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

Факультет гуманитарных наук
Школа лингвистики

Рабочая программа дисциплины
«Методы компьютерной обработки лингвистических данных»

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

Разработчики программы
Архангельский Т.А., канд. филол. наук, tarkhangelskiy@hse.ru
Орехов Б. В., канд. филол. наук, borekhov@hse.ru
Мустакимова Э.Г., cclikespizza@gamil.com

Одобрена на заседании Школы лингвистики ФГН «30» мая 2016 г.
Руководитель Школы лингвистики Е.В. Рахилина

Рекомендована Академическим советом образовательной программы
«01» июня 2016 г., № протокола 10

Утверждена «01»июня 2016 г.
Академический руководитель образовательной программы
Ю.А. Ландер

Москва, 2016 г.

1 Область применения и нормативные ссылки

Настоящая программа учебной дисциплины устанавливает требования к образовательным результатам и результатам обучения студента и определяет содержание и виды учебных занятий и отчетности.

Программа предназначена для преподавателей, ведущих дисциплину «Методы компьютерной обработки лингвистических данных», учебных ассистентов и студентов направления 45.03.03 «Фундаментальная и прикладная лингвистика» факультета гуманитарных наук.

Программа учебной дисциплины разработана в соответствии с:

1. Образовательным стандартом НИУ ВШЭ;
2. Образовательной программой направления 45.03.03 «Фундаментальная и компьютерная лингвистика» подготовки бакалавра;
3. Рабочим учебным планом НФ НИУ-ВШЭ на 2016/2017 по направлению подготовки 45.03.03 «Фундаментальная и прикладная лингвистика», утвержденным в 2016 году.

2 Цели освоения дисциплины

Цель курса — научить слушателей применять компьютерные технологии (в первую очередь, язык программирования Python) для решения возникающих на практике лингвистических задач: автоматическая обработка и анализ текстовых данных, поиск информации и др. Часть курса посвящена изучению программирования на языке Python, алгоритмов и модулей. Также, одной из целей освоения дисциплины «Методы компьютерной обработки лингвистических данных» является знакомство с форматами лингвистических данных, средствами их хранения и предоставления открытого доступа к ним.

3 Компетенции обучающегося, формируемые в результате освоения дисциплины

В результате освоения дисциплины студент должен:

- *Знать:*
- методы представления результатов исследования в виде баз данных и доступных в интернете ресурсов;
- способы хранения информации на электронных носителях;
- методы автоматической обработки информации с помощью языка программирования Python;
- основы работы с unix;
- форматы HTML, XML, JSON, используемые для хранения текстовых данных.

- *Уметь:*
- публиковать свои данные на веб-сайте;
- пользоваться редактором Notepad++ и программами сравнения текстов для ручной обработки текстовых данных;
- строить алгоритмы для решения практических задач;
- использовать средства языка Python для реализации алгоритмов;
- пользоваться консолью unix, работать с файловой системой, ставить пакеты;
- подключаться к серверу по ssh;
- пользоваться англоязычной документацией языка Python.

- *Иметь навыки (приобрести опыт):*
- работы с материалом, собранным в сети Интернет с помощью Python;
- работы с программами морфологического анализа (Mystem);

- сбора и первичной обработки данных с использованием Python;
- представления материала в виде баз данных;
- построения алгоритмов для решения практических задач;
- реализации алгоритмов средствами языка Python;
- использования языка регулярных выражений.

В результате освоения дисциплины студент осваивает следующие компетенции:

Компетенция	Код по ФГОС/ НИУ	Дескрипторы – основные признаки освоения (показатели достижения результата)	Формы и методы обучения, способствующие формированию и развитию компетенции
Способен учиться, приобретать новые знания, умения, в том числе в области, отличной от профессиональной	УК-1	понимает постановку задачи лингвистического исследования с точки зрения использования возможностей соответствующих электронных ресурсов для сбора лингвистического исследования; умеет применять компьютерные инструменты для сбора лингвистических данных и их обработки	- чтение специальной литературы -выполнение самостоятельных заданий по сбору данных - анализ полученных данных с использованием Python
способен проводить формализацию лингвистических знаний, анализ и синтез лингвистических структур, квантитативный анализ лингвистических данных с использованием математических знаний и методов	ПК-2		
способен участвовать в создании представительных текстовых массивов, корпусов текстов, корпусов звучащей речи, мультимодальных корпусов, лингвистических исоциолингвистических баз данных и пользоваться этими ресурсами	ПК-11		
способен проектировать системы анализа и синтеза естественного языка, анализа и синтеза мультимодальных язы-	ПК-12		

Компетенция	Код по ФГОС/ НИУ	Дескрипторы – основные признаки освоения (показатели достижения результата)	Формы и методы обучения, способствующие формированию и развитию компетенции
ковых 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 входит в базовую часть профессионального цикла (модуль «Программирование»).

При изучении дисциплины используются знания и навыки, полученные в результате освоения дисциплины «Компьютерные инструменты лингвистического исследования» и «Программирование (язык Python)» (1 курс).

Основные положения дисциплины должны быть использованы в дальнейшем при изучении следующих дисциплин: Программирование (язык Python) (курсы 2, 3 и 4), Базы данных, Автоматическая обработка естественного языка (курсы 3 и 4), Информационный поиск и извлечение данных, Компьютерная лингвистика, Онтологии и семантические технологии (курс 2).

5 Тематический план учебной дисциплины

[Тематический план отражает содержание дисциплины (перечень разделов), структурированное по видам учебных занятий с указанием их объемов в соответствии с ОУП]

№	Название раздела	Всего часов	Аудиторные часы	Самостоятельная работа
			Практические занятия	
1	Сбор и обработка текстовых данных с помощью Python	46	20	26
2	Форматы и хранение лингвистических данных	36	16	20
3	Модули и другие инструменты Python	35	15	20

4	Введение в создание веб-приложений	35	15	20
	ИТОГО	**152**	**66**	**86**

6 Формы контроля знаний студентов

Тип контроля	Форма контроля	2 курс				Параметры **
		1	2	3	4	
Текущий	Контрольная работа	*	*			письменная работа 80 минут
	Домашнее задание (проект)	*			*	1 модуль: работа по созданию корпуса текстов 4 модуль: групповая работа по написанию технического задания и выполнению работы по полученному ТЗ
Итоговый	Экзамен		*		*	письменный экзамен, 120 минут

7 Критерии оценки знаний, навыков

- Выполненные домашние задания студенты загружают в свои репозитории на веб-сервисе https://github.com/. Домашние задания, если явно не указано иное, необходимо выложить в репозиторий до 12:00 дня, предшествующего следующему семинару.
- При оценивании программы в первую очередь обращается внимание на то, насколько её работа соответствует требованиям, описанным в задании. Программа, не запускающаяся из-за синтаксических ошибок, не может получить оценку выше 4 баллов. Баллы могут сниматься, в частности, за неточное выполнение задания и отсутствие разбора случаев, из-за которых при исполнении программы может произойти ошибка. Во вторую очередь могут оцениваться оптимальность решения (в смысле времени работы программы и количества строк кода) и стиль.
- На контрольной работе в 1 модуле проверяется умение пользоваться модулем urllib, обрабатывать html-файлы с помощью Python: читать, извлекать информацию, создавать новые файлы.
- На контрольной работе в 2 модуле проверяется знание функций языка Python, предназначенных для обработки текста, знание инструментов Python для работы с базами данных.
- В проектной работе в 1 модуле студенты должны создать корпус газетных текстов. Каждый студент должен обработать тексты на веб-сайте одного публицистического издания. Итогом работы является набор проанализированных текстов в специальном формате.
- В 4 модуле предполагается групповая проектная работа. Каждая группа должна а) составить план проекта лингвистического ресурса и написать техническое задание (ТЗ) для другой группы, б) получить от другой группы ТЗ и создать на основе этого ТЗ полноценный ресурс.
- На экзамене проверяются все знания и умения, приобретённые во время изучения настоящей дисциплины к моменту проведения экзамена.
- Все контрольные работы, зачёты и экзамен проводятся в письменном виде; все практические задания выполняются на компьютере.
- Основной частью задания контрольной работы и экзамена 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.

8 Содержание дисциплины

1. **Сбор и обработка текстовых данных с помощью Python.**
   Модуль urllib. Язык разметки HTML. Язык разметки XML. Модули html и lxml. Использование Mystem. Создание газетного корпуса.
2. **Форматы и хранение лингвистических данных.**
   Введение в SQL. Работа с базами данных в Python. Формат JSON.
3. **Модули и другие инструменты Python.**
   Генераторы списков и словарей. Модуль time. Увеличение скорости работы программы. Структуры данных: кортежи (tuples) и множества (sets).
4. **Введение в создание веб-приложений.**
   Создание HTML-форм. Запросы GET и POST. Основы Flask. Основы unix: работа с консолью, установка пакетов, логин в сервер по ssh.

9 Образовательные технологии

Для изучения дисциплины необходим компьютер и следующее программное обеспечение: редактор электронных таблиц MS Excel или OpenOffice Calc; 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).

10 Оценочные средства для текущего контроля и аттестации студента

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

Примерный список типов вопросов к контрольным и экзаменам по курсу:

- 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 программу для сбора текстов с определенного интернет-ресурса.
- Составить таблицу на основе полученного XML-файла.

11 Порядок формирования оценок по дисциплине

Преподаватель или учебный ассистент каждую неделю оценивает самостоятельную работу студентов, проверяя домашние и проектные работы. Оценки за самостоятельную работу студента выставляются в рабочую ведомость. Накопленная оценка по десятибалльной шкале за самостоятельную работу определяется перед промежуточным или итоговым контролем — $О_{сам.\ р.}$.

Накопленная оценка за текущий контроль равна среднему арифметическому оценок за контрольные работы. Таким образом, в конце курса текущий контроль считается по формуле:

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

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

$$О_{экзамен} = 1/2 \cdot О_{экз\ 1} + 1/2 \cdot О_{экз\ 2}.$$

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

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

- экзамены — 35% (по 17,5% на каждый экзамен)
- текущий контроль — 35% (по 17,5% на каждую контрольную работу)
- самостоятельная работа и проекты— 30%

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

12 Учебно-методическое и информационное обеспечение дисциплины

12.1 Базовый учебник

Курс лекций.

12.2 Основная литература

Джеффри Фридл. Регулярные выражения (3-е издание). Символ-плюс: М., 2008 (главы из книги)

12.3 Дополнительная литература

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

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/

Документация по Mystem: https://tech.yandex.ru/mystem/doc/index-docpage/

Симулятор консоли Linux: http://bellard.org/jslinux/

12.4 Программные средства

- редактор электронных таблиц MS Excel или OpenOffice Calc;
- текстовый редактор Notepad++ или любой другой, поддерживающий подсветку синтаксиса, переключение между разными кодировками и поиск с использованием регулярных выражений;
- интерпретатор языка Python (http://www.python.org/download/).

13 Материально-техническое обеспечение дисциплины

Для проведения семинаров необходим компьютерный класс. Для проведения лекций и семинаров необходим проектор.