**Белорусский государственный университет**

**УТВЕРЖДАЮ**
Проректор по учебной работе и образовательным инновациям
_____ О.Г. Прохоренко
1 декабря 2023 г.

Регистрационный № УД-790/м.

# ТЕХНОЛОГИИ ОБРАБОТКИ ТЕКСТОВ

**Учебная программа учреждения высшего образования по учебной дисциплине для специальности**

**7-06-0533-05 Прикладная математика и информатика**
Профилизация: Интеллектуальные системы

2023 г.

Учебная программа составлена на основе ОСВО 7-06-0533-05-2023 (№ 160 от 18.05.23), примерного учебного плана 7-06-05-016/пр от 18.01.2023 и учебного плана М53-5.3-79/уч. от 11.04.2023 г.

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

**РЕЦЕНЗЕНТЫ:**
**Б.А. Железко** – доцент кафедры маркетинга факультета маркетинга и предпринимательства БНТУ, кандидат технических наук

**РЕКОМЕНДОВАНА К УТВЕРЖДЕНИЮ:**
Кафедрой информационных систем управления
Белорусского государственного университета
(протокол № 3 от 19.10.2023 г.).

Научно-методическим Советом БГУ
(протокол № 3 от 30.11.2023 г.)

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** и программ по дисциплинам первой ступени высшего образования «Интеллектуальные информационные системы», «Вычислительная лингвистика», и дисциплин второй ступени высшего образования «Компьютерная лингвистка» и «Нейроносетевая обработка данных».

## 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 промежуточной 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. Прикладные средства анализа текстов

### *Тема 4.1. Методы классификации*

Общее описание методов классификации. Оценка качества классификации. Вероятностные методы классификации. Методы классификации на основе расстояний. Методы классификации на основе правил. Комбинированные методы классификации. Метод комбинированной иерархической классификации.

### *Тема 4.2. Методы кластерного анализа*

Линейная регрессия, Логистическая регрессия. Линейный дискриминантный анализ. Деревья принятия решений. Метод опорных векторов. Методы аугментации тренировочного корпуса ограниченного объёма для решения прикладных лингвистических задач. Общее описание методов кластерного анализа. Оценка качества кластерного Общее описание методов кластерного анализа. Оценка качества кластерного анализа. Вероятностные методы кластерного анализа. Структурные методы кластерного анализа. Интерпретация результатов кластерного анализа. Сравнительный анализ методов кластерного анализа.

## УЧЕБНО-МЕТОДИЧЕСКАЯ КАРТА УЧЕБНОЙ ДИСЦИПЛИНЫ

Очная форма получения углубленного высшего образования с применением дистанционных образовательных технологий (ДОТ)

| Номер раздела, темы | Название раздела, темы | Количество аудиторных часов | | Форма контроля знаний |
|---|---|---|---|---|
| | | Лекции | Практические занятия | |
| 1 | 2 | 3 | 4 | 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** | **6** | Выполнение тестов. Расчетно-графическая работа №1 |
| 2.1 | Представление лингвистических данных | 2 | 2 | |
| 2.2 | Архитектура инструментальных ЕЯ-систем | 2 | 2 | |
| 2.3 | Системы обработки ЕЯ-текстов | 2 | 2 | |
| **3** | **Векторное представление текста** | **8** | **6** | Коллоквиум. Расчетно-графическая работа №2 |
| 3.1 | Моделирование языка | 2 | 2 | |
| 3.2 | Методы векторного представления | 2 | 2 | |
| 3.3 | Задачи, решаемые с помощью векторного представления слов | 4 | 2 | |
| **4** | **Прикладные средства анализа текстов** | **4** | **8** | Расчетно-графическая работа №3 Контрольная работа |
| 4.1 | Методы классификации | 2 | 4 | |
| 4.2 | Методы кластерного анализа | 2 | 4 | |
| | **ВСЕГО:** | **20** | **20** | |

## ИНФОРМАЦИОННО-МЕТОДИЧЕСКАЯ ЧАСТЬ

### Перечень основной литературы


1. Хобсон, Л. Обработка естественного языка в действии = Natural Language Processing in Action / Л. Хобсон, Х. Ханнес, Х. Коул; [пер. с англ. И. Пальти, С. Черников]. – Санкт-Петербург [и др.]: Питер, 2020. – 575 с. – URL: https://ibooks.ru/bookshelf/371695.
2. Гольдберг, Й. Нейросетевые методы в обработке естественного языка / Йоав Гольдберг; пер. с англ. А. А. Слинкина. – Москва: ДМК Пресс, 2019. – 281 с.
3. Макшанов, А. В. Технологии интеллектуального анализа данных: учебное пособие / А. В. Макшанов, А. Е. Журавлев. – Санкт-Петербург; Москва; Краснодар: Лань, 2018. – 208 с. – URL: https://e.lanbook.com/book/206711.


### Перечень дополнительной литературы


1. 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.
2. Васильев В. Г., Кривенко М. П. Методы автоматизированной обработки текстов. – М.: ИПИ РАН, 2008. –305 с.
3. Всеволодова А.В.Компьютерная обработка лингвистических данных. Изд.2 – 2007. – 96 с.
4. Ландэ Д.В., Снарский А.А., Безсуднов И.В. Интернетика: Навигация в сложных сетях: модели и алгоритмы, 2009.
5. И.В. Совпель. Инженерно-лингвистические принципы, методы и алгоритмы автоматической переработки текста. –Мн., Вышэйшая школа, 1991.
6. 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.
7. 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.
8. Технологии Яндекса. – Режим доступа: https://yandex.ru/dev/
9. Проект АОТ (Автоматическая Обработка Текста). – Режим доступа: https://AOT.ru

## 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.
2. Письменная форма: расчетно-графические работы, контрольные работы.
3. Устно-письменная форма: отчеты по домашним практическим заданиям с их устной защитой, выполнение тестов.

Формой промежуточной 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):

- расчетно-графические работы – 40 %;
- контрольные работы – 40 %;
- устный опрос, колоквиум, тесты– 20%.

Отметка «зачет» выставляется магистранту, имеющему отметку за текущую успеваемость не ниже 4 («четырех») баллов.

## Примерная тематика практических занятий

*Занятие № 1.* Лингвистический процессор и его функциональность.

*Занятие № 2.*Методы машинного перевода.

*Занятие № 3.* Исследование популярных ИПС, изучение расширенной функциональности для поиска документов и веб-страниц.

*Занятие № 4.* Сравнительный анализ результатов работы ИПС.

*Занятие № 5.* Нейросетевая обработка текста.

*Занятие № 6.* Векторное представление текста.

*Занятия № 7-8.*Задачи, решаемые с помощью векторного представления слов.

*Занятие № 9.* Методы классификации.

*Занятие № 10.* Методы кластеризации.

## Рекомендуемая тематика расчетно-графических работ

*Работа № 1.* Автоматизация обнаружения ошибок в текстах.

*Работа № 2.* Исследование популярных ИПС, изучение расширенной функциональности для поиска документов и веб-страниц.

*Работа № 3.* Сравнительный анализ результатов работы ИПС.

## Рекомендуемая тематика контрольной работы

*Контрольная работа*. 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

Для организации самостоятельной работы студентов магистратуры по учебной дисциплине используется образовательный портал БГУ https://edufpmi.bsu.by, на котором размещаются комплекс учебных и учебно-методических материалов (учебно-программные материалы, учебные издания 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 зачету, 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. Процессы обработки текста
6. Системы на базе разметки.
7. Системы на базе аннотаций.
8. Системы интеграции поверхностной и глубокой обработки.
9. Системы, развивающие отдельные аспекты обработки текста.
10. Подходы к моделированию языка и обучению представлений в обработке естественного языка.
11. Модели векторного представления слов для различных конечных задач обработки естественного языка.
12. Нейронные сети.
13. Методы снижения размерности.
14. Задачи, решаемые с помощью векторного представления слов.
15. Алгоритмы машинного обучения.
16. Методы классификации.
17. Методы кластеризации.

## 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) |
|---|---|---|---|
| Компьютерная лингвистика | Информационных систем управления | Нет | Оставить содержание учебной дисциплины без изменения, (протокол № 3 от 19 октября 2023 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 _____ 202_ г.)

Заведующий кафедрой

_____ _____ _____

(степень, звание) (подпись) (И.О.Фамилия)

УТВЕРЖДАЮ
Декан факультета

_____ _____ _____

(степень, звание) (подпись) (И.О.Фамилия)