annif tutorial

A little bit about algorithms

Two kinds of approaches

1640 THE NATIONAL LIBRARY OF FINLAND

ZBW Leibniz-Informationszentrum Wirtschaft Leibniz Information Centre for Economics

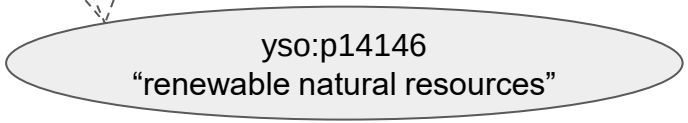
Lexical vs. associative algorithms for subject indexing

Lexical vs. associative algorithms for subject indexing

lexical approaches (e.g.: MLLM, stwfsa)

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 approaches need comparatively little training data.

Lexical vs. associative algorithms for subject indexing

lexical approaches (e.g.: MLLM, stwfsa)

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"

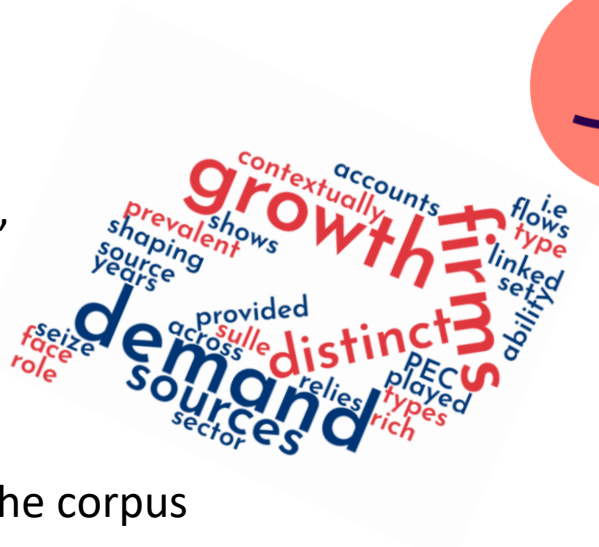
Lexical approaches need comparatively little training data.

associative approaches (e.g.: TF-IDF, fastText, Omikuji)

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

Associative approaches need a lot more training data in order to cover each subject.

TF-IDF



- “term frequency – inverse document frequency” (or TF-IDF) is based on the assumption that
  a term which does not occur frequently
  in general (i.e., in the entire corpus)
  but occurs frequently in a certain document of the corpus
  could indicate a subject that is relevant to the content of the document

- TF-IDF similarity as a way to compare new documents to known documents is a very simple numerical statistic which can be used to establish a baseline that more advanced machine learning methods have to meet

Algorithms used in Annif

lexical

MLLM (improved reimplementation of a lexical tool called Maui)

stwfsa (improved reimplementation of an algorithm based on finite-state automata)

MLLM and stwfsa are based on lexical algorithms for automated indexing

associative

TF-IDF similarity (implemented with the Gensim Python library)

baseline bag-of-words similarity measure and vector space model

fastText (by Facebook Research)

uses word embeddings and simulates a deep neural network architecture

Parabel and Bonsai (implemented with the Omikuji Python library)

tree-based algorithms for extreme multi-label classification (i.e., when the set of subjects is huge)

implemented as Annif backends – see the Annif wiki documentation for details about each backend