



Правительство Российской Федерации

**Нижегородский филиал
Федерального государственного автономного образовательного учреждения высшего профессионального образования
"Национальный исследовательский университет
"Высшая школа экономики"**

Факультет Бизнес информатики и прикладной математики

Программа научно-исследовательского семинара «Методология и методы исследований в компьютерной лингвистике»

для направления 010400.68 «Прикладная математика и информатика» подготовки
магистра
для магистерских программ "Прикладная математика и информатика", «Компьютерная лингвистика»

Автор программы: Карпов Н.В., к.т.н., nkarpov@hse.ru

Одобрена на заседании кафедры
прикладной математики и информатики «__»_____ 2013 г
Зав. кафедрой Калягин В.А. _____

Рекомендована секцией УМС Секция «Прикладная математика» «__»_____ 2013 г
Председатель Калягин В.А. _____

Утверждена УМС НИУ ВШЭ – Нижний Новгород «__»_____ 2013 г.
Председатель Петрухин Н.С. _____



Область применения и нормативные ссылки

Настоящая программа НИС устанавливает минимальные требования к знаниям и умениям студента и определяет содержание и виды учебных занятий и отчетности.

Программа предназначена для преподавателей, ведущих данный семинар, учебных ассистентов и студентов направления подготовки 010400.68 «Прикладная математика и информатика», проходящих научно-исследовательский семинар «Методология и методы исследований в компьютерной лингвистике».

Программа разработана в соответствии с:

- ОрОС НИУ ВШЭ по направлению «Прикладная математика и информатика», степень — магистр прикладной математики и информатики.
- Образовательной программой 010400.68 «Прикладная математика и информатика».
- Рабочими учебными планами университета по направлению подготовки 010400.68 "Прикладная математика и информатика", магистерские программы «Прикладная математика и информатика», «Компьютерная лингвистика», утвержденными в 2012г.

1. Цели освоения НИС

Целями освоения научно-исследовательский семинара «Методология и методы исследований в компьютерной лингвистике» являются освоение магистрами современных математических и алгоритмических методов обработки естественного языка.

2. Компетенции обучающегося, формируемые в результате освоения НИС

В результате освоения НИС студент должен:

- Знать основные понятия обработки сигналов и понимать теоретическую базу компьютерного анализа речи.
- Уметь применять стандартные методы и модели к решению задач распознавания, синтеза и обработки речевых сигналов и получать обоснованные выводы.
- Иметь навыки (приобрести опыт) работы с основными принципами и методами обработки речевых данных, овладеть навыками применения речевых технологий и программ для анализа звука на ПЭВМ.

В результате освоения НИС студент осваивает следующие компетенции:

Компетенции	Код по НИУ	Дескрипторы – основные признаки освоения (показатели достижения результата)	Формы и методы обучения, способствующие формированию и развитию компетенции
Способен создавать новые теории, изобретать новые способы и инструменты профессиональной деятельности.	СК-М2	Применяет полученные знания для совершенствования и развития своего интеллектуального и общекультурного уровня с целью нравственного и физического совершенствования своей личности	Самостоятельная работа
Способен использовать в профессиональной деятельности знания в области естественных наук, математики и информатики, понимание основных фактов, концепций, принципов теорий, связанных с прикладной математикой и	ИК-М7.1пми	Применяет углубленные знания в области естественных наук, математики и информатики при оценке последствий своей профессиональной деятельности.	Чтение лекций, проведение практических занятий, самостоятельная работа.



информатикой.			
Способен создавать междисциплинарные тексты с использованием языка и аппарата прикладной математики	ИК-М2.2пми	Применяет полученные знания для анализа естественного мира и формулирует результаты в виде отчета	Проведение практических занятий, самостоятельная работа
Способен строить профессиональную деятельность, бизнес и делать выбор, руководствуясь принципами социальной ответственности.	СЛК-М7	представляет связи между областями прикладной математики и информационных технологий по профильной направленности ООП магистратуры	Чтение лекций, проведение практических занятий, самостоятельная работа
Способен организовать многостороннюю коммуникацию и управлять ею	СК-М7	применяет полученные знания в самостоятельных проектах по тематике специализации	Чтение лекций, проведение практических занятий, самостоятельная работа
Способен использовать социальные и мультикультурные различия для решения проблем в профессиональной и социальной деятельности.	СЛК-М3	представляет связи между профессиональными сетевыми сообществами по конкретным направлениям	Чтение лекций, проведение практических занятий, самостоятельная работа

3. Место НИС в структуре образовательной программы

Настоящий научно исследовательский семинар относится к циклу дисциплин программы и блоку дисциплин, М.3 – Практики и научно-исследовательская работа.

Освоение материала базируется на следующих дисциплинах:

- Современные методы анализа данных

Для освоения НИС, студенты должны владеть следующими знаниями и компетенциями:

- Знание основных принципов статистического анализа языка.
- Ориентация в современных технологиях программирования.

Основные положения НИС должны быть использованы в дальнейшем в научных исследованиях естественного языка и при самостоятельной применении технологий обработки речи.

4. Тематический план НИС

№	Название раздела	Всего часов	Аудиторные часы			Самостоятельная работа
			Лекции	Семинары	Практические занятия	
1	Применение интеллектуального анализа данных в самостоятельном исследовании и проектировании собственных программных продуктов.	109		30		79
2	Использование методов интеллектуального анализа данных в перспективных направлениях развития информационных технологии	107		28		79
	Итого	216		58		158



5. Формы контроля знаний студентов

Тип контроля	Форма контроля	1 год			Параметры
		2	3		
Текущий (неделя)	Контрольная работа	8			Выполнение четырех домашних лабораторных работ
Итоговый	Зачет			*	Устный зачет, 120 мин.

Критерии оценки знаний, навыков

При написании контрольной работы студент должен продемонстрировать умение самостоятельно проводить поиск информации в открытых источниках и делать критический анализ работ по выбранному направлению.

На зачете студент должен продемонстрировать владение терминологией, умение правильно ставить задачи и формулировать цели исследования, решать задачи, способность отвечать на поставленные вопросы.

Оценки по всем формам текущего контроля выставляются по 10-ти балльной шкале.

6. Содержание НИС

1. Ручное аннотирование текста
2. Сравнение результатов аннотирования текста
3. Использование цепочки обрабатывающих ресурсов
4. Создание и применение словаря
5. Применение JAPE для автоматического аннотирования
6. Использование Searchable DataStore
7. Аннотирование текста с использованием JAVA
8. Извлечение информации из коллекции гомогенных текстов.

7. Образовательные технологии

При реализации учебной работы предполагается использовать разбор практических задач.

8. Оценочные средства для текущего контроля и аттестации студента

8.1. Тематика заданий текущего контроля

Примеры заданий для контрольной работы

Задание 1

Реализовать одно задание из лабораторной работы № 5 ([Применение JAPE для автоматического аннотирования](#)) используя Java API в правой части выражения JAPE. Для тестирования использовать публицистический корпус, сформированный в предыдущих работах.

Задание 2

Разработать независимое Java-приложение, использующее традиционные ANNIE ресурсы из библиотеки Gate для выделения в английском тексте следующих элементов: Person - Location - Date - QuotedText - Money - Address.

Задание 3

Разработать плагин для Gate Developer, который производит автоматическое аннотирование дат в русскоязычном формате.



8.2. Вопросы для оценки качества освоения НИС

1. Аннотация, Разметка, Корпус, Документ, Приложение
2. Annotation Diff
3. Метрики сравнения аннотаций
4. Recall, Precision
5. Error rate
6. F-Мера
7. Corpus Quality Assurance
8. Corpus Benchmark
9. Использование Java в правой части выражений JAPE
10. Разработка Java-приложения на основе технологии Gate Embedded
11. Разработка обрабатывающего ресурса CREOLE (a Collection of REusable Objects for Language Engineering).

9. Порядок формирования оценок по НИС

Контроль знаний студентов включает формы текущего и итогового контроля. Текущий контроль осуществляется в течение двух модулей. По курсу предусмотрены: текущий контроль знаний студентов, контроль индивидуальной и практической работы студентов. Каждая форма текущего контроля оценивается 10-балльной оценкой, которая выставляется в рабочую ведомость преподавателя. По результатам текущего контроля организуются индивидуальные консультации либо в рамках второй половины рабочего дня преподавателя, либо по электронной почте. Предоставленные студентами результаты индивидуальных и лабораторных работ обсуждаются на семинарских занятиях и оцениваются лектором и преподавателем, ведущим практические занятия.

Формы итогового контроля:

По окончании второго модуля курса предусмотрена письменная контрольная работа в виде отчета по лабораторной работе, которая оценивается по 10-балльной шкале.

По окончании всего учебного курса (окончание курса) предусмотрен письменный дифференцированный зачет по темам 6-10, который оценивается по 10-балльной шкале. Продолжительность зачета — 120 мин.

В течение одного учебного модуля курса проводится публичная защита и обсуждение результатов индивидуальных работ и домашних практических работ, которые оцениваются по 10-балльной шкале. Защита включает в себя презентацию материала, ответы на вопросы аудитории и преподавателя.

Для получения накопленной оценки $O_{\text{накопл.}}$ используются следующие весовые множители:

0,6 – для оценки $O_{\text{контр.раб}}$ за выполнение и презентацию индивидуальных и лабораторных работ студентов,

0,4 – для оценки $O_{\text{аудит}}$, за работу на занятиях.

$$O_{\text{накопл.}} = 0,4 * O_{\text{аудит.}} + 0,6 * O_{\text{контр.раб.}}$$

Результирующая оценка за итоговый контроль в форме зачета выставляется по следующей формуле, где $O_{\text{зачет}}$ - оценка за работу непосредственно на зачете:

$$O_{\text{итоговый}} = 0,3 * O_{\text{зачет}} + 0,7 * O_{\text{накопл.}}$$

Способ округления накопленной оценки и оценки за зачет – арифметический.

Таблица соответствия оценок по десятибалльной системе и результата зачета.

По десятибалльной шкале	Результат
1 – неудовлетворительно	



2 – очень плохо 3 – плохо	«не зачтено»
4 – удовлетворительно 5 – весьма удовлетворительно	«зачтено»
6 – хорошо 7 – очень хорошо	«зачтено»
8 – почти отлично 9 – отлично 10 – блестяще	«зачтено»

10. Учебно-методическое и информационное обеспечение НИС

10.1. Базовый учебник

1. Большакова Е.И. и др. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учеб. пособие — М.: МИЭМ, 2011. — 272 с.

10.2. Основная литература

2. Jurafsky, Daniel, and James H. Martin. Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. 2nd edition. Prentice-Hall. 2009

10.3. Дополнительная литература

3. П. Хоровиц, У. Хилл. Искусство схемотехники: В 2-х т. Пер. с англ. — М: Мир, 1984.
4. Л.В. Бондарко. Звуковой строй современного русского языка. М.: Просвещение, 1997.
5. Паклин Н.В., Орешков В. Бизнес-аналитика. От данных к знаниям (+ CD-ROM) – СПб: Питер. 2010
6. Кацко И. А., Паклин Н. Б. Практикум по анализу данных на компьютере. - М: КолосС, 2009

10.4. Программные средства

Для успешного освоения дисциплины, студент использует следующие программные средства:

- Matlab
- Microsoft Excel

11. Материально-техническое обеспечение НИС

Для выполнения и демонстрации лабораторных работ предполагается использовать ресурсы вычислительного кластера НИУ ВШЭ – Нижний Новгород.

Автор программы

Н.В. Карпов