Introduccion

Problema

Queremos poder usar urls del siguiente tipo: - http:// .tw/ - https:// .la2

Sin embargo, las urls solo pueden tener caracteres ASCII.

Solucion

Es necesario encontrar una forma de representar los caracteres que no pertenecen a ASCII. Para ello se usa el **URL encoding**

URL Encoding

El URL encoding consiste en reemplazar los caracteres que no son ASCII usando un encoding especifico. Esto se realiza usando "Percent encoding"

Percent Encoding

El percent encoding se utiliza para representar un octeto de datos en un componente cuando el carácter correspondiente a ese octeto está fuera del conjunto permitido o se utiliza como delimitador, o dentro del componente. Un octeto codificado porcentualmente se codifica como un conjunto de caracteres, compuesto por el carácter porcentaje "%" seguido de los dos dígitos hexadecimales que representan el valor numérico de ese octeto. Por ejemplo, "%20" es la codificación porcentual para el octeto binario "00100000", que en US-ASCII corresponde al carácter de espacio (SP).

Fuente: rfc3986

Tipos de caracteres en una URL

Caracteres seguros:

Son los caracteres alfanuméricos, es decir, 0-9, a-z y A-Z, caracteres especiales $, -, _, ., +, !, *, ', (,), son caracteres reservados que tienen funciones especificas.

Estos caracteres no necesitan ser codificados.

Caracteres de control ASCII:

Incluye los caracteres que van desde 00-1F en hex (0-31 decimal) y 7F (127 decimal).

Estos caracteres deben ser codificados.

Caracteres de control no ASCII:

Incluye 80-FF en hex (128-255 decimal).

Estos caracteres deben ser codificados.

Caracteres reservados:

Estos caracteres se utilizan para fines especiales y requieren codificación.

Caracteres inseguros:

Estos caracteres pueden ser malinterpretados dentro de las URL por varias razones. Por lo tanto, requieren codificación.

Ejemplo Los caracteres < y > ya que se utilizan como delimitadores alrededor de las URL en texto libre