

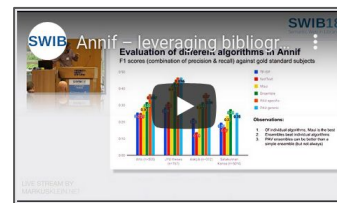
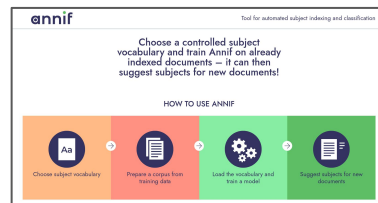
Introduction to the online hands-on tutorial



1

Understand what Annif is

Study the website annif.org,
watch a presentation about it,
or read the LIBER Quarterly [paper](#).



Annif: DIY automated subject indexing using multiple algorithms

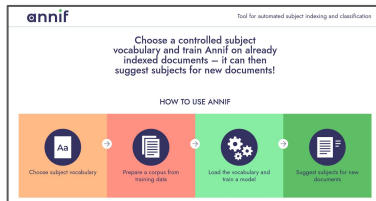
Abstract

Manually indexing documents for subject-based access is a labor-intensive process. We propose using metadata gathered from bibliographic databases to train algorithms that assist librarians in their work. We have developed Annif, an open-source tool and ecosystem for automated subject indexing. After training it with a subject vocabulary and existing metadata, Annif can be used to suggest subject headings for new documents. We have tested Annif with different document collections including scientific papers, old scanned books and contemporary e-books. Old papers from an "early discovery" service, Project Gutenberg, and the archives of a local newspaper. The results of analyzing scientific papers and current books have been promising, while other types of documents have proved to be more challenging. The current version is based on a combination of existing natural language processing and machine learning tools. By combining multiple approaches and relying on open source algorithms, Annif can build on the strengths of individual algorithms and adapt to different settings. With built-in support for regular subject indexing and classification processes especially for electronic documents as well as collections that otherwise would not be indexed at all.

1

Understand what Annif is

Study the website annif.org, watch a presentation about it, or read the LIBER Quarterly [paper](#).



Annif: DIY automated subject indexing using multiple algorithms

Abstract

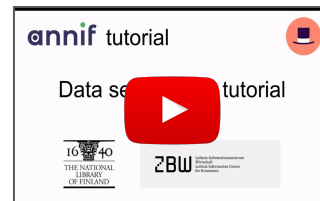
Manually indexing documents for subject-based access is a labor-intensive process. We propose using metadata gathered from bibliographic databases to train algorithms that assist librarians in their work. We have developed *annif*, an open-source tool and ecosystem for automated subject indexing. After making it with a subject vocabulary and training materials, *annif* can be used to suggest subject headings for new documents. We have tested *annif* with different document collections including scientific papers, old newspaper books and contemporary e-books. Our paper focuses on "old newspaper" service. From Wikipedia, and the archives of a local newspaper. The results of analyzing scientific papers and current books have been promising, while other types of documents have proved to be more challenging. The current version is based on a combination of existing natural language processing and machine learning tools. By combining multiple approaches and relying on open source algorithms, *annif* can build on the strengths of individual algorithms and adapt to different settings. With *annif*, we expect to improve subject indexing and classification processes especially for electronic documents as well as collections that otherwise would not be indexed at all.

2

Complete this hands-on tutorial

Watch the videos, install Annif, and complete the exercises as far as you can, on your own time.

you
are
here



Exercise 2: Set up and train a TFIDF project

Annif requires you to set up one or more **projects** before you can use it. A project is a set of configuration settings and (usually) some data files, such as a trained machine learning model. A project is identified by a **project-id**, which is typically a short string such as "you-tilde".

Projects are defined in a configuration file called **projects.cfg**. *Annif* looks for a **projects.cfg** file in the current working directory unless you tell it otherwise using either the `-c` command line option or the `ANNIF_PROJECTS` environment variable. For the rest of this tutorial, we will be using a working directory.

In this lesson, we will set up the simplest kind of *Annif* project, which uses a TFIDF model that needs to be trained on example documents (and/or metadata records before it can be used).

You need to choose which data set you want to use: either the **year100** data set from the National Library of Finland or the **low-dow** data set from ZBW. Either one will work, but you need to do something.

1. Create a projects.cfg file

Use a text editor to create a **projects.cfg** file within the *Annif* tutorial directory.

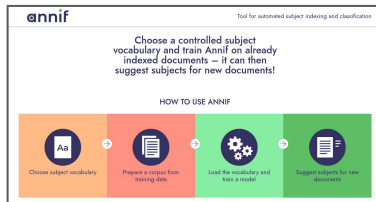
If you use the **year100** data set, use the following contents:

```
[year100-tilde]
name=year100-tilde
language=fi
```

1

Understand what Annif is

Study the website annif.org, watch a presentation about it, or read the LIBER Quarterly [paper](#).



Annif: DIY automated subject indexing using multiple algorithms

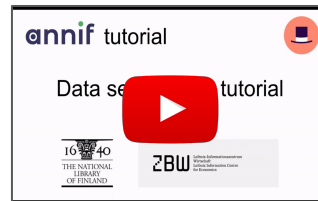
Abstract

Manually indexing documents for subject-based access is a labor-intensive process. We propose using metadata gathered from bibliographic databases to train algorithms that automate this task. We have developed Annif, an open-source tool and ecosystem for automated subject indexing. After making it with a subject vocabulary and training corpora, Annif can be used to suggest subject headings for new documents. We have tested Annif with different document collections including scientific papers, old newspaper books and contemporary e-books. Our paper-based 'old newspaper' results, Francis & Taylor, and the online of a local newspaper. The results of analyzing internet papers and current books have been promising, adding other types of documents have proved to be more challenging. The current version is based on a combination of existing natural language processing and machine learning tools. By extending multiple algorithms and adding open source algorithms, Annif can build on the strengths of individual algorithms and adapt to different settings. With Annif, we expect to improve subject indexing and classification processes especially for electronic documents as well as collections that otherwise would not be indexed at all.

2

Complete this hands-on tutorial

Watch the videos, install Annif, and complete the exercises as far as you can, on your own time.



Exercise 2: Set up and train a TFIDF project

Annif requires you to set up one or more projects before you can use it. A project is a set of configuration settings and (usually) some data files. Even an internal machine learning model. A project is identified by a unique id, which is typically a url string such as 'you:tfidf'.

Projects are defined in a configuration file called `projects.cfg`. Annif looks for a `projects.cfg` file in the current working directory (unless you tell it otherwise using either the `-c` command line option or the `ANNIF_PROJECTS` environment variable). For the rest of this tutorial, we will be using a working directory.

In this lesson, we will set up the simplest kind of Annif project, which uses a TFIDF model that needs to be trained on example documents (and/or metadata records before it can be used).

You need to choose which data set you want to use: either the `project` data set from the National Library of Finland or the `low-dim` data set from ZBW. Either one will work, but you need to do something.

1. Create a `projects.cfg` file

Use a text editor to create a `projects.cfg` file within the Annif tutorial directory.

If you use the 'you:tfidf' data set, use the following contents:

```
[you:tfidf-vec]
name=you:tfidf-vec
language=...
```

you
are
here

3

Join an online session (optional)

In the online sessions, you can ask questions, get help and discuss what you've learned. Registration required.



Annif-tutorial GitHub repository

the main resource for this tutorial

NatLibFi / **Annif-tutorial**

Unwatch ▾

10

Star

6

Fork

2

<> Code

Issues

Pull requests

Actions

Projects

Wiki

Security

Insights

Settings

master ▾

3 branches

0 tags

Go to file

Add file ▾

Code ▾



osma minor wordsmithing

586b131 2 days ago

189 commits

data-sets	Delete metadata for 2 fulltext docs that have been removed from Ec...	2 months ago
exercises	minor wordsmithing	2 days ago
presentations	Add a slide about Omikujii; other updates	6 months ago
.dockerignore	Move Dockerfile to master branch	6 months ago
.gitignore	Ignore mauidata directory	10 months ago
Dockerfile	Build tutorial image based on Annif 0.49; add "tutorial" image tag	21 days ago
LICENSE.txt	Add CC By 4.0 International license	10 months ago
README.md	Add instructions for cloning/downloading the repo	9 months ago

About



Instructions, exercises and example data sets for Annif hands-on tutorial

Readme

CC-BY-4.0 License

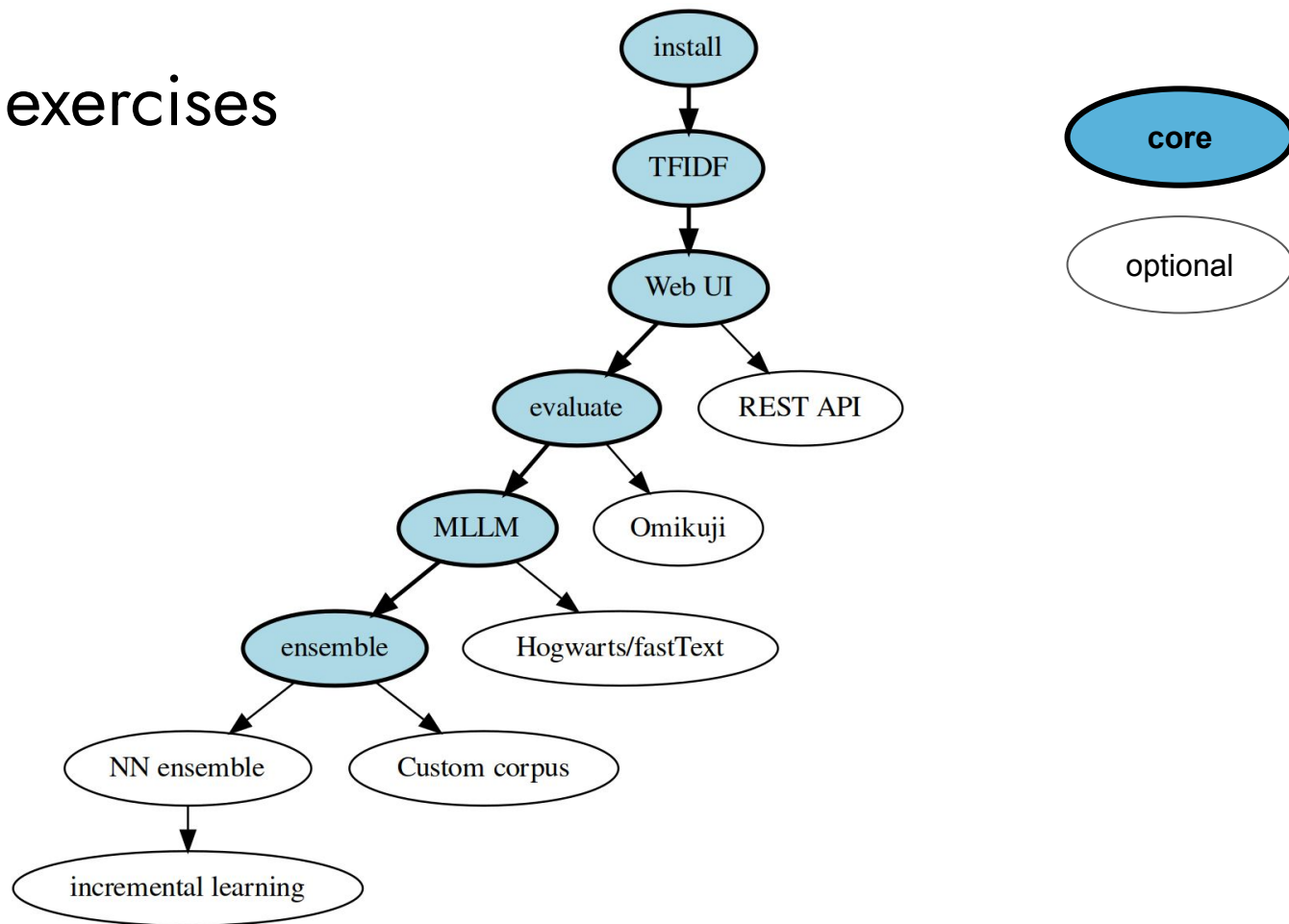
Releases

No releases published
[Create a new release](#)

Packages

No packages published
[Publish your first package](#)

Overview of exercises



Annif installation types

VirtualBox install

Recommended for most people, as it's the easiest way of getting Annif running so you can work on the tutorial exercises.

Need to install VirtualBox software - available for Windows, macOS and Linux

Docker install

If you know Docker, this is a good way of getting Annif set up, with all the dependencies included in a pre-built container.

Need to install Docker software - available for Windows, macOS and Linux

Linux local install

If you're an experienced Linux user and used to working with Python packages, a local install allows maximum flexibility.

Needs Python 3.6, 3.7 or 3.8 and support for virtual environments.

Accessing Annif

Command line interface

- setup and administration
- training models
- testing and evaluating models
- bulk indexing of documents

Web user interface

- interactive testing of models

REST API

- integrating Annif services to other systems

Apply Annif on your own data!



Choose subject vocabulary



Prepare a corpus from
training data



Load the vocabulary and
train a model



Suggest subjects for new
documents