Санкт-Петербургский филиал федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский университет "Высшая школа экономики"»

Факультет Санкт-Петербургская школа
физико-математических и компьютерных наук
Департамент информатики

Рабочая программа дисциплины
Веб-поиск и ранжирование

для образовательной программы «Прикладная математика и информатика»
направления подготовки 01.03.02 «Прикладная математика и информатика»
уровень бакалавриат

Разработчик: Кузнецов Антон Михайлович, amkuznetsov@hse.ru

Утверждена Академическим руководителем образовательной программы

«31» августа 2018 г.

А.В. Омельченко

Санкт-Петербург, 2018

Настоящая программа не может быть использована другими подразделениями университета и другими вузами без разрешения подразделения-разработчика программы.

1. Область применения и нормативные ссылки

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

Программа предназначена для преподавателей, ведущих дисциплину «Веб-поиск и ранжирование», учебных ассистентов и студентов направления 01.03.02 «Прикладная математика и информатика» подготовки бакалавра, обучающихся по бакалаврской программе «Прикладная математика и информатика» и изучающих дисциплину «Веб-поиск и ранжирование».

Рабочая программа дисциплины разработана в соответствии с:

- Образовательным стандартом НИУ ВШЭ по направлению подготовки 01.03.02 «Прикладная математика и информатика» (уровень бакалавриата), утвержденным ученым советом Национального исследовательского университета «Высшая школа экономики», протокол от 03.03.2017 №02.
- Основной профессиональной образовательной программой «Прикладная математика и информатика» направления подготовки 01.03.02 «Прикладная математика и информатика»;
- Объединенным учебным планом университета по образовательной программе «Прикладная математика и информатика», утвержденным в 2018 г.

2. Цели освоения дисциплины

Целями освоения дисциплины «Веб-поиск и ранжирование» являются формирование у студентов теоретических знаний и практических навыков веб-поиска и ранжирования данных.

Студенты получат об основных алгоритмах веб поиска, создадут собственный веб-краулер и проведут оценку качества собранных результатов.

3. Компетенции обучающегося, формируемые в результате освоения дисциплины

В результате освоения дисциплины студент должен:

- Знать технологии оценки качества поиска.
- Уметь собирать данные из веб-ресурсов.
- Иметь навыки (приобрести опыт) использования прямых методов ранжирования и способов проведения ранжирования с использованием машинного обучения.

В результате освоения дисциплины студент осваивает следующие компетенции:

Компетенция	Код по ОС НИУ 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 решать проблемы в профессиональной деятельности на основе анализа и синтеза	УК-3	РБ СД	Знает основные методы веб-поиска. Обрабатывает, анализирует и синтезирует информацию.	Лекции, подготовка к практическим занятиям, работа на практических занятиях,	Письменный курсовой проект, устный экзамен

		МЦ	Решает проблему, используя выбранный метод.	самостоятельная работа	
Способен осуществлять производственную или прикладную деятельность в международной среде	УК-10	РБ СД МЦ	Знает английскую номенклатуру и умеет обсуждать веб-поиск на английском языке. Применяет принципы веб-поиска ранжирования к прикладным задачам. Владеет навыками веб-поиска на английском языке при решении практических задач.	Лекции, подготовка к практическим занятиям, работа на практических занятиях, самостоятельная работа	Письменный курсовой проект, устный экзамен
Способен использовать и адаптировать существующие математические методы и системы программирования для разработки и реализации алгоритмов решения прикладных задач	ОПК-2	РБ СД МЦ	Знает основные алгоритмы машинного обучения, применяемых в области веб-поиска. Строит решающие деревья и нейронные сети для решения задач ранжирования. Проводит обучение и тестирует модели машинного обучения, применяемых в веб-поиске и ранжировании.	Лекции, подготовка к практическим занятиям, работа на практических занятиях, самостоятельная работа	Письменный курсовой проект, устный экзамен
Способен применять и модифицировать математические модели для решения задач в области профессиональной деятельности	ОПК-3	РБ СД МЦ	Знает основные алгоритмы веб-краулинга. Выбирает подходящую архитектуру и программную модель краулера. Разрабатывает и использует веб-краулеры для решения практических задач.	Лекции, подготовка к практическим занятиям, работа на практических занятиях, самостоятельная работа	Письменный курсовой проект, устный экзамен
Способен решать задачи профессиональной деятельности с	ОПК-4	РБ	Знает современные методы, используемые в	Лекции, подготовка к практическим занятиям, работа	Письменный курсовой проект, устный экзамен

использованием существующих информационно-коммуникационных технологий и с учетом основных требований информационной безопасности		СД МЦ	задачах веб-поиска и ранжирования. Обрабатывает базы данных, веб-поиска и ранжирования. Работает с современными пакетами программ, применяемых в задачах веб-поиска и ранжирования.	на практических занятиях, самостоятельная работа	
Способен вести письменную и устную коммуникацию на русском и иностранном языках в рамках профессионального и научного общения	ОПК-5	РБ СД МЦ	Знает основную терминологию и способы построения технических текстов на русском и английском языке. Анализирует статьи, книги, материалы конференций и другие ресурсы для выявления необходимых знаний по методам веб-поиска. Пишет технические отчеты по выполненному проекту.	Лекции, подготовка к практическим занятиям, работа на практических занятиях, самостоятельная работа	Письменный курсовой проект, устный экзамен
Способен собирать, обрабатывать и интерпретировать данные современных научных исследований в области математики и компьютерных наук, необходимых для формирования выводов по соответствующим научным исследованиям	ПК-1	РБ СД МЦ	Знает основную литературу, а также научные журналы и другие ресурсы, содержащие актуальную информацию о методах ранжирования. Умеет анализировать и интерпретировать информацию из различных источников, необходимую для решения научных и профессиональных задач. Критически отбирает методы решения поставленных задач в области веб-	Лекции, подготовка к практическим занятиям, работа на практических занятиях, самостоятельная работа	Письменный курсовой проект, устный экзамен

			поиска, умеет аргументированно обосновать выбор метода веб-поиска в соответствии со спецификой задачи.		
Способен разрабатывать и реализовывать в виде программного модуля алгоритм решения поставленной теоретической или прикладной задачи на основе математической модели	ПК-2	РБ СД МЦ	Знает основные способы оценки качества поиска. Умеет создавать ранжирующие системы. Использует кликовые модели.	Лекции, подготовка к практическим занятиям, работа на практических занятиях, самостоятельная работа	Письменный курсовой проект, устный экзамен
Способен анализировать, писать и редактировать академические и технические тексты на русском и иностранном языках для решения задач профессиональной и научной деятельности в области математики и компьютерных наук	ПК-4	РБ СД МЦ	Знает принципы написания сопровождающей технической документации. Анализирует имеющуюся техническую документацию к программам на русском и английском языках. Создаёт документацию и анализирует академические тексты для решения задач.	Лекции, подготовка к практическим занятиям, работа на практических занятиях, самостоятельная работа	Письменный курсовой проект, устный экзамен

4. Место дисциплины в структуре образовательной программы

Для образовательной программы «Прикладная математика и информатика» направления подготовки 01.03.02 «Прикладная математика и информатика» настоящая дисциплина относится к базовой профильной части блока дисциплин.

Для освоения дисциплины необходимы компетенции, полученные в ходе изучения дисциплин:

- Машинное обучение;
- Теория вероятностей и математическая статистика.

5. Тематический план учебной дисциплины

Курс рассчитан на 60 часов аудиторной нагрузки, из них 30 часов лекций и 30 часов практических занятий, общим объемом 5 зачетных единиц (190 часов).

№	Название раздела	Всего	Аудиторные часы	Самостоя-тельная

		часов	Лек-ции	Семи-нары	Практи-ческие занятия	работа
1	Оценка качества информационного поиска	51	8	0	8	35
2	Подготовка данных для поиска, обработка запроса	44	7	0	7	30
3	Классические подходы к ранжированию, применение семантических методов и машинного обучения	63	10	0	10	43
4	Федеративный поиск, кликовые модели	32	5	0	5	22
ИТОГО		190	30	0	30	130

6. Содержание дисциплины

Раздел 1 Оценка качества информационного поиска	
Тема 1	Тестовые коллекции. Меры качества. Поведение пользователей. A\B тестирование. Интерливинг.
Раздел 2 Подготовка данных для поиска, обработка запроса	
Тема 1	Получение и подготовка данных для поиска Обход веба (crawling). Спам. Обработка текста: токенизация, нормализация. Индексация.
Тема 2	Обработка запросов Исправление опечаток. Расширение запроса. Предложение альтернативных запросов. Автодополнение запроса.
Раздел 3 Классические подходы к ранжированию, применение семантических методов и машинного обучения	
Тема 1	Классические подходы к ранжированию Векторная модель, tf-idf. Вероятностная модель, BM25. Языковые модели.
Тема 2	Ранжирование с помощью семантических методов Латентно-семантический анализ. Тематические модели. Распределенные представления для слов и документов. PageRank, HITS.
Тема 3	Машинное обучение для ранжирования Машинное обучение для ранжирования: point-wise, pair-wise, list-wise. Обзор современных проблем и задач в области информационного поиска.
Раздел 4 Федеративный поиск, кликовые модели	
Тема 1	Федеративный поиск Распределенный поиск (distributed IR). Агрегированный поиск (aggregated search).
Тема 2	Кликовые модели Каскадные и позиционные модели. Модели на основе нейронных сетей

7. Оценочные средства

7.1. Формы контроля знаний студентов

Тип контроля	Форма контроля	4 год		Параметры
		1 модуль	2 модуль	
Текущий	Курсовой проект	*	*	Письменный курсовой проект
Итоговый	Устный экзамен		*	Экзамен в устной форме

7.2. Критерии и шкалы оценки, примеры заданий

7.2.1. Текущий контроль

Текущий контроль успеваемости по дисциплине осуществляется с помощью следующих оценочных средств.

КУРСОВОЙ ПРОЕКТ

В качестве курсового проекта обучающиеся должны будут реализовать проект, использующий современные методы веб-поиска и ранжирования для решения прикладной задачи. Срок выполнения домашнего задания - 5 недель.

Примерная тематика исследовательских работ/проектов/докладов:

Тема 1. Поиск статьи для цитирования по отрывку текста

Тема 2. Программа-помощник для написания кода с возможностью поиска по специализированным ресурсам

Тема 3. Анализ научных публикация для предсказания рейтинга свежих статей

Критерии оценивания и шкала оценки курсовой работы

Оценка	Критерии выставления оценки
«Отлично» (8-10)	Создана функционирующая система для решения поставленной задачи или продемонстрирована невозможность решения поставленной задачи современными методами веб-поиска и ранжирования.
«Хорошо» (6-7)	Создана система для решения поставленной задачи, однако она не учитывает всех особенностей поставленной задачи
«Удовлетворительно» (4-5)	Создана система для решения поставленной задачи, хотя задача не может быть решена современными методами веб-поиска и ранжирования; создана система поиска, которая при этом не решает поставленную задачу
«Неудовлетворительно» (0-3)	Не создана функционирующая система для решения поставленной задачи и не продемонстрирована невозможность решения поставленной задачи современными методами веб-поиска и ранжирования.

7.2.2. Итоговый контроль по дисциплине

Проверка качества освоения дисциплины производится в форме устного экзамена.

УСТНЫЙ ЭКЗАМЕН

Устный экзамен проводится в форме ответов на вопросы экзаменационного билета. Экзаменационный билет содержит два вопроса из перечня вопросов к экзамену.

На подготовку ответа выделяется 2,5 часа.

Примерный перечень вопросов к экзамену:

1. Меры качества. A\B тестирование. Интерливинг.
2. Обработка текста: токенизация, нормализация.
3. Индексация.
4. tf-idf, BM25.
5. Латентно-семантический анализ. Тематические модели.
6. Ранжирование на основе ссылок. PageRank, HITS.
7. Point-wise алгоритмы машинного обучения для ранжирования.
8. Pair-wise алгоритмы машинного обучения для ранжирования.
9. List-wise алгоритмы машинного обучения для ранжирования.
10. Распределенный поиск.
11. Агрегированный поиск.
12. Кликовые модели. Каскадные модели.
13. Кликовые модели. Позиционные модели.
14. Кликовые модели. Нейронные модели.

Критерии оценивания и шкала оценки устного экзамена

Оценка	Критерии выставления оценки
«Отлично» (8-10)	Дан развернутый ответ на поставленный вопрос. Материал изложен последовательно. Имеются логичные и аргументированные выводы.
«Хорошо» (6-7)	Дан развернутый ответ на поставленный вопрос. Материал изложен непоследовательно. Имеются логичные и аргументированные выводы.
«Удовлетворительно» (4-5)	Ответ на вопрос не является полным. Материал изложен непоследовательно. Отсутствуют выводы.
«Неудовлетворительно» (0-3)	Ответ на вопрос является неверным. Материал изложен непоследовательно. Отсутствуют выводы.

7.3 Порядок формирования оценок по дисциплине

Преподаватель учитывает оценку за текущий контроль (домашние задания).

Результирующая оценка за дисциплину рассчитывается следующим образом:

$$О_{Результирующая} = 0{,}5 * О_{курсовая} + 0{,}5 * О_{экзамен}$$

Действует следующий способ округления: при значениях от 0,1 до 0,4 оценка округляется в меньшую сторону, от 0,5 до 0,9 – в большую.

На экзамене студенту не предоставляется возможность получить дополнительный балл для компенсации оценки за текущий контроль.

8. Образовательные технологии

Основными образовательными технологиями являются: интерактивные лекции, работа в группах на практических занятиях.

9. Учебно-методическое и информационное обеспечение дисциплины

9.1 Основная литература

1. Search Engines for Children: Search User Interfaces and Information-Seeking Behaviour / Gossen, Tatiana; Schindler. Springer Fachmedien Wiesbaden. 2015
2. Modern Algebra of Information Retrieval (The information retrieval series) / Dominich, Sándor; Gerstner. Springer Berlin Heidelberg. 2008

9.2 Дополнительная литература

1. Advances in Information Retrieval / Pasi; Hofmann. Springer International Publishing. 2018
2. An Introduction to Search Engines and Web Navigation, Second Edition / Mark Levene. John Wiley & Sons Incorporated. 2010
3. Information Retrieval Technology / Hwee Tou Ng; Mun-Kew Leong; Min-Yen Kan; Donghong Ji. Springer Verlag.2006

10. Рекомендации для самостоятельной работы студентов

Самостоятельная работа может рассматриваться как организационная форма обучения - система педагогических условий, обеспечивающих управление учебной деятельностью по освоению знаний и умений в области учебной деятельности без посторонней помощи. Студенту нужно четко понимать, что самостоятельная работа – не просто обязательное, а необходимое условие для получения знаний по дисциплине и развитию компетенций, необходимых в будущей профессиональной деятельности.

Самостоятельная работа проводится с целью:

- систематизации и закрепления полученных на лекциях теоретических знаний;
- углубления и расширения теоретических знаний;
- формирования умений использовать нормативную, правовую, справочную документацию и специальную литературу;
- развития познавательных способностей и активности студентов: творческой инициативы, самостоятельности, ответственности и организованности;
- формирования самостоятельности мышления, способностей к саморазвитию, самосовершенствованию и самореализации;
- формирования практических (общеучебных и профессиональных) умений и навыков;
- развития исследовательских умений;
- получения навыков эффективной самостоятельной профессиональной (практической и научно-теоретической) деятельности.

В учебном процессе выделяют два вида самостоятельной работы:

- аудиторная;
- внеаудиторная.

Аудиторная самостоятельная работа по дисциплине выполняется на учебных занятиях под непосредственным руководством преподавателя и по его заданию.

Внеаудиторная самостоятельная работа - планируемая учебная работа студентов, выполняемая во внеаудиторное время по заданию и при методическом руководстве преподавателя, но без его непосредственного участия.

Самостоятельная работа, не предусмотренная программой учебной дисциплины, раскрывающей и конкретизирующей ее содержание, осуществляется студентом инициативно, с целью реализации собственных учебных и научных интересов.

Для более эффективного выполнения самостоятельной работы по дисциплине преподаватель рекомендует источники для работы, характеризует наиболее рациональную методику самостоятельной работы, демонстрирует ранее выполненные студентами работы и т. п.

Виды заданий для внеаудиторной самостоятельной работы, их содержание и характер могут иметь вариативный и дифференцированный характер, учитывать индивидуальные особенности студента.

Самостоятельная работа может осуществляться индивидуально или группами студентов online и на занятиях в зависимости от цели, объема, конкретной тематики самостоятельной работы, уровня сложности.

Контроль результатов внеаудиторной самостоятельной работы осуществляется в пределах времени, отведенного на обязательные учебные занятия по дисциплине на практических занятиях.

11. Материально-техническое обеспечение дисциплины и информационные технологии, используемые при осуществлении образовательного процесса по дисциплине, включая перечень программного обеспечения информационных справочных систем (при необходимости).

Для проведения всех занятий используется проектор и компьютер для проекции слайдов. Для самостоятельной работы необходимо следующее ПО: GNU C++, Oracle Java, Python.

12. Особенности организации обучения для лиц с ограниченными возможностями здоровья

В случае необходимости, обучающимся из числа лиц с ограниченными возможностями здоровья (по заявлению обучающегося) могут предлагаться следующих варианты восприятия учебной информации с учетом их индивидуальных психофизических особенностей, в том числе с применением электронного обучения и дистанционных технологий:

1) *для лиц с нарушениями зрения:* в печатной форме увеличенным шрифтом; в форме электронного документа; в форме аудиофайла (перевод учебных материалов в аудиоформат); индивидуальные консультации с привлечением тифлосурдопереводчика; индивидуальные задания и консультации.

2) *для лиц с нарушениями слуха*: в печатной форме; в форме электронного документа; видеоматериалы с субтитрами; индивидуальные консультации с привлечением сурдопереводчика; индивидуальные задания и консультации.

3) *для лиц с нарушениями опорно-двигательного аппарата*: в печатной форме; в форме электронного документа; в форме аудиофайла; индивидуальные задания и консультации.