

Challenge

Bij webontwikkeling en web scraping kan het vaak handig zijn om links uit een HTML-pagina te extraheren. Dit kan gedaan worden met een HTML-parser zoals BeautifulSoup, maar als oefening willen we dat je dit probleem oplost met behulp van reguliere expressies.

Opdracht

Schrijf een Python-script dat een HTML-bestand leest en alle URL's uit de `<a>`-tags (hyperlinks) extraheert en afdruckt. De URL's zijn de waarden van het `href` attribuut in `<a>`-tags. Hier is een voorbeeld van een `<a>`-tag:

```
<a href="https://www.voorbeeld.nl">Een voorbeeld link</a>
```

In dit voorbeeld zou je script "<https://www.voorbeeld.nl>" moeten extraheren en afdrukken.

Tips

1. Je kunt de `re.findall()` functie gebruiken om alle overeenkomsten van een patroon in een string te vinden.
2. Vergeet niet om rekening te houden met de verschillende manieren waarop een attribuut kan worden geschreven in HTML. Bijvoorbeeld, het kan worden omringd door dubbele aanhalingstekens (`href="url"`), enkele aanhalingstekens (`href='url'`), of het kan geen aanhalingstekens hebben (`href=url`).

Inleveren

1. Jouw eigen test bestand waarin meerdere links staan. De links staan tussen `"` en `"` (dubbele quote en enkele quotes).
2. Jouw code.
3. Een screendump waarin je laat zien dat jouw code werkt.
Laat in de screendump ook de datum en tijd van Windows zien (dit is wat je rechtsonder in beeld ziet).

Succes!

Revision #3

Created 26 July 2023 18:09:24 by Max

Updated 31 July 2023 14:03:23 by Max