Annif
leveraging bibliographic metadata for automated subject indexing and classification

Osma Suominen

SWIB18, Bonn, 28 November 2018
Extrahlad till ÅBO UNDERRÄTTSELER

Finlands oavhängighet.

Nationens rökningar rörande saken att Finlands oavhängighet ska vidare förlängas och att Finlands stora pinse pollineras på ett annat område än landets östra delar. Stora framsteg med tillkomsten av 1800-talets sista år till den alltmer oavhän- get från de nuvarande europeiska störkrafterna.

[Image of a map of the Baltic region]
Extrablad till ÅBO UNDERRÄTTELSER

Bö 2.

Finlands oavhängighet.

Landet, sedan dess regeringen nått ett oavhängigt ställning, och med det, att det förbättringar i regeringens program för framställning av landets nu tillstånd. Sedan extrablad med 300 eljest skickas till det, att detta är smädelkännande förändring. Sträv.
Idea of Annif
We have a lot of LAM metadata, e.g. 15M records in Finna.fi discovery service
Machine learning using library data

Finna.fi metadata
15M titles + subjects

Fulltext docs
Annif prototype (2017)
# Annif prototype vs. new Annif

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>architecture</strong></td>
<td>loose collection of scripts</td>
<td>Flask web application</td>
</tr>
<tr>
<td><strong>coding style</strong></td>
<td>quick and dirty</td>
<td>solid software engineering</td>
</tr>
<tr>
<td><strong>backends</strong></td>
<td>Elasticsearch index</td>
<td>TF-IDF, fastText, Maui ...</td>
</tr>
<tr>
<td><strong>language support</strong></td>
<td>Finnish, Swedish, English</td>
<td>any language supported by NLTK</td>
</tr>
<tr>
<td><strong>vocabulary support</strong></td>
<td>YSO, GACS ...</td>
<td>YSO, YKL, others coming</td>
</tr>
<tr>
<td><strong>REST API</strong></td>
<td>minimal</td>
<td>extended (e.g. list projects)</td>
</tr>
<tr>
<td><strong>user interface</strong></td>
<td>web form for testing</td>
<td><a href="http://dev.annif.org">http://dev.annif.org</a></td>
</tr>
<tr>
<td><strong>mobile app</strong></td>
<td>HTML/CSS/JS based</td>
<td>native Android app</td>
</tr>
<tr>
<td><strong>open source license</strong></td>
<td>CC0</td>
<td>Apache License 2.0</td>
</tr>
</tbody>
</table>
Algorithms for automated subject indexing
Lexical vs. Associative approaches for subject indexing

Lexical approaches

Match the terms in a document to terms in a controlled vocabulary.

“Renewable resources are a part of Earth's natural environment and the largest components of its ecosphere.“

Associative approaches

Learn which concepts are correlated with which terms in documents, based on training data.

For more information, see:
Algorithms used in Annif

Statistical / Associative

- TF-IDF similarity
  Baseline bag-of-words similarity measure. Implemented with the `Gensim` library.

- **fastText** by Facebook Research
  Machine learning algorithm for text classification.
  Uses word embeddings (similar to `word2vec`) and resembles a neural network architecture.
  Promises to be good for e.g. library classifications (DDC, UDC, YKL…)

Lexical

- **Maui** using MauiService REST API
  MauiService is a microservice wrapper around the Maui automated indexing tool.
  Based on traditional Natural Language Processing techniques - finds terms within text.
Algorithms may be used *alone*, or in combinations, *ensembles*
Algorithms make silly mistakes

Some reasons for mistakes:

- errors and skew in training data
- correlation $\neq$ causation
- homonyms (e.g. rock)
- misinterpreted names (e.g. Smith, AIDS)
- random noise
In an ensemble, each algorithm makes different mistakes

Solution: If we have some more training documents, we can perform second order learning!

Isotonic regression, implemented using the Pool Adjacent Violators (PAV) algorithm, is a good way of assessing trustworthiness of individual algorithms and turning raw scores into final probability estimates.


Anrif Fusion experiment demonstrates PAV
Evaluation of algorithms
Test corpora for evaluating algorithms

Full text documents indexed with YSA/YSO for training and evaluation

1. **Arto**: Articles from Arto database (n=6287)
   Both scientific research papers and less formal publications. Many disciplines.

2. **JYU theses**: Master’s and Doctoral theses from University of Jyväskylä (n=7400)
   Long, in-depth scientific documents. Many disciplines.

3. **AskLib**: Question/Answer pairs from an Ask a Librarian service (n=3150)
   Short, informal questions and answers about many different topics.

4. **Satakunnan Kansa**: Digital archives of Satakunnan Kansa regional newspaper.
   Over 100k documents, of which 50 have been indexed independently by 4 librarians.

Corpora 1-3 available on GitHub: [https://github.com/NatLibFi/Annif-corpora](https://github.com/NatLibFi/Annif-corpora)
(for 1-2, only links to PDFs are provided for copyright reasons)
Evaluation of different algorithms in Annif
F1 scores (combination of precision & recall) against gold standard subjects

Observations:
1. Of individual algorithms, Maui is the best
2. Ensembles beat individual algorithms
3. PAV ensembles can be better than a simple ensemble (but not always)
Software architecture
**Annif Architecture**

- **Any metadata / document management system**
- **Mobile app**
- **Flask/Connexion web app**
- **REST API**
- **CLI**
- **Fusion module**
  - TF-IDF model
  - fastText model
  - HTTP backend
- **MauiService**
  - Microservice around Maui
- **Finna.fi metadata**
- **Fulltext docs**

Diagram:
- Annif
  - admin
  - CLI
  - REST API
- Any metadata / document management system
- Mobile app
- OCR
- Any metadata / document management system

Train data can be added in future, e.g. neural network, fastXML, StarSpace

---

**Key Components**

- **Annif**
  - Fusion module
  - Flask/Connexion web app
  - REST API

- **MauiService**
  - Microservice around Maui

- **Finna.fi metadata**
- **Fulltext docs**

Any metadata / document management system can be added in future.
Try Annif!

Text to analyze:

SWIB conference (Semantic Web in Libraries) is an annual conference, being held for the 10th time, focusing on Linked Open Data (LOD) in libraries and related organizations. It is well established as an event where IT staff, developers, librarians, and researchers from all over the world meet and mingle and learn from each other. The topics of talks and workshops at SWIB revolve around opening data, linking data and creating tools and software for LOD production scenarios. These areas of focus are supplemented by presentations of research projects in applied sciences, industry applications and LOD activities in other areas.

Project (vocabulary and language):

YSO ensemble English

Results

- open data
- semantic web
- linked open data
- semantics
- data
- applied sciences
- librarians
- data science
- libraries
- Works
Form for testing at annif.org

Try Annif!

Text to analyze:

SWIB conference (Semantic Web in Libraries) is an annual conference, being held for the 10th time, focusing on Linked Open Data (LOD) in libraries and related organizations. It is well established as an event where IT staff, developers, librarians, and researchers from all over the world meet and mingle and learn from each other. The topics of talks and workshops at SWIB revolve around opening data, linking data and creating tools and software for LOD production scenarios. These areas of focus are supplemented by presentations of research projects in applied sciences, industry applications and LOD activities in other areas.

Results

Lod
big data
librarian
library science
data mining
data
Semantic Web
library
Library of Congress
data structure
Command line interface

Load a vocabulary to be used by one or more models:
$ annif loadvoc tfidf-en yso-en.tsv

Train a model:
$ annif train tfidf-en yso-finna-en.tsv.gz

Analyze a document:
$ annif analyze tfidf-en <berries.txt
<http://www.yso.fi/onto/yso/p772> strawberry 0.39644203283656165
<http://www.yso.fi/onto/yso/p18109> wild strawberry 0.37539359094384245
<http://www.yso.fi/onto/yso/p25548> stolons 0.3261554545369906
<http://www.yso.fi/onto/yso/p6749> berry cultivation 0.2394291077460799
<http://www.yso.fi/onto/yso/p10631> questionnaire survey 0.22714475653823335
<http://www.yso.fi/onto/yso/p6821> farms 0.2172524306795587
<http://www.yso.fi/onto/yso/p3294> customers 0.216395821347059
<http://www.yso.fi/onto/yso/p1834> work motivation 0.21612376226244975
<http://www.yso.fi/onto/yso/p8531> customership 0.21536113638508098
<http://www.yso.fi/onto/yso/p19047> corporate clients 0.21412270159920782

Evaluate a model using several measures (e.g. recall, precision, F1 score, NDCG):
$ annif eval tfidf-en directory-with-gold-standard-docs/
REST API access example

“The quick brown fox jumped over the lazy dog.”

Analyze this!

```json
results=[
  {uri="<http://www.yso.fi/onto/yso/p2228>"", score=0.2595, label="red fox"},
  {uri="<http://www.yso.fi/onto/yso/p5319>"", score=0.2039, label="dog"},
  {uri="<http://www.yso.fi/onto/yso/p8122>"", score=0.1946, label="laziness"},
  {uri="<http://www.yso.fi/onto/yso/p25726>"", score=0.1285, label="brown"},
  {uri="<http://www.yso.fi/onto/yso/p4760>"", score=0.1220, label="triple jump"}
]
```
What can you do with Annif?
JYX repository, University of Jyväskylä

Students upload their Master’s and doctoral theses, Annif suggests subjects

## Keywords

<table>
<thead>
<tr>
<th>Keyword suggestions</th>
<th>□ information management systems [YSO]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>□ metadata [YSO]</td>
</tr>
<tr>
<td></td>
<td>□ connections (technical systems) [YSO]</td>
</tr>
<tr>
<td></td>
<td>□ content management [YSO]</td>
</tr>
<tr>
<td></td>
<td>□ multimedia (information technology) [YSO]</td>
</tr>
<tr>
<td></td>
<td>□ digital libraries [YSO]</td>
</tr>
<tr>
<td></td>
<td>□ XML [YSO]</td>
</tr>
<tr>
<td></td>
<td>□ semantic web [YSO]</td>
</tr>
<tr>
<td></td>
<td>□ open source code [YSO]</td>
</tr>
<tr>
<td></td>
<td>□ open data [YSO]</td>
</tr>
<tr>
<td></td>
<td>□ user-centeredness [YSO]</td>
</tr>
<tr>
<td></td>
<td>□ archives (memory organisations) [YSO]</td>
</tr>
<tr>
<td></td>
<td>□ seeking [YSO]</td>
</tr>
<tr>
<td></td>
<td>□ Works [YSO]</td>
</tr>
<tr>
<td></td>
<td>□ cloud services [YSO]</td>
</tr>
<tr>
<td></td>
<td>□ electronic publications [YSO]</td>
</tr>
</tbody>
</table>

**Implemented using DSpace & GLAMpipe**

by Ari Häyrinen

---

**Your own keywords**

- keyword 1, keyword 2
Indexing Wikipedia by topics

Finnish Wikipedia has 410 000 articles (620 MB as raw text)
Automated subject indexing took 7 hours on a laptop, using the Annif prototype
1-3 topics per article (average ~2)
Indexing Wikipedia by topics

Finnish Wikipedia has 410 000 articles (620 MB as raw text)
Automated subject indexing took 7 hours on a laptop
1-3 topics per article (average ~2)

Examples: (random sample)

<table>
<thead>
<tr>
<th>Wikipedia article</th>
<th>YSO topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ahvenuslammi (Urjala)</td>
<td>shores</td>
</tr>
<tr>
<td>Brasilian Grand Prix 2016</td>
<td>race drivers, formula racing, karting</td>
</tr>
<tr>
<td>Guy Topelius</td>
<td>folk poetry researcher, saccharin</td>
</tr>
<tr>
<td>HMS Laforey</td>
<td>warships</td>
</tr>
<tr>
<td>Liigacup</td>
<td>football, football players</td>
</tr>
<tr>
<td>Pää Kii</td>
<td>ensembles (groups), pop music</td>
</tr>
<tr>
<td>RT-21M Pioneer</td>
<td>missiles</td>
</tr>
<tr>
<td>Runoja</td>
<td>pop music, recording (music recordings), compositions (music)</td>
</tr>
<tr>
<td>Sjur Røthe</td>
<td>skiers, skiing, Nordic combined</td>
</tr>
<tr>
<td>Veikko Lavi</td>
<td>lyricists, comic songs</td>
</tr>
</tbody>
</table>
Most common topics in Finnish Wikipedia

- football players
- ice hockey players
- sports teams
- warships
- football
- pop music
- ensembles (groups)
- recording (music recording)
- Olympics
- Finnish championships
- heavy rock
- solemnising a marriage
- matches (sports contests)
- Nordic combined
- summer Olympics
- track and field athletes
- formula racing
- goalkeepers
- film directors
- ice hockey
Most common topics in Finnish Wikipedia

football players
ice hockey players
sports teams
warships
football
pop music
ensembles (groups)
recording (music recording)
Olympics
Finnish championships
heavy rock
solemnising a marriage
matches (sports contests)
Nordic combined
summer Olympics
track and field athletes
formula racing
goalkeepers
film directors
ice hockey

Image credits:
Petteri Lehtonen [CC BY-SA 3.0]
Hockeybroad/Cheryl Adams [CC BY-SA 3.0]
Tomisti [CC BY-SA 3.0]
Tuomas Vitikainen [CC BY-SA 3.0]
Prototype web app, ocr.space cloud OCR
m.annif.org

Prototype Android app with OCR on the device
(by Okko Vainonen)
Finna Recommends Chrome browser extension

Finna Recommends Chrome browser extension

Analyze selected text from any web page using Annif API and recommends books from Finna.fi

Created during WIDE hackathon by Yazan Alhalabi Samuel Akangbe Steven Nebo

Thick-billed parrot

From Wikipedia, the free encyclopedia

This page is about the species of parrot. For the genus of parrots, see Rhynchopsitta.

The thick-billed parrot (Rhynchopsitta pachyrhyncha) is a medium-sized green and red parrot found in Mexico, that formerly ranged into the southwestern United States. Its position in parrot phylogeny is the subject of ongoing discussion; it is sometimes referred to as thick-billed macaw or thick-billed conure. In Mexico, it is locally called guacamaya ("macaw") or cotorra serrana ("mountain parrot"). Classified internationally as Endangered through IUCN,[1] the thick-billed parrot's decline has been central to multiple controversies over wildlife management.

Contents [hide]

1 Taxonomy
2 Description
Getting Annif
Annif on GitHub

Python 3.5+ code base
Apache License 2.0

Fully unit tested (98% coverage)
PEP8 style guide compliant
Usage documentation in the wiki

https://github.com/NatLibFi/Annif
Annif on PyPI

Installing into a virtualenv:

```
pip install annif
```

https://pypi.org/project/annif/
Apply Annif on your own data!

- Choose an indexing vocabulary
- Prepare a corpus from your existing metadata
- Load the corpus into Annif
- Use it to index new documents
Community group on DIY automated subject indexing?

To discuss applications, algorithms, API standards, corpora, formats etc.

Contact me if interested!
Thank you!
Questions?

osma.suominen@helsinki.fi  - @OsmaSuominen

Website: http://annif.org
API: http://api.annif.org

These slides: https://tinyurl.com/annif-swib