Taller de Python
aplicado a la extracción de datos de sitios Web

Instructores: José Coto - Noam López

jose.coto@pucp.edu // lopez.noam@pucp.edu.pe

18 de agosto de 2015

1. Introducción:

El internet pone a nuestro alcance mucha información. Gobiernos, empresas y ciudadanos usan internet para compartir e intercambiar información. Una ventaja fundamental para los futuros científicos sociales o estudiantes de humanidades será el poder acceder, limpiar y analizar esta data. Python es un lenguaje de programación que nos permite lograr estos objetivos de manera sencilla. Como todo idioma (sea `R`, matemáticas, francés, inglés, etc.), Python tiene una curva de aprendizaje (¡aunque no tan inclinada!). Como todo en la vida, los primeros pasos son los más difíciles. La buena noticia es que equivocarse es gratis y está bien. Es parte del proceso de aprendizaje.

En las cuatro sesiones que dura el curso, no pretendo que se conviertan en expertos programadores. Mi objetivo es que tengan las herramientas básicas para que puedan lanzarse a explorar por su cuenta. Es por eso que el curso hace énfasis en una de las tantas aplicaciones para las que puede usarse Python: la extracción de datos web. La idea es que, más allá de aprender aspectos básicos del lenguaje, puedan apreciar su utilidad. Me doy por satisfecho si el curso despierta en ustedes el deseo por aprender más Python. Me iré más contento aún si descubren cómo sus nuevas habilidades de programación les permite retomar o plantearse proyectos de investigación.

2. Objetivos:

Este taller está centrado en dos objetivos:

1. Introducir a los participantes al lenguaje de programación Python.
2. Utilizar los conceptos aprendidos para la extracción de datos de sitios Web.

3. Temario:

Sesión	Fecha	Temas
Primeros pasos - Tipos de datos en Python	Sábado 5 de Setiembre	- Manejo de enteros, floats, strings y métodos básicos. - Listas: creación, índices, segmentación (slicing), ingreso y borrado de registros, etc. - Diccionarios: creación, acceso a registros, ingreso y borrado de información, etc.
Condicionales, funciones, loops, errores y excepciones	Sábado 12 de Setiembre	- Condicionales: if, elif, else. - Funciones: definición, utilidad, aspectos prácticos. - For loop: nociones básicas. Loops de listas y diccionarios. - While loops: nociones básicas y uso. - Manejo de errores: uso de try y except.
Extrayendo información de Páginas Web	Sábado 19 de Setiembre	- Estructura básica de una página web (HTML, CSS) - HTML - Relaciones de parentesco entre elementos. - BeautifulSoup - Importación de módulo y métdos básicos. - Manos a la obra. Ejercicio práctico.
Automatizando el proceso de Extracción - Guardando información	Sábado 26 de Setiembre	- Robobrowser: módulo para automatización de extracción de datos. - CSV: módulo para guardar y leer información de datos separados por comas. - Manos a la obra. Ejercicio Práctico.

4. Software:

Python es gratutio. Para el taller utilizaremos la versión Python 3.X.[1] Por favor, verifica que tu computadora cuente con cualquiera de esas versiones. El taller usará IPython notebooks. Los notebooks permiten trabajar archivos de texto y código a la misma vez. Tienen también una serie de ventajas adicionales a trabajar con Python en una consola. Es una gran herramienta de aprendizaje.

Para poder trabajar con notebooks, necesitarás instalar una serie de paquetes a tu computadora. La manera más sencilla de asegurarte que tienes todos los paquetes necesarios es usar la distribución Anaconda. Puedes encontrarla en http://continuum.io/downloads#py34. Asegúrate de marcar que quieres la opción de Python 3.4. **Por favor, asegúrate de haber instalado todos los componentes necesarios antes de la primera sesión. Por motivos de tiempo, no me puedo**

[1] Cualquiera de las versiones de Python 3.0, en adelante. Los usuarios de MAC tienen Python instalado en su sistema operative. Sin embargo, la versión es 2.7. Por favor, comprueben que tienen instalado y están trabajando con la versión 3.

detener a revisar los errores de instalación de último minuto que puedan surgir.

5. Referencias útiles:

Parte del aprendizaje es que le encuentres el gusto a Python. Esto es algo que irás adquiriendo a medida que practiques, practiques y practiques. A continuación te dejo algunas páginas web que te pueden servir en este proceso.

- *Learn Python the Hard Way.* Zed A. Shaw.
  Libro muy completo y guía básica para aprender Python.
  Disponible en http://learnpythonthehardway.org/book/.

- Trinket - Básico.
  Disponible en https://trinket.io/.
  Esta es una buena página web para iniciarse en Python. Hay cursos elementales con actividades que pueden ser completadas desde la propia página web.

- *Checkio().* - Intermedio.
  Disponible en http://www.checkio.org/.
  Puedes crear un usuario y pasar los diferentes retos de programación propuestos por los demás usuarios.

- *Documentación*
  Disponible en https://docs.python.org/3/.
  Página oficial de Python con información sobre los módulos básicos incluidos en el programa.

- *Stackoverflow*
  Disponible en http://stackoverflow.com/questions/tagged/python.
  Si tienes alguna duda o dificultad respecto a Python, lo más probable es que otra persona se haya topado con la misma piedra. Stackoverflow reúne todas las respuestas a los problemas que tiene la gente con Python. Así disfrutarás otro aspecto de este lenguaje: el que tiene una comunidad muy amigable y dispuesta a ayudar.