PR0MiDAT

IBEROAMERICANO

Programa Iberoamericano de Formación en Minería de Datos

ML3006

Web Mining en Python

(506) 4030.1205 - (506) 4030.1114 info@promidat.com facebook.com/promidat Twitter.com/promidat www.promidat.com

Índice

Duración 2

Descripción 2

Objetivos 2

Metodología 2

Contenido 3

Evaluación 4

Bibliografía 4

Duración

Cuatro semanas.

Descripción

En este curso se presentarán los conceptos y tecnicas necesarios para extraer datos de la web. Se estudian los fundamentos de HTML, CSS y JavaScript para entender el funcionamiento de una página web. Se estudiará el uso de expresiones regulares y su aplicación en la minería de texto para la correcta extración de datos de la web. Se aprenderá a extraer datos tanto de páginas web estáticas como dinámicas, lo cual incluye redes sociales.

Objetivos

En este curso el estudiante será capaz de:

1. Entender los elementos básicos de un página web.
2. Aprender a extraer información de páginas web estáticas.
3. Aprender a extraer información de páginas web dinámicas.
4. Aprender a extraer información de Redes Sociales.
5. Manejar páginación y navegación de páginas web de forma automatizada.
6. Aplicar expresiones regulares para la limpieza y extracción de datos.
7. Analizar información obtenida de páginas web.

Metodología

Basado en la teoría y en la aplicación directa de los conceptos aprendidos. Para esto se dispondrán de las siguientes herramientas:

1. Una vídeo conferencia semanal, las cuales quedarán grabadas en Zoom, para que los alumnos la puedan acceder en cualquier momento.
2. Trabajos prácticos semanales.
3. Foros para plantear dudas al tutor y compañeros.

4. Aula virtual en Moodle.

Luego de este curso el estudiante será capaz de:

Desarrollar y programar proyectos de Machine Learning que involucren extraer, limpiar, corregir y analizar información de sitios web utilizando código automatizado en `Python`.

Contenido

1. **Conceptos básicos de una página web.**
    a. ¿Qué es una página web?
    b. Diferencias entre una página web estática y una dinámica.
    c. ¿Qué es Web Mining, Web Scraping y Text Mining?
    d. `HTML` (Hyper Text Markup Language).
    e. `CSS` (Cascading Style Sheets).
    f. `JavaScript`.
2. **Extracción de datos de la web (Web Scraping).**
    a. XPath como lenguaje de consultas para documentos WEB.
    b. `BeautifulSoup` como herramienta para extraer datos de páginas web estáticas.
    c. `Scrapy-Splash` como herramientas para extraer datos de páginas web dinámicas.
    d. Manejar paginación y navegación.
    e. Minando las Redes Sociales.
3. **Minería de texto (Text Mining).**
    a. Expresiones regulares.
        I. Clases de caracteres.
        II. Cuantificadores.
        III. Aserciones.
    b. Manipulación y limpiaza de textos.
        I. Detección de patrones de texto.
        II. Extracción de patrones de texto.
        III. Eliminación de patrones de texto.
        IV. Modificación de patrones de texto.
4. **Análisis de datos obtenidos de la web.**

a. Preparar los datos para un análisis de métodos no supervizados.

b. Preparar los datos para un análisis de métodos supervizados.

c. Nubes de palabras (wordclouds).

d. Análisis de Sentimientos.

Evaluación

El curso se evalúa con 4 tareas, una por semana, cada tarea tiene un valor de 25 puntos. La nota mínima de aprobación es de 70.

Bibliografía

[1] Seppe Vanden Broucke and Bart Baesens. *Practical web scraping for data science: best practices and examples with Python*. In collab. with Seppe Vanden Broucke and Bart Baesens. CreateSpace, 2017. 306 pp. URL: https://eprints.soton.ac.uk/425855/ (visited on 10/10/2022).

[2] Markus Hofmann and Andrew Chisholm. *Text Mining and Visualization: Case Studies Using Open-Source Tools*. Google-Books-ID: JfQYCwAAQBAJ. CRC Press, Jan. 5, 2016. 337 pp. ISBN: 978-1-4822-3758-0.

[3] Dimitrios Kouzis-Loukas. *Learning Scrapy*. Google-Books-ID: EF8dDAAAQBAJ. Packt Publishing Ltd, Jan. 30, 2016. 270 pp. ISBN: 978-1-78439-091-4.

[4] Bing Liu. *Web Data Mining*. Berlin, Heidelberg: Springer, 2011. ISBN: 978-3-642-19459-7 978-3-642-19460-3. DOI: 10.1007/978-3-642-19460-3. URL: http://link.springer.com/10.1007/978-3-642-19460-3 (visited on 10/10/2022).

[5] *Mining the Social Web, 3rd Edition [Book]*. ISBN: 9781491985045. URL: https://www.oreilly.com/library/view/mining-the-social/9781491973547/ (visited on 10/10/2022).

[6] Ryan Mitchell. *Web Scraping with Python: Collecting More Data from the Modern Web*. Google-Books-ID: TYtSDwAAQBAJ. "O'Reilly Media, Inc.", Mar. 21, 2018. 329 pp. ISBN: 978-1-4919-8552-6.