



Wikification

Machine Learning WS16/17

Thipeesan Thirunavukkarasu, Dennis Bingenheimer



Agenda

1. Idee
2. Architektur Wikification
3. Konzept Wikification
4. Kategorisierung
5. Messung
6. Fazit
7. Demo



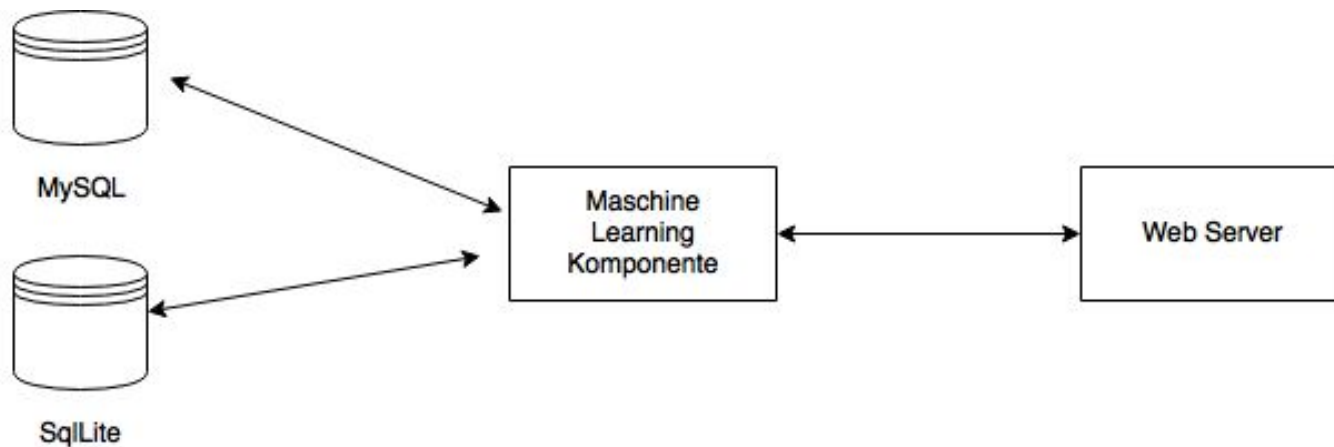
Idee



Architektur Wikification



Architektur



Speicher Größe: MySQL: 20 GB SQLite: 25GB



Datenquellen

- PetScan
- Beschaffung der HTML-Dateien über Wikipedia-API für Python
- 15.906 Sport-Artikel:
 - 14.231 wurde zum Training verwendet (89,5%)
 - 1.675 wurden für den Test verwendet (10,5%)
- 15.296 Information-System-Artikel
 - 13.021 Artikel wurden für das Training verwendet (85%)
 - 2.275 Artikel wurden für den Test verwendet (15%)
- categorie und categorie_link SQL dump



Konzept Wikification



Commonness

- Extraktion der Links aus den HTML-Dateien
- Filterung der Links auf sinnvolle Inhalte
- Zählen des Auftretens der verschiedenen Links für die einzelnen Wörter

`<a href="/wiki/Baseball_positioning" title="Baseball positioning">defensive positioning</a>`



Link Probability

- Extraktion der reinen Texte aus den HTML-Dateien
- Aufbereitung der Texte
 - Ersetzen von Sonderzeichen
 - Bigrams
 - Stemming
- Erstellung der Bag of Words über alle Trainingsdaten

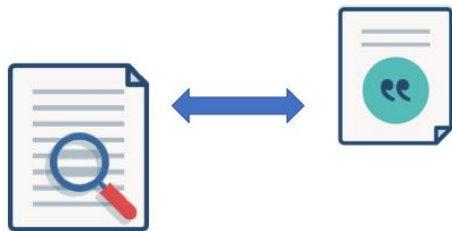
Berechnung:

$\#(\text{Auftreten des Wortes als Link})$

$\#(\text{Auftreten des Wortes in den Texten})$



Relatedness



- semantische relatedness
- eindeutige und mehrdeutige link kandidaten werden betrachtet

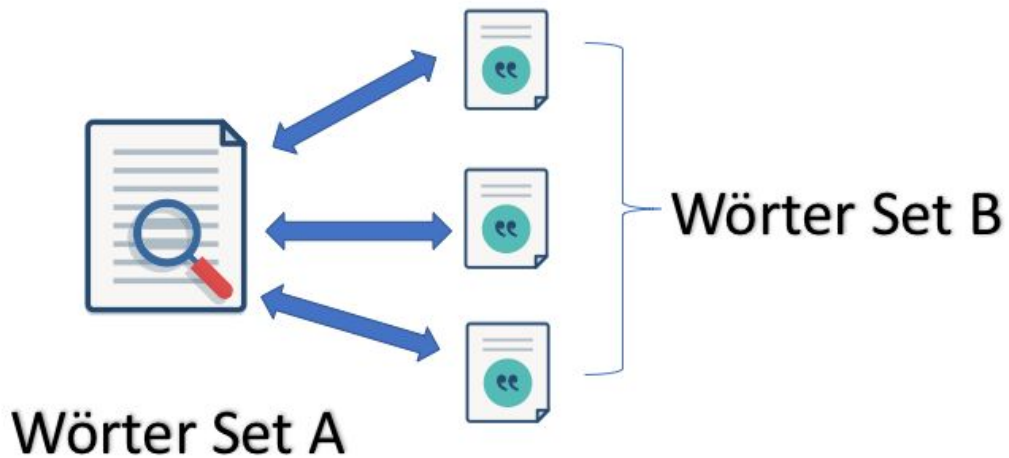
Berechnung:

$\#(\text{für alle Wörter } \mathbf{Ausgangsdatei} - \text{ in } \mathbf{Link datei})$

$\#(\text{ alle Wörter } \mathbf{Ausgangsdatei})$



Content Quality



Berechnung:

#(Wörter Set A_in Wörter Set B)

#(Wörter Set A)



Wikification Classifier

- Link Detection:
 - Feste Schwelle von 10% Link Probability
- Disambiguation Classifier:
 - Logistic Regression Classifier
 - Samples:

Feature-Nr	Commoness	Releatedness	Content Quality	Label
1	0.9	0.1	0.2	1
2	0.3	0.8	0.2	0
3	0.8	0.7	0.6	1
...				



Kategorisierung



Umsetzung

categorie dumb sql

SPALTEN : cat_title (string) cat_subcats (int)

categorie_links dumb sql

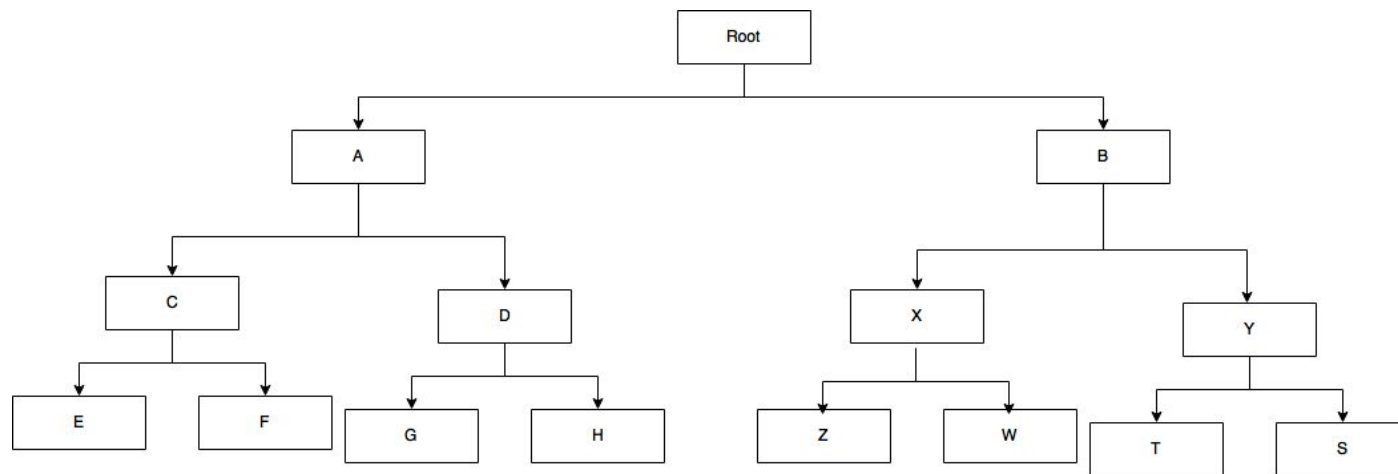
SPALTEN : cL_from (int) cl_to (string)

Info: Kategorien (ca 1,5 Mio) ,

Kategorien ohne Subkategorien (ca 700 000)



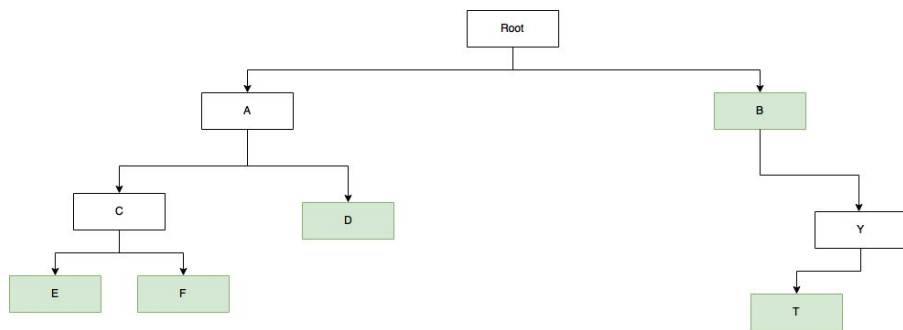
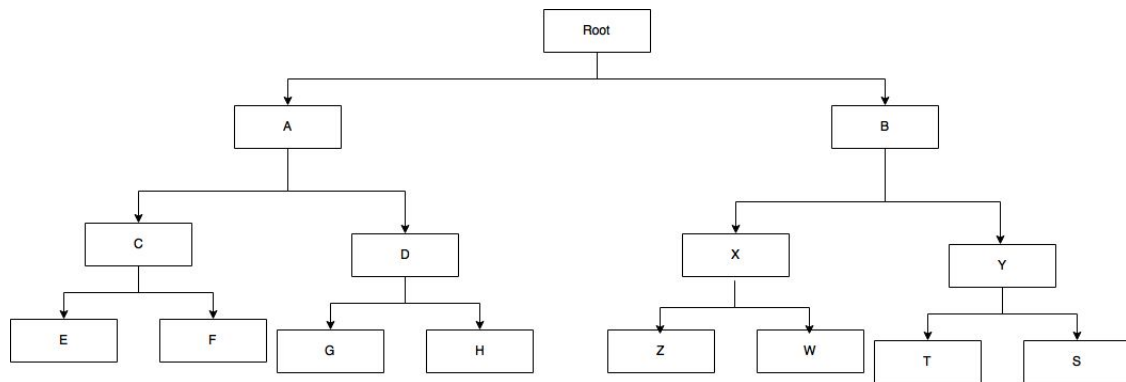
Umsetzung



Info: Baum basiert auf 130 Subkategorien



Umsetzung





Messung



Eckdaten

- Berechnung der Kennzahlen ausschließlich auf den Trainingsdaten der Sport-Dateien
- Training des Disambiguation Classifiers auf 30.000 Samples



Ergebnisse (Training auf Sportdaten)

	Trainingsdaten		Testdaten	
Commonness-beschränkung	25	50	25	50
Wikifictionzeit	8 min	14 min	4 min	6 min
Precision	88%	88%	83%	84%
Recall	55%	56%	44%	45%



Eckdaten

- Berechnung der Kennzahlen auf den gesamten Trainingsdaten der Sport- sowie Information-Systems-Dateien
- Training des Disambiguation Classifiers auf 100.000 Samples



Ergebnisse (Training auf allen Daten)

	Trainingsdaten		Testdaten	
Commonness-beschränkung	25	50	25	50
Wikifictionzeit	3 min	6 min	3 min	5 min
Precision	92%	92%	89%	90%
Recall	48%	49%	46%	47%



Machine learning

From Wikipedia, the free encyclopedia

For the journal, see [Machine Learning \(journal\)](#).

Machine learning is the subfield of [computer science](#) that gives computers the ability to learn without being explicitly programmed ([Arthur Samuel](#) 1959).^[1] Evolved from the study of [pattern recognition](#) and [computational learning theory](#) in [artificial intelligence](#)^[2] machine learning explores the study and construction of [algorithms](#) that can [learn](#) from and make predictions on [data](#)^[3] – such algorithms overcome following strictly static [program instructions](#) by making data driven predictions or decisions,^{[4]:2} through building a [model](#) from sample inputs. Machine learning is employed in a range of computing tasks where designing and programming explicit [algorithms](#) is infeasible; example applications include [spam filtering](#), detection of network intruders or malicious insiders working towards a data breach,^[5] [optical character recognition](#) (OCR),^[6] [search engines](#) and [computer vision](#).

Wikification Result

All Links reference to Wikipedia Pages

[Machine learning](#) is the subfield of computer science that gives computers the ability to learn without being explicitly programmed ([Arthur Samuel](#) 1959). Evolved from the study of [pattern recognition](#) and computational [learning theory](#) in [artificial intelligence](#), machine learning explores the study and construction of [algorithms](#) that can [learn](#) from and make predictions on [data](#) – such algorithms overcome following strictly static [program instructions](#) by making data driven predictions or decisions, through building a [model](#) from sample inputs. [Machine learning](#) is employed in a range of computing tasks where designing and programming explicit [algorithms](#) is infeasible; example applications include [spam filtering](#), detection of network intruders or malicious insiders working towards a data breach, [optical character recognition](#) (OCR), [search engines](#) and [computer vision](#).



Demo



Fazit