

**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:	Введение в обработку и распознавание документов
по направлению:	Прикладная математика и информатика
профиль подготовки:	А1360: Передовые методы искусственного интеллекта Физтех-школа Прикладной Математики и Информатики кафедра интеллектуальной обработки документов
курс:	3
квалификация:	бакалавр

Семестр, формы промежуточной аттестации: 6 (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

Количество контрольных работ, заданий: 1

Программу составил: К.В. Анисимович, заведующий кафедрой

Программа обсуждена на заседании кафедры интеллектуальной обработки документов 16.06.2022

Аннотация

Для задач, связанных с автоматическим анализом естественного языка, необходимо ознакомление с применяемыми в этой области современными методами и лингвистическими технологиями. Настоящий курс рассматривает использование этих методов при создании компьютерных систем обработки текстов в научно-практической области исследований «компьютерная лингвистика». Это позволяет студентам понимать качественную составляющую процессов обработки языкового материала, изучаемых в других курсах.

Материалом курса служат тексты на естественном языке, дифференцированные по своим лингвистическим свойствам, а также по задачам обработки текстов в связи с социальными запросами общества.

Программой курса предусмотрена как вводная часть, рассчитанная на ознакомление с историей дисциплины и её основными характеристиками, так и знакомство с классификацией лингвистических технологий (по уровням лингвистической разметки; по конкретным задачам пользователя; по способам формализации естественно-языковых единиц). В программе рассматриваются следующие важные темы: машинный перевод, морфологический, синтаксический и семантический этапы анализа, виды представления языковых структур.

В курсе уделяется внимание не только задачам анализа готового текстового материала, но и задачам генерации текстов на естественном языке.

Методы, рассматриваемые в курсе, нацелены на обучение элементарным практическим навыкам по применению компьютерно-лингвистических методов к языковому материалу и использованию лингвистических технологий.

1. Цели и задачи

Цель дисциплины

обзор задач, возникающих в области распознавания.

Задачи дисциплины

- обработки документов;
- связанные с формированием признакового пространства.

2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

Код и наименование компетенции	Индикаторы достижения компетенции
УК-1 Способен осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач	УК-1.2 Находит, критически анализирует и выбирает информацию, необходимую для решения поставленной задачи
	УК-1.1 Анализирует задачу, выделяя этапы ее решения, действия по решению задачи
	УК-1.3 Рассматривает различные варианты решения задачи, оценивает их преимущества и недостатки
	УК-1.4 Грамотно, логично, аргументированно формирует собственные суждения и оценки
	УК-1.5 Определяет и оценивает практические последствия возможных вариантов решения задачи
УК-2 Способен определять круг задач в рамках поставленной цели и выбирать оптимальные способы их решения, исходя из действующих правовых норм, имеющихся ресурсов и ограничений	УК-2.2 Проектирует решение конкретной задачи проекта, выбирая оптимальный способ ее решения, исходя из действующих правовых норм и имеющихся ресурсов и ограничений
	УК-2.1 Формулирует совокупность взаимосвязанных задач в рамках поставленной цели работы, обеспечивающих ее достижение. Определяет ожидаемые результаты решения поставленных задач
ОПК-2 Способен использовать современные информационные технологии и	ОПК-2.2 Знает и умеет применять численные математические методы и прикладное программное обеспечение для решения научных задач в профессиональной области

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-2.3 Знает основные требования информационной безопасности
ОПК-4 Способен осуществлять сбор и обработку научно-технической и (или) технологической информации для решения фундаментальных и прикладных задач	ОПК-4.1 Владеет методами научного поиска и интеллектуального анализа информации при решении задач профессиональной деятельности
	ОПК-4.2 Знает основные источники научно-технической и (или) технологической информации в области профессиональной деятельности
	ОПК-4.3 Умеет составлять аннотации, рефераты, библиографические перечни и обзоры информации в области своей профессиональной деятельности
	ОПК-4.4 Владеет навыками работы с компьютером и компьютерными сетями с целью получения, хранения и обработки научной (технической, технологической) информации
ПК-1 Способен ставить, формализовывать и решать задачи, в том числе разрабатывать и исследовать математические модели изучаемых явлений и процессов, системно анализировать научные проблемы, получать новые научные результаты	ПК-1.1 Способен находить, анализировать и обобщать информацию об актуальных результатах исследований в рамках тематической области своей профессиональной деятельности
	ПК-1.2 Способен выдвигать гипотезы, строить математические модели для описания изучаемых явлений и процессов, оценивать качество разработанной модели
	ПК-1.3 Способен применять теоретические и (или) экспериментальные методы исследований к конкретной научной задаче и интерпретировать полученные результаты
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:

- задачи и методы обработки изображений;
- основные методы и алгоритмы выделения признаков на изображении;
- различные методы распознавания.

уметь:

- использовать методы обработки изображений.

владеть:

- навыками самостоятельной работы в Интернете;
- навыками реализации методов обработки изображений.

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.	3	2		
2	Image Processing.	2	2		
3	Document Analysis.	3	3		
4	Recognizer.	2	2		
5	Synthesis.	2	3		
6	Нейронные сети.	3	3		15
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: 6 (Весенний)

1. Введение.

Краткая информация о курсе. Для чего нужен OCR. FineReader use cases. Основные этапы распознавания. Анонс статьи Яна Лекуна. Введение в машинное обучение.

2. Image Processing.

Подготовка изображения к распознаванию. Бинаризация, выравнивание. Повышение качества изображения для распознавания. Бинаризация и сравнение с groundtruth.

3. Document Analysis.

Распознавание структуры документа. Поиск текста, изображений, таблиц. Классификация фрагмента (текст, не текст). Описание глобальных задач. Разбор статей из ICDAR.

4. Recognizer.

Распознавание изображения блока текста. Разбивка на строки, фрагменты. ГЛД. Символьные классификаторы: генеративные и дискриминативные. Контекст (в упрощенной форме). Подходы без эвристической сегментации (LSTM, HMM). Соревнование по символьным классификаторам на Kaggle (генеративные модели).

5. Synthesis.

Какие задачи выполняет синтез. Краткое описание архитектуры синтеза и технологий вокруг него: страничный и документный синтезы, входные - выходные данные синтеза, режимы экспорта. Организация работы в синтезе: тесты, логи, настройка механизмов.

6. Нейронные сети.

Основные подходы. Применение в задаче классификации фрагмента текста.

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

Учебная аудитория, оснащенная компьютером и мультимедийным оборудованием (проектор, звуковая система).

6. Перечень рекомендуемой литературы

Основная литература

1. Lomakina-Rumyantseva E, Voronin P, Kropotov D, Vetrov D, Konushin A. Video tracking and behavior segmentation of laboratory rodents. // Pattern Recognition and Image Analysis. Pattern Recognition and Image Analysis. 2010;19(4):616-22.
2. Moiseyev B, Konev A, Chigorin A, Konushin A. Evaluation of Traffic Sign Recognition Methods Trained on Synthetically Generated Data. In: Advanced Concepts for Intelligent Vision Systems (Springer LNCS, Vol. 8192).; 2013. p. 576-83.
3. Sindeyev M, Konushin A, Rother C. Alpha-flow for video matting. In: ACCV 2012 (Springer LNCS, Vol. 7726).; 2012. p. 438-52.

Дополнительная литература

1. LeCun Y., Gradient-Based Learning Applied to Document Recognition, Proc. of the Ieee, 1988.

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 лекционных занятиях используются мультимедийные технологии, включая демонстрацию презентаций.

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

Методические рекомендации позволяют студенту оптимальным образом организовать процесс обучения. В рабочей программе приведено примерное распределение часов аудиторной и внеаудиторной нагрузки по различным темам данной дисциплины.

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 интеллектуальной обработки документов
курс: 3
квалификация: бакалавр

Семестр, формы промежуточной аттестации: 6 (весенний) - Дифференцированный зачет

Разработчик: К.В. Анисимович, заведующий кафедрой

1. Компетенции, формируемые в процессе изучения дисциплины

Код и наименование компетенции	Индикаторы достижения компетенции
УК-1 Способен осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач	УК-1.2 Находит, критически анализирует и выбирает информацию, необходимую для решения поставленной задачи
	УК-1.1 Анализирует задачу, выделяя этапы ее решения, действия по решению задачи
	УК-1.3 Рассматривает различные варианты решения задачи, оценивает их преимущества и недостатки
	УК-1.4 Грамотно, логично, аргументированно формирует собственные суждения и оценки
	УК-1.5 Определяет и оценивает практические последствия возможных вариантов решения задачи
УК-2 Способен определять круг задач в рамках поставленной цели и выбирать оптимальные способы их решения, исходя из действующих правовых норм, имеющихся ресурсов и ограничений	УК-2.2 Проектирует решение конкретной задачи проекта, выбирая оптимальный способ ее решения, исходя из действующих правовых норм и имеющихся ресурсов и ограничений
	УК-2.1 Формулирует совокупность взаимосвязанных задач в рамках поставленной цели работы, обеспечивающих ее достижение. Определяет ожидаемые результаты решения поставленных задач
ОПК-2 Способен использовать современные информационные технологии и программные средства при решении задач профессиональной деятельности, соблюдая требования информационной безопасности	ОПК-2.2 Знает и умеет применять численные математические методы и прикладное программное обеспечение для решения научных задач в профессиональной области
	ОПК-2.1 Способен применять современные вычислительную технику и сервисы сети Интернет в области (сфере) профессиональной деятельности
	ОПК-2.3 Знает основные требования информационной безопасности
ОПК-4 Способен осуществлять сбор и обработку научно-технической и (или) технологической информации для решения фундаментальных и прикладных задач	ОПК-4.1 Владеет методами научного поиска и интеллектуального анализа информации при решении задач профессиональной деятельности
	ОПК-4.2 Знает основные источники научно-технической и (или) технологической информации в области профессиональной деятельности
	ОПК-4.3 Умеет составлять аннотации, рефераты, библиографические перечни и обзоры информации в области своей профессиональной деятельности
	ОПК-4.4 Владеет навыками работы с компьютером и компьютерными сетями с целью получения, хранения и обработки научной (технической, технологической) информации
ПК-1 Способен ставить, формализовывать и решать задачи, в том числе разрабатывать и исследовать математические модели изучаемых явлений и процессов, системно анализировать научные проблемы, получать новые научные результаты	ПК-1.1 Способен находить, анализировать и обобщать информацию об актуальных результатах исследований в рамках тематической области своей профессиональной деятельности
	ПК-1.2 Способен выдвигать гипотезы, строить математические модели для описания изучаемых явлений и процессов, оценивать качество разработанной модели
	ПК-1.3 Способен применять теоретические и (или) экспериментальные методы исследований к конкретной научной задаче и интерпретировать полученные результаты

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 «Введение в обработку и распознавание документов» обучающийся должен:

знать:

- задачи и методы обработки изображений;
- основные методы и алгоритмы выделения признаков на изображении;
- различные методы распознавания.

уметь:

- использовать методы обработки изображений.

владеть:

- навыками самостоятельной работы в Интернете;
- навыками реализации методов обработки изображений.

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

1. Основные этапы распознавания;
2. подготовка изображения к распознаванию;
3. распознавание структуры документа;
4. Распознавание изображения блока текста;
5. Организация работы в синтезе;
6. Нейронные сети.

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

1. Общая постановка задачи распознавания.
2. Обучение с учителем или без.
3. Особенности задачи распознавания изображений.
4. Методы обработки изображений в задачах распознавания.
5. Способы представления изображения.
6. Математическая морфология.
7. Сужение, расширение, открытие, закрытие.
8. Бинаризация изображений.
9. Разные виды бинаризации.
10. Свертка.
11. Подсчет характеристик на изображении.
12. Структурное описание изображения.
13. Генерация и подтверждение гипотез.
14. Практические применения теории распознавания изображений.

Критерии оценивания

отлично (10) - выставляется студенту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины и умение уверенно применять их на практике при решении конкретных задач, свободное и правильное обоснование принятых решений.

отлично (9) - выставляется студенту, показавшему свободное оперирование знаниями учебной программы дисциплины, выполнение заданий творческого характера.

отлично (8) - выставляется студенту, показавшему владение программным учебным материалом с наличием несущественных ошибок в действиях, самостоятельно исправляемых учащимся.

хорошо (7) - выставляется студенту, если он твердо знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но допускается в ответе или в решении задач некоторые неточности.

хорошо (6) - выставляется студенту если он осознает воспроизведение программного учебного материала, в том числе и различной степени сложности, с несущественными ошибками, затруднения в применении отдельных навыков.

хорошо (5) - выставляется студенту если теоретическое содержание освоено не полностью, некоторые практические навыки сформированы недостаточно, в некоторых случаях были допущены ошибки.

удовлетворительно (4) - выставляется студенту, показавшему фрагментарный, разрозненный характер знаний, недостаточно правильные формулировки базовых понятий, нарушения логической последовательности в изложении программного материала, но при этом он владеет основными разделами учебной программы, необходимыми для дальнейшего обучения и может применять полученные знания по образцу в стандартной ситуации.

удовлетворительно (3) - выставляется студенту в случае большого количества недочетов и неправильных ответов, а также пассивной работе в ходе занятий, многие учебные задания не выполнены.

неудовлетворительно (2) - выставляется студенту, который не знает большей части основного содержания учебной программы дисциплины, допускает грубые ошибки в формулировках основных понятий дисциплины и не умеет использовать полученные знания при решении типовых практических задач.

неудовлетворительно (1) - выставляется студенту, который не освоил теоретическое и практическое содержание курса, все выполненные учебные задания содержат грубые ошибки.

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

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