Министерство образования и науки Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Кубанский государственный университет»

Факультет компьютерных технологий и прикладной математики
Кафедра вычислительных технологий

УТВЕРЖДАЮ:
Проректор по учебной работе, качеству образования – первый проректор
_____ Хагуров Т.А.
подпись
«27» 04 2018 г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ
Б1.В.ДВ.01.02 «ОБРАБОТКА ЕСТЕСТВЕННО-ЯЗЫКОВЫХ ТЕКСТОВ»

Направление подготовки/специальность 02.03.02 **Фундаментальная информатика и информационные технологии**
(код и наименование направления подготовки/специальности)

Направленность (профиль) / специализация Вычислительные технологии
(наименование направленности (профиля) специализации)

Программа подготовки академический бакалавриат
(академическая /прикладная)

Форма обучения очная
(очная, очно-заочная, заочная)

Квалификация (степень) выпускника бакалавр
(бакалавр, магистр, специалист)

Краснодар 2018

Рабочая программа дисциплины Б1.В.ДВ.01.02 «Обработка естественно-языковых текстов» составлена в соответствии с федеральным государственным образовательным стандартом высшего образования (ФГОС ВО) по направлению подготовки 02.03.02 Фундаментальная информатика и информационные технологии

Программу составил(а):

Приходько Татьяна Александровна, доцент, к. т. н. _____
Ф.И.О. , должность, ученая степень, ученое звание подпись

Рабочая программа дисциплины Б1.В.ДВ.01.02 «Обработка естественно-языковых текстов » утверждена на заседании кафедры Вычислительных Технологий протокол № 7 «03 » апреля 2018 г.

Заведующий кафедрой (разработчика) Миков А. И. _____
фамилия, инициалы подпись

Рабочая программа обсуждена на заседании кафедры Вычислительных Технологий протокол № 7 «03 » апреля 2018 г.

Заведующий кафедрой (выпускающей) Миков А. И _____
фамилия, инициалы подпись

Утверждена на заседании учебно-методической комиссии факультета Компьютерных Технологий и Прикладной Математики протокол № 1 от «20» апреля 2018 г

Председатель УМК факультета Малыхин К. В. _____
фамилия, инициалы подпись

Рецензенты:

Гаркуша О.В., доцент кафедры информационных технологий ФБГОУ ВО «Кубанский государственный университет», кандидат физико-математических наук.

Зайков В.П. Ректор НЧОУ ВО «Кубанский институт информзащиты» д.экон. наук, к.т.н., доцент.

1. ЦЕЛИ И ЗАДАЧИ ОСВОЕНИЯ ДИСЦИПЛИНЫ

1.1 Цель освоения дисциплины

Целью дисциплины «Обработка естественно-языковых текстов» является обучение передовым методам, моделям, средствам и технологиям компьютерной обработки текстов на естественных языках.

1.2. Задачи дисциплины

Основными задачами при этом являются:

- получение теоретических знаний и практических навыков обработки естественно-языковых текстов;
- знание сложностей, связанных с применением существующих методов обработки естественно-языковых текстов;
- умение использовать полученные знания по разработке, адаптации и использованию новейших средств информатики для обработки текстов на естественных языках.

1.3 Место дисциплины в структуре образовательной программы

Дисциплина «Обработка естественно-языковых текстов» относится к вариативной части блока Б1 дисциплин бакалавриата. Для изучения дисциплины необходимо знание основ объектно-ориентированного проектирования и программирования, операционных систем, компьютерных сетей, баз данных, нечеткой логики, нейронных сетей и др. методов ИИ.

Знания, получаемые при изучении технологий обработки естественно-языковых текстов, используются при изучении других дисциплин учебного плана бакалавриата, а также при работе над выпускной работой студента.

1.4 Перечень планируемых результатов обучения по дисциплине (модулю), соотнесенных с планируемыми результатами освоения образовательной программы.

Изучение данной учебной дисциплины направлено на формирование у обучающихся следующих **профессиональных компетенций:**

Код компетенции	Формулировка компетенции
ПК-8	способностью применять на практике международные и профессиональные стандарты информационных технологий, современные парадигмы и методологии, инструментальные и вычислительные средства
ОК-5	способностью к коммуникации в устной и письменной формах на русском и иностранном языках для решения задач межличностного и межкультурного взаимодействия

Таблица 1. Профессиональные компетенции студента

Компетенция	знать	уметь	владеть

ПК-8	Международные и профессиональные стандарты информационных технологий, современные парадигмы и методологии, инструментальные и вычислительные средства	Применять на практике международные и профессиональные стандарты информационных технологий, современные парадигмы и методологии, инструментальные и вычислительные средства	Способностью применять на практике международные и профессиональные стандарты информационных технологий, современные парадигмы и методологии, инструментальные и вычислительные средства
ОК-5	Способы налаживания профессионального взаимодействия в устной и письменной формах на русском и иностранном языках для решения задач межличностного и межкультурного взаимодействия	Устанавливать коммуникации профессиональной сфере деятельности в устной и письменной формах на русском и иностранном языках для решения задач межличностного и межкультурного взаимодействия	Способностью к коммуникации в устной и письменной формах на русском и иностранном языках для решения задач межличностного и межкультурного взаимодействия

2. СТРУКТУРА И СОДЕРЖАНИЕ ДИСЦИПЛИНЫ

2.1 Распределение трудоемкости дисциплины по видам работ

Общая трудоемкость дисциплины составляет 5 зач.ед. (180 часов), их распределение по видам работ представлено в таблице *(для студентов ОФО)*

Вид учебной работы	Всего часов	Семестры (часы)			
		7			
Контактная работа в том числе:					
Аудиторные занятия (всего):	96,3	96,3			
В том числе:					
Занятия лекционного типа	36	36			
Занятия семинарского типа (семинары, практ. занятия)					
Лабораторные занятия	54	54			
Иная контрольная работа					
Контроль самостоятельной работы	6	6			
Промежуточная аттестация (ИКР)	0,3	0,3			
Самостоятельная работа (всего)					
В том числе:					
Курсовая работа					
Проработка учебного (теоретического) материала	20	20			
Выполнение индивидуальных заданий (подготовка сообщений, презентаций)	12	12			
Реферат					
Подготовка к текущему контролю	16	16			

Контроль:						
Подготовка к экзамену:		35,7	35,7			
Общая трудоемкость	час	180	180			
	в т.ч. контактная работа	96,3	96,3			
	зач. ед.	5	5			

2.2 Структура дисциплины:

Распределение видов учебной работы и их трудоемкости по разделам дисциплины.
Разделы дисциплины, изучаемые в _7___семестре *(очная форма)*

№	Наименование разделов	Количество часов				
		Всего	Аудиторная работа			Внеаудиторная работа
			Л	КСР	ЛР	СРС
1	2	3	4	5	6	7
1.	Раздел 1. Введение в обработку естественно-языковых текстов. Разновидности языковых групп и особенности их обработки.	42	8	2	16	16
2.	Раздел 2. Методы обработки естественных языков. Нормализация, лематизация, стемминг.	46	12	2	16	16
3.	Раздел 3. Программирование и проектирование систем обработки естественных языков.	56	16	2	22	16
	Итого:	144	36	6	54	48
	Контроль	35,7				
	ИКР	0,3				
	Итого по дисциплине:	180				

Примечание: Л – лекции, КСР – контрольные и самостоятельные работы, ЛР – лабораторные занятия, СРС – самостоятельная работа студента

2.3 Содержание разделов дисциплины:

2.3.1 Занятия лекционного типа

№ раз-дела	Наименование раздела	Содержание раздела	Форма текущего контроля
1	2	3	4
1	Раздел 1. Введение в обработку естественно-языковых текстов. Разновидности языковых групп и особенности их обработки.	Лингвистика как наука о языке. Представление об уровнях представления языка – фонетика, морфология, синтаксис, семантика. Лингвистика и прагматика. Лингвистическое моделирование. Действующие модели языка. Теория «Смысл – Текст» как фундамент для построения систем автоматической обработки текста.	ЛР

2	Раздел 2. Методы обработки естественных языков..	Анализ и синтез текста. Морфологический и синтаксический анализ. Парсинг. Различные подходы к синтаксическому анализу: анализ «сверху вниз» и «снизу вверх». Языковая неоднозначность как принципиальное свойство языка и методы ее разрешения при автоматической обработке текста. Интерактивное разрешение лексической и синтаксической неоднозначности. Правиловые и статистические подходы к автоматической обработке текста. Алгоритм синтаксического анализа. Синтаксические отношения. Синтагмы. Синтаксическая структура предложения.	
3	Раздел 3. Программирование и проектирование систем обработки естественных языков.	Задачи морфологического анализа, морфологический разбор, стемминг, лемматизация. Понятия лексемы, словоформы, леммы, морфемы, псевдо-основы и псевдо-окончания. Грамматические категории. Словоизменительная парадигма. Морфотактика. Структура данных морфологического словаря, лексикона. Грамматические модели русского языка в контексте автоматической обработки. Минимальное расстояние редактирования. Алгоритм подсчета расстояния Левенштейна. Практика по подсчету минимального расстояния Левенштейна. Понятие статистической языковой модели. Области применения. N-грамы.	ЛР

2.3.3 Лабораторные занятия

№ работы	№ раздела дисциплины	Наименование лабораторных работ
1	1	Понятие токенизации. Основные подходы. Автоматические разбиение текста на предложения и другие структурные блоки. Основные проблемы и методы. Особенности сегментации русскоязычных текстов.
2	1	Разработка сегментатора текста. Работа с UIMA CAS Editor. Создание небольшого тестового документа Написание программы оценки качества сегментатора.
3	2	Простая модель n-грам без сглаживания и еѐ недостатки. Подсчет частоты слов в корпусе. Оценка по методу максимального правдоподобия.

4-6	2	Задачи автоматического снятия морфологической неоднозначности: постановка и применение. Подход, основанный на правилах. Подход, основанный на скрытой марковской модели. Методы оценивания параметров скрытой марковской модели. Алгоритм Витерби. Подход, основанный на трансформациях, получаемых автоматически. Тэггер Брилла. Особенности решения задачи для русского языка
7	3	Нюансы программной реализации языковых моделей. Обзор программных инструментариев для реализации и применения языковых моделей. Сбор экспериментального корпуса. Практическая работа с инструментарием SRILM. Консультация по задачам, заданными для самостоятельной работы.
8	3	Эксперименты по автоматическому снятию морфологической неоднозначности. Корпус OpenCorpora. Практика с применением программных компонент Apache UIMA.

2.3.4 Примерная тематика курсовых работ (проектов)

Учебным планом не предусмотрены.

2.3.5 Расчетно-графические задания

Учебным планом не предусмотрены.

2.4 Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине

№	Вид СРС	Перечень учебно-методического обеспечения дисциплины по выполнению самостоятельной работы
1	2	3
1	Методы хранения словарей.	Источники основной и дополнительной литературы
2	Машинный перевод и другие прикладные задачи компьютерной лингвистики.	Источники основной и дополнительной литературы
3	Методы машинного обучения в задачах прикладной лингвистики.	Источники основной и дополнительной литературы

Учебно-методические материалы для самостоятельной работы обучающихся из числа инвалидов и лиц с ограниченными возможностями здоровья (ОВЗ) предоставляются в формах, адаптированных к ограничениям их здоровья и восприятия информации:

Для лиц с нарушениями зрения:

– в печатной форме увеличенным шрифтом,

– в форме электронного документа,

Для лиц с нарушениями слуха:

– в печатной форме,

– в форме электронного документа.

Для лиц с нарушениями опорно-двигательного аппарата:

– в печатной форме,

– в форме электронного документа,

Данный перечень может быть конкретизирован в зависимости от контингента обучающихся.

3. ОБРАЗОВАТЕЛЬНЫЕ ТЕХНОЛОГИИ

Семестр	Вид занятия (Л, ПР, ЛР)	Используемые интерактивные образовательные технологии	Количество часов
7	Л	Компьютерные презентации и обсуждение	36
	ЛР	Разбор конкретных ситуаций (задач), тренинги по решению задач, компьютерные симуляции (программирование алгоритмов)	54
7	КРС	Контрольная работа	6
Итого:			96

Для лиц с ограниченными возможностями здоровья предусмотрена организация консультаций с использованием электронной почты.

4. ОЦЕНОЧНЫЕ СРЕДСТВА ДЛЯ ТЕКУЩЕГО КОНТРОЛЯ УСПЕВАЕМОСТИ И ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ

4.1 Фонд оценочных средств для проведения текущего контроля

Фонд оценочных средств дисциплины состоит из средств текущего контроля выполнения лабораторных работ, контрольной работы, средств для итоговой аттестации (экзамена в 7 семестре).

Оценка успеваемости осуществляется по результатам:

- выполнения лабораторных работ - компьютерных программ, сопровождаемой вопросами по теоретической части предмета;

-контрольной работы;

- ответа на экзамене (для выявления знания и понимания теоретического материала дисциплины).

Текущий контроль включает контрольную работу по итогам первой половины курса.

Пример задания для контрольной работы:

1. Перечислите трудности автоматизации обработки естественного языка в интеллектуальных системах.
2. Разработайте программу синтаксического анализа текста.

Перечень вопросов, которые выносятся на экзамен в 7 семестре

3. Трудности автоматизации обработки естественного языка в интеллектуальных системах.
4. Этапы анализа предложений на естественном языке.
5. Синтаксически-ориентированный и семантически-ориентированный анализ ЕЯ.
6. Основные методы анализа ЕЯ: шаблоны.
7. Основные методы анализа ЕЯ: семантические грамматики.
8. Основные методы анализа ЕЯ: падежные фреймы.

9. Деревья анализа и свободно-контекстные грамматики.
10. Синтаксический анализ естественного языка.
11. Основные подходы к решению задач в интеллектуальных системах. Поиск в пространстве состояний.
12. Основные подходы к решению задач в интеллектуальных системах – логический вывод.
13. Основные подходы к решению задач в интеллектуальных системах – сопоставление с образцом и ассоциативный поиск.
14. Психолингвистический подход к анализу естественного языка.
15. Интеллектуальные системы, использующие естественный язык.
16. Задачи морфологического анализа, морфологический разбор, стемминг, лемматизация.
17. Понятия лексемы, словоформы, леммы, морфемы, псевдо-основы и псевдо-окончания.
18. Грамматические категории. Словоизменительная парадигма. Морфотактика. Структура данных морфологического словаря, лексикона.
19. Грамматические модели русского языка в контексте автоматической обработки.
20. Минимальное расстояние редактирования. Алгоритм подсчета расстояния Левенштейна.
21. Понятие статистической языковой модели. Области применения. N-грамы.

Примеры экзаменационных билетов

Экзаменационный билет № _

1. Перечислите этапы анализа предложений на естественном языке.
2. Минимальное расстояние редактирования. Алгоритм подсчета расстояния Левенштейна.
3. Разработать упрощенную программу стемминга.

Критерии оценивания к экзамену:

- 84-100 баллов (оценка «отлично») - изложенный материал фактически верен, наличие глубоких исчерпывающих знаний в объеме пройденной программы дисциплины в соответствии с поставленными программой курса целями и задачами обучения; правильные, уверенные действия по применению полученных знаний на практике, грамотное и логически стройное изложение материала при ответе, усвоение основной и знакомство с дополнительной литературой; Практические задания выполнены в срок и в полном объеме.

- 67-83 баллов (оценка «хорошо») - наличие твердых и достаточно полных знаний в объеме пройденной программы дисциплины в соответствии с целями обучения, правильные действия по применению знаний на практике, четкое изложение материала, допускаются отдельные логические и стилистические погрешности. Практические задания выполнены в срок в объеме не менее 80%.

- 50-66 баллов (оценка удовлетворительно) - наличие твердых знаний в объеме пройденного курса в соответствии с целями обучения, изложение ответов с отдельными ошибками, уверенно исправленными после дополнительных вопросов; правильные в целом действия по применению знаний на практике; Практические задания выполнены в объеме не менее 60%.

- 0-49 баллов (оценка неудовлетворительно) - ответы не связаны с вопросами, наличие грубых ошибок в ответе, непонимание сущности излагаемого вопроса, неумение применять знания на практике, неуверенность и неточность ответов на дополнительные и наводящие вопросы». Практические задания выполнены в объеме менее 50%.

Оценочные средства для инвалидов и лиц с ограниченными возможностями здоровья выбираются с учетом их индивидуальных психофизических особенностей.

– при необходимости инвалидам и лицам с ограниченными возможностями здоровья

предоставляется дополнительное время для подготовки ответа на экзамене;

– при проведении процедуры оценивания результатов обучения инвалидов и лиц с ограниченными возможностями здоровья предусматривается использование технических средств, необходимых им в связи с их индивидуальными особенностями;

– при необходимости для обучающихся с ограниченными возможностями здоровья и инвалидов процедура оценивания результатов обучения по дисциплине может проводиться в несколько этапов.

Процедура оценивания результатов обучения инвалидов и лиц с ограниченными возможностями здоровья по дисциплине предусматривает предоставление информации в формах, адаптированных к ограничениям их здоровья и восприятия информации:

Для лиц с нарушениями зрения:

– в печатной форме увеличенным шрифтом,

– в форме электронного документа.

Для лиц с нарушениями слуха:

– в печатной форме,

– в форме электронного документа.

Для лиц с нарушениями опорно-двигательного аппарата:

– в печатной форме,

– в форме электронного документа.

Данный перечень может быть конкретизирован в зависимости от контингента обучающихся.

5. ПЕРЕЧЕНЬ ОСНОВНОЙ И ДОПОЛНИТЕЛЬНОЙ УЧЕБНОЙ ЛИТЕРАТУРЫ, НЕОБХОДИМОЙ ДЛЯ ОСВОЕНИЯ ДИСЦИПЛИНЫ

5.1 Основная литература:

1. Павлов, С.И. Системы искусственного интеллекта : учебное пособие / С.И. Павлов. - Томск : Томский государственный университет систем управления и радиоэлектроники, 2011. - Ч. 2. - 194 с. - ISBN 978-5-4332-0014-2 ; То же [Электронный ресурс]. - URL: http://biblioclub.ru/index.php?page=book&id=208939
2. Кокорина, И.В. Основы математической обработки информации в филологии: комбинаторика, теория вероятностей и математическая статистика : учебно-методическое пособие / И.В. Кокорина ; Министерство образования и науки Российской Федерации, Федеральное государственное автономное образовательное учреждение высшего профессионального образования Северный (Арктический) федеральный университет им. М.В. Ломоносова. - Архангельск : ИД САФУ, 2014. - 115 с. : ил. - Библиогр. в кн. - ISBN 978-5-261-00928-3 ; То же [Электронный ресурс]. - URL: http://biblioclub.ru/index.php?page=book&id=312317

5.2. Дополнительная литература

1. Костюкова, Н.И. Комбинаторные алгоритмы для программистов / Н.И. Костюкова. - 2-е изд/, исправ/. - Москва : Национальный Открытый Университет «ИНТУИТ», 2016. - 217 с. : ил. ; То же [Электронный ресурс]. - URL: http://biblioclub.ru/index.php?page=book&id=429067
2. Гуслякова, А.В. Информационные технологии и лингвистика XXI века : учебное пособие / А.В. Гуслякова ; Министерство образования и науки Российской Федерации. - Москва : МПГУ, 2016. - 96 с. : ил. - Библиогр. в кн. - ISBN 978-5-4263-0398-0 ; То же [Электронный ресурс]. - URL: http://biblioclub.ru/index.php?page=book&id=469675

5.3. Интернет - ресурсы

1. Автоматическая обработка текста (материалы сайта). URL: http://www.aot.ru.
2. Национальный корпус русского языка. URL: http://www.ruscorpora.ru

3. Аношкина Ж.Г. Словарь омонимичных словоформ русского языка. М: Машинный фонд русского языка Института русского языка РАН, 2001.
4. Апресян Ю. Д., Богуславский И. М., Иомдин Б. Л. и др. Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы. Национальный корпус русского языка: 2003-2005. М.: Индрик, 2005. http://ruscorpora.ru/sbornik2005/12apresyan.pdf
5. Кобзарева Т. Ю., Афанасьев Р. Ю. Универсальный модуль предсинтаксического анализа омонимии частей речи в русском языке на основе словаря диагностических ситуаций. "Компьютерная лингвистика и интеллектуальные технологии".

Для освоения дисциплины инвалидами и лицами с ограниченными возможностями здоровья имеются издания в электронном виде в электронно-библиотечных системах

1. ЭБС Издательства «Лань» http://e.lanbook.com ,
2. ЭБС «Университетская библиотека онлайн» www.biblioclub.ru ,
3. ЭБС «Юрайт» http://www.biblio-online.ru ,
4. ЭБС «ZNANIUM.COM» www.znanium.com,
5. ЭБС «BOOK.ru» https://www.book.ru.

Методические указания для обучающихся по освоению дисциплины

По курсу предусмотрено проведение лекционных занятий, на которых дается основной систематизированный материал для получения теоретических сведений, для выполнения лабораторных работ и подготовки к экзамену.

Важнейшим этапомкурса является самостоятельная работа по дисциплине с использованием указанных литературных источников и методических указаний автора курса.

Виды и формы СР, сроки выполнения, формы контроля приведены выше в данном документе.

Для лучшего освоения дисциплины при защите ЛР студент должен ответить на несколько вопросов из лекционной части курса.

В освоении дисциплины инвалидами и лицами с ограниченными возможностями здоровья большое значение имеет индивидуальная учебная работа (консультации) – дополнительное разъяснение учебного материала.

Индивидуальные консультации по предмету являются важным фактором, способствующим индивидуализации обучения и установлению воспитательного контакта между преподавателем и обучающимся инвалидом или лицом с ограниченными возможностями здоровья.

6. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю)

6.1 Перечень информационных технологий.

– Проверка домашних заданий и консультирование посредством электронной почты.

– Использование электронных презентаций при проведении лекций и практических занятий.

6.2 Перечень необходимого программного обеспечения

Программное обеспечение

1. OS Windows, MS Office
2. NetBeans+ MPJ (JAVA)
3. Программы для демонстрации и создания презентаций («Microsoft Power Point»).

3.1 Перечень информационных справочных систем:

1. Электронная библиотечная система eLIBRARY.RU (http://www.elibrary.ru)/

8. Материально-техническая база, необходимая для осуществления образовательного процесса по дисциплине

№	Вид работ	Материально-техническое обеспечение дисциплины (модуля) и оснащенность
1.	Лекционные занятия	Лекционная аудитория, оснащенная презентационной техникой (проектор, экран, компьютер/ноутбук) и соответствующим программным обеспечением (ПО) PowerPoint. ауд. 129, 131, А305.
2.	Лабораторные занятия	Лаборатория, укомплектованная специализированными техническими средствами обучения – компьютерный класс, с возможностью подключения к сети «Интернет», программой экранного увеличения и обеспеченный доступом в электронную информационно-образовательную среду университета. (лаб. 102-106.).
3.	Групповые (индивидуальные) консультации	Аудитория, (кабинет) – компьютерный класс
4.	Текущий контроль, промежуточная аттестация	Аудитория, приспособленная для письменного ответа при промежуточной аттестации.
5.	Самостоятельная работа, контрольная работа	Кабинет для самостоятельной работы, оснащенный компьютерной техникой с возможностью подключения к сети «Интернет», программой экранного увеличения и обеспеченный доступом в электронную информационно-образовательную среду университета.