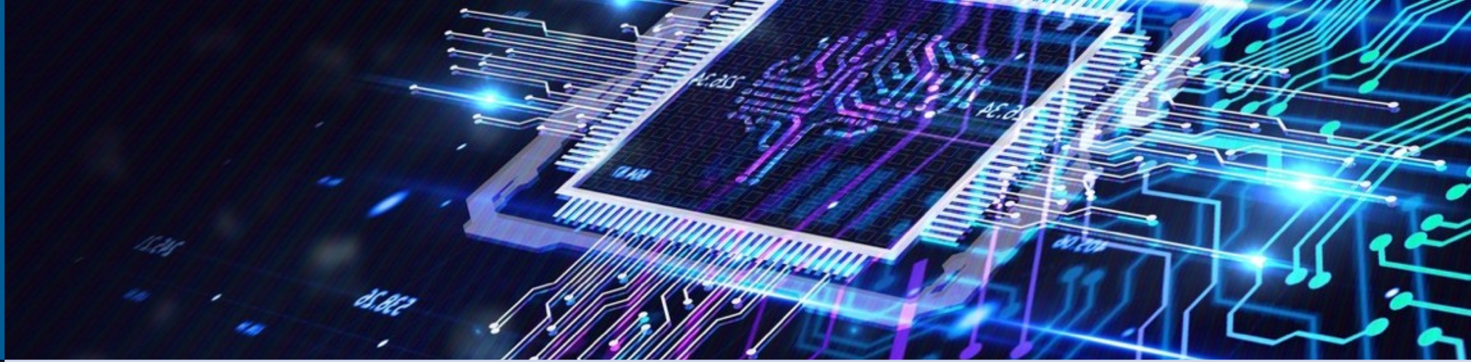
CSC

ICT Solutions for Brilliant Minds

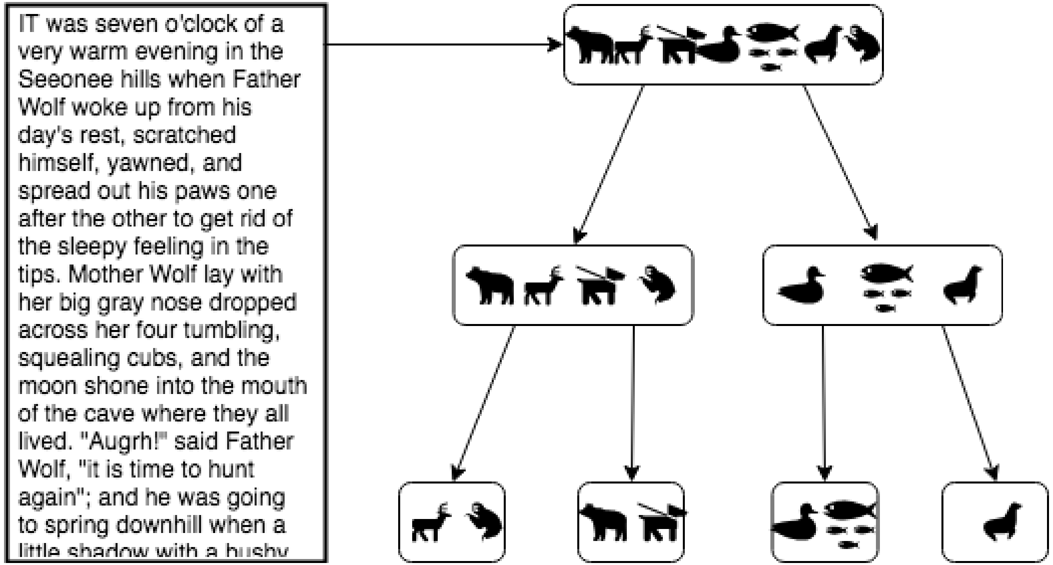


Tree-based methods for XMTC

Mats Sjöberg

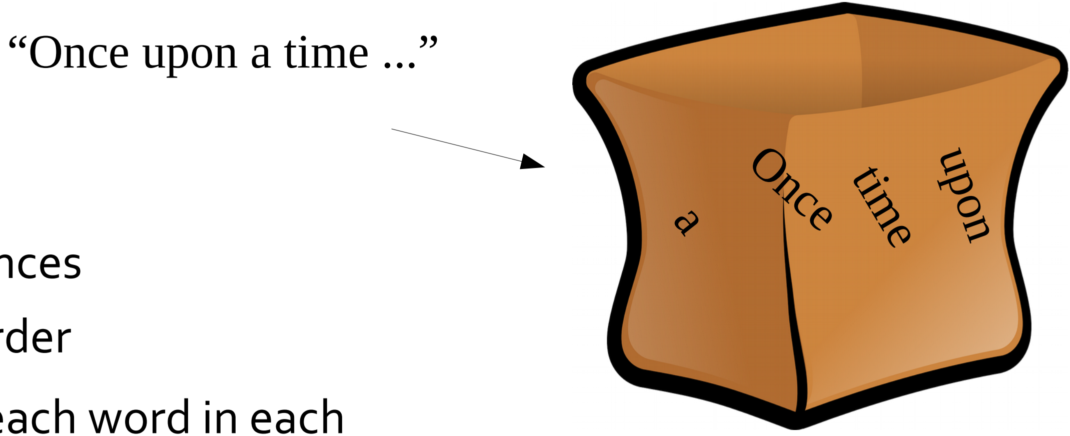
Tree-based methods for XMTC

- Recursively divide the space of labels or features ***based on data***
- Many methods: FastXML, PFastreXML, Parabel, Bonsai, SwiftXML, CraftML, …



IT was seven o'clock of a very warm evening in the Seeonee hills when Father Wolf woke up from his day's rest, scratched himself, yawned, and spread out his paws one after the other to get rid of the sleepy feeling in the tips. Mother Wolf lay with her big gray nose dropped across her four tumbling, squealing cubs, and the moon shone into the mouth of the cave where they all lived. "Augrh!" said Father Wolf, "it is time to hunt again"; and he was going to spring downhill when a little shadow with a bushy

tf-idf features



- Bag-of-words model
  - Just count word occurrences
  - Don’t care about their order
- Instead of raw word count, each word in each document is represented by a tf-idf weight
  - tf-idf = tf × idf
  - tf = term frequency = how often word occurs in a document
  - idf = inverse document frequency = N / document frequency
    - N = total number of documents
    - document frequency = proportion of documents where this word occurs
    - idf measure of how rare a word is

tf-idf example

D1: this is a cat

D2: this is a dog

D3: this is also a cat

N=3

idf = N/df = 3/df

tf-idf = tf × idf

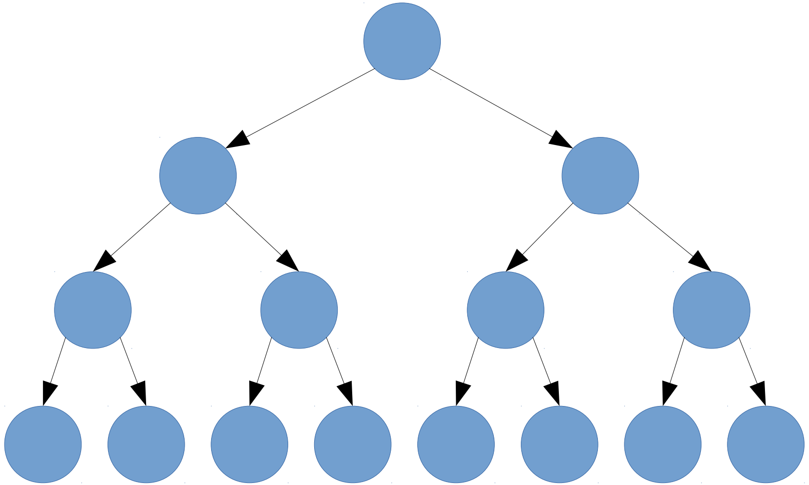
tf	a	also	cat	dog	is	this
D1	1	0	1	0	1	1
D2	1	0	0	1	1	1
D3	1	1	1	0	1	1

	a	also	cat	dog	is	this
df	3	1	2	1	3	3
idf	1.0	3.0	1.5	3.0	1.0	1.0

tf-idf	a	also	cat	dog	is	this
D1	1.0	0	1.5	0	1.0	1.0
D2	1.0	0	0	3.0	1.0	1.0
D3	1.0	3.0	1.5	0	1.0	1.0

Parabel

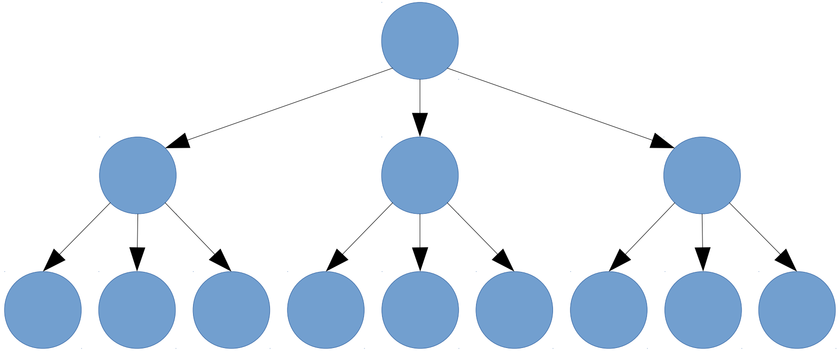
- Partitions label space recursively using 2-means clustering → *binary label-tree*
- *Balanced*: (nearly) same number of labels in each cluster
- One-vs-all classifier at leaf nodes
- Suffers from error propagation and cascading effect
  - error at top cascades down in tree
  - poor performance for uncommon labels



Y. Prabhu, A. Kag, S. Harsola, R. Agrawal, and M. Varma, "Parabel: Partitioned Label Trees for Extreme Classification with Application to Dynamic Search Advertising," WWW Conference 2018

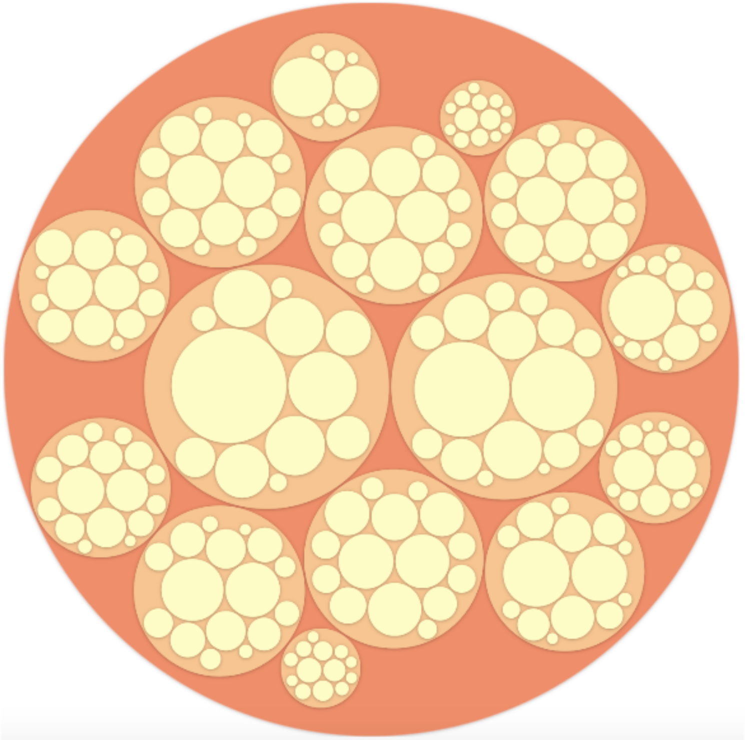
Bonsai

- *Shallow tree*
  - K-means, with K ≥ 2, often K ≥ 100
  - Less levels needed than in Parabel
- Clusters *don't need to be balanced*
- *Generalized label representation*
  1) Labels represented in feature space (input space)
  2) Co-occurrence of labels (output space)
  3) Combination of both

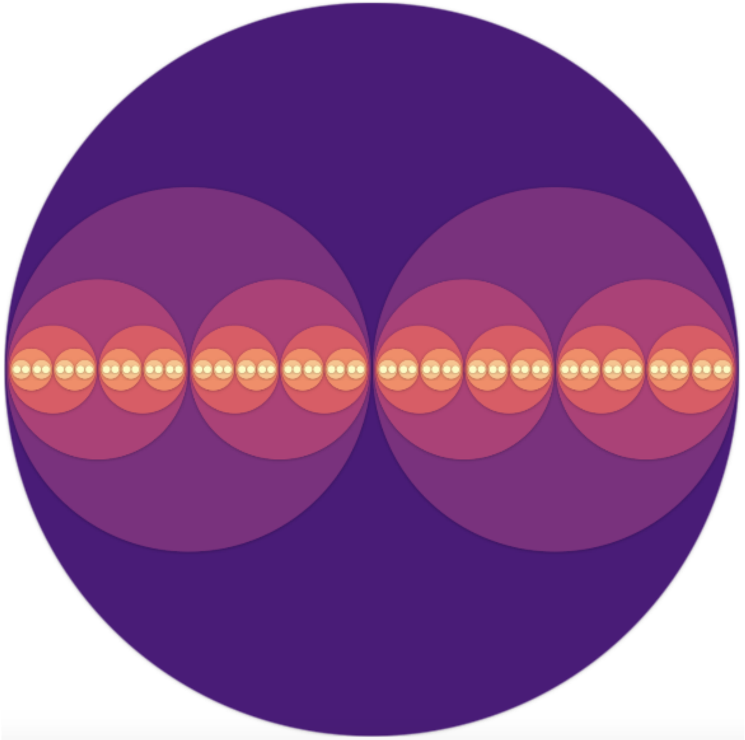


S. Khandagale, H. Xiao, and R. Babbar, "Bonsai - Diverse and Shallow Trees for Extreme Multi-label Classification," Machine Learning 2020

Parabel vs Bonsai



Bonsai : $K = 16$, tree depth 2



Parabel : $K = 2$, tree depth 6

Omikuji

- Efficient implementation of Parabel: https://github.com/tomtung/omikuji
- Also supports:
  - Unbalanced trees
  - Clustering with K≥2
  - Layer collapsing
    - removing adjacent layers to transform binary tree to more wide and shallow tree

Similar to *Bonsai* (except for generalized label representation)

Similar to *AttentionXML* (except for deep learning part…)

Some other tree-based methods

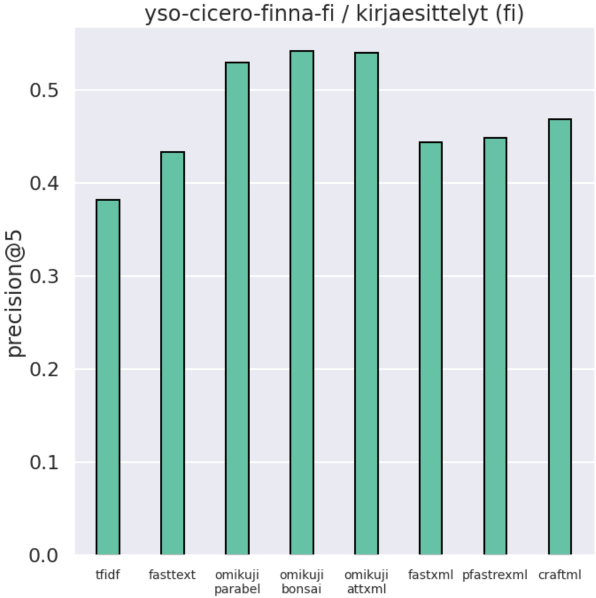
- FastXML: learns an ensemble of trees which partitions feature space by directly optimizing an nDCG based ranking loss function

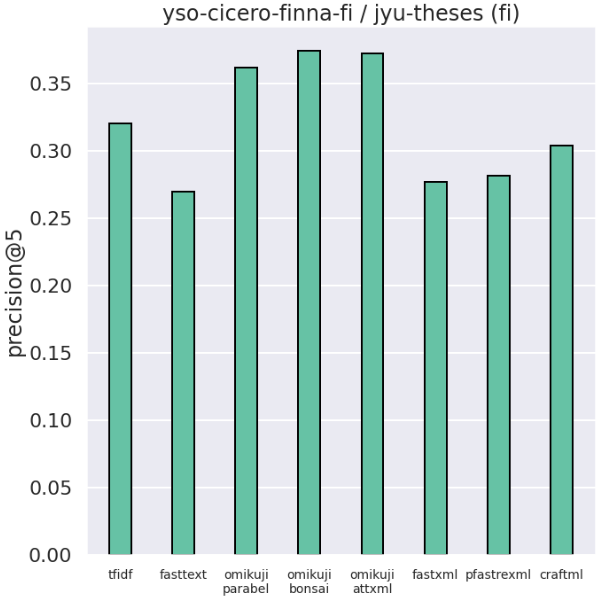
  Y. Prabhu and M. Varma. "FastXML: a fast, accurate and stable tree-classifier for extreme multi-label learning.", ACM KDD 2014

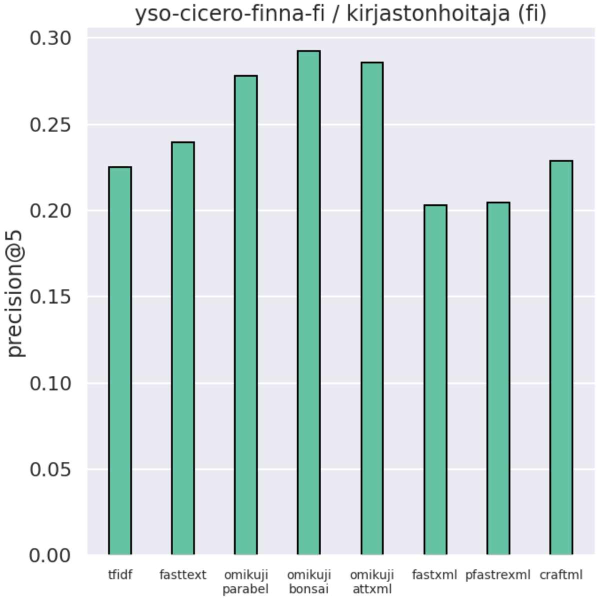
- CraftML: similar to FastXML, but uses Random Forest-style ensembling with random sampling of label and feature space

  W. Siblini, P. Kuntz, F. Meyer, "CRAFTML, an Efficient Clustering-based Random Forest for Extreme Multi-label Learning", PMLR 2018.

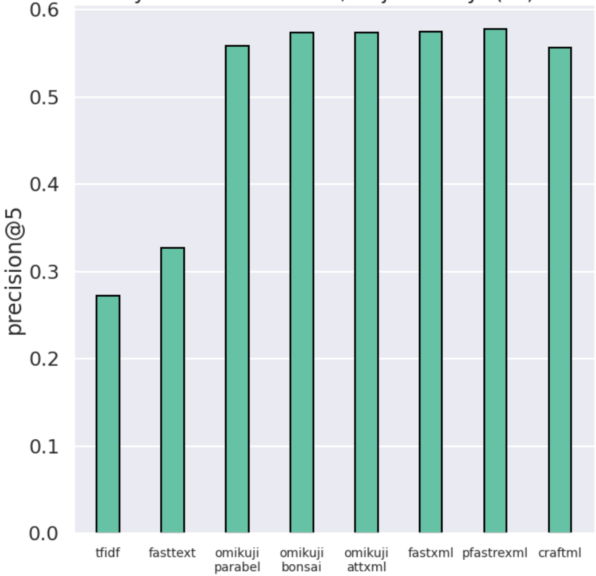
Some results: Finnish

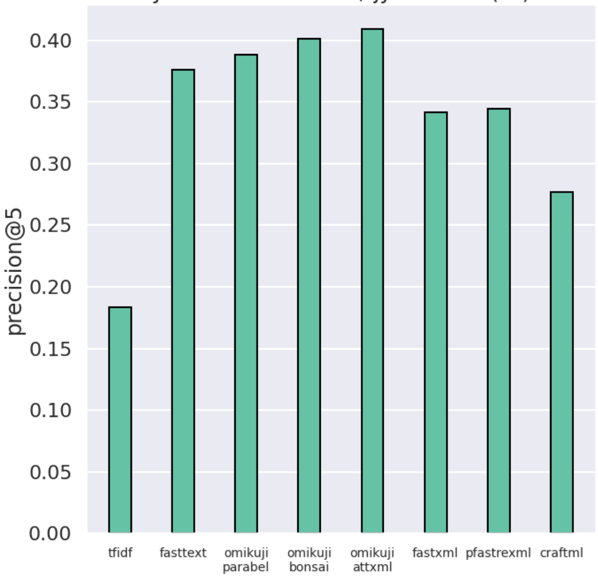

yso-cicero-finna-fi / kirjaesittelyt (fi)

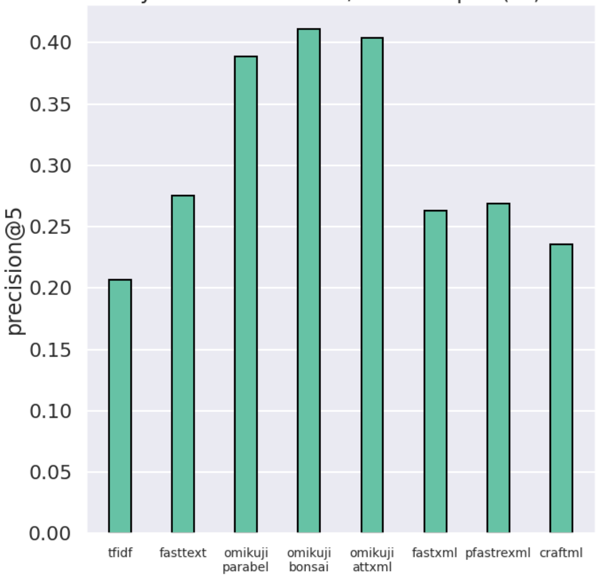

yso-cicero-finna-fi / jyu-theses (fi)


yso-cicero-finna-fi / kirjastonhoitaja (fi)

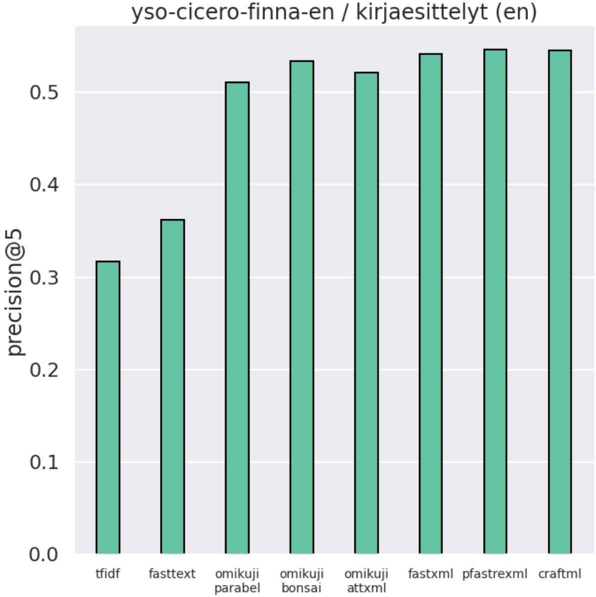
Some results: Swedish

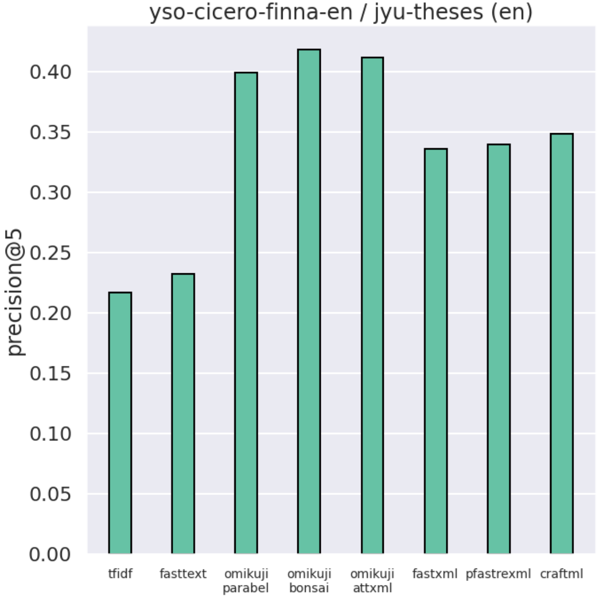

yso-cicero-finna-sv / kirjaesittelyt (sv)

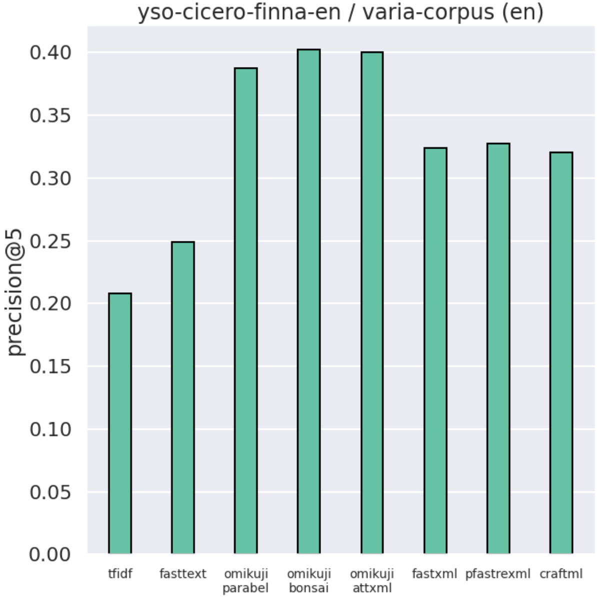

yso-cicero-finna-sv / jyu-theses (sv)


yso-cicero-finna-sv / varia-corpus (sv)

Some results: English


yso-cicero-finna-en / kirjaesittelyt (en)


yso-cicero-finna-en / jyu-theses (en)


yso-cicero-finna-en / varia-corpus (en)

CSC

Mats Sjöberg

mats.sjoberg@csc.fi

http://staff.csc.fi/msjoberg/

facebook.com/CSCfi

twitter.com/CSCfi

linkedin.com/company/csc---it-center-for-science

github.com/CSCfi

Kuvat CSC:n arkisto, Adobe Stock ja Thinkstock