Федеральное государственное автономное образовательное учреждение высшего образования
"Национальный исследовательский университет
"Высшая школа экономики"

Факультет Компьютерных наук

Департамент Анализа Данных и Искусственного Интеллекта

Рабочая программа дисциплины «Автоматическая обработка текстов»

для образовательной программы «Прикладная математика и информатика»
направления подготовки 01.03.02. Прикладная математика и информатика
уровень бакалавриат

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

Одобрена на заседании департамента анализа данных и искусственного интеллекта
«_____»_____2017 г.
Руководитель департамента анализа данных и искусственного интеллекта
С.О. Кузнецов_____

Утверждена Академическим советом образовательной программы
«_____»_____2017 г., № протокола_____

Академический руководитель образовательной программы А.С. Конушин____

Москва, 2017

1 Область применения и нормативные ссылки

Настоящая программа учебной дисциплины «Автоматическая обработка текстов» устанавливает минимальные требования к результатам обучения студента и определяет содержание и виды учебных занятий и отчетности.

Программа предназначена для преподавателей, ведущих данную дисциплину, учебных ассистентов и студентов четвертого года обучения в бакалавриате по направлению 01.03.02 «Прикладная математика и информатика».

Программа учебной дисциплины разработана в соответствии с:

- Образовательным стандартом ВПО ГОБУ НИУ ВШЭ;
- Образовательной программой подготовки бакалавра по направлению 01.03.02 «Прикладная математика и информатика»;
- Рабочим учебным планом подготовки бакалавра по направлению 01.03.02, утвержденным в 2017 г.

2 Цели освоения дисциплины

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. Знания принципов компьютерной обработки текстов и навыки использования соответствующих программных средств необходимы в профессиональной деятельности специалистов по прикладной математике и информатике.

3 Компетенции обучающегося, формируемые в результате освоения дисциплины

В результате изучения дисциплины студенты должны:

- Знать основные особенности неструктурированных текстов на ЕЯ и принципы их графематического, морфологического, синтаксического и статистического анализа;
- Понимать ограничения существующих 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 АОТ;
- Уметь применять готовые программные модули анализа текстов и открытые лингвистические ресурсы для решения частных задач АОТ

В результате освоения дисциплины студент развивает и осваивает компетенции:

Компетенция	Код по ОС ВШЭ	Уровень формирования	Дескрипторы – основные признаки освоения (показатели достижения результата)	Формы и методы обучения, способствующие формированию и развитию компетенции	Форма контроля уровня сформированности
Способен понимать, совершенствовать и применять современный математический аппарат	ПК-3	СД, МЦ	Студент понимает особенности формальных моделей текста на естественном языке, принципы их построения и оценки	Лекции и семинары, включающие анализ принципов работы компьютерных моделей АОТ на примерах известных программных модулей и систем	Самостоятельные аудиторные работы, письменная контрольная работа

Компетенция	Код по ОС ВШЭ	Уровень формирования	Дескрипторы – основные признаки освоения (показатели достижения результата)	Формы и методы обучения, способствующие формированию и развитию компетенции	Форма контроля уровня сформированности
Способен провести сбор, обработку и анализ данных с использованием существующих методов машинного обучения	ПК-7	РБ, СД	Студент демонстрирует знание основных видов лингвистических ресурсов и принципов их применения для построения моделей АОТ на базе машинного обучения	Домашние практические задания по изучению открытых лингвистических ресурсов и их применению для анализа текстовых данных	Оценка домашних практических заданий
Способен разработать математическую модель и провести её анализ для поставленной теоретической или прикладной задачи	ПК-8	РБ,СД	Студент владеет основными моделями и методами АОТ, используемыми в прикладных информационных системах, а также навыками их экспериментальной оценки	Лекции, семинары, самостоятельные аудиторные работы и домашние практические задания по построению моделей АОТ для конкретных прикладных задач	Самостоятельные аудиторные работы, письменная контрольная работа
Способен разработать и реализовать в виде программного модуля алгоритм решения поставленной теоретической или прикладной задачи на основе математической модели	ПК-9	СД	Студент составляет программы на языке высокого уровня на базе библиотечных модулей для проведения вычислений по моделям автоматической обработки текста	Домашние практические задания по реализации в виде программ различных моделей обработки текстов, их тестировани, а также проведение с их помощью экспериментов	Оценка домашних практических заданий

4 Место дисциплины в структуре образовательной программы

Учебная дисциплина «Автоматическая обработка текстов» является обязательной дисциплиной специализации «Анализ данных и интеллектуальные системы» в учебной программе подготовки бакалавра направления 01.03.02 «Прикладная математика и информатика».

Изучение данной дисциплины требует предварительных знаний по дисциплинам:

- Дискретная математика;
- Алгебра;
- Теория вероятностей и математическая статистика;
- Основы программирования;
- Алгоритмы и структуры данных;
- Компьютерные системы;
- Технологии программирования.

Для освоения учебной дисциплины студенты должны знать основы формальных логических систем, принципы получения вероятностных и статистических оценок, владеть навыками формализации и анализа простых формальных моделей, 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.

5 Тематический план учебной дисциплины

№	Название темы	Всего часов по дисциплине	Аудиторные часы		Самостоятельная работа
			Лекции	Семинары	
1	Введение	14	2	2	10
2	Начальные этапы обработки текста	30	4	4	22
3	Статистические характеристики текстов и корпусная лингвистика	28	4	2	22
4	Подходы к автоматическому анализу синтаксиса и семантики текстов	34	4	6	24
5	Лингвистические ресурсы	30	6	4	20
6	Прикладные задачи АОТ	54	10	12	32
	Итого	190	30	30	130

6 Формы контроля знаний студентов

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

7 Критерии оценки знаний, навыков

На текущем и итоговом контроле студент должен продемонстрировать владение основными понятиями и навыками по пройденным темам дисциплины.

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.

8 Содержание дисциплины

Тема 1. Введение

1. Автоматическая обработка текстов на естественном языке (ЕЯ): основные задачи и особенности направления, связь со смежными научными дисциплинами. 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 (машинный перевод, информационный поиск, классификация текстов, реферирование и аннотирование, извлечение информации и знаний, анализ тональности, автоматизация редактирования текстов).

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, Глава 1.
2. Прикладная и компьютерная лингвистика / Под ред. Николаева И.С. и др. – М.: ЛЕНАНД, 2016.

Дополнительная литература

1. Васильев В. Г., Кривенко М. П. Методы автоматизированной обработки текстов. – М.: ИПИ РАН, 2008.
2. Касевич В.Б. Элементы общей лингвистики. – М., Наука, 1977.
3. Леонтьева Н. Н. Автоматическое понимание текстов: Системы, модели, ресурсы: Учебное пособие – М.: Академия, 2006.
4. Мельчук И.А. Язык: от смысла к тексту – М.: Языки славянской культуры, 2012.
5. Bolshakov, I.A., Gelbukh A. Computational Linguistics. Models, Resources, Applications. Mexico, IPN, 2004.

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 морфологического анализа: стемминг, лемматизация, полный морфоанализ. Принципы морфоанализа на базе словаря основ или словаря словоформ. Морфологические процессоры для русского языка.

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, Глава 2.
2. Прикладная и компьютерная лингвистика / Под ред. Николаева И.С. и др. – М.: ЛЕНАНД, 2016, Глава 1.

Дополнительная литература

1. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Большакова Е.И. и др. – М.: МИЭМ, 2011.
2. Болховитянов А.В., Гусев А.В., Чеповский А.М. Морфологические модели компьютерной лингвистики: учеб. пособие – М. МГУП, 2010.
3. Васильев В. Г., Кривенко М. П. Методы автоматизированной обработки текстов. – М.: ИПИ РАН, 2008
4. Лингвистический энциклопедический словарь / Гл. ред. В.Н.Ярцева, 2-ое изд., дополненное – М.: Научное издательство "Большая Российская энциклопедия", 2002.
5. Jurafsky D., Martin J. Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Prentice Hall, 2000.

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. Обзор корпусов. 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. Меры устойчивости сочетаний.

Основная литература

1. Пиотровский Р.Г. , Бектаев К.Б., Пиотровская А.А. Математическая лингвистика. – М.: Высшая школа, 1977.
2. Прикладная и компьютерная лингвистика / Под ред. Николаева И.С. и др. – М.: ЛЕНАНД, 2016, Глава 6.
3. Jurafsky D., Martin J. Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Prentice Hall, 2000.

Дополнительная литература

1. Национальный Корпус Русского Языка. http://ruscorpora.ru
2. Открытый корпус русского языка OpenCorpora. http://opencorpora.org
3. Чатуев М.Б., Чеповский А.М. Частотные методы в компьютерной лингвистике: учеб. пособие – М. МГУП, 2011.
4. Biber, D., Conrad S., and Reppen D. Corpus Linguistics. Investigating Language Structure and Use. Cambridge University Press, Cambridge, 1998.

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. Понятия синтаксического предиката и модели управления. 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. Смысловая и синтаксическая связность. Анафорические ссылки, лексические повторы, дискурсивные слова. Сверхфразовые единства. Композиционные и дискурсивные особенности текстов, их учет в задачах АОТ.

Основная литература

1. Прикладная и компьютерная лингвистика / Под ред. Николаева И.С. и др. – М.: ЛЕНАНД, 2016, Глава 2.
2. Васильев В. Г., Кривенко М. П. Методы автоматизированной обработки текстов. – М.: ИПИ РАН, 2008.
3. Леонтьева Н. Н. Автоматическое понимание текстов: Системы, модели, ресурсы: Учебное пособие – М.: Академия, 2006.

Дополнительная литература

1. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Большакова Е.И. и др. – М.: МИЭМ, 2011.
2. Апресян Ю.Д. и др. Лингвистическое обеспечение системы ЭТАП-2. М.: Наука, 1989.
3. Лингвистический энциклопедический словарь / Гл. ред. В.Н.Ярцева, 2-ое изд., дополненное – М.: Научное издательство "Большая Российская энциклопедия", 2002.

4. Bolshakov, I.A., Gelbukh A. Computational Linguistics. Models, Resources, Applications. Mexico, IPN, 2004.
5. Jurafsky D., Martin J. Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Prentice Hall, 2000.

Тема 5. Лингвистические ресурсы

1. Словари для автоматической обработки текстов. Виды словарей. Терминологические словари и автоматизация их построения.

2. Смысловые (парадигматические) отношения лексических единиц. Синонимия и лексическая многозначность. Тезаурус как словарь с семантическими связями единиц. Информационно-поисковые тезаурусы и рубрикаторы, их применение в АОТ.

3. Понятие онтологии. Лингвистические онтологии. Лексические ресурсы WordNet и EvroNet. Методологии создания онтологий.

Основная литература

1. Лукашевич Н.В. Тезаурусы в задачах информационного поиска. – М.: Изд-во Московского университета, 2011.
2. Прикладная и компьютерная лингвистика / Под ред. Николаева И.С. и др. – М.: ЛЕНАНД, 2016, Глава 3.
3. Word Net: an Electronic Lexical Database. /Edit. by Christiane Fellbaum. Cambridge, MIT Press, 1998.

Дополнительная литература

1. Кобозева И.М. Лингвистическая семантика. – М., 2009.
2. Леонтьева Н. Н. Автоматическое понимание текстов: Системы, модели, ресурсы: Учебное пособие – М.: Академия, 2006.

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 текстов как типичная задача обработки текстов в области Text Mining. Обзор методов машинной классификации. Выбор признаков и метрик. Особенности кластеризации текстов. Рубрицирование текстовых документов. Обзор задач АОТ, решаемых на основе классификации текстов.

3. Автоматическое реферирование и аннотирование документов как смежные задачи информационного поиска. Основные стратегии сжатия текста. Типы аннотаций. Обзорное реферирование. Оценка качества аннотаций.

4. Машинный перевод. Стратегии машинного перевода, основанного на лингвистических правилах. Статистический машинный перевод: особенности и виды. Принципы создания статистического переводчика.

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

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

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, Главы 1,3,4.
2. Прикладная и компьютерная лингвистика / Под ред. Николаева И.С. и др. – М.: ЛЕНАНД, 2016, Часть 2.

3. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Большакова Е.И. и др. – М.: МИЭМ, 2011, Часть V.
4. Маннинг К., Рагхаван П., Шютце Х. Введение в информационный поиск. — Вильямс, 2011.

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. Bolshakov, I.A., Gelbukh A. Computational Linguistics. Models, Resources, Applications. Mexico, IPN, 2004.
5. Jurafsky D., Martin J. Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Prentice Hall, 2000.

9 Образовательные технологии

В преподавании данной дисциплины сочетаются:

- лекции в форме презентаций (которые высылаются студентам для их самостоятельной работы и подготовки к письменной контрольной работе);
- самостоятельные аудиторные работы и практические занятия по изучению интернет-ресурсов по АОТ, а также прикладного программного обеспечения;
- домашние задания на применение изученных ресурсов и методов автоматической обработки текстов.

10 Оценочные средства для текущего контроля и аттестации студента

10.1 Оценочные средства для оценки качества освоения дисциплины в ходе текущего контроля

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 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.

10.2 Примеры заданий промежуточной (итоговой) аттестации

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.

11 Порядок формирования оценок по дисциплине

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 $O_{д/з}$ и $O_{сам.\ аудит.\ работа}$, 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:

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

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

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

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

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

12 Учебно-методическое и информационное обеспечение дисциплины

12.1 Базовый учебник – ридер «Автоматическая обработка текста», составленный по следующим источникам:

1. Автоматическая обработка текстов на естественном языке и анализ данных: учеб. пособие / Большакова Е.И. и др. – М.: Изд-во НИУ ВШЭ, 2017 – 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 – http://clschool.miem.edu.ru/uploads/swfupload/files/98e8cdfb0288b275a3197626ffe06e277a03d43d.pdf
3. Прикладная и компьютерная лингвистика / Под ред. Николаева И.С. и др. – М.: ЛЕНАНД, 2016.

12.2 Дополнительная литература

1. Апресян Ю.Д. и др. Лингвистическое обеспечение системы ЭТАП-2. М.: Наука, 1989.
2. Барсегян А.А. и др. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP – 2-е изд. – СПб.: БХВ-Петербург, 2008.
3. Болховитянов А.В., Гусев А.В., Чеповский А.М. Морфологические модели компьютерной лингвистики: учеб. пособие – М. МГУП, 2010.
4. Васильев В. Г., Кривенко М. П. Методы автоматизированной обработки текстов. – М.: ИПИ РАН, 2008.
5. Ингерсолл Г.С., Мортон Т.С., Фэррис Э.Л. Обработка неструктурированных текстов. Поиск, организация и манипулирование / Пер. с англ. – М.: ДМК Пресс, 2015.
6. Касевич В.Б. Элементы общей лингвистики. — М., Наука, 1977.
7. Кобозева И.М. Лингвистическая семантика. – М., 2009.
8. Леонтьева Н. Н. Автоматическое понимание текстов: Системы, модели, ресурсы: Учебное пособие – М.: Академия, 2006.
9. Лукашевич Н.В. Тезаурусы в задачах информационного поиска. – М.: Изд-во Московского университета, 2011.
10. Маннинг К., Рагхаван П., Шютце Х. Введение в информационный поиск. — Вильямс, 2011.
11. Мельчук И.А. Язык: от смысла к тексту – М.: Языки славянской культуры, 2012.
12. Пиотровский Р.Г. , Бектаев К.Б., Пиотровская А.А. Математическая лингвистика. – М.: Высшая школа, 1977.
13. Чатуев М.Б., Чеповский А.М. Частотные методы в компьютерной лингвистике: учеб. пособие – М. МГУП, 2011.
14. Bolshakov, I.A., Gelbukh A. Computational Linguistics. Models, Resources, Applications. Mexico, IPN, 2004.
15. Jurafsky D., Martin J. Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Prentice Hall, 2000.

12.3 Справочники, словари, энциклопедии

Лингвистический энциклопедический словарь / Гл. ред. В.Н.Ярцева, 2-ое изд., дополненное – М.: Научное издательство "Большая Российская энциклопедия", 2002.

13 Материально-техническое обеспечение дисциплины

Для лекций и практических занятий по темам дисциплины используется проектор и компьютеры с выходом в сеть Интернет.

Автор программы: _____/ Большакова Е.И. /