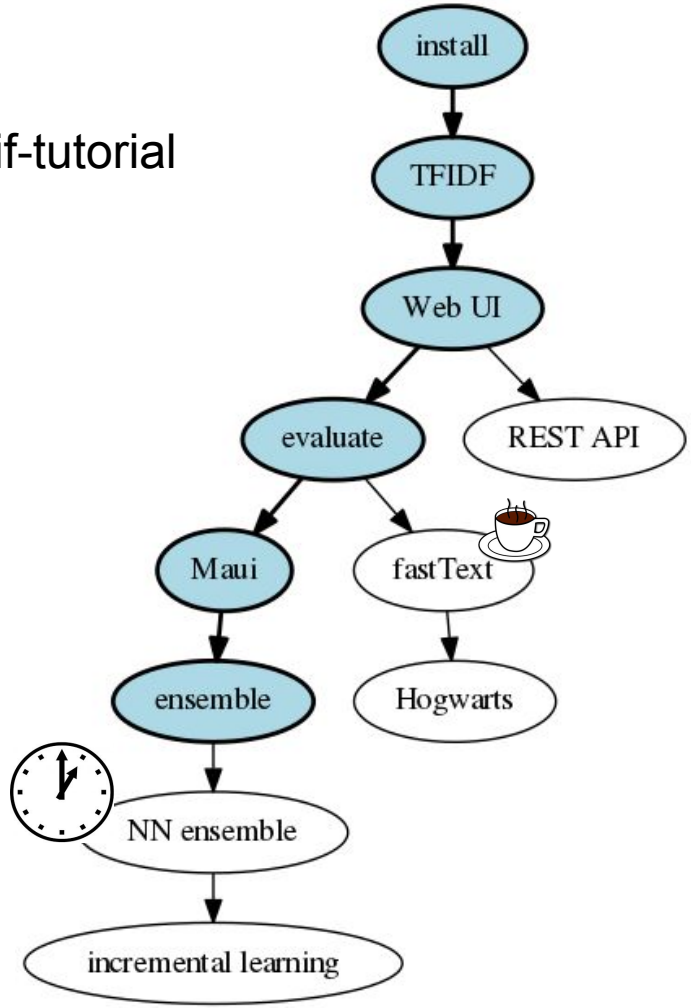
Annif tutorial @ SWIB19

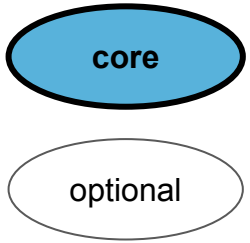
1640
THE NATIONAL LIBRARY OF FINLAND

Exercises

ZBW
Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

[Exercises in GitHub]
https://github.com/NatLibFi/Annif-tutorial





Exercise 1: Installation

VirtualBox install

1. Install VirtualBox software
2. Copy the “VirtualBox VMs / annif-tutorial” folder from the USB drive to your VirtualBox VMs folder
3. Add the VM in VirtualBox
4. Start the VM

Docker install

1. Install Docker software
2. Copy the “Annif-tutorial” folder from the USB drive to your home folder
3. Open a terminal and follow the detailed instructions

Local install (Linux only)

1. Create a Python venv
2. Activate the venv
3. `pip install annif`
4. Follow the detailed instructions

Please write on a post-it note and stick it on your laptop:

- your operating system
- type of install (VirtualBox, Docker or local install)

Data sets for this tutorial

yso-nlf and stw-zbw

Data set of the National Library of Finland (NLF)

The folder yso-nlf in our GitHub repository contains

- the trilingual General Finnish Ontology YSO plus YSO-Places
- a training data set constructed from metadata records
  from the Finna.fi discovery service, and
- 2.066 English language Master's and doctoral theses
  published in the years 2010 to 2017
  from the University of Jyväskylä (JYX repository).

Data set of ZBW

The folder stw-zbw contains

- the STW thesaurus for economics
- a training data set constructed from metadata records from the EconBiz discovery service
- 4.192 working papers in economics from the ZBW open access repository EconStor

Contents of short text training data sets

	Title	Descriptors (URIs)		
1	Principles of orchestration : with musical examples draw	<http://www.yso.fi/onto/yso/p12833>	<http://www.yso.fi/onto/yso/p12833>	
2	Proceedings of the 10th World Clean Air Congress, held	<http://www.yso.fi/onto/yso/p11516>	<http://www.yso.fi/onto/yso/p5393>	<http:/
3	Audit of the University of Eastern Finland 2017	<http://www.yso.fi/onto/yso/p10895>	<http://www.yso.fi/onto/yso/p7413>	<http:/
4	The Evangelical-Lutheran Church in Finland. 1984-1987	<http://www.yso.fi/onto/yso/p11817>	<http://www.yso.fi/onto/yso/p94426>	<http:/
5	The power of appreciative inquiry : a practical guide to p	<http://www.yso.fi/onto/yso/p272>	<http://www.yso.fi/onto/yso/p277>	<http:/
6	Market society : markets and modern social theory	<http://www.yso.fi/onto/yso/p10825>	<http://www.yso.fi/onto/yso/p16572>	<http:/
7	Lean supply chain management essentials : a framework	<http://www.yso.fi/onto/yso/p944>	<http://www.yso.fi/onto/yso/p9140>	<http:/
8	Deciding where to live : an interdisciplinary approach to	<http://www.yso.fi/onto/yso/p1797>	<http://www.yso.fi/onto/yso/p7432>	<http:/
9	Molecular basis of colorectal cancer predisposition	<http://www.yso.fi/onto/yso/p5937>	<http://www.yso.fi/onto/yso/p147>	<http:/

1	Demographic and labour force analysis based on Euro	<http://zbw.eu/stw/descriptor/11271-0>	<http://zbw.eu/stw/descriptor/15912-3	<http://
2	Agriculture and the GATT : rewriting the rules	<http://zbw.eu/stw/descriptor/18008-1>	<http://zbw.eu/stw/descriptor/10713-6	<http://
3	Below-replacement fertility in industrial societies : cau	<http://zbw.eu/stw/descriptor/15941-3>	<http://zbw.eu/stw/descriptor/10173-5	<http://
4	Spatial differentiation in the impact of technology on s	<http://zbw.eu/stw/descriptor/10470-6>	<http://zbw.eu/stw/descriptor/19073-6	<http://
5	Private interests, public policy, and American agricultu	<http://zbw.eu/stw/descriptor/10968-1>	<http://zbw.eu/stw/descriptor/11801-4	<http://
6	Rural development and population: institutions and p	<http://zbw.eu/stw/descriptor/10575-6>	<http://zbw.eu/stw/descriptor/13454-3	<http://
7	An integrated study of desertification : applications of	<http://zbw.eu/stw/descriptor/12021-4>	<http://zbw.eu/stw/descriptor/17677-5>	
8	At the very least she pays the rent : women and Germ	<http://zbw.eu/stw/descriptor/11284-5>	<http://zbw.eu/stw/descriptor/11313-3	<http://
9	The United States and Germany : a vital partnership	<http://zbw.eu/stw/descriptor/16441-4>	<http://zbw.eu/stw/descriptor/17829-1	<http://

Contents of fulltext training data sets

1 INTRODUCTION

Humans possess the ability to perceptually parse ongoing streams into discrete, meaningful events. This perceptual operation, which is called segmentation, makes it possible to understand continuous information or activities th… volve sound and movement, just like it is possible, in a messy re… each of its objects (Zacks & Swallow, 2007). Besides h… regarding human perception and cognition th… nentation has central importance f… s needed for languag… y music l… r high … y that … songs. … arger … rake … of

<http://www.yso.fi/onto/yso/p1808>	music
<http://www.yso.fi/onto/yso/p7302>	structure
<http://www.yso.fi/onto/yso/p5293>	perception (activity)
<http://www.yso.fi/onto/yso/p18246>	segmentation
<http://www.yso.fi/onto/yso/p277>	change
<http://www.yso.fi/onto/yso/p10670>	musicology
<http://www.yso.fi/onto/yso/p21685>	music research
<http://www.yso.fi/onto/yso/p9106>	listening

College Major Choice and the Gender Gap
Basit Zafar
Federal Reserve Bank of New York Staff Reports, no. 364
February 2009
JEL classification: D8, I2, J1, Z1

Abstract

Males and fe… ifferent in their choice of college major. Two main reasons hav… differences in innate abilities and differences … how college majors are chosen, foc… be consistent with many … Northwes… about ch… major is … realizati… finding … import… prefere… Nonpecuniary outcom… hile pecuniary outcomes realized at the … e for males. I decompose the gender gap into differenc… der

<http://zbw.eu/stw/descriptor/11295-0>	Occupational choice
<http://zbw.eu/stw/descriptor/11378-3>	Students
<http://zbw.eu/stw/descriptor/19756-6>	Gender discrimination
<http://zbw.eu/stw/descriptor/11327-6>	Wage structure
<http://zbw.eu/stw/descriptor/19516-5>	Returns to education
<http://zbw.eu/stw/descriptor/17829-1>	United States

Both data sets at a glance

	vocabulary (languages) (#concepts, terms)	**short texts training docs**	**fulltexts (#; train, validate, test)**
NLF	YSO version 2019.3 Cicero (Finnish, Swedish, English) 32.265 concepts, 168.456 terms	~2 Mio. (~100.000 for testing)	Master's & doctoral theses (2.066; 1.417, 349, 300)
ZBW	STW version 9.06 (German, English) 5.746 concepts, 32.272 terms	~1 Mio. (~100.000 for testing)	articles / working papers (4.192; 2.939, 628, 625)

Choose one data set and use it for the rest of the tutorial

Exercise 2: TFIDF project

- The “Hello World” algorithm of automated subject indexing: quick to set up and test, but not the final say!

So, what are the alternatives?

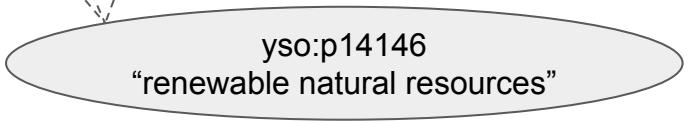
A little bit about algorithms

Lexical vs. Associative algorithms for subject indexing

Lexical approaches: e.g. Maui

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.“*



Lexical vs. Associative algorithms for subject indexing

Lexical approaches: e.g. Maui

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."*

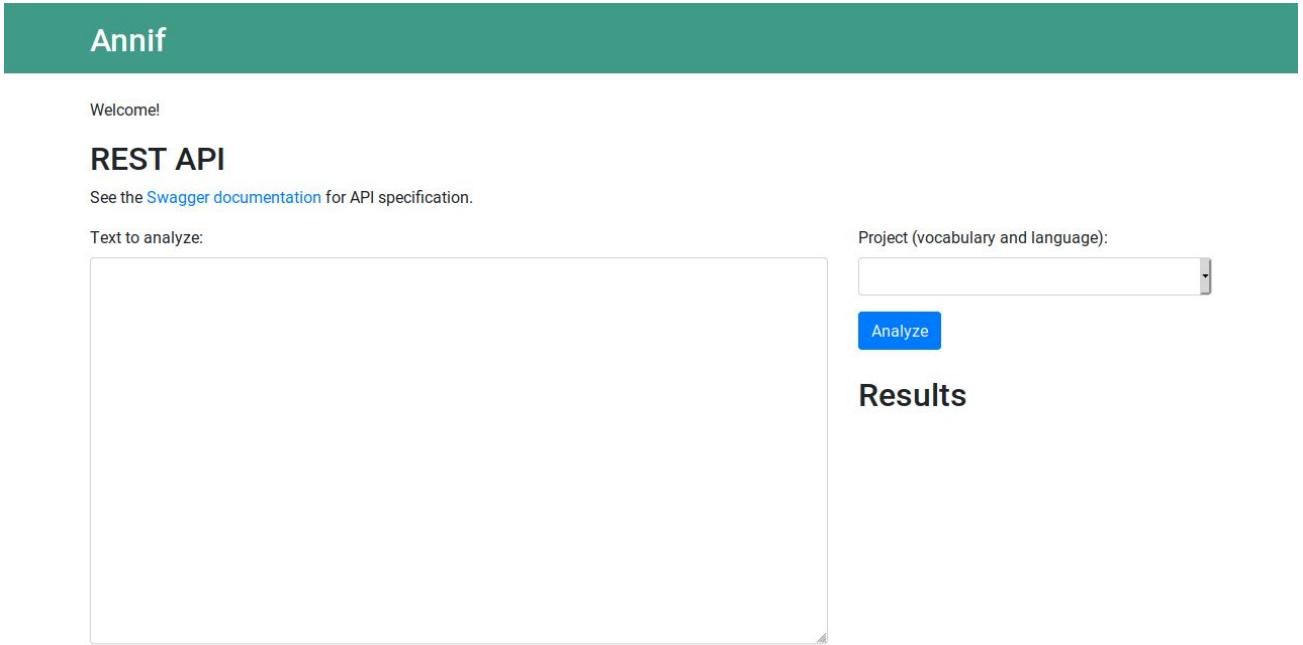
yso:p14146
"renewable natural resources"

Associative approaches (TFIDF, fastText …)

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

Exercise 3: Web user interface

- Good for quick interactive testing on example documents



Metrics – what do we need them for?

In order to assess and compare the quality of the output
of our Annif projects, we need to fix criteria to do so.

To that end, we use metrics from machine learning / information retrieval because they provide numeric values that can be compared easily.

In this tutorial, we will consider the following:

- precision & recall
- F1 score
- Normalized Discounted Cumulative Gain (NDCG)

Precision, recall and F1 score

- **Precision:** fraction of correct descriptors among the descriptors suggested
  “How many of the suggested ones are actually correct?”
- **Recall:**
  fraction of the total amount of correct descriptors that were actually suggested
  “How many of those that the machine should suggest
  have actually been suggested?”
- The **F1 score** is the harmonic mean between precision and recall
  (i.e., a way of combining precision and recall values into
  an average measure between 0.0 – *worst*, and 1.0 – *best*).

NDCG – Normalized Discounted Cumulative Gain

The NDCG is a ranking-based measure, i.e., the order of the descriptors suggested by the machine is significant for its usefulness:

Getting the top ranked (highest score) result right will matter more than getting the 2nd or 3rd right.

Just like precision, recall, and the F1 score, the NDCG is also a value between 0.0 – *worst*, and 1.0 – *best*.

No need to decide up front whether to use top 5 or 10 or …

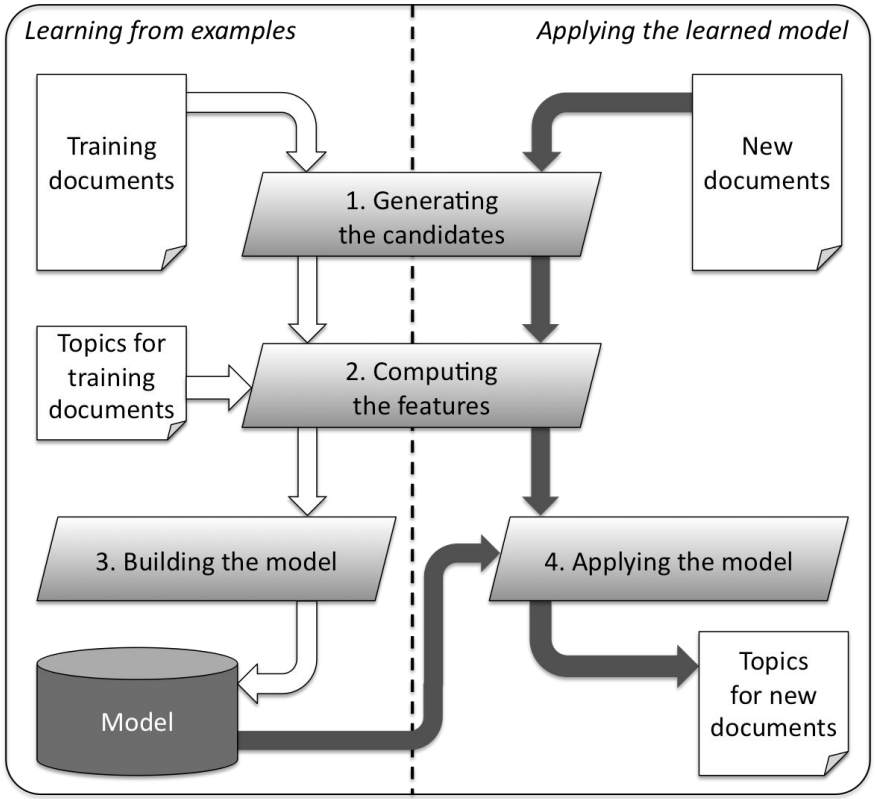
Exercise 4: Evaluate on document collections

- When you need to have firm numbers for quality
- Write down the F1@5 and NDCG scores you get, for all kinds of projects

Algorithms: Maui

Maui was developed at the University of Waikato (**Medelyan, 2009**).

It's lexical, but it does use heuristics for determining best possible matches between vocabulary terms and words in a document. It uses machine learning to better decide between available heuristics.



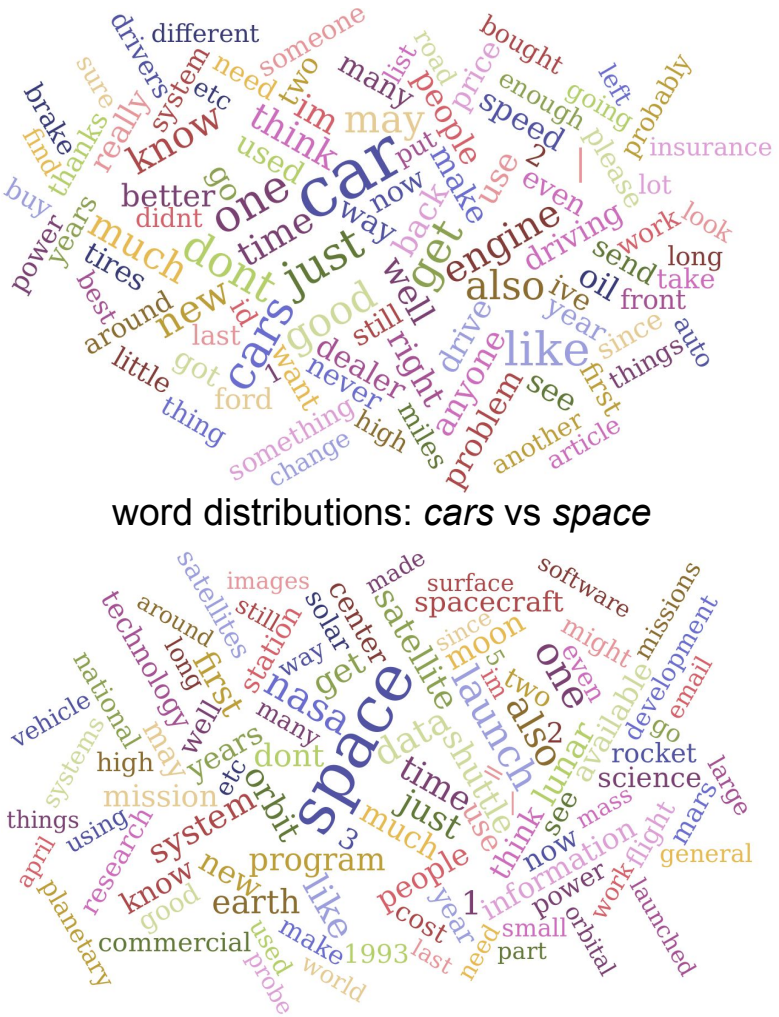
Operation of Maui (Medelyan, 2009)

Algorithms: TFIDF

Representative set of text is formed for each subject in the vocabulary from documents that have been manually indexed with that subject.

The *term frequencies* and *inverse document frequencies* are then calculated for all words appearing in those sets and these TF-IDF values are stored as vectors in an index.

For new documents, similar vectors are calculated and compared to the ones in the index.



Algorithms: fastText

fastText (**Joulin et al., 2017**) is a machine learning algorithm for text classification and representation created at Facebook Research.

1. transforms text into vectors (using bags of words or n-grams)
2. creates a neural network style model with a hidden (embedding) layer
3. adjusts the weights of the network based on training examples

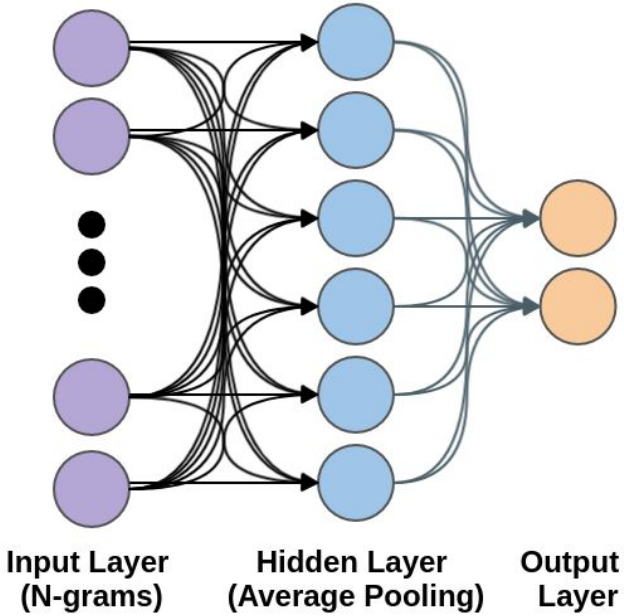


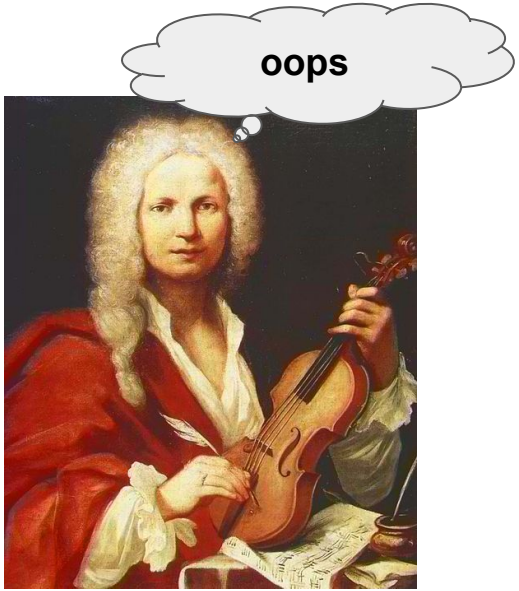
Image source: FastText for Sentence Classification by Austin G. Walters

Exercise 5: Maui project

- Maui is very good for detecting rare subjects mentioned by name
- Technical setup is a bit challenging – Maui Server is an external service

Algorithms may be used **alone**, or in combinations, **ensembles**

Algorithms make silly mistakes



Some reasons for mistakes:

- errors and skew in training data
- correlation ≠ causation
- homonyms (e.g. rock)
- misinterpreted names (e.g. Smith, AIDS)
- random noise

In an ensemble, **each algorithm makes different mistakes**

In an ensemble, **each algorithm makes different mistakes**

In an ensemble, **each algorithm makes different mistakes**

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

The three ensembles

Simple ensemble

Averages the scores given by different backends for all subjects.

No training of the ensemble

PAV ensemble

Applies *isotonic regression* to estimate the relationship between given scores and probability of relevance of a subject.

Must be trained

Neural network ensemble

A lot like PAV. Starts off like a simple averaging ensemble, but fine-tunes the scores based on training.

Must be trained

Can learn further after training

Wilbur, W. J., & Kim, W. (2014). **Stochastic Gradient Descent and the Prediction of MeSH for PubMed Records.** *AMIA Annual Symposium proceedings. AMIA Symposium*, *2014*, 1198-207.

Exercise 6: Ensemble project

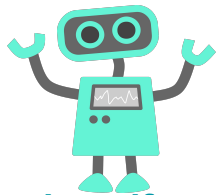
- Let’s set up a simple ensemble which combines results from the projects set up in previous exercises.
- In general, ensemble models should perform better than individual algorithms

Exercise 7: REST API

- Annif can be integrated to other systems via a simple RESTful API

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

Analyze this!



api.annif.org

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**”}
]

Exercise 8: fastText project

- A real machine learning algorithm!
- It can give good results, but it’s very sensitive to hyperparameters

Exercise 9: Hogwarts

- You can implement a Harry-Potter-style Sorting Hat with Annif!
- Character n-grams can be useful in other use cases as well…

By Suzelfe [CC BY-SA 4.0], Wikimedia Commons

Exercise 10: Neural network ensemble

- Let’s try a more intelligent ensemble based on neural networks
- It’s trainable and dynamic

Exercise 11: Incremental learning

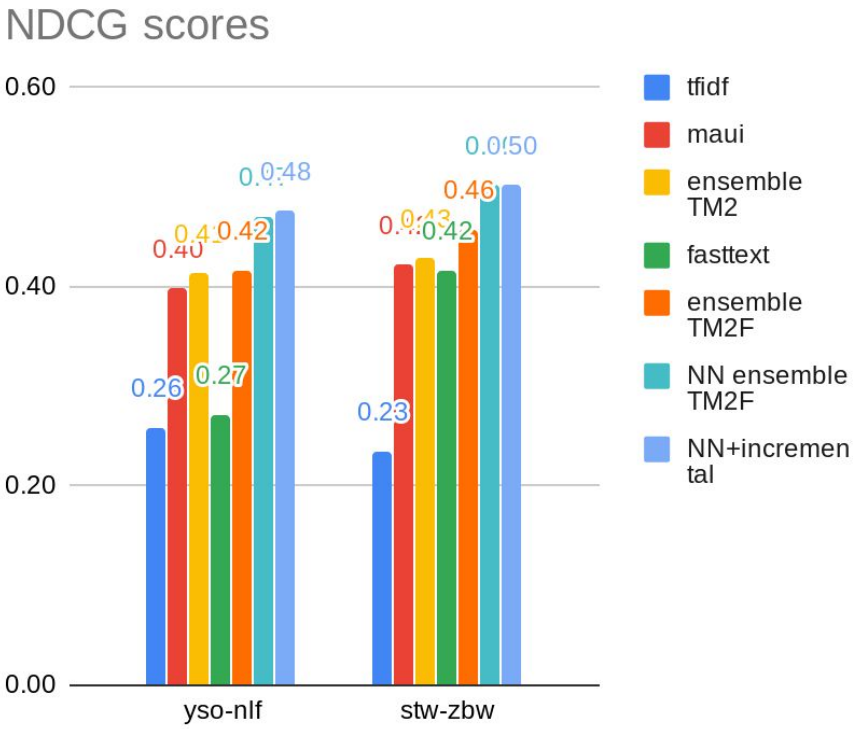
- Now to fine-tune the neural ensemble model from the previous exercise

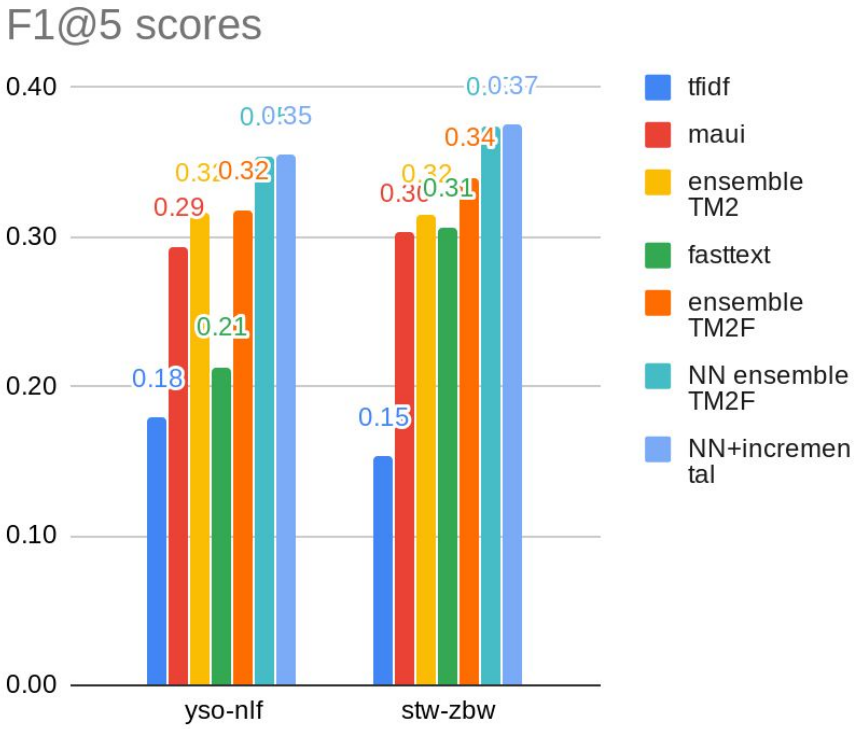
Closing words

- To summarize
  - … that’s how you use Annif
  - Annif can utilize several different backends to index subjects
  - It can easily be integrated to other systems (and modified or tweaked)

Expected results

Scores for different backends by data sets





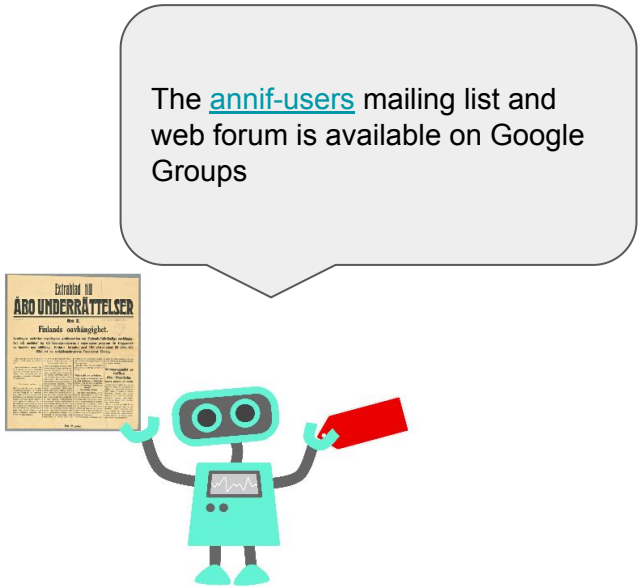
Future improvements to Annif

(no promises - contributions welcome!)

New algorithms	- e.g. Omikuji backend
Optimizations	- caching preprocessed training data - better scalability by spooling data to disk (LMDB?) - more use of parallel processing
Easier administration	- automated hyperparameter search - administration through REST API and Web UI

Closing words

- Fill in the feedback form at https://tinyurl.com/swib19annif
- Find out more at https://github.com/NatLibFi/Annif or annif.org




Thank you!