинарские). Самостоятельные внеаудиторные занятия.	Итоговый экзамен
Способен писать, оформлять, отлаживать и оптимизировать программный код.	ПК-5	СД	Умеет реализовать и оптимизировать несложную поисковую систему.	Стандартные (лекционно- семинарские). Самостоятельные внеаудиторные занятия.	Итоговый экзамен
Способен разработать математическую модель и провести её анализ для поставленной теоретической или прикладной задачи	ПК-8	СД	Умеет формулировать задачи в виде задач комбинартной оптимизации. В случае линейных программ, умеет анализировать ицелочисленность, строить двойственные программы и использовать их для решения исходной задачи.	Стандартные (лекционно- семинарские). Самостоятельные внеаудиторные занятия.	Итоговый экзамен
Способен разработать и реализовать в виде программного модуля алгоритм решения поставленной теоретической или прикладной задачи на основе	ПК-9	СД			

Компетенция	Код по ОС ВШЭ	Уровень формирования компетенции	Дескрипторы – основные признаки освоения (показатели достижения результата)	Формы и методы обучения, способствующие формированию и развитию компетенции	Форма контроля уровня сформированности компетенции
математической модели					

Виды и задачи профессиональной деятельности:

Научно-исследовательские	
Исследование и разработка математических моделей и методов, алгоритмов и программного обеспечения по тематике проводимых научно-исследовательских проектов;	НИД-3
Проектные и производственно-технологические	
Разработка и исследование алгоритмов, вычислительных моделей и моделей данных для реализации элементов новых (или известных) систем информационных технологий	ПД-2
Разработка архитектуры, алгоритмических и программных решений системного и прикладного программного обеспечения	ПД-3
Разработка программного и информационного обеспечения компьютерных систем, автоматизированных систем вычислительных комплексов, сервисов, операционных систем и распределенных баз данных	ПД-4

4 Место дисциплины в структуре образовательной программы

Настоящая дисциплина относится к профессиональному циклу, базовой части вариативного профиля.

5 Тематический план учебной дисциплины

№	Название раздела	Всего часов	Аудиторные часы				Самостоятельная работа
			Лекции	Семинары	Практические занятия	Другие виды работы[1]	
1	Введение. Простые модели документа	24	2	4			18
2	Глубинное обучение в задачах информационного поиска	24	2	4			18
3	Поисковый робот	26	4	6			16
4	Поиск дубликатов	24	4	6			14
5	Построение индекса	18	2	4			12
6	Обход документов	18	2	4			12
7	Learning to Rank	18	4	4			10
	Итого	152	20	32			100

[1] Указать другие виды аудиторной работы студентов, если они применяются при изучении данной дисциплины.

6 Формы контроля знаний студентов

Тип контроля	Форма контроля	1 год		Параметры **
		3	4	
	Контрольное задание	*	*	Решение практических задач в учебной аудитории
	Домашнее задание	*		Решение сложных задач информационного поиска
Итоговый	Экзамен		*	Теоретический

7 Критерии оценки знаний, навыков

Для прохождения контроля студент должен продемонстрировать понимание архитектуры поисковых систем, умение пользоваться машинным обучением в задачах информационного поиска, умение строить поисковый индекс и обнаруживать дубликаты, а также умение применять все изученные концепции и инструменты на практике.

Студент должен уверенно владеть следующим:

1. Уметь строить векторные представления документов.
2. Уметь использовать модели глубинного обучения в задачах инфопоиска.
3. Уметь написать поисковой робот.
4. Уметь находить дубликаты документов.
5. Уметь эффективно индексировать документы.
6. Уметь эффективно решать задачу ранжирования.
7. Уметь оценивать поисковые системы в оффлайн и в онлайн режиме.

8 Содержание дисциплины

Программа лекций:

1. Введение. Простые модели документа. Обобщённая векторная модель. Языковые модели. Тематическое моделирование. Word2Vec и Doc2Vec.

2. Глубинное обучение для задач информационного поиска. Representation Strategies. Shift Invariant Neural Architectures. Autoencoder vs Siamese Network. Модель DSSM. Interaction-based Networks.

3. Поисковый робот.

5. Обнаружение дубликатов. Виды дубликатов. Шинглы, Min-Hash. Odd Sketch. SimHash. Вероятностный поиск.

6. Обход страниц. Архитектурные аспекты. Качество обхода. Обход свежих страниц.

7. Построение и использование индекса. Сжатие.

8. Learning to Rank. Метрики ранжирования. Алгоритмы ранжирования.

Программа семинаров

1. Модель LSI.
2. Глубинное обучение для задач информационного поиска.
3. Поисковый робот.
5. SimHash.
6. Вероятностный поиск SimHash.

7. VarInt
9. Обход документов.
8. Learning to Rank.

9 Оценочные средства для текущего контроля и аттестации студента. Порядок формирования оценок по дисциплине

Контроль:

- Контрольное мероприятие с решение практических задач в учебной аудитории.
- Контрольное мероприятие с решение теоретических задач в учебной аудитории.
- Практическое домашнее задание на решение сложных задач информационного поиска как с использованием, так и без использования сторонних инструментов.
- Теоретический экзамен по программе курса.

Оценка за каждую из частей контроля выставляется исходя из 10 баллов, при этом количество баллов за контрольные работы и домашнюю работу может быть не целым. Итоговая оценка выставляется исходя из формулы:

$$O_{\text{накопленная}} = O_{\text{ДЗ}}$$

$$O_{\text{итог}} = 0.5\, O_{\text{накопленная}} + 0.5\, O_{\text{экз}}$$

Способ округления накопленной оценки по учебной дисциплине: округление вверх.
Способ округления результирующей оценки по учебной дисциплине: арифметический.

10 Учебно-методическое и информационное обеспечение дисциплины

Основная литература

- B. Croft, D. Metzler, T. Strohman Search Engines: Information Retrieval in Practice.
- C.D. Manning, P. Raghavan, H. Schütze Introduction to Information Retrieval.
- S. Büttcher, C.L.A. Clarke, G.V. Cormack Information Retrieval: Implementing and Evaluating Search Engines.
- R.A. Baeza-Yates, B. Ribeiro-Neto Modern Information Retrieval (2nd ed.).
- C. Olston, M. Najork Web Crawling.
- F. Silvestri Mining Query Logs: Turning Search Usage Data into Knowledge.
- Ian H. Witten, Alistair Moffat, Timothy C. Bell Managing gigabytes: compressing and indexing documents and images.
- M. Sanderson Test Collection Based Evaluation of Information Retrieval Systems

Дополнительная литература:

- D. Chakrabarti, R. Kumar, K. Punera A Graph-Theoretic Approach to Webpage Segmentation.
- D. Cai, S. Yu, J-R. Wen, W-Y. Ma Extracting Content Structure for Web Pages based on Visual Representation.