

МИНОБРНАУКИ РОССИИ



Федеральное государственное бюджетное образовательное учреждение
высшего образования

**«Российский государственный гуманитарный университет»
(ФГБОУ ВО «РГГУ»)**

ОТДЕЛЕНИЕ ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМ В ГУМАНИТАРНОЙ СФЕРЕ
Кафедра математики, логики и интеллектуальных систем в гуманитарной сфере

ВЫЧИСЛИТЕЛЬНАЯ ЛИНГВИСТИКА

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

45.03.04 Интеллектуальные системы в гуманитарной сфере

Разработка и программирование интеллектуальных систем
Уровень высшего образования: бакалавриат

Форма обучения очная

РПД адаптирована для лиц
с ограниченными возможностями
здоровья и инвалидов

Москва 2024

Вычислительная лингвистика

Рабочая программа дисциплины

Составитель:

Доктор физико-математических наук, профессор

М. Р. Пентус

.....

УТВЕРЖДЕНО

Протокол заседания кафедры МЛиИС

№ 6 от 08.02.2024

ОГЛАВЛЕНИЕ

1. Пояснительная записка

1.1 Цель и задачи дисциплины

1.2. Перечень планируемых результатов обучения по дисциплине, соотнесенных с индикаторами достижения компетенций

1.3. Место дисциплины в структуре образовательной программы

2. Структура дисциплины

3. Содержание дисциплины

4. Образовательные технологии

5. Оценка планируемых результатов обучения

5.1. Система оценивания

5.2. Критерии выставления оценок

5.3. Оценочные средства (материалы) для текущего контроля успеваемости, промежуточной аттестации обучающихся по дисциплине

6. Учебно-методическое и информационное обеспечение дисциплины

6.1. Список источников и литературы

6.2. Перечень ресурсов информационно-телекоммуникационной сети «Интернет»

7. Материально-техническое обеспечение дисциплины

8. Обеспечение образовательного процесса для лиц с ограниченными возможностями здоровья

9. Методические материалы

9.1. Планы практических (семинарских, лабораторных) занятий

9.2. Методические указания для обучающихся по освоению дисциплины

Приложения

Приложение 1. Аннотация дисциплины

Приложение 2. Лист изменений

1. Пояснительная записка

1.1. Цель и задачи дисциплины

Цель дисциплины: познакомить студентов с возможностями вычислительной лингвистики в компьютерных системах.

Задача дисциплины: освоение базовых методов и средств вычислительной лингвистики и их использования в информационных и интеллектуальных системах.

1.2. Перечень планируемых результатов обучения по дисциплине, соотнесенных с индикаторами достижения компетенций:

Компетенция (код и наименование)	Индикаторы компетенций (код и наименование)	Результаты обучения
ПК-7. Способен к участию в разработке архитектур информационных и интеллектуальных систем	ПК-7.1. Знает способы представления архитектуры информационных и интеллектуальных систем и примеры типичных архитектур информационных и интеллектуальных систем; ПК-7.2. Умеет применять CASE-технологии для разработки и наглядного представления архитектуры информационных и интеллектуальных систем; ПК-7.3. Имеет практический опыт участия в разработке архитектуры интеллектуальных и информационных систем.	Знать: основные виды моделей, используемых для формализации морфологии, синтаксиса и семантики естественных и искусственных языков; подходы к построению языковых процессоров Уметь: разрабатывать и отлаживать программы, связанные с обработкой текста. Владеть: навыками применения основных видов языковых ресурсов при разработке приложений вычислительной лингвистики.

1.3. Место дисциплины в структуре образовательной программы

Дисциплина «Вычислительная лингвистика» входит в состав части, формируемой участниками образовательных отношений, учебного плана дисциплин по выбору.

Для освоения дисциплины необходимы знания, умения и владения, сформированные в ходе изучения следующих дисциплин и прохождения практик: высшей алгебры, математической логики, информатики, программирования, практического курса английского языка.

В результате освоения дисциплины формируются компетенции, необходимые для изучения следующих дисциплин и прохождения практик: Методология проектирования интеллектуальных систем.

2. Структура дисциплины

Общая трудоёмкость дисциплины составляет 3 з.е., 108 академических часов.

Структура дисциплины для очной формы обучения

Объем дисциплины в форме контактной работы обучающихся с педагогическими работниками и (или) лицами, привлекаемыми к реализации образовательной программы на иных условиях, при проведении учебных занятий:

Семестр	Тип учебных занятий	Количество часов
7	Лекции	10
7	Семинары/лабораторные работы	32
Всего:		42

Объем дисциплины в форме самостоятельной работы обучающихся составляет 66 академических часов.

3. Содержание дисциплины

№	Наименование раздела дисциплины	Содержание
1	Краткое введение в проблематику компьютерной лингвистики	Роль автоматической обработки текста и связанных с ней приложений; компьютерная лингвистика и автоматический перевод Вычислительная лингвистика и искусственный интеллект, связь вычислительной лингвистики с лингвистикой, информатикой, математикой.
2	Подходы к формальному описанию естественных и искусственных языков	Аналитические и порождающие грамматики; контекстно-свободные грамматики, категориальные грамматики, трансформационные грамматики, многоуровневая двунаправленная модель Смысл ↔ Текст. Валентности и актанты, GPSG, HPSG, унификационная грамматика, деревья и грамматики зависимостей
3	Лексический, морфологический и синтаксический анализ языков	Лексический и морфологический анализ и конечные автоматы; различные подходы к синтаксическому анализу, в частности, применимые к языкам, описываемым контекстно-свободными и другими грамматиками. Уровни модели и уровни языкового строения в лингвистической теории
4	Подходы к описанию семантики языков	Использование логики предикатов, семантических сетей для формального представления смыслов текстов. Более общие средства: модальные и интенциональные логики
5	Статистические методы обработки текстов	

		Вероятностные грамматики; частотные словари; метод n-грамм; байесовские модели методы кластеризации, векторные машины
6	Машинное обучение	Обучение по примерам; индуктивное логическое программирование; метод Вапника-Червоненкиса, использование байесовых моделей и скрытых Марковских цепей, деревьев решений и векторных машин
7	Ресурсы и приложения компьютерной лингвистики	Различные виды лингвистических корпусов с разметкой; информационный поиск, аннотирование и извлечение информации Использование естественного языка в интерфейсе, анализ и генерация текста, машинный перевод.

4. Образовательные технологии

Образовательные технологии

<i>№ п/ п</i>	<i>Наименование раздела</i>	<i>Виды учебной работы</i>	<i>Информационные и образовательные технологии</i>
<i>1</i>	<i>2</i>	<i>3</i>	<i>5</i>
1	Краткое введение в проблематику компьютерной лингвистики	Лекция 1 Прак. зан. 1-2 Самостоятельная работа	Вводная лекция-беседа. Семинар-обсуждение Работа с электронным конспектом и интернет-ресурсами. Консультирование посредством электронной почты
2	Подходы к формальному описанию естественных и искусственных языков	Лекция 2 Прак. зан 3-4 Самостоятельная работа	Теоретическая лекция. Семинар-обсуждение. Работа с электронным конспектом и интернет-ресурсами..
3	Лексический, морфологический и синтаксический анализ языков <Смысл<=>Текст>	Лекция 3 Прак. зан 5-6 Самостоятельная работа	Теоретическая лекция. Работа на компьютере с использованием системы NLTK. Работа с электронным конспектом и интернет-ресурсами.
4	Подходы к описанию семантики языков	Лекция 4 Прак. зан. 7-8 лаб 5 Самостоятельная работа	Теоретическая лекция. Работа на компьютере с использованием системы NLTK. Работа на компьютере с использованием системы NLTK. Работа с электронным конспектом и интернет-ресурсами. Консультирование и приём домашних заданий посредством электронной почты

5	Статистические методы обработки текстов	Лекция 5 Прак. зан. 9-10 Самостоятельная работа	Теоретическая лекция. Работа на компьютере с использованием системы NLTK. Работа с электронным конспектом и интернет-ресурсами. Консультирование и приём домашних заданий посредством электронной почты
6	Машинное обучение	Лекция 6 Прак. зан. 11-14 Самостоятельная работа	Теоретическая лекция. Работа на компьютере с использованием системы NLTK. Консультирование и приём домашних заданий посредством электронной почты
7	Ресурсы и приложения компьютерной лингвистики	Лекция 7-8 Прак. зан. 15-20 Самостоятельная работа	Теоретическая лекция. Работа на компьютере с использованием системы NLTK. Консультирование и приём домашних заданий посредством электронной почты

В период временного приостановления посещения обучающимися помещений и территории РГГУ. для организации учебного процесса с применением электронного обучения и дистанционных образовательных технологий могут быть использованы следующие образовательные технологии:

- видео-лекции;
- онлайн-лекции в режиме реального времени;
- электронные учебники, учебные пособия, научные издания в электронном виде и доступ к иным электронным образовательным ресурсам;
- системы для электронного тестирования;
- консультации с использованием телекоммуникационных средств.

5. Оценка планируемых результатов обучения

5.1. Система оценивания

<i>Форма контроля</i>	<i>Макс. количество баллов</i>	
	<i>За одну работу</i>	<i>Всего</i>
Текущий контроль:		
• Опрос (1—5)	5 баллов	20 баллов
• дом. задание (темы 1—5)	5 баллов	20 баллов
• контр. работа (темы 1—3)	20 баллов	20 баллов
Промежуточная аттестация экзамен)		40 баллов
Итого за семестр (дисциплину)		100 баллов
Текущий контроль:		
• опрос (6—8)	5 баллов	20 баллов
• дом. задание (темы 6—8)	5 баллов	20 баллов
• контр. работа (темы 6—7)	20 баллов	20 баллов

Промежуточная аттестация (экзамен)		40 баллов
Итого за семестр (дисциплину)		100 баллов

Оценка «зачтено» выставляется обучающемуся, набравшему не менее 50 баллов в результате суммирования баллов, полученных при текущем контроле и промежуточной аттестации. Полученный совокупный результат (максимум 100 баллов) конвертируется в традиционную шкалу оценок и в шкалу оценок Европейской системы переноса и накопления кредитов (European Credit Transfer System; далее – ECTS) в соответствии с таблицей:

<i>100-балльная шкала</i>	<i>Традиционная шкала</i>		<i>Шкала ECTS</i>
95 – 100	отлично	зачтено	A
83 – 94			B
68 – 82	хорошо		C
56 – 67	удовлетворительно		D
50 – 55			E
20 – 49	неудовлетворительно	не зачтено	FX
0 – 19			F

Полученный совокупный результат конвертируется в традиционную шкалу оценок и в шкалу оценок Европейской системы переноса и накопления кредитов (European Credit Transfer System; далее – ECTS) в соответствии с таблицей:

100-балльная шкала	Традиционная шкала		Шкала ECTS
95 – 100	отлично	зачтено	A
83 – 94			B
68 – 82	хорошо		C
56 – 67	удовлетворительно		D
50 – 55			E
20 – 49	неудовлетворительно	не зачтено	FX
0 – 19			F

5.2. Критерии выставления оценки по дисциплине

Баллы/ Шкала ECTS	Оценка по дисциплине	Критерии оценки результатов обучения по дисциплине
100-83/ А,В	«отлично»/ «зачтено (отлично)»/ «зачтено»	<p>Выставляется обучающемуся, если он глубоко и прочно усвоил теоретический и практический материал, может продемонстрировать это на занятиях и в ходе промежуточной аттестации.</p> <p>Обучающийся исчерпывающе и логически стройно излагает учебный материал, умеет увязывать теорию с практикой, справляется с решением задач профессиональной направленности высокого уровня сложности, правильно обосновывает принятые решения.</p> <p>Свободно ориентируется в учебной и профессиональной литературе.</p> <p>Оценка по дисциплине выставляется обучающемуся с учётом результатов текущей и промежуточной аттестации.</p> <p>Компетенции, закреплённые за дисциплиной, сформированы на уровне – «высокий».</p>
82-68/ С	«хорошо»/ «зачтено (хорошо)»/ «зачтено»	<p>Выставляется обучающемуся, если он знает теоретический и практический материал, грамотно и по существу излагает его на занятиях и в ходе промежуточной аттестации, не допуская существенных неточностей.</p> <p>Обучающийся правильно применяет теоретические положения при решении практических задач профессиональной направленности разного уровня сложности, владеет необходимыми для этого навыками и приёмами.</p> <p>Достаточно хорошо ориентируется в учебной и профессиональной литературе.</p> <p>Оценка по дисциплине выставляется обучающемуся с учётом результатов текущей и промежуточной аттестации.</p> <p>Компетенции, закреплённые за дисциплиной, сформированы на уровне – «хороший».</p>
67-50/ D,E	«удовлетвори- тельно»/ «зачтено (удовлетвори- тельно)»/ «зачтено»	<p>Выставляется обучающемуся, если он знает на базовом уровне теоретический и практический материал, допускает отдельные ошибки при его изложении на занятиях и в ходе промежуточной аттестации.</p> <p>Обучающийся испытывает определённые затруднения в применении теоретических положений при решении практических задач профессиональной направленности стандартного уровня сложности, владеет необходимыми для этого базовыми навыками и приёмами.</p> <p>Демонстрирует достаточный уровень знания учебной литературы по дисциплине.</p>

		Оценка по дисциплине выставляются обучающемуся с учётом результатов текущей и промежуточной аттестации. Компетенции, закреплённые за дисциплиной, сформированы на уровне – «достаточный».
49-0/ F,FX	«неудовлетворительно»/ не зачтено	Выставляется обучающемуся, если он не знает на базовом уровне теоретический и практический материал, допускает грубые ошибки при его изложении на занятиях и в ходе промежуточной аттестации. Обучающийся испытывает серьёзные затруднения в применении теоретических положений при решении практических задач профессиональной направленности стандартного уровня сложности, не владеет необходимыми для этого навыками и приёмами. Демонстрирует фрагментарные знания учебной литературы по дисциплине. Оценка по дисциплине выставляются обучающемуся с учётом результатов текущей и промежуточной аттестации. Компетенции на уровне «достаточный», закреплённые за дисциплиной, не сформированы.

5.3. Оценочные средства (материалы) для текущего контроля успеваемости, промежуточной аттестации обучающихся по дисциплине

5.3.1. Образцы заданий для самостоятельного выполнения

- Как построить токенизатор для русскоязычных текстов?
- Применить токенизатор из системы NLTK для разбиения некоторого естественнорусского текста на лексические единицы.
- Перечислить по крайней мере два метода синтаксического анализа для контекстно-свободных языков с кратким описанием.
- Какие синтаксические анализаторы входят в состав NLTK?
- Для некоторой простой категориальной грамматики построить эквивалентную контекстно-свободную грамматику.
- Использовать систему NLTK для вычисления частотных характеристик некоторого естественнорусского текста.
- Найти в тексте первые 10 слов с наиболее частым употреблением.
- Описать способы доступа к корпусам в системе NLTK

5.3.2. Образцы заданий для контрольной работы (экзамена)

Вариант 1

Написать программу перевода арифметических выражений в префиксную форму, используя Кс-грамматику для них

Вариант 2

Написать программу, вычисляющую значение арифметических выражений, используя Кс-грамматику для них

6. Учебно-методическое и информационное обеспечение дисциплины

6.1. Список источников и литературы

о

а) Основная литература

1. Леонтьева Н.Н. Автоматическое понимание текстов : системы, модели, ресурсы : [учеб. пособие] : для студентов лингвист. фак. вузов / . - М. : Академия, 2006. - 302 с. - (Высшее профессиональное образование. Языкознание). - ISBN 5-7695-1842-1 .
2. Вычислительные технологии в лингвистике и филологии : учеб.-метод. пособие / [В. П. Захаров и др.] ; под общ. ред. В. П. Захарова ; С.-Петербург. гос. ун-т, Филол. фак. - , 2012-2013.
3. В. П. Захаров, С. Ю. Богданова Корпусная лингвистика : учебник для студентов направления "Лингвистика" / М-во образования и науки Рос. Федерации, ФГБОУ ВПО "С.-Петербург. гос. ун-т". - Санкт-Петербург : Филол. фак. СПбГУ, 2013.
ISBN 978-5-8465-1343-3.<https://e.lanbook.com/book/139576>
4. Мельчук, И. А. Опыт теории лингвистических моделей «СМЫСЛ - ТЕКСТ» : сборник научных трудов / И. А. Мельчук. — Москва : Языки славянских культур, 1999. — 346 с. — ISBN 5-7859-0078-5. — Текст : электронный // Лань : электронно-библиотечная система. — URL: <https://e.lanbook.com/book/137010>.
5. С.Рассел, П.Норвиг. Искусственный интеллект. Современный подход. Изд. "Вильямс", Москва-Санкт-Петербург-Киев, 2006

б) Дополнительная литература

1. Зубов А.В., Зубова И.И. Основы искусственного интеллекта для лингвистов : учеб. пособие для студентов вузов, обучающихся по специальности "Теорет. и прикладная лингвистика", М. : Логос, 2007. - 319 с.
2. Alpaydin E. Introduction to Machine Learning. MIT Press, 2004.
3. Попов Э.В. Общение с ЭВМ на естественном языке. М. УРСС, 2004.

6.2 Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимый для освоения дисциплины

[https://ru.wikipedia.org/wiki/Вычислительная лингвистика](https://ru.wikipedia.org/wiki/Вычислительная_лингвистика)
<https://www.nltk.org>
[https://ru.wikipedia.org/wiki/Теория «СМЫСЛ-ТЕКСТ»](https://ru.wikipedia.org/wiki/Теория_«СМЫСЛ-ТЕКСТ»)
<https://morphix-nlp.berlios.de>.

7. Материально-техническое обеспечение дисциплины

Учебный класс с хорошей доской, с экраном, мультимедийным проектором и доступом к Интернету.

Состав программного обеспечения:

1. Windows
2. Microsoft Office
3. Python
4. spaCy
5. NLTK

8. Обеспечение образовательного процесса для лиц с ограниченными возможностями здоровья и инвалидов

В ходе реализации дисциплины используются следующие дополнительные методы обучения, текущего контроля успеваемости и промежуточной аттестации обучающихся в зависимости от их индивидуальных особенностей:

- для слепых и слабовидящих:
 - лекции оформляются в виде электронного документа, доступного с помощью компьютера со специализированным программным обеспечением;
 - письменные задания выполняются на компьютере со специализированным программным обеспечением, или могут быть заменены устным ответом;
 - обеспечивается индивидуальное равномерное освещение не менее 300 люкс;
 - для выполнения задания при необходимости предоставляется увеличивающее устройство; возможно также использование собственных увеличивающих устройств;
 - письменные задания оформляются увеличенным шрифтом;
 - экзамен и зачёт проводятся в устной форме или выполняются в письменной форме на компьютере.
- для глухих и слабослышащих:
 - лекции оформляются в виде электронного документа, либо предоставляется звукоусиливающая аппаратура индивидуального пользования;
 - письменные задания выполняются на компьютере в письменной форме;
 - экзамен и зачёт проводятся в письменной форме на компьютере; возможно проведение в форме тестирования.
- для лиц с нарушениями опорно-двигательного аппарата:
 - лекции оформляются в виде электронного документа, доступного с помощью компьютера со специализированным программным обеспечением;
 - письменные задания выполняются на компьютере со специализированным программным обеспечением;
 - экзамен и зачёт проводятся в устной форме или выполняются в письменной форме на компьютере.

При необходимости предусматривается увеличение времени для подготовки ответа.

Процедура проведения промежуточной аттестации для обучающихся устанавливается с учётом их индивидуальных психофизических особенностей. Промежуточная аттестация может проводиться в несколько этапов.

При проведении процедуры оценивания результатов обучения предусматривается использование технических средств, необходимых в связи с индивидуальными особенностями обучающихся. Эти средства могут быть предоставлены университетом, или могут использоваться собственные технические средства.

Проведение процедуры оценивания результатов обучения допускается с использованием дистанционных образовательных технологий.

Обеспечивается доступ к информационным и библиографическим ресурсам в сети Интернет для каждого обучающегося в формах, адаптированных к ограничениям их здоровья и восприятия информации:

- для слепых и слабовидящих:
 - в печатной форме увеличенным шрифтом;
 - в форме электронного документа;

- в форме аудиофайла.
- для глухих и слабослышащих:
- в печатной форме;
- в форме электронного документа.
- для обучающихся с нарушениями опорно-двигательного аппарата:
- в печатной форме;
- в форме электронного документа;
- в форме аудиофайла.

Учебные аудитории для всех видов контактной и самостоятельной работы, научная библиотека и иные помещения для обучения оснащены специальным оборудованием и учебными местами с техническими средствами обучения:

- для слепых и слабовидящих:
 - устройством для сканирования и чтения с камерой SARA CE;
 - дисплеем Брайля PAC Mate 20;
 - принтером Брайля EmBraille ViewPlus;
- для глухих и слабослышащих:
 - автоматизированным рабочим местом для людей с нарушением слуха и слабослышащих;
 - акустический усилитель и колонки;
- для обучающихся с нарушениями опорно-двигательного аппарата:
 - передвижными, регулируемые эргономическими партами СИ-1;
 - компьютерной техникой со специальным программным обеспечением.

9. Методические материалы

9.1. Планы семинарских занятий

Тема 1 (2 ч.) Краткое введение в проблематику компьютерной лингвистики

Цель занятия: Общее ознакомление с проблематикой, изучаемой в вычислительной лингвистике. Форма проведения – обсуждение, опрос.

Вопросы для обсуждения:

Естественный язык как сложная система

Список источников и литературы:

Леонтьева Н.Н. Автоматическое понимание текстов : системы, модели, ресурсы : [учеб. пособие] : для студентов лингвист. фак. вузов / . - М. : Академия, 2006. - 302 с. - (Высшее профессиональное образование. Языкознание).

Перечень ресурсов информационно-телекоммуникационной сети «Интернет»:

[https://ru.wikipedia.org/wiki/Вычислительная лингвистика](https://ru.wikipedia.org/wiki/Вычислительная_лингвистика)

Материально-техническое обеспечение занятия: доска, графические распечатки.

Тема 2 (2 ч.) Подходы к формальному описанию естественных и искусственных языков

Цель занятия: ознакомление с основными подходами к построению формальных моделей языков

Форма проведения – обсуждение, опрос.

Вопросы для обсуждения:

Уровни описания языка: морфология, синтаксис, семантика, прагматика

Список источников и литературы:

Мельчук И. А. Опыт теории лингвистических моделей «Смысл-Текст». Языки русской культуры, 1999..

Перечень ресурсов информационно-телекоммуникационной сети «Интернет»:

[https://ru.wikipedia.org/wiki/Вычислительная лингвистика](https://ru.wikipedia.org/wiki/Вычислительная_лингвистика)

www.nltk.org

[https://ru.wikipedia.org/wiki/Теория «СМЫСЛ-ТЕКСТ»](https://ru.wikipedia.org/wiki/Теория_«СМЫСЛ-ТЕКСТ»)

Материально-техническое обеспечение занятия: доска, графические распечатки.

Тема 3 (4 ч.) Лексический, морфологический и синтаксический анализ языков Цель занятия: Ознакомление с многоуровневой формальной моделью языков Форма проведения –обсуждение, опрос.

Содержание работы: Установка системы NLTK, ориентированной на обработку текстов естественных языков; ознакомление с методами и средствами, которые используются в NLTK для токенизации и синтаксического анализа

Вопросы для обсуждения:

Уровни описания языка: морфологический, синтаксический, семантический

Список источников и литературы:

Мельчук И. А. Опыт теории лингвистических моделей «Смысл-Текст». Языки русской культуры, 1999..

Перечень ресурсов информационно-телекоммуникационной сети «Интернет»:

[https://ru.wikipedia.org/wiki/Вычислительная лингвистика](https://ru.wikipedia.org/wiki/Вычислительная_лингвистика)

www.nltk.org

[https://ru.wikipedia.org/wiki/Теория «СМЫСЛ-ТЕКСТ»](https://ru.wikipedia.org/wiki/Теория_«СМЫСЛ-ТЕКСТ»)

Материально-техническое обеспечение занятия: компьютерный класс.

Тема 4 (4 ч.) Подходы к описанию семантики языков(4 часа)

Цель: изучение средств представления смысла текстов: логика предикатов, семантические сети.

Содержание работы: выявление смысла некоторых простых текстов и его представление в логических терминах.

С.Рассел, П.Норвиг. Искусственный интеллект. Современный подход. Изд. "Вильямс", Москва-Санкт-Петербург-Киев, 2006

[https://ru.wikipedia.org/wiki/Теория «СМЫСЛ-ТЕКСТ»](https://ru.wikipedia.org/wiki/Теория_«СМЫСЛ-ТЕКСТ»)

Тема 5 (8 ч.) Статистические методы обработки текстов

Содержание работы: изучение частотных и вероятностных методов работы с текстами. Использование средств системы NLTK для вычисления частот слов, словосочетаний и n-грамм в текстах. Вероятностные грамматики, байесовы модели.

С.Рассел, П.Норвиг. Искусственный интеллект. Современный подход. Изд. "Вильямс", Москва-Санкт-Петербург-Киев, 2006

www.nltk.org/book

Тема 6 (6 ч.) Машинное обучение

Содержание работы: изучение основных подходов обучения по примерам: метод Вапника-Червоненкиса, использование байесовых моделей.

С. Рассел, П. Норвиг. Искусственный интеллект. Современный подход. Изд. "Вильямс", Москва-Санкт-Петербург-Киев, 2006

Тема 7 (6 ч.) Ресурсы и приложения в вычислительной лингвистике

Цель занятия: приобрести навыки работы с локальными и глобальными лингвистическими ресурсами

Форма проведения – опрос, решение задач.

Вопросы для обсуждения:

1. Виды электронных словарей и лингвистических корпусов с разметкой
2. Доступ к ресурсам для разных языков: национальный корпус русского языка, корпус Gutenberg.

Список источников и литературы:

1. В. П. Захаров, С. Ю. Богданова Корпусная лингвистика : учебник для студентов направления "Лингвистика" / М-во образования и науки Рос. Федерации, ФГБОУ ВПО "С.-Петерб. гос. ун-т". - Санкт-Петербург : Филол. фак. СПбГУ, 2013. ISBN 978-5-8465-1343-3.
2. Вычислительные технологии в лингвистике и филологии : учеб.-метод. пособие / [В. П. Захаров и др.] ; под общ. ред. В. П. Захарова ; С.-Петерб. гос. ун-т, Филол. фак. - , 2012-2013.

Перечень ресурсов информационно-телекоммуникационной сети «Интернет»:

<http://nltk.org>

https://ru.wikipedia.org/wiki/Корпусная_лингвистика

Материально-техническое обеспечение занятия: компьютерный класс.

Методические указания для обучающихся по освоению дисциплины

Методические рекомендации по организации самостоятельной работы.

Освоение дисциплины предполагает активную самостоятельную работу студента. Самостоятельная работа студента состоит из:

подготовки к лекциям и семинарам (чтения и усвоения соответствующей литературы);

выполнения домашних заданий;

выполнения домашних индивидуальных контрольных работ;

подготовки к контрольным работам и экзамену.

Самостоятельная работа студента является важным компонентом обучения.

Студент обязан приходить на лекции и семинары предварительно подготовившись уже по пройденным темам, которые используются в текущих лекциях и семинарах.

АННОТАЦИЯ ДИСЦИПЛИНЫ

Дисциплина «Вычислительная лингвистика» входит в состав части, формируемой участниками образовательных отношений, блока Б1 дисциплин по выбору для подготовки студентов по направлению подготовки 45.03.04 «Интеллектуальные системы в гуманитарной сфере». Дисциплина реализуется на Отделении интеллектуальные системы в гуманитарной сфере кафедрой математики, логики и интеллектуальных систем в гуманитарной сфере в 7-м семестре.

Цель дисциплины: познакомить студентов с возможностями вычислительной лингвистики в компьютерных системах.

Задача дисциплины: освоение базовых методов и средств вычислительной лингвистики и их использования в информационных и интеллектуальных системах.

В результате освоения дисциплины (*модуля*) обучающийся должен:

В результате изучения дисциплины студент должен:

Знать:

- основные виды моделей, используемых для формализации морфологии, синтаксиса и семантики естественных и искусственных языков;
- подходы к построению языковых процессоров.

Уметь:

- разрабатывать и отлаживать программы, связанные с обработкой текста.

Владеть:

- навыками применения основных видов языковых ресурсов при разработке приложений вычислительной лингвистики.

По дисциплине предусмотрены следующие виды контроля: текущий контроль успеваемости в форме устных ответов у доски, выполнения домашних заданий на компьютере; промежуточная аттестация в форме зачета.

Общая трудоемкость освоения дисциплины составляет 3 зачетные единицы.