МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное учреждение
высшего профессионального образования
"Казанский (Приволжский) федеральный университет"
Отделение татарской филологии и культуры имени Габдуллы Тукая

УТВЕРЖДАЮ
Проректор
по образовательной деятельности КФУ
Проф. Таюрский Д.А.

_____
"___"_____20___ г.

Программа дисциплины

Методы автоматического анализа языка Б3.ДВ.6

Направление подготовки: 032700.62 - Филология
Профиль подготовки: Прикладная филология (Татарский язык и литература, информационные технологии)
Квалификация выпускника: бакалавр
Форма обучения: очное
Язык обучения: русский

Автор(ы):

Гильмуллин Р.А.

Рецензент(ы):

СОГЛАСОВАНО:

Заведующий(ая) кафедрой: Салехова Л. Л.
Протокол заседания кафедры No ___ от "____" _____ 201__г
Учебно-методическая комиссия Института филологии и межкультурной коммуникации (отделение татарской филологии и культуры имени Габдуллы Тукая):
Протокол заседания УМК No ____ от "____" _____ 201__г

Регистрационный No

Казань
2016

Содержание

Программу дисциплины разработал(а)(и) Гильмуллин Р.А.

1. Цели освоения дисциплины

Основная цель курса - познакомить студентов с основными методами автоматической обработки естественного языка (NLP - Natural Language Processing) с использованием современных информационных технологий. Данная дисциплина призвана объяснить основные причины и условия применения методов автоматического анализа языка в научных лингвистических исследованиях, прикладной, в том числе переводческой деятельности лингвиста; познакомить студентов с существующими программными продуктами для осуществления профессиональной деятельности.

2. Место дисциплины в структуре основной образовательной программы высшего профессионального образования

Данная учебная дисциплина включена в раздел " Б3.ДВ.6 Профессиональный" основной образовательной программы 032700.62 Филология и относится к дисциплинам по выбору. Осваивается на 2, 3 курсах, 4, 5 семестры.

Данная учебная дисциплина включена в раздел " ? Общепрофессиональный" основной образовательной программы 032700.62 Филология и относится к базовой (общепрофессиональной) части. Осваивается на ? курсе, ? семестр.

Данная учебная дисциплина входит в базовую часть естественнонаучного цикла ФГОС ВПО по направлению подготовки 032700.62 Филология. Дисциплина логически связана с курсами "Компьютерные технологии в лингвистике" и "Использование современных информационных технологий в лингвистике".

3. Компетенции обучающегося, формируемые в результате освоения дисциплины /модуля

В результате освоения дисциплины формируются следующие компетенции:

Шифр компетенции	Расшифровка приобретаемой компетенции
ОК-10 (общекультурные компетенции)	способность понимать сущность и значение информации в развитии современного информационного общества, сознавать опасности и угрозы, возникающие в этом процессе, соблюдать основные требования информационной безопасности, в том числе защиты государственной тайны
ОК-11 (общекультурные компетенции)	владение основными методами, способами и средствами получения, хранения, переработки информации, навыки работы с компьютером как средством управления информацией
ОК-12 (общекультурные компетенции)	способность работать с информацией в глобальных компьютерных сетях
ПК-13 (профессиональные компетенции)	владение базовыми навыками доработки и обработки (корректура, редактирование, комментирование, реферирование и т. п.) различных типов текстов
ПК-2 (профессиональные компетенции)	владение базовыми навыками сбора и анализа языковых и литературных фактов с использованием традиционных методов и современных информационных технологий

В результате освоения дисциплины студент:

1. должен знать:

теоретические основы применения методов автоматического анализа языка в профессиональной, прикладной, научно-исследовательской и образовательной деятельности лингвиста;

- основные понятия и термины, относящиеся к сфере информатизации общества, науки и образования;

- основные математико-статистические методы обработки лингвистической информации;

- принципы работы специализированных программных продуктов, в том числе созданных для решения переводческих задач.

2. должен уметь:

- использовать компьютерные технические средства и стандартное программное обеспечение в профессиональных, исследовательских и образовательных целях;

- работать с основными типами профессиональных, прикладных, научно-исследовательских и учебных компьютерных программ.

3. должен владеть:

- использовать компьютерные технические средства и стандартное программное обеспечение в профессиональных, исследовательских и образовательных целях;

- работать с основными типами профессиональных, прикладных, научно-исследовательских и учебных компьютерных программ.

работать с компьютером как средством получения, обработки и управления информацией

4. Структура и содержание дисциплины/ модуля

Общая трудоемкость дисциплины составляет 2 зачетных(ые) единиц(ы) 72 часа(ов).

Форма промежуточного контроля дисциплины отсутствует в 4 семестре; зачет в 5 семестре.

Суммарно по дисциплине можно получить 100 баллов, из них текущая работа оценивается в 50 баллов, итоговая форма контроля - в 50 баллов. Минимальное количество для допуска к зачету 28 баллов.

86 баллов и более - "отлично" (отл.);

71-85 баллов - "хорошо" (хор.);

55-70 баллов - "удовлетворительно" (удов.);

54 балла и менее - "неудовлетворительно" (неуд.).

4.1 Структура и содержание аудиторной работы по дисциплине/ модулю

Тематический план дисциплины/модуля

N	Раздел Дисциплины/ Модуля	Семестр	Неделя семестра	Виды и часы аудиторной работы, их трудоемкость (в часах)			Текущие формы контроля
				Лекции	Практические занятия	Лабораторные работы	
1.	Тема 1. Компьютерные технологии в филологии. Введение в методы автоматического анализа языка.						

Ресурсы автоматической обработки текстов естественного языка.

4	1-3	0	6	0		

N	Раздел Дисциплины/ Модуля	Семестр	Неделя семестра	Виды и часы аудиторной работы, их трудоемкость (в часах)			Текущие формы контроля
				Лекции	Практические занятия	Лабораторные работы	
2.	Тема 2. Компьютерная лексикография. Формализация структуры словаря. Работа с лексикографической базой данных.	4	4-6	0	6	0	
3.	Тема 3. Корпусная лингвистика. Работа с электронными корпусами языков.	4	7-9	0	6	0	
4.	Тема 4. Статистический анализ текста. Количественные методы в исследовании текстов. Построение частотного словаря.	5	1-3	0	6	0	
5.	Тема 5. Информационно-поисковые системы. Работа с ИПС. Общие принципы индексации и ранжирования документов. Анализ релевантности результатов запроса.	5	4-6	0	6	0	
6.	Тема 6. Системы машинного перевода. Методы создания машинного перевода. Проблема многозначности.	5	7-9	0	6	0	
.	Тема . Итоговая форма контроля	5		0	0	0	зачет
	Итого			0	36	0	

4.2 Содержание дисциплины

Тема 1. Компьютерные технологии в филологии. Введение в методы автоматического анализа языка. Ресурсы автоматической обработки текстов естественного языка.

практическое занятие (6 часа(ов)):

Введение. Филологические направления, в которых активно задействуются компьютерные технологии. Экскурс в проблемы автоматической обработки текста, необходимой для работы программ, анализирующих и преобразующих текстовые данные.

Тема 2. Компьютерная лексикография. Формализация структуры словаря. Работа с лексикографической базой данных.

практическое занятие (6 часа(ов)):

Общие сведения. Формализация структуры словаря. Устройство базы данных словаря. Типы информации в словаре и базе данных (БД). Объекты БД: таблицы и формы, фильтры, запросы, отчеты, макропрограммы. Пользовательская работа с объектами базы в лексикографической практике.

Тема 3. Корпусная лингвистика. Работа с электронными корпусами языков.

практическое занятие (6 часа(ов)):

Корпусная лингвистика (КЛ). Общие соображения. Понятия КЛ. Требования к корпусу. Специфика разметки языковых данных. Проблемы снятия неоднозначностей в корпусах текстов. Достижения КЛ. Современные проекты. Корпуса текстов on-line. Проблемы современной корпусной лингвистики.

Тема 4. Статистический анализ текста. Количественные методы в исследовании текстов. Построение частотного словаря.

практическое занятие (6 часа(ов)):

Лингвистические принципы автоматического выделения информации из текста. Выделение терминов из корпуса текстов: графический уровень, словообразовательный уровень, лексический уровень, синтаксический уровень, текстовый уровень. Проблемы автоматического реферирования документов. Формализация филологических моделей художественного текста. Лексическая статистика и идиостиль автора. Количественные методы в применении к структуре сюжета. Статистические исследования стихотворного ритма.

Тема 5. Информационно-поисковые системы. Работа с ИПС. Общие принципы индексации и ранжирования документов. Анализ релевантности результатов запроса.

практическое занятие (6 часа(ов)):

Информационно-поисковые системы. Поиск информации как лингвистическая проблема. Современные ИПС (Google, Яндекс, Rambler и др.). Возможности расширенного поиска в ИПС. Синтаксис запросов. Общие принципы индексации и ранжирования документов.

Тема 6. Системы машинного перевода. Методы создания машинного перевода. Проблема многозначности.

практическое занятие (6 часа(ов)):

Проблемы машинного перевода. Перевод как прикладная лингвистическая дисциплина. Комбинирование различных методов уровневого лингвистического анализа при переводе. Методы автоматического разрешения многозначности при переводе.

4.3 Структура и содержание самостоятельной работы дисциплины (модуля)

N	Раздел Дисциплины	Семестр	Неделя семестра	Виды самостоятельной работы студентов	Трудоемкость (в часах)	Формы контроля самостоятельной работы
1.	Тема 1. Компьютерные технологии в филологии. Введение в методы автоматического анализа языка. Ресурсы автоматической обработки текстов естественного языка.	4	1-3	Анализ систем и технологий в области автоматического анализа языка.	6	Домашнее задание. Реферат.
2.	Тема 2. Компьютерная лексикография. Формализация структуры словаря. Работа с лексикографической базой данных.	4	4-6	Анализ структуры татарско-русского общелексического словаря. Выделение основных зон, оформление стат	6	Домашнее задание.
3.	Тема 3. Корпусная лингвистика. Работа с электронными корпусами языков.	4	7-9	Анализ значений определенной лексики с использованием национального корпуса русского языка.	6	Домашнее задание. Реферат.
4.	Тема 4. Статистический анализ текста. Количественные методы в исследовании текстов. Построение частотного словаря.	5	1-3	Построение частотного словаря для определенных произведений.	6	Домашнее задание.
5.	Тема 5. Информационно-поисковые системы. Работа с ИПС. Общие принципы индексации и ранжирования документов. Анализ релевантности результатов запроса.	5	4-6	Анализ поисковых систем. Выполнение запросов с использованием разных поисковых систем. Исследование	6	Домашнее задание.

N	Раздел Дисциплины	Семестр	Неделя семестра	Виды самостоятельной работы студентов	Трудоемкость (в часах)	Формы контроля самостоятельной работы
6.	Тема 6. Системы машинного перевода. Методы создания машинного перевода. Проблема многозначности.	5	7-9	Исследование результатов перевода текстов с использованием различных систем машинного перевода.	6	Реферат.
	Итого				36	

5. Образовательные технологии, включая интерактивные формы обучения

Освоение курса "Методы автоматического анализа языка" предполагает использование как традиционных, так и инновационных образовательных технологий. Традиционные образовательные технологии подразумевают использование в учебном процессе таких методов работ, как лекция, семинар, практическое занятие и др. Инновационные образовательные технологии обуславливают внедрение в учебный процесс таких методов и приемов, как различные формы тренингов, деловые игры, дискуссия, моделирование ситуаций и др.

6. Оценочные средства для текущего контроля успеваемости, промежуточной аттестации по итогам освоения дисциплины и учебно-методическое обеспечение самостоятельной работы студентов

Тема 1. Компьютерные технологии в филологии. Введение в методы автоматического анализа языка. Ресурсы автоматической обработки текстов естественного языка.

Домашнее задание. Реферат. , примерные вопросы:

Анализ систем и технологий в области автоматического анализа языка. Назначение систем. Указание источника ресурсов.

Тема 2. Компьютерная лексикография. Формализация структуры словаря. Работа с лексикографической базой данных.

Домашнее задание. , примерные вопросы:

Анализ структуры татарско-русского общелексического словаря. Выделение основных зон, оформление статьи с использованием языка разметки. Результат - формальное представление статьи словаря.

Тема 3. Корпусная лингвистика. Работа с электронными корпусами языков.

Домашнее задание. Реферат. , примерные вопросы:

Анализ значений определенной лексики с использованием национального корпуса русского языка. Исследование всех значений лексики в определенном периоде времени. Построение графика.

Тема 4. Статистический анализ текста. Количественные методы в исследовании текстов. Построение частотного словаря.

Домашнее задание. , примерные вопросы:

Построение частотного словаря для определенного произведений. С помощью специальных функций Microsoft Excel построить частотный словарь для произвольного литературного произведения.

Тема 5. Информационно-поисковые системы. Работа с ИПС. Общие принципы индексации и ранжирования документов. Анализ релевантности результатов запроса.

Домашнее задание. , примерные вопросы:

Анализ поисковых систем. Выполнение запросов с использованием разных поисковых систем. Исследование результатов обработки.

Тема 6. Системы машинного перевода. Методы создания машинного перевода. Проблема многозначности.

Реферат. , примерные вопросы:

Исследование результатов перевода текстов с использованием различных систем машинного перевода.

Тема . Итоговая форма контроля

Примерные вопросы к зачету:

Примерные вопросы к зачету /экзамену:

1) Лингвистические компьютерные технологии.

2) История компьютерной лингвистики.

3) Компьютерные методы лингвистических исследований.

4) Автоматический анализ текста.

5) Лингвистические модели.

6) Формализация языковой структуры.

7) Компьютерная лексикография, электронные словари.

8) Квантитативная лингвистика. Частотные словари.

9) Корпусная лингвистика.

10) Использование корпусов текстов в научных исследованиях.

11) Лингвистические ресурсы и поиск в Интернет.

12) Лингвистические технологии информационного поиска.

13) Системы машинного перевода.

7.1. Основная литература:

Марчук Ю.Н. Компьютерная лингвистика. Учебное пособие. - М.: Восток-Запад, 2007. - 317 с. (1 экз.).

Захаров В.П., Богданова С.Ю. Корпусная лингвистика. - Иркутск: Издательство ИГЛУ, 2011. - 161 с. (1 экз.).

Зубов А.В. Информационные технологии в лингвистике: учебное пособие для студентов вузов. - М.: Academia, 2004. - 205 с. (26 экз.).

Хроленко А.Т., Денисов А.В. Современные информационные технологии для гуманитария: практическое руководство. - М.: Флинта: Наука, 2008. - 128 с. (1 экз.).

7.2. Дополнительная литература:

Сулейманов Д.Ш., Хадиев Р.М., Якушев Р.С. Компьютерные информационные технологии. - Казань: КГУ, 2004. - 191 с. (12 экз.).

Термины информатики и информационных технологий: Англо-татарско-русский толковый словарь. - Казань: Магариф, 2006. - 383 с. (2 экз.).

Гладкий А.В., Мельчук И.А. Элементы математической лингвистики. - М.: Наука, 1969. - 192 с. (3 экз.).

Кибрик А.Е. Очерки по общим и прикладным вопросам языкознания: (универсальное, типовое и специфическое в языке). - М.: Изд-во МГУ, 1992. - 335 с. (2 экз.).

7.3. Интернет-ресурсы:

Британский национальный корпус - http://www.natcorp.ox.ac.uk

Компания 'Аби' - http://abbyy.ru

Лаборатории общей и компьютерной лексикографии МГУ - http:// lexigraph.nm.ru/library.htm

Национальный корпус русского языка - http://www.ruscorpora.ru

НИИ 'Прикладная семиотика' АН РТ - http://ips.antat.ru

Филологический факультет МГУ - http://www.philol.msu.ru/~lex/main.htm

8. Материально-техническое обеспечение дисциплины(модуля)

Освоение дисциплины "Методы автоматического анализа языка" предполагает использование следующего материально-технического обеспечения:

Компьютерный класс, представляющий собой рабочее место преподавателя и не менее 15 рабочих мест студентов, включающих компьютерный стол, стул, персональный компьютер, лицензионное программное обеспечение. Каждый компьютер имеет широкополосный доступ в сеть Интернет. Все компьютеры подключены к корпоративной компьютерной сети КФУ и находятся в едином домене.

Освоение дисциплины "Методы автоматического анализа языка" предполагает использование следующего материально-технического обеспечения:

Мультимедийная аудитория, вместимостью более 60 человек. Мультимедийная аудитория состоит из интегрированных инженерных систем с единой системой управления, оснащенная современными средствами воспроизведения и визуализации любой видео и аудио информации, получения и передачи электронных документов. Типовая комплектация мультимедийной аудитории состоит из: мультимедийного проектора, автоматизированного проекционного экрана, акустической системы, а также интерактивной трибуны преподавателя, включающей тач-скрин монитор с диагональю не менее 22 дюймов, персональный компьютер (с техническими характеристиками не ниже Intel Core i3-2100, DDR3 4096Mb, 500Gb), конференц-микрофон, беспроводной микрофон, блок управления оборудованием, интерфейсы подключения: USB,audio, HDMI. Интерактивная трибуна преподавателя является ключевым элементом управления, объединяющим все устройства в единую систему, и служит полноценным рабочим местом преподавателя. Преподаватель имеет возможность легко управлять всей системой, не отходя от трибуны, что позволяет проводить лекции, практические занятия, презентации, вебинары, конференции и другие виды аудиторной нагрузки обучающихся в удобной и доступной для них форме с применением современных интерактивных средств обучения, в том числе с использованием в процессе обучения всех корпоративных ресурсов. Мультимедийная аудитория также оснащена широкополосным доступом в сеть интернет. Компьютерное оборудованием имеет соответствующее лицензионное программное обеспечение.

Компьютерный класс, представляющий собой рабочее место преподавателя и не менее 15 рабочих мест студентов, включающих компьютерный стол, стул, персональный компьютер, лицензионное программное обеспечение. Каждый компьютер имеет широкополосный доступ в сеть Интернет. Все компьютеры подключены к корпоративной компьютерной сети КФУ и находятся в едином домене.

Учебно-методическая литература для данной дисциплины имеется в наличии в электронно-библиотечной системе "КнигаФонд", доступ к которой предоставлен студентам. Электронно-библиотечная система "КнигаФонд" реализует легальное хранение, распространение и защиту цифрового контента учебно-методической литературы для вузов с условием обязательного соблюдения авторских и смежных прав. КнигаФонд обеспечивает широкий законный доступ к необходимым для образовательного процесса изданиям с использованием инновационных технологий и соответствует всем требованиям новых ФГОС ВПО.

Для изучения данной дисциплины необходима компьютерный класс, оборудованный мультимедийными компьютерами с доступом в Интернет, проектор, экран, интерактивная доска, принтер, сканер, копир.

Программа составлена в соответствии с требованиями ФГОС ВПО и учебным планом по направлению 035700.62 "Лингвистика" и профилю подготовки Перевод и переводоведение (английский и второй иностранные языки) .

Программа составлена в соответствии с требованиями ФГОС ВПО и учебным планом по направлению 032700.62 "Филология" и профилю подготовки Прикладная филология (Татарский язык и литература, информационные технологии) .

Автор(ы):
Гильмуллин Р.А. _____
"__" _____ 201 __ г.

Рецензент(ы):
"__" _____ 201 __ г.