

**Федеральное государственное автономное образовательное
учреждение высшего образования
«Московский физико-технический институт
(национальный исследовательский университет)»**

УТВЕРЖДЕНО

**Директор физтех-школы
прикладной математики и
информатики**

А.М. Райгородский

	Рабочая программа дисциплины (модуля)
по дисциплине:	Автоматическая обработка естественного языка
по направлению:	Прикладная математика и информатика
профиль подготовки:	А1360: Передовые методы искусственного интеллекта Физтех-школа Прикладной Математики и Информатики кафедра алгоритмов и технологий программирования
курс:	4
квалификация:	бакалавр

Семестр, формы промежуточной аттестации: 8 (весенний) - Дифференцированный зачет

Аудиторных часов: 30 всего, в том числе:

лекции: 15 час.

семинары: 15 час.

лабораторные занятия: 0 час.

Самостоятельная работа: 15 час.

Всего часов: 45, всего зач. ед.: 1

Программу составил: Е.Л. Артемова, канд. техн. наук, доцент

Программа обсуждена на заседании кафедры алгоритмов и технологий программирования 12.02.2024

Аннотация

Курс «Автоматическая обработка текстов» является вводным в проблематику компьютерной лингвистики и построения программных систем для обработки текстов на естественном языке. Изучаются основные методы автоматической обработки текста (АОТ), а также виды необходимых для этого лингвистических ресурсов. Обзорно рассматриваются современные приложения в области АОТ и принципы их построения. Теоретический материал курса, дополняется практическими занятиями по изучению соответствующих интернет-ресурсов и прикладного программного обеспечения, а также домашними заданиями по их применению.

1. Цели и задачи

Цель дисциплины

Изучение современных алгоритмов интеллектуального анализа и обработки изображений.

Задачи дисциплины

- изучение моделей формирования, представления и искажения изображений;
- освоение математического аппарата обработки изображений;
- освоение основных алгоритмов цифровой обработки, восстановления, анализа, классификации и распознавания изображений.

2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

Код и наименование компетенции	Индикаторы достижения компетенции
ОПК-2 Способен использовать современные информационные технологии и программные средства при решении задач профессиональной деятельности, соблюдая требования информационной безопасности	ОПК-2.1 Способен применять современные вычислительную технику и сервисы сети Интернет в области (сфере) профессиональной деятельности
	ОПК-2.2 Знает и умеет применять численные математические методы и прикладное программное обеспечение для решения научных задач в профессиональной области
	ОПК-2.3 Знает основные требования информационной безопасности
ПК-2 Способен самостоятельно или в качестве члена (руководителя) малого коллектива организовывать и проводить научные исследования и их апробацию	ПК-2.1 Знает принципы построения научной работы, методы сбора и анализа полученного материала, способы аргументации
	ПК-2.2 Способен планировать и проводить научные исследования самостоятельно или в качестве члена (руководителя) малого научного коллектива
	ПК-2.3 Способен проводить апробацию результатов научно-исследовательской работы посредством публикации научных статей и участия в конференциях

3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны

знать:

- постановку задач морфологического, синтаксического анализа;
- методы решения этих задач.

уметь:

- формулировать задачи классификации текстов, предложений или их элементов для выделения структурированной информации;
- реализовывать подходящий алгоритм классификации текстов;
- решать задачи выделения ключевых слов и определения тональности.

владеть:

- основными программными системами для выделения скрытых тем и снижения размерности векторных моделей.

4. Содержание дисциплины (модуля), структурированное по темам (разделам) с указанием отведенного на них количества академических часов и видов учебных занятий

4.1. Разделы дисциплины (модуля) и трудоемкости по видам учебных занятий

№	Тема (раздел) дисциплины	Трудоемкость по видам учебных занятий, включая самостоятельную работу, час.			
		Лекции	Семинары	Лаборат. работы	Самост. работа
1	Введение в обработку текстов	1	1		1
2	Методы сбора и хранения данных	1	1		1
3	Частотный анализ текстов	1	1		1
4	Морфологический анализ и разрешение неоднозначности	1	1		1
5	Синтаксический анализ. Универсальные зависимости	1	1		1
6	Выделение ключевых слов и словосочетаний	1	1		1
7	Векторная модель текста и слова, методы снижения размерности	1	1		1
8	Классификация текстов	1	1		1
9	Языковые модели	1	1		1
10	Классификация последовательностей	1	1		1
11	Суммаризация текстов, вопросно-ответные системы	1	1		1
12	Исправление опечаток	1	1		1
13	Обработка речи, речевые технологии	1	1		1
14	Информационный поиск	1	1		1
15	Мультимодальная обработка текстов	1	1		1
Итого часов		15	15		15
Подготовка к экзамену		0 час.			
Общая трудоёмкость		45 час., 1 зач.ед.			

4.2. Содержание дисциплины (модуля), структурированное по темам (разделам)

Семестр: 8 (Весенний)

1. Введение в обработку текстов

Основные задачи обработки и анализа текстов. Актуальность обработки и анализа текстов. Краткий исторический экскурс по обработке и анализу текстов. Обзор существующих систем обработки и анализа текстов. Классификация систем обработки и анализа текстов.

2. Методы сбора и хранения данных

Форматы данных, способы хранения, принципы работы интернета. Краулинг. Regexp. Unicode.

3. Частотный анализ текстов

Модель мешка слов. Закон Ципфа. Закон Хипса. Векторное представление текстов. Релевантность в векторной модели. Расширения модели мешка слов. Реализация модели мешка слов в библиотеках Gensim и NLTK.

4. Морфологический анализ и разрешение неоднозначности

Задача морфологического анализа. Типы языков. Алгоритмы морфологического разбора. Морфологическая разметка. Омонимия и неоднозначность. Алгоритм разрешения омонимии. Скрытые Марковские модели. Декодирование в скрытых Марковских моделях.

5. Синтаксический анализ. Универсальные зависимости

Задача синтаксического разбора предложений. Модель составляющих. Вероятностные контекстно-свободные грамматики. Модель зависимостей. Универсальные зависимости. Парсинг зависимостей. Архитектура SyntaxNet.

6. Выделение ключевых слов и словосочетаний

Лексический анализ. Словари и тезаурусы. Поиск синонимов. Частотные методы выделения ключевых слов и словосочетаний. Метрики совместной встречаемости. Выделение ключевых словосочетаний по морфологическим шаблонам. Выделение ключевых словосочетаний по синтаксическим шаблонам. Алгоритмы RAKE и TextRank. Программные средства для выделения ключевых слов: NLTK, Томита-парсер.

7. Векторная модель текста и слова, методы снижения размерности

Векторная модель документа, векторная модель слова. Поиск похожих текстов. Косинусная мера близости. Методы снижения размерности в векторной модели документа: сингулярное разложение, латентный семантический анализ. Связь с моделями скрытых тем. Латентное размещение Дирихле (LDA). Параметры модели. Выбор числа скрытых тем. Расширения модели LDA. Дистрибутивная семантика, векторная модель слова. Построение матрицы PPMI. Поиск близких слов по значению. Снижение размерности и факторизация матрицы PPMI. Эмбединги: word2vec, GloVe, AdaGram. Обучение моделей word2vec. Отрицательное сэмплирование.

8. Классификация текстов

Задачи классификации текстов и предложений по теме, тональности и жанру. Метод наивного Байеса, метод максимальной энтропии. Сверточные нейронные сети. Архитектура FastText.

9. Языковые модели

Счетные языковые модели. Проблема нулевых вероятностей. Преобразование Лапласа, преобразование Гуд-Тьюринга. Вероятностные нейронные языковые модели. Генерация текстов. Рекуррентные нейронные сети.

10. Классификация последовательностей

Задача классификации последовательностей. Частеречная разметка, определение семантических ролей, извлечение именованных сущностей. IOB разметка, IOBES разметка. Условные случайные поля.

11. Суммаризация текстов, вопросно-ответные системы

Абстрактивная и генеративная суммаризация текстов. Алгоритм TextRank. Вопросно-ответные системы. Архитектура энкодера-декодера для вопросно-ответных систем и чат-ботов.

12. Исправление опечаток

Модель зашумленного канала. Исправление опечаток по правилам. Редакционное расстояние.

13. Обработка речи, речевые технологии

Распознавание речи. Генерация речи.

14. Информационный поиск

Понятие релевантности. Использование векторной модели в задаче поиска. Косинусная мера релевантности. Использование языковой модели в задаче поиска. Обучение ранжированию. A|B - тестирование.

15. Мультиязычная обработка текстов

Связь обработки текстов с обработкой изображений. Генерация изображения по тексту. Поиск изображения по описанию.

5. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине (модулю)

Учебная аудитория, оснащенная мультимедийным оборудованием (проектор или плазменная панель), доской.

6. Перечень рекомендуемой литературы

Основная литература

Литература кафедры:

Плас Дж. Python для сложных задач. Наука о данных и машинное обучение. - СПб.: Питер, 2017. - 576 с.

Николаев И.С., Митренина О.В., Ландо Т.М. Прикладная и компьютерная лингвистика. - М.: URSS, 2017. - 320 с.

Маннинг К., Рагхаван П., Шютце Х. Введение в информационный поиск. - М.: Вильямс, 2014. - 528 с.

Дополнительная литература

7. Перечень ресурсов информационно-телекоммуникационной сети "Интернет", необходимых для освоения дисциплины (модуля)

Не используются

8. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), включая перечень необходимого программного обеспечения и информационных справочных систем (при необходимости)

На занятиях используются мультимедийные технологии, включая демонстрацию презентаций.

9. Методические указания для обучающихся по освоению дисциплины (модуля)

Студент, изучающий дисциплину, должен с одной стороны, овладеть общим понятийным аппаратом, а с другой стороны, должен научиться применять теоретические знания на практике.

Успешное освоение дисциплины требует:

- посещения студентом всех видов аудиторных занятий;
- ведения конспекта в ходе лекционных занятий;
- качественной самостоятельной подготовки к практическим занятиям, активной работы на них;
- активной самостоятельной и аудиторной работы студента;

- своевременной сдачи преподавателю заданий по аудиторным видам работ.

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

по направлению: Прикладная математика и информатика
профиль подготовки: АІ360: Передовые методы искусственного интеллекта
Физтех-школа Прикладной Математики и Информатики
кафедра алгоритмов и технологий программирования
курс: 4
квалификация: бакалавр

Семестр, формы промежуточной аттестации: 8 (весенний) - Дифференцированный зачет

Разработчик: Е.Л. Артемова, канд. техн. наук, доцент

1. Компетенции, формируемые в процессе изучения дисциплины

Код и наименование компетенции	Индикаторы достижения компетенции
ОПК-2 Способен использовать современные информационные технологии и программные средства при решении задач профессиональной деятельности, соблюдая требования информационной безопасности	ОПК-2.1 Способен применять современные вычислительную технику и сервисы сети Интернет в области (сфере) профессиональной деятельности
	ОПК-2.2 Знает и умеет применять численные математические методы и прикладное программное обеспечение для решения научных задач в профессиональной области
	ОПК-2.3 Знает основные требования информационной безопасности
ПК-2 Способен самостоятельно или в качестве члена (руководителя) малого коллектива организовывать и проводить научные исследования и их апробацию	ПК-2.1 Знает принципы построения научной работы, методы сбора и анализа полученного материала, способы аргументации
	ПК-2.2 Способен планировать и проводить научные исследования самостоятельно или в качестве члена (руководителя) малого научного коллектива
	ПК-2.3 Способен проводить апробацию результатов научно-исследовательской работы посредством публикации научных статей и участия в конференциях

2. Показатели оценивания компетенций

В результате изучения дисциплины «Автоматическая обработка естественного языка» обучающийся должен:

знать:

- постановку задач морфологического, синтаксического анализа;
- методы решения этих задач.

уметь:

- формулировать задачи классификации текстов, предложений или их элементов для выделения структурированной информации;
- реализовывать подходящий алгоритм классификации текстов;
- решать задачи выделения ключевых слов и определения тональности.

владеть:

- основными программными системами для выделения скрытых тем и снижения размерности векторных моделей.

3. Перечень типовых (примерных) вопросов, заданий, тем для подготовки к текущему контролю

1. Методы распознавания именованных сущностей в текстах
2. Автоматическое выявление терминов и терминологических связей в тексте
3. Исследование современных языковых моделей для задач семантической классификации
4. Исследование методов создания вопросно-ответных систем: IR based QA vs KB based QA
5. Что такое токенизация в обработке естественного языка и какие методы токенизации существуют?
6. Объясните, что такое стемминг и лемматизация. В чем их различия?
7. Что такое частеречная разметка (POS-тегирование) и какие задачи она решает?
8. Что такое морфологический анализ текста и какие инструменты используются для его выполнения?
9. Объясните, что такое синтаксический анализ текста. Какие методы синтаксического анализа существуют?
10. Что представляет собой модель Word2Vec и какие задачи она решает в обработке естественного языка?
11. Что такое машинный перевод и какие методы используются для его реализации?

12. Объясните, что такое анализ тональности текста и какие подходы применяются для определения тональности.
13. В чем заключается задача исследования информационного извлечения из текста (Information Extraction)?
14. Какие методы машинного обучения применяются в задачах обработки естественного языка?

4. Перечень типовых (примерных) вопросов и тем для проведения промежуточной аттестации обучающихся

1. Приведите примеры использования технологий обработки текстов, с которыми вы сталкиваетесь в повседневной жизни
2. Напишите регулярное выражение для извлечения дат
3. Сформулируйте законы Хипса и Ципфа
4. Дайте определение скрытой Марковской цепи и покажите, как она может быть использована для разрешения морфологической омонимии
5. Перечислите несколько коэффициентов считаемости биграмм
6. Почему возникает необходимость в снижении размерности в векторной модели
7. Как устроен алгоритм обучения word2vec?
8. Что такое латентное размещение Дирихле?
9. Методы классификации текстов
10. Как формулируется задача классификации последовательности?
11. Как сгенерировать текст с помощью счетной языковой модели?
12. Как сгенерировать текст с помощью нейронной вероятностной языковой модели?
13. Как формулируется задача заполнения слотов для чат-ботов?
14. Связь модели зашумленного канала и языковой модели
15. Назовите несколько мер релевантности в задаче поиска

Критерии оценивания

Оценка "Отлично" (10) - полностью и вовремя решены все задачи без ошибок. Продemonстрирован грамотный подход к решению задач, реализованы оптимальные алгоритмы, код оформлен в едином удобочитаемом стиле.

Оценка "Отлично" (9) - полностью и вовремя решены все задачи без ошибок. Продemonстрирован грамотный подход к решению задач, реализованы оптимальные алгоритмы.

Оценка "Отлично" (8) - полностью и вовремя решены все задачи без ошибок. Продemonстрирован грамотный подход к решению задач.

Оценка "Хорошо" (7) - полностью решены все задачи. Допущены несущественные ошибки.

Оценка "Хорошо" (6) - полностью решено большинство задач. В некоторых задачах допущены и не исправлены ошибки, либо некоторые задачи решены частично.

Оценка "Хорошо" (5) - полностью решено две трети задач. В некоторых задачах допущены и не исправлены ошибки, либо некоторые задачи решены частично.

Оценка "Удовлетворительно" (4) - полностью решено более половины задач. В остальных задачах допущены и не исправлены ошибки, либо некоторые задачи решены частично.

Оценка "Удовлетворительно" (3) - полностью решено более половины задач.

Оценка "Неудовлетворительно" (2) - решено менее половины задач.

Оценка "Неудовлетворительно" (1) - не решено ни одной задачи.

5. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности

Дифференцированный зачет может проводиться по итогам текущей успеваемости и сдачи заданий и других видов работ, предусмотренных программой дисциплины и (или) путем организации специального опроса, проводимого в устной и (или) письменной форме.

При проведении устного дифференцированного зачета обучающемуся предоставляется 30 минут на подготовку. Опрос обучающегося не должен превышать одного астрономического часа.

Во время проведения дифференцированного зачета обучающиеся могут пользоваться программой дисциплины, а также справочной литературой, конспектами лекций или другими материалами.