WDEI Stage 5, Exercise 3:

Q-gram-based Tree Similarity Algorithm

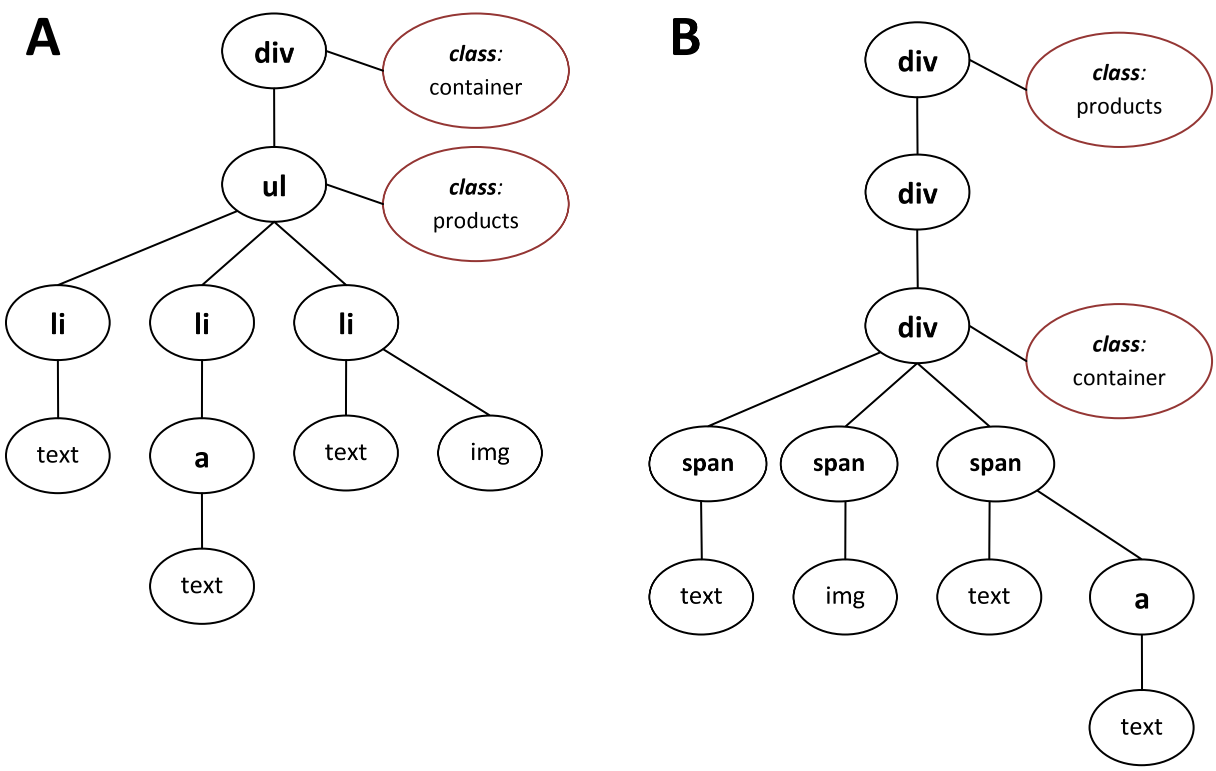
Our algorithm is specifically designed for comparing HTML fragments.

It is based on the following principles:

1. Take the semantics of HTML-elements into account.
2. Don’t compare tree A and tree B directly. Instead, turn both A and B into (simpler) “abstract trees”. Use these abstract trees for the actual similarity analysis.
3. Optional: Perform further simplifications on the abstract trees.
4. Compute q-grams (e.g. with q=2) for each level of $A_{abstract}$ and $B_{abstract}$.
5. Use q-grams together with a penalty-system to compute the actual degree of similarity between A and B.
   E.g.
   …each matching q-gram results in a bonus of +1, each non-matching q-gram results into a score of -1.
   …if the “id”- or “class”-attributes of two elements on the same level match, this gives a bonus of +10.

Description

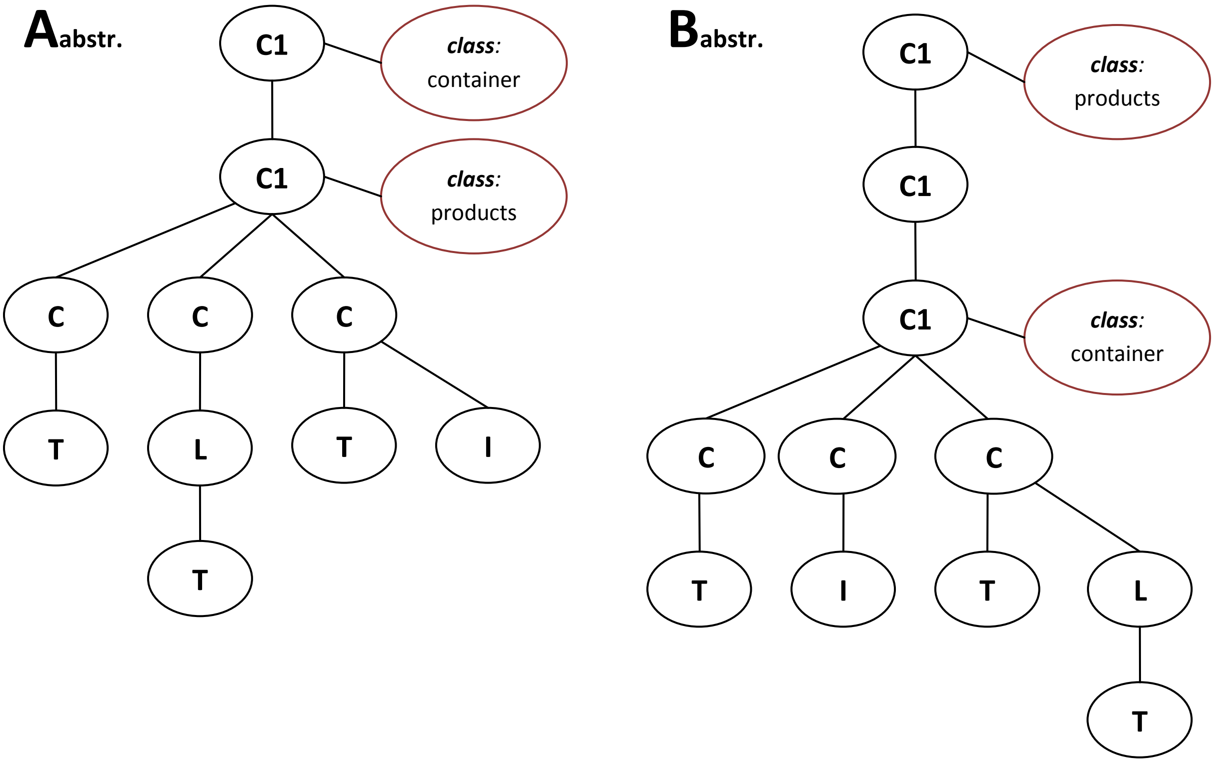
For the explanation of our algorithm, we use the following example trees:



1 – Turn A and B into abstract trees.

This is done by replacing each concrete HTML tag with its category. Categories are user-defined and group similar tags (e.g. “span” and “div”) together.

The resulting abstract trees look like this:



Categories used are:

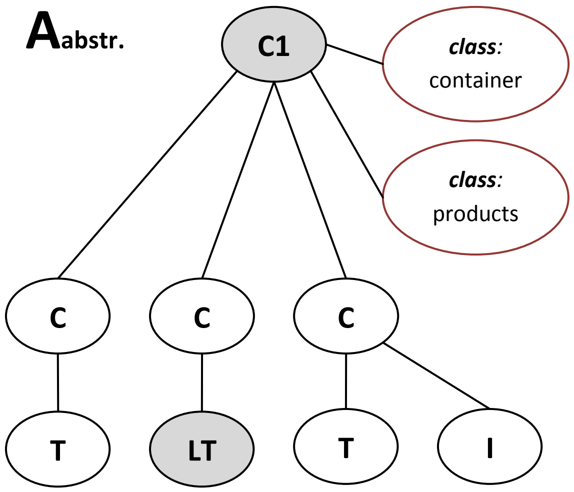
- C1 ...top-level container (body, div, table, ul, ol)
- C ...container (span, li, tr, th, td)
- L ...links (a)
- T ...text “text content”

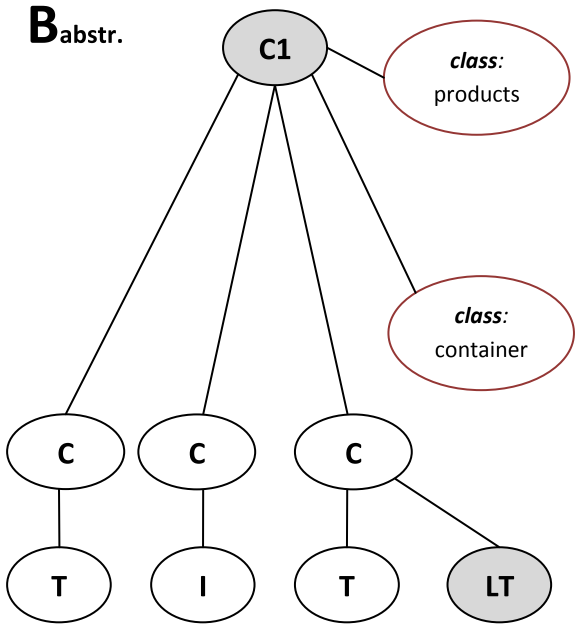
2 – Perform simplifications on the abstract trees

Starting from the root nodes of A*abstract* and B*abstract*, rules are applied that make the abstract trees even simpler. These rules exploit general knowledge about the structure of content on the web. As an example, we provide 2 rules:

1. *If a top-level-container contains nothing but another top-level-container, merge those two containers and all of their attributes.*
   This rule aims at nested containers that are only there for styling purposes.
2. *If a link element has only text as its single child, merge these two elements into a single element “LinkText” (LT).*

If these rules are applied, the abstract trees now look like this:





3 – Compute q-grams for each level of A and B

For q=2, the result is:

Level	Tree A	Tree B
1	{ C1 }	{ C1 }
2	{ (C,C) }	{ (C,C) }
3	{ (T,LT), (LT,T), (T,I) }	{ (T,I), (I,T), (T,LT) }

4 – Compute the degree of similarity using q-grams and a penalty system

At the beginning, the global variable score is set to 0.

Iterating over each *row of the q-gram* table:

- each matching q-gram results into score + 1
- each non-matching q-gram results into score - 1

Iterating over the *nodes* of each level of $A_{abstract}$ and $B_{abstract}$:

- each matching “id” attribute results into score + 5
- each matching “class” attribute results into score + 3

At the end, “score” directly indicates the degree of similarity between A and B - meaning, a high positive value indicates high similarity and vice versa.

In the example, the resulting score is 0 (+ 1) (+ 1) (+ 1 -1 +1) (+3 +3) = 9

Since the max. number of nodes is 13, a value of 9 indicates a rather high level of similarity.

5 – Conclusion

The algorithm is clearly targeted at trees that represent HTML code.

By defining custom tag-categories and rules that take domain knowledge of the structure of HTML pages into account, it is quite flexible and extendable.

Since the algorithm replaces the actual HTML-tags with their (custom defined) categories, it works on a “higher” level of abstraction. This might, however, also lead to oversimplification.

The major disadvantage of the proposed algorithm is that it computes and evaluates q-grams per level. Even though there is a simplification step that takes care of things like nested divs, a simple structural difference on level n between otherwise very similar trees A and B might lead to an unintended result (because, level n of A should actually be compared to level n+1 of B).

To avoid this, the evaluation of q-grams could be done in a more complex way. For example, level n is compared to levels n, n+1 and n+2. Then, the maximum value is taken.