**Санкт-Петербургский государственный университет**

**Р А Б О Ч А Я П Р О Г Р А М М А**

**УЧЕБНОЙ ДИСЦИПЛИНЫ**

Основы информационного поиска

Fundamentals of Information Retrieval

**Язык(и) обучения**

русский

Трудоемкость в зачетных единицах: 3

Регистрационный номер рабочей программы: 045769

Санкт-Петербург

2020

**Раздел 1. Характеристики учебных занятий**

**1.1. Цели и задачи учебных занятий**

Цель курса — ознакомление обучающихся со структурой и базовыми принципами работы современных поисковых систем.

Идея курса в том, чтобы ознакомить обучающихся с реалиями работы информационно-поисковой системы общего назначения. С теми задачами, которые стоят перед разработчиками подобной системы, с методами, которые применяются для их решения, и принципами на основе которых построены механизмы принятия решений.

Подобные знания позволят слушателям работать с поисковой системой не по принципу черного ящика, а с пониманием происходящего в рамках четырех базовых ролей: пользователь поисковой системы, web-мастер, рекламодатель и разработчик. В курсе рассмотрены базовые понятия информационного поиска (инвертированный индекс, текстовая релевантность, анализ веб-графа) с точки зрения их реализации при построении поисковой системы большого размера. Также приводятся удачные примеры использования прикладной математики в рамках решения практических задач, встающих перед разработчиками поисковой системы.

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

Ожидается, что слушатели знакомы с базовыми понятиями алгебры и аналитической геометрии (операции с матрицами, векторное пространство), статистики (распределение случайной величины), умеют программировать.

**1.3. Перечень результатов обучения (learning outcomes)**

**1.3.1. Компетенции, развиваемые дисциплиной**

Компетенции, которые формируются у обучающихся в процессе изучения дисциплины:

* ПКП-4 Способен применять современные информационные технологии при проектировании, реализации, оценке качества и анализа эффективности программного обеспечения для решения задач в различных предметных областях
* ПКП-5 Способен использовать основные методы и средства автоматизации проектирования, реализации, испытаний и оценки качества при создании конкурентоспособного программного продукта и программных комплексов, а также способен использовать методы и средства автоматизации, связанные с сопровождением, администрированием и модернизацией программных продуктов и программных комплексов
* ПКП-6 Способен использовать знания направлений развития компьютеров с традиционной (нетрадиционной) архитектурой; современных системных программных средств: операционных систем, операционных и сетевых оболочек, сервисных программ; тенденции развития функций и архитектур проблемно-ориентированных программных систем и комплексов в профессиональной деятельности.

**1.3.2. Знания, умения и навыки, формируемые дисциплиной**

После прохождения курса обучающиеся должны:

* понимать структуру информационно-поисковой системы (ИПС) общего назначения;
* знать базовые механизмы и модели лежащие в основе различных частей ИПС;
* уметь оценивать качество ИПС;
* иметь представление о работе с поисковыми системами с точки зрения: пользователя, web-мастера, рекламодателя;

**1.4. Перечень и объём активных и интерактивных форм учебных занятий**

Активные формы учебных занятий — лекции, предполагающие обсуждение с преподавателем (4 академических часа).

**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 | | | | | | | | | | | | | | | | | | |
| Код модуля в составе дисциплины,  практики и т. п. | 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 | | | | | | | | | | | | | | | | | | |
| Форма обучения: очная | | | | | | | | | | | | | | | | | | |
| Семестр 6 | 45 |  | 2 |  |  |  |  |  | 2 |  |  |  | 30 |  | 29 |  | 4 | 3 |
|  | 1-100 |  | 1-100 |  |  |  |  |  | 1-100 |  |  |  | 1-1 |  | 1-1 |  |  |  |
| ИТОГО | 45 |  | 2 |  |  |  |  |  | 2 |  |  |  | 30 |  | 29 |  | 4 | 3 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Виды, формы и сроки текущего контроля успеваемости и промежуточной аттестации | | | | | | |
| Код модуля в составе дисциплины, практики и т.п. | Формы текущего контроля успеваемости | | Виды промежуточной аттестации | | Виды итоговой аттестации  (только для программ итоговой аттестации и дополнительных образовательных программ) | |
| Формы | Сроки | Виды | Сроки | Виды | Сроки |
| ОСНОВНАЯ ТРАЕКТОРИЯ | | | | | | |
| Форма обучения: очная | | | | | | |
| Семестр 6 |  |  | экзамен, устно, традиционная форма | по графику промежуточной аттестации |  |  |

**2.2. Структура и содержание учебных занятий**

|  |  |  |  |
| --- | --- | --- | --- |
| № п/п | Наименование темы (раздела, части) | Вид учебных занятий | Количество часов |
| I | Введение | лекции | 15 |
| по методическим материалам | 10 |
| II | Построение индексов | лекции | 15 |
| по методическим материалам | 10 |
| III | Обработка запросов | лекции | 15 |
| по методическим материалам | 10 |
|  | Промежуточная аттестация | самостоятельная работа | 29 |
| консультации | 2 |
| экзамен | 2 |
| **Итого** | | | **108** |

**2.2.1 Содержание учебных занятий**

1. Введение. Принципиальные подходы к поиску. Общая схема работы полнотекстовой поисковой системы общего назначения в сети
2. Оценка качества поиска. 3 цели оценки, вертикаль метрик. Виды пользовательской оценки: а/б тестирование, слепое тестирование, постановочные ответы. Экспертная оценка: понятие релевантности и его структура, метрика DCG, ее построение в виде Pfound и подбор ее параметров, классические метрики (множественные, MAP)
3. Текстовый индекс: понятие инвертированного индекса, обзор основных текстовых факторов (BM\*, цитаты, кучность, ML подход к текстовой релевантности), фильтрация и pruning, сжатие текстового индекса. Term-sharded подход.
4. Ссылочный индекс. Гиперссылки. Нормализация. Структура Web-графа, подходы к оценке качества страницы (эксперты, оригинальность, популярность), моделирование распространения качества страницы по web-графу (PageRank, ТИЦ, etc.). Поисковый спам.
5. Пользовательские факторы: позиционное смещение, сессионные факторы проблема обратной связи и перетекания сигнала в пользовательские факторы. Поисковый suggest. Сжатие кликового индекса.
6. Построение индексов: MR подход к обработке информации, CAP теорема и ее практические следствия на примере построения индексов. Возможная схема построения поискового индекса.
7. Обработка запросов. Сегментирование, лемматизация и гео-классификация запросов. Построение синонимов. Word2vec идея и ее критика.
8. Машинное обучение в ранжировании. Возможные постановки задачи. Известные реализации MLR (LambdaMart, MatrixNet -F)
9. Смешивание результатов разных источников. Основные принципы и метрики для оценки смешивания. Эффективный опрос источников и временной бюджет смешивания. Несколько способов организации смешивания.
10. Реклама в поисковых системах. Основные принципы построения аукциона. Вторая цена, игровая модель и VCG. Как построить рекламную кампанию.

**Раздел 3. Обеспечение учебных занятий**

**3.1. Методическое обеспечение**

**3.1.1 Методические указания по освоению дисциплины**

Успешное освоение дисциплины возможно благодаря посещению лекций и практических занятий, участию в обсуждении рассматриваемых вопросов, самостоятельной работе, включающей в себя чтение специальной литературы по разделам темы.

**3.1.2 Методическое обеспечение самостоятельной работы**

Самостоятельная работа обучающегося, как вид деятельности, стимулирующий активность, самостоятельность, познавательный интерес с целью поиска необходимой информации, приобретения знаний, использования этих знаний для решения учебных, научных и профессиональных задач, представляет собой важную составляющую учебного процесса, которой отводится не менее половины учебного времени при очной форме обучения. Время, отводимое на самостоятельную работу, должно использоваться обучающимися для наиболее полного освоения учебной дисциплины. Следовательно, организация эффективной внеаудиторной самостоятельной работы в процессе обучения требует, с одной стороны, создание условий, призванных обеспечить рациональное и планомерное управление учебной деятельностью, протекающей в отсутствие преподавателя, и тщательной подготовки целого ряда учебных пособий, снабженных методическими указаниями, с другой стороны.

При самостоятельном изучении теоретического материала, выполнении практических заданий и во время подготовки доклада целесообразно использовать рекомендованную основную и дополнительную литературу.

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

**3.1.3.1. Методика проведения текущего контроля успеваемости и промежуточной аттестации**

Обучающиеся получают 4 задания, которые могут быть обзорными докладами по публикациям по тематике курса, докладами по текущему состоянию собственной исследовательской работы, или заданиями на реализацию рассматриваемых в курсе алгоритмов.

Экзамен проводится в письменной и устной форме. Обучающиеся получают задания в форме билетов из трех вопросов по курсу или другое задание, и должны его выполнить письменно в течение 30 минут. По окончании выполнения задания отправляются на проверку. Если задание не выполнено в установленный срок, оценка может быть снижена. В процессе выполнения задания допускается использование ресурсов из сети интернет. После проверки преподаватель в устной части экзамена задает вопросы по выполненному заданию, а также может задавать дополнительные вопросы, не требующие длительной подготовки, по всему материалу курса.

По желанию преподавателя на экзамен допустимо приглашать других преподавателей с квалификацией не ниже изложенной в п. 3.2.1 как для независимого оценивания ответов обучающихся, так и для коллегиального. В последнем случае оценка за экзамен ставится на основании голосования простого большинства. В спорных ситуациях преподаватель, ведущий дисциплину, имеет право принятия окончательного решения.

**3.1.3.2. Критерии оценивания итогового процента освоения дисциплины**

За каждое задание ставится не более 10 баллов, а за экзамен до 60 баллов. Дополнительная форма – краткие опросы в конце занятия. За ответы на вопросы могут быть проставлены до 4 баллов. За задания, сданные после установленного срока, максимальная оценка 5 баллов. Оценка может быть снижена на 5 баллов за неточное или неполное выполнение.

Итоговая оценка рассчитывается из суммы всех полученных баллов согласно таблице:

|  |  |  |
| --- | --- | --- |
| Итоговый процент  выполнения, % | Оценка СПбГУ при  проведении экзамена | Оценка  ECTS |
| 90-100 | отлично | A |
| 80-89 | хорошо | B |
| 70-79 | хорошо | C |
| 61-69 | удовлетворительно | D |
| 50-60 | удовлетворительно | E |
| менее 50 | неудовлетворительно | F |

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

**3.1.4.1. Формируемые дисциплиной компетенции**

**ПКП-4** Способен применять современные информационные технологии при проектировании, реализации, оценке качества и анализа эффективности программного обеспечения для решения задач в различных предметных областях.

□ Формируется дисциплиной.

✓ Развивается дисциплиной.

□ Полностью сформирована по результатам освоения дисциплины.

Шкала оценивания: линейная, определяется долей успешно выполненных заданий, проверяющих данную компетенцию.

**ПКП-5** Способен использовать основные методы и средства автоматизации проектирования, реализации, испытаний и оценки качества при создании конкурентоспособного программного продукта и программных комплексов, а также способен использовать методы и средства автоматизации, связанные с сопровождением, администрированием и модернизацией программных продуктов и программных комплексов.

□ Формируется дисциплиной.

✓ Развивается дисциплиной.

□ Полностью сформирована по результатам освоения дисциплины.

Шкала оценивания: линейная, определяется долей успешно выполненных заданий, проверяющих данную компетенцию.

**ПКП-6** Способен использовать знания направлений развития компьютеров с традиционной (нетрадиционной) архитектурой; современных системных программных средств: операционных систем, операционных и сетевых оболочек, сервисных программ; тенденции развития функций и архитектур проблемно-ориентированных программных систем и комплексов в профессиональной деятельности.

□ Формируется дисциплиной.

✓ Развивается дисциплиной.

□ Полностью сформирована по результатам освоения дисциплины.

Шкала оценивания: линейная, определяется долей успешно выполненных заданий, проверяющих данную компетенцию.

**3.1.4.2. Контрольно-измерительные материалы (примеры)**

*Примеры заданий.*

* «Реализация построения инвертированного индекса средствами Java»
* «Сравнительное тестирование алгоритмов лемматизации на примере документов wikipedia»
* «Доклад о способах сжатия инвертированного индекса».

*Вопросы к экзамену*

1. Принципиальные подходы в поиске, различие Yahoo vs. Altavista

2. Узкие места в производительности ИПС. Схема прохождения поискового запроса.

3. Схема построения поискового индекса.

4. Виды логов пользователей. Способы объединения. Идентификация пользователей.

5. Критерии оценки поисковой системы. Вертикаль метрик.

6. Виды тестирования, требования к метрикам каждого вида.

7. А/Б тестирование. Проверка однородности выборки.

8. Team Draft Interleaving. Обосновать почему результаты равномерны.

9. Слепое тестирование. Слепое тестирование сложных объектов.

10. Кренфилдский подход. В чем заключается. Бинарные и 5-бальные оценки релевантности.

11. Согласованность оценок и моделирование роста качества оценок с ростом перекрытия.

12. Многомерная модель релевантности.

13. DCG метрики. Нормализация и ее критика. Pfound

14. Нахождение оптимальных параметров Pfound

15. Классические метрики оценки: P, R, MAP, MRR

16. Инвертированный индекс и его использование для исполнения полнотекстовых запросов

17. Фильтрация и pruning

18. Методы сжатия инвертированного индекса от RLE до p4

19. Вычисление BM\* и полных цитат на инвертированном индексе

20. Оконные факторы на примере кучности

21. Постановка задачи ML для текстовой релевантности

22. Способы шардирования инвертированного индекса

23. Постановка задачи Term Sharded поиска

24. Обновляемые в реальном времени текстовые индексы

25. Гиперссылки: структура, проблемы и нормализация

26. Структура Web-графа

27. Алгоритмы определения дублирования контента на примере шинглов

28. Зеркала, определение зеркалов и подход к автоматическому структурированию ресурсов

29. Оценка оригинальности страницы на основе шинглов

30. PageRank модель. PageRank, ТИЦ.

31. Использование PR-like моделей для распространения понятия качества

32. Поисковый спам: простейшие виды, методы борьбы. Виды санкций на примере Яндекс.

33. Сессии запросов пользователей. Постановка задачи сессионного сегментирования.

34. Поисковый suggest: оценка качества, подходы к построению

35. Позиционное смещение и методы борьбы с ним

36. Методы сбора информации о поведении пользователя на SERP

37. Механизм вычисления пользовательских факторов на примере CTR, pSkip, etc.

38. Обратная связь через пользовательские факторы, перетекание сигнала в пользовательские факторы

39. CAP теорема и ее следствия для поисковой системы

40. MapReduce парадигма. Разложение построения инвертированного индекса в MR.

41. Виды отказов в распределенных системах

42. Схема построения MR обработки на примере Hadoop стека

43. Реализация надежной систематической обработки информации в MR системах

44. Язык запросов поисковой системы на примере языка запросов Яндекс

45. Лемматизация слов на примере Porter Stemming. Обзор существующих решений для русского языка. Риски, связанные с лемматизацией.

46. Синонимы. Постановка задачи построения синонимов на основе коллекции документов.

47. Сегментация запросов. Простейшая сегментация на основе mutual information

48. Гео-классификация запросов. Возможные подходы к автоматизации.

49. Word2Vec. Построение пространства на основе LSA.

50. Современные постановки Word2Vec. Понятие семантического словаря на примере идей проф. Тузова.

51. Постановки задачи обучения ранжированию.

52. Подход к прямой оптимизации DCG. LambdaMart

53. MatrixNet -F

54. Смешивание различных аспектов релевантности в единой формуле

55. Понятие вертикального поиска и постановка задачи смешивания вертикалей

56. Оценка качества смешивания

57. Временной бюджет исполнения запроса и его распределение по фазам поиска

58. Подход к организации смешивания на основе классификации аспектов запроса

59. Понятие контекстной рекламы и ее виды

60. Способы задания рекламной кампании на примере Я.Директ

61. Аукцион второй цены

62. Теоретико-игровая модель аукциона, равновесный аукцион на примере VCG

63. Способы анализа поискового трафика и оптимизация рекламной кампании

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

Для оценки обучающимися содержания и качества учебного процесса может применяться анкетирование в соответствии с методикой и графиком, утвержденными в установленном порядке.

**3.2. Кадровое обеспечение**

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

К преподаванию дисциплины могут быть допущены преподаватели, имеющие диплом о высшем образовании по соответствующему направлению.

**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**

Стандартная аудитория для проведения занятий.

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

Стандартное аудиторное оборудование.

**3.3.3 Характеристики специализированного оборудования**

Не требуется.

**3.3.4 Характеристики специализированного программного обеспечения**

Не требуется.

**3.3.5 Перечень и объёмы требуемых расходных материалов**

Мел, маркеры для досок, губка.

**3.4. Информационное обеспечение**

**3.4.1 Список обязательной литературы**

1. Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.

**3.4.2 Список дополнительной литературы**

1. Ricardo Baeza-Yates, Berthier Ribeiro-Neto, Modern Information Retrieval (the concepts and technology behind search)

**3.4.3 Перечень иных информационных источников**

Нет

**Раздел 4. Разработчики программы**

Кураленок Игорь Евгеньевич. ikuralenok@gmail.com