Правительство Российской Федерации

Федеральное государственное автономное образовательное учреждение высшего профессионального образования «Национальный исследовательский университет „Высшая школа экономики“»

Школа лингвистики Гуманитарного факультета

Программа дисциплины

«Программирование (язык Python)»

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

подготовки бакалавра

2 курс

Автор программы: Т. А. Архангельский, к. ф. н. (tarkhangelskiy@hse.ru)

Одобрена на заседании школы лингвистики «12» мая 2015 г.
Руководитель школы Е.В. Рахилина _____ [подпись]

Рекомендована Академическим советом образовательной программы
«19» мая 2015 г., Протокол № 4

Утверждена «21» мая 2015 г.
Академический руководитель образовательной программы
Ю.А. Ландер _____ [подпись]

Москва, 2015 г.

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 отчётности.

Программа предназначена для преподавателей, ведущих данную дисциплину, учебных ассистентов и студентов направления подготовки 45.03.03 «Фундаментальная и прикладная лингвистика», обучающихся по программе подготовки бакалавров «Фундаментальная и прикладная лингвистика», изучающих дисциплину «Программирование (язык Python)».

Программа разработана в соответствии с:

- Образовательным стандартом государственного образовательного бюджетного учреждения высшего профессионального образования Высшей школы экономики, в отношении которого установлена категория «национальный исследовательский университет» (ГОБУ ВПО НИУ—ВШЭ), протокол от 02.07.2010
- Образовательной программой направления «Фундаментальная и прикладная лингвистика» подготовки бакалавра;
- Рабочим учебным планом НФ НИУ—ВШЭ на 2015/2016 гг. по направлению подготовки 45.03.03 «Фундаментальная и прикладная лингвистика», утверждённым в 2015 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. Курс является продолжением одноимённого курса первого года обучения. В рамках курса рассматриваются, во-первых, приёмы программирования и специализированные модули языка Python, а во-вторых, другие инструменты для хранения и обработки лингвистических данных.

3 Компетенции обучающегося, формируемые в результате освоения дисциплины

В результате освоения дисциплины студент должен:

- иметь представление об HTML, XML и других форматах, используемых для хранения текстовых данных;
- уметь пользоваться редактором Notepad++ и программами сравнения текстов для ручной обработки текстовых данных;
- уметь строить алгоритмы для решения практических задач;
- уметь использовать средства языка Python для реализации алгоритмов;
- знать язык регулярных выражений, используемый в языке Python, и уметь его применять для решения задач;
- иметь представление о базах данных и владеть языком SQL на начальном уровне;
- уметь представлять результаты своей работы в виде сервисов, использующих HTML/CSS;

- уметь анализировать и строить GET-запросы для получения информации с серверов, в том числе с помощью модуля urllib.request;
- знать английские эквиваленты всех используемых в курсе терминов и понятий, уметь пользоваться документацией языка Python на английском языке.

4 Место дисциплины в структуре образовательной программы

Настоящая дисциплина входит в базовую часть профессионального цикла (раздел «Программирование»).

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

Основные положения дисциплины и приобретённые навыки должны быть использованы в дальнейшем при изучении следующих дисциплин: Программирование (язык Python) (курсы 3 и 4), Базы данных, Теория автоматов и формальных языков, Автоматическая обработка естественного языка, Информационный поиск и извлечение данных.

5 Тематический план учебной дисциплины

№	Название темы	Всего часов по дисциплине	Аудиторные часы		Самостоятельная работа
			Лекции	Сем. и практ. занятия	
1	Программирование на языке Python: структуры данных, работа с файловой системой, модуль urllib.request	18	0	16	32
2	Инструменты хранения и обработки лингвистических данных	40	0	16	32
	Итого:	**96**	**0**	**32**	**64**

6 Формы контроля знаний студентов

Тип контроля	Форма контроля	1 курс				
		1	2	3	4	
Текущий	Контрольная работа	*				письменная работа, 75 минут
Итоговый	Экзамен	*				письменный экзамен, 120 минут

6.1 Критерии оценки знаний, навыков

- Домашние задания, если явно не указано иное, необходимо присылать на корпоративную почту преподавателя 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 экзамене проверяются все знания и умения, приобретённые во время изучения настоящей дисциплины.
- Все контрольные мероприятия 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.

7 Содержание дисциплины

1. Программирование на языке Python.
1.1. Работа с файловой системой: обход дерева каталогов, создание директорий.
1.2. Множества и операции над ними.
1.3. Модуль urllib2: загрузка веб-страниц и файлов. Использование регулярных выражений для извлечение информации из HTML. Краулеры. Структура GET-запроса, анализ и составление GET-запросов.
1.4. Функции как «объекты первого класса». Аргументы по умолчанию, keyword arguments.
1.5. Списочные и словарные включения.

2. Инструменты хранения и обработки лингвистических данных.
2.1. Реляционные базы данных, СУБД (на примере SQLite). Таблицы, первичный ключ, представление корпусных данных в виде БД, нормализация данных. Работа с БД через графический интерфейс. Основы SQL (SELECT, INSERT, UPDATE).
2.2. HTML и CSS для обеспечения доступа к лингвистическим ресурсам. Веб-формы и GET-запросы.
2.3. Инструменты разметки корпусов, форматы представления корпусных данных и их преобразование.

8 Образовательные технологии

Для изучения дисциплины необходим компьютер и следующее программное обеспечение: графический интерфейс для баз данных SQLite; 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 3.x. Домашние задания необходимо присылать электронной почтой на адрес tarkhangelskiy@hse.ru.

9 Оценочные средства для текущего контроля и аттестации студента

9.1 Вопросы для оценки качества освоения дисциплины

Примерный список типов вопросов к контрольным и зачётам по курсу:

- 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 десятибалльной шкале за самостоятельную работу определяется перед промежуточным или итоговым контролем — $О_{сам.\ р.}$.

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

$$О_{итоговый} = 0{,}4 \cdot О_{экзамен} + 0{,}3 \cdot О_{текущий} + 0{,}3 \cdot О_{сам.\ р.}$$

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

- экзамен — 40%
- текущий контроль — 30%
- самостоятельная работа — 30%

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

11.1 Базовый учебник

Курс лекций.

11.2 Основная литература

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

11.3 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/

11.4 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/);
- графический интерфейс для работы с БД SQLite (например, SQLite DB Browser).

12 Материально-техническое обеспечение дисциплины

Для проведения семинаров необходим компьютерный класс с проектором.