Программа учебной дисциплины "Автоматическая обработка текста"

Утверждена
Академическим советом ООП
Протокол № от «__»_____20__ г.

Автор	Большакова Е.И.
Число кредитов	5
Контактная работа (час.)	60
Самостоятельная работа (час.)	130
Курс	4
Формат изучения дисциплины	Без использования онлайн курса

I. ЦЕЛЬ, РЕЗУЛЬТАТЫ ОСВОЕНИЯ ДИСЦИПЛИНЫ И ПРЕРЕКВИЗИТЫ

Программа предназначена для студентов четвертого года обучения в бакалавриате по направлению 01.03.02 «Прикладная математика и информатика».

Главная цель изучения учебной дисциплины «Автоматическая обработка текстов» – освоение основ автоматической обработки текстов (АОТ) на естественном языке (ЕЯ), что предполагает также овладение базовыми навыками работы с существующими программными средствами АОТ и лингвистическими ресурсами.

В результате изучения дисциплины студенты должны:

- Знать основные особенности неструктурированных текстов на ЕЯ и принципы их графематического, морфологического, синтаксического и статистического анализа;
- Понимать ограничения компьютерных моделей автоматической обработки текстов (АОТ);
- Знать типичные прикладные системы в области АОТ и их архитектурные особенности;
- Иметь представление о видах лингвистических ресурсов, используемых в различных системах обработки текстов;
- Уметь применять готовые программные модули анализа текстов и открытые лингвистические ресурсы для решения частных задач АОТ.

Изучение данной дисциплины требует предварительных знаний по дисциплинам: Дискретная математика; Теория вероятностей и математическая статистика; Основы программирования; Алгоритмы и структуры данных; Технологии программирования. Для освоения учебной дисциплины студенты должны уметь уверенно программировать модули на языке высокого уровня с использованием инструментальных средств.

Основные положения дисциплины используются в дальнейшем при изучении следующих дисциплин учебных программ бакалавра и магистра: Методы машинного обучения и разработки данных; Современные методы анализа данных; Компьютерная лингвистика.

II. СОДЕРЖАНИЕ УЧЕБНОЙ ДИСЦИПЛИНЫ

Тема 1. Введение

1. Автоматическая обработка текстов на естественном языке (ЕЯ): основные задачи и особенности направления. Естественный язык как сложная система языковых знаков. Уровни языковой системы. Феномены ЕЯ: полисемия, синонимия, омонимия.

2. Лингвистические процессоры и лингвистические ресурсы. Этапы анализа текста. Обзор основных приложений АОТ.

Тема 2. Начальные этапы обработки текста

1. Графематический анализ и сегментация текста. Токенизация и разбиение на предложения. Виды токенов, обработка сложных случаев.

2. Основные понятия морфологии: словоформа, морфема, аффикс, корень, основа, флексия. Словоизменительная парадигма и морфологические параметры. Словарные и бессловарные модели морфологии.

3. Автоматический морфологический анализ и синтез. Виды морфоанализа: стемминг, лемматизация, полный морфоанализ. Принципы морфоанализа на базе словаря основ или словаря словоформ. Морфологические процессоры для русского языка.

Тема 3. Статистические характеристики текстов и корпусная лингвистика

1. Статистика словоупотреблений в текстах на ЕЯ. Абсолютные и относительные частоты словоформ и лексем. Закон Ципфа-Мандельброта и его интерпретация. Соотношение длины слова и его частоты. Глоттохронология.

2. Статистика встречаемости символов и буквосочетаний: биграмм и триграмм, N-грамм. Задачи АОТ, решаемые на базе статистики символов.

3. Задачи корпусной лингвистики. Коллекции и корпуса текстов. Характеристики и состав типичного корпуса. Национальный корпус русского языка.

4. Статистика N-грамм для слов. Понятие статистической языковой модели. Применение статистической (вероятностной) модели для разрешения морфологической омонимии. Использование статистики для автоматического выделения устойчивых словосочетаний языка.

Тема 4. Подходы к автоматическому анализу синтаксиса и семантики текста

1. Задачи синтаксического анализа ЕЯ. Синтаксические деревья непосредственных составляющих и деревья зависимостей. Синтаксические связи слов. Понятие модели управления слова-предиката. Синтаксический разбор на базе контекстно-свободных грамматик. Примеры синтаксических парсеров.

2. Частичный синтаксический анализ. Понятие синтаксической сегментации текста. Автоматическое выделение словосочетаний (именных, предложных групп).

3. Основные способы представления смысла текста и модели представления знаний в искусственном интеллекте: семантические сети, язык предикатов. Семантический анализ текста на основе семантико-синтаксических моделей управления.

4. Связный текст (дискурс), его особенности.

Тема 5. Лингвистические ресурсы

1. Словари для автоматической обработки текстов. Виды словарей. Тезаурус как словарь с семантическими связями единиц. Информационно-поисковые тезаурусы и рубрикаторы.

2. Понятие онтологии. Класификация онтологий. Лингвистическая онтология WordNet.

3. Дистрибутивная семантика и технология Word2Vec.

Тема 6. Прикладные задачи АОТ

1. Подходы к разработке приложений АОТ: инженерный подход, основанный на лингвистических правилах, и подход, основанный на машинном обучении. Основные показатели качества работы систем АОТ: точность, полнота, F-мера.

2. Информационный поиск в массивах полнотекстовых документов: основные понятия. Индексирование текстов для информационного поиска. Векторная модель документа. Булевский поиск, ранжированный поиск. Оценка релевантности документа. Поиск в сети Интернет, принципы работы поисковых машин.

3. Классификация и кластеризация текстов как задачи в области Text Mining. Обзор методов машинной классификации. Особенности кластеризации текстов. Обзор задач АОТ, решаемых на основе классификации текстов.

4. Автоматическое реферирование и аннотирование документов как смежные задачи информационного поиска. Основные стратегии сжатия текста. Типы аннотаций.

5. Машинный перевод. Стратегии машинного перевода, основанного на правилах. Статистический машинный перевод, принципы создания статистического переводчика.

6. Извлечение информации и знаний из текстов: особенности задачи и типы извлекаемых объектов. Понятие лингвистического шаблона для извлечения информации. Инструментальные средства для построения систем извлечения информации из текстов.

7. Автоматический анализ тональности текстов и извлечение мнений из текстов: особенности и подходы к решению. Анализ тональности как задача классификации.

III. ОЦЕНИВАНИЕ

Курс «Автоматическая обработка текстов» читается в 1 и 2 модуле.

Тип контроля	Форма контроля	Параметры
Текущий контроль (1 и 2 модуль)	Контрольная работа	Письменная работа 80 минут
	Самостоятельная аудиторная работа	Письменная работа 10-15 минут
	Домашнее практическое задание	Выдается для выполнения в течение 2-х недель
Итоговый контроль во 2 модуле	Экзамен	Письменная работа 80 минут

Текущий контроль включает самостоятельные аудиторные работы по текущим темам дисциплины; письменную контрольную, проводимую во втором модуле и состоящую из нескольких вопросов и задач по пройденному материалу; а также домашние практические задания на изучение, тестирование и анализ компьютерных моделей АОТ и лингвистических ресурсов.

Домашние работы высылаются по электронной почте и оцениваются дистанционно. Б'ольшая часть вариантов домашних заданий требует программной реализации моделей. Самостоятельные аудиторные и домашние практические работы оцениваются суммарно, в баллах, и исходя из набранной суммы выставляется результирующая оценка по каждому виду работ.

Итоговый контроль проводится в форме письменного экзамена, включающего несколько вопросов и задач по темам дисциплины: каждый вопрос/задача оценивается в баллах, общая оценка определяется как доля набранных баллов по отношению к максимально возможному числу баллов.

В первом и втором модуле преподаватели оценивают домашние практические и самостоятельные аудиторные работы студентов, выставляя в итоге за каждый вид работ суммарную оценку по десятибалльной системе. Таким образом, конце 2-го модуля определяются соответствующие результирующие оценки $О_{д/з}$ и $О_{сам.\ аудит.\ работа}$, рассчитанные по десятибалльной системе на основе нормированной суммы баллов, полученных за все домашние задания и самостоятельные работы соответственно.

Накопленная оценка за первый и второй модуль рассчитывается (с округлением до целого арифметическим способом) по формуле:

$$О_{накопленная} = 0{,}3 \cdot О_{к/р} + 0{,}4 \cdot О_{д/з} + 0{,}3 \cdot О_{сам.\ аудит.\ работа}$$

где $О_{к/р}$ – оценка письменной контрольной работы (по десятибалльной системе).

В диплом выставляется ***результирующая оценка*** по данной учебной дисциплине, согласно следующей формуле (округление арифметическое):

$$О_{дисциплина} = 0{,}8 \cdot О_{накопленная} + 0{,}2 \cdot О_{экзамен}$$

где $О_{экзамен}$ – оценка по десятибалльной системе за письменную работу непосредственно на экзамене.

Оценка за курс не включает блокирующие элементы.

Пересдачи проводятся в виде письменной работы на 80 минут, состоящей из 10-15 вопросов и задач по темам дисциплины.

IV. ПРИМЕРЫ ОЦЕНОЧНЫХ СРЕДСТВ

Примеры домашних практических заданий

1) Реализовать на одном из языков программирования собственный графематический анализатор для русскоязычных текстов и протестировать его на реальных текстах.
2) Провести сравнительный анализ функциональных возможностей двух морфопроцессоров для русского языка.
3) Выполнить статистический анализ двух текстов на русском языке, программно вычислив 5-7 его статистических характеристик (общестатистических, морфологических, лексических) на базе предварительного морфологического анализа слов текста.
4) Для заданного слова русского языка исследовать временные изменения частоты его употребления слова и смысла, рассмотрев его значения и толкования в различных толковых словарях, в Национальном корпусе русского языка, а также в яндекс-новостях.

Примеры самостоятельных аудиторных работ

1. Для заданной словоформы найти результат лемматизации. Указать также результат полного морфологического анализа.
2. Для заданного предложения текста указать количество словоупотреблений, число различных, число различных лемм.
3. Для заданного предложения построить синтаксическое дерево зависимостей и дерево составляющих.

Вопросы для оценки качества освоения дисциплины

Тема 1.

1. Укажите основные особенности и сложности естественных языков.
2. В чем суть явления полисемии? омонимии? Приведите примеры.
3. Перечислите основные этапы обработки текста в системах АОТ.
4. Какие лингвистические ресурсы используются в лингвистических процессорах?
5. Укажите типичные приложения методов автоматической обработки текстов.

Тема 2.

6. В чем заключается этап графематического анализа текста?
7. Что такое морфема? аффикс?
8. Чем основа слова отличается от корня? Приведите примеры.
9. Что такое словоизменительная парадигма?
10. В чем заключается лемматизация?
11. Приведите пример морфологической омонимии.
12. Чем лемма отличается от лексемы?
13. Назовите виды морфологического анализа.

Тема 3.

14. Как определяется статистика словоупотреблений в текстах?
15. Что такое биграмма? триграмма?
16. Объясните смысл закона Ципфа-Мальдельброта.
17. В чем отличие коллекции текстов от корпуса?
18. Какие бывают типы разметки в корпусе текстов?
19. Что такое статистическая языковая модель?
20. Какие статистические меры применяются для извлечения словосочетаний?

Тема 4.

21. Что такое синтаксическое дерево?
22. В чем отличие деревьев составляющих от деревьев зависимостей?
23. Что такое модель управления слова-предиката? Приведите примеры.
24. Какие методы синтаксического разбора вы знаете?
25. В чем состоит синтаксическая сегментация текста?
26. Укажите особенности семантической сети как способа представления смысла текста.
27. Что такое семантический падеж?
28. Как семантические падежи используются при анализе текста?

29. Назовите отличительные характеристики связного текста.
30. Что такое анафорическая ссылка?

Тема 5.

31. Какие виды смысловых связей лексических единиц вы знаете?
32. Что такое тезаурус?
33. Охарактеризуйте понятие онтологии. Приведите пример.
34. Какие виды онтологий бывают?
35. Какие семантические связи представлены в системе WordNet?

Тема 6.

36. Укажите приложения АОТ, в которых нужен морфологический анализ.
37. В каких приложениях АОТ применяется синтаксический анализ?
38. Что такое индексация текста?
39. Какие модели информационного поиска вы знаете?
40. В чем заключается задача классификации текстов?
41. Чем классификация текстов отличается от кластеризации?
42. Что такое рубрицирование текстов?
43. Какие бывают стратегии машинного перевода?
44. Укажите особенности задачи извлечения информации из текстов.

Примеры вопросов в письменной контрольной/экзаменационной работе:

1. Определите, есть ли в предложении омонимичные словоформы, и если есть, укажите для одной из них все варианты леммы и морфологические характеристики:

   Пила – инструмент со множеством резцов
2. Покажите на примере, чем словоупотребление отличается от словоформы.
3. Что такое N-грамма? Перечислите все символьные триграммы в словосочетании *на бал*.
4. Объясните понятие валентности слова-предиката. Приведите примеры трех слов-предикатов разных частей речи с указанием для них валентностей.
5. На примере 2-5 слов русского языка покажите возможные виды связей лексем в лингвистических онтологиях.
6. Охарактеризуйте модель булевского поиска в массиве документов.
7. Поясните смысл показателей *idf* и *tf.idf*.
8. Сравните два основных подхода к машинный переводу (на основе лингвистических правил, статистический перевод).
9. Укажите основные этапы обработки текста при извлечении информации в подходе, основанном на правилах.
10. Назовите и кратко охарактеризуйте две задачи АОТ, которые можно решать с помощью информации о частотах употребления слов

Примеры задач в письменной контрольной/экзаменационной работе:

1) Для заданной фразы построить возможные синтаксические деревья зависимостей.
2) По заданному фрагменту текста на русском языке построить семантическую сеть, отражающую смысл фразы.

V. РЕСУРСЫ

1. Основная литература

1. Автоматическая обработка текстов на естественном языке и анализ данных: учеб. пособие / Большакова Е.И. и др. – М.: Изд-во НИУ ВШЭ, 2017 – URL: https://miem.hse.ru/clschool/the_book
2. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Большакова Е.И. и др. – М.: МИЭМ, 2011 – URL: http://clschool.miem.edu.ru/uploads/swfupload/files/98e8cdfb0288b275a3197626ffe06e277a03d43d.pdf

2. Дополнительная литература

1. Барсегян А.А. и др. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP – 2-е изд. – СПб.: БХВ-Петербург, 2008.
2. Васильев В. Г., Кривенко М. П. Методы автоматизированной обработки текстов. – М.: ИПИ РАН, 2008.
3. Ингерсолл Г.С., Мортон Т.С., Фэррис Э.Л. Обработка неструктурированных текстов. Поиск, организация и манипулирование / Пер. с англ. – М.: ДМК Пресс, 2015.
4. Леонтьева Н. Н. Автоматическое понимание текстов: Системы, модели, ресурсы: Учебное пособие – М.: Академия, 2006.
5. Лингвистический энциклопедический словарь / Гл. ред. В.Н.Ярцева, 2-ое изд., дополненное – М.: Научное издательство "Большая Российская энциклопедия", 2002.
6. Лукашевич Н.В. Тезаурусы в задачах информационного поиска. – М.: Изд-во Московского университета, 2011.
7. Маннинг К., Рагхаван П., Шютце Х. Введение в информационный поиск. — Вильямс, 2011.
8. Мельчук И.А. Язык: от смысла к тексту – М.: Языки славянской культуры, 2012.
9. Прикладная и компьютерная лингвистика / Под ред. Николаева И.С. и др. – М.: ЛЕНАНД, 2016.
10. Jurafsky D., Martin J. Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Prentice Hall, 2000.

3. Программное обеспечение

№ п/п	Наименование	Условия доступа
1.	Microsoft Windows 7 Professional RUS Microsoft Windows 10 Microsoft Windows 8.1 Professional RUS	*Из внутренней сети университета (договор)*
2.	Microsoft Office Professional Plus 2010	*Из внутренней сети университета (договор)*

4. Интернет-ресурсы (электронные образовательные ресурсы)

№ п/п	Наименование	Условия доступа
	Интернет-ресурсы	
1.	ДИАЛИНГ-АОТ	URL: http://www.aot.ru
2.	Система WordNet	URL:http://wordnetweb.princeton.edu/perl/webwn
3.	Тезаурус РУТЕЗ	URL:http://www.labinform.ru/pub/ruthes/index.htm

5. Материально-техническое обеспечение дисциплины

Учебные аудитории для лекционных и семинарских занятий по дисциплине обеспечивают использование и демонстрацию тематических иллюстраций по программе дисциплины в составе:

- ПЭВМ с доступом в Интернет (операционная система, офисные программы);
- мультимедийный проектор с дистанционным управлением.

Автор программы: Большакова Е.И., канд. физ.-мат. наук, доцент, eibolshakova@hse.ru