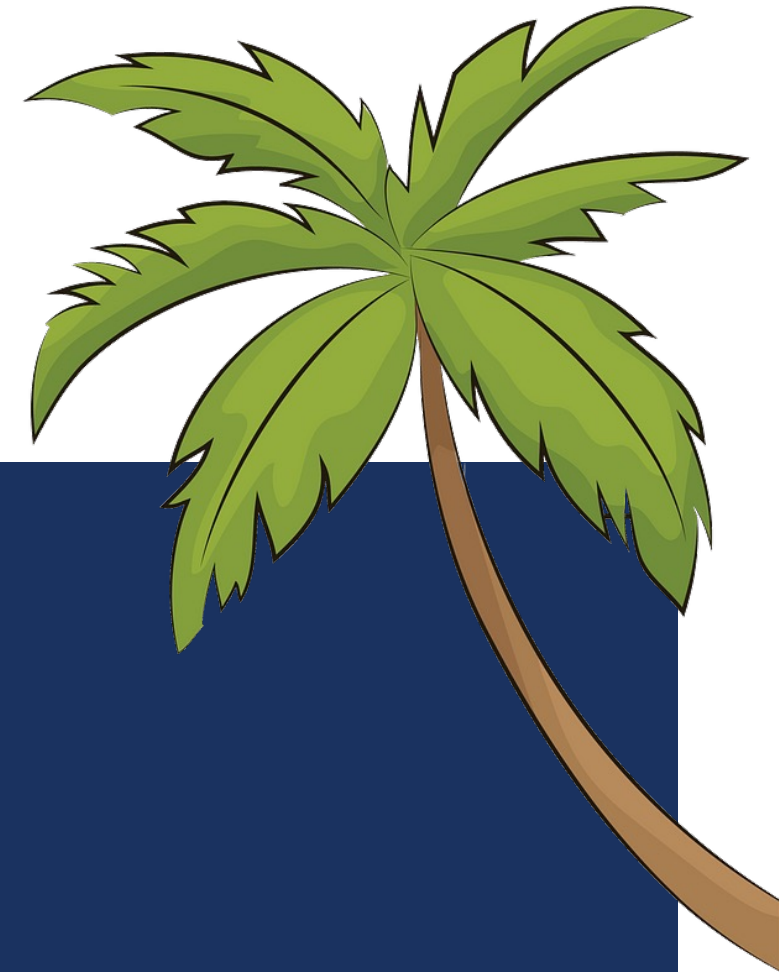


LET'S TALK ABOUT PALM LEAVES FROM MINIMAL DATA TO TEXT UNDERSTANDING

MAGNUS BENDER¹, MARCEL GEHRKE¹, TANYA BRAUN²



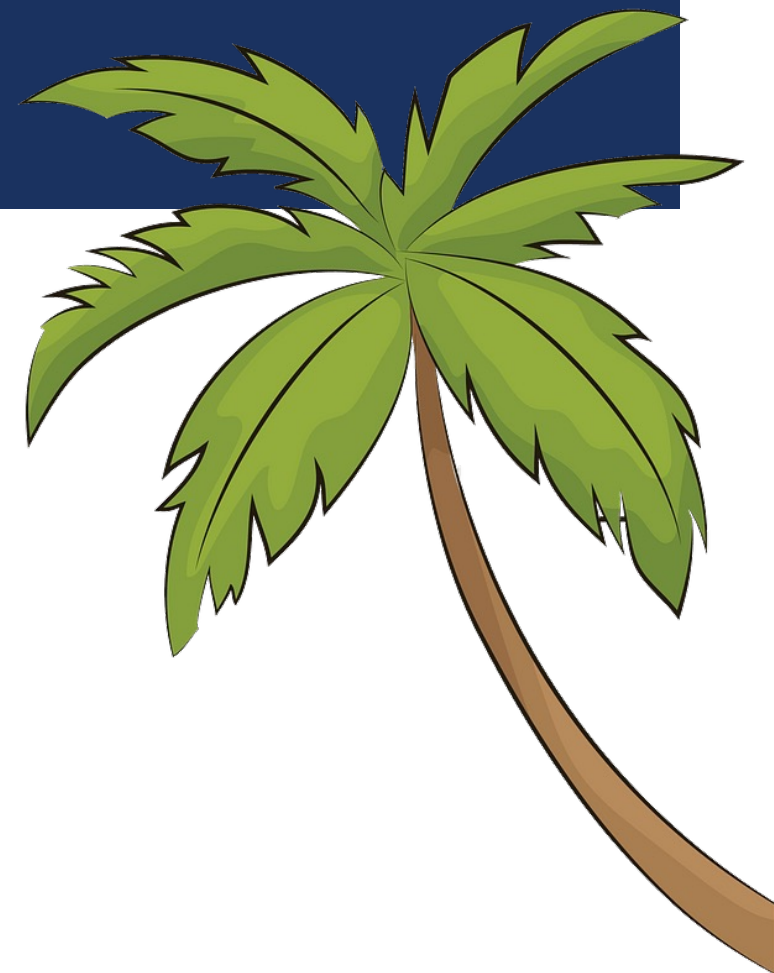
UNIVERSITÄT ZU LÜBECK

¹Institute of Information Systems, University of Lübeck
²Computer 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]



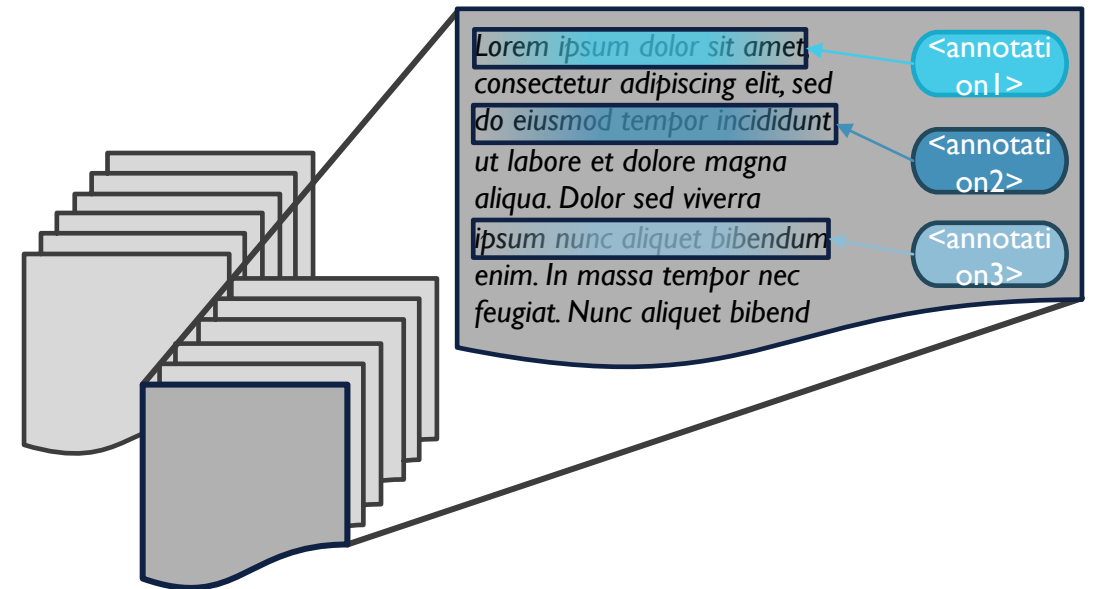
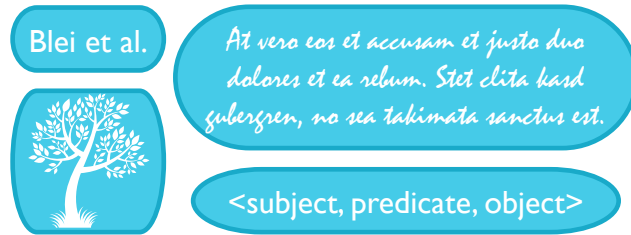
SCDS

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 \triangleq *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

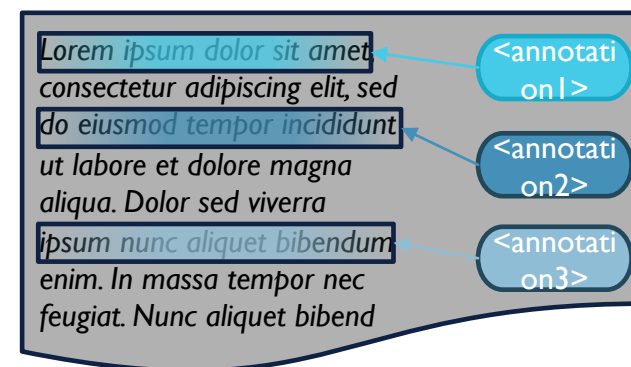


Proposition 1:
Annotations generate the words in a document

CONSTRUCTING THE SCD-WORD DISTRIBUTION MATRIX

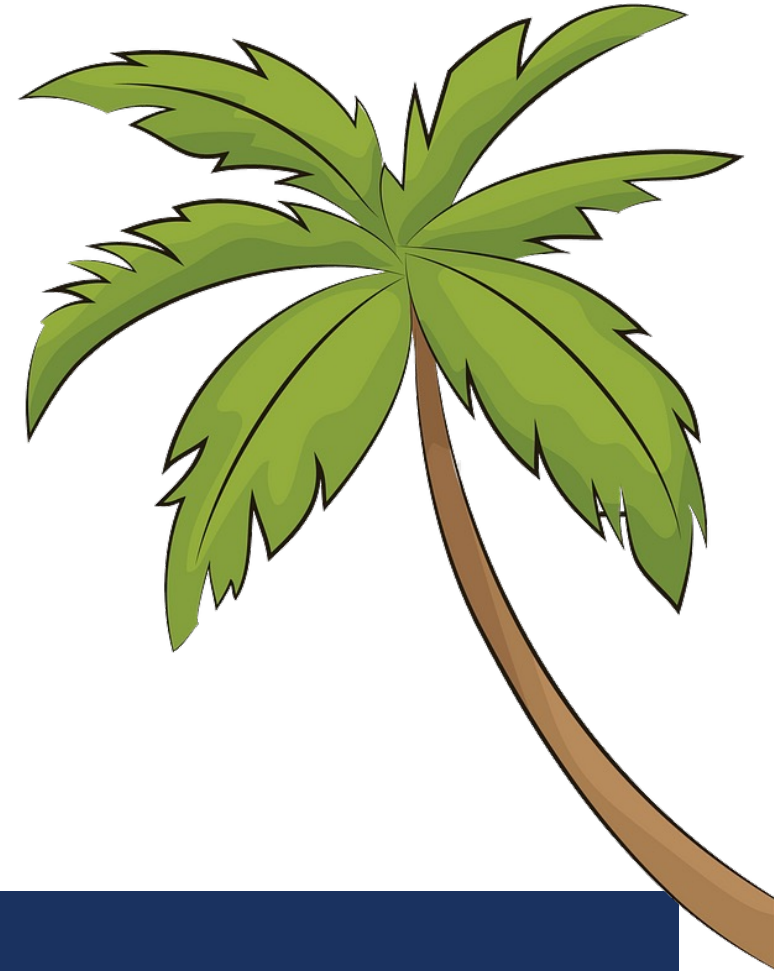
- Particular section annotated
 - Area around section (context) also crucial for annotation
- Annotations assumed to be SCDs of the text
- Construct a matrix for SCD-Word Distribution
- Each row corresponds to an annotation (SCD) and contains the word distribution for that SCD
- Each column corresponds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

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



CORPUS ENRICHMENT

SUPERVISED LEARNING



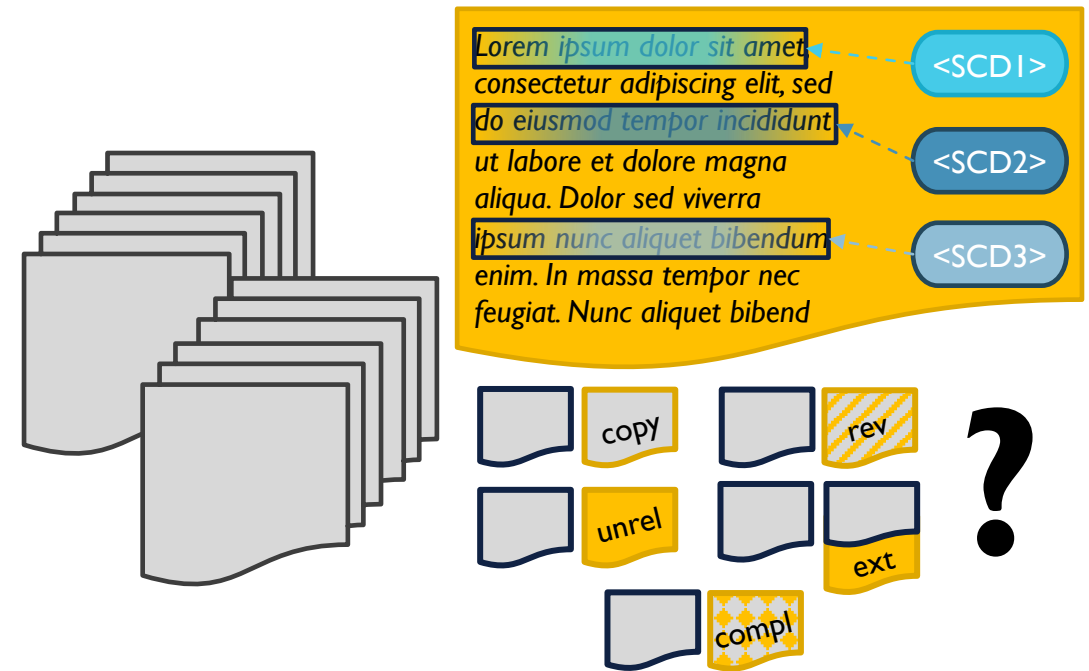
CORPUS ENRICHMENT: TASK

- Another 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?

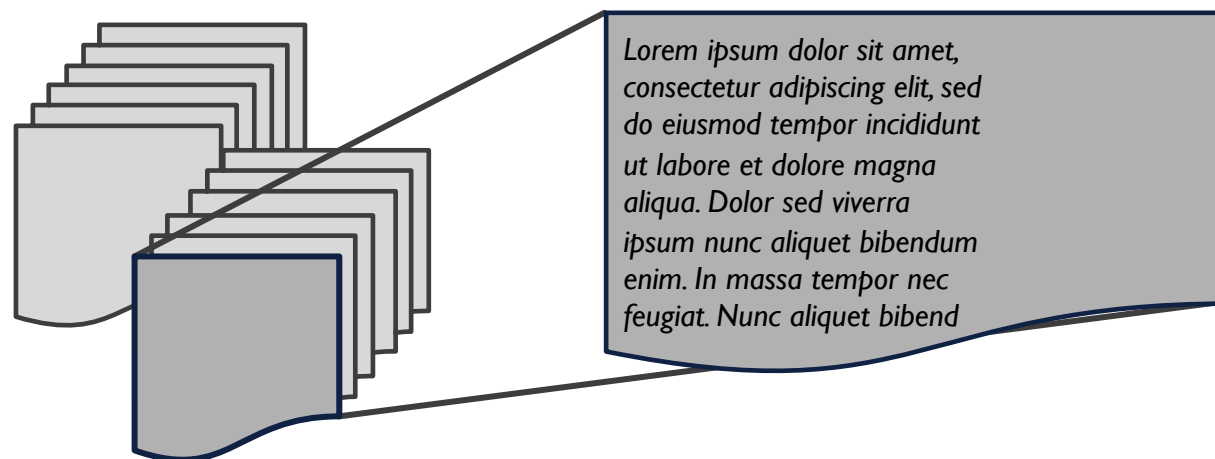


WHEN TO ADD A NEW DOCUMENT TO THE CORPUS?

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?



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 generated 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).

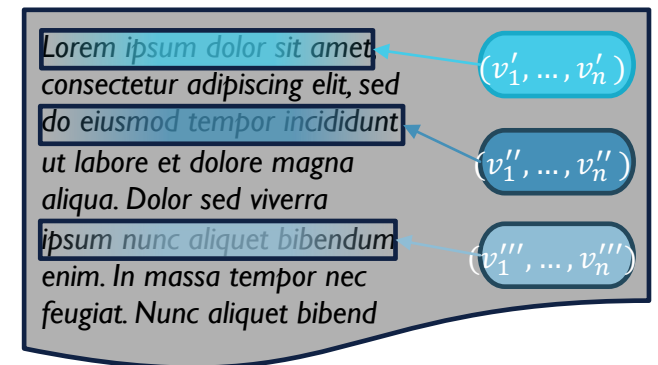
HOW TO COMPARE A NEW DOCUMENT AGAINST A CORPUS

- New document: for word chunks, build vector representation of the words occurring in the chunk
- Use cosine similarity to find annotation whose vector representation is most similar to the words of a chunk:

$$\begin{matrix} & w_1 & \dots & w_n \\ t_1 & [v_{1,1} & \dots & v_{1,n}] \\ \vdots & \vdots & \ddots & \vdots \\ t_m & [v_{m,1} & \dots & v_{m,n}] \end{matrix}$$

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

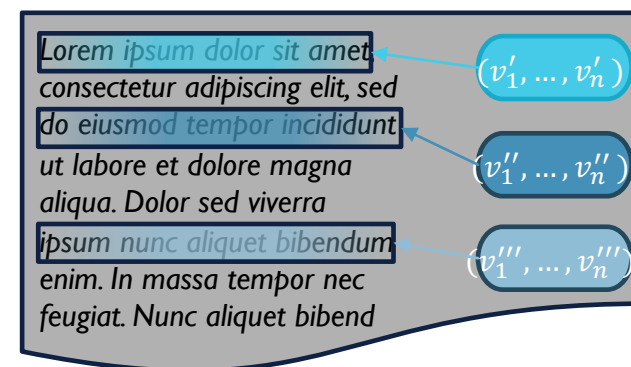
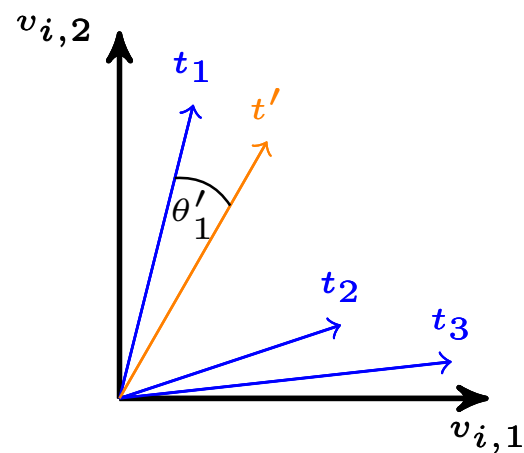
- Identify most probable SCD (MPSCD)
 - Sometimes also called most probably suited SCD (MPS²CD)



HOW TO COMPARE A NEW DOCUMENT AGAINST A CORPUS

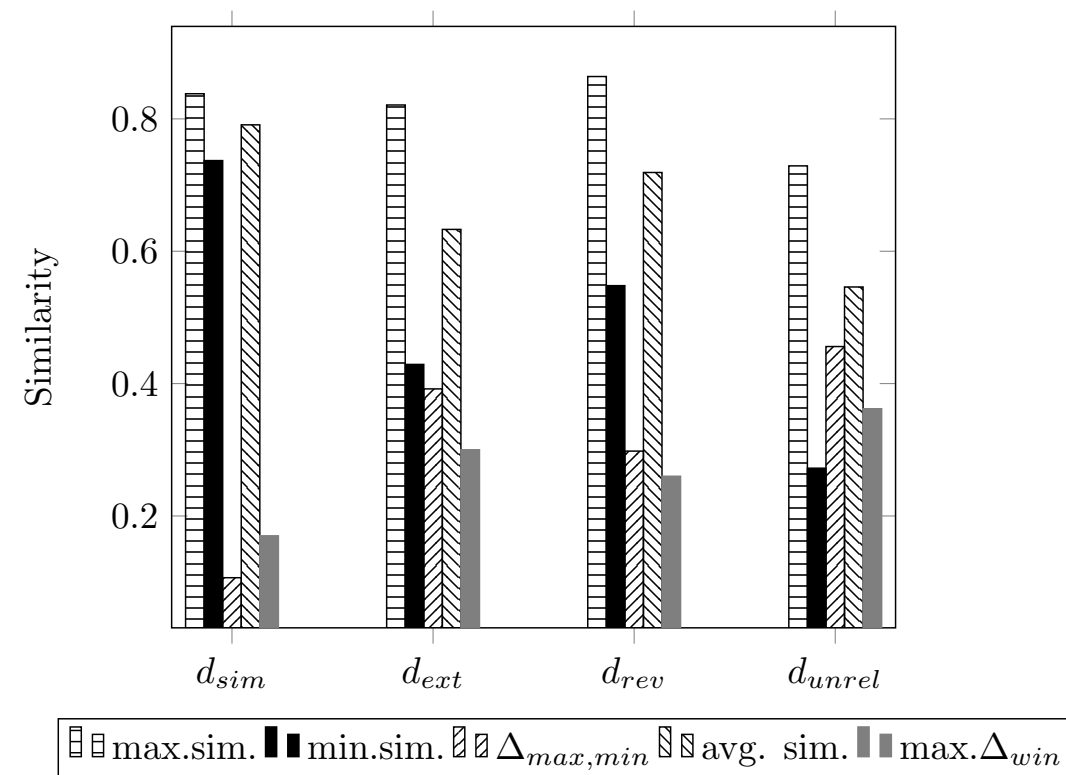
- Simplified representation of corpus annotations t_i with two words in the vocabulary
 - Representation of vector representation of word chunk t'
 - Angle θ'_1 between t_1 and t' smallest compared to t_2, t_3
- Find t_i with smallest angle for each word chunk

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



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!



DISCRETISES MEASURES

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

“+”: $I \geq 0.7$, “-”: $I \leq 0.3$, “○”: $0.3 < I < 0.7$

IDEA: USE A HIDDEN MARKOV MODEL FOR CLASSIFICATION

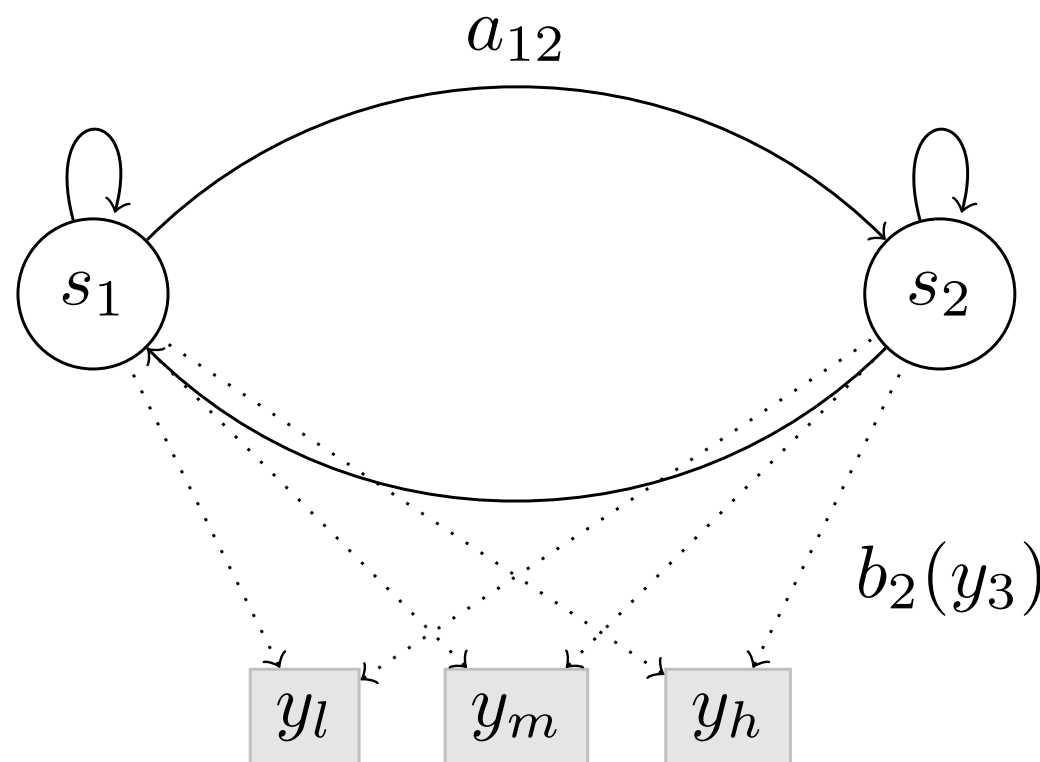
- Hidden states given by $\Omega = \{s_1, \dots, s_n\}$, where $n = 2$, with state s_1 representing related content and s_2 representing unrelated content
- An observation alphabet $\Delta = \{y_1, \dots, y_m\}$, where each observation symbol represents a range of MPSCD similarity values
- A transition 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 $\pi = \pi_0$

ENSEMBLE OF HMMS

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

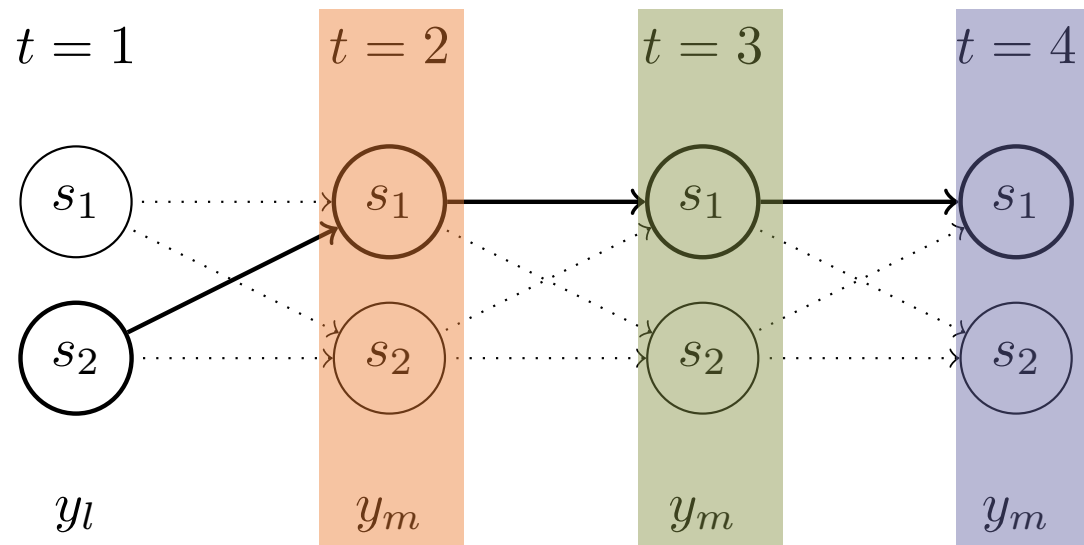
- d_{sim} : known
- d_{ext} : extended
- d_{rev} : revised
- d_{unrel} : unrelated

Using discretised similarity values



IDENTIFYING THE DOCUMENT TYPE

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



>Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Dolor sed viverra ipsum nunc aliquet bibendum enim. In massa tempor nec feugiat. Nunc aliquet bibend

<annotation2>

<annotation3>

<annotation1>

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

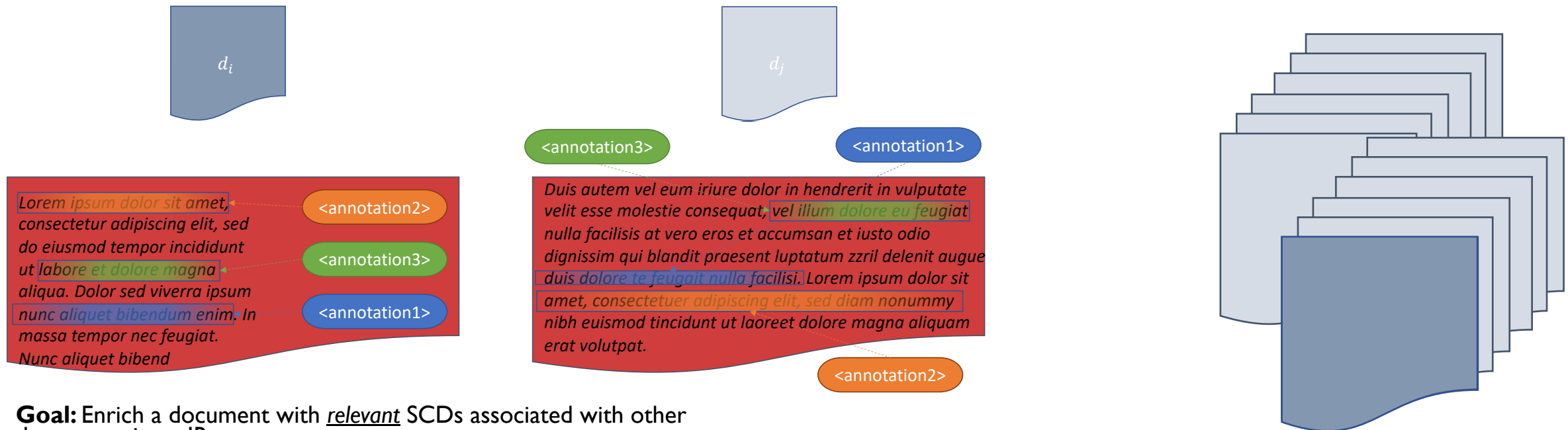
Document Type	president corpus		
	Precision	Recall	F1-Score
d_{sim}	0.77	0.71	0.74
d_{unrel}	1.00	0.96	0.98
d_{ext}	0.91	0.84	0.87
d_{rev}	0.72	0.58	0.64

ANNOTATION ENRICHMENT

SUPERVISED LEARNING



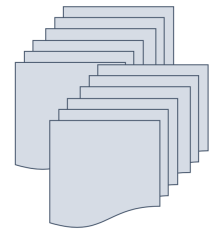
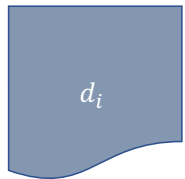
CORPUS-DRIVEN DOCUMENT ENRICHMENT USING SCDS



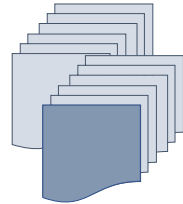
Goal: Enrich a document with relevant SCDS associated with other documents in an IR-corpus.

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 one
- until no more SCDS are assigned to document d .



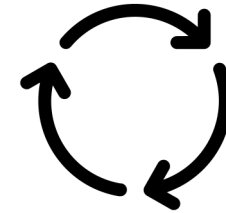
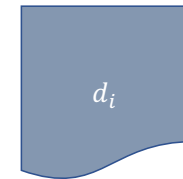
IR-corpora



d_i - related documents



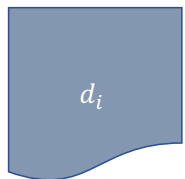
Topic similarity
SCD similarity
Frequency



d_i - related documents

$related_documents(d_i, IR_corpus)$

- Subset of *IR-corpora*
 - *topic-similar* documents whose
 - SCDs are *SCD-similar* to d_i



d_i - related documents

$expected_relevance(t, d_i)$

- 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

$mean_expected_relevance(d_i)$

average expected relevance value of SCDs in d_i -related documents

$enrich(d_i, IR_corpus)$

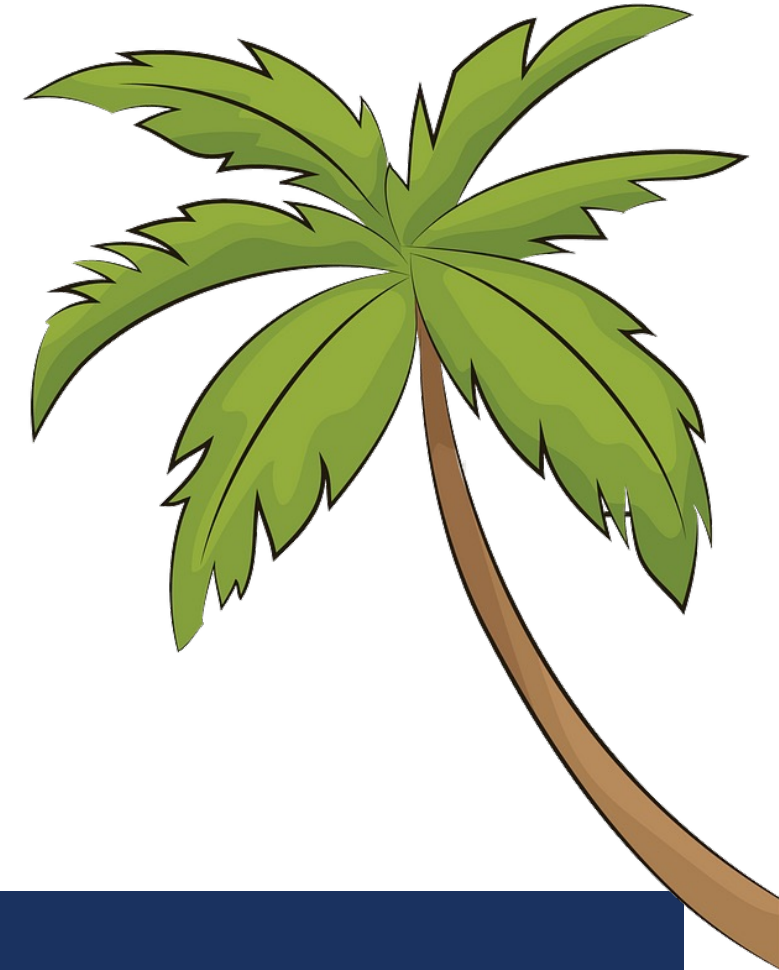
- Add SCD t to d_i if

$expected_relevance(t, d_i) > mean_expected_relevance(d_i)$

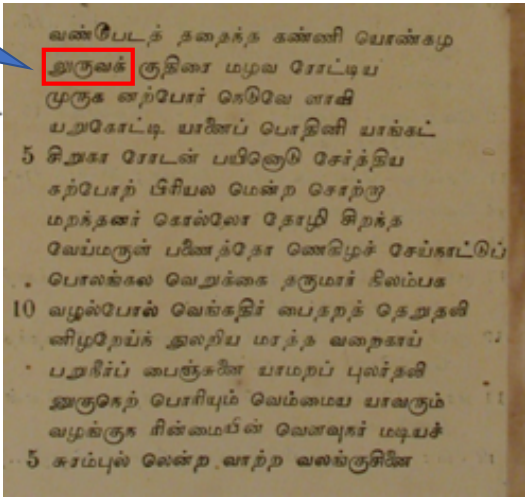
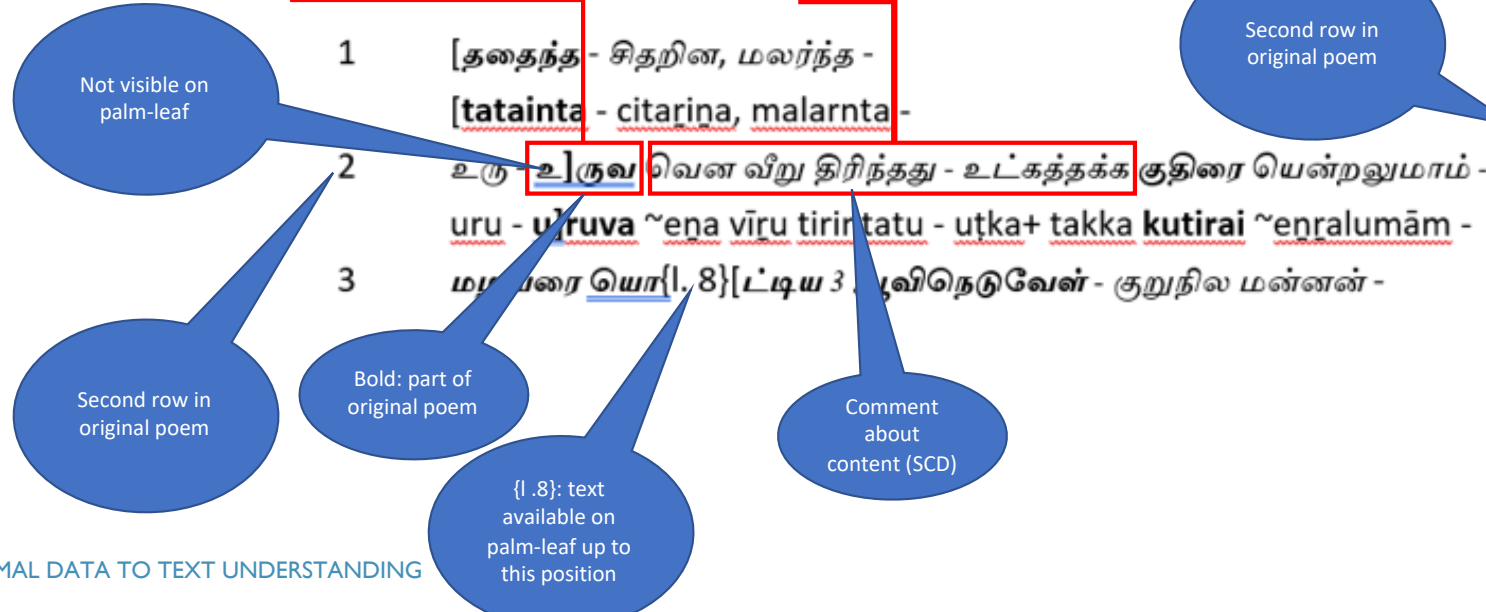
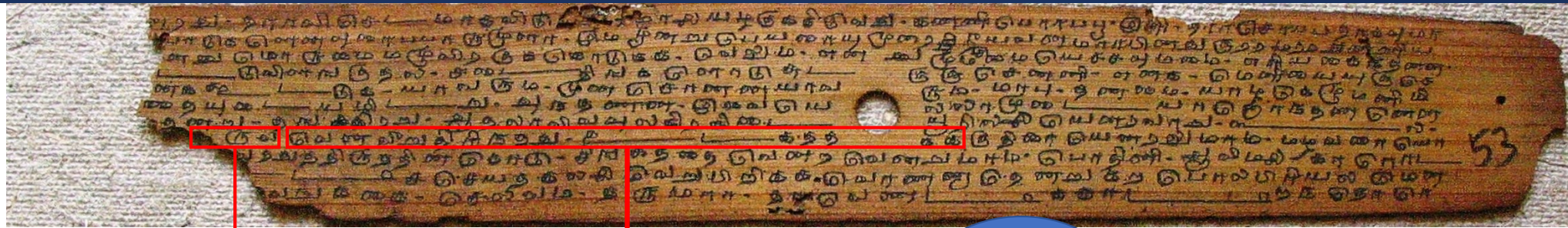
- Iterative enrichment process
 - Related documents changes with enriched SCDs
- Terminating enrichment process
 - Value of SCD similarity of d_i to related documents increases in a negligible way

DETECTING IN-LINE COMMENTS

SUPERVISED LEARNING



WHAT ARE COMMENTS WITHIN A TEXT?



FROM MINIMAL DATA TO TEXT UNDERSTANDING

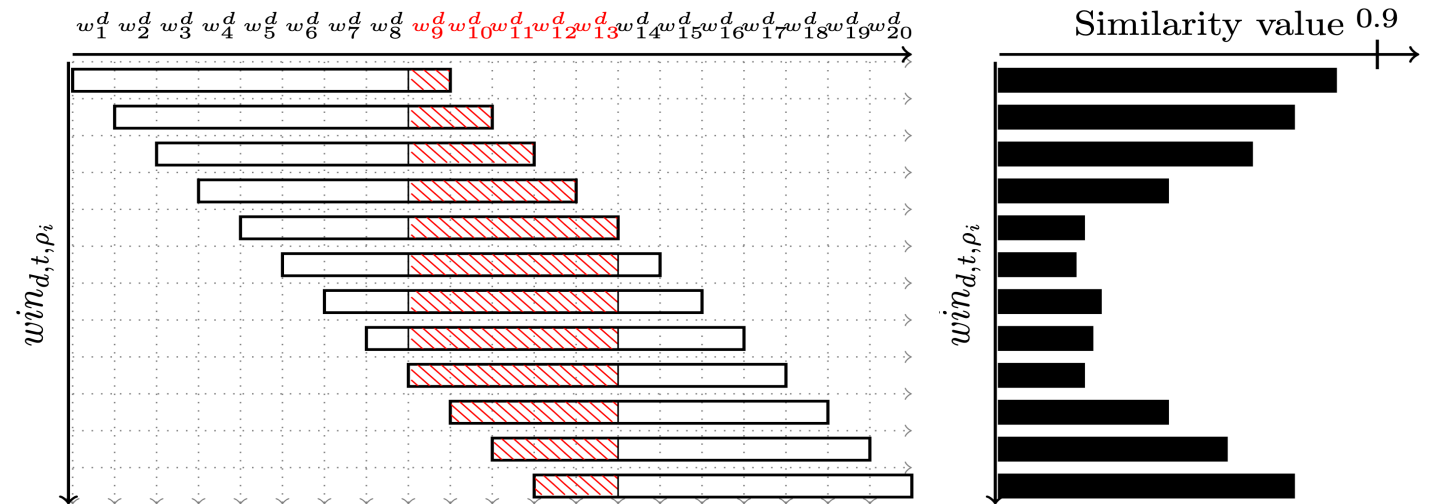
Supervised Learning

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 $d = (w_1^d, \dots, w_D^d)$, $w_i^d \in (\mathcal{V}_D \cup \mathcal{V}_{g(\mathcal{D})})$
 - Document d belongs to the same context as \mathcal{D}
 - Vocabulary \mathcal{V}_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



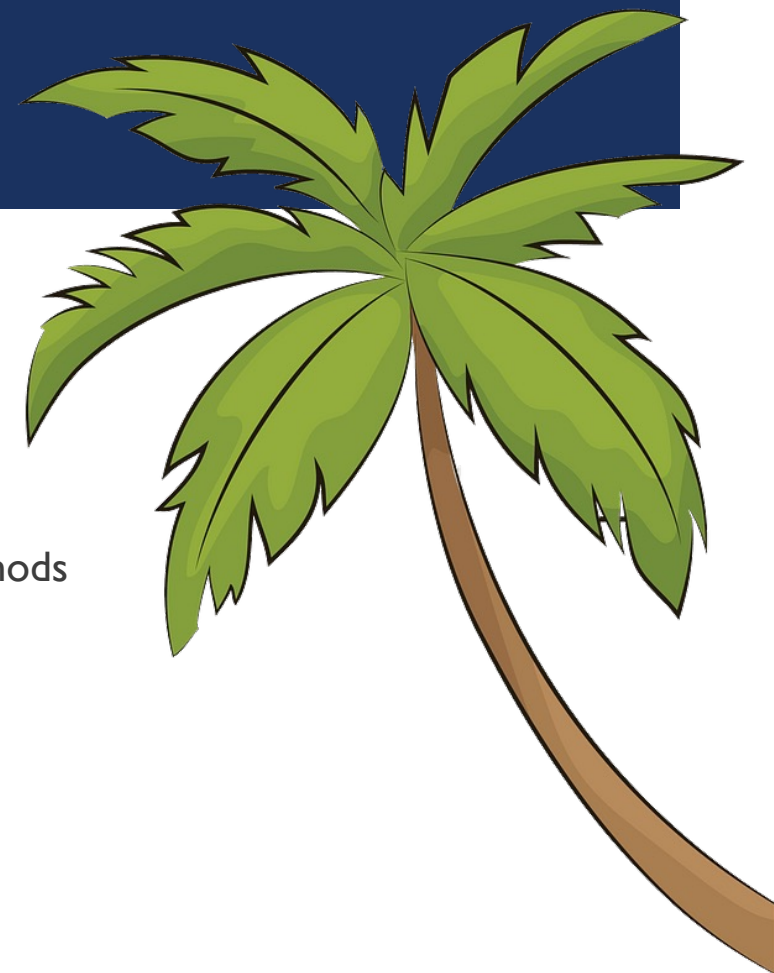
ESTIMATING ISCDs USING MPSCDS

- 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

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]

