ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«БЕЛГОРОДСКИЙ ГОСУДАРСТВЕННЫЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ»
(Н И У « Б е л Г У »)

УТВЕРЖДАЮ
Директор института инженерных и цифровых технологии

К.А. Польщиков
18.05.2022

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Информационный поиск и обработка текстов на естественном языке
наименование дисциплины

Программа составлена в соответствии с требованиями ФГОС ВО по направлению подготовки

Направление подготовки 09.04.03 Прикладная информатика
Профиль подготовки Искусственный интеллект и наука о данных

Автор: Профессор, д.ф.-м.н. профессор Тулупьев Александр Львович, доцент, к.псх.н. доцент Тулупьева Татьяна Валентиновна, доцент, к.т.н. Абрамов Максим Викторович, Корепанова Анастасия Андреевна
должность, ученая степень, ученое звание, инициалы и фамилия

Программа одобрена Кафедрой прикладной информатики информационных технологий

Протокол заседания кафедры от 06.04.2022 № 8
дата

Программа согласована Кафедрой прикладной информатики и информационных технологий

Протокол заседания кафедры от 06.04.2022 № 8
дата

Раздел 1. Характеристики учебных занятий

1.1. Цели и задачи учебных занятий

Целью освоения дисциплины «Информационный поиск и обработка текстов на естественном языке» является ознакомление слушателей с методами обработки текста на естественном языке, а также методами обработки слабоструктурированных данных и извлечения информации. Предполагается знакомство с методами извлечения отношений, анализа тональности, аннотирования и кластеризации текстов, а также с существующими программными реализациями этих методов.

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

Программа дисциплины в первую очередь предназначена для обучающихся 1-го курса магистратуры (2 семестр) по направлению подготовки 09.04.03 «Прикладная информатика», хотя может быть использована и на других курсах магистратуры. Максимальная эффективность дисциплины будет обеспечена при следующем условии: обучающийся владеет базовыми математическими понятиями и навыками программирования на языке высокого уровня, имеет представление о принципах проектной работы и работе с системами управления базами данных.

1.3. Перечень результатов обучения (learning outcomes)

Код и наименование компетенции	Планируемые результаты обучения, обеспечивающие формирование компетенции	Код индикатора и индикатор достижения универсальной компетенции
ПКП-3-ИИР-ОПК-3. Способен использовать методы научных исследований и математического моделирования в области проектирования и управления системами искусственного интеллекта	знает: · логические методы и приемы научного исследования; методологические принципы современной науки, направления, концепции, источники знания и приемы работы с ними; основные особенности научного метода познания; программно-целевые методы решения научных проблем; основы моделирования управленческих решений; динамические оптимизационные модели; математические модели оптимального управления для непрерывных и дискретных процессов, их сравнительный анализ; многокритериальные методы принятия решений в профессиональной деятельности умеет: · применять логические методы и приемы научного исследования; методологические принципы современной науки, концепции, источники знания и приемы работы с ними; основные метода научного познания; программно-целевые методы решения научных проблем; основы моделирования	ПКП-3-ИИР-ОПК-3.1. Применяет логические методы и приемы научного исследования, методологические принципы современной науки, направления, концепции, источники знания и приемы работы с ними, основные особенности научного метода познания, программно-целевые методы решения научных проблем в профессиональной деятельности

	управленческих решений; динамические оптимизационные модели; математические модели оптимального управления для непрерывных и дискретных процессов, их сравнительный анализ; многокритериальные методы принятия решений в профессиональной деятельности	
ПКП-5-ИИР-ПК-2. Способен выбирать, разрабатывать и проводить экспериментальную проверку работоспособности программных компонентов систем искусственного интеллекта по обеспечению требуемых критериев эффективности и качества функционирования	знает: · основные критерии эффективности и качества функционирования системы, основанной на знаниях: точность, релевантность, достоверность, целостность, быстрота решения задач, надежность, защищенность функционирования систем, основанных на знаниях · методы, языки и программные средства разработки программных компонентов систем, основанных на знаниях · Знает методы постановки задач, проведения и анализа тестовых и экспериментальных испытаний работоспособности систем, основанных на знаниях умеет: · выбирать, адаптировать, разрабатывать и интегрировать программные компоненты систем, основанных на знаниях, с учетом основных критериев эффективности и качества функционирования · ставить задачи и проводить тестовые и экспериментальные испытания работоспособности систем, основанных на знаниях, анализировать результаты и вносить изменения	ПКП-5-ИИР-ПК-2.1. Выбирает и разрабатывает программные компоненты систем искусственного интеллекта ПКП-5-ИИР-ПК-2.2. Проводит экспериментальную проверку работоспособности систем искусственного интеллекта
ПКП-1-ИИР-ОПК-1 Способен разрабатывать алгоритмы и программные средства для решения задач в области создания и применения искусственного интеллекта	знает: · инструментальные среды, программно-технические платформы для решения профессиональных задач умеет: · применять инструментальные среды, программно-технические платформы для решения профессиональных задач	ПКП-1-ИИР-ОПК-1.1. Применяет инструментальные среды, программно-технические платформы для решения задач в области создания и применения искусственного интеллекта

1.4. Перечень и объём активных и интерактивных форм учебных занятий

Активные формы учебных занятий — лекции, предполагающие обсуждение материала с преподавателем, 10 ак.ч.

Раздел 2. Организация, структура и содержание учебных занятий

2.1. Организация учебных занятий

2.1.1 Основной курс

Трудоёмкость, объёмы учебной работы и наполняемость групп обучающихся																		
Код модуля в составе дисциплины, практики и т.п.	Контактная работа обучающихся с преподавателем												Самостоятельная работа				Объём активных и интерактивных	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	текущий контроль (сам.раб.)	промежуточная аттестация (сам.раб.)	итоговая аттестация (сам.раб.)		
ОСНОВНАЯ ТРАЕКТОРИЯ																		
Форма обучения: очная																		
Семестр 2	30		2	16					2				64		30		10	4
	2-25		2-25	2-25					2-25				1-1		1-1			
ИТОГО	30		2	16					2				64		30		10	4

Виды, формы и сроки текущего контроля успеваемости и промежуточной аттестации						
Код модуля в составе дисциплины, практики и т.п.	Формы текущего контроля успеваемости		Виды промежуточной аттестации		Виды итоговой аттестации (только для программ итоговой аттестации и дополнительных образовательных программ)	
	Формы	Сроки	Виды	Сроки	Виды	Сроки
ОСНОВНАЯ ТРАЕКТОРИЯ						
Форма обучения: очная						
Семестр 2			экзамен, устно, традиционная форма	по графику промежуточной аттестации		

2.2. Структура и содержание учебных занятий

№ п/п	Наименование темы (раздела, части)	Вид учебных занятий	Количество часов
I	Введение в обработку естественного языка	лекции	7
		практические занятия	4
		по методическим материалам	16
II	Классификация и кластеризация текстов	лекции	7
		практические занятия	4
		по методическим материалам	16
III	Информационный поиск	лекции	7
		практические занятия	4
		по методическим материалам	16
IV	Введение в машинный перевод	лекции	9
		практические занятия	4
		по методическим материалам	16
	Промежуточная аттестация	самостоятельная работа	2
		консультации	2
		экзамен	30
Итого			

2.2.1 Содержание учебных занятий

Темы для изучения и обсуждения:

1. Введение в обработку естественного языка. Этапы анализа текста. Обзор основных приложений автоматического анализа текста (АОТ) (машинный перевод, информационный поиск, и т.д.). Слова, фразы, предложения, корпусы. Языковые модели. Автоматический морфологический анализ и синтез. Виды морфологического анализа: стемминг, лемматизация, полный морфоанализ. Принципы морфоанализа на базе словаря основ или словаря словоформ. Морфологические процессоры для русского языка

2. Классификация и кластеризация текстов. Классификация текстов как типичная задача обработки текстов в области TextMining. Обзор методов машинной классификации. Выбор признаков и метрик. Особенности кластеризации текстов. Рубрицирование текстовых документов. Обзор задач АОТ, решаемых на основе классификации текстов. Модели и методы автоматической классификации и кластеризации текстовой информации. Иерархические и вероятностные подходы. Интеллектуальный анализ данных.

3. Информационный поиск. Индексирование текстов для информационного поиска. Векторная модель документа. Булевский поиск, ранжированный поиск. Оценка релевантности документа. Поиск в сети Интернет, принципы работы поисковых машин. Автоматическое реферирование и аннотирование документов как смежные задачи информационного поиска. Основные стратегии сжатия текста. Типы аннотаций. Обзорное реферирование. Оценка качества аннотаций

4. Введение в машинный перевод. Стратегии машинного перевода, основанного на лингвистических правилах. Статистический машинный перевод: особенности и виды. Принципы создания статистического переводчика.

Раздел 3. Обеспечение учебных занятий

3.1. Методическое обеспечение

3.1.1 Методические указания по освоению дисциплины

Успешное освоение дисциплины возможно благодаря посещению лекций и практических занятий, участию в обсуждении рассматриваемых вопросов, самостоятельной работе, включающей в себя чтение специальной литературы по разделам темы.

3.1.2 Методическое обеспечение самостоятельной работы

При самостоятельном изучении теоретического материала, выполнении практических заданий и во время подготовки доклада целесообразно использовать рекомендованную основную и дополнительную литературу.

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

Экзамен в устной форме.

Экзамен проводится в устной форме. Билет состоит из двух вопросов, на подготовку ответа на которые даётся не менее одного академического часа. После ответа на вопросы билета преподаватель вправе задать дополнительные вопросы по любой теме из списка вопросов, вынесенных на экзамен. Количество и содержание дополнительных вопросов – на усмотрение преподавателя, принимающего экзамен. Максимальный балл за ответ на каждый из трех вопросов билета и на дополнительные вопросы — 100 баллов. Результаты усредняются.

Перевод баллов в оценку (набранные баллы округляются до десятых):

До 50 – 2 (**F** оценка в системе ECTS);

от 50 до 60 – 3 (**E** оценка в системе ECTS);

от 61 до 69 – 3 (**D** оценка в системе ECTS);

от 70 до 79 – 4 (**C** оценка в системе ECTS);

от 80 до 89 – 4 (**B** оценка в системе ECTS);

от 90 – 5 (**A** оценка в системе ECTS).

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

№	Код индикатора и индикатор достижения универсальной компетенции	Контрольно-измерительные материалы (КИМ) (тестовые вопросы, контрольные задания, кейсы и пр.)
1	ПКП-3-ИИР-ОПК-3.1. Применяет логические методы и приемы научного исследования, методологические принципы современной науки, направления, концепции, источники знания и приемы работы с ними, основные особенности научного метода познания, программно-целевые методы решения научных проблем в профессиональной деятельности	Экзамен.

2	ПКП-5-ИИР-ПК-2.1. Выбирает и разрабатывает программные компоненты систем искусственного интеллекта	Экзамен.
3	ПКП-5-ИИР-ПК-2.2. Проводит экспериментальную проверку работоспособности систем искусственного интеллекта	Экзамен.
4	ПКП-1-ИИР-ОПК-1.1. Применяет инструментальные среды, программно-технические платформы для решения задач в области создания и применения искусственного интеллекта	Экзамен.

3.1.4.1 Формируемые дисциплиной компетенции

- ПКП-3-ИИР-ОПК-3. Способен использовать методы научных исследований и математического моделирования в области проектирования и управления системами искусственного интеллекта
- ПКП-5-ИИР-ПК-2. Способен выбирать, разрабатывать и проводить экспериментальную проверку работоспособности программных компонентов систем, основанных на знаниях, по обеспечению требуемых критериев эффективности и качества функционирования
- ПКП-1-ИИР-ОПК-1 Способен разрабатывать алгоритмы и программные средства для решения задач в области создания и применения искусственного интеллекта

□ **Формируется дисциплиной.**

✓ **Развивается дисциплиной.**

□ **Полностью сформирована по результатам освоения дисциплины.**

Шкала оценивания: линейная, определяется долей успешно выполненных заданий, проверяющих данную компетенцию.

3.1.4.2 Контрольно-измерительные материалы (примеры)

Примерный список вопросов к экзамену:

1. Этапы анализа текста. Обзор основных приложений автоматического анализа текста (АОТ) (машинный перевод, информационный поиск, и т.д.). Слова, фразы, предложения, корпусы.
2. Классификация текстов как типичная задача обработки текстов в области TextMining. Обзор методов машинной классификации.
3. Выбор признаков и метрик. Особенности кластеризации текстов. Рубрицирование текстовых документов.
4. Обзор задач АОТ, решаемых на основе классификации текстов. Модели и методы автоматической классификации и кластеризации текстовой информации.
5. Индексирование текстов для информационного поиска. Векторная модель документа.
6. Булевский поиск, ранжированный поиск. Оценка релевантности документа.
7. Оценка релевантности документа. Поиск в сети Интернет, принципы работы поисковых машин.
8. Автоматическое реферирование и аннотирование документов как смежные задачи информационного поиска. Основные стратегии сжатия текста. Типы аннотаций. Обзорное реферирование. Оценка качества аннотаций.

9. Стратегии машинного перевода, основанного на лингвистических правилах. Статистический машинный перевод: особенности и виды.
10. Принципы создания статистического переводчика.

Примеры тестовых заданий:

1) Что такое лемматизация:
   a) приведение слова в нормальную морфологическую форму; +
   b) разбиение текста на слова;
   c) нахождение основы слова для заданного исходного слова;
   d) доказательство леммы.
2) Векторное представление текста, не учитывающее грамматику и порядок слов, но сохраняющее информацию об их колличестве:
   a) мешок слов; +
   b) мешок N-грамм;
   c) word2vec;
   d) fastText.
3) Что относится к недостаткам байесовской классификации:
   a) неспособность учитывать зависимость результата классификации от сочетания признаков; +
   b) низкая скорость работы;
   c) сложная реализация алгоритма;
   d) сложная интерпретируемость результатов работы алгоритма.
4) Как вычисляется формула TF-IDF:
   a) TF*IDF; +
   b) TF+IDF;
   c) TF-IDF;
   d) IDF-TF.
5) Что такое стемминг:
   a) приведение слова в нормальную морфологическую форму;
   b) разбиение текста на слова;
   c) нахождение основы слова для заданного исходного слова; +
   d) морфологический разбор слова.
6) Что не входит в этапы машинного анализа текста:
   a) графематический анализ;
   b) морфологический анализ;
   c) лингвистический анализ; +
   d) прагматический анализ.
7) Что не входит в задачу анализа тональности по аспектам:
   a) выделение аспектных терминов;
   b) классификация аспектных терминов в аспектные категории;
   c) автоматическое определение аспектов по отношению к выделенным категориям;
   d) ранжирование аспектных категорий. +
8) Что такое Text Mining:
   a) интеллектуальный анализ текстов; +
   b) извлечение текстов из информационного источника;
   c) создание новых блоков базы транзакций ради вознаграждения в различных криптовалютах;
   d) выделение в тексте предложений и словоформ, точнее токенов.
9) Стемминг, лемматизация – это:
   a) Виды морфологического анализа;
   b) виды лингвистического анализа;

c) виды синтаксического анализа;
d) виды фонетического анализа.

10) Что относится к обучению без учителя:
a) кластеризация;
b) классификация;
c) регрессия;
d) всё вышеперечисленное.

***Проверяемые компетенции*:** ПКП-1-ИИР-ОПК-1, ПКП-3-ИИР-ОПК-3, ПКП-5-ИИР-ПК-2
Проверяемые индикаторы: все, в соответствии с компетенциями
***Критерии оценивания*:** обучающемуся даётся два билета с одним вопросом и задаётся несколько дополнительных вопросов по курсу. Ответы на вопросы экзамена оцениваются по шкале от 0 (обучающийся не может ответить на вопрос) до 100 (ответ на вопрос является исчерпывающим) с последующим усреднением.

3.1.5 Методические материалы для оценки обучающимися содержания и качества учебного процесса

Для оценки обучающимися содержания и качества учебного процесса может применяться анкетирование в соответствии с методикой и графиком, утвержденными в установленном порядке.

3.2. Кадровое обеспечение

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

К чтению лекций должны привлекаться преподаватели, имеющие ученую степень доктора или кандидата наук (в том числе степень PhD, прошедшую установленную процедуру признания и установления эквивалентности) и/или ученое звание профессора или доцента.

3.2.2 Обеспечение учебно-вспомогательным и (или) иным персоналом

Учебно-вспомогательный и инженерно-технический персонал должен иметь соответствующее образование и обладать навыками организации работы с пользовательскими программными продуктами в локальной сети компьютерного класса и в Интернете.

3.3. Материально-техническое обеспечение

3.3.1 Характеристики аудиторий (помещений, мест) для проведения занятий

Учебные аудитории для проведения учебных занятий, оснащенные стандартным оборудованием, используемым для обучения в СПбГУ в соответствии с требованиями материально-технического обеспечения.

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

В аудиториях, где проводятся лекционные занятия, необходимо наличие досок и средств письма на них. Для показа слайдов необходим компьютер с установленным программным обеспечением для работы со слайдами в форматах PDF, PPT, PPTX и подключенный к нему мультимедийный проектор с экраном.

3.3.3 Характеристики специализированного оборудования

Нет.

3.3.4 Характеристики специализированного программного обеспечения

В рамках изучения дисциплины выполнения практических заданий обучающимся могут потребоваться средства, среды разработок для языков программирования Python 3, R.

3.3.5 Перечень и объёмы требуемых расходных материалов

Для аудиторий с маркерными досками необходимы стирающиеся маркеры в объёме, достаточном для проведения дисциплины. Для аудиторий с меловыми досками необходим мел в объёме, достаточном для проведения дисциплины. Канцелярские принадлежности в объёме, достаточном для проведения дисциплины.

3.4. Информационное обеспечение

3.4.1 Список литературы

1. Голицына, О. Л. Информационные системы и технологии : учебное пособие / О.Л. Голицына, Н.В. Максимов, И.И. Попов. — Москва : ФОРУМ : ИНФРА-М, 2021. — 400 с. – ISBN 978-5-00091-592-9. ЭБС Знаниум по подписке СПбГУ:
https://proxy.library.spbu.ru/login?url=http://new.znanium.com/bookread2.php?book=374442

2. Ингерсолл, Г. С. Обработка неструктурированных текстов. Поиск, организация и манипулирование / Г. С. Ингерсолл, Т. С. Мортон, Э. Л. Фэррис. — Москва : ДМК Пресс, 2015. — 414 с. — ISBN 978-5-97060-144-0. ЭБС ЛАНЬ по подписке СПбГУ: https://proxy.library.spbu.ru/login?url=https://e.lanbook.com/reader/book/73069#1

3 Основы Data Science и Big Data. Python и наука о данных С. Дэви, М. Арно, А. Мохамед. 2017. СПб: Питер. 336 с. ISBN: 9785496025171 ЭБС «Айбукс» по подписке СПбГУ:
https://proxy.library.spbu.ru/login?url=http://ibooks.ru/reading.php?productid=354390

3.4.2 Перечень иных информационных источников, в том числе современных профессиональных баз данных и информационных справочных систем

Электронные ресурсы Научной библиотеки им. М. Горького СПбГУ

Сайт Научной библиотеки им. М. Горького СПбГУ:
http://www.library.spbu.ru/

Электронный каталог Научной библиотеки им. М. Горького СПбГУ:
http://www.library.spbu.ru/cgi-bin/irbis64r/cgiirbis_64.exe?C21COM=F&I21DBN=IBIS&P21DBN=IBIS

Перечень электронных ресурсов, находящихся в доступе СПбГУ:
http://cufts.library.spbu.ru/CRDB/SPBGU/

Перечень ЭБС, на платформах которых представлены российские учебники, находящиеся в доступе СПбГУ:
http://cufts.library.spbu.ru/CRDB/SPBGU/browse?name=rures&resource_type=8

Математика: тематическая рубрика

http://cufts.library.spbu.ru/CRDB/SPBGU/browse?subject=1

Информатика: тематическая рубрика
http://cufts.library.spbu.ru/CRDB/SPBGU/browse?subject=93

Раздел 4. Разработчики программы

Фамилия, имя, отчество	Учёная степень	Учёное звание	Должность	Контактная информация
Тулупьев Александр Львович	д.ф.-м.н,	профессор	профессор.	a.tulupyev@spbu.ru alt@dscs.pro +7 (931) 288-31-77
Тулупьева Татьяна Валентиновна	к.псих.н.	доцент	доцент	t.tulupyeva@spbu.ru tvt@dscs.pro +7(921)753-54-88
Абрамов Максим Викторович	к.т.н.		доцент	m.abramov@spbu.ru mva@dscs.pro +7(981) 680-99-29
Корепанова Анастасия Андреевна				aak@dscs.pro