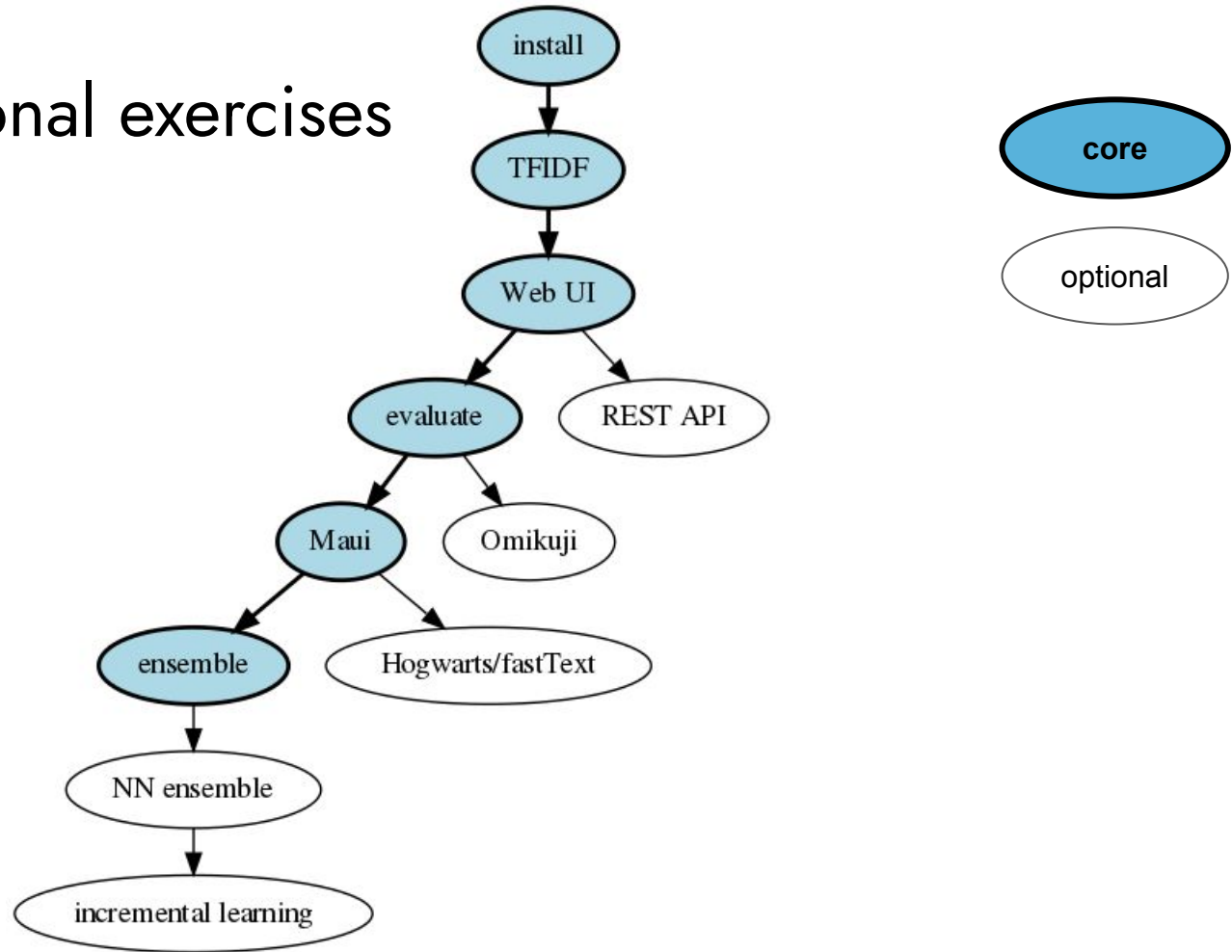


Closing the tutorial



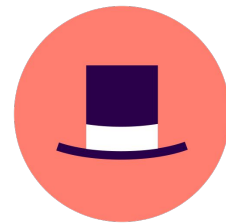
Leibniz-Informationszentrum
Wirtschaft
Leibniz Information Centre
for Economics

Core and optional exercises

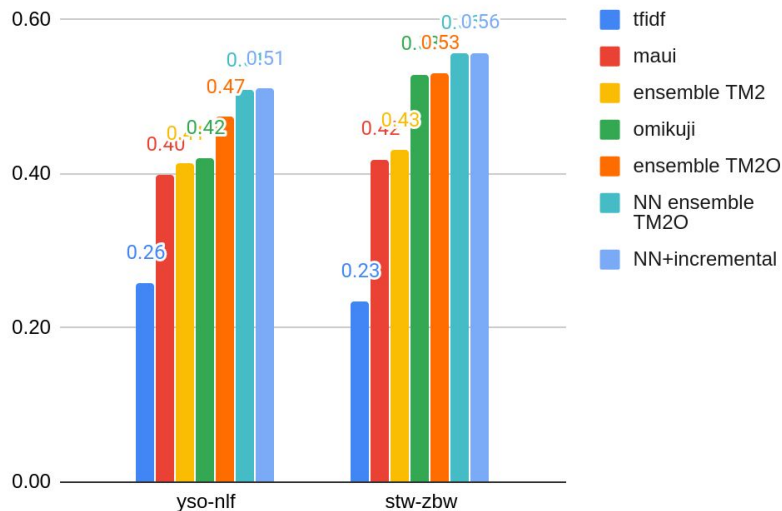


Expected results

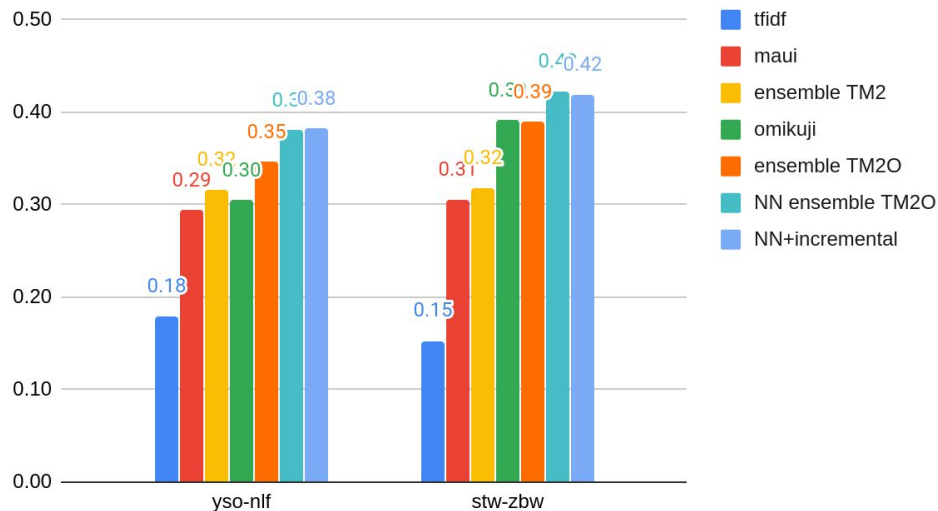
Scores for different backends by data sets



NDCG scores



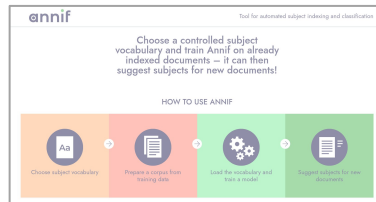
F1@5 scores



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

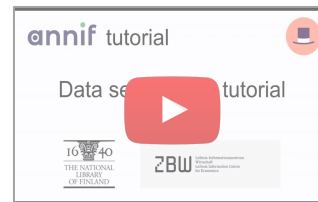
Abstract

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 the process. We have developed Annif, an open-source tool and environment for automated subject indexing. After training it with a subject vocabulary and training documents, 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 and 'old document' results, based on Wikipedia, and the outcomes of a blind comparison. The results of applying subject papers and recent books have been promising, indicating that subject 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 working with 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. Annif uses an internal machine learning model. Annif is identified by a project ID, which is typically a short string such as 'my-first'.

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 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 or other 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 `libris` 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 `project` data set, use the following settings:

```
[core:tfidf:web]
sourcepath: /tmp/annif
targetpath: /tmp/annif
```

3

Join an online session (optional)

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



See the Annif-tutorial GitHub repository

Join the user group



The annif-users mailing list and web forum is available on Google Groups

