Framework for Extraction of Wikipedia Articles Content

Diplomová práce

Oleksandr Husiev
Vedoucí práce: Ing. Milan Dojčinovski, Ph. D.

Fakulta informačních technologií
České vysoké učení technické v Praze

07. 12. 2021

1. Motivation
2. Thesis goals
3. Requirements
4. Implementation
5. Tools and Libraries
6. Dynamic Language Support
7. Testing
8. End-to-End Testing - Output validation
9. End-to-End Testing - Scale Testing
10. End-to-End Testing - Language Support

Motivation

- Knowledge bases are growing up in importance as a Web and enterprise search engine.
- Currently knowledge bases cover only specific niches and are not useful outside of their primary purpose.
- Formatted article data from the Wikipedia in DBPedia is not updated on a regular basis

Thesis goals

DBpedia is a crowd-sourced community effort that aims at extraction of information from Wikipedia. The main goal of the thesis is to develop a framework for extraction of Wikipedia articles content , structure and annotations which can be further split into those subgoals:

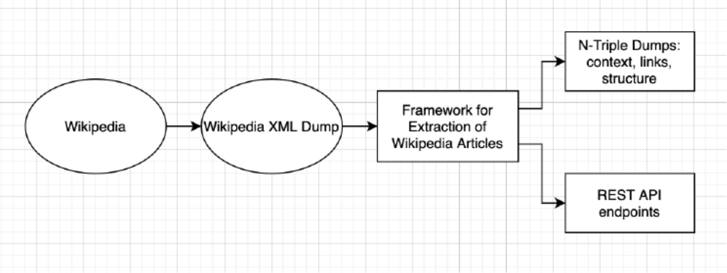
- Accept and process input data in the form of Wikipedia XML dumps
- Extract Wikipedia context, page structure, and links.
- Provide outputs for context, links and page structure in the form of N-Triples.
- Implement language extensibility.
- Provide a user interface to interact with the framework.

Requirements to the Framework

- Accept input data - support XML dump with up to 20GB of data;
- Language extensibility - parse XML in English and at least 4 other languages of choice;
- Provide outputs - print all the outputs in the NIF triples, concatenating processed data from all articles in a single XML input file and writing the data to .nt output file.

Data Workflow

General Extraction Framework data workflow



Wikipedia

Wikipedia XML Dump

Framework for Extraction of Wikipedia Articles

N-Triple Dumps: context, links, structure

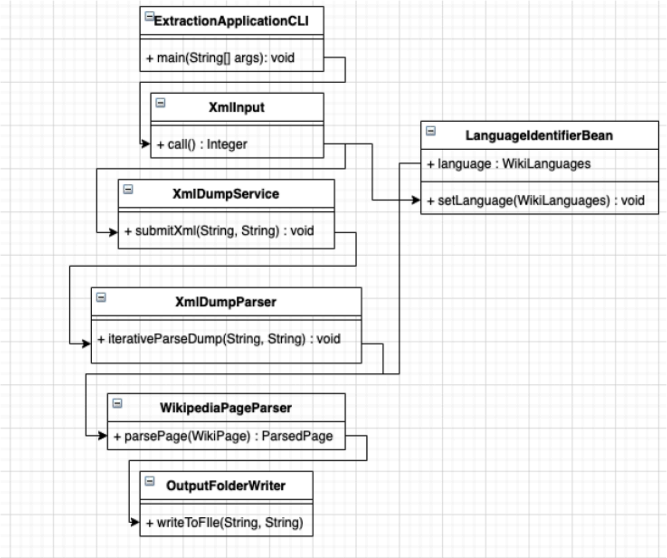
REST API endpoints

Implementation

Project Architecture

- Interface Layer - used for CLI and REST interface
- Logic Layer - used for XML processing of the incoming data
- Output Layer - used for tuning parameters for file output.

Framework Class Diagram



ExtractionApplicationCLI

+ main(String[] args): void

XmlInput

+ call() : Integer

XmlDumpService

+ submitXml(String, String) : void

XmlDumpParser

+ iterativeParseDump(String, String) : void

WikipediaPageParser

+ parsePage(WikiPage) : ParsedPage

OutputFolderWriter

+ writeToFIle(String, String)

LanguageIdentifierBean

+ language : WikiLanguages

+ setLanguage(WikiLanguages) : void

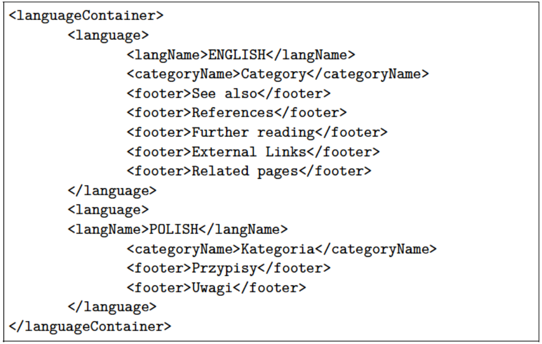
Tools and Libraries

- Java(with Spring Boot and Spring Dependency Injection)
- Java Jackson XML Library for XML parsing
- picocli for a command-line interface
- JUnit for unit testing of the framework

Dynamic Language Support

Support

Before the framework starts, it processes the configuration file language list.xml that is stored in the configuration folder and instantiates objects during the runtime:



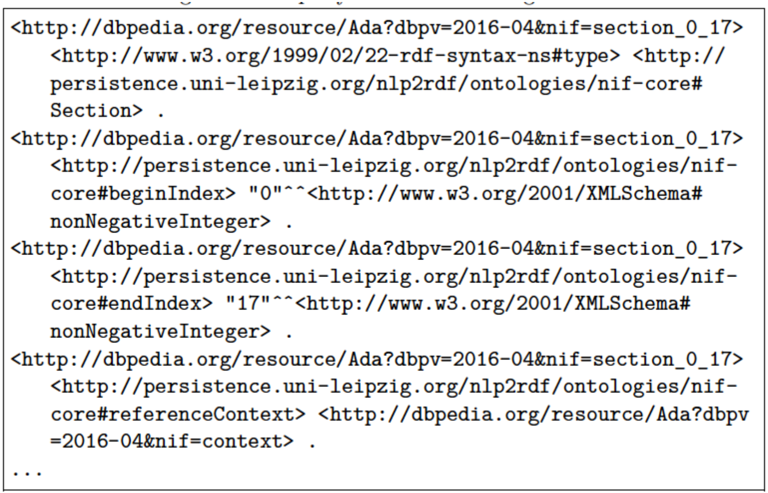
```
<languageContainer>
        <language>
                <langName>ENGLISH</langName>
                <categoryName>Category</categoryName>
                <footer>See also</footer>
                <footer>References</footer>
                <footer>Further reading</footer>
                <footer>External Links</footer>
                <footer>Related pages</footer>
        </language>
        <language>
        <langName>POLISH</langName>
                <categoryName>Kategoria</categoryName>
                <footer>Przypisy</footer>
                <footer>Uwagi</footer>
        </language>
</languageContainer>
```

Testing

The framework has been tested during the implementation, using several methodologies

- Unit Testing - to provide test coverage for classes in an isolated environment;
- End-to-End Testing: output validation;
- End-to-End Testing: scale testing.

Processed Page Structure Output



```
<http://dbpedia.org/resource/Ada?dbpv=2016-04&nif=section_0_17>
    <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://
    persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#
    Section> .
<http://dbpedia.org/resource/Ada?dbpv=2016-04&nif=section_0_17>
    <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-
    core#beginIndex> "0"^^<http://www.w3.org/2001/XMLSchema#
    nonNegativeInteger> .
<http://dbpedia.org/resource/Ada?dbpv=2016-04&nif=section_0_17>
    <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-
    core#endIndex> "17"^^<http://www.w3.org/2001/XMLSchema#
    nonNegativeInteger> .
<http://dbpedia.org/resource/Ada?dbpv=2016-04&nif=section_0_17>
    <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-
    core#referenceContext> <http://dbpedia.org/resource/Ada?dbpv
    =2016-04&nif=context> .
...
```

End-to-End Testing - Scale Testing

- Total pages parsed: 258. Success rate: 84.88%. Seconds passed: 31(rate of 8 articles per second).
- Total pages parsed: 2087. Success rate: 88.55%. Seconds passed: 109(19 articles per second).
- Total pages parsed: 6738. Success rate: 87.90%. Seconds passed: 237(28 articles per second).

End-to-End Testing - Language Support

After implementing the dynamic language support, I have picked the most popular languages on Wikipedia:

- English: 2,567,509 articles, 22.5% of the total number of articles;
- German: 808,044 articles, 7.1%;
- French: 709,312 articles, 6.2%;
- Polish: 539,688 articles, 4.7%;
- Japanese: 523,629 articles, 4.6%.

Conclusions

- Accept and process input data in the form of Wikipedia XML dumps. The Wikipedia XML Dump parsing was achieved. The statistics show that the parsing success rate averages on 88% over the large amounts of articles.
- Extract context. Context is extracted and stored in the form of NTriples. Some of the contexts might still contain traces of the original XML code.
- Extract page structure. Page structure is extracted and recursively built in the form of N-Triples.
- Extract links. Links are extracted, URLs that link them to the page structure are created.
- Provide outputs for context, links and page structure in the form of N-Triples. Output is printed.
- Implement language extensibility. Language extensibility mechanism is implemented, new languages can be added in the form of an XML configuration that is parsed when the application is starting
- Provide a user interface. User interface is provided in two different forms

The end

Thanks for your attention!

Děkuji za pozornost!

Otázky oponenta

Otázka první: Proč?

Odpověď: Prostě proto.

Otázky oponenta

Otázka druhá: Proč?

Odpověď: Prostě proto.