Domain-specific Intelligent Search

Degree programme: BSc in Computer Science | Specialisation: Computer Perception and Virtual Reality
Thesis advisor: Prof. Dr. Bernhard Anrig, Prof. Dr. Erik Graf
Expert: Xavier Monnat (Post CH AG)

The Swiss web landscapes feature a great number of sources of information. This thesis aims to focus on an arbitrary domain and develop a search solution aimed at increasing the accessibility of knowledge via the combination of techniques from Information Retrieval (IR), Natural Language Processing (NLP), and Machine Learning (ML).

Introduction

The idea is to create a system that collects news articles through public channels of Swiss universities, extracts their content, identifies named-entities and finally indexes the results of the process in a search engine.

Data collection

Data is collected in the process of crawling, a web crawler is fine-tuned in order to optimally crawl relevant pages in efficient time. Crawling begins from root URLs called 'seeds' and then continues the navigation following a strategy to avoid irrelevant pages.

Article extraction and Named-Entity Recognition

We exploit Boilerpipe and Stanford CoreNLP, two very powerful tools, to apply robust article content extraction and Named-Entity Recognition to each HTML page crawled.
In particular, we are interested in locating universities, organisations and people.

Search

The search is backed up by Elasticsearch. It is famous for its high flexibility and scalability, aspects that are key for this solution.
All information extracted is mapped following the configuration of the index and is then indexed and stored within the search engine, becoming searchable.

Extract, Transform, Load

The tasks of data collection, article extraction, Named-Entity Recognition and indexing, need to be integrated into a single sequence of data (stream), also known as a pipeline. Apache Camel allows us exactly to realise this integration, implementing an Extract, Transform, Load (ETL) system.
The framework also allows us to develop an independent and autonomous system. The idea in fact, is to have the process automatically run itself based on a definable and arbitrary schedule.

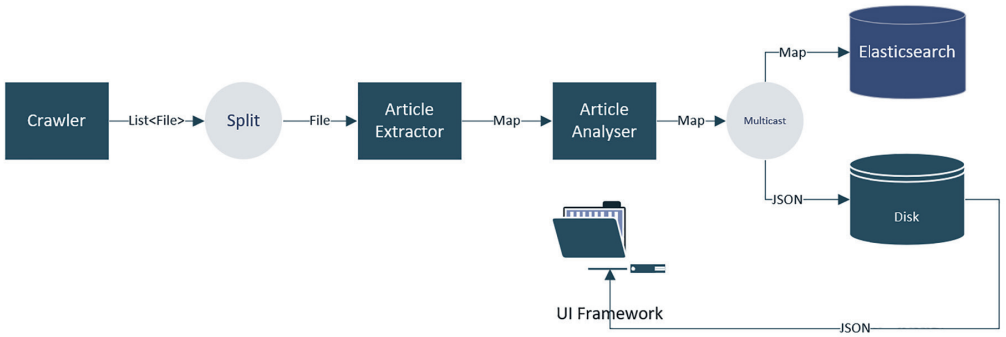
User Interface

The User Interface provides the possibility to search data with the traditional keyword-based search or with facets, that is filters based on the named-entities recognised.

Results

The final solution features an autonomous system that can be arbitrarily scheduled to collect news web pages, extract the content of the articles, identify and locate named-entities and incrementally index the results to augment the searchable data. All data is searchable via the User Interface.

Nicolas Di Vittorio
nrdivittorio@gmail.com



Pipeline design in Apache Camel