Р А Б О Ч А Я П Р О Г Р А М М А
У Ч Е Б Н О Й Д И С Ц И П Л И Н Ы

Информационный поиск в неструктурированных данных
Information Retrieval in Unstructured Documents

Язык(и) обучения
русский

Трудоёмкость (границы трудоёмкости) в зачетных единицах: 2

Регистрационный номер рабочей программы: 001127

Санкт-Петербург
2018

Раздел 1. Характеристики учебных занятий

1.1. Цели и задачи учебных занятий

Цель дисциплины — изучение базовых задач и методов информационного поиска, обработки естественного языка, классификации и кластеризации текстовых данных.

Задачами дисциплины являются: изучение основных принципов обработки естественного языка, изучение методов предварительной обработки текстовых данных, способов хранения текстовой информации и поиска по ней, знакомство с существующими платформами полнотекстового поиска, изучение моделей представления текстов, изучение принципов и алгоритмов обработки текстовых данных, в том числе изучение алгоритмов классификации и кластеризации текстовых документов, знакомство с тематических моделированием, дистрибутивной семантикой и средствами семантического анализа документов, анализ влияния на качество работы изучаемых алгоритмов процессов предварительной обработки текстов, получение практических навыков по обработке текстовых данных в программах, в том числе с применением современных библиотек обработки информации, выработка навыков эффективного использования существующих решений в области обработки естественного языка для задач информационного поиска.

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

Линейная алгебра и теория вероятностей, алгоритмы и структуры данных, программирование, базы данных.

1.3. Перечень результатов обучения (learning outcomes)

По результатам обучения формируются:
Знание базовых методов автоматического анализа больших объемов текстовой информации (первичная обработка, хранение, поиск, кластеризация и классификация).
Умение использовать современные инструменты обработки, хранения и анализа текстов на естественном языке.
В результате прохождения курса учащийся должен *владеть* базовыми технологиями и практическими навыками, полученными при работе с реальными коллекциями уровня 20NewsGroup, Reuters-21578.
Изучение программы способствует формированию следующих *компетенций*:
ПК-1 способность применять в профессиональной деятельности современные языки программирования и языки баз данных, методологии системной инженерии, системы автоматизации проектирования, электронные библиотеки и коллекции, сетевые технологии, библиотеки и пакеты программ, современные профессиональные стандарты информационных технологий в соответствии с профилем подготовки
ПК-2 способность профессионально решать задачи производственной и технологической деятельности с учетом современных достижений науки и техники, включая: разработку алгоритмических и программных решений в области системного и прикладного программирования; разработку математических, информационных и имитационных моделей по тематике выполняемых исследований; создание информационных ресурсов глобальных сетей, образовательного контента, прикладных баз данных; разработку тестов и средств тестирования систем и средств на соответствие стандартам и исходным требованиям; разработку эргономичных человеко- машинных интерфейсов в соответствии с профилем подготовки
ПК-3 способность разрабатывать и реализовывать процессы жизненного цикла информационных систем, программного обеспечения, сервисов систем информационных технологий, а также методы и механизмы оценки и анализа функционирования средств и

систем информационных технологий; способность разработки проектной и программной документации, удовлетворяющей нормативным требованиям
КП-02.2 способность ставить задачи на создание и применение баз данных; уметь разрабатывать приложения баз данных, владеть алгоритмами и технологиями анализа данных
ОКМ-2 Готов использовать знание современных достижений науки и образования при решении образовательных и профессиональных задач
ОКМ-4 Готов самостоятельно приобретать с помощью информационных технологий и использовать в практической деятельности новые знания и умения, в том числе в новых областях

1.4. Перечень активных и интерактивных форм учебных занятий

Семинары и практические занятия, презентации выполненных проектов и работающих программ - 16 ч. Практические занятия проводятся с привлечением интерактивных методов: групповая, научная дискуссия и обсуждение материалов курса, выполнение исследовательских проектов малыми группами, представление самостоятельно выполненных проектов и коллективное обсуждение полученных результатов. Семинары частично проводятся в диалоговом режиме в активной или интерактивной форме, частично являются проблемными лекциями.

Раздел 2. Организация, структура и содержание учебных занятий

2.1. Организация учебных занятий

2.1.1. Профиль Технологии баз данных

Трудоёмкость, объёмы учебной работы и наполняемость групп обучающихся																	
Период обучения (модуль)	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																		
очная форма обучения																		
Семестр 2		12		16				4	2				28	4	6		16	2
ИТОГО		**12**		**16**				**4**	**2**				**28**	**4**	**6**		**16**	**2**

Формы текущего контроля успеваемости, виды промежуточной и итоговой аттестации

Период обучения (модуль)	Формы текущего контроля успеваемости	Виды промежуточной аттестации	Виды итоговой аттестации (только для программ итоговой аттестации и дополнительных образовательных программ)
ОСНОВНАЯ ТРАЕКТОРИЯ			
очная форма обучения			
Семестр 2	Презентация	Зачет	

2.2. Структура и содержание учебных занятий

Период обучения (модуль): **Семестр 2**

№ п/п	Наименование темы (раздела, части)	Вид учебных занятий	Количество часов
1	Методы индексирования текстовых данных, структура индекса, методика оценки качества поиска	Семинары практические занятия по методическим материалам	2 0 4
2	Платформы полнотекстового поиска, обработка текстовой коллекции	Семинары Практические занятия По методическим материалам	1 4 4
3	Модель векторного пространства, вероятностная модель информационного поиска, языковые модели для информационного поиска	Семинары Практические занятия По методическим материалам	2 0 4
4	Классификация текстов, обработка текстовой коллекции	Семинары Практические занятия По методическим материалам	2 4 4
5	Кластеризация текстов, обработка текстовой коллекции	Семинары Практические занятия По методическим материалам	2 4 4
6	Тематическое моделирование, обработка текстовой коллекции (по выбору)	Семинары Практические занятия По методическим материалам	2 2 4
7	Дистрибутивная семантика, обработка текстовой коллекции (по выбору)	Семинары Практические занятия По методическим материалам	2 2 2

8	Тезарусы и онтологии в информационном поиске	Семинары Практические занятия По методическим материалам	1 0 2

Методы индексирования текстовых данных, структура индекса, методика оценки качества поиска. Примеры информационного поиска, инвертированный индекс, обработка булевых запросов, лексикон терминов, токенизация, стемминг, лемматизация, указатели пропусков, фразовые запросы, двухсловные и координатные индексы, распределенное индексирование и другие подходы к построению индекса, методика оценки качества поиска.

Платформы полнотекстового поиска, обработка текстовой коллекции. Обзор современных систем полнотекстового поиска, более подробное знакомство с одной из систем (могут быть рекомендованы: Solr (Apache Lucene), Elasticsearch, Terrier, Sphinx выполнение тестового задания с использованием рассмотренной системы и тестовой текстовой коллекции.

Модель векторного пространства, вероятностная модель информационного поиска, языковые модели для информационного поиска. Ранжирование документов, взвешивание терминов, модель векторного пространства для ранжирования, варианты функций tf-idf, принцип вероятностного ранжирования, языковые модели для информационного поиска.

Классификация текстов, обработка текстовой коллекции. Классификация текстов, выбор признаков, оценка качества классификации, наивная байесовская классификация текстов, модель Бернулли, классификация в векторном пространстве, метод опорных векторов, классификация тестовой коллекции.

Кластеризация текстов, обработка текстовой коллекции. Задача кластеризации, плоская кластеризация и метод k-средних, иерархическая кластеризация, кластеризация тестовой коллекции.

Тематическое моделирование, обработка текстовой коллекции. Латентный семантический анализ, вероятностный латентный семантический анализ, латентное размещение Дирихле.

Дистрибутивная семантика, обработка текстовой коллекции. Нейронные языковые модели в дистрибутивной семантике

Тезарусы и онтологии в информационном поиске. Семантический поиск, онтологии, тезарусы.

Раздел 3. Обеспечение учебных занятий

3.1. Методическое обеспечение

3.1.1 Методические указания по освоению дисциплины

Данный курс состоит из двух основных частей – теоретическая часть, состоящая в изучении ряда алгоритмов информационного поиска и обработки естественного языка, классификации и кластеризации текстов, а так же практической части, в рамках которой для работы с текстами используются системы полнотекстового поика и библиотеки обработки данных. После изучения каждой группы алгоритмов в рамках теоретической

части студенты получат возможность практического применения некоторых из изученных алгоритмов. Студенты не ограничены в выборе языка программирования и платформы для разработки тестовых программ (Python и Java могут быть рекомендованы).

Самостоятельная работа студентов включает в себя изучение дополнительного материала, освоение (если не было сделано ранее) библиотек обработки данных, выполнение тестовых и небольших исследовательских проектов как в группах (по 2 человека) так и индивидуально. Языки программирования и библиотеки, содержащие готовые реализации алгоритмов выбираются студентами самостоятельно. Обязательным является понимание и знание алгоритмов, готовые реализации которых используются. Время и место самостоятельной работы выбираются студентом самостоятельно в соответствии со сроками сдачи проектов, запланированных в рамках курса (всего 4 проекта). Самостоятельная работа включает в себя повторение лекционного материала, подготовку вопросов по пройденному материалу в случае их возникновения после семинара.

Подготовка к семинару заключается в следующем: повторение материала предыдущей лекции, составление схемы, отражающей смысл и связи основных понятий изучаемого материала, изучение дополнительного материала в соответствии с рекомендациями преподавателя, подготовка дополнительных вопросов по пройденному материалу.

Подготовка к практическим занятиям и презентации выполненных проектов. Рекомендуется заранее сделать обзор по предполагаемым к использованию библиотекам обработки данных и обсудить с преподавателем выбор конкретной из них. Необходимо хорошо разобрать и освоить алгоритмы, изучаемые на семинаре, понимать как они устроенны и работают в выбранных готовых решениях (библиотеках), рекомендуется узнать тему предстоящей практики (по тематическому плану) и попробовать предварительно самостоятельно поработать с выбранными инструментами (библиотеками, системами полнотекстового поиска), включая их установку и настройку, а также подготовить список возникших вопросов.

Подготовленные списки вопросов отправляются на почту преподавателю не позднее чем за 2 дня до проведения занятия.

Зачет выставляется по результатам выполнения проектов. Должны быть зачтены 4 проекта. Защита проектов выполняется в виде презентаций в указанные сроки. В случае, если один или более проектов сдаются вне указанных сроков, для получения зачета дополнительно пишется проверочная работа, количество заданий в которой пропорционально числу не сданных своевременно проектов. Подготовка к письменной работе включает в себя разбор всего изученного в рамках курса материала, а также выполнение упражнений из указанных источников.

Рекомендации по подготовке презентации: презентация должна быть подготовлена и отрепетирована заранее, включать в себя все требующиеся пункты и занимать по времени 15 минут.

3.1.2 Методическое обеспечение самостоятельной работы

Обеспечение самостоятельной работы: ссылки на ресурсы в сети интернет — коллекция текстов (Reuters-21578, 20 Newsgroups, BY.web 2007), ссылки на учебные материалы и рекомендуемые библиотека машинного обучения (например: Scikit-learn, Weka, Mllib), консультации во время аудиторных занятий.

Самостоятельная работа выполняется в группах до 2х человек или индивидуально в соответствии с описанием задания и сроками. Комплект заданий для проектов студенты получают на первой паре, в это же время фиксируются сроки сдачи.

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

Зачет выставляется по результатам выполнения проектов. Должны быть зачтены 4 проекта. Защита проектов выполняется в виде презентаций в указанные сроки. В случае, если один или более проектов сдаются вне указанных сроков, для получения зачета дополнительно пишется проверочная работа, количество заданий в которой пропорционально числу не сданных своевременно проектов. В случае сдачи письменной работы для получения зачета должно быть выполнено не менее 70% заданий, сдача 4 проектов в этом случае остается обязательной. Для зачета выполняется презентация работающих программ (4 презентации) по темам 1) информационного поиска, 2) классификации текстов, 3) кластеризации текстов, 4) дистрибутивной семантики / тематическому моделированию (тема выбирается по согласованию с преподавателем) . Зачет выставляется в том случае, если все презентации выполнены в срок и поставленные задачи решены полностью. Оценка ‘неудовлетворительно’ выставляется во всех остальных случаях. Преподаватель имеет право предоставить информацию о задолженностях студента в аттестационную комиссию.

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

Описание требований к презентации: должны быть описаны данные (набор текстовых документов), процедура их индексирования, алгоритм, его имплементация, описание экспериментов и их результаты. Презентация должна быть заранее подготовлена и отрепетирована. Продолжительность презентации 15 минут.

Пример проектного задания:
Провести классификацию/кластеризацию выбранной тестовой коллекции.
В презентации описать предпринятые шаги, использованные настройки и влияние параметров алгоритма на качество обработки данных, возникшие проблемы и их решение. Показать как на качество обработки текстов влияют использование стемминга/лемматизации, удаление стоп слов, удаление редко частотных слов (с документной частотой 1, 2,...n), удаление слов с наибольшей документной частотой. Полученные результаты представить в виде таблиц с оценкой качества. Для оценки качества используются меры основанные на: Precision, Recall, F1-score. Сделать интерпретацию полученных результатов. В презентации должны быть описаны данные (набор текстовых документов), алгоритмы и их имплементация, эксперименты и их результаты.

Классификация: используются наивный байесовский классификатор и метод опорных векторов

Кластеризация: используются k-средних и иерархическая кластеризация (метод single link, complete link, average link)

Выполнение задание предполагает использование готовых решений и библиотек в соответствии с выбранным языком программирования. Например: Scikit-learn (http://scikit-learn.org/stable/), Weka (https://www.cs.waikato.ac.nz/ml/weka/), MLlib и другие библиотеки. Выбор библиотеки рекомендуется согласовать с преподавателем.

Тестовые коллекции
Reuters-21578
http://www.daviddlewis.com/resources/testcollections/reuters21578/
или
20 Newsgroups
http://qwone.com/~jason/20Newsgroups/

по согласованию с преподавателем может быть выбрана другая тестовая коллекция (разрешаются коллекции, используемые в исследовательском мире IR и представленные в научных публикациях)

При необходимости модификации, изменения или уточняющих вопросов по заданиям — обсуждаем эти изменения и вопросы с преподавателем.

3.1.5 Методические материалы для оценки обучающимися содержания и качества учебного процесса

Для оценки обучающимися содержания и качества учебного процесса используется анкета-отзыв установленная локальными актами СПбГУ.

3.2. Кадровое обеспечение

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

Знание программирования и баз данных

3.2.2 Обеспечение учебно-вспомогательным и (или) иным персоналом

Нет специальных требований

3.3. Материально-техническое обеспечение

3.3.1 Характеристики аудиторий (помещений, мест) для проведения занятий

Компьютерный класс с интернетом и проектором

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

Интернет

3.3.3 Характеристики специализированного оборудования

Нет специальных требований

3.3.4 Характеристики специализированного программного обеспечения

JDK, Apache Solr

3.3.5 Перечень и объёмы требуемых расходных материалов

Фломастеры для доски

3.4. Информационное обеспечение

3.4.1 Список обязательной литературы

1. Маннинг, Кристофер Д. Введение в информационный поиск : научное издание / К. Д. Маннинг, П. Рагхаван, Х. Шютце ; пер. с англ. П. И. Клюшин. - М. ; СПб. ; Киев : Вильямс, 2011. - 520 с. : ил. - (Профессионалам от профессионалов). - Библиогр.: с. 473-505. - Предм. указ.: с. 506-520. - 1000 экз.. - ISBN 978-5-8459-1623-5 (в пер.)

3.4.2 Список дополнительной литературы

1. Международная конференция по компьютерной лингвистике. http://www.dialog-21.ru/
2. V. Dobrynin. Document Clustering.

   http://www.apmath.spbu.ru/cbu/2008/dobrynin_clustering2.pdf

3.4.3 Перечень иных информационных источников

Отсутствует.

Раздел 4. Разработчики программы

Попова Светлана Владимировна, старший преподаватель, кафедра технологии программирования, svp@list.ru, +79214388077