Правительство Российской Федерации

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

Факультет гуманитарных наук

Школа лингвистики

Рабочая программа дисциплины
Программирование и лингвистические данные

для образовательной программы «Фундаментальная и компьютерная лингвистика»
направления 45.03.03 «Фундаментальная и прикладная лингвистика»
подготовки бакалавра

Разработчики программы:
Б. В. Орехов, канд. филол. наук, borekhov@hse.ru
О. Н. Ляшевская, канд. филол. наук, olesar@gmail.com
К. Л. Маланчев, канд. наук, kmalanchev@hse.ru

Одобрена на заседании школы лингвистики «05» июня 2018 г.
Руководитель школы Е.В. Рахилина _____

Рекомендована Академическим советом образовательной программы
«05» июня 2018 г., Протокол № 15

Утверждена «05» июня 2018 г.
Академический руководитель образовательной программы
Ю.А. Ландер _____]

Москва, 2018

1. Область применения и нормативные ссылки

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

Программа предназначена для преподавателей, ведущих данную дисциплину, учебных ассистентов и бакалавров направления подготовки 45.03.03 «Фундаментальная и прикладная лингвистика» факультета гуманитарных наук.

Программа разработана в соответствии с:

1. Образовательным стандартом бакалавриата НИУ ВШЭ по направлению 45.03.03 «Фундаментальная и прикладная лингвистика».
2. Базовым учебным планом по направлению 45.03.03 «Фундаментальная и прикладная лингвистика» подготовки бакалавра
3. Рабочим учебным планом НИУ ВШЭ по направлению 45.03.03 «Фундаментальная и прикладная лингвистика», утвержденным в 2018 году.

2. Цели освоения дисциплины

Целями освоения дисциплины «Программирование и компьютерные инструменты лингвистических исследований» являются знакомство с основными компьютерными инструментами и ресурсами, применяемыми в лингвистических исследованиях. Курс закладывает теоретическую и практическую базу для использования различного инструментария для компьютеризации сбора, обработки и анализа лингвистического материала, а также для представления данных и результатов исследований в виде компьютерных ресурсов. Часть курса посвящена изучению программирования на языке Python, алгоритмов, регулярных выражений.

3. Компетенции обучающегося, формируемые в результате освоения дисциплины

В результате освоения дисциплины студент должен:

Знать:

- основные типы компьютерных лингвистических ресурсов, используемых для сбора материала исследований;
- базовые принципы работы с лингвистическими корпусами и ресурсами;
- основные типы запросов к корпусам для поиска материала в соответствии с различными типами задач лингвистических исследований;
- основные методы работы с материалом с использованием различных опций MicrosoftExcel.
- методы обработки материала с помощью специальных компьютерных инструментов, таких как конкордансеры;
- способы хранения информации на электронных носителях;
- методы автоматической обработки информации с помощью языка программирования Python;
- читать и записывать файлы средствами языка Python;
- форматы HTML, XML, используемые для хранения текстовых данных.

Уметь:

- работать с простыми средствами обработки текстов: многофункциональными текстовыми редакторами типа Notepad++и редакторами электронных таблиц, таких как Excel;
- осуществлять оценку различных типов современных корпусных ресурсов и выбирать ресурсы, подходящие для выполнения тех или иных исследовательских и производственных задач;

- осуществлять поиск в корпусах, в том числе и с использованием специальных языков запросов, в соответствии с исследовательской гипотезой в области грамматики и лексикографических исследований;
- работать с различными типами программ обработки текстов: конкордансерами, программами для поиска коллокаций, создания частотных списков и т.п., корпусными менеджерами, программами для документации языков, включающих поморфемную аннотацию текстов и составление словарей;
- публиковать свои данные на веб-сайте;
- работать с различными кодировками текстовых файлов, конвертировать тексты и пользоваться программами сравнения текстов для ручной обработки текстовых данных;
- строить алгоритмы для решения практических задач;
- использовать средства языка Python для реализации алгоритмов;
- пользоваться англоязычной документацией языка Python.

Иметь навыки (приобрести опыт):

- работы с материалом, собранным с использованием корпусных ресурсов;
- работы с программами первичной обработки текста;
- работы с текстовыми редакторами и электронными таблицами;
- сбора материала с использованием корпусов;
- построения алгоритмов для решения практических задач;
- реализации алгоритмов средствами языка Python;
- форматирование строк средствами языка Python;
- использования языка регулярных выражений.

В результате освоения дисциплины студент осваивает следующие компетенции:

Компетенция	Код по ФГОС/ НИУ
Способен учиться, приобретать новые знания, умения, в том числе в области, отличной от профессиональной	УК-1
Способен выявлять научную сущность проблем в профессиональной области.	УК-2
Способен решать проблемы в профессиональной деятельности на основе анализа и синтеза	УК-3
Способен вести исследовательскую деятельность, включая анализ проблем, постановку целей и задач, выделение объекта и предмета исследования, выбор способа и методов исследования, а также оценку его качества	УК-6
способен использовать основные понятия и категории современной лингвистики в своей профессиональной деятельности	ПК-1
способен проводить формализацию лингвистических знаний, анализ и синтез лингвистических структур, квантитативный анализ лингвистических данных с использованием математических знаний и методов	ПК-2
способен дать описание и провести формальную репрезентацию денотативной, концептуальной, коммуникативной и прагматической информации, содержащейся в тексте на естественном языке	ПК-8
способен участвовать в создании представительных текстовых массивов, корпусов текстов, корпусов звучащей речи, мультимодальных корпусов, лингвистических и социолингвистических баз данных и пользоваться этими ресурсами	ПК-11
способен проектировать системы анализа и синтеза естественного языка,	ПК-12

Компетенция	Код по ФГОС/ НИУ
анализа и синтеза мультимодальных языковых систем, в том числе лингвистических компонентов интеллектуальных и информационных электронных систем	
способен провести квалифицированное тестирование эффективности лингвистически ориентированного программного продукта	ПК-13
способен гибко адаптироваться к различным профессиональным ситуациям, проявлять творческий подход, инициативу и настойчивость в достижении целей профессиональной деятельности и личных	ПК-23

4. Место дисциплины в структуре образовательной программы

Настоящая дисциплина относится к циклу профессиональных дисциплин, обязательных для изучения.

Для освоения учебной дисциплины, студенты должны владеть следующими знаниями и компетенциями:

- владеть базовыми представлениями о грамматических категориях и анализе языковых единиц;
- владеть базовыми навыками работы с компьютером.

Основные положения дисциплины должны быть использованы в дальнейшем при изучении следующих дисциплин: Программирование и теория алгоритмов, Компьютерная лингвистика и информационные технологии, Базы данных, Автоматическая обработка естественного языка (курсы 3 и 4), Информационный поиск и извлечение данных, Компьютерная лингвистика, Онтологии и семантические технологии, Теория языка, научно-исследовательские семинары по различным лингвистическим задачам.

5. Тематический план учебной дисциплины

№	Название раздела	Всего часов	Аудиторные часы			Самостоятельная работа
			Лекции	Семинары	Практические занятия	
1	Компьютерные инструменты лингвистического исследования и электронные лингвистические ресурсы	56	8		20	28
2	Программирование	96		46		50
	Итого	152	8	46	20	78

6. Формы контроля знаний студентов

Тип контроля	Форма контроля	1 курс				Параметры
		1	2	3	4	
Текущий	Контрольная работа		*	*		Письменная работа, 75 минут
Итоговый	Экзамен	*			*	1 модуль: экзамен в виде выполнения итогового проектного задания (письменная

						и устная часть) 4 модуль: письменная работа, 120 минут

a. Критерии оценки знаний, навыков

- Оценки по всем формам текущего контроля выставляются по десятибалльной шкале.
- Выполненные домашние задания блоку «Программирование» студенты загружают в свои репозитории на веб-сервисе https://github.com/.
- Экзаменационное задание за 1 модуль представляет собой групповой проект. Проекты выполняются в группах по 4 человека. Участники выступают с групповой презентацией проекта на экзамене. Обязательной частью экзамена являются а) веб-страница с письменным описанием проекта (около 4 страниц, выполняется заранее дома) и б) презентация в PowerPoint или аналогичных форматах. Возможные темы проектов: 1) Мини-исследование истории слова или грамматического явления в корпусе; 2)Мини-исследование корпуса живой русской речи; 3) Характеристика корпуса некоторого языка (языков); 4) Характеристика лингвистической базы данных; 5) Характеристика словарного ресурса
- На контрольной работе в 2 модуле проверяется знание синтаксиса, основных функций и операторов языка Python и умение их применять при решении простых задач.
- На контрольной работе в 3 модуле проверяется знание функций языка Python, предназначенных для обработки текста, владение языком регулярных выражений и умение их использовать в языке Python.
- На экзамене в 4 модуле проверяются все знания и умения, приобретённые во время изучения настоящей дисциплины.
- Все практические задания выполняются на компьютере.
- Основной частью задания контрольной работы и экзамена за 4 модуль является задача, состоящая из 2-3 частей разного уровня сложности. Для получения положительной оценки необходимо решить задачу, написав программу на языке Python. Во время контрольных мероприятий разрешается пользоваться любыми источниками информации (если явным образом не оговорено иное).
- При обнаружении плагиата в домашнем или контрольном задании это задание получает оценку 0 баллов.

7. Содержание дисциплины

	Название	Лекции	Практические занятия	Литература или сетевые ресурсы по разделу
Раздел Лингвистические ресурсы				
1.	Типы лингвистических ресурсов. Специальные базы данных, корпуса, лексикографические ресурсы	1		Плунгян В. А. Зачем нужен Национальный корпус русского языка? Неформальное введение // Национальный корпус русского языка: 2003—2005. М.: Индрик, 2005, 6—20 Савчук С. О. Метатекстовая разметка в Национальном корпусе русского языка: базовые принципы и основные функции // Национальный корпус русского языка: 2003-2005. Результаты и перспективы. — М., 2005, 62—88 http://www.ruscorpora.ru/corpora-parameter.html - о метаразметке
2.	Особенности поиска в Национальном корпусе русского языка		2	
3.	Работа с корпусами		1	

	английского языка (в первую очередь, Araneum). Коллокации. Mutual information.			О. Н. Ляшевская, С. А. Шаров, Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник, 2009. Ресурсы и корпуса: http://ruscorpora.ru/ http://corpus.byu.edu/coca/ http://wordnetweb.princeton.edu/perl/webwn http://visuwords.com/ http://thesaurus.com/ http://dict.ruslang.ru/ http://starling.rinet.ru/babel.php?lan=ru http://www.ethnologue.com/ http://ucts.uniba.sk/
4.	Excel: различные текстовые функции и их комбинации, полезные для работы со словарными списками, фильтры и сортировка, сводные таблицы		3	Ресурсы и программы с прилагающейся на сайтах документацией: http://office.microsoft.com/ru-ru/excel-help/
5.	Учебный корпус живой разговорной речи. Сбор данных. Метаразметка. Расшифровка. Техническая подготовка текстов. XML-формат. Морфологическая аннотация и снятие грамматической омонимии.		3	

Раздел Программирование

1. Основы программирования на языке Python.

- Принципы работы системы контроля версий. Работа с git и github. XML и HTML.
- Введение. Принцип работы программ. Типы языков программирования. Понятия программы, компилятора, интерпретатора. Процесс разработки программы, отладка. Возможные ошибки в программе (синтаксические ошибки, ошибки во время исполнения). Использование интерпретатора языка Python: интерактивный режим, интерпретация программного кода в файле. Комментарии в программном коде.
- Основные понятия и синтаксические конструкции. Переменная, значение, присваивание, оценка выражения. Операторы. Арифметические операторы, логические операторы, операции над строками, порядок выполнения операций. Встроенные математические функции. Типы данных.
- Операторы для реализации нелинейных алгоритмов. Условный оператор, операторы цикла while и for. Прерывание цикла. Оформление блоков кода в Python и других языках.
- Функции, аргументы, возвращаемые значения. Рекурсия. Модули и библиотеки в Python и других языках.

- Функции для работы со строками и кодировками, поиск подстроки в строке. Чтение и запись файлов. Токенизация и первичная очистка текста.
- Структуры данных. Списки, индексы, функции для работы со списками, срезы. Строка как индексируемый объект. Словари (ассоциативные массивы), способы их задания в Python и функции для работы с ними.
- Функции для работы с файловой системой: обход дерева директорий с помощью os.walk, копирование, перемещение, удаление файлов, создание директорий и операции над ними.
- Обработка исключений.
- Способы отладки программы.

2. Регулярные выражения и их использование для поиска и обработки информации.

- Понятие регулярного выражения, примеры регулярных выражений с объяснением принципов их действия. Функции языка Python для работы с регулярными выражениями. Использование регулярных выражений для поиска и замены в Python и редакторе Notepad++.
- Основные элементы регулярных выражений: диапазоны символов, итерация (оператор
- «звёздочка»), количество повторений, дизъюнкция, escape-последовательности для обозначения
- системных символов и групп символов (\w, \d и т. п.), символы начала и конца строки. «Жадные» и «нежадные» квантификаторы.
- Группы и их использование для анализа данных. Нумерация групп в выражении. Токенизация строк с помощью регулярных выражений и функций Python. Прогрессивные и регрессивные незахватывающие группы. Обратные ссылки (backreferences). Использование групп для замены.
- Использование регулярных выражений для извлечения информации из текстов на естественном языке, структурированных данных в форматах CSV, HTML, XML. Примеры задач на сбор и обработку информации из XML-файлов Национального корпуса русского языка.

8. Образовательные технологии

Для изучения дисциплины необходим компьютер и следующее программное обеспечение: редактор электронных таблиц MS Excel или OpenOfficeCalc; текстовый редактор Notepad++ или любой другой, поддерживающий подсветку синтаксиса, переключение между разными кодировками и поиск с использованием регулярных выражений; интерпретатор языка Python.

Рекомендуемые образовательные технологии включают лекции, практические занятия, самостоятельную работу студентов (выполнение практических домашних заданий с использованием специализированного компьютерного инструментария).

При проведении занятий рекомендуется использование интерактивных форм занятий (проектных методик, разбор конкретных ситуаций, включение в лекционный курс интерактивного общения с аудиторией, презентаций, контрольных вопросов на понимание) в сочетании с внеаудиторной работой. Удельный вес занятий, проводимых в интерактивных формах, должен составлять не менее 40% аудиторных занятий.

9. Оценочные средства для текущего контроля и аттестации студента

a. Вопросы для оценки качества освоения дисциплины

- Какие типы лингвистических данных вам известны?
- Какие требования предъявляются к составлению корпусов?
- Каковы методы оценки частотности слова в корпусе?
- Какие типы корпусов Вы знаете?
- Дать определение одному из встретившихся в курсе понятий (программа, алгоритм и т. п.).

- Дано текстовое описание алгоритма и его блок-схема или реализация на языке Python с ошибкой. Найти и исправить ошибку.
- Реализовать на языке Python алгоритм средней сложности (предполагаемая длина менее 100 строк кода) по текстовому описанию.
- Написать регулярное выражение для поиска или замены определённой информации в тексте.
- Использовать регулярные выражения и язык Python для обработки текста (например, разбить текст на предложения; посчитать количество слов, начинающихся с гласной, в XML-файле и т. п.).

10.Порядок формирования оценок по дисциплине

Преподаватель или учебный ассистент каждую неделю оценивает самостоятельную работу студентов, проверяя домашние работы. Оценки за самостоятельную работу студента выставляются в рабочую ведомость. Накопленная оценка по 10-балльной шкале за самостоятельную работу определяется перед промежуточным или итоговым контролем – $О_{сам.р}$. Оценка за курс складывается из двух независимых частей. Оценка за 1 модуль считается по формуле:

$$О_{итоговый} = 0,3 \cdot О_{экзамен} + 0,7 \cdot О_{сам.р.}$$

Накопленная оценка за самостоятельную работу считается следующим образом:

$$О_{сам.\ р.} = 0,8 \cdot О_{д.з.} + 0,2 \cdot О_{семинары}.$$

Таким образом, оценка за первый модуль складывается из следующих компонентов:

- экзамен – 30%
- домашние задания – 56%
- ответы на семинарах – 14%.

Оценка со 2 по 4 модуль считается следующим образом. Оценка за контрольные работы ($О_{к/р}$) равна среднему арифметическому оценок за две контрольные работы:

$$О_{к/р} = 1/2 \cdot О_{к/р1} + 1/2 \cdot О_{к/р2}.$$

Результирующая оценка за итоговый контроль в форме экзамена выставляется по следующей формуле, где $О_{экзамен}$ — оценка за работу непосредственно на экзамене:

$$О_{итоговый} = 0,35 \cdot О_{экзамен} + 0,3 \cdot О_{к/р} + 0,35 \cdot О_{сам.р.}$$

Таким образом, в процентном отношении вклад имеющихся форм контроля выглядит так:

- экзамен — 35%
- текущий контроль — 30% (по 15% на каждую контрольную работу)
- самостоятельная работа — 35%

При подсчёте итоговой оценки промежуточные оценки (среднее арифметическое оценок за контрольные работы и среднее арифметическое оценок за домашние работы) не округляются.

11.Учебно-методическое и информационное обеспечение дисциплины

a. Основная литература

Плунгян В. А. Зачем нужен Национальный корпус русского языка? Неформальное введение // Национальный корпус русского языка: 2003—2005. М.: Индрик, 2005, 6—20

Савчук С. О. Метатекстовая разметка в Национальном корпусе русского языка: базовые принципы и основные функции // Национальный корпус русского языка: 2003-2005. Результаты и перспективы. — М., 2005, 62—88

b. Дополнительная литература

http://studiorum.ruscorpora.ru/index.php?option=com_docman&Itemid=111 — примеры корпусных исследований лексики в исторической перспективе
http://office.microsoft.com/ru-ru/excel-help/
Захаров В.П., Хохлова М.В. Анализ эффективности статистических методов выявления коллокаций в текстах на русском языке // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 26-30 мая 2010 г.). Вып. 9 (16). URL: http://www.dialog-21.ru/digests/dialog2010/materials/pdf/22.pdf
Christopher Manning and Hinrich Schütze. Foundations of Statistical Natural Language Processing. Chapter 5. Collocations. MIT Press. 1999. URL: http://nlp.stanford.edu/fsnlp/promo/colloc.pdf (URL: http://www.alingva.ru/articles/collocation_rus.pdf — русский перевод главы 5)
Марк Лутц. Изучаем Питон (4-е издание). Символ-плюс: М., 2011
Томас Кормен, Чарльз Лейзерсон, Рональд Ривест, Клиффорд Штайн. Алгоритмы: построение и анализ. Вильямс: М., 2011
Рейтц К., Шлюссер Т. Автостопом по Python. Питер: 2017

10.3 Интернет-ресурсы

Документация по языку Python: http://docs.python.org/
Steven Bird, Ewan Klein, Edward Loper. Natural Language Processing with Python: http://www.nltk.org/

10.4 Программные средства

Для успешного освоения дисциплины студент должен использовать следующие программные средства/ресурсы:

- Национальный корпус русского языка (http://www.ruscorpora.ru/)
- средства Microsoft Office/ OpenOffice
- текстовый редактор Notepad++ или его аналог для Mac / Linux
- интерпретатор языка Python (http://www.python.org/download/).

В программе также используются база данных «Частотного словаря русского языка (на материалах НКРЯ)», база данных Грамматического словаря русского языка А. А. Зализняка.

12.Материально-техническое обеспечение дисциплины

Для проведения практических занятий необходимы компьютерные классы, для проведения лекций — проектор.