Программа разработана экспертами
Федерального учебно-методического объединения
высшего образования по укрупненной группе
специальностей и направлений подготовки
45.00.00 Языкознание и литературоведение

Утверждена на заседании ФУМО
25 мая 2021 года

Примерная программа учебной дисциплины

АВТОМАТИЧЕСКАЯ ОБРАБОТКА ТЕКСТА

Уровень высшего образования:

МАГИСТРАТУРА

Направление подготовки:

45.03.03 «ФУНДАМЕНТАЛЬНАЯ И ПРИКЛАДНАЯ ЛИНГВИСТИКА»

Москва 2021

Раздел 1. Характеристики учебных занятий

1.1 Цели и задачи учебных занятий

Целью данного курса является ознакомление студентов с современными лингвистическими технологиями в задачах автоматической обработки текстов. Результатом занятий должно стать приобретение студентами навыков самостоятельного подбора метода решения лингвистической задачи и навыков применения статистических методов к языковым данным.

1.2 Место дисциплины (модуля) в структуре образовательной программы, связь с другими дисциплинами (модулями) программы

Относится к вариативной части ОПОП ВО.

1.3 Требования подготовленности обучающегося к освоению содержания учебных занятий (пререквизиты)

Устанавливаются образовательной организацией.

1.4 Перечень результатов обучения

По окончании курса студент должен знать современные лингвистические технологии, достигаемый ими уровень качества в задачах автоматической обработки текстов; уметь выбирать подходящее решение среди имеющихся подходов, инструментов анализа текстов для задач автоматической обработки текстов; владеть приемами обоснования принятых решений, приемами тестирования систем автоматической обработки текстов.

1.5 Перечень рекомендуемых образовательных технологий

В преподавании дисциплины «Автоматическая обработка текста» используются разнообразные образовательные технологии как традиционного, так и инновационного характера, учитывающие смешанный, теоретико- и практикоориентированный характер дисциплины:

- лекции;
- практические занятия;
- дискуссии;
- выступления с докладами и сообщениями;
- аудиторные контрольные работы;
- внеаудиторные контрольные работы;
- тестирование.

Степень необходимости образовательной среды и ее выбор определяется образовательной организацией. Формы текущей аттестации определяются образовательной организацией.

1.6 Объем дисциплины (модуля) в зачетных единицах

2 з.е.

Раздел 2. Организация, структура и содержание учебных занятий

2.1 Организация учебных занятий

Предусмотрены учебные занятия с использованием дистанционных технологий.

2.2 Краткая аннотация содержания дисциплины (модуля)

Наименование темы (раздела, части)	Вид учебных занятий	Кол-во часов
1. Автоматическая обработка текстов в технологиях искусственного интеллекта. Задачи автоматической обработки текстов. Типы данных, используемых в компьютерной лингвистике.	Лекции	2
2. Языковые статистические модели: n-граммы. Понятие языковой модели. Сглаживание, аддитивное сглаживание. Интерполяция и откат в языковой модели. Метод Уиттена-Белла. Проблемы энграммных моделей, модификации энграммных моделей.	Практические занятия	4
3. Современные методы морфологического анализа. Теория формальных языков: регулярные выражения, конечные автоматы, преобразователи. Явные и скрытые марковские модели. Особенности применения скрытых марковских моделей к морфологическому анализу. Состояния модели. Лексические вероятности. Недостатки скрытых марковских моделей.	Практические занятия	6
4. Современные методы синтаксического анализа. Грамматики зависимостей. Синтаксические структуры, проективность. Алгоритм Нивре синтаксического анализа для проективных и непроективных деревьев. Реализация алгоритма Нивре на основе машинного обучения. Грамматики непосредственно составляющих. Введение в теорию компиляторов и интерпретаторов	Практические занятия	4
5. Современные методы разрешения анафоры и кореференции Задача разрешения анафоры: модель generate-filter-rank. Признаки для классификации. Задача разрешения кореферентности: модель mention-pair, модель entity-mention, недостатки моделей. Признаки. Способы оценки качества.	Практические занятия	2
Итого:		18

Раздел 3. Обеспечение учебных занятий

3.1 Методические указания по освоению дисциплины

Преподавание дисциплины осуществляется в форме лекций и практических занятий. Во время занятий обучающиеся выполняют практические задания, иллюстрирующие основные задачи и методы автоматической обработки текстов. Для закрепления пройденного материала предлагаются домашние задания по каждой из тем. Успешное овладение содержанием дисциплины «Автоматическая обработка текстов» предполагает работу обучающихся в группах в аудитории, а также их самостоятельную работу.
Дополнительные методические указания устанавливаются образовательной организацией.

3.2 Примерный перечень учебно-методического обеспечения самостоятельной работы обучающихся по дисциплине (модулю), в том числе примерный перечень учебной литературы и ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины (модуля)

Самостоятельная работа студентов должна включать усвоение теоретического материала, подготовку к практическим занятиям, выполнение творческих заданий, работу с электронным учебно-методическим комплексом, подготовку к текущему контролю знаний, к промежуточной аттестации (зачету).

Список рекомендованной литературы

Jurafsky D. & Martin J. H. 2000. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. Upper Saddle River, N.J: Prentice Hall.

Witten, I & Bell, T. 1991. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. In *IEEE Transactions on Information Theory*. 37. 1085-1094. 10.1109/18.87000.

Баранов А. Н. 2017. Введение в прикладную лингвистику: [учебник]. Московский государственный университет им. М.В. Ломоносова. Изд. 5-е. Москва: URSS, Москва: ЛЕНАНД.

Иомдин Л. Л. 1990. Автоматическая обработка текста на естественном языке: модель согласования. М.: Наука, 1990.

Кибрик А. Е. и др. 2019. Введение в науку о языке. Раздел 5: Прикладная и компьютерная лингвистика. М.: Буки-Веди. С. 455-534.

Кибрик А. А. 2019. Дискурс / Введение в науку о языке / Под ред. С. Г. Татевосова, О. В. Федоровой. М: Буки Веди. 126–163.

Кобзарева Т. Ю. 2015. В поисках синтаксической структуры: автоматический анализ русского предложения с опорой на сегментацию. М.: РГГУ.

Корочков А. В. 2006. О количественной оценке адекватности лингвистических правил (на материале правил чтения для английского языка). *Вопросы языкознания*, №5. 78-91

Леонтьева Н. Н. 2006. Автоматическое понимание текстов: системы, модели, ресурсы: учеб. пособие для студ. линг. фак. вузов. М.: Издательский центр "Академия".

Падучева Е. В. 1986. О референции языковых выражений с непредметным значением. *Научно-техническая информация*, сер.2, N 1.

Толдова С. Ю., Бонч-Осмоловская А.А. 2019. Автоматическая обработка текста. // А.Е. Кибрик и др. Введение в науку о языке. М.: Буки-Веди. 513-527.

Шаврина Т. О. 2017. Методы обнаружения и исправления опечаток: исторический обзор. *Вопросы языкознания*, № 4. 115-134

Описание материально-технической базы, рекомендуемой для осуществления образовательного процесса по дисциплине (модулю)

Учебная аудитория с мультимедийным комплексом.

Описание материально-технической базы (в т.ч. программного обеспечения), рекомендуемой для адаптации электронных и печатных образовательных ресурсов для обучающихся из числа инвалидов и лиц с ОВЗ

Устанавливается образовательной организацией.

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

Для контроля усвоения данной дисциплины предусмотрен зачет. Контрольные мероприятия по текущему контролю знаний обучающихся проводятся в часы, отведенные для изучения дисциплины.
В течение семестра студентами выполняются практические и контрольные работы.
Порядок проведения зачета определяется ВУЗом.

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

Примерные вопросы для самоконтроля:

1. Понятие языковой модели.
2. Проблемы энграммных моделей, модификации энграммных моделей.
3. Скрытая марковская модель. Особенности применения скрытых марковских моделей к морфологическому анализу. Состояния модели.
4. Особенности применения скрытых марковских моделей к морфологическому анализу. Лексические вероятности.
5. Недостатки скрытых марковских моделей.
6. Условные случайные поля. Отличие от скрытых марковских моделей.
7. Синтаксические структуры, проективность
8. Алгоритм Нивре синтаксического анализа. Основные операции.

Примерные практические задания:

1. Найти оптимальное исправление опечатки с помощью энграммной модели.
2. Определить параметры скрытой марковской модели на основании корпусных данных.
3. Построить синтаксическое дерево с помощью алгоритма Нивре.

Примерный перечень вопросов к зачету (экзамену) по всему курсу:

1. Типы данных, используемых в компьютерной лингвистике. Способы разметки данных.
2. Понятие языковой модели. Сглаживание, аддитивное сглаживание.
3. Интерполяция и откат в языковой модели.
4. Метод Уиттена-Белла.
5. Проблемы энграммных моделей, модификации энграммных моделей.
6. Теория формальных языков: регулярные выражения, конечные автоматы, преобразователи.
7. Скрытая марковская модель. Основная идея, алгоритм Витерби поиска наиболее вероятной последовательности состояний.
8. Особенности применения скрытых марковских моделей к морфологическому анализу. Состояния модели.
9. Особенности применения скрытых марковских моделей к морфологическому анализу. Лексические вероятности.
10. Недостатки скрытых марковских моделей.
11. Условные случайные поля. Основная идея. Отличие от скрытых марковских моделей.
12. Синтаксические структуры, проективность.
13. Проект Универсальные зависимости. Основная идея. Универсализм морфологических признаков.
14. Алгоритм Нивре синтаксического анализа. Основные операции.

15. Вариант алгоритма Нивре для непроективных деревьев.
16. Реализация алгоритма Нивре на основе машинного обучения. Типы признаков.
17. Грамматика непосредственно составляющих. Базовые элементы теории компиляторов/интерпретаторов.
18. Задача разрешения анафоры: модель generate-filter-rank. Признаки для классификации.
19. Задача разрешения кореферентности: модель mention-pair, модель entity-mention, недостатки моделей. Признаки.

3.5 Материально-техническое обеспечение

Минимально необходимый для реализации курса перечень материально-технического обеспечения включает лекционные аудитории (с компьютерным и видеопроекционным оборудованием для презентаций, средствами звуковоспроизведения и экраном, с выходом в Интернет). Количество индивидуальных рабочих станций должно соответствовать количество студентов.

3.6 Информационное обеспечение

Рекомендуемая основная литература

Кибрик А. А. 2019. Дискурс // Введение в науку о языке / Под ред. С. Г. Татевосова, О. В. Федоровой. М: Буки Веди. С. 126–163.

Баранов А. Н. 2017. Введение в прикладную лингвистику: [учебник]. Московский государственный университет им. М.В. Ломоносова. Изд. 5-е. Москва: URSS, Москва: ЛЕНАНД.

Толдова С. Ю., Бонч-Осмоловская А. А. 2019. Автоматическая обработка текста. // А.Е. Кибрик и др. Введение в науку о языке. М.: Буки-Веди. 513-527.

Jurafsky D. & Martin J. H. 2000. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. Upper Saddle River, N.J: Prentice Hall.

Рекомендуемая дополнительная литература

Witten I & Bell T. 1991. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. In *IEEE Transactions on Information Theory*. 37. 1085-1094.

Иомдин Л. Л. 1990. Автоматическая обработка текста на естественном языке: модель согласования. М.: Наука.

Кибрик А. Е. и др. 2019. Введение в науку о языке. Раздел 5: Прикладная и компьютерная лингвистика. М.: Буки-Веди. 455-534.

Кобзарева Т. Ю. 2015. В поисках синтаксической структуры: автоматический анализ русского предложения с опорой на сегментацию. М.: РГГУ.

Корочков А. В. 2006. О количественной оценке адекватности лингвистических правил (на материале правил чтения для английского языка). *Вопросы языкознания*, №5. 78-91

Леонтьева Н. Н. 2006. Автоматическое понимание текстов: системы, модели, ресурсы: учеб. пособие для студ. линг. фак. вузов. М.: Издательский центр "Академия".

Падучева Е. В. 1986. О референции языковых выражений с непредметным значением. *Научно-техническая информация*, сер.2, N 1.

Шаврина Т.О. 2017. Методы обнаружения и исправления опечаток: исторический обзор. *Вопросы языкознания*, № 4. 115-134

Рекомендуемый перечень иных информационных источников

Портал материалов по машинному обучению machinelearning.ru

Раздел 4. Разработчики программы

Сорокин Алексей Андреевич, кандидат физико-математических наук, ассистент.
Рабочая группа ФУМО 45.00.00 по проблемам искусственного интеллекта в языкознании и литературоведении.