invenio-classifier Documentation

Release 0.1.0

CERN

August 19, 2015

Contents

1 Features 3

2 Keyword extraction is simple 5

3 Thesaurus 7

4 Keyword extraction 9

5 Contents: 11

6 Indices and tables 13

Invenio module for record classification.

This is an experimental development preview release.

- Free software: GPLv2 license
- Documentation: https://invenio-classifier.readthedocs.org.

CHAPTER 1

Features

Classifier automatically extracts keywords from fulltext documents. The automatic assignment of keywords to textual documents has clear benefits in the digital library environment as it aids catalogization, classification and retrieval of documents.

CHAPTER 2

Keyword extraction is simple

Dependencies. Classifier requires Python RDFLib in order to process the RDF/SKOS taxonomy.

In order to extract relevant keywords from a document `fulltext.pdf` based on a controlled vocabulary `thesaurus.rdf`, you would run Classifier as follows:

```
$ inveniomanage classifier extract -k thesaurus.rdf -f fulltext.pdf
```

Launching `inveniomanage classifier --help` shows the options available.

As an example, running classifier on document nucl-th/0204033 using the high-energy physics RDF/SKOS taxonomy (`HEP.rdf`) would yield the following results (based on the HEP taxonomy from October 10th 2008):

```
Input file: 0204033.pdf

Author keywords:
Dense matter
Saturation
Unstable nuclei

Composite keywords:
10  nucleus: stability [36, 14]
6  saturation: density [25, 31]
6  energy: symmetry [35, 11]
4  nucleon: density [13, 31]
3  energy: Coulomb [35, 3]
2  energy: density [35, 31]
2  nuclear matter: asymmetry [21, 2]
1  n: matter [54, 36]
1  n: density [54, 31]
1  n: mass [54, 16]

Single keywords:
61  K0
23  equation of state
12  slope
4  mass number
4  nuclide
3  nuclear model
3  mass formula
2  charge distribution
2  elastic scattering
2  binding energy
```

CHAPTER 3

Thesaurus

Classifier performs an extraction of keywords based on the recurrence of specific terms, taken from a controlled vocabulary. A controlled vocabulary is a thesaurus of all the terms that are relevant in a specific context. When a context is defined by a discipline or branch of knowledge then the vocabulary is said to be a *subject thesaurus*. Various existing subject thesauri can be found here.

A subject thesaurus can be expressed in several different formats. Different institutions/disciplines have developed different ways of representing their vocabulary systems. The taxonomy used by classifier is expressed in RDF/SKOS. It allows not only to list keywords but to specify relations between the keywords and alternative ways to represent the same keyword.

```
<Concept rdf:about="http://cern.ch/thesauri/HEP.rdf#scalar">
 <composite rdf:resource="http://cern.ch/thesauri/HEP.rdf#Composite.fieldtheoryscalar"/>
 <prefLabel xml:lang="en">scalar</prefLabel>
 <note xml:lang="en">nostandalone</note>
</Concept>

<Concept rdf:about="http://cern.ch/thesauri/HEP.rdf#fieldtheory">
 <composite rdf:resource="http://cern.ch/thesauri/HEP.rdf#Composite.fieldtheoryscalar"/>
 <prefLabel xml:lang="en">field theory</prefLabel>
 <altLabel xml:lang="en">QFT</altLabel>
 <hiddenLabel xml:lang="en">/field theor\w*/</hiddenLabel>
 <note xml:lang="en">nostandalone</note>
</Concept>

<Concept rdf:about="http://cern.ch/thesauri/HEP.rdf#Composite.fieldtheoryscalar">
 <compositeOf rdf:resource="http://cern.ch/thesauri/HEP.rdf#scalar"/>
 <compositeOf rdf:resource="http://cern.ch/thesauri/HEP.rdf#fieldtheory"/>
 <prefLabel xml:lang="en">field theory: scalar</prefLabel>
 <altLabel xml:lang="en">scalar field</altLabel>
</Concept>
```

In RDF/SKOS, every keyword is wrapped around a *concept* which encapsulates the full semantics and hierarchical status of a term - including synonyms, alternative forms, broader concepts, notes and so on - rather than just a plain keyword.

The specification of the SKOS language and various manuals that aid the building of a semantic thesaurus can be found at the SKOS W3C website. Furthermore, Classifier can function on top of an extended version of SKOS, which includes special elements such as key chains, composite keywords and special annotations.

CHAPTER 4

Keyword extraction

Classifier computes the keywords of a fulltext document based on the frequency of thesaurus terms in it. In other words, it calculates how many times a thesaurus keyword (and its alternative and hidden labels, defined in the taxonomy) appears in a text and it ranks the results. Unlike other similar systems, Classifier does not use any machine learning or AI methodologies - a just plain phrase matching using regular expressions: it exploits the conformation and richness of the thesaurus to produce accurate results. It is then clear that Classifier performs best on top of rich, well-structured, subject thesauri expressed in the RDF/SKOS language.

Happy hacking and thanks for flying Invenio-Classifier.

Invenio Development Team

Email: info@invenio-software.org

IRC: #invenio on irc.freenode.net

Twitter: http://twitter.com/inveniosoftware

GitHub: https://github.com/inveniosoftware/invenio-classifier

URL: http://invenio-software.org

CHAPTER 5

Contents:

CHAPTER 6

Indices and tables

- genindex
- modindex
- search