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

Факультет компьютерных наук
Департамент программной инженерии

Рабочая программа дисциплины
Обработка текстов

для образовательной программы «Программная инженерия»
направления подготовки 09.03.04 «Программная инженерия»
уровень - бакалавр

Разработчики программы
Турдаков Д.Ю., к.ф.-м.н., turdakov@ispras.ru

Одобрена на заседании департамента программной инженерии «___»_____ 2017 г.
Руководитель департамента Авдошин С.М. _____

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

Академический руководитель образовательной программы
Шилов В.В. _____

Москва, 2017

1. Область применения и нормативные ссылки

Настоящая программа учебной дисциплины устанавливает минимальные требования к знаниям и умениям студента и определяет содержание и виды учебных занятий и отчетности.

Программа предназначена для преподавателей, ведущих данную дисциплину, учебных ассистентов и студентов образовательной программы «Программная инженерия» направления подготовки 09.03.04 «Программная инженерия», изучающих дисциплину "Обработка текстов".

Программа разработана в соответствии с образовательным стандартом Национального исследовательского университета «Высшая школа экономики» по направлению 09.03.04 «Программная инженерия».

2. Цели освоения дисциплины

Целью освоения дисциплины является получение базовых знаний в области обработки текстов на естественном языке, а также приобретение навыков решения задач, возникающих при разработке систем текстового анализа.

3. Компетенции обучающегося, формируемые в результате освоения дисциплины

В результате освоения дисциплины студент должен:

1. Знать:

- Фундаментальные понятия и идеи в области компьютерной обработки текстов, современные направления исследований в данной области и основные проблемы, возникающие при обработке текстов.

2. Уметь:

- решать задачи из области обработки текстов, проектировать системы для анализа отдельных текстовых документов и коллекций текстовых документов, применять методы статистического анализа и машинного обучения для решения прикладных задач области.

3. Иметь навыки (приобрести опыт):

- освоения большого объема информации;
- самостоятельной работы с современными технологиями и программными инструментами для обработки текстов.
- культурой разработки и реализации системного программного обеспечения современных компьютеров;

В результате освоения дисциплины студент должен обладать следующими компетенциями:

Универсальные компетенции:

- Способен решать проблемы в профессиональной деятельности на основе анализа и синтеза (УК-3)
- Способен оценивать потребность в ресурсах и планировать их использование при решении задач в профессиональной деятельности (УК-4)
- Способен работать с информацией: находить, оценивать и использовать информацию из различных источников, необходимую для решения научных и профессиональных задач (в том числе на основе системного подхода) (УК-5)
- Способен вести исследовательскую деятельность, включая анализ проблем, постановку целей и задач, выделение объекта и предмета исследования, выбор способа и методов исследования, а также оценку его качества (УК-6)

Профессиональные компетенции:

- Способен использовать методы и инструментальные средства исследования объектов профессиональной деятельности (ПК-3)
- Способен обосновать принимаемые проектные решения, осуществлять постановку и выполнение экспериментов по проверке их корректности и эффективности (ПК-4)
- Способен проектировать, конструировать и тестировать программные продукты (ПК-10)
- Способен читать, понимать и выделять главную идею прочитанного исходного кода, документации (ПК-11)
- Способен использовать различные технологии разработки программного обеспечения (ПК-16)
- Способен применять основные методы и инструменты разработки программного обеспечения (ПК-17)

4. Место дисциплины в структуре образовательной программы

Настоящая дисциплина относится к вариативной части профессионального цикла дисциплин, является дисциплиной по выбору студентов.

Изучение данной дисциплины базируется на знаниях, полученных студентами при освоении учебных дисциплин:

- «Программирование»,
- «Data Analysis»,
- «Построение и анализ алгоритмов».

5. Тематический план учебной дисциплины

№	Название раздела	Всего часов	Аудиторные часы			Самостоятельная работа
			Лекции	Семинары	Практические занятия	
1	Задачи обработки текстов.	34	2		2	30
2	Базовые инструменты обработки текстов	39	3		6	30
3	Морфологический анализ текстов	41	5		4	32
4	Синтаксический анализ текстов	48	4		6	38
5	Семантический анализ текстов	52	5		6	41
6	Прикладное применение методов анализа текстов	52	5		6	41
		266	24		30	212

6. Формы контроля знаний студентов

Тип контроля	Форма контроля	4 год			
		1 модуль	2 модуль	3 модуль	4 модуль
Текущий (месяц)	Домашнее задание	1-6 неделя	1-6 неделя		
	Контрольная рабо-		*		

	та				
Итого-вый	Экзамен			*	

Комплексное индивидуальное практическое задание (домашнее задание)

Большинство современных задач обработки текстов не имеют точного решения, а точность и полнота существующих алгоритмов зависит от данных, с которыми они работают. В качестве практического задания студентам предлагается попробовать самостоятельно решить одну из таких задач, применяя знания, полученные в теоретической части курса.

Каждый год студентам предлагается решить новую задачу. Примеры задач:

- Определение авторства текста
- Определение ключевых слов для научно-технической литературы
- Определение эмоциональной окраски текста
- Определение эмоциональной окраски сообщений микроблогов Twitter.

Студентам рассказывается постановка задачи, метод оценки решения и наиболее простое, но неточное решение. Далее студенты должны предложить собственное решение, которое показывает более точные результаты. Для проверки задания используется автоматическая система на удаленном сервере. Таким образом, студенты не знают на каких именно данных проводится тестирование и не могут искусственно подобрать точное решение.

При этом в каждом задании присутствует творческая составляющая: чем лучше студент предложит решение, тем больше он получит технических баллов, которые повлияют на итоговую оценку. Это мотивирует студентов применять полученные теоретические знания.

Срок выполнения задания – 12 недель. Каждые 4 недели выставляются технические баллы промежуточного контроля. В конце семестра выставляются итоговые технические баллы.

Критерии оценки знаний, навыков

Оценки по всем формам текущего контроля выставляются по 10-ти балльной шкале.

Порядок формирования оценок по дисциплине

Оценка по курсу состоит из оценки за выполнение домашнего задания $O_{прак.}$ (10 баллов), контрольной работы $O_{конт}$ (10 баллов) и оценки за итоговый устный экзамен (10 баллов). В диплом выставляет результирующая оценка по учебной дисциплине, которая формируется по следующей формуле:

$$O_{результ} = 0{,}4*O_{прак.} + 0{,}2*O_{конт} + 0{,}4*O_{экз}$$

7. Содержание дисциплины

№ п/п	Разделы и темы лекционных занятий	Содержание
1	Задачи обработки текстов.	Задачи обработки текста. Многозначность при обработке текста. Проблема понимания. Тест Тьюринга. Китайская комната.
2	Базовые	Регулярные выражения и конечные авто-

	инструменты обработки текстов	маты. Распознавание языка с помощью КА. Построение КА для регулярных выражений. Токенизация и сегментация. Стемминг и лемматизация. Определение границ предложений. Методы поиска словосочетаний. Общая схема. Методы поиска кандидатов. Проверка статистических гипотез. Методы классификации документов. Наивный байесовский классификатор. Логистическая регрессия. Модель максимальной энтропии.
3	Морфологический анализ текстов	Языковые модели и задача определения частей речи. Модель N-грамм. Оценка вероятности высказывания. Методы сглаживания. Оценка качества. Тренировочный и проверочный корпуса. Задача определения частей речи. Существующие подходы. Алгоритмы, основанные на правилах. Алгоритмы, основанные на трансформации. Скрытые марковские модели. Вероятность последовательности. Прямой алгоритм. Наиболее правдоподобное объяснение. Использование скрытой марковской модели для определения частей речи. Алгоритм Витерби.
4	Синтаксический анализ текстов	Контекстно-свободные грамматики и синтаксический анализ. Типы грамматик. Грамматика составляющих. Грамматика зависимостей. Категориальная грамматика. Контекстно-свободные грамматики. КС грамматики и регулярные языки. Банк деревьев. синтаксический разбор. Разбор сверху вниз и снизу вверх. Алгоритм Кока-Янгера-Касами (CKY parsing). Эквивалентность КС грамматик. Группировка (chunking) Статистические методы синтаксического анализа. Стохастические контекстно-свободные грамматики. Разрешение синтаксической многозначности. Моделирование языка. Обучение стохастических КС грамматик. Вероятностная версия алгоритма Кока-Янгера-Касами. Оценка качества. Проблемы стохастический КС грамматик. Алгоритм Коллинза.
5	Семантический анализ текстов	Лексическая семантика. WordNet. Значения слов. Разрешение лексической многозначности. Алгоритмы классификации. Самонастройка. Методы основанные на словарях и тезаурусах. Варианты алгоритма Леска. Методы оценки качества. Семантическая близость слов. Подходы на основе тезаурусов. Подходы на

		основе статистик. Методы оценки качества.
6	Прикладное применение методов анализа текстов	Информационный поиск. Ранжирование документов. Векторная модель. Взвешивание терминов. Индексирование. Инвертированный индекс. Запросы с джокером. Исправление опечаток. Вопросно-ответные системы: общая архитектура. Обработка запроса. Извлечение фрагментов текста. Обработка ответа. Автоматическое реферирование: общая архитектура. Машинный перевод. Классические подходы. Статистический машинный перевод. Модель зашумленного канала. Модель перевода на основе фраз. Выравнивание фраз. Декодирование. Выравнивание слов. Модель IBM Model 1. Тренировка моделей выравнивания. Тематическое моделирование. Вероятностная латентная семантическая модель. Латентное размещение Дирихле. Робастные модели.

8. Оценочные средства для текущего контроля и аттестации студента

Перечень вопросов для экзамена

1. Задачи обработки текста. Многозначность при обработке текста. Проблема понимания. Тест Тьюринга. Китайская комната
2. Регулярные выражения
3. Конечные автоматы, распознавание языка с помощью КА
4. Регулярные языки и конечные автоматы. Построение КА для регулярных выражений
5. Методы поиска словосочетаний. Общая схема. Методы поиска кандидатов
6. Методы поиска словосочетаний. Проверка статистических гипотез
7. Модель N-грамм. Оценка вероятности высказывания
8. Модель N-грамм. Сглаживание (Лапласа и Откат)
9. Модель N-грамм. Оценка качества. Тренировочный и проверочный корпуса
10. Задача определения частей речи. Существующие подходы. Алгоритмы, основанные на правилах. Алгоритмы, основанные на трансформации.
11. Использование скрытой марковской модели для определения частей речи.
12. Скрытые марковские модели. Вероятность последовательности. Прямой алгоритм
13. Скрытые марковские модели. Наиболее правдоподобное объяснение. Алгоритм Витерби
14. Модели классификации. Наивный байесовский классификатор
15. Модели классификации. Логистическая регрессия
16. Модели классификации. Модель максимальной энтропии
17. Модели классификации. Марковская модель максимальной энтропии
18. Типы грамматик. Грамматика составляющих. Грамматика зависимостей. Категориальная грамматика
19. Контекстно-свободные грамматики. КС грамматики и регулярные языки. Банк деревьев.
20. Синтаксический разбор. Разбор сверху вниз и снизу вверх

21. Синтаксический разбор. Алгоритм Кока-Янгера-Касами (CKY parsing). Эквивалентность КС грамматик
22. Синтаксический разбор. Группировка (chunking)
23. Стохастические контекстно-свободные грамматики. Разрешение синтаксической многозначности
24. Моделирование языка. Обучение стохастических КС грамматик
25. Вероятностная версия алгоритма Кока-Янгера-Касами. Оценка качества
26. Проблемы стохастический КС грамматик. Алгоритм Коллинза. Оценка качества
27. Лексическая семантика. WordNet. Значения слов
28. Разрешение лексической многозначности. Алгоритмы классификации. Самонастройка. Методы оценки качества
29. Разрешение лексической многозначности. Методы основанные на словарях и тезаурусах. Варианты алгоритма Леска. Методы оценки качества
30. Семантическая близость слов. Подходы на основе тезаурусов. Методы оценки качества
31. Семантическая близость слов. Подходы на основе статистик. Методы оценки качества
32. Информационный поиск. Ранжирование документов. Векторная модель. Взвешивание терминов. TF-IDF
33. Информационный поиск. Индексирование. Инвертированный индекс. Запросы с джокером. Исправление опечаток.
34. Вопросно-ответные системы. Общая архитектура. Обработка запроса
35. Вопросно-ответные системы. Общая архитектура. Извлечение фрагментов текста
36. Вопросно-ответные системы. Общая архитектура. Обработка ответа
37. Автоматическое реферирование. Общая архитектура
38. Машинный перевод. Классические подходы
39. Статистический машинный перевод. Модель зашумленного канала. Модель перевода на основе фраз. Выравнивание фраз. Декодирование
40. Статистический машинный перевод. Выравнивание слов. Модель IBM Model 1
41. Статистический машинный перевод. Выравнивание слов. Тренировка моделей выравнивания
42. Статистический машинный перевод. Методы оценки качества. BLUE

9. Учебно-методическое и информационное обеспечение дисциплины

9.1. Базовый учебник

Daniel Jurafsky and James H. Martin. 2008. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Second Edition. Prentice Hall.

9.2. Основная литература

- Christopher D. Manning and Hinrich Schütze. 1999. Foundations of Statistical Natural Language Processing. MIT Press.
- Steven Bird, Ewan Klein, and Edward Loper. Natural Language Processing with Python. O'Reilly Media, 2009 (http://www.nltk.org/book)

9.3. Дополнительная литература

- Стюарт Рассел, Питер Норвиг, ”Искусственный интеллект: современный подход”. Вильямс, 2015 г.

- Кристофер Д. Маннинг, Прабхакар Рагхаван, Хайнрих Шютце, “Введение в информационный поиск”. Вильямс, 2011 г.
- Тоби Сегаран, “Программируем коллективный разум”, Символ-Плюс, 2008