Language identification for creating national web archives

Tommi Jauhiainen tommi.jauhiainen@helsinki.fi
Department of Modern Languages
National Library of Finland
Heidi Jauhiainen heidi.jauhiainen@helsinki.fi
Department of Modern Languages
Petteri Veikkolainen petteri.veikkolainen@helsinki.fi
National Library of Finland

IIPC Web Archiving Conference (WAC) 2017

The Finno-Ugric Languages and The Internet

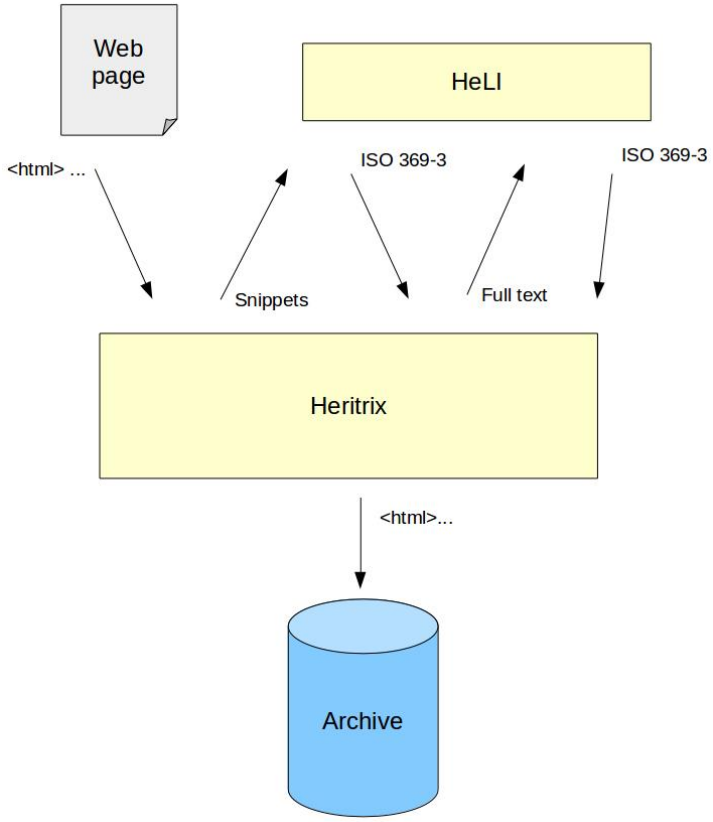
- Project started at the beginning of 2013 as part of the Kone Foundation Language Programme.
- It is situated in the University of Helsinki and is part of the international CLARIN cooperation.
- The main focus of the project is on gathering texts written in small Uralic languages from the internet using language identification methods developed within the project.
- **http://suki.ling.helsinki.fi/wanca/**
- Partly funded by the Finnish National Library, the project also targets to identify Finnish websites outside of the .fi domain.

Finding Finnish web pages

- Heritrix crawl on .ru, .ee, .se, and .no domains and extracted outlinks from .fi domain.
- On the fly identification based on three 100 character snippets.
- Analysis for the full page if even one of the snippets is identified to be Finnish.
- Once identified the page can be archived.

- All of the Common Crawl open repository of web crawl data from 2014 has also been processed. All pages identified as Finnish have been harvested from the internet and identified again.

Identifying and archiving Finnish web pages using HeLI and Heritrix



Top tier language identifier

$$R_{HeLI}(g, M) = \frac{\sum_{i=1}^{l_{T(M)}} v_g(t_i)}{l_{T(M)}}$$

- Helsinki language identification method (HeLI).
- Word-based method, which backs off to character n-grams if needed.
- Probabilistic, every word is considered equally important.
- A state of the art method developed within the project.
- Shared 1st place in the closed track of the Discriminating Between Similar Languages shared task in 2016.
- Support for identifying multilingual documents.

Written in Java

- Java application based on the method is available on GitHub.
- https://github.com/tosaja/HeLI
- Requires training data for the desired languages.
- Merely a command line tool to identify languages from a plain text file.
- Full language identifier service and the Heritrix (3.1) modifications will be released as open source before the end of the project (2018).

Try it!

- It works for your language as well!
- Even though the full version is not out yet, ask for it via email if you want to try it. (heidi.jauhiainen@helsinki.fi)

Thanks!

Tommi Jauhiainen, tommi.jauhiainen@helsinki.fi

Heidi Jauhiainen, heidi.jauhiainen@helsinki.fi

Petteri Veikkolainen, petteri.veikkolainen@helsinki.fi