Программа учебной дисциплины «Программирование и компьютерные инструменты лингвистического исследования»

Утверждена
Академическим советом ООП
Протокол № 15 от «28» июня 2018 г.

Автор	Орехов Борис Валерьевич, Бородин Ростислав Алексеевич, Дереза Оксана Владимировна
Число кредитов	4
Контактная работа (час.)	66
Самостоятельная работа (час.)	48
Курс	2
Формат изучения дисциплины	Без использования онлайн-курса

I. ЦЕЛЬ, РЕЗУЛЬТАТЫ ОСВОЕНИЯ ДИСЦИПЛИНЫ И ПРЕРЕКВИЗИТЫ

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.

В результате освоения дисциплины студент должен:

1. **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.
2. **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.

3. **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.

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» (1 курс).

Основные положения дисциплины должны быть использованы в дальнейшем при изучении следующих дисциплин:

- «Базы данных»
- «Автоматическая обработка естественного языка»
- «Информационный поиск и извлечение данных»
- «Компьютерная лингвистика»
- «Программирование и теория алгоритмов»
- «Онтологии и семантические технологии»
- Подготовка и защита выпускной квалификационной работы

II. СОДЕРЖАНИЕ УЧЕБНОЙ ДИСЦИПЛИНЫ

1. **Сбор и обработка текстовых данных с помощью Python.**
   Модуль urllib. Язык разметки HTML. Язык разметки XML. Модули html и lxml. Краулеры и создание веб-корпусов. Bootstrap.
2. **Форматы и хранение лингвистических данных.**
   Введение в SQL. Работа с базами данных в Python. Формат JSON.
3. **Структуры данных и стандартные методы Python.**
   Структуры данных: кортежи (tuples), множества (sets), словари (dictionaries). Генераторы списков и словарей. Обработка исключений.
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. Heroku.
5. **Обработка естественного языка**
   Основы компьютерной обработки текстовых данных. Частотные списки, стоп-слова, закон Ципфа. N-граммы. Корпуса текстов. Лемматизация. Частеречная раметка. Морфологический анализатор Mystem. Библиотеки pymystem3 и pymorphy2. Дистрибутивная семантика и word2vec.
6. **Взаимодействие с внешними сервисами**
   VK API. Telegram API. Twitter API. GitHub API.
7. **Визуализация данных**

Библиотеки matplotlib и seaborn. Динамические графики в вебе. Построение и визуализация графов, библиотека network.

III. ОЦЕНИВАНИЕ

1. 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 23:59 дня, предшествующего следующему семинару.
2. 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 3 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.
3. Каждая домашняя работа состоит из 3-5 задач разного уровня сложности. Для получения положительной оценки необходимо решить задачи, написав программу на языке Python.
4. При обнаружении плагиата в домашнем или контрольном задании это задание получает оценку 0 баллов.
5. Накопленная оценка равна среднему арифметическому оценок за все домашние задания. Итоговая оценка выставляется следующей формуле.

$$O_{нак} = mean(O_{дз})$$
$$O_{итог} = O_{нак} \times 0{,}6 + O_{экз} \times 0{,}4$$

Итоговая оценка округляется в пользу студента.

IV. ПРИМЕРЫ ОЦЕНОЧНЫХ СРЕДСТВ

Оценочные средства для текущего контроля

1. Написать программу, которая получает на вход список пользователей гитхаба и умеет делать следующее:

    a. Выбрать какого-то одного пользователя из полученного списка и распечатать список его репозиториев (name) и их описания (description). Выбор пользователя должен осуществляться с помощью ввода с клавиатуры (функция `input()`).
    b. Распечатать список языков (language) выбранного пользователя и количество репозиториев, в котором они используются.
    c. Узнать, у кого из пользователей в списке больше всего репозиториев.
    d. Узнать, какой язык самый популярный среди пользователей списка.
    e. Узнать, у кого из пользователей списка больше всего фолловеров? (фолловеров можно достать по ссылке `https://api.github.com/users/username/followers`, где вместо username — имя пользователя)

2. Напишите сайт-анкету для полевой работы с информантом. На сайте должны быть:

a. Главная страница (127.0.0.1), на которой показывается анкета с полями. Данные, которые будут вводиться в анкету, должны записываться в файл (лучше всего csv).
b. Страница статистики (127.0.0.1/stats), на которой результаты должны систематизироваться и в удобном виде выводиться на экран (это могут быть таблицы, какие-то подсчеты и тд).
c. Страница с выводом всех данных (127.0.0.1/json), на которой возвращается json со всеми введенными на сайте данными. Этот json должен выводиться на веб-странице.
d. Страница поиска (127.0.0.1/search) и результатов поиска (127.0.0.1/results) . В ней нужно сделать минимум два поля поиска (например, текстовый ввод для поиска конкретного слова в ответах и чекбокс, где можно будет выбрать пол информанта). На странице должно быть описано, по каким данным ведется поиск.

V. РЕСУРСЫ

5.1 Основная литература

1. Курс лекций.
2. Документация по языку Python [Электронный ресурс]. — Режим доступа: http://docs.python.org/
3. Документация по морфологическому анализатору Mystem [Электронный ресурс]. — Режим доступа: https://tech.yandex.ru/mystem/doc/index-docpage/

5.2 Дополнительная литература

1. Лутц М. Изучаем Python. Символ-плюс: М., 2011 или 2014
2. Фридл, Дж. Регулярные выражения (3-е издание). Символ-плюс: М., 2008

5.3 Программное обеспечение

№ п/п	Наименование	Условия доступа
1.	Microsoft Windows 7 Professional RUS Microsoft Windows 10 Microsoft Windows 8.1 Professional RUS	*Из внутренней сети университета (договор)*
2.	Microsoft Office Professional Plus 2010	*Из внутренней сети университета (договор)*
3.	Текстовый редактор Notepad++, Atom или любой другой, поддерживающий подсветку синтаксиса, переключение между разными кодировками и поиск с использованием регулярных выражений.	*Свободно распространяемое ПО*
4.	Интерпретатор языка Python 3.6 и основные библиотеки в сборке Anaconda	*Свободно распространяемое ПО*

5.4 Материально-техническое обеспечение дисциплины

Учебные аудитории для лекционных занятий по дисциплине обеспечивают использование и демонстрацию тематических иллюстраций, соответствующих программе дисциплины в составе:

– ПЭВМ с доступом в Интернет (операционная система, офисные программы, антивирусные программы);

– мультимедийный проектор с дистанционным управлением.

Учебные аудитории для лабораторных и самостоятельных занятий по дисциплине оснащены ПЭВМ, с возможностью подключения к сети Интернет и доступом к электронной информационно-образовательной среде НИУ ВШЭ.