Tipología y ciclo de vida de los datos

Máster en Ciencia de datos

PRA-1

Web Scraping para estándares publicados por ETSI

apulidod

Alejandro J Pulido Duque

1. Contexto

Los estándares de Telecomunicaciones son documentos donde se define una tecnología, bien sea su arquitectura, sus requerimientos, su testeo, casos de uso, o guías de desarrollo.

ETSI: European Telecommunication Standard Institute, es la agencia europea de estandarización para las telecomunicaciones [1] cuyos estándares de tecnología tienen alcance global, más específicamente, ETSI desarrolló los primeros estándares para las redes GSMA, y sus posteriores evoluciones hasta alcanzar 5G mediante el grupo 3GPP. Estos estándares se realizan mediante las contribuciones de las empresas miembros de ese comité.

En este caso concreto, se va a recolectar la información de las publicaciones de ETSI correspondientes a tecnología SIM (UICC/eUICC) cuya página puede ser consultada en el siguiente enlace:

https://www.etsi.org/deliver/etsi_ts/102200_102299/

Las especificaciones correspondientes de ETSI se organizan en Releases, actualmente se encuentran desarrollando los Releases 16 y 17 que corresponden con las actualizaciones de 5G para tecnologías “Smart Card”, tal y como se publica en su página [2].

Las publicaciones se realizan periódicamente, existiendo aproximadamente unos 70 documentos, con 3 publicaciones de media por Release, y teniendo en cuenta que antes del reléase 5 todas las especificaciones se agrupan en uno, estaríamos hablando de unos 700 documentos.

Por lo anterior, una herramienta que sea capaz de extraer las nuevas publicaciones de su base de datos de manera automatizada es bastante útil.

De hecho, el programa podría cargar la información de cualquier tanda de especificaciones.

2. Datos generados

Los datos recopilados por el scraper se recogen en el archivo csv generado:
Specs200_299

El dataset recogerá la siguiente información:

- Spec: es la especificación ETSI (102 221, 102 204, etc)
- Date: Fecha en la que se subió la versión
- Release: Release al que pertenece la versión

- Version: versión subida

En definitiva, el dataset tendrá la estructura siguiente:

	Spec	Date	Release	Version
0	102204	8/28/2003	1	01.04_60
1	102207	8/28/2003	1	01.03_60
…	…	…	…	…

3. Agradecimientos

Los datos se han recolectado del repositorio de publicaciones de ETSI de la cual soy miembro y contribuyente activo.

Para la realización de la herramienta se ha utilizado el lenguaje de programación Python y técnicas de web scraping de la asignatura Tipología y ciclo de vida del dato [3], así como de “Web Scraping with Python” [4]

4. Inspiración

Teniendo en cuenta las limitaciones que ofrece el repositorio de ETSI y la poca claridad de su contenido, resulta útil tener alguna herramienta que haga un resumen de las especificaciones, además, siendo una herramienta que podría programarse, podría ofrecer las actualizaciones periódicas de las especificaciones.

5. Licencia

6. Proceso de creación

La herramienta ha sido desarrollada por Alejandro J Pulido Duque, en el proceso de creación se han seguido los siguientes pasos:

Investigación: Se ha investigado acerca de la utilidad y realización de esta herramienta

Redacción de respuestas: Se crea el presente documento para tal fin

Desarrollo de código: El código Python se encuentra disponible GitHub:

www.github.com/ETSIScraper.com

Referencias

[1] https://es.wikipedia.org/wiki/European_Telecommunications_Standards_Institute
[2] https://www.3gpp.org/release-15
[3] Subirats, L., Calvo, M. (2019). Web Scraping. Editorial UOC.
[4] Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.