Белорусский государственный университет

УТВЕРЖДАЮ
Проректор по учебной работе и
образовательным инновациям
_____ О.Н.Здрок
«30» ноября 2020 г.

Регистрационный № УД-9248/уч.

ТЕХНОЛОГИИ ОБРАБОТКИ ТЕКСТОВ

Учебная программа учреждения высшего образования по учебной дисциплине для специальности

1-31 80 09 Прикладная математика и информатика

Профилизация: Интеллектуальные системы

2020 г.

Учебная программа составлена на основе ОСВО 1-31 80 09-2019 и учебного плана G31-128/уч. от 11.04.2019 г.

СОСТАВИТЕЛИ:

Н.К. Рубашко – старший преподаватель кафедры информационных систем управления факультета прикладной математики и информатики Белорусского государственного университета

РЕЦЕНЗЕНТЫ:

А.А. Дудкин – заведующий лабораторией идентификации систем ОИПИ НАН Беларуси,доктор технических наук, профессор.

РЕКОМЕНДОВАНА К УТВЕРЖДЕНИЮ:

Кафедрой информационных систем управления
(протокол № 9 от 20 марта 2020 года);

Научно-методическим Советом БГУ

(протокол № 4 от 25 марта 2020 года).

Заведующий кафедрой
информационных систем управления _____ В.В. Краснопрошин

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Цели и задачи учебной дисциплины

Учебная дисциплина «Технологии обработки текстов» знакомит студентов магистратуры с теоретическими основами анализа и разработки методов, алгоритмов и технологий для обработки текстов естественного языка (ЕЯ), дает основы фундаментальной и прикладной подготовки в области автоматической обработки текста с целью решения широкого круга актуальных задач, так или иначе связанных с документооборотом, автоматизацией инженерии знаний и построения основанных на них инновационных решений, прежде всего, в виде текстовых документов, в том числе и в социальных сетях, автоматизации принятия решений.

Дисциплина базируется на современных достижениях в области информационных технологий и ориентирована на решение прикладных задач обработки текстов на естественном языке.

Цель учебной дисциплины – дальнейшее развитие у студентов магистратуры умений и навыков в области использования компьютерных технологий для обработки текстов естественного языка.

Задачи учебной дисциплины:

1. формирование компетентности в области использования возможностей современных компьютерных технологий для решения как теоретических, так и практических задач обработки текстов;
2. освоение практических методов обработки и анализа текста, повышения эффективности человеко-машинного взаимодействия.

Место учебной дисциплины в системе подготовки специалиста с высшим образованием (магистра).

Учебная дисциплина относится **к модулю** «Анализ описательной информации» компонента учреждения высшего образования.

Программа составлена с учетом **межпредметных связей** с учебными дисциплинами. Основой для изучения являются учебные дисциплины первой ступени высшего образования по дискретной математике и математической логике, теории вероятностей и математической статистике, теории алгоритмов и программированию и дисциплина второй ступени высшего образования «Компьютерная лингвистка».

Требования к компетенциям

Освоение учебной дисциплины «Технологии обработки текстов» должно обеспечить формирование следующих специализированных и углубленных профессиональных компетенций:

специализированные компетенции:

СК–12. Владеть основными подходами к разработке эффективных алгоритмов обработки текстов и построению индексных структур для коллекций текстовых документов.

СК–13. Уметь использовать научные и технические достижения для разработки эффективных алгоритмов решения прикладных задач.

углубленные профессиональные компетенции:

УПК–4. Оценивать эффективность алгоритмов решения прикладных задач.

В результате изучения дисциплины студент магистратуры должен:

знать:

– место и роль естественного языка в современных информационных технологиях;

– методы анализа языкового материала;

– общую технологию решения задач автоматической обработки текста;

уметь:

– использовать технологию автоматической обработки текстовой информации для анализа естественного языка;

– реализовывать различные алгоритмы обработки естественного языка для решения прикладных задач;

– разрабатывать, в том числе с использованием существующих стандартных средств, программное обеспечение систем автоматической обработки текста;

владеть:

– основными методами и приемами исследовательской и практической работы в области обработки естественного языка;

– методикой использования компьютерных технологий при обработке текста;

– технологией разработки прикладных систем автоматической обработки текста, от постановки задачи до создания программного образца.

Структура учебной дисциплины

Дисциплина изучается во 2 семестре. Всего на изучение учебной дисциплины «Технологии обработки текстов» отведено:

– для очной формы получения высшего образования – 126 часов, в том числе 40 аудиторных часов, из них: лекции – 20 часов, практические занятия – 20 часов.

Трудоемкость учебной дисциплины составляет 3 зачетные единицы.

Форма текущей аттестации по учебной дисциплине – зачет.

СОДЕРЖАНИЕ УЧЕБНОГО МАТЕРИАЛА

Раздел 1. Краткое введение в проблематику

Тема 1.1. Технологии обработки естественного языка в науке и промышленности

Основы обработки неструктурированных (текстовых) данных в информационных системах и современных веб-приложениях. Ввод речи (текста) в компьютер. Человеко-компьютерное взаимодействие

Раздел 2. Инструментальные системы разработки приложений по автоматической обработке текстов на естественном языке

Тема 2.1. Представление лингвистических данных

Подходы к представлению данных. Лингвистическая разметка. Лингвистические аннотации. Представления, основанные на абстракции. Недоспецифицированные представления.

Тема 2.2. Архитектура инструментальных ЕЯ-систем

Компонентная организация. Процессы обработки текста

Тема 2.3. Системы обработки ЕЯ-текстов

Системы на базе разметки. Системы на базе аннотаций. Системы интеграции поверхностной и глубокой обработки. Системы, развивающие отдельные аспекты обработки текста

Раздел 3. Векторное представление текста

Тема 3.1. Моделирование языка

Подходы к моделированию языка и обучению представлений в обработке естественного языка. Анализ и сравнение моделей векторного представления слов для различных конечных задач обработки естественного языка.

Тема 3.2. Методы векторного представления

Обзор методов изменения векторных пространств и их применения для решения прикладных лингвистических задач . Нейронные сети, методы снижения размерности. Использование векторных представлений слов и фраз для улучшения качества работы методов автоматической обработки естественного языка.

.***Тема 3.3. Задачи, решаемые с помощью векторного представления слов***

Использование векторного представления слов (текста) для расшифровки акронимов . Подбор синонимов при помощи векторных представлений слов. Исправление опечаток с использованием векторных

представлений слов. Поиск при помощи векторных представлений слов в базе вопросов и ответов.Автоматическое обнаружение токсичных комментариев.

Раздел 4. Машинное обучение

Тема 4.1. Задача машинного обучения

Понятие машинного обучения в искусственном интеллекте. Классификация задач машинного обучения. Открытые наборы данных для обучения. Глубокое обучение в обработке текстов. Модели глубокого обучения для определения смысла слов. Применение задач машинного обучения с их использованием на стадии предобучения с целью повышения качества векторного представления текстовых документов.

Тема 4.2. Алгоритмы машинного обучения

Линейная регрессия, Логистическая регрессия. Линейный дискриминантный анализ. Деревья принятия решений. Метод опорных векторов. Методы аугментации тренировочного корпуса ограниченного объёма для решения прикладных лингвистических задач.

УЧЕБНО-МЕТОДИЧЕСКАЯ КАРТА УЧЕБНОЙ ДИСЦИПЛИНЫ

Дневная форма получения образования с применением дистанционных образовательных технологий

Номер раздела, темы	Название раздела, темы	Количество аудиторных часов			Форма контроля знаний
		Лекции	Практические занятия	Количество часов УСР	
1	2	3	4	5	6
1	**КРАТКОЕ ВВЕДЕНИЕ В ПРОБЛЕМАТИКУ**	**2**			Устный опрос.
1.1	Технологии обработки естественного языка в науке и промышленности	2			
2	**ИНСТРУМЕНТАЛЬНЫЕ СИСТЕМЫ РАЗРАБОТКИ ПРИЛОЖЕНИЙ ПО АВТОМАТИЧЕСКОЙ ОБРАБОТКЕ ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ**	**6**	**8**		Доклад. Отчеты по домашним практическим заданиям с их устной защитой. Выполнение тестов. Контрольная работа 1
2.1	Представление лингвистических данных	2	4		
2.2	Архитектура инструментальных ЕЯ-систем	2	2		
2.3	Системы обработки ЕЯ-текстов	2	2		
3	**ВЕКТОРНОЕ ПРЕДСТАВЛЕНИЕ ТЕКСТА**	**6**	**6**		Доклад. Отчеты по домашним практическим заданиям с их устной защитой. Выполнение тестов.
3.1	Моделирование языка	2	2		
3.2	Методы векторного представления	2	2		
3.3	Задачи, решаемые с помощью векторного представления слов	2	2		
4	**МАШИННОЕ ОБУЧЕНИЕ**	**6**	**6**		Устный опрос. Отчет по лабораторным работам. Выполнение тестов. Контрольная работа 2
3.1	Задача машинного обучения	4	2		
3.2	Алгоритмы машинного обучения	2	4		
	ВСЕГО:	**20**	**20**		

ИНФОРМАЦИОННО-МЕТОДИЧЕСКАЯ ЧАСТЬ

Перечень основной литературы

1. Хобсон Лейн, Ханнес Макс Хапке, Коул Ховард. Обработка естественного языка в действии – С.Пб: Издательский дом «Питер», 2020. – 576 с.
2. Николенко С. И., Кадурин А. А., Архангельская Е. О. Глубокое обучение – С.Пб: Издательский дом «Питер», 2021. – 480 с.
3. И.В. Совпель. Инженерно-лингвистические принципы, методы и алгоритмы автоматической переработки текста. – Мн., Вышэйшая школа, 1991.
4. Автоматическая обработка текста на естественном языке и компьютерная лингвистика: учебное пособие / Большакова Е.И., Клышинский Э.С., Ландэ Д.В. , Носков А.А., Песков О.В., Ягунова Е.В. – М.: МИЭМ, 2011.
5. Jurafsky, D. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition / D. Jurafsky, J. H. Martin. – New Jersey: Prentice Hall PTR, 2000. – 934 p.

Перечень дополнительной литературы

1. Всеволодова А.В.Компьютерная обработка лингвистических данных. Изд.2 2007. 96 с.
2. Ландэ Д.В., Снарский А.А., Безсуднов И.В. Интернетика: Навигация в сложных сетях: модели и алгоритмы, 2009.
3. Fastus: A cascaded finite-state transducer for extracting information from natural-language text / D. Israel [et al.] // Finite State Devices for Natural Language Processing / ed. by Roche, Schabes. – Cambridge, MA, USA: MIT Press, 1996. – P. 383–406.
4. Технологии Яндекса. – Режим доступа: https://yandex.ru/dev/
5. Проект АОТ (Автоматическая Обработка Текста). – Режим доступа: https://AOT.ru
6. Nooj – программное обеспечение для среды лингвистической разработки, – Режим доступа: http://www.nooj-association.org/

Перечень рекомендуемых средств диагностики и методика формирования итоговой оценки

Для диагностики компетенции в рамках учебной дисциплины рекомендуется использовать следующие формы:

1. Устная форма: устный опрос, коллоквиум, выступление с докладом на семинаре.
2. Письменная форма: контрольные работы.
3. Устно-письменная форма: отчеты по домашним практическим заданиям с их устной защитой, оценивание на основе проектного метода, выполнение тестов.

Формой текущей аттестации по дисциплине «Технологии обработки текстов» учебным планом предусмотрен зачет.

При формировании итоговой оценки используется рейтинговая оценка знаний студента, дающая возможность проследить и оценить динамику процесса достижения целей обучения. Рейтинговая оценка предусматривает использование весовых коэффициентов для текущего контроля знаний студентов по дисциплине.

Примерные весовые коэффициенты, определяющие вклад текущего контроля знаний в рейтинговую оценку (формирование оценки за текущую успеваемость):

- отчёты по практическим домашним заданиям с их устной защитой – 40 %;
- контрольные работы – 20 %;
- тестирование – 10%;
- устный опрос – 10%;
- выступление с докладом – 20%.

Рейтинговая оценка по дисциплине рассчитывается на основе оценки текущей успеваемости и зачетной оценки с учетом их весовых коэффициентов. Вес оценки по текущей успеваемости составляет 30 %, зачетной оценки – 70 %.

Примерная тематика практических занятий

Занятие № 1. NOOJ - работа со словарями.

Занятие № 2. NOOJ - грамматики.

Занятие № 3. NOOJ - статистика и конкордансы.

Занятие № 4. Исследование популярных ИПС, изучение расширенной функциональности для поиска документов и веб-страниц.

Занятие № 5. Сравнительный анализ результатов работы ИПС.

Рекомендуемая тематика контрольных работ

Контрольная работа №1. Понятие обработки текстов ЕЯ.
Контрольная работа №2. Векторное представление слов.

Текущий контроль знаний проводится в соответствии с учебно-методической картой дисциплины.

Описание инновационных подходов и методов к преподаванию учебной дисциплины

При организации образовательного процесса большинства практических занятий используется *практико-ориентированный подход*, который предполагает освоение содержания учебного материала через решение практических задач, а также приобретение навыков эффективного выполнения разных видов профессиональной деятельности.

Кроме этого, при организации образовательного процесса используется комбинация таких методов *креативного обучения*, как *методы группового обучения*, *проектного обучения* и *учебной дискуссии*. Комбинация методов предполагает ориентацию на генерирование идей, приобретение навыков для решения исследовательских, творческих и коммуникационных задач, появление нового уровня понимания изучаемой темы, применение знаний (теорий, концепций) при решении проблем, определение способов их решения.

Методические рекомендации по организации самостоятельной работы обучающихся, подготовка к экзамену

Для организации самостоятельной работы студентов магистратуры по учебной дисциплине следует использовать информационно-коммуникационные технологии:

– образовательный портал БГУ https://edufpmi.bsu.by;
– образовательный портал InsightRunnerhttps://acm.bsu.by;
– систему AnyTaskhttps://anytask.org/school/bsu;

разместить в сетевом доступе комплекс учебных и учебно-методических материалов (учебно-программные материалы, учебное издание для теоретического изучения дисциплины, презентации лекций, методические указания к практическим занятиям, электронные версии домашних заданий, материалы текущего контроля и текущей аттестации, позволяющие определить соответствие учебной деятельности обучающихся требованиям образовательных стандартов высшего образования и учебно-программной документации, в том числе вопросы для подготовки к экзамену, задания,

вопросы для самоконтроля, список рекомендуемой литературы, информационных ресурсов и др.).

Примерный перечень вопросов к зачету

1. Ввод речи (текста) в компьютер.
2. Человеко-компьютерное взаимодействие
3. Лингвистическая разметка.
4. Лингвистические аннотации.
5. Представления, основанные на абстракции.
6. Недоспецифицированные представления.
7. Процессы обработки текста
8. Системы на базе разметки.
9. Системы на базе аннотаций.
10. Системы интеграции поверхностной и глубокой обработки.
11. Системы, развивающие отдельные аспекты обработки текста
12. Подходы к моделированию языка и обучению представлений в обработке естественного языка.
13. Модели векторного представления слов для различных конечных задач обработки естественного языка.
14. Нейронные сети.
15. Методы снижения размерности.
16. .Задачи, решаемые с помощью векторного представления слов
17. Понятие машинного обучения в искусственном интеллекте.
18. Классификация задач машинного обучения.
19. Модели глубокого обучения для определения смысла слов.
20. Алгоритмы машинного обучения.

ПРОТОКОЛ СОГЛАСОВАНИЯ УЧЕБНОЙ ПРОГРАММЫ УВО

Название учебной дисциплины, с которой требуется согласование	Название кафедры	Предложения об изменениях в содержании учебной программы учреждения высшего образования по учебной дисциплине	Решение, принятое кафедрой, разработавшей учебную программу (с указанием даты и номера протокола)
Интеллектуальные системы мониторинга	Информационных систем управления	Нет	Оставить содержание учебной дисциплины без изменения, (протокол № 9 от 20 марта 2020 г.)

ДОПОЛНЕНИЯ И ИЗМЕНЕНИЯ К УЧЕБНОЙ ПРОГРАММЕ

на _____/_____ учебный год

№№ Пп	Дополнения и изменения	Основание

Учебная программа пересмотрена и одобрена на заседании кафедры информационных систем управления (протокол № ____ от _____ 20__ г.)

Заведующий кафедрой
Д.т.н., профессор _____ В.В.Краснопрошин
(ученая степень, звание) (подпись) (И.О. Фамилия)

УТВЕРЖДАЮ
Декан факультета
Д.т.н., доцент _____ А.М. Недзведзь
(ученая степень, звание) (подпись) (И.О.Фамилия)