TO EXTEND OR NOT TO EXTEND? COMPLEMENTARY DOCUMENTS

MAGNUS BENDER[1], FELIX KUHR[1], TANYA BRAUN[2]

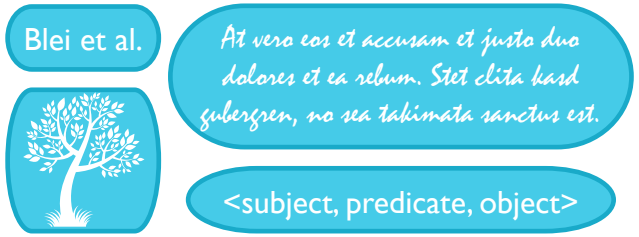
UNIVERSITÄT ZU LÜBECK

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

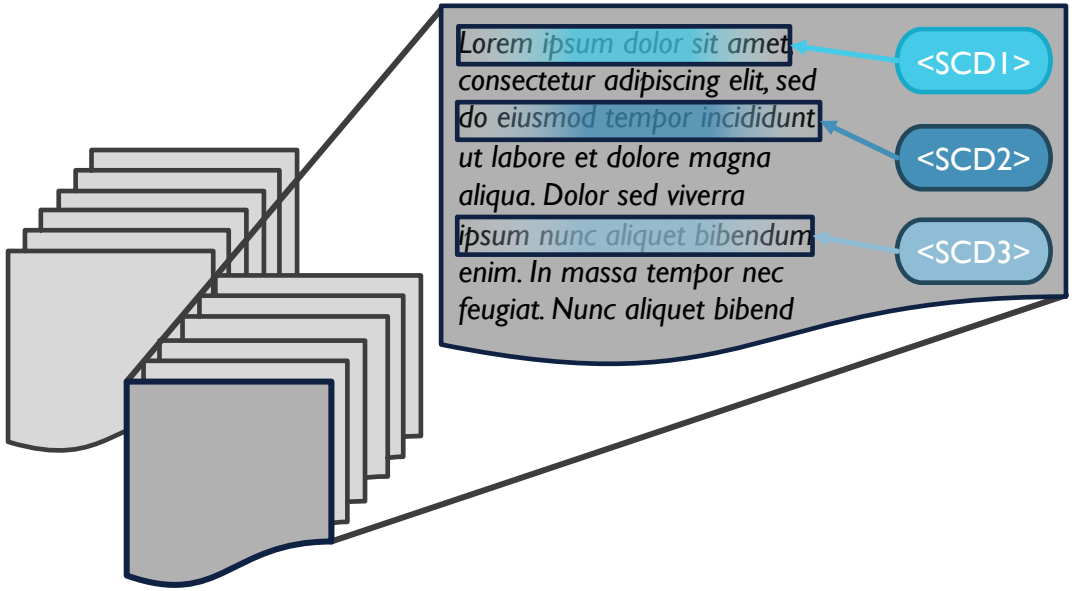
WWU MÜNSTER

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



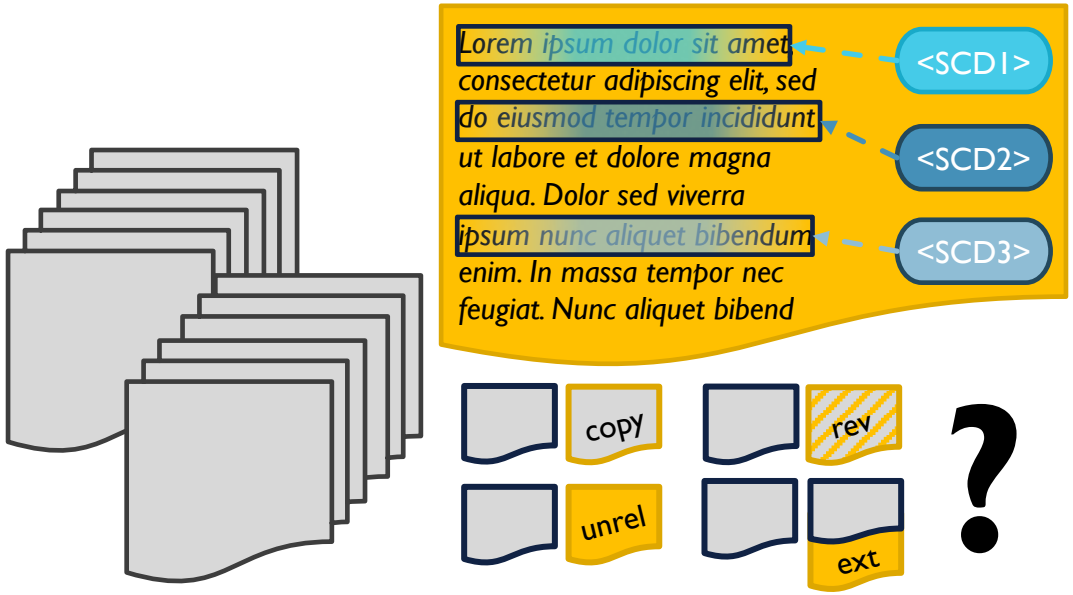
Proposition 1:
SCDs generate the words in a document

TASKS: DOCUMENT RETRIEVAL & CORPUS ENRICHMENT

- Document retrieval user-driven
  - External task
  - Well-rounded corpus needed for high-quality retrieval

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

  - Internal task
  - Classification problem
    - Input: new document d, corpus $\mathcal{D}$
    - Classify d as: quasi-copy, *revision*, *extension*, unrelated
      - *Documents with assumed added value*

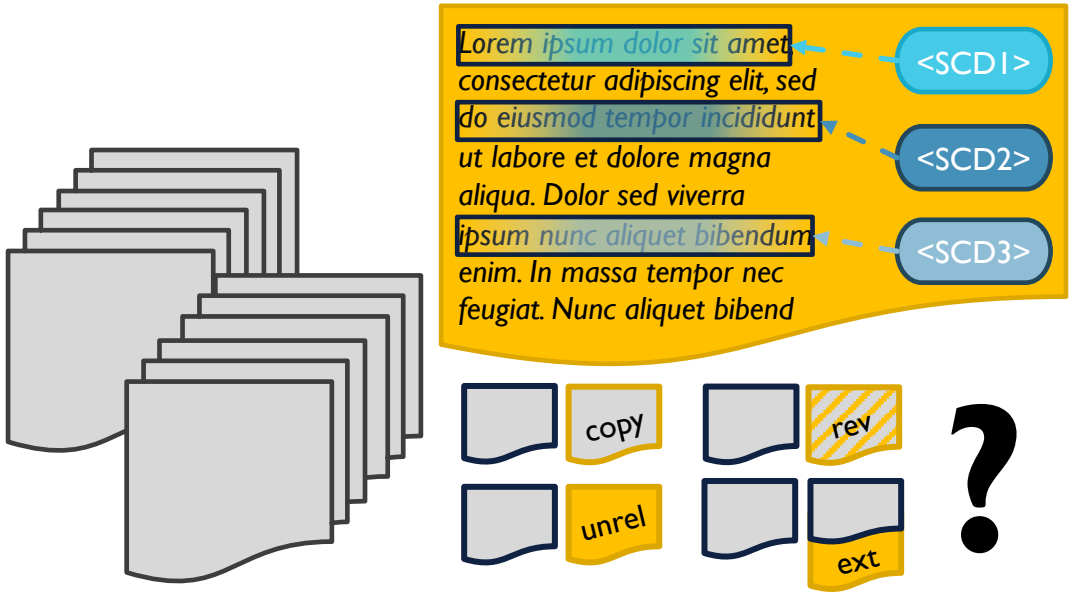


Based on Proposition 1:
How much of d can SCDs of $\mathcal{D}$ generate?

PROBLEM: SIMILARITY AS A MEASURE

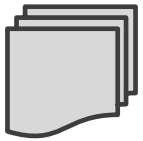
- Solution approach to corpus enrichment uses cosine similarity at its core
  - Sequence of similarity values between vector representations of SCDs and the words in the new document
- Also applies to many document retrieval approaches: return documents similar in some regard
  - Topic distribution similar, entities match (equality), etc.

Problem: Similarity-based approaches might only provide more of the same



DREAM SOLUTION: COMPLEMENTARITY

- Goal: identify documents that are complementary to a corpus / a document in a corpus
  - Binary classification problem: $Complement = true$ or $Complement = false$

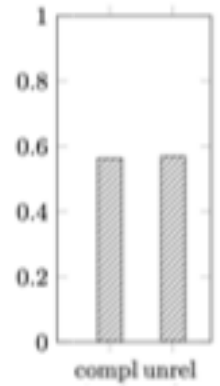


- Corpus on sporting events
  - Olympics 2020, UEFA Euro 2020
- Complementary documents on
  - Covid-19 spread in cities

- *BUT: Numbers-based representations make it hard to grasp complementarity on a* semantic *level*

New Problem: How do we *formally* define complementarity accounting for semantics?

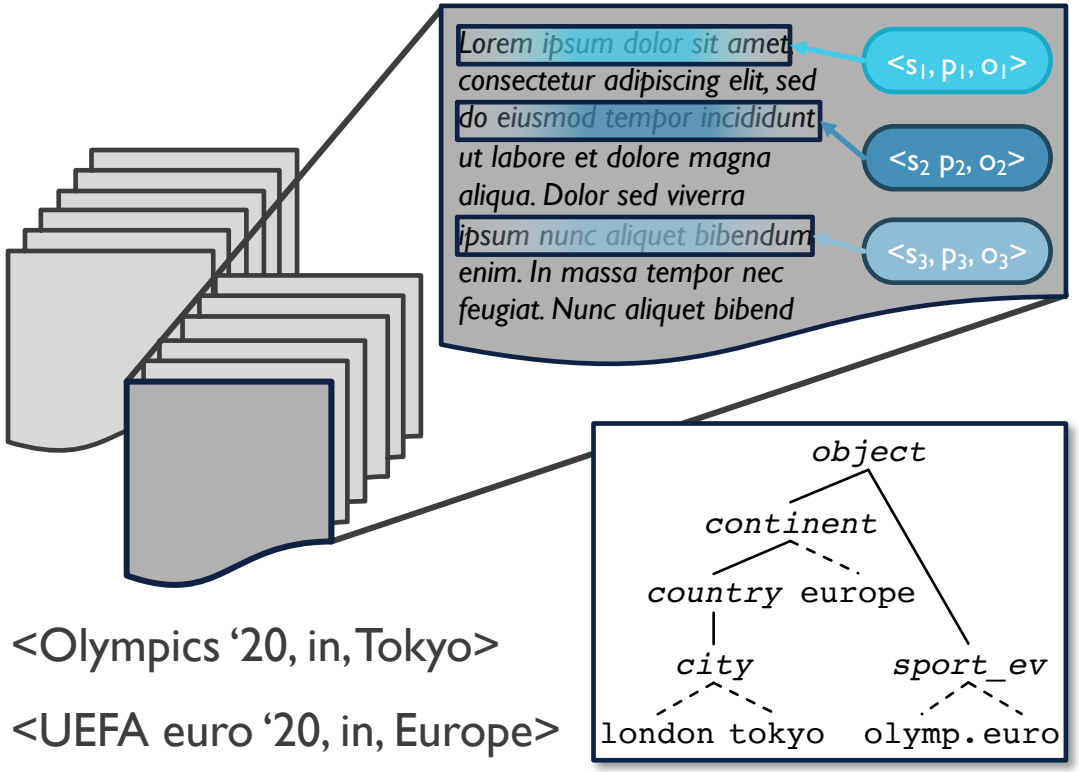
- Words very different compared to corpus
  - Different (topic / SCD) distributions
  - Likely to be classified as unrelated
    - Fig.: Similarity values of complements and unrelated documents for corpus enrichment



NEW PROBLEM: HOW TO GRASP COMPLEMENTARITY ON A FORMAL LEVEL?

- Use SCDs specifically in the form of SPO triples
  - SPO triple: <subject, predicate, object>
  - Extract for any document, e.g., with OpenIE tools

  together with a taxonomy
  - Hierarchy of concepts
  - Dictionary of synonyms
- Allow for grasping complementarity on a semantic level by
  - Looking at shared concepts in the SPO triples
  - While also accounting for hierarchy and synonyms



- <Olympics '20, in, Tokyo>
- <UEFA euro '20, in, Europe>

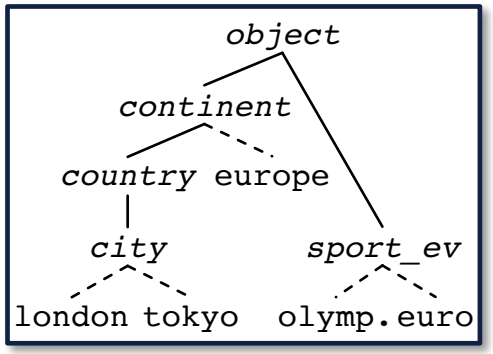
A FORMAL DEFINITION: COMPLEMENTARY SCDS

- Let $x^{\uparrow}$ refer to the concept or meaning of x
- Seven types of complementarity between SCDs t_i, t_j

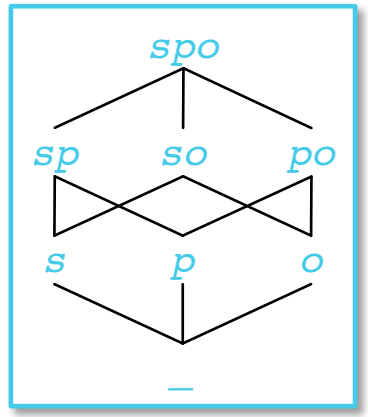
1. *s* $t_i = \langle s^{\uparrow}, p, o \rangle, t_j = \langle s^{\uparrow}, p, o \rangle$
2. *p* $t_i = \langle s, p^{\uparrow}, o \rangle, t_j = \langle s, p^{\uparrow}, o \rangle$
3. *o* $t_i = \langle s, p, o^{\uparrow} \rangle, t_j = \langle s, p, o^{\uparrow} \rangle$
4. *sp* $t_i = \langle s^{\uparrow}, p^{\uparrow}, o \rangle, t_j = \langle s^{\uparrow}, p^{\uparrow}, o \rangle$
5. *so* $t_i = \langle s^{\uparrow}, p, o^{\uparrow} \rangle, t_j = \langle s^{\uparrow}, p, o^{\uparrow} \rangle$
6. *po* $t_i = \langle s, p^{\uparrow}, o^{\uparrow} \rangle, t_j = \langle s, p^{\uparrow}, o^{\uparrow} \rangle$
7. *spo* $t_i = \langle s^{\uparrow}, p^{\uparrow}, o^{\uparrow} \rangle, t_j = \langle s^{\uparrow}, p^{\uparrow}, o^{\uparrow} \rangle$

- Types get less strict → Order in lattice

- t_1: <Olympics '20, in, Tokyo>
- t_2: <UEFA euro '20, in, Europe>
- t_3: <Covid-19, in, Tokyo>
- t_4: <Covid-19, in, London>



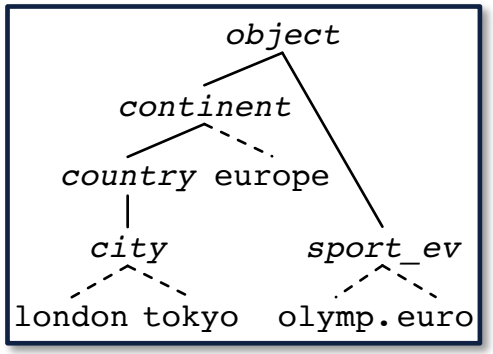
- t_1, t_3 *s*-complementary
  - s_1, s_3 share *object*; $p_1 = p_3$; $o_1 = o_3$
  - and *sp*, *so*, *spo*-complementary
- t_1, t_4 *so*-complementary (+*spo*)
  - Same holds for t_2, t_3 and t_2, t_4
- If deleting *object* in taxonomy, no complementarity!



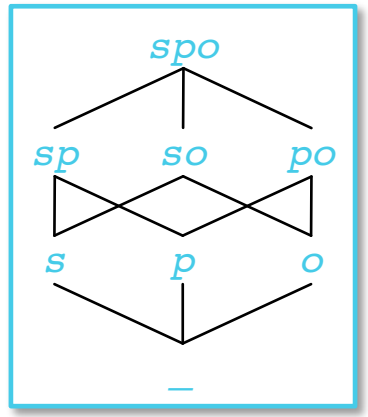
A FORMAL DEFINITION: COMPLEMENTARY DOCUMENTS – PRELIMINARIES

- Let $\mathfrak{C}_x(t_i, t_j), x \in \mathcal{X} = \{s, p, o, sp, so, po, spo\}$ be an indicator function
  - Returns 1 if t_i, t_j x-complementary; otherwise 0
  - $\mathfrak{C}_x$ is symmetric, i.e.,$\mathfrak{C}_x(t_i, t_j) = \mathfrak{C}_x(t_j, t_i)$
- Assign weights $w_x, \sum_{x \in \mathcal{X}} w_x = 1$, to complementarity types x to encode which complementarity interested in
  - Depends on corpus composition and desired outcome
  - $w_x = 0$ if x-complementarity uninteresting
    - $w_{sp} = 1$, rest 0: complementary SCDs specific to object o
  - $w_x \neq 0$ only for types of same level in lattice sensible
    - $w_x = \frac{1}{3}$ for $x \in \{sp, so, po\}$, rest 0

- t_1: <Olympics '20, in, Tokyo>
- t_2: <UEFA euro '20, in, Europe>
- t_3: <Covid-19, in, Tokyo>
- t_4: <Covid-19, in, London>



- Pairs $(t_1, t_3), (t_1, t_4), (t_2, t_3), (t_2, t_4)$
  - $\mathfrak{C}_s(t_1, t_3) = 1$, for other pairs 0
  - $\mathfrak{C}_{sp}(t_1, t_3) = 1$, for other pairs 0
  - $\mathfrak{C}_{so} = \mathfrak{C}_{spo} = 1$ for all pairs
  - $\mathfrak{C}_{po} = 0$ for all pairs



A FORMAL DEFINITION: COMPLEMENTARY DOCUMENTS – PRELIMINARIES

- Complementarity value between documents d', d:
  - Sum over all pairs of SCDs $t_i \in g(d'), t_j \in g(d)$, indicating if t_i, t_j are x-complementary:

$$\mathfrak{c}(d', d) = \sum_{(t_i,t_j)\in g(d')\times g(d)} \sum_{x\in\mathcal{X}} w_x \mathfrak{C}_x(t_i, t_j)$$

  - $\mathfrak{c}$ is symmetric, i.e., $\mathfrak{c}(d', d) = \mathfrak{c}(d, d')$
  - E.g., with $w_{sp} = 1$, rest 0:

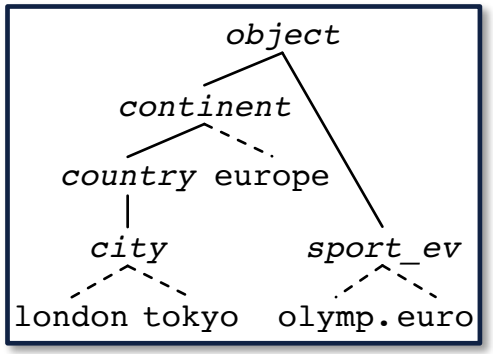
$$\mathfrak{c}(d_1, d_2) = 1 + 0 + 0 + 0 = 1$$

  - E.g., with $w_x = \frac{1}{3}$ for $x \in \{sp, so, po\}$, rest 0:

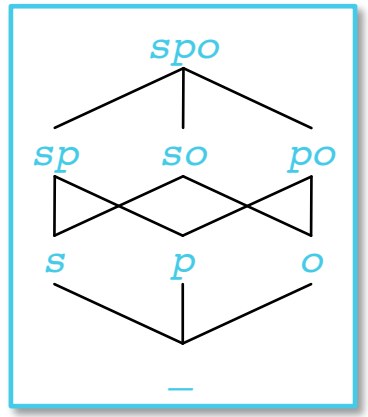
$$\mathfrak{c}(d_1, d_2) = \frac{1}{3}\cdot 1 + \frac{1}{3}\cdot 1 + \frac{1}{3}\cdot 0 + 3\left(\frac{1}{3}\cdot 0 + \frac{1}{3}\cdot 1 + \frac{1}{3}\cdot 0\right) = 1.667$$

d_1
- t_1: <Olympics '20, in, Tokyo>
- t_2: <UEFA euro '20, in, Europe>

d_2
- t_3: <Covid-19, in, Tokyo>
- t_4: <Covid-19, in, London>



- Pairs $(t_1, t_3), (t_1, t_4), (t_2, t_3), (t_2, t_4)$
  - $\mathfrak{C}_s(t_1, t_3) = 1$, for other pairs 0
  - $\mathfrak{C}_{sp}(t_1, t_3) = 1$, for other pairs 0
  - $\mathfrak{C}_{so} = \mathfrak{C}_{spo} = 1$ for all pairs
  - $\mathfrak{C}_{po} = 0$ for all pairs



A FORMAL DEFINITION: COMPLEMENTARY DOCUMENTS – DEFINITION

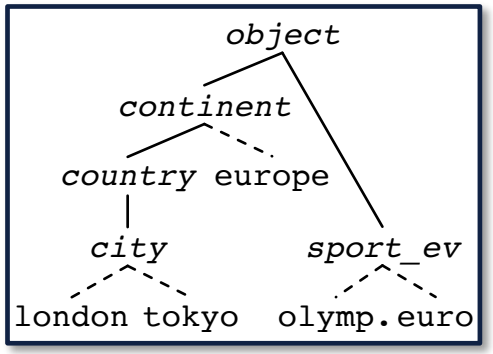
- Given a taxonomy ξ
- Given weights $w_x, x \in \mathcal{X}$
- Let $\mathfrak{C}_x(t_i, t_j), x \in \mathcal{X}$, be an indicator function
- Let $\mathfrak{c}(d', d)$ denote the complementarity value between documents d', d with SCDs $g(d'), g(d)$

- Then, given a threshold θ_d, d' is a complement to d if

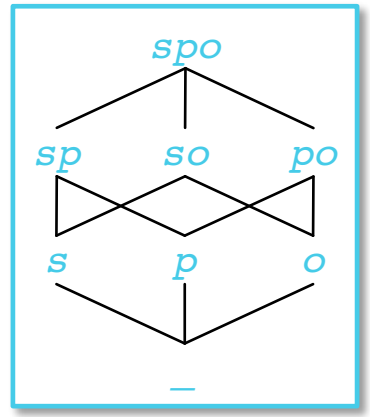
$$\mathfrak{c}(d', d) > \theta_d$$

d_1
- t_1: <Olympics '20, in, Tokyo>
- t_2: <UEFA euro '20, in, Europe>

d_2
- t_3: <Covid-19, in, Tokyo>
- t_4: <Covid-19, in, London>



- d_2 complement to d_1? ($\theta_d = 1.5$)
  - $w_{sp} = 1$:
    $\mathfrak{c}(d_2, d_1) = 1$ ✗
  - $w_x = \frac{1}{3}, x \in \{sp, so, po\}$:
    $\mathfrak{c}(d_2, d_1) = 1.667$ ✓



USING THE NEW DEFINITION: CORPUS ENRICHMENT – THE COMPLEMENTARITY VERSION

- Inputs:
  - Corpus $\mathcal{D}$
    - Includes SCDs including SPO triples, a taxonomy ξ
  - New document d'
  - Threshold $\theta_{\mathcal{D}}$
  - Weights $\{w_x\}_{x \in \mathcal{X}}$
- Returns $Complement = true$ or $Complement = false$
- Goes through all documents $d \in \mathcal{D}$ and adds up the complementary value $\mathfrak{c}(d', d)$ to a corpus value c

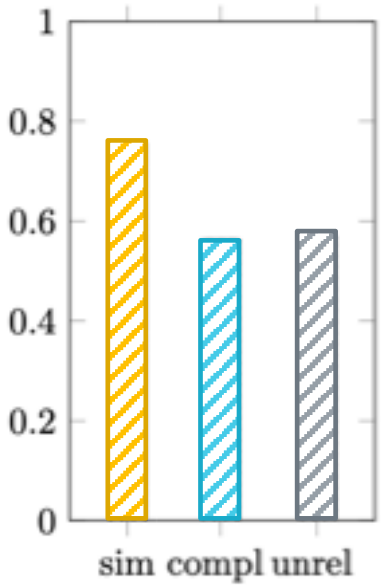
Adapt threshold to corpus setting!

function extendComplement($\mathcal{D}, d', \theta_{\mathcal{D}}, \{w_x\}_{x \in \mathcal{X}}$)
    if $g(d') = \emptyset$ **then**
        Add SCDs to d' using OpenIE
    $c \leftarrow 0$
    for each $t_i \in g(d')$ **do**
        for each $d \in \mathcal{D}$ **do**
            for each $t_j \in g(d')$ **do**
                for each $x \in \mathcal{X}$ **do**
                $c \leftarrow c + w_x \mathfrak{C}_x(t_i, t_j)$
    if $c > \theta_{\mathcal{D}}$ **then**
        return true
    return false

USING THE NEW DEFINITION: CORPUS ENRICHMENT – THE COMPLEMENTARITY VERSION

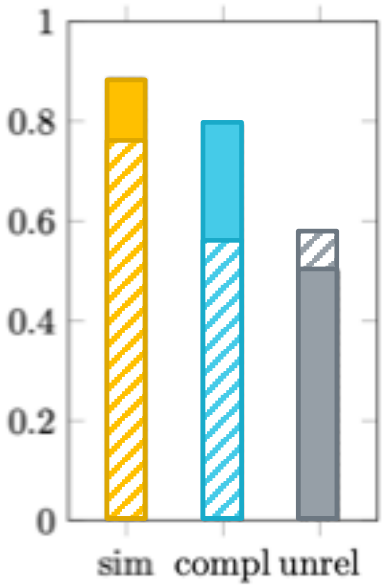
Similarity-based method

- Classifies both complements and unrelated documents as unrelated
- Identifies quasi-copies (sim)



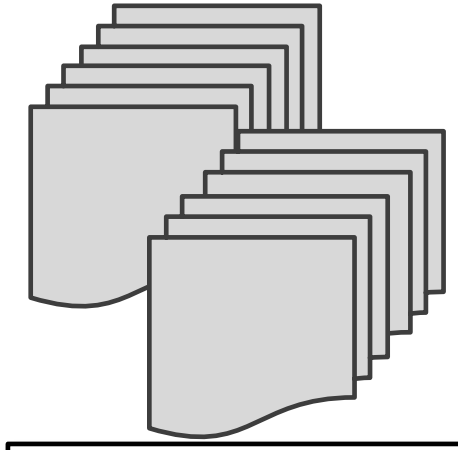
Complement-based method

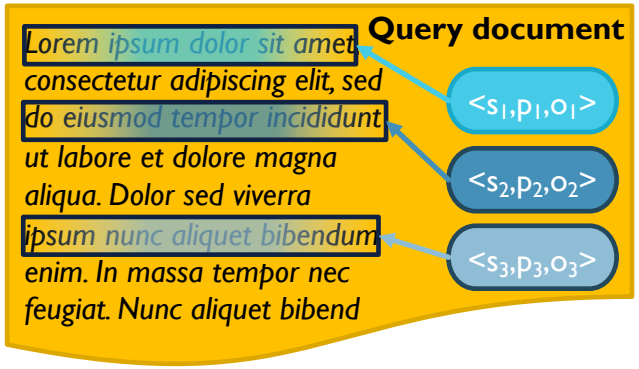
- Classifies complements as complements and unrelated documents as non-complements
- Quasi-copies (sim) have even higher complementarity values than complements
  - *Rationale*: Equal SPO triples match all complementarity types



USING THE NEW DEFINITION: DOCUMENT RETRIEVAL – THE COMPLEMENTARITY VERSION

- Input:
  - Corpus $\mathcal{D}$
    - Includes SCDs including SPO triples, a taxonomy ξ
  - Query document d'
  - Weights $\{w_x\}_{x \in \mathcal{X}}$
- Extract SPO triples for d' if none exist
- Compute $\mathfrak{c}(d', d)$ for each document in $d \in \mathcal{D}$
- Return
  - Given threshold θ_d: all documents d with $\mathfrak{c}(d', d) > \theta_d$
  - Given a number k: k documents with highest $\mathfrak{c}(d', d)$





Given threshold θ_d:

d_1: $\mathfrak{c}(d', d_1) > \theta_d$ → return
d_2: $\mathfrak{c}(d', d_2) \not> \theta_d$
d_3: $\mathfrak{c}(d', d_3) \not> \theta_d$
d_4: $\mathfrak{c}(d', d_4) > \theta_d$ → return
⋮

Ranking, $k = 3$:

d_1: $\mathfrak{c}(d', d_1)$ → return
d_{12}: $\mathfrak{c}(d', d_{12})$ → return
d_4: $\mathfrak{c}(d', d_4)$ → return
d_3: $\mathfrak{c}(d', d_3)$
⋮

Also allows for indicating **positions of interest** in a returned document where complementary SCDs clustered

CONCLUSION: A FORMAL DEFINITION OF COMPLEMENTARITY

- Formal definition of complementary documents based on SCDs in the form of SPO triples
  - Grasping complementarity on a semantic level
- Solve tasks using complementarity definition
  - Corpus enrichment, document retrieval
- Future work
  - Include "degree of ancestry" in indicator function
    - Also to distinguish better between copies and complements
  - Deal with uncertainty / unknown concepts