Developing An Imprintcrawler – A Rule-Based Approach

Background

The German „Telemediengesetz“ and „Medienstaatsvertrag“ mandates the content of an imprint of a website.

The information, contained within an imprint document, can be used to get more insight about the provider and also the general integrity.

Example: a set of websites can be analyzed and by the information about the address contained in the imprint, different websites can be set into relation.

The Imprint

The content of a website's imprint depends on it's usage. Which can be of the following way:

commercial vs. ***non commercial***
&
editorial content vs. ***non editorial content***

The information that needs to be provided ranges from the name and address of the firm, or person in charge.
Up to information of the still held capital when a firm is in the process of liquidation.

Overall, there are 10 different fields, that can be divided into smaller subfields.

Beside the information regarding the content of an imprint, the legislative also gives insight about the accessibility of an imprint document.

The imprint needs to be accessible behind a navigation tab which is named „Kontakt“ or „Impressum“ and it needs to be accessible from every point of a website by at most two clicks.

Approach

The basis for the overall apporoach was to define rules and translate them into code. The rules look like this:

$$X \longrightarrow Y$$

And can be interpreted as an implication: *when „X“ can be observed, „Y“ can be assumed* (Sarawagi, 2008). To define these rules in a first step, domain specific knowledge was collected through research and observation.

This domain knowledge can be separated into three main categories:

- ***Content specific structural information***
  Example: A postcode in germany consists of 5 digits.
- ***Html/text specific structural information***
  Example: At the end of a postcode-city combination comes a tag, after a street is mentioned a postcode follows.
- ***Content specific information***
  Example: A street name contains a typical word that gives indication if a street name is present.

Note: While the different fields are often displayed in a specific order, their occurrence differs enough that such observations can not be taken into account to formulate rules. So an overall structural approach would be deemed to fail.

Implementation

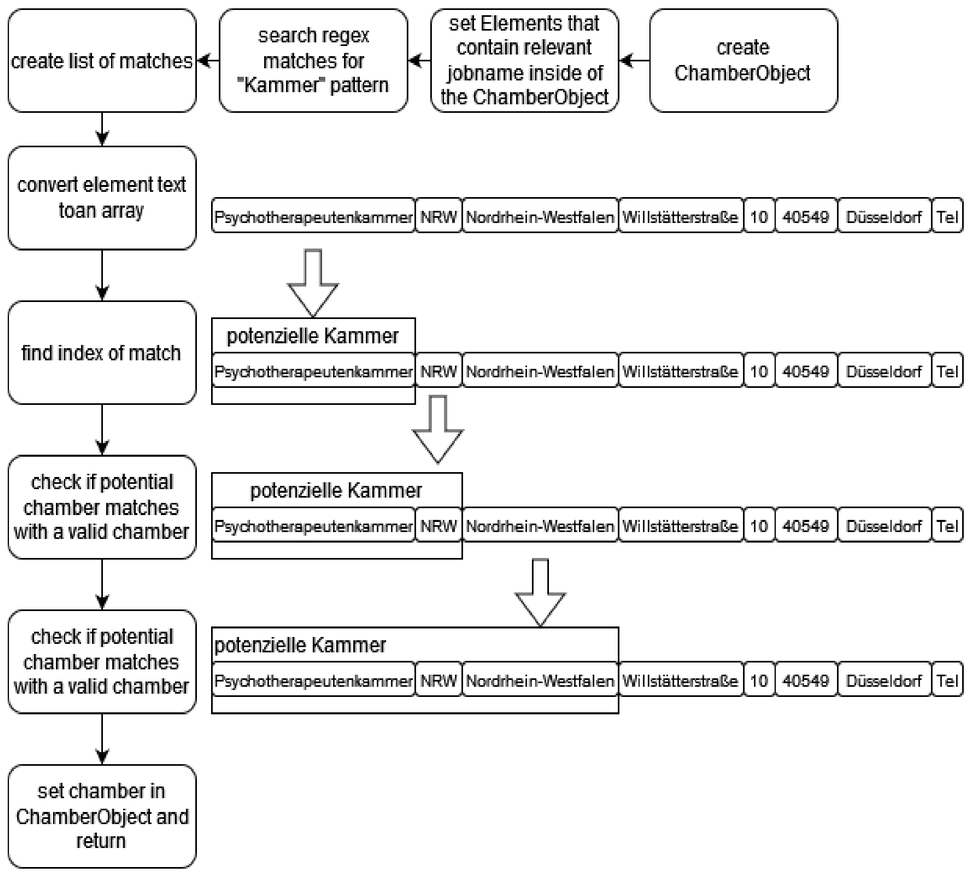
With the help of manually curated lists, web scraping of useful information and observation the formulated rules were translated into different algorithms for the field extraction.

Some fields could be extracted through the use of a regular expression:

[\s>][a-z][a-z](\s?\d){9}

Regular expression used for finding an „Umsatzsteueridentifikationsnummer“

While other needed more complex approaches:



Visualization of the algorithm that finds the different chambers that need to be listed by specific job professions

Results

While for some fields, the used rules lead to a very good extraction rate, some showed unsatisfactory results. But the unsatisfactory ones are probably due to a too small set of rules and not because of an in general failed approach.

Especially the fields that were tackled by „wholemeal“ approaches and had a high heterogeneity in the way they appeared had problems in achieving high precision. But the approaches that worked on the other fields could also be applicable to these fields.

With more rules, we should be able to give a definitive answer to the question if the approach reaches its limits or if these specific fields are more complex and hence need more rules to be adequately captured.

But even with the small set of rules, some fields could be extracted with an accuracy of nearly 100% and the general approach seems to be effective.

Overall it could be shown that the formulated rules can be translated into a functioning programmatic context and the approach seems promising.

Finally, a short assessment in regard to the performance:

Good performance:
- Ust-Number
- Contact Infos
- Address
- Chamber
- Editorial indicator

Mediocre performance:
- Legal form
- Representative
- Editor info

Quellen:
MStV. (2023). https://www.medienanstalt-nrw.de/zum-nachlesen/rechtsgrundlagen/staatsvertraege.html
Sarawagi, S. (2008). Information Extraction. *Foundations and Trends in Databases*, 1, 261–377. https://doi.org/10.1561/1500000003
TMG. (2007). https://www.gesetze-im-internet.de/tmg/__5.html

UNIVERSITÄT DUISBURG ESSEN
Open-Minded