



# Annif tutorial



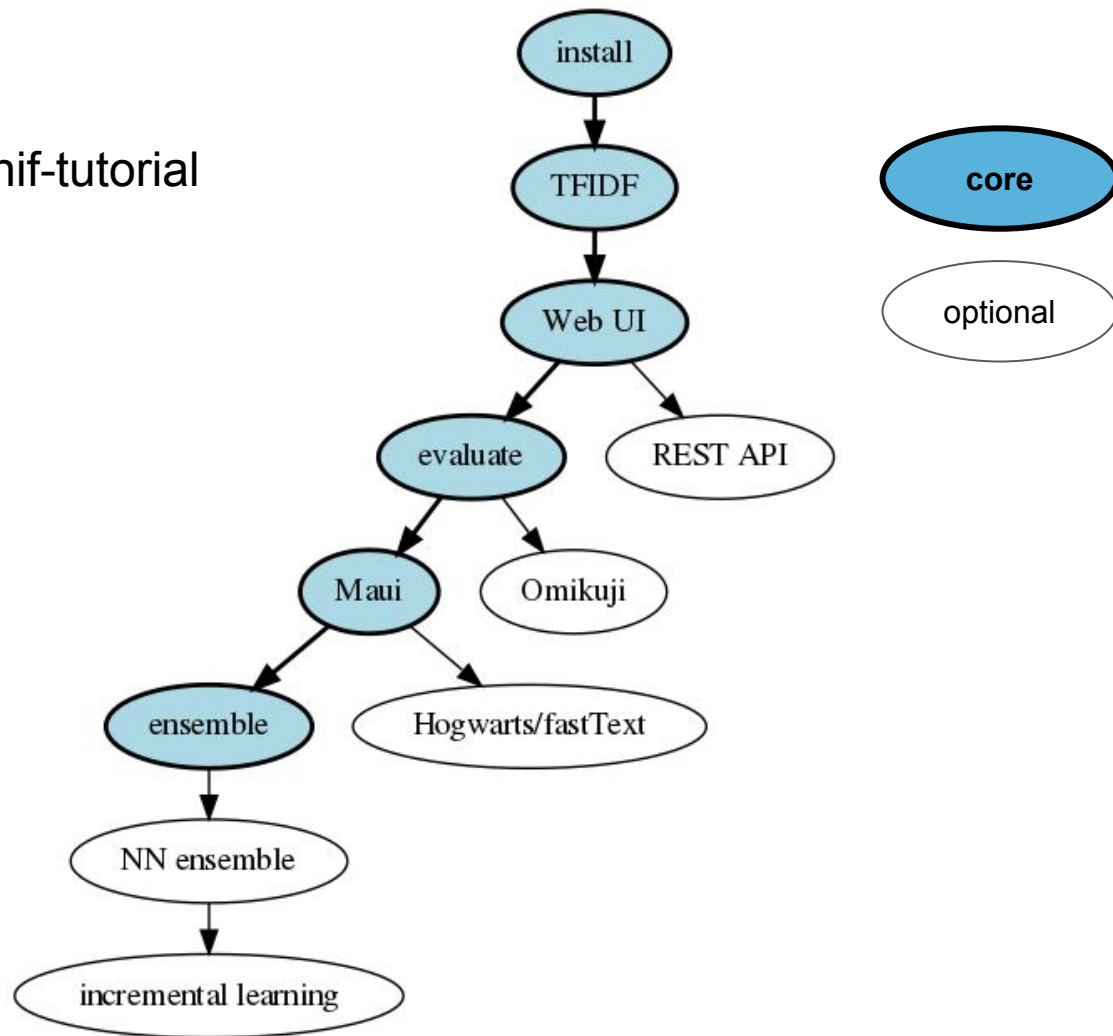
Exercises



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 drawn from the repertoire of the Finnish National Opera	<http://www.yso.fi/onto/yso/p12833>	<http://www.yso.fi/onto/yso/p12833>	
2	Proceedings of the 10th World Clean Air Congress, held in Helsinki, Finland, 1988	<http://www.yso.fi/onto/yso/p11516>	<http://www.yso.fi/onto/yso/p5393>	<http://www.yso.fi/onto/yso/p11516>
3	Audit of the University of Eastern Finland 2017	<http://www.yso.fi/onto/yso/p10895>	<http://www.yso.fi/onto/yso/p7413>	<http://www.yso.fi/onto/yso/p10895>
4	The Evangelical-Lutheran Church in Finland. 1984-1987	<http://www.yso.fi/onto/yso/p11817>	<http://www.yso.fi/onto/yso/p94426>	<http://www.yso.fi/onto/yso/p11817>
5	The power of appreciative inquiry : a practical guide to using appreciative inquiry in the workplace	<http://www.yso.fi/onto/yso/p272>	<http://www.yso.fi/onto/yso/p277>	<http://www.yso.fi/onto/yso/p272>
6	Market society : markets and modern social theory	<http://www.yso.fi/onto/yso/p10825>	<http://www.yso.fi/onto/yso/p16572>	<http://www.yso.fi/onto/yso/p10825>
7	Lean supply chain management essentials : a framework for lean thinking	<http://www.yso.fi/onto/yso/p944>	<http://www.yso.fi/onto/yso/p9140>	<http://www.yso.fi/onto/yso/p944>
8	Deciding where to live : an interdisciplinary approach to housing choice	<http://www.yso.fi/onto/yso/p1797>	<http://www.yso.fi/onto/yso/p7432>	<http://www.yso.fi/onto/yso/p1797>
9	Molecular basis of colorectal cancer predisposition	<http://www.yso.fi/onto/yso/p5937>	<http://www.yso.fi/onto/yso/p147>	<http://www.yso.fi/onto/yso/p5937>

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

# 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 that involve sound and movement, just like it is possible, in a messy room, to identify each of its objects (Zacks & Swallow, 2007). Besides the role of segmentation in human perception and cognition, it also has central importance for language processing. This paper discusses the role of segmentation in music perception and cognition.

1	<a href="http://www.yso.fi/onto/yso/p1808">http://www.yso.fi/onto/yso/p1808</a>	music
2	<a href="http://www.yso.fi/onto/yso/p7302">http://www.yso.fi/onto/yso/p7302</a>	structure
3	<a href="http://www.yso.fi/onto/yso/p5293">http://www.yso.fi/onto/yso/p5293</a>	perception (activity)
4	<a href="http://www.yso.fi/onto/yso/p18246">http://www.yso.fi/onto/yso/p18246</a>	segmentation
5	<a href="http://www.yso.fi/onto/yso/p277">http://www.yso.fi/onto/yso/p277</a>	change
6	<a href="http://www.yso.fi/onto/yso/p10670">http://www.yso.fi/onto/yso/p10670</a>	musicology
7	<a href="http://www.yso.fi/onto/yso/p21685">http://www.yso.fi/onto/yso/p21685</a>	music research
8	<a href="http://www.yso.fi/onto/yso/p9106">http://www.yso.fi/onto/yso/p9106</a>	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 females differ in their choice of college major. Two main reasons have been suggested: differences in innate abilities and differences in innate preferences. This paper shows that college majors are chosen, for the most part, for reasons that are consistent with many of the findings of the literature on the gender gap in the United States. I decompose the gender gap into differences in innate abilities and preferences. I find that the gender gap is larger for majors that require higher cognitive skills and for majors that require higher social skills. The gender gap is also larger for majors that require higher social skills and for majors that require higher cognitive skills. The gender gap is also larger for majors that require higher social skills and for majors that require higher cognitive skills.

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



# Both data sets at a glance

	<b>vocabulary (languages) (#concepts, terms)</b>	<b>short texts training docs</b>	<b>fulltexts (#; train, validate, test)</b>
<b>NLF</b>	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)
<b>ZBW</b>	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?*

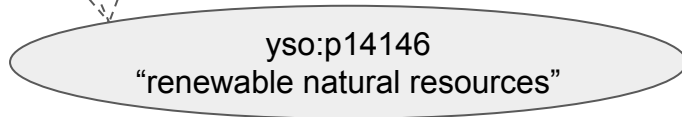
# A little bit about algorithms

# 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

Annif

Welcome!

## REST API

See the [Swagger documentation](#) for API specification.

Text to analyze:

Project (vocabulary and language):

Analyze

Results

# 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 ...*



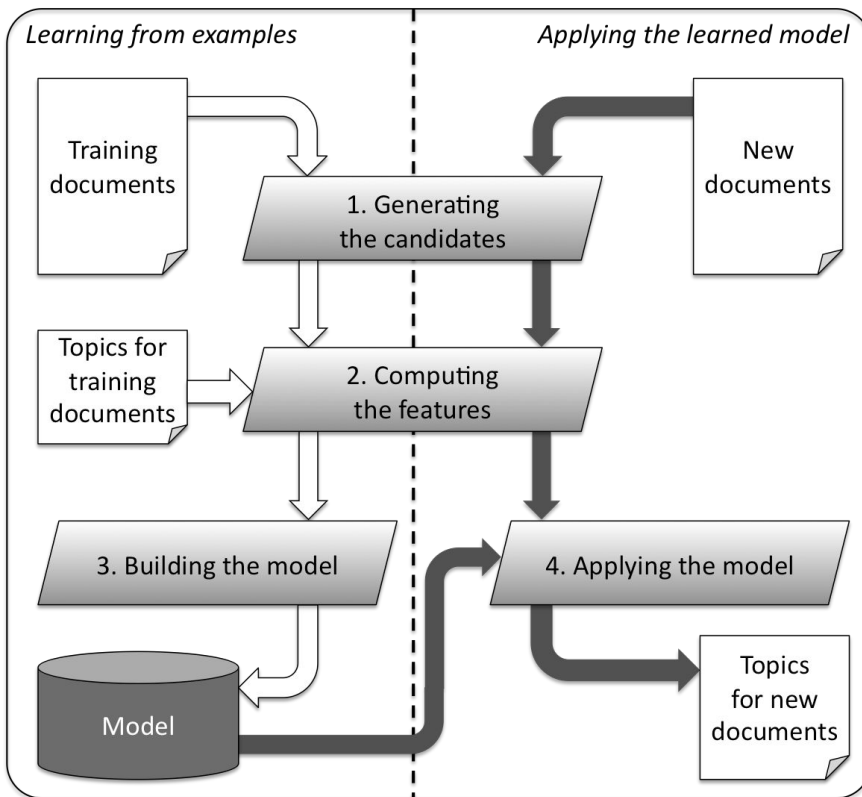
## Exercise 4: Evaluate on document collections

- 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

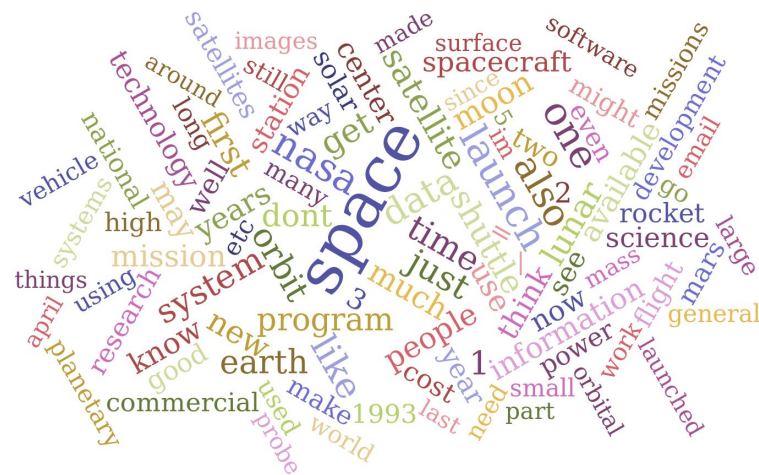
Representative set of text is formed for each subject in the vocabulary from documents that have been manually indexed with that subject.

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.



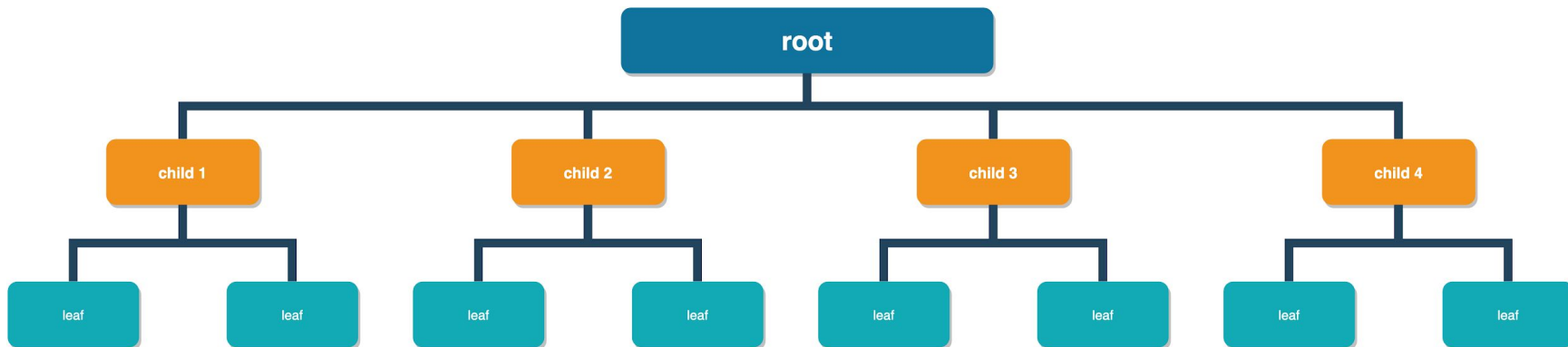
word distributions: *cars* vs *space*



# Algorithms: Omikuji

[Omikuji](#) is an implementation of a family of tree-based machine learning algorithms for extreme multilabel classification. With Omikuji it is possible to build models that emulate multiple algorithms, for example

- [Parabel](#)
- [Bonsai](#)
- [AttentionXML](#)



# 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

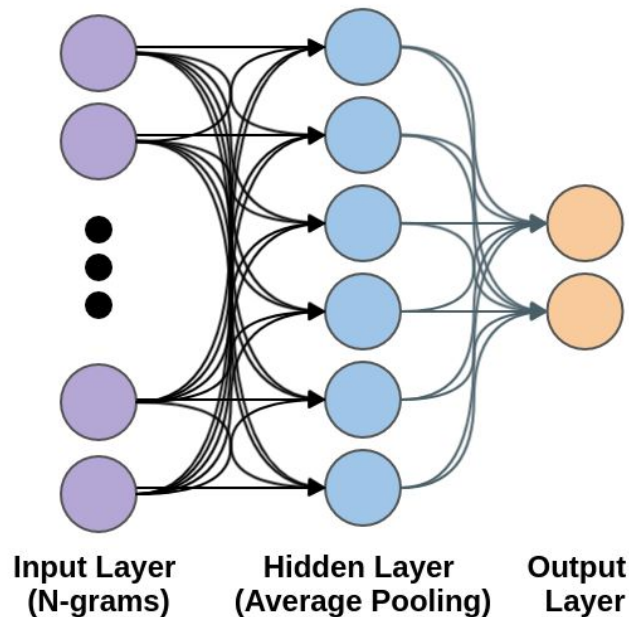


Image source: [FastText for Sentence Classification](#) by Austin G. Walters

## Exercise 5: [Maui project](#)

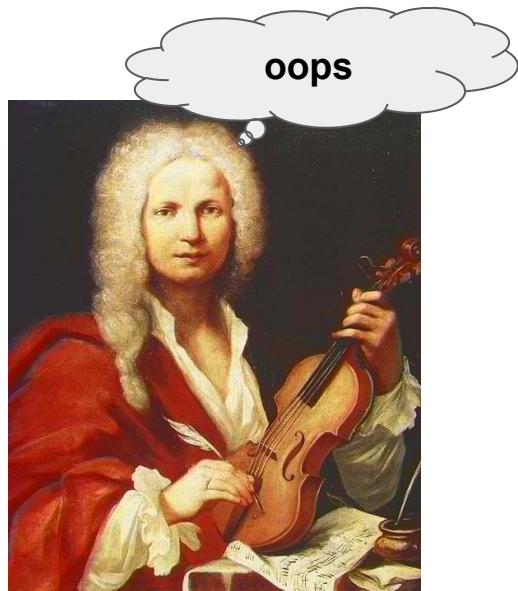
- Maui is very good for detecting rare subjects mentioned by name
- Technical setup is a bit challenging – Maui Server is an external service



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



# Algorithms make silly mistakes



## Some reasons for mistakes:

- errors and skew in training data
- correlation  $\neq$  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!



```
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"}  
]
```



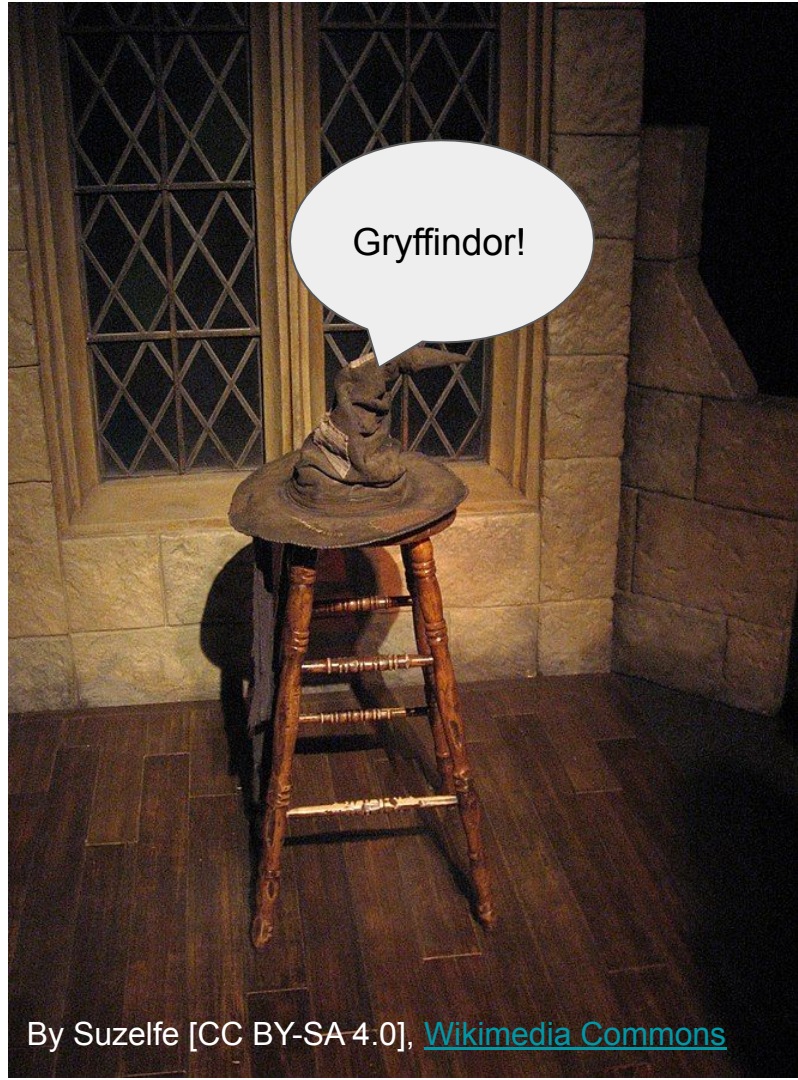
## Exercise 8: Omikuji project

- Set up a model that emulates the Parabel algorithm (by using the Omikuji backend in Annif with default parameter values)
- Very good results can be expected

## Exercise 9: Hogwarts - Sorting Hat with fastText!

fastText library:

- A real machine learning algorithm!
- It can give good results, but it's very sensitive to hyperparameters
- Character n-grams can be useful in other use cases as well...
- “[...a sentient hat at Hogwarts that magically determines which of the four school Houses each new student belongs most to.](#)”



## 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

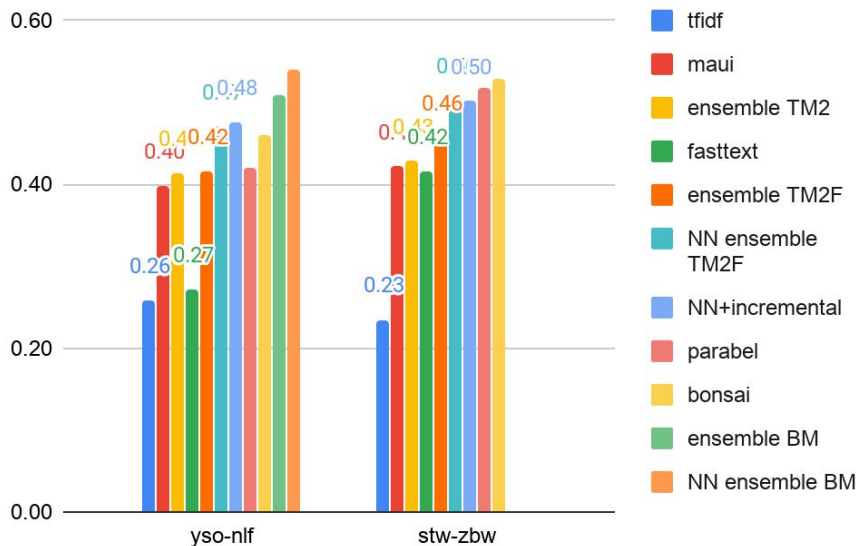
# Closing words

- 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)

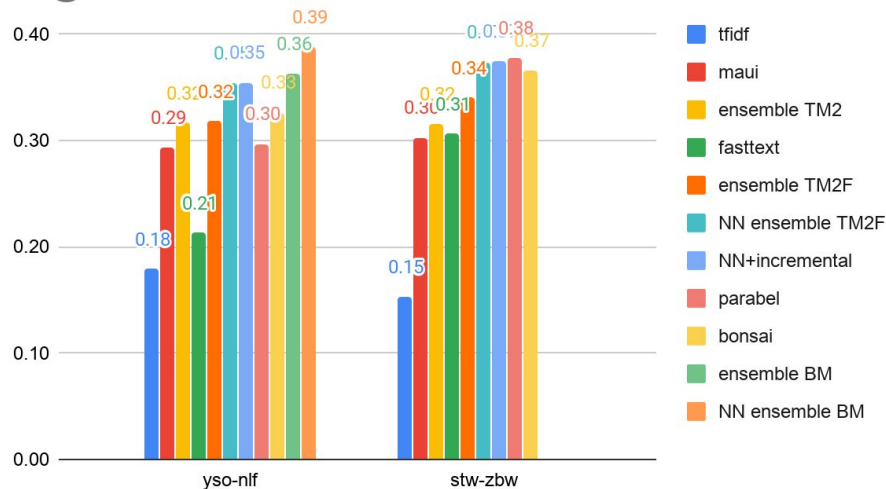
# Expected results

Scores for different backends by data sets

NDCG scores



F1@5 scores



# Future improvements to Annif

(no promises - contributions welcome!)

## **New algorithms**

## **Optimizations**

- more use of parallel processing

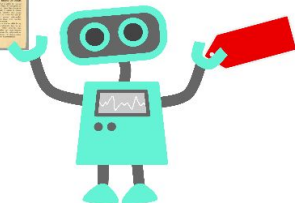
## **Easier administration**

- automated hyperparameter search
- administration through REST API and Web UI

# Closing words

- Fill in the feedback form at TBD
- Find out more at <https://github.com/NatLibFi/Annif> or [annif.org](https://annif.org)

The [annif-users](#) mailing list and web forum is available on Google Groups





*Thank you!*

