

Vorprojekt: Markierung eines durch einen Hash bestimmten Text einer Webseite

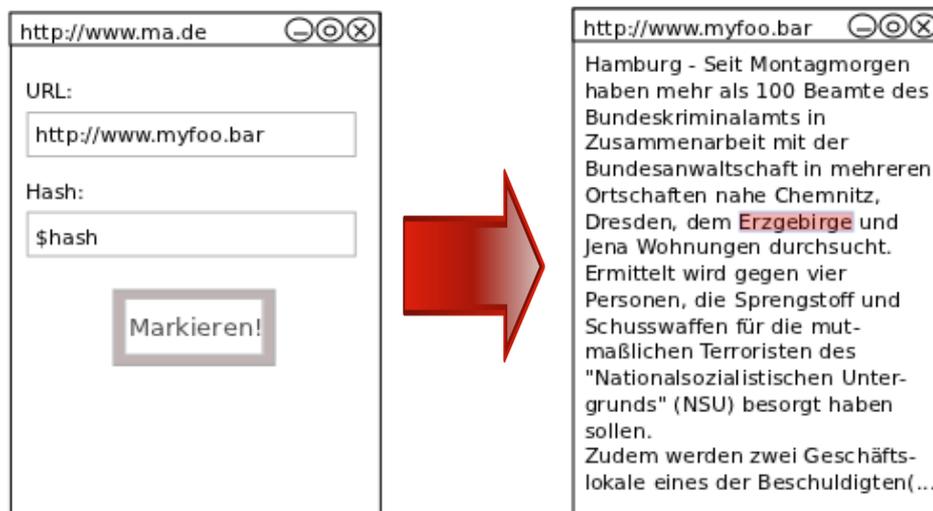
Aufgabenstellung:

Das NLP Interchange Format, oder kurz NIF definiert unter anderem ein hash-basiertes Schema um Elemente in Hypertext zu identifizieren und Annotationen in Dokumenten mit Hilfe von Fragment Identifier zu verankern ([siehe Recherchebericht](#)).

In diesem Vorprojekt soll nun ein Webservice entstehen, der solch ein Hash-basierter Fragment Identifier und eine HTML-DATEI/URL entgegennimmt, die durch den Hash beschriebene Textstelle im HTML Code findet und die Webseite mit der deutlich gekennzeichneten Textstelle ausgibt.



Sei dies zum Beispiel eine beliebige HTML-Seite und das Wort Erzgebirge soll markiert werden. Das Bestimmen der entsprechenden Fragment-ID ist nicht Gegenstand dieses Projekts. Daher nehmen wir diese als gegeben und bezeichnen sie mit \$hash.



Die geplante Funktionsweise unseres Vorprojekts

Begründung des Projekts:

Die hier zu implementierende Funktion der Bestimmung einer Textstelle zu einer Fragment-ID wird auch später im Projekt essentiell sein, um eine Annotation einer exakten Textstelle auf einer Webseite zuordnen zu können und wird daher auch direkt übertragbar sein.

Die Funktionsweise entspricht auch, stark vereinfacht, unserem Projektvorschlag und würde den unerfahreneren Teammitgliedern einen ersten Einblick in die genutzten Werkzeuge und Technologien ermöglichen.

Vergleich mit dem LiveURLs-Projekt (<http://liveurls.mozdev.org/>):

Liveurls ist ein Firefox-Plugin, mit dem Nutzer Textstellen markieren können und diese Markierungen untereinander austauschen können. Hierbei wird die URL der zu markierenden Seite durch Parameter ergänzt, die vom Plugin ausgewertet werden und so die zu markierende Textstelle identifizieren.

Als Parameter genutzt werden:

- Länge des ersten Wortes der Selektion
- das erste Wort der Selektion
- Gesamtlänge der Markierung
- eine Prüfsumme der Selektion

Die Parameter entsprechen demnach grob denen im NIF definierten Standard einer hash-basierten URI, mit dem Unterschied, dass im NIF noch zusätzlich der Kontext des markierten Textes beachtet wird.

So ist es bei unserer Herangehensweise möglich, mehrere identische Textstellen auf einer Webseite anhand ihres Kontextes (mehrere Zeichen vor und nach des markierten Textes) zu unterscheiden und nur passende zu markieren, während bei LiveURLs zwangsläufig jedes Vorkommen markiert wird.