АННОТАЦИЯ
рабочей программы дисциплины

Информационный поиск и обработка естественного языка

1. Общая трудоёмкость

Трудоёмкость дисциплины составляет 6 зачётных единиц (216 часов), из них 18 часов лекционных занятий, 36 часов практических занятий.

2. Место дисциплины в структуре образовательной программы

Дисциплина относится к модулю профессиональных дисциплин, формируемому участниками образовательных отношений, части образовательной программы, формируемой участниками образовательных отношений.

Для изучения данной дисциплины необходимы знания, умения и навыки, формируемые предшествующими элементами образовательной программы: Методы машинного обучения; Современные проблемы и методы прикладной информатики.

Результаты обучения, формируемые данной дисциплиной, потребуются при освоении следующих элементов образовательной программы: выполнение и защита выпускной квалификационной работы; производственная практика, преддипломная практика.

3. Цель изучения дисциплины

Ознакомление обучающихся с методами обработки текста на естественном языке, а также методами обработки слабоструктурированных данных и извлечения информации. Предполагается знакомство с методами извлечения отношений, анализа тональности, аннотирования и кластеризации текстов, а также с существующими программными реализациями этих методов.

4. Содержание дисциплины

Модуль 1. Информационный поиск

Тема 1. Введение в информационный поиск

Введение в информационный поиск: основные понятия, практическая значимость, задачи. Булев поиск: пример информационного поиска, первая попытка создать инвертированный индекс, обработка булевых запросов. Лексикон и список словопозиций: схематизация документа и декодирование последовательности символов, определение лексикона.

Тема 2. Словари и нечёткий поиск

Словари и нечёткий поиск: поисковые структуры для словарей, запросы с джокером, исправление опечаток, фонетические исправления. Построение индекса: блочное индексирование, основанное на сортировке, однопроходное индексирование в оперативной памяти. Сжатие индекса: статистические характеристики терминов в информационном поиске, сжатие словаря, сжатие инвертированного индекса.

Тема 3. Ранжирование, взвешивание терминов и модель векторного пространства

Ранжирование, взвешивание терминов и модель векторного пространства: параметрические и зонные индексы, частота термина и взвешивание. Ранжирование в полнофункциональной поисковой системе: эффективное ранжирование, компоненты информационно-поисковой системы.

Тема 4. Оценка информационного поиска

Оценка информационного поиска: оценка информационно-поисковой системы, стандартные текстовые коллекции, оценка неранжированных результатов поиска, оценка ранжированных результатов поиска, оценка релевантности, качество системы и её полезность для пользователя, снипеты. Обратная связь по релевантности и расширение запроса: обратная связь по релевантности и псевдорелевантность, глобальные методы для переформулирования запроса.

Модуль 2. Обработка естественного языка

Тема 5. Введение в обработку естественно- языковых текстов

Лингвистика как наука о языке. Представление об уровнях представления языка – фонетика, морфология, синтаксис, семантика. Лингвистика и прагматика. Лингвистическое моделирование. Действующие модели языка. Теория «Смысл – Текст» как фундамент для построения систем автоматической обработки текста.

Тема 6. Методы обработки естественных языков

Анализ и синтез текста. Морфологический и синтаксический анализ. Парсинг. Различные подходы к синтаксическому анализу: анализ «сверху вниз» и «снизу вверх». Языковая неоднозначность как принципиальное свойство языка и методы ее разрешения при автоматической обработке текста. Интерактивное разрешение лексической и синтаксической неоднозначности. Правиловые и статистические подходы к автоматической обработке текста. Алгоритм синтаксического анализа. Синтаксические отношения. Синтагмы. Синтаксическая структура предложения.

Тема 7. Вопросно-ответные системы

Вопросно-ответные системы: основы вопросно-ответной системы, архитектуры вопросно-ответной системы, установление смысла вопроса и порождение ответов. Распознавание имён людей, географических названий и других сущностей, различные подходы к распознаванию именованных сущностей

Тема 8. Программирование и проектирование систем обработки естественных языков

Задачи морфологического анализа, морфологический разбор, стемминг, лемматизация. Понятия лексемы, словоформы, леммы, морфемы, псевдо-основы и псевдо-окончания. Грамматические категории. Словоизменительная парадигма. Морфотактика. Структура данных морфологического словаря, лексикона. Грамматические модели русского языка в контексте автоматической обработки. Минимальное расстояние редактирования. Алгоритм подсчета расстояния Левенштейна. Практика по подсчету минимального расстояния Левенштейна. Понятие статистической языковой модели. Области применения. N-грамы.

5. Дополнительная полезная информация

Дисциплина предназначена для формирования элементов следующих компетенций образовательной программы:

ПК-1. Способен адаптировать и применять методы и алгоритмы машинного обучения для решения прикладных задач в различных предметных областях.

Форма промежуточной аттестации: дифференцированный зачёт.

Наименование оценочного средства: практические работы № 1-6 (собеседование по результатам выполнения практических работ); индивидуальное проектное задание; контрольные работы.