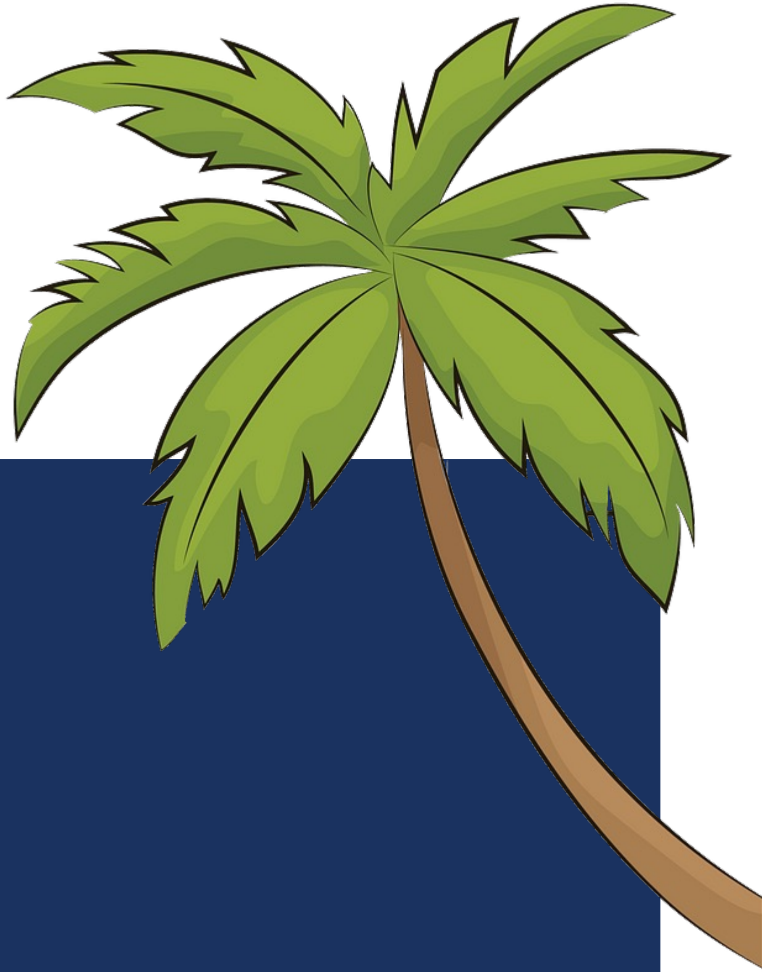
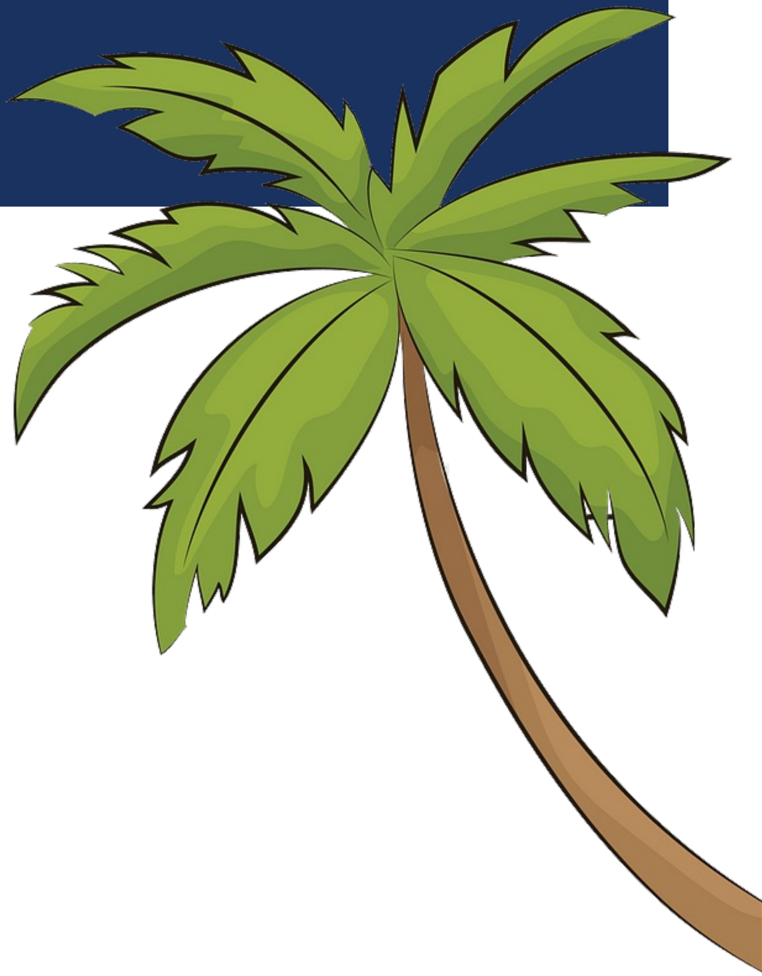
LET’S TALK ABOUT PALM LEAVES FROM MINIMAL DATA TO TEXT UNDERSTANDING

MAGNUS BENDER[1], MARCEL GEHRKE[1], TANYA BRAUN[2]

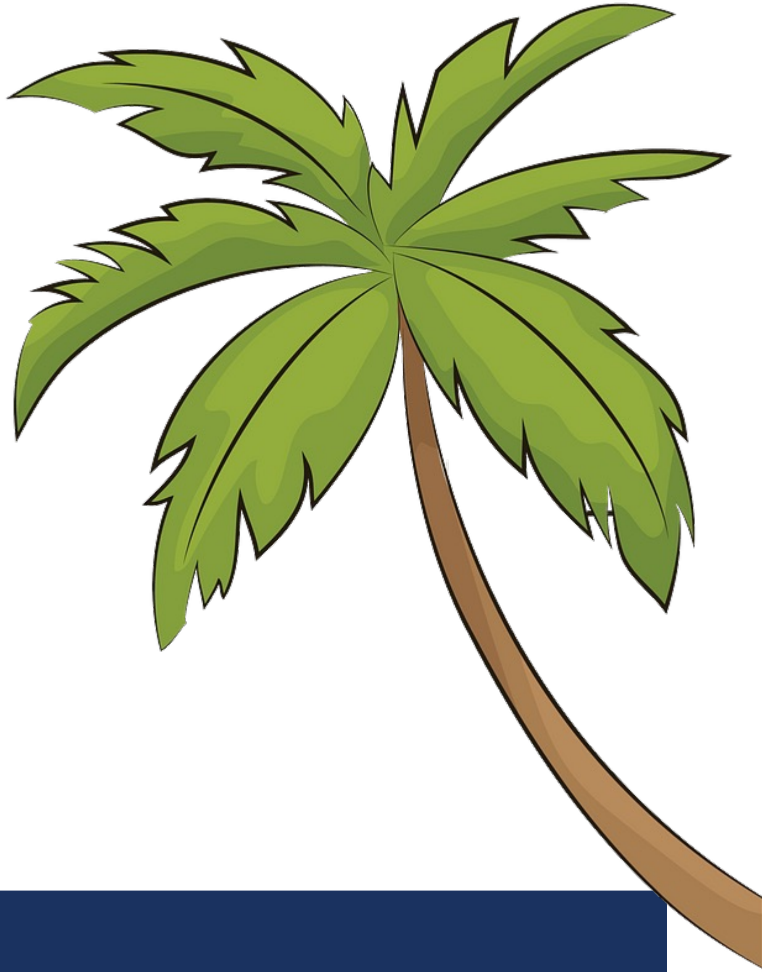


[1]Institute of Information Systems, University of Lübeck
[2]Computer Science Department, University of Münster

AGENDA



1. Introduction to Semantic Systems [Tanya]
2. Supervised Learning [Marcel]
   - Subjective content descriptions
   - Corpus enrichment
   - Inline annotations (🌴)
3. Transition to Unsupervised and Relational Learning [Magnus]
4. Summary [Tanya]

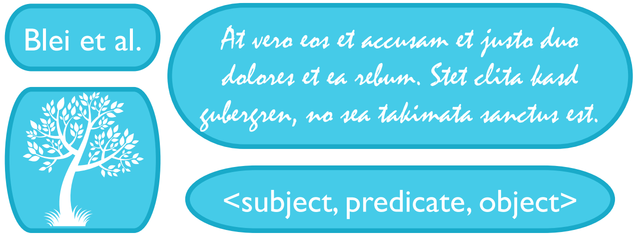


SCDS

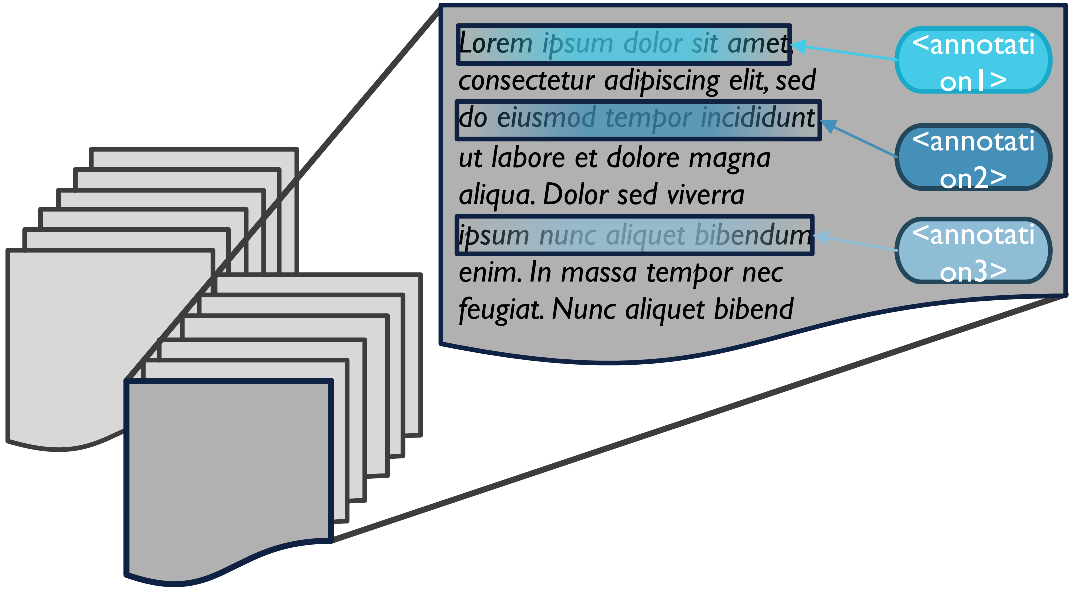
SUPERVISED LEARNING

THE SETTING: A CORPUS OF DOCUMENTS AND ANNOTATIONS

- Corpus = set of documents $\mathcal{D}$
- Each document d has a set of annotations $g(d)$
  - Annotation ≙ *subjective content description* (SCD)
  - Reflect the *context* of the purpose of the corpus
- Types of SCDs can be manifold
  - Figures, notes, references, …



- Each SCD associated with words at specific location
  - Assumption: Words closer to location, influence higher



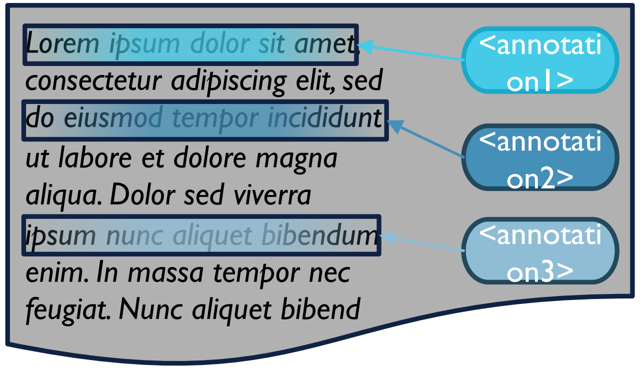
Proposition 1:
Annotations generate the words in a document

Felix Kuhr, Tanya Braun, Magnus Bender, Ralf Möller: To Extend or not to Extend? Context-specific Corpus Enrichment. Proceedings of the AI 2019: Advances in Artificial Intelligence (AI-19), 2019

CONSTRUCTING THE SCD-WORD DISTRIBUTION MATRIX

- Particular section annotated
  - Area around section (context) also crucial for annotation
- Annotations assumed to be SCDs of the text
- Construct a matrix for SCD-Word Distribution
- Each row corresponds to an annotation (SCD) and contains the word distribution for that SCD
- Each column corresponds to a word in our corpus

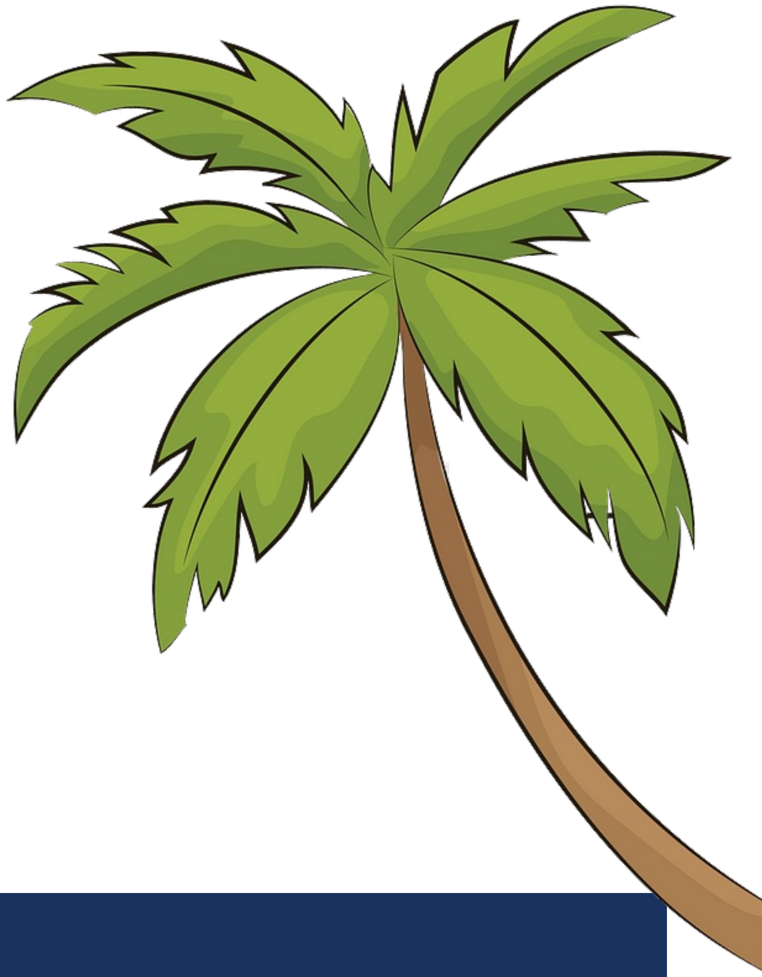
$$
\begin{array}{cc}
 & \begin{array}{ccc} w_1 & \cdots & w_n \end{array} \\
\begin{array}{c} t_1 \\ \vdots \\ t_m \end{array} & \begin{bmatrix} v_{1,1} & \cdots & v_{1,n} \\ \vdots & \ddots & \vdots \\ v_{m,1} & \cdots & v_{m,n} \end{bmatrix}
\end{array}
$$



n words in corpus

m annotations associated to corpus

Felix Kuhr, Tanya Braun, Magnus Bender, Ralf Möller: To Extend or not to Extend? Context-specific Corpus Enrichment. Proceedings of the AI 2019: Advances in Artificial Intelligence (AI-19), 2019



CORPUS ENRICHMENT

SUPERVISED LEARNING

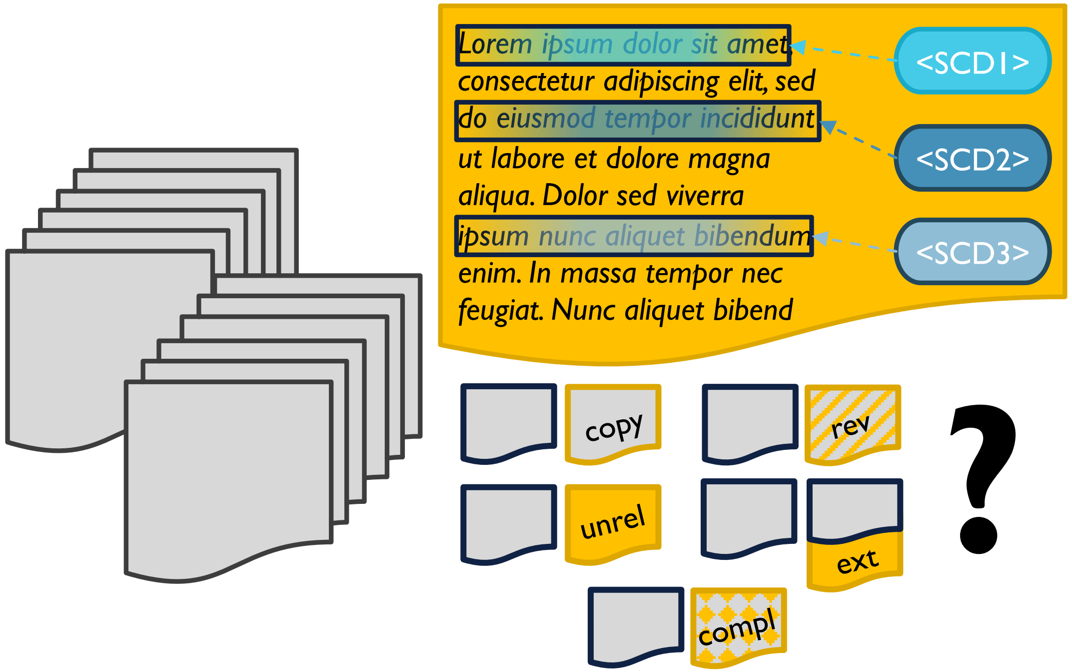
CORPUS ENRICHMENT:
TASK

- An important aspect:

  Well-rounded corpus needed
  for high-quality information retrieval

→ Corpus enrichment to extend corpus with documents that provide *added* value in task context

- From system perspective: Internal task
- A classification problem
  - Input: new document d, corpus $\mathcal{D}$
  - Possible classes?
    - Quasi-copy, *revision*, *extension*, unrelated, *complementary*?



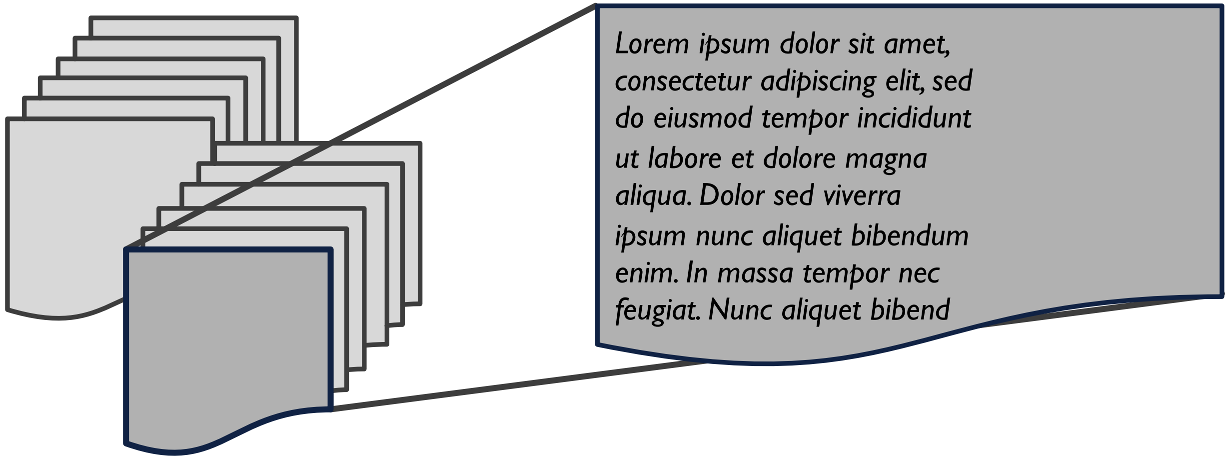
Felix Kuhr, Tanya Braun, Magnus Bender, Ralf Möller: To Extend or not to Extend? Context-specific Corpus Enrichment. Proceedings of the AI 2019: Advances in Artificial Intelligence (AI-19), 2019
Felix Kuhr, Tanya Braun, Ralf Möller: Augmenting and Automating Corpus Enrichment. Proceedings of the 14th IEEE International Conference on Semantic Computing (ICSC-20), 2020
Felix Kuhr, Magnus Bender, Tanya Braun, Ralf Möller: Augmenting and Automating Corpus Enrichment. Int. J. Semantic Computing, 2020

WHEN TO ADD A NEW DOCUMENT TO THE CORPUS?

Extend a corpus with a new document
only if the document

provides additional data relevant for a
given task, i.e., adds value in a given context.

- Make decision based on
  - words, BUT: not context-specific
  - topics, BUT: possibly inconclusive
  - annotations?



Felix Kuhr, Tanya Braun, Magnus Bender, Ralf Möller: To Extend or not to Extend? Context-specific Corpus Enrichment. Proceedings of the AI 2019: Advances in Artificial Intelligence (AI-19), 2019
Felix Kuhr, Tanya Braun, Ralf Möller: Augmenting and Automating Corpus Enrichment. Proceedings of the 14th IEEE International Conference on Semantic Computing (ICSC-20), 2020
Felix Kuhr, Magnus Bender, Tanya Braun, Ralf Möller: Augmenting and Automating Corpus Enrichment. Int. J. Semantic Computing, 2020

WHEN DOES A NEW DOCUMENT PROVIDE NEW INSIGHTS

- SCDs reflect the context of the annotated area
  - Decide to extend the corpus based on how much of the new document can be gerated given SCDs in corpus

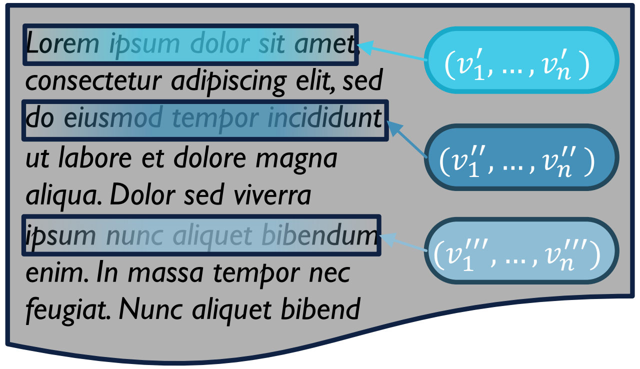
- Based on answer to how much is generated with high probability: decide extension (IN/OUT)
  - Generate large part with high probability: OUT (→ known).
  - Probability low: OUT (→ unrelated).
  - Generate only some parts with high probability: IN (→ extension).

Felix Kuhr, Tanya Braun, Magnus Bender, Ralf Möller: To Extend or not to Extend? Context-specific Corpus Enrichment. Proceedings of the AI 2019: Advances in Artificial Intelligence (AI-19), 2019
Felix Kuhr, Tanya Braun, Ralf Möller: Augmenting and Automating Corpus Enrichment. Proceedings of the 14th IEEE International Conference on Semantic Computing (ICSC-20), 2020
Felix Kuhr, Magnus Bender, Tanya Braun, Ralf Möller: Augmenting and Automating Corpus Enrichment. Int. J. Semantic Computing, 2020

HOW TO COMPARE A NEW DOCUMENT AGAINST A CORPUS

- New document: for word chunks, build vector representation of the words occurring in the chunk

$$
\begin{array}{cc}
 & \begin{array}{ccc} w_1 & \cdots & w_n \end{array} \\
\begin{array}{c} t_1 \\ \vdots \\ t_m \end{array} & \begin{bmatrix} v_{1,1} & \cdots & v_{1,n} \\ \vdots & \ddots & \vdots \\ v_{m,1} & \cdots & v_{m,n} \end{bmatrix}
\end{array}
$$



- Use cosine similarity to find annotation whose vector representation is most similar to the words of a chunk:

$$\mathrm{sim}(A, B) = \cos(\angle A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \cdot \|B\|}$$

- Identify most probable SCD (MPSCD)
  - Sometimes also called most probably suited SCD (MPS^2CD)

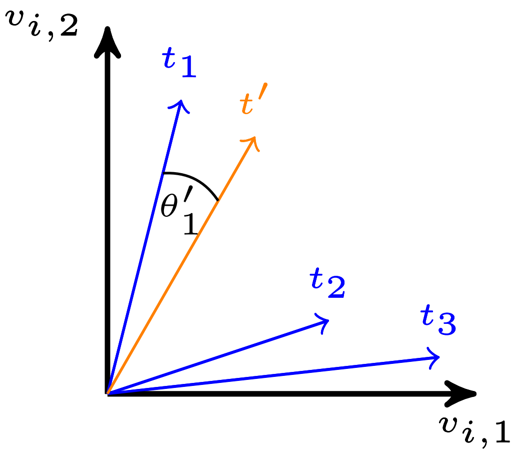
Felix Kuhr, Tanya Braun, Magnus Bender, Ralf Möller: To Extend or not to Extend? Context-specific Corpus Enrichment. Proceedings of the AI 2019: Advances in Artificial Intelligence (AI-19), 2019
Felix Kuhr, Tanya Braun, Ralf Möller: Augmenting and Automating Corpus Enrichment. Proceedings of the 14th IEEE International Conference on Semantic Computing (ICSC-20), 2020
Felix Kuhr, Magnus Bender, Tanya Braun, Ralf Möller: Augmenting and Automating Corpus Enrichment. Int. J. Semantic Computing, 2020

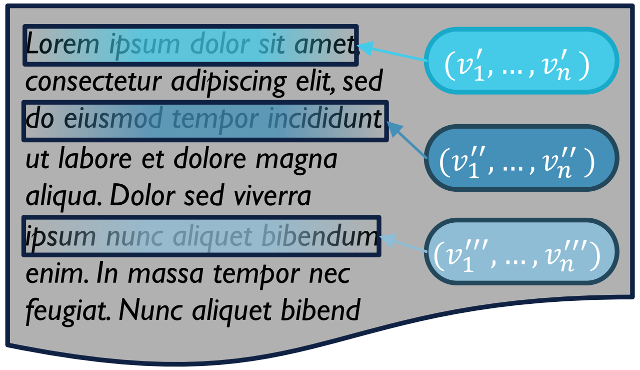
HOW TO COMPARE A NEW DOCUMENT AGAINST A CORPUS

- Simplified representation of corpus annotations t_i with two words in the vocabulary
- Representation of vector representation of word chunk t'
- Angle $\theta_1{}'$ between t_1 and t' smallest compared to t_2, t_3

→ Find t_i with smallest angle for each word chunk

Use set of t_i's for all word chunks t' in the new document and their similarities for decision

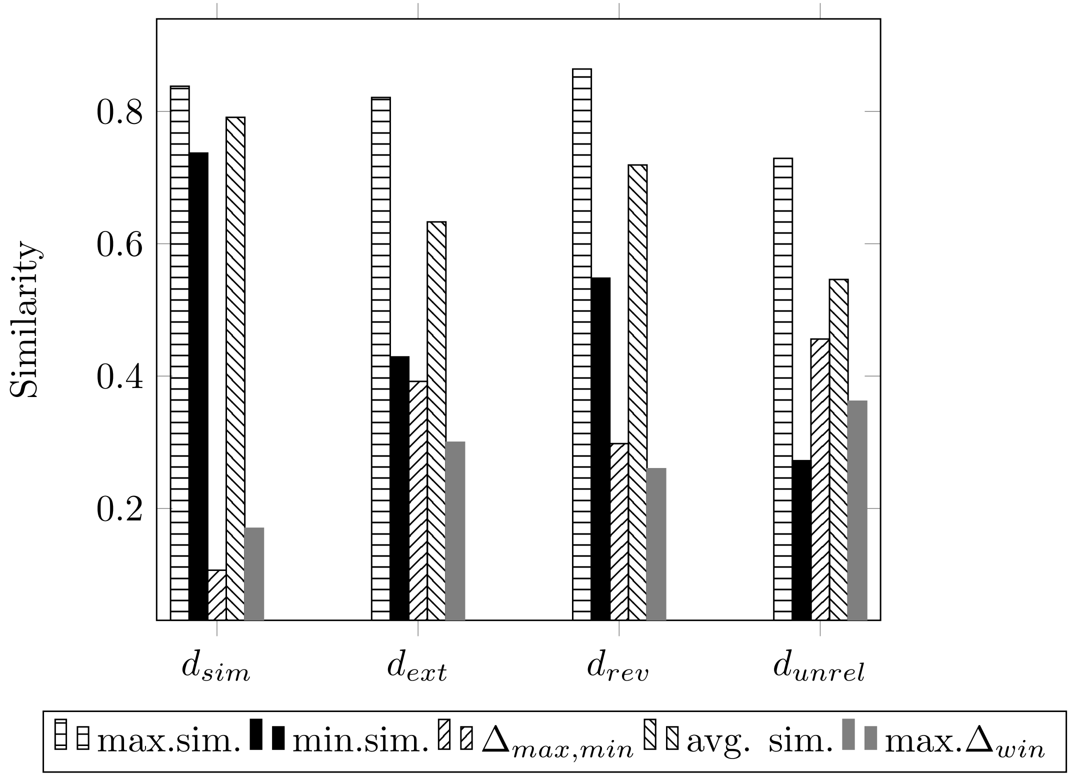




Felix Kuhr, Tanya Braun, Magnus Bender, Ralf Möller: To Extend or not to Extend? Context-specific Corpus Enrichment. Proceedings of the AI 2019: Advances in Artificial Intelligence (AI-19), 2019
Felix Kuhr, Tanya Braun, Ralf Möller: Augmenting and Automating Corpus Enrichment. Proceedings of the 14th IEEE International Conference on Semantic Computing (ICSC-20), 2020
Felix Kuhr, Magnus Bender, Tanya Braun, Ralf Möller: Augmenting and Automating Corpus Enrichment. Int. J. Semantic Computing, 2020

HOW SIMILAR ARE UNKNOWN DOCUMENTS?

- New document:
  - d_{sim}: known
  - d_{ext}: extended
  - d_{rev}: revised
  - d_{unrel}: unrelated
- Influencing factors:
  - Corpus size
  - Quality of annotations
  - Indicators
  - → No single indicator to rule them all!
  - → Limited transfer between corpora!



Felix Kuhr, Tanya Braun, Magnus Bender, Ralf Möller: To Extend or not to Extend? Context-specific Corpus Enrichment. Proceedings of the AI 2019: Advances in Artificial Intelligence (AI-19), 2019
Felix Kuhr, Tanya Braun, Ralf Möller: Augmenting and Automating Corpus Enrichment. Proceedings of the 14th IEEE International Conference on Semantic Computing (ICSC-20), 2020
Felix Kuhr, Magnus Bender, Tanya Braun, Ralf Möller: Augmenting and Automating Corpus Enrichment. Int. J. Semantic Computing, 2020

DISCRETISES MEASURES

	city corpus				president corpus			
Indicator I	d_{sim}	d_{ext}	d_{rev}	d_{unrel}	d_{sim}	d_{ext}	d_{rev}	d_{unrel}
Max Sim.	$+$	$+$	$+$	$\circ$	$+$	$+$	$+$	$\circ$
Min Sim.	$+$	$\circ$	$\circ$	$-$	$\circ$	$\circ$	$\circ$	$-$
$\Delta_{max,min}$	$-$	$\circ$	$-$	$\circ$	$-$	$\circ$	$-$	$\circ$
Avg. Sim.	$+$	$\circ$	$+$	$\circ$	$+$	$+$	$+$	$\circ$
Max.Δ_{win}	$-$	$\circ$	$-$	$\circ$	$-$	$\circ$	$-$	$\circ$

$$\text{“}+\text{”}: I \geq 0.7, \text{“}-\text{”}: I \leq 0.3, \text{“}\circ\text{”}: 0.3 < I < 0.7$$

Felix Kuhr, Tanya Braun, Magnus Bender, Ralf Möller: To Extend or not to Extend? Context-specific Corpus Enrichment. Proceedings of the AI 2019: Advances in Artificial Intelligence (AI-19), 2019
Felix Kuhr, Tanya Braun, Ralf Möller: Augmenting and Automating Corpus Enrichment. Proceedings of the 14th IEEE International Conference on Semantic Computing (ICSC-20), 2020
Felix Kuhr, Magnus Bender, Tanya Braun, Ralf Möller: Augmenting and Automating Corpus Enrichment. Int. J. Semantic Computing, 2020

IDEA: USE A HIDDEN MARKOV MODEL FOR CLASSIFICATION

- Hidden states given by $\Omega = \{s_1, \dots, s_n\}$, where $n = 2$, with state s_1 representing related content and s_2 representing unrelated content
- An observation alphabet $\Delta = \{y_1, \dots, y_m\}$, where each observation symbol represents a range of MPSCD similarity values
- A transition probability matrix A representing the probability between all possible state transitions $a_{i,j}$ between the two hidden states $s_1, s_2 \in \Omega$
- An emission probability matrix B representing the probability to emit a symbol from observation alphabet Δ for each possible hidden state in Ω
- An initial state distribution vector $\pi = \pi_0$

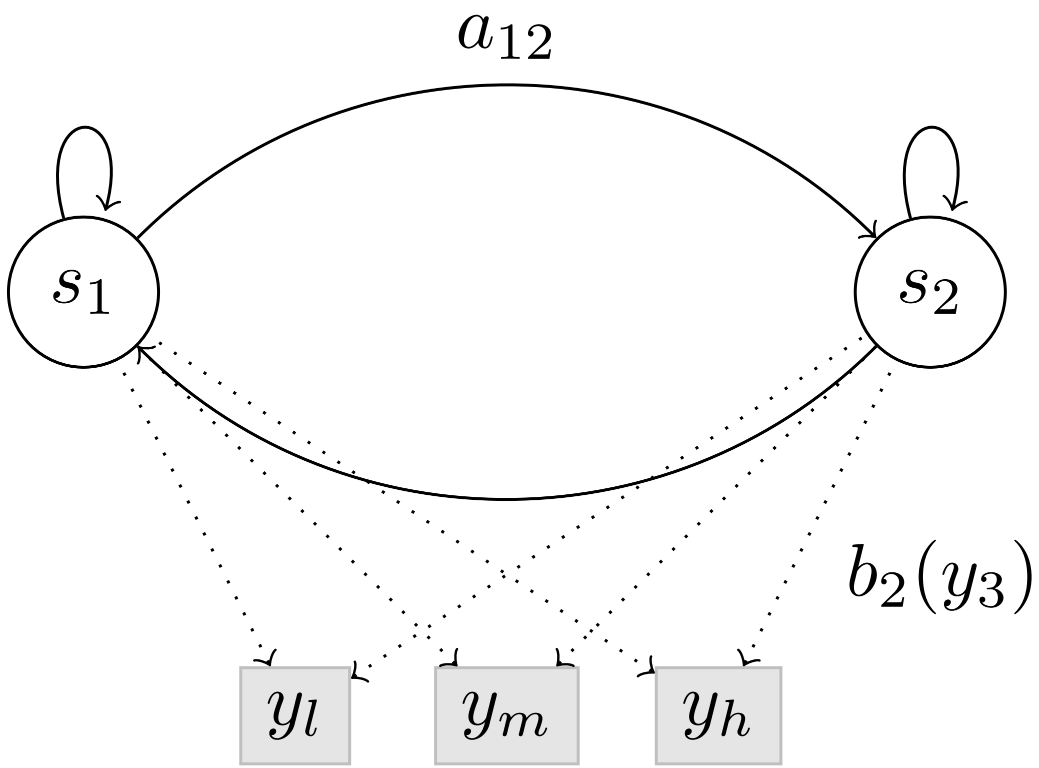
Felix Kuhr, Tanya Braun, Magnus Bender, Ralf Möller: To Extend or not to Extend? Context-specific Corpus Enrichment. Proceedings of the AI 2019: Advances in Artificial Intelligence (AI-19), 2019
Felix Kuhr, Tanya Braun, Ralf Möller: Augmenting and Automating Corpus Enrichment. Proceedings of the 14th IEEE International Conference on Semantic Computing (ICSC-20), 2020
Felix Kuhr, Magnus Bender, Tanya Braun, Ralf Möller: Augmenting and Automating Corpus Enrichment. Int. J. Semantic Computing, 2020

ENSEMBLE OF HMMS

Learn an ensemble of HMMs using Baum-Welch algorithm for:

- d_{sim}: known
- d_{ext}: extended
- d_{rev}: revised
- d_{unrel}: unrelated

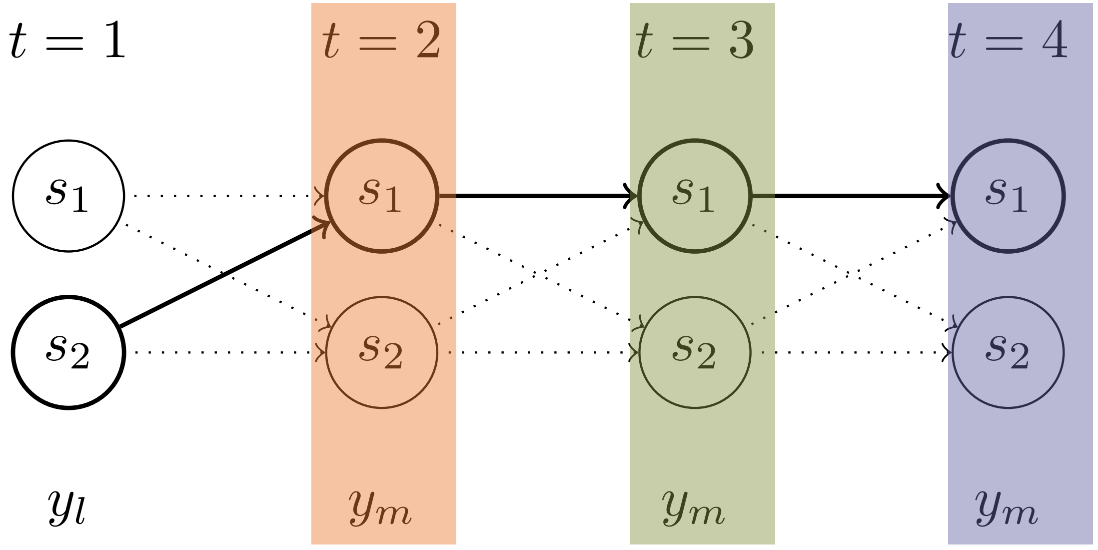
Using discretised similarity values

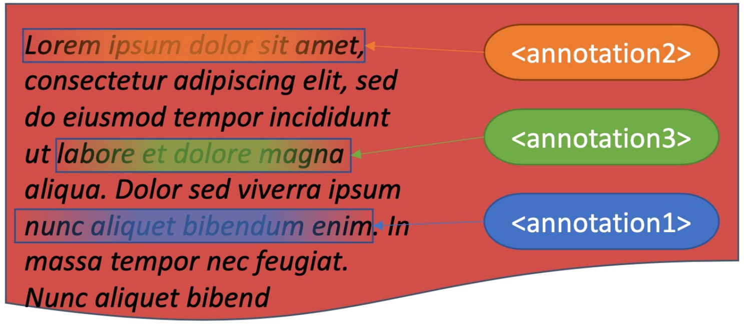


Felix Kuhr, Tanya Braun, Magnus Bender, Ralf Möller: To Extend or not to Extend? Context-specific Corpus Enrichment. Proceedings of the AI 2019: Advances in Artificial Intelligence (AI-19), 2019
Felix Kuhr, Tanya Braun, Ralf Möller: Augmenting and Automating Corpus Enrichment. Proceedings of the 14th IEEE International Conference on Semantic Computing (ICSC-20), 2020
Felix Kuhr, Magnus Bender, Tanya Braun, Ralf Möller: Augmenting and Automating Corpus Enrichment. Int. J. Semantic Computing, 2020

IDENTIFYING THE DOCUMENT TYPE

- Calculate most likely sequence of hidden states for each HMM
- Select document type from HMM with most likely sequence





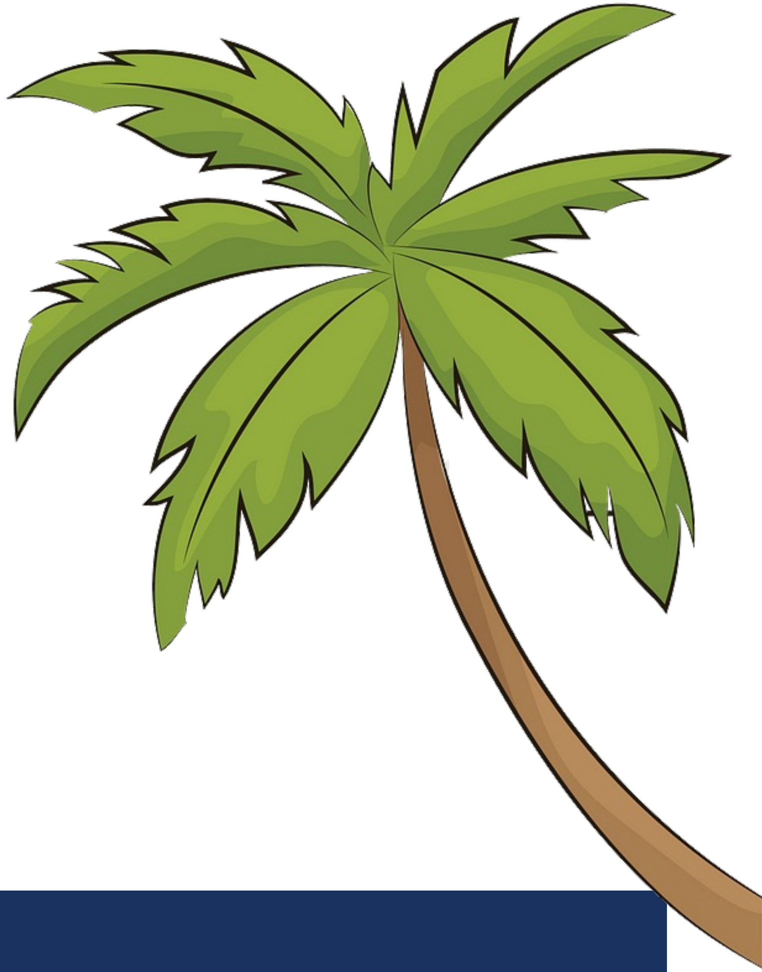
Felix Kuhr, Tanya Braun, Magnus Bender, Ralf Möller: To Extend or not to Extend? Context-specific Corpus Enrichment. Proceedings of the AI 2019: Advances in Artificial Intelligence (AI-19), 2019
Felix Kuhr, Tanya Braun, Ralf Möller: Augmenting and Automating Corpus Enrichment. Proceedings of the 14th IEEE International Conference on Semantic Computing (ICSC-20), 2020
Felix Kuhr, Magnus Bender, Tanya Braun, Ralf Möller: Augmenting and Automating Corpus Enrichment. Int. J. Semantic Computing, 2020

RESULTS

Document Type	city corpus Precision	Recall	F1-Score
d_{sim}	0.72	0.65	0.68
d_{unrel}	1.00	1.00	1.00
d_{ext}	0.93	0.86	0.89
d_{rev}	0.70	0.41	0.52

Document Type	president corpus Precision	Recall	F1-Score
d_{sim}	0.77	0.71	0.74
d_{unrel}	1.00	0.96	0.98
d_{ext}	0.91	0.84	0.87
d_{rev}	0.72	0.58	0.64

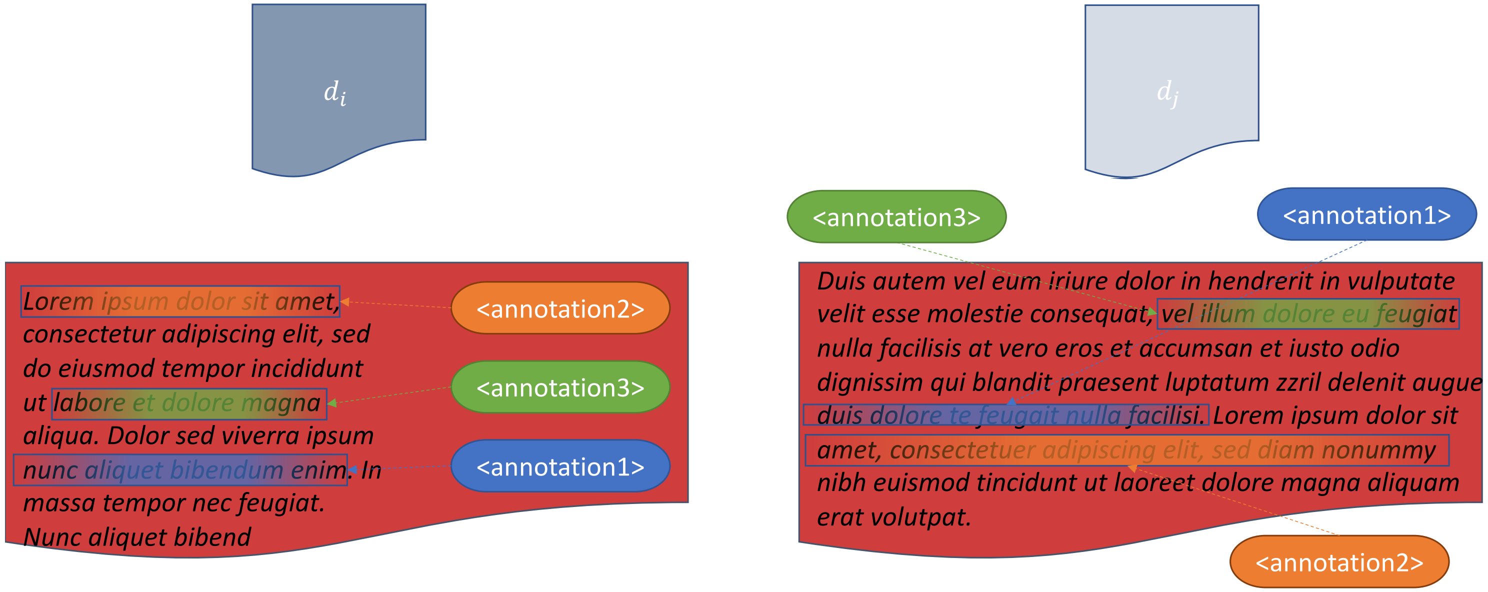
Felix Kuhr, Tanya Braun, Magnus Bender, Ralf Möller: To Extend or not to Extend? Context-specific Corpus Enrichment. Proceedings of the AI 2019: Advances in Artificial Intelligence (AI-19), 2019
Felix Kuhr, Tanya Braun, Ralf Möller: Augmenting and Automating Corpus Enrichment. Proceedings of the 14th IEEE International Conference on Semantic Computing (ICSC-20), 2020
Felix Kuhr, Magnus Bender, Tanya Braun, Ralf Möller: Augmenting and Automating Corpus Enrichment. Int. J. Semantic Computing, 2020

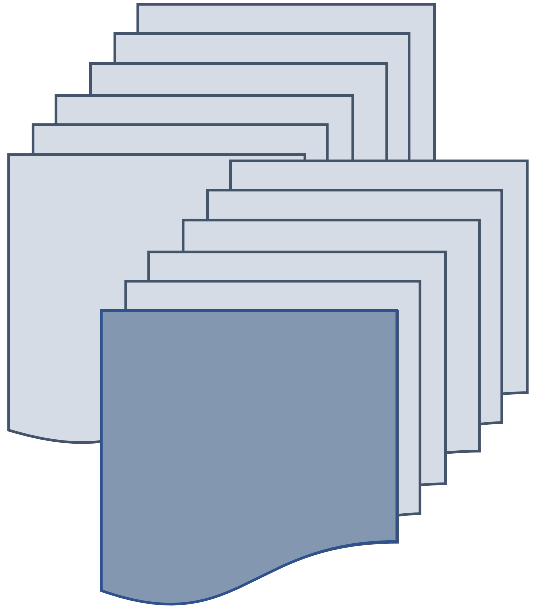


ANNOTATION ENRICHMENT

SUPERVISED LEARNING

CORPUS-DRIVEN DOCUMENT ENRICHMENT USING SCDS



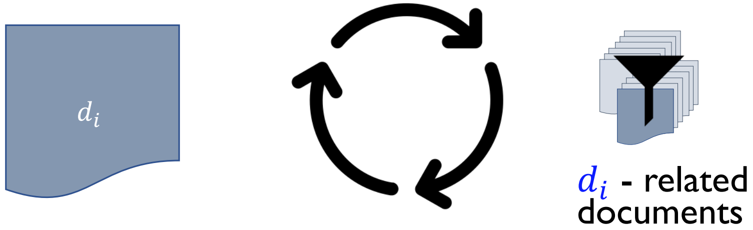


Goal: Enrich a document with *relevant* SCDs associated with other documents in an IR-corpus.

Fixed-point iteration procedure:

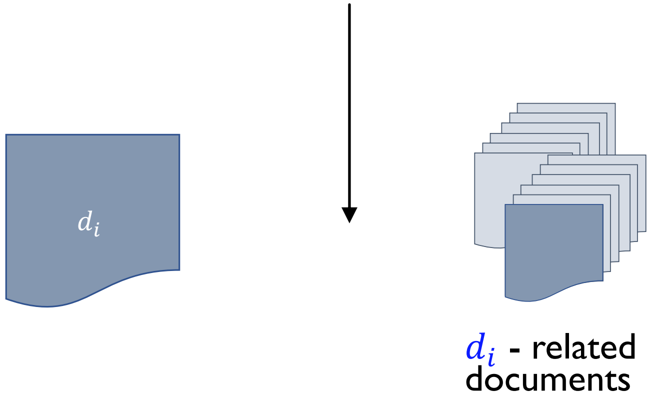
- determine the expected related documents in corpus D,
- determine the set of SCDs T from D that are newly added to d, then
- determine the expected related documents D again, and so on
- until no more SCDs are assigned to document d.

Felix Kuhr, Bjarne Witten, Ralf Möller: Corpus-Driven Annotation Enrichment. Proceedings of the 13th IEEE International Conference on Semantic Computing (ICSC-19), 2019



$related\text{-}documents(d_i, IR\text{-}corpus)$

- Subset of $IR\text{-}corpus$
  - $topic\text{-}similar$ documents whose
  - SCDs are $SCD\text{-}similar$ to d_i



$expected\text{-}relevance(t, d_i)$

- estimates relevance of t w.r.t. d by document d_i:
  - Mean topic similarity of related documents containing SCD t
  - Mean SCD similarity to related documents containing SCD t
  - Number of related documents in which SCD t occurs

$mean\text{-}expected\text{-}relevance(d_i)$

average expected relevance value of SCDs in d_i -related documents

$enrich(d_i, IR\text{-}corpus)$

- Add SCD t to d_i if

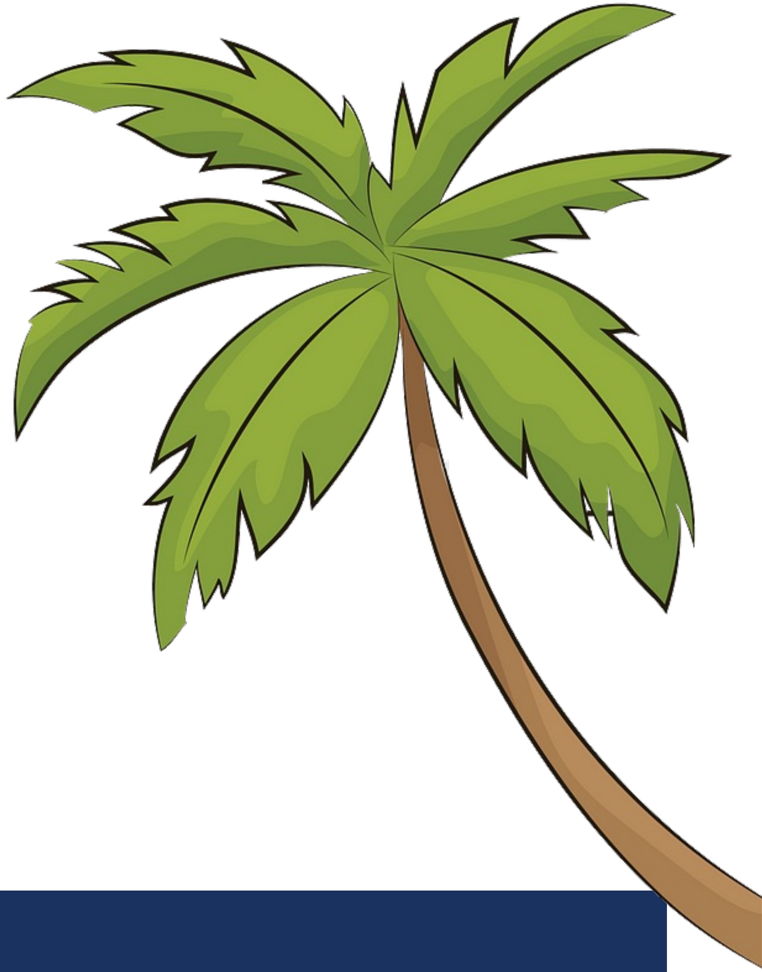
$$expected\text{-}relevance(t, d_i) > mean\text{-}expected\text{-}relevance(d_i)$$

- Iterative enrichment process
  - Related documents changes with enriched SCDs
- Terminating enrichment process
  - Value of SCD similarity of d_i to related documents increases in a negligible way

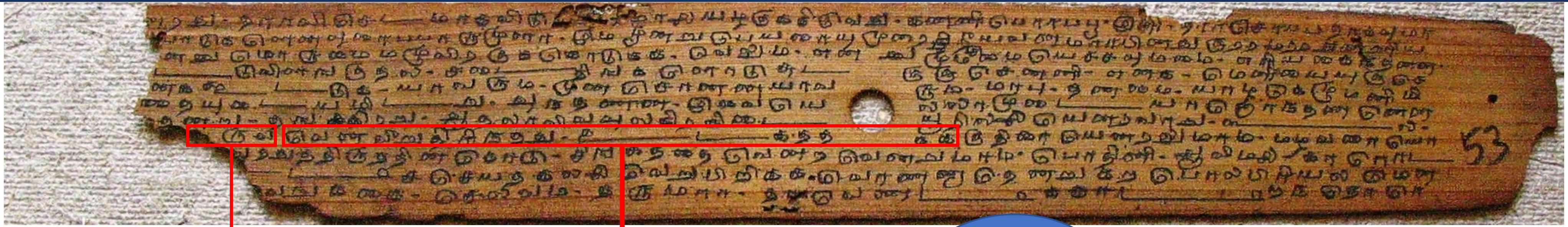
Felix Kuhr, Bjarne Witten, Ralf Möller: Corpus-Driven Annotation Enrichment. Proceedings of the 13th IEEE International Conference on Semantic Computing (ICSC-19), 2019

DETECTING IN-LINE COMMENTS

SUPERVISED LEARNING



WHAT ARE COMMENTS WITHIN A TEXT?



1 [**தைந்த** - சிதறின, மலர்ந்த -

[**tatainta** - citarina, malarnta -

2 உரு - உ]ருவ வென வீறு திரிந்தது - உட்கத்தக்க **குதிரை** யென்றலுமாம் -

uru - **u]ruva** ~ena vīru tirintatu - utka+ takka **kutirai** ~enralumām -

3 மபு…வரை யொர{l. 8}[ட்டிய 3 …விநெடுவேள் - குறுநில மன்னன் -

Not visible on palm-leaf

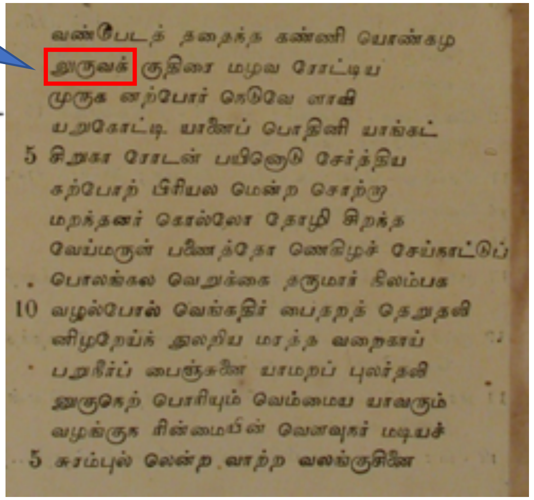
Second row in original poem

Bold: part of original poem

{l .8}: text available on palm-leaf up to this position

Comment about content (SCD)

Second row in original poem

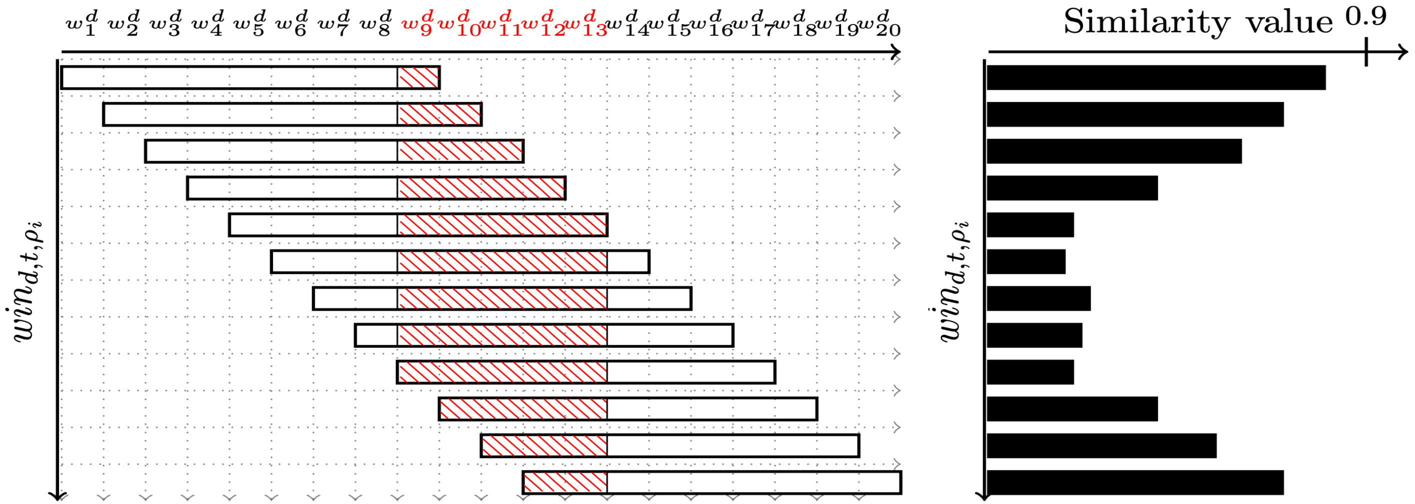


Magnus Bender, Tanya Braun, Marcel Gehrke, Felix Kuhr, Ralf Möller, Simon Schiff: Identifying Subjective Content Descriptions Among Texts. Proceedings of the 15th IEEE International Conference on Semantic Computing (ICSC-21), 2021

Magnus Bender, Tanya Braun, Marcel Gehrke, Felix Kuhr, Ralf Möller, Simon Schiff : Identifying and Translating Subjective Content Descriptions Among Texts. Int. J. Semantic Computing, 2020

CAN WE USE SIMILARITIES IN WORD COOCCURRENCES FOR EVEN MORE?

- An agent does not know which subsequences of words are content and which are iSCDs for a document $d = (w_1^d, \ldots, w_D^d), w_i^d \in (\mathcal{V}_\mathcal{D} \cup \mathcal{V}_{g(\mathcal{D})})$
  - Document d belongs to the same context as $\mathcal{D}$
  - Vocabulary $\mathcal{V}_\mathcal{D}$ or the words occurring together in a window of an associated SCD differ from vocabulary $\mathcal{V}_{g(\mathcal{D})}$ or the words occurring together in the SCD
- Identify the iSCDs



Magnus Bender, Tanya Braun, Marcel Gehrke, Felix Kuhr, Ralf Möller, Simon Schiff: Identifying Subjective Content Descriptions Among Texts. Proceedings of the 15th IEEE International Conference on Semantic Computing (ICSC-21), 2021

Magnus Bender, Tanya Braun, Marcel Gehrke, Felix Kuhr, Ralf Möller, Simon Schiff : Identifying and Translating Subjective Content Descriptions Among Texts. Int. J. Semantic Computing, 2020

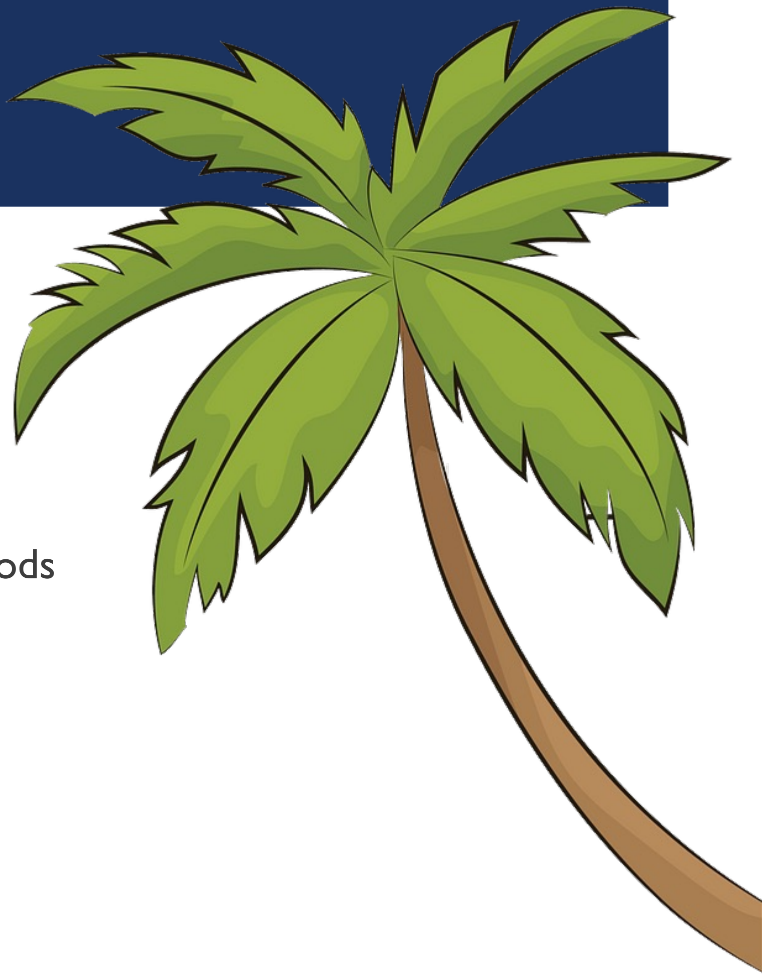
ESTIMATING ISCDS USING MPSCDS

- Given: SCD word distribution, trained HMM to detect *inline* SCDs in text
- Estimate iSCDs by using HMMs and analyse sequence of corresponding SCD similarity values (MPSCD)
  - Small similarity values – different content – possibly new SCDs in text
  - New SCD = Content of window
  - New SCDs represent new row in SCD word matrix
- Apply Viterbi on the HMM given the text
  - Obtain most likely sequence of content and comment

Magnus Bender, Tanya Braun, Marcel Gehrke, Felix Kuhr, Ralf Möller, Simon Schiff: Identifying Subjective Content Descriptions Among Texts. Proceedings of the 15th IEEE International Conference on Semantic Computing (ICSC-21), 2021

Magnus Bender, Tanya Braun, Marcel Gehrke, Felix Kuhr, Ralf Möller, Simon Schiff : Identifying and Translating Subjective Content Descriptions Among Texts. Int. J. Semantic Computing, 2020

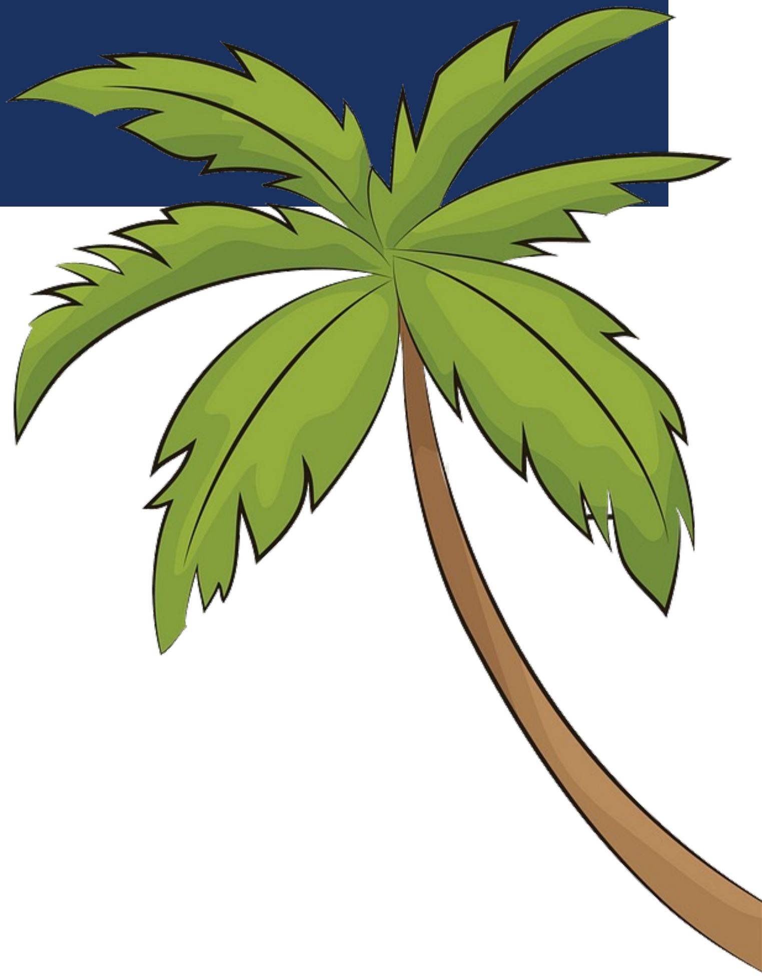
INTERIM SUMMARY



Information retrieval having only minimal data

- Annotations help to guide the search
- Annotations generate the text around the annotation
  - Using this assumption, we can tackle the following challenges with well established methods
- Enrich corpus
  - Should we add a new document to our corpus?
  - Can we enrich our corpus?
- Detecting switches between content and comments

AGENDA



1. Introduction to Semantic Systems [Tanya]
2. Supervised Learning [Marcel]
   - Subjective content descriptions
   - Corpus enrichment
   - Inline annotations (🌴)
3. Transition to Unsupervised and Relational Learning [Magnus]
4. Summary [Tanya]