Федеральное государственное бюджетное
образовательное учреждение высшего образования
«Государственный институт русского языка им. А.С. Пушкина»

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Автоматическая обработка специальных текстов

наименование дисциплины (модуля)

Уровень высшего образования:

магистратура

указывается: бакалавриат, магистратура

Направление подготовки:

45.04.01 – «Филология»

(код и название направления подготовки)

Направленность (профиль) образования:

Прикладная филология

Форма обучения:

очная

очная, очно-заочная, заочная

Рабочая программа дисциплины (модуля) разработана в соответствии с федеральным государственным образовательным стандартом высшего образования по направлению подготовки 45.04.01 – «Филология», утвержденным приказом Минобрнауки России № 980 от 12.08.2020

1. Место дисциплины (модуля) в структуре ОПОП

В соответствии с учебным планом по направлению подготовки 45.04.01 – «Филология», дисциплина «Автоматическая обработка специальных текстов» относится к части, формируемой участниками образовательного процесса программы бакалавриата Б1.В.ДВ.06.01

2. Цель и задачи освоения дисциплины (модуля)

Целью дисциплины является сформировать представление о теоретических основах, методах и технологиях автоматической обработки текстов на естественном языке, а также развить практические умения и навыки применения инструментов компьютерной лингвистики для анализа, обработки и решения прикладных филологических задач на материале специальных текстов (научных, технических, юридических, деловых).

3. Планируемые результаты обучения по дисциплине (модулю), соотнесенные с требуемыми компетенциями выпускников

Компетенции выпускников (коды)	**Индикаторы достижения компетенций**	**Планируемые результаты обучения по дисциплине (модулю), соотнесенные с индикаторами достижения компетенций**
ПК-1. Способен проводить разноаспектное описание лексических единиц для решения теоретических и прикладных задач	ПК-1.2. Применяет цифровые и корпусные технологии для работы с лексическими данными	Знать: - основы и принципы работы цифровых и корпусных технологий обработки лексических данных, а также возможности современных инструментов для анализа специальных текстов. Уметь: - применять цифровые и корпусные технологии для извлечения, систематизации и анализа лексических данных специальных текстов. Владеть: - навыками использования программных средств для анализа лексической сочетаемости, извлечение ключевых слов и

		визуализацию результатов корпусного анализа.
ПК-2. Способен осуществлять деятельность по созданию и редактированию текстового контента в соответствии с задачами профессиональной деятельности	ПК-2.1. Редактирует, трансформирует, реферирует различные типы текстов в зависимости от задач профессиональной деятельности ПК-2.2. Владеет приемами и навыками создания эффективных текстов в зависимости от задач профессиональной деятельности	Знать: - методы и инструменты автоматической обработки текстов, позволяющие редактировать, трансформировать и реферировать специальные тексты различных типов в соответствии с задачами профессиональной филологической деятельности Уметь: - применять технологии автоматической обработки текстов для редактирования и автоматического реферирования специальных текстов с учётом их типа и поставленных профессиональных задач. Владеть: - навыками обработки специальных текстов с применением инструментов автоматической обработки (частотный анализ, извлечение ключевых слов, тематическое моделирование, оценка читаемости) в зависимости от задач профессиональной деятельности.

4. Объем дисциплины 2 з.е., в том числе 32 академических часа на контактную работу обучающихся с преподавателем; 38 академических часов на самостоятельную работу обучающихся.

5. Содержание дисциплины (модуля), структурированное по темам (разделам) с указанием отведенного на них количества академических часов

Наименование и краткое содержание разделов и тем дисциплины (модуля). **Форма промежуточной аттестации по дисциплине (модулю)**	**Всего (часы)**	В том числе							
		Контактная работа (работа во взаимодействии с преподавателем) ***Виды контактной работы, часы***					**Самостоятельная работа обучающегося** ***Виды самостоятельной работы, часы***		
		Занятия лекционного типа	Занятия семинарского типа (практические/ лабораторные)	Групповые консультации	Индивидуальная работа с обучающимися	**Всего**	**УО, Д.**	**Тест**	**Всего**
Тема 1. Введение в автоматическую обработку специальных текстов	10	2	2			4	4	2	6
Тема 2. Этапы автоматической обработки текста	10	2	2			4	4	2	6
Тема 3. Морфологический анализ и лемматизация	8	2	2			4	4		4
Тема 4. Синтаксический анализ текста	8	2	2			4	4		4
Тема 5. Статистические методы в обработке текстов	10	2	2			4	4	2	6
Тема 6. Семантический анализ и тематическое моделирование	8	2	2			4	4		4

Тема 7. Автоматическая классификация и кластеризация текстов	8	2	2			4	4		4
Тема 8. Прикладные аспекты автоматической обработки специальных текстов	8	2	2			4	4		4
Промежуточная аттестация *(указывается форма проведения)*	*2*								
Итого	72								

6. Фонд оценочных материалов по дисциплине (модулю)

Типовые задания и иные материалы, необходимые для оценки результатов обучения

6.1. **Тематика и требования к структуре и содержанию докладов**

Примерная тематика докладов:

1. История развития автоматической обработки естественного языка: от первых экспериментов до современных нейросетевых моделей.
2. Проблемы токенизации специальных текстов: аббревиатуры, формулы, числа и нестандартные символы.
3. Лемматизация vs стемминг: эффективность применения в задачах обработки специальных текстов.
4. Синтаксический анализ научных текстов: особенности построения деревьев зависимостей.
5. Закон Ципфа и его проявление в специальных текстах различных жанров.
6. Построение частотных словарей лемм как инструмент лексического анализа специального текста.
7. Тематическое моделирование как метод выявления скрытой тематической структуры коллекции специальных текстов.
8. Извлечение ключевых слов из научных и технических текстов: методы и инструменты.

9. Автоматическая классификация специальных текстов по жанрам и отраслям знания.
10. Кластеризация документов: алгоритмы и оценка качества на материале специальных текстов.
11. Методы автоматического реферирования: экстрактивный и абстрактивный подходы.
12. Атрибуция текстов: статистические и стилометрические методы.
13. Оценка читаемости специальных текстов с помощью автоматических метрик.
14. Корпусные технологии в исследовании специальных текстов: создание и разметка корпусов.
15. Перспективы развития автоматической обработки специальных текстов: большие языковые модели и их применение в филологии.

Требования к структуре и содержанию доклада:

Доклад – это продукт самостоятельной работы студента, публичное выступление по представлению полученных результатов решения определенной учебно-практической, учебно-исследовательской или научной темы

Критерии оценки доклада

Критерии	Показатели
Актуальность темы, корректность предлагаемых решений.	-умение формулировать проблему исследования; -выдвигать корректные гипотезы; -умение намечать пути исследования; -умение отбирать материал;
Качество доклада. Полнота раскрытия темы.	- изложение полученных знаний полное; -объем и полнота разработок, - самостоятельность, законченность, подготовленность к докладу, -уровень творчества: оригинальность раскрытия темы, подходов, предлагаемых решений. -эрудиция, наличие межпредметных (междисциплинарных) связей.
Критерий «Выбор и использование средств решения проблемы, адекватных целям»	- **Умение представлять подготовленный доклад.** - Качество презентации: оформление, соответствие стандартным требованиям, качество схем, рисунков. - Научный стиль речи: навыки публичного выступления; - аттрактивность, - чувство времени; - Культура речи.

Глубина проработки материала, умение держать внимание аудитории	Ответы на вопросы: полнота, аргументированность, убежденность, стремление раскрыть тему и сильные стороны работы.

Шкала оценивания доклада:
«отлично» – все критерии соответствуют показателям в полном объеме.
«хорошо» – 1-2 критерия не в полном объеме соответствуют показателям.
«удовлетворительно» – все критерии реализованы не в полном объеме.
«неудовлетворительно» – результаты работы не соответствует всем или нескольким критериям

6.2. Вопросы для подготовки к контрольным работам

1. Что принято понимать под «специальным текстом» и какие типы специальных текстов выделяют в современной лингвистике?
2. Каковы основные направления и задачи автоматической обработки естественного языка на современном этапе?
3. В чём заключается специфика автоматической обработки специальных текстов по сравнению с обработкой художественных и разговорных текстов?
4. Какие этапы включает полный цикл автоматической обработки текста и какова последовательность их выполнения?
5. В чём состоят основные проблемы токенизации специальных текстов и какими способами они решаются?
6. Какие инструменты токенизации и сегментации предложений применяются для русскоязычных текстов и каковы критерии их выбора?
7. Каковы задачи морфологического анализа текста и какие методы лежат в его основе?
8. В чём различие между лемматизацией и стеммингом и в каких задачах обработки специальных текстов предпочтителен каждый из подходов?
9. Какие инструменты морфологического анализа существуют и каковы их сравнительные преимущества и ограничения?
10. Какие типы синтаксических структур (деревья зависимостей и деревья составляющих) используются при автоматическом анализе текста?
11. Каковы основные этапы и методы синтаксического парсинга предложений?

12. Какие инструменты синтаксического анализа применяются для русскоязычных специальных текстов и как визуализируются результаты разбора?
13. В чём суть частотного анализа текста и какие лингвостатистические параметры он позволяет вычислить?
14. Как формулируется закон Ципфа и какова его роль в анализе специальных текстов?
15. Как строится частотный словарь лемм и какие выводы о специальном тексте можно сделать на его основе?
16. Как работает метод тематического моделирования и каковы условия его эффективного применения?
17. Какие меры семантической близости и методы извлечения ключевых слов применяются при анализе специальных текстов?
18. Какие методы машинного обучения применяются для автоматической классификации текстов и в чём особенности наивного байесовского классификатора?
19. В чём различие между классификацией и кластеризацией текстов и какие алгоритмы кластеризации документов используются?
20. Как оценивается качество автоматической классификации и кластеризации специальных текстов?
21. Какие подходы к автоматическому реферированию (экстрактивный и абстрактивный) существуют и каковы их возможности при работе со специальными текстами?
22. Какие задачи решает извлечение информации из специальных текстов и какие методы при этом используются?

6.3. Типовые задания для самостоятельной проверочной работы.

Задание 1. Токенизация и нормализация специального текста

Условие: Дан фрагмент научной статьи (5–7 предложений), содержащий аббревиатуры, формулы, числовые обозначения и специальные символы.

Выполните токенизацию текста с помощью библиотеки NLTK и spaCy.

Сравните результаты токенизации, выявите расхождения.

Предложите способы устранения ошибок токенизации (обработка аббревиатур, формул, десятичных дробей).

Оформите отчёт с таблицей сравнения результатов.

Задание 2. Морфологический анализ и лемматизация

Условие: Дан фрагмент юридического текста (постановление, договор) объёмом 500–700 слов.

Задание:
Проведите морфологический анализ текста с помощью специального программного средства.
Выполните лемматизацию текста.
Постройте морфологический профиль текста (распределение частей речи).
Сравните результаты двух инструментов, сделайте выводы о качестве их работы на юридическом тексте.

Задание 3. Синтаксический анализ предложений
Условие: Даны 5 предложений из технической документации (инструкция, руководство пользователя).
Задание:
Выполните синтаксический разбор предложений с помощью spaCy.
Постройте и визуализируйте деревья зависимостей.
Охарактеризуйте типичные синтаксические конструкции, характерные для технического текста.
Сделайте вывод об особенностях синтаксиса технической документации.
Критерии оценивания: корректность разбора, качество визуализации, обоснованность лингвистических выводов.

Задание 4. Частотный анализ и статистика текста
Условие: Дан корпус специальных текстов одного жанра (например, 5 научных статей по одной тематике) общим объёмом не менее 3000 слов.
Задание:
Постройте частотный словарь лемм корпуса.
Проверьте соответствие распределения частот закону Зипфа (постройте график в логарифмических координатах).
Вычислите постоянную Ципфа для данного корпуса.
Выделите ядро лексики специального текста (первые 50 лемм) и охарактеризуйте его.

Задание 5. Автоматическое реферирование текста
Условие: Дан специальный текст объёмом 1500–2000 слов (научная статья или аналитический отчёт).
Задание:
Выполните экстрактивное реферирование текста (выделение наиболее значимых предложений).
Составьте аннотацию текста объёмом 5–7 предложений.

Сравните полученный автоматический реферат с аннотацией, составленной вручную.

Оцените качество автоматического реферирования и сделайте выводы.

Критерии оценивания: корректность применения метода реферирования, качество составленной аннотации, обоснованность сравнения и выводов.

6.4. Примерные вопросы для подготовки к зачету

1. Предмет, цели и задачи автоматической обработки специальных текстов.
2. Понятие «специальный текст»: определение, типология и лингвистические особенности.
3. Виды специальных текстов (научные, технические, юридические, деловые) и их характеристики.
4. История развития автоматической обработки естественного языка: основные этапы и достижения.
5. Основные направления современной компьютерной лингвистики и обработки естественного языка.
6. Специфика автоматической обработки специальных текстов по сравнению с другими типами текстов.
7. Этапы автоматической обработки текста: общая характеристика и последовательность.
8. Графематический анализ и нормализация текста.
9. Токенизация текста: задачи, проблемы и методы решения.
10. Сегментация предложений и инструменты токенизации для русскоязычных текстов.
11. Морфологический анализ текста: задачи, методы и инструменты.
12. Лемматизация и стемминг: сравнительный анализ подходов.
13. Инструменты морфологического анализа русского языка.
14. Синтаксический анализ текста: типы синтаксических структур (деревья зависимостей и составляющих).
15. Методы и инструменты синтаксического парсинга предложений.
16. Синтаксические особенности специальных текстов и их автоматический анализ.
17. Частотный анализ текста: задачи и лингвостатистические параметры.
18. Закон Зипфа: формулировка, проверка и применение в анализе специальных текстов.
19. Построение частотных словарей лемм и их использование в исследовании специальных текстов.
20. Тематическое моделирование текстов и его применения.
21. Автоматическая классификация текстов: методы машинного обучения и наивный байесовский классификатор.
22. Автоматическое реферирование специальных текстов: экстрактивный и абстрактивный подходы.
23. Атрибуция текстов: статистические и стилометрические методы.

24. Машинный перевод специальных текстов: особенности, проблемы терминологии и точности.

25. Перспективы развития автоматической обработки специальных текстов: большие языковые модели и их применение в филологии.

6.5. Шкала и критерии оценивания ***(шкала и критерии оценивания могут быть едиными (типовыми) для всех дисциплин (модулей), входящих в основную образовательную программу)***

Знания обучающего определяются следующими отметками: «отлично», «хорошо», «удовлетворительно», «неудовлетворительно». К промежуточной аттестации допускаются обучающиеся, принявшие участие во всех формах текущего контроля по дисциплине. В случае наличия задолженности по формам текущего контроля обучающийся отрабатывает пропущенные элементы текущего контроля по согласованию с преподавателем.

Промежуточная аттестация по дисциплине «Морфология» осуществляется в форме экзамена, а именно, в виде ответа на теоретические вопросы билета и выполнения практического задания.

Оценочные материалы для проведения промежуточной аттестации обучающихся приведены в приложении к рабочей программе дисциплины.

Критерии оценивания результатов обучения по дисциплине при проведении промежуточной аттестации:

<table>
<tr><td rowspan="3">Планируемые результаты обучения по дисциплине</td><td colspan="4">Критерии оценивания результатов обучения по дисциплине</td></tr>
<tr><td>«неудовлетворительно»</td><td>«удовлетворительно»</td><td>«хорошо»</td><td>«отлично»</td></tr>
<tr><td colspan="2">«не зачтено»</td><td colspan="2">«зачтено»</td></tr>
<tr><td>Сформированность знаний (знать)</td><td>Уровень знаний ниже минимальных требований.

Имели место грубые ошибки.

Невозможно оценить полноту знаний вследствие отказа обучающегося от ответа.</td><td>Минимально допустимый уровень знаний.

Допущено много негрубых ошибок.</td><td>Уровень знаний в объеме, соответствующем программе подготовки.

Допущено несколько несущественных ошибок.</td><td>Уровень знаний в объеме, соответствующем программе подготовки, или в объеме, превышающем программу подготовки, продемонстрирован без ошибок и погрешностей.</td></tr>
</table>

Сформированность умений (уметь)	Отсутствие минимальных умений. При решении стандартных задач не продемонстрированы основные умения. Имели место грубые ошибки. Невозможно оценить наличие умений вследствие отказа обучающегося от ответа.	Продемонстрированы основные умения. Решены типовые задачи с негрубыми ошибками. Выполнены все задания, но не в полном объеме.	Продемонстрированы все основные умения. Решены все основные задачи, но некоторые с негрубыми ошибками. Выполнены все задания, в полном объеме, но некоторые с недочетами.	Продемонстрированы все основные умения. Решены все основные задачи. Выполнены все задания, в полном объеме без недочетов.
Сформированность навыков (владеть)	При решении стандартных задач не продемонстрированы базовые навыки. Имели место грубые ошибки. Невозможно оценить наличие навыков вследствие отказа обучающегося от ответа.	Продемонстрирован минимальный набор навыков. Решены стандартные задачи с некоторыми недочетами.	Продемонстрированы базовые навыки. Решены стандартные задачи, но некоторые с недочетами.	Продемонстрировано всестороннее владение навыками. Задачи, в том числе нестандартные, решены без ошибок и недочетов.

7. Методические материалы

7.1. Перечень основной и дополнительной учебной литературы и иные материалы:

7.1.1. Основная литература

1. Компьютерная лингвистика и автоматическая обработка текстов : учебное пособие / Н.В. Лукашевич, А.А. Сорокин. – Москва : МАКС Пресс, 2025. – 608 с.
2. Баранов А.Н. Введение в прикладную лингвистику. — М.: ЛЕНАНД, 2020.
3. Большакова Е.И. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика. — М.: МИЭМ, 2011.

7.1.2. Дополнительная литература

1. Jurafsky D., Martin J.H. Speech and Language Processing. — Pearson, 2020.
2. Manning C.D., Schütze H. Foundations of Statistical Natural Language Processing. — MIT Press, 1999.
3. Лукашевич Н.В. Тезаурусы в задачах информационного поиска. — М.: Изд-во МГУ, 2011.

7.2. Перечень лицензионного программного обеспечения

1. Microsoft Windows 10,
2. Microsoft Office 2016 стандартная,
3. Big Blue Button

8. Ресурсное обеспечение

8.1. Перечень профессиональных баз данных и информационных справочных систем:

Наименование	Ссылка на ресурс	Доступность (свободный доступ/ ограниченный доступ)
Информационно-справочные системы		
Интернет-портал «Русский язык»:	http://www.gramota.ru	свободный доступ
Портал «Образование на русском»	https://pushkininstitute.ru	ограниченный доступ, необходима регистрация
Электронно-библиотечные системы		
ЭБС Юрайт	https://urait.ru	ограниченный доступ, необходима регистрация
ЭБС IPR Smart	https://iprbookshop.ru	ограниченный доступ, необходима регистрация
Профессиональные базы данных		
Национальный корпус русского языка.	http://www/ruscorpora.ru/corpora-morph.html	свободный доступ
Научная электронная библиотека и Российский индекс научного цитирования	http://www.elibrary.ru	свободный доступ

8.2. Перечень ресурсов информационно-телекоммуникационной сети «Интернет»:

Наименование	Ссылка на ресурс	Доступность

		(свободный доступ/ ограниченный доступ)
Электронно-библиотечные системы		
ЭБС Юрайт	https://urait.ru	ограниченный доступ, необходима регистрация
ЭБС IPR Smart	https://iprbookshop.ru	ограниченный доступ, необходима регистрация

9. Язык преподавания

Русский

10. Разработчики программы

Катышев П.А., доктор филологических наук, профессор, заведующий кафедрой общего и русского языкознания