Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение высшего образования
«Сибирский государственный индустриальный университет»

Кафедра прикладных информационных технологий и программирования

УТВЕРЖДАЮ
Директор института
информационных технологий и
автоматизированных систем
_____ Л.Д. Павлова
подпись
«____» _____ 20___ г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Анализ текстовых данных

09.04.03 «Прикладная информатика»
(направленность (профиль): «Прикладная информатика»)

Квалификация выпускника
Магистр

Форма обучения
Очная форма

Срок обучения: 2 года

Год начала подготовки 2024

Новокузнецк
2024

1 Цели и задачи освоения учебной дисциплины

Целями учебной дисциплины являются:

– изучение подходов к решению основных задач автоматической обработки текстов на основе классического машинного обучения и глубоких нейронный сетей.

Задачами учебной дисциплины являются:

– освоение методов предобработки текстовых данных;

– применение на практике основных методов классификации и кластеризации текстов, методов поиска и / или генерации ответа на вопрос и базовых методов машинного перевода.

2 Место учебной дисциплины в структуре ООП по направлению подготовки (специальности)

Учебная дисциплина относится к учебным дисциплинам части, формируемой участниками образовательных отношений **Блока 1 «Дисциплины (модули)»** ООП по направлению подготовки (специальности) 09.04.03 «Прикладная информатика».

Учебная дисциплина базируется на предварительном усвоении обучающимися учебных дисциплин:

– Математические и инструментальные методы анализа данных;

– Машинное обучение;

– Программирование глубоких нейронных сетей на Python;

– Наука о данных и аналитика больших данных.

Учебная дисциплина дополняет знания и умения, получаемые по одновременно изучаемым и последующим учебным дисциплинам:

– Наука о данных и аналитика больших данных;

– Методы и инструменты цифровой трансформации.

3 Планируемые результаты обучения по учебной дисциплине

Процесс изучения учебной дисциплины направлен на формирование следующих компетенций:

– Профессиональные компетенции

Наименование категории (группы) ПК	Код и наименование ПК	Код и наименование индикатора достижения ПК	Планируемые результаты обучения
	ПК-1: Способен участвовать в управлении работами по получению, хранению и обработки	ПК-1.1 Принимает участие в разработке моделей данных, проводит анализ больших объемов данных, строит модели на основе данных	– знать: основные математические и алгоритмические модели систем, методы их имитационного моделирования, основы построения компьютерных

	больших объемов данных		дискретно-математических моделей. – уметь: проводить необходимые статистические расчёты в рамках построенной экономической модели, выбирать современные методы принятия экономических решений в информационных системах.
	ПК-2: Способен к проведению работ по обработке и анализу научно-технической информации и результатов исследований	ПК-2.1 Собирает и изучает научно-техническую информацию по теме исследований и разработок	– знать: способы и средства сбора научно-технической и статистической информации по тематике исследования. – уметь: применять информационные технологии и статистические методы для сбора, обработки и анализа научно-технической информации по тематике исследования.
		ПК-2.2 Проводит анализа научных данных, результатов экспериментов и наблюдений	– знать: основные положения теории статистики, основные методы анализа информационных процессов в сложных экономических системах. – уметь: решать задачи теоретического и прикладного характера из различных разделов математики, статистики и теории систем, строить модели объектов и понятий.

4 Объем и содержание учебной дисциплины

Учебные занятия по учебной дисциплине проводятся в форме контактной работы и в форме самостоятельной работы обучающихся.

Контактная работа включает в себя занятия лекционного типа (лекции), занятия семинарского типа (семинары, практические занятия,

практикумы), промежуточную аттестацию обучающихся и иные формы взаимодействия обучающихся с педагогическими работниками и (или) лицами, привлекаемыми к реализации ООП на иных условиях, в том числе при проведении промежуточной аттестации обучающихся. Контактная работа может проводиться с применением электронного обучения, дистанционных образовательных технологий.

Объем учебной дисциплины

Семестр / курс		***ИТОГО***	***5 семестр***
Форма промежуточной аттестации			*экзамен*
Трудоёмкость	*академ. час.*	***180***	*180*
	зачетных единиц	***5***	*5*
Лекции, *академ. час.*		***16***	*16*
в форме практической подготовки		***0***	*0*
Лабораторные работы, *академ. час.*		***0***	*0*
в форме практической подготовки		***0***	*0*
Практические занятия, *академ. час.*		***32***	*32*
в форме практической подготовки		***0***	*0*
Курсовая работа / проект, *академ. час.*		***0***	*0*
в форме практической подготовки		***0***	*0*
Консультации, *академ. час.*		***0***	*0*
в форме практической подготовки		***0***	*0*
Самостоятельная работа, *академ. час.*		***114***	*114*
в форме практической подготовки		***0***	*0*
Контроль, *академ. час.*		***18***	*18*
в форме практической подготовки		***0***	*0*

Содержание учебной дисциплины

Раздел 1 Методы текстового анализа;

Тема 1.1 Введение в анализ текстов, базовые методы предобработки и выделения признаков (Задачи анализа текстов, специфика, история. Базовая предобработка текстов. Простейшие текстовые признаки: “мешок слов”, TF-IDF. Предобработка текстовых данных: регулярные выражения);

Тема 1.2 Неглубокие векторные представления слов (Идея векторных представлений, one-hot-векторы, SVD. Модель word2vec, методы её обучения. Оптимизации обучения word2vec: SGNS и иерархический softmax. Модель GloVe. Модель и библиотека FastText, приём Hashing Trick);

Тема 1.3 Классификация текстов (Задача классификация текстов. Логистическая регрессия на счётчиках и TF-IDF. Неглубокие векторные представления документов. Библиотека FastText для классификации текстов.CNN для классификации текстов. Работа с обучающими данными);

Тема 1.4 Разметка последовательности (Счетные языковые модели. Морфологический анализ, скрытые Марковские модели.

Нейросетевые языковые модели. Рекуррентные нейронные сети (RNN). Генерация текстов. Извлечение именованных сущностей (NER). Перекрестное обучение);

Раздел 2 Приложения анализа текстовых данных;

Тема 2.1 Машинный перевод (Модели класса кодировщик-декодировщик. Механизм внимания. Модель Трансформер. Метрики качества в машинном переводе. Улучшение качества машинного перевода);

Тема 2.2 Предобученные языковые модели. (Векторное представление предложений. Модель ELMo. Модель BERT. Модель GPT2. Оценка качества моделей. BERTоведение.Модель XLNet. Сжатие языковых моделей. Потомки модели BERT. Мультимодальные модели. Мультиязычные модели. Zero-shot мультиязычная классификация текстов.
Мультиязычное распознавание именованных сущностей);

Тема 2.3 Синтаксис в рамках грамматики зависимостей (Автоматический синтаксический анализ. Теоретические подходы. Парсинг на основе грамматики зависимостей. Метрики и соревнования. Инструменты. Синтаксический парсинг);

Тема 2.4 Тематическое моделирование (Постановка задачи тематического моделирования. Модель PLSA. EM-алгоритм. Модель ARTM. Модель LDA. Модель M-ARTM. Технические аспекты обучения TM);

Тема 2.5 Суммаризация и симплификация текстов (Суммаризация текстов. Абстрактивная суммаризация. Упрощение текстов);

Тема 2.6 QA-системы (Основы разработки чат-ботов. Инструменты разработки чат-ботов. Виртуальные ассистенты. Вопросно-ответные системы. Машинное чтение.
Вопросно-ответные системы в индустрии);

Тема 2.7 Графы знаний (Методы извлечения графа знаний. Машинное обучение для извлечения графа знаний. Разрешение многозначности).

5 Перечень тем лекций

№ раздела / темы дисциплины	Темы лекций	Трудоемкость, *академ.час*	
		всего	**в форме практической подготовки**
Раздел 1.	Методы текстового анализа	6	
Раздел 2.	Приложения анализа текстовых данных	10	
Итого:		**16**	**0**

6 Перечень тем практических занятий (семинаров)

№ раздела / темы дисциплины	Темы практических занятий (семинаров)	Трудоемкость, *академ.час*	
		всего	**в форме практической подготовки**
Тема 1.1.	Выделение признаков из текстов	3	
Тема 1.2.	Модели word2vec и fastText	3	
Тема 1.3.	Word2vec и fastText для классификации текстов.CNN для классификации текстов	4	
Тема 1.4.	Языковые модели для генерации текстов. Извлечение именованных сущностей	3	
Тема 2.1.	Модель кодировщик-декодировщик. Модель Трансформер	4	
Тема 2.2.	Классификация текстов.	3	
Тема 2.4.	Тематическое моделирование	4	
Тема 2.5.	Алгоритм TextRank	4	
Тема 2.7.	Библиотека CoreNLP	4	
Итого:		**32**	**0**

7 Перечень тем лабораторных работ

№ раздела / темы дисциплины	Темы лабораторных работ	Трудоемкость, *академ.час*	
		всего	**в форме практической подготовки**
	Отсутствуют		
Итого:		**0**	**0**

8 Перечень тем курсовых работ (проектов)

№ раздела / темы дисциплины	Темы курсовых работ (проектов)	Трудоемкость, *академ.час*	
		всего	**в форме практической подготовки**
	Отсутствуют		
Итого:		**0**	**0**

9 Виды самостоятельной работы

№ раздела / темы дисциплины	Виды самостоятельной работы	Трудоемкость, *академ.час*	
		всего	**в форме практической подготовки**

Раздел 1.	1. Изучение лекционного материала; 2. Подготовка к практическому занятию; 3. Прохождение тестирования.	57	
Раздел 2.	1. Изучение лекционного материала; 2. Подготовка к практическому занятию; 3. Прохождение тестирования.	57	
Контроль	*Подготовка к экзамену*	18	
Итого:		**132**	**0**

10 Учебно-методическое и информационное обеспечение учебной дисциплины

а) литература:

1 Рабчевский, А. Н. Синтетические данные и развитие нейросетевых технологий : учебное пособие для вузов / А. Н. Рабчевский. — Москва : Издательство Юрайт, 2024. — 187 с. — (Высшее образование). — ISBN 978-5-534-17716-9. – URL: https://urait.ru/bcode/545036 (дата обращения: 01.03.2024);

2 Анализ данных : учебник для вузов / В. С. Мхитарян [и др.] ; под редакцией В. С. Мхитаряна. — Москва : Издательство Юрайт, 2024. — 490 с. — (Высшее образование). — ISBN 978-5-534-00616-2. – URL: https://urait.ru/bcode/536007 (дата обращения: 01.03.2024).

б) ресурсы информационно-телекоммуникационной сети «Интернет»:

1 Консультант студента : электронно-библиотечная система / ООО «КОНСУЛЬТАНТ СТУДЕНТА». – Москва, [200 –]. – URL: http://www.studentlibrary.ru. – Режим доступа: для авторизир. пользователей;

2 ЛАНЬ : электронно-библиотечная система : [коллекция «Инженерно-технические науки»] / ООО «Издательство ЛАНЬ». – Санкт-Петербург, [200 –]. – URL: http://e.lanbook.com. – Режим доступа: для авторизир. пользователей;

3 НАУЧНАЯ ЭЛЕКТРОННАЯ БИБЛИОТЕКА eLIBRARY.RU : база данных / ООО «НЭБ». – Москва, [200 –]. – URL: http://elibrary.ru. – Режим доступа: по подписке;

4 Образовательная платформа ЮРАЙТ / ООО «Электронное издательство ЮРАЙТ». – Москва, [200 –]. – URL: https://urait.ru. – Режим доступа: для авторизир. пользователей;

5 Университетская библиотека онлайн : электронно-библиотечная система / ООО «Директ-Медиа». – Москва, [200 –]. – URL:

https://biblioclub.ru. – Режим доступа: для авторизир. пользователей. – URL: http://www.biblioclub.ru;

6 Электронная библиотека // Научно-техническая библиотека СибГИУ : сайт. – Новокузнецк, [200 –]. – URL: http://library.sibsiu.ru/LibrELibraryFullText.asp. – Режим доступа: для авторизир. пользователей. – URL: https://library.sibsiu.ru/LibrELibraryFullText.asp;

7 Электронные периодические издания ИВИС : универсальная база данных / ООО «ИВИС». – Москва, [200 –]. – URL: http://eivis.ru. – Режим доступа: по подписке;

8 Электронный каталог : сайт / Научно-техническая библиотека СибГИУ. – Новокузнецк, [199 –]. – URL: http://libr.sibsiu.ru. – URL: https://libr.sibsiu.ru.

в) лицензионное и свободно распространяемое программное обеспечение:

– 7-Zip;
– Adobe Acrobat Reader;
– Astra Linux Special Edition;
– Kaspersky Endpoint Security;
– Microsoft Office;
– Microsoft Windows;
– OnlyOffice;
– PyCharm;
– Р7-Офис.

г) базы данных и информационно-справочные системы:

1 ГАРАНТ : справочно-правовая система / ООО «Правовой центр «Гарант». – Кемерово, [200 –]. – Режим доступа: компьютерная сеть Сиб. гос. индустр. ун-та.;

2 КонсультантПлюс : справочно-правовая система / ООО «Информационный центр АНВИК». – Новокузнецк, [199 –]. – Режим доступа: компьютерная сеть библиотеки Сиб. гос. индустр. ун-та.;

3 Техэксперт : информационно-справочная система / ООО «Группа компаний «Кодекс». – Кемерово, [200 –]. – Режим доступа: компьютерная сеть Сиб. гос. индустр. ун-та.

11 Материально-техническое обеспечение учебной дисциплины

Материально-техническое обеспечение учебной дисциплины включает учебные аудитории, оснащенные оборудованием, компьютерной техникой, и техническими средствами обучения, в том числе:

- учебную аудиторию для проведения занятий лекционного типа, оборудованную учебной доской, экраном и мультимедийным проектором;
- учебную аудиторию для проведения занятий семинарского типа

(практических занятий);
- учебную аудиторию (помещения) для групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации;
- помещения для самостоятельной работы, оснащенные компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно-образовательную среду, научно-техническую библиотеку СибГИУ.

Рабочая программа дисциплины составлена в соответствии с требованиями ФГОС ВО по направлению подготовки (специальности) 09.04.03 «Прикладная информатика».

Составитель(и):

доцент Маслова Елена Владимировна (кафедра прикладных информационных технологий и программирования).

Рабочая программа дисциплины рассмотрена и утверждена на заседании кафедры.

Приложение

Аннотация
рабочей программы дисциплины «Анализ текстовых данных»

по направлению подготовки (специальности)
09.04.03 «Прикладная информатика»
(направленность (профиль): «Прикладная информатика»)
форма обучения – Очная форма

1 Цели и задачи освоения учебной дисциплины

Целями учебной дисциплины являются:

– изучение подходов к решению основных задач автоматической обработки текстов на основе классического машинного обучения и глубоких нейронный сетей.

Задачами учебной дисциплины являются:

– освоение методов предобработки текстовых данных;

– применение на практике основных методов классификации и кластеризации текстов, методов поиска и / или генерации ответа на вопрос и базовых методов машинного перевода.

2 Место учебной дисциплины в структуре ООП по направлению подготовки (специальности)

Учебная дисциплина относится к учебным дисциплинам части, формируемой участниками образовательных отношений **Блока 1 «Дисциплины (модули)»** ООП по направлению подготовки (специальности) 09.04.03 «Прикладная информатика».

Учебная дисциплина базируется на предварительном усвоении обучающимися учебных дисциплин:

– Математические и инструментальные методы анализа данных;

– Машинное обучение;

– Программирование глубоких нейронных сетей на Python;

– Наука о данных и аналитика больших данных.

Учебная дисциплина дополняет знания и умения, получаемые по одновременно изучаемым и последующим учебным дисциплинам:

– Наука о данных и аналитика больших данных;

– Методы и инструменты цифровой трансформации.

3 Планируемые результаты обучения по учебной дисциплине

Процесс изучения учебной дисциплины направлен на формирование следующих компетенций:

– Профессиональные компетенции

Наименование категории (группы) ПК	Код и наименование ПК	Код и наименование индикатора достижения ПК	Планируемые результаты обучения

	ПК-1: Способен участвовать в управлении работами по получению, хранению и обработки больших объемов данных	ПК-1.1 Принимает участие в разработке моделей данных, проводит анализ больших объемов данных, строит модели на основе данных	– знать: основные математические и алгоритмические модели систем, методы их имитационного моделирования, основы построения компьютерных дискретно-математических моделей. – уметь: проводить необходимые статистические расчёты в рамках построенной экономической модели, выбирать современные методы принятия экономических решений в информационных системах.
	ПК-2: Способен к проведению работ по обработке и анализу научно-технической информации и результатов исследований	ПК-2.1 Собирает и изучает научно-техническую информацию по теме исследований и разработок	– знать: способы и средства сбора научно-технической и статистической информации по тематике исследования. – уметь: применять информационные технологии и статистические методы для сбора, обработки и анализа научно-технической информации по тематике исследования.
		ПК-2.2 Проводит анализа научных данных, результатов экспериментов и наблюдений	– знать: основные положения теории статистики, основные методы анализа информационных процессов в сложных экономических системах. – уметь: решать задачи теоретического и прикладного характера из различных разделов математики, статистики и теории систем, строить модели

			объектов и понятий.

4 Объем учебной дисциплины

Семестр / курс		***ИТОГО***	***5 семестр***
Форма промежуточной аттестации			*экзамен*
Трудоёмкость	*академ. час.*	***180***	*180*
	зачетных единиц	***5***	*5*
Лекции, *академ. час.*		***16***	*16*
в форме практической подготовки		***0***	*0*
Лабораторные работы, *академ. час.*		***0***	*0*
в форме практической подготовки		***0***	*0*
Практические занятия, *академ. час.*		***32***	*32*
в форме практической подготовки		***0***	*0*
Курсовая работа / проект, *академ. час.*		***0***	*0*
в форме практической подготовки		***0***	*0*
Консультации, *академ. час.*		***0***	*0*
в форме практической подготовки		***0***	*0*
Самостоятельная работа, *академ. час.*		***114***	*114*
в форме практической подготовки		***0***	*0*
Контроль, *академ. час.*		***18***	*18*
в форме практической подготовки		***0***	*0*

5 Краткое содержание учебной дисциплины

В структуре учебной дисциплины выделяются следующие основные разделы (темы):

Раздел 1 Методы текстового анализа;

Тема 1.1 Введение в анализ текстов, базовые методы предобработки и выделения признаков (Задачи анализа текстов, специфика, история. Базовая предобработка текстов. Простейшие текстовые признаки: “мешок слов”, TF-IDF. Предобработка текстовых данных: регулярные выражения);

Тема 1.2 Неглубокие векторные представления слов (Идея векторных представлений, one-hot-векторы, SVD. Модель word2vec, методы её обучения. Оптимизации обучения word2vec: SGNS и иерархический softmax. Модель GloVe. Модель и библиотека FastText, приём Hashing Trick);

Тема 1.3 Классификация текстов (Задача классификация текстов. Логистическая регрессия на счётчиках и TF-IDF. Неглубокие векторные представления документов. Библиотека FastText для классификации текстов.CNN для классификации текстов. Работа с обучающими данными);

Тема 1.4 Разметка последовательности (Счетные языковые модели. Морфологический анализ, скрытые Марковские модели. Нейросетевые языковые модели. Рекуррентные нейронные сети (RNN). Генерация текстов. Извлечение именованных сущностей (NER). Перекрестное обучение);

Раздел 2 Приложения анализа текстовых данных;

Тема 2.1 Машинный перевод (Модели класса кодировщик-декодировщик. Механизм внимания. Модель Трансформер. Метрики качества в машинном переводе. Улучшение качества машинного перевода);

Тема 2.2 Предобученные языковые модели. (Векторное представление предложений. Модель ELMo. Модель BERT. Модель GPT2. Оценка качества моделей. BERTоведение.Модель XLNet. Сжатие языковых моделей. Потомки модели BERT. Мультимодальные модели. Мультиязычные модели. Zero-shot мультиязычная классификация текстов.
Мультиязычное распознавание именованных сущностей);

Тема 2.3 Синтаксис в рамках грамматики зависимостей (Автоматический синтаксический анализ. Теоретические подходы. Парсинг на основе грамматики зависимостей. Метрики и соревнования. Инструменты. Синтаксический парсинг);

Тема 2.4 Тематическое моделирование (Постановка задачи тематического моделирования. Модель PLSA. EM-алгоритм. Модель ARTM. Модель LDA. Модель M-ARTM. Технические аспекты обучения TM);

Тема 2.5 Суммаризация и симплификация текстов (Суммаризация текстов. Абстрактивная суммаризация. Упрощение текстов);

Тема 2.6 QA-системы (Основы разработки чат-ботов. Инструменты разработки чат-ботов. Виртуальные ассистенты. Вопросно-ответные системы. Машинное чтение.
Вопросно-ответные системы в индустрии);

Тема 2.7 Графы знаний (Методы извлечения графа знаний. Машинное обучение для извлечения графа знаний. Разрешение многозначности).

6 Составитель(и):

доцент Маслова Елена Владимировна (кафедра прикладных информационных технологий и программирования).