Context is the Key:

Context-aware Corpus Annotation Using Subjective Content Descriptions

Colloquium

Felix Kuhr

Institute of Information Systems
University of Lübeck

March 24, 2022

- Information retrieval (IR) is the task of finding documents that are relevant to a user's need for information

- Information retrieval (IR) is the task of finding documents that are relevant to a user's need for information
- Algorithms estimate relevance of displayed documents to searched queries:

- Information retrieval (IR) is the task of finding documents that are relevant to a user's need for information
- Algorithms estimate relevance of displayed documents to searched queries:
  - Compare words in a query with content of documents

- Information retrieval (IR) is the task of finding documents that are relevant to a user's need for information
- Algorithms estimate relevance of displayed documents to searched queries:
  - Compare words in a query with content of documents

An information retrieval system can be characterized by:

- Information retrieval (IR) is the task of finding documents that are relevant to a user's need for information
- Algorithms estimate relevance of displayed documents to searched queries:
  - Compare words in a query with content of documents

An information retrieval system can be characterized by:

- **Corpus** of documents

- Information retrieval (IR) is the task of finding documents that are relevant to a user's need for information
- Algorithms estimate relevance of displayed documents to searched queries:
  - Compare words in a query with content of documents

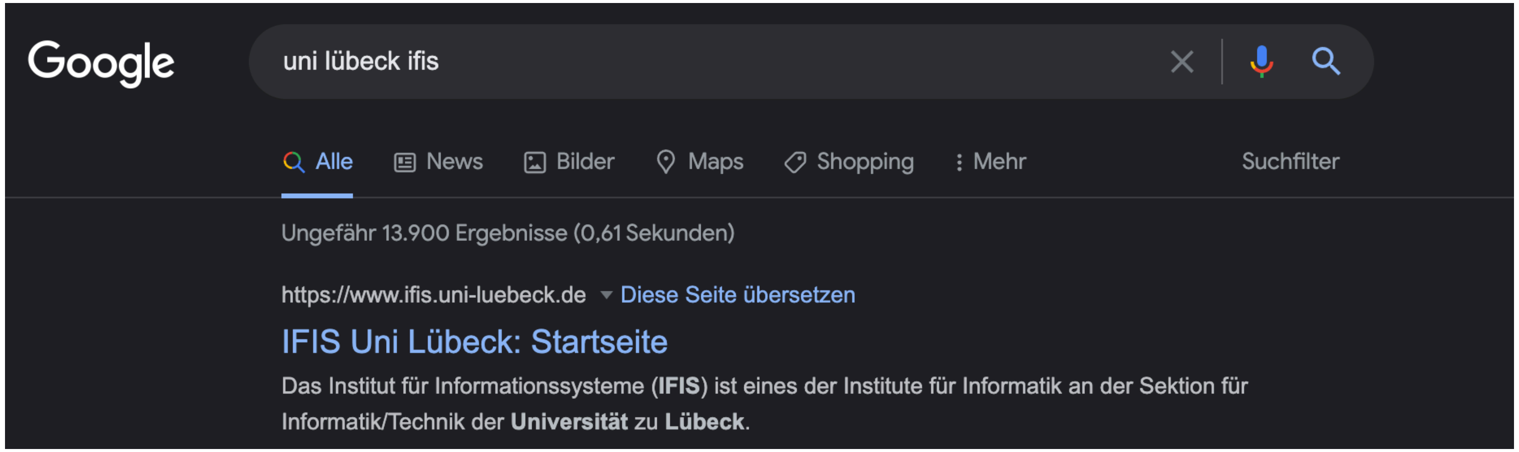
An information retrieval system can be characterized by:

- **Corpus** of documents
- **Queries** posed in a query language

- Information retrieval (IR) is the task of finding documents that are relevant to a user's need for information
- Algorithms estimate relevance of displayed documents to searched queries:
  - Compare words in a query with content of documents

An information retrieval system can be characterized by:

- **Corpus** of documents
- **Queries** posed in a query language



Ungefähr 13.900 Ergebnisse (0,61 Sekunden)

https://www.ifis.uni-luebeck.de ▾ Diese Seite übersetzen

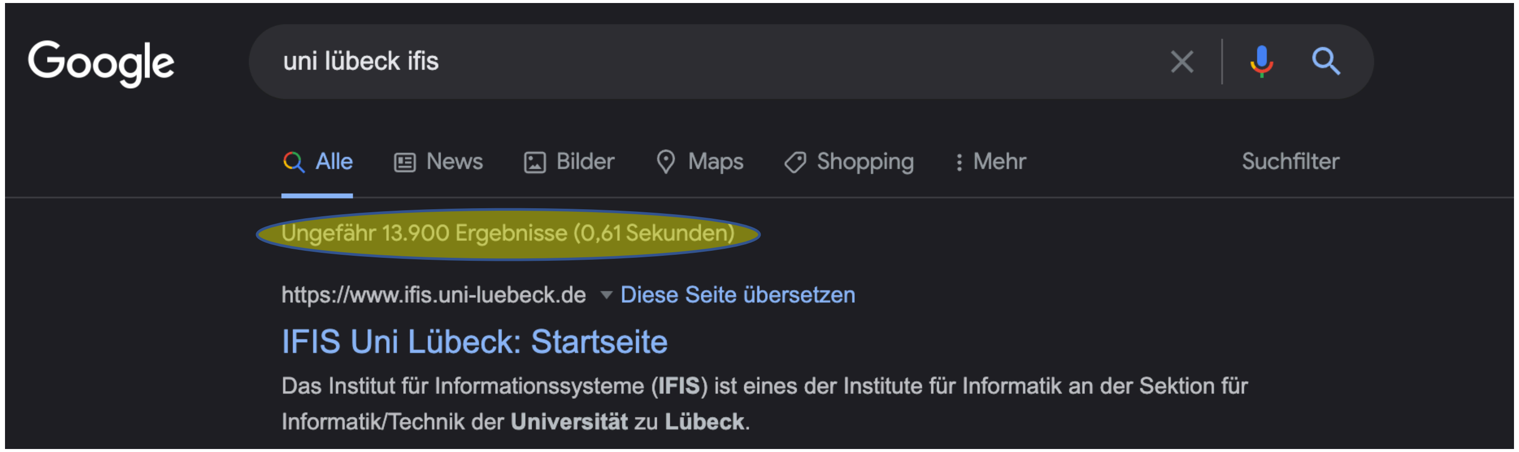
IFIS Uni Lübeck: Startseite

Das Institut für Informationssysteme (**IFIS**) ist eines der Institute für Informatik an der Sektion für Informatik/Technik der **Universität** zu **Lübeck**.

- Information retrieval (IR) is the task of finding documents that are relevant to a user's need for information
- Algorithms estimate relevance of displayed documents to searched queries:
  - Compare words in a query with content of documents

An information retrieval system can be characterized by:

- **Corpus** of documents
- **Queries** posed in a query language



Ungefähr 13.900 Ergebnisse (0,61 Sekunden)

https://www.ifis.uni-luebeck.de ▾ Diese Seite übersetzen

IFIS Uni Lübeck: Startseite

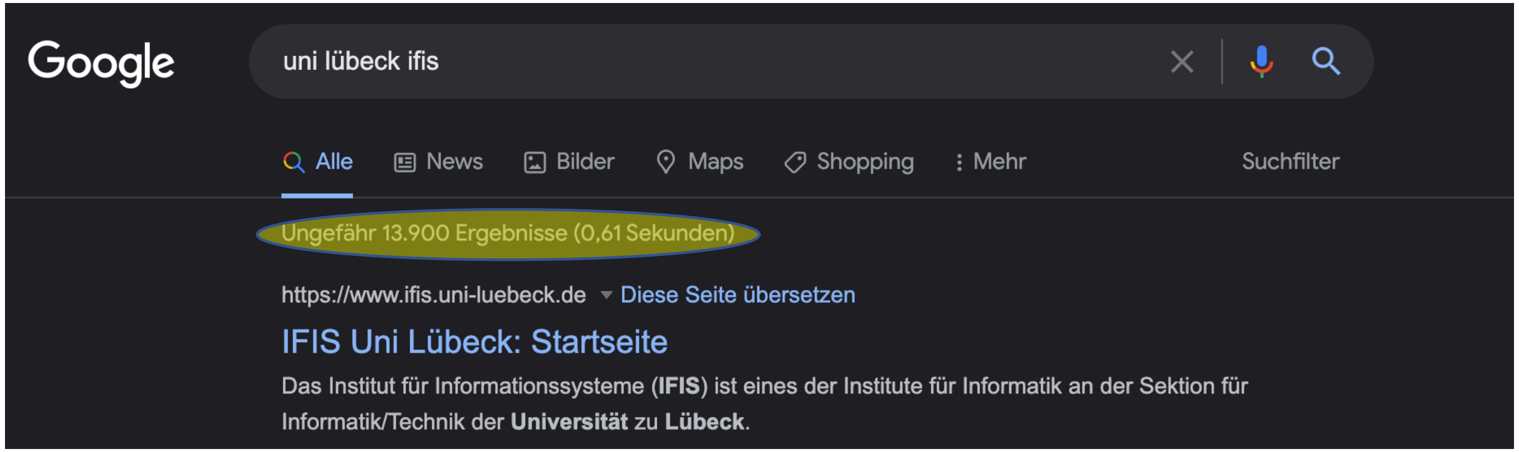
Das Institut für Informationssysteme (**IFIS**) ist eines der Institute für Informatik an der Sektion für Informatik/Technik der **Universität** zu **Lübeck**.

- Information retrieval (IR) is the task of finding documents that are relevant to a user's need for information
- Algorithms estimate relevance of displayed documents to searched queries:
  - Compare words in a query with content of documents

An information retrieval system can be characterized by:

- **Corpus** of documents
- **Queries** posed in a query language

Are *documents* in the result set relevant to the information need of a user?



Ungefähr 13.900 Ergebnisse (0,61 Sekunden)

https://www.ifis.uni-luebeck.de ▾ Diese Seite übersetzen

IFIS Uni Lübeck: Startseite

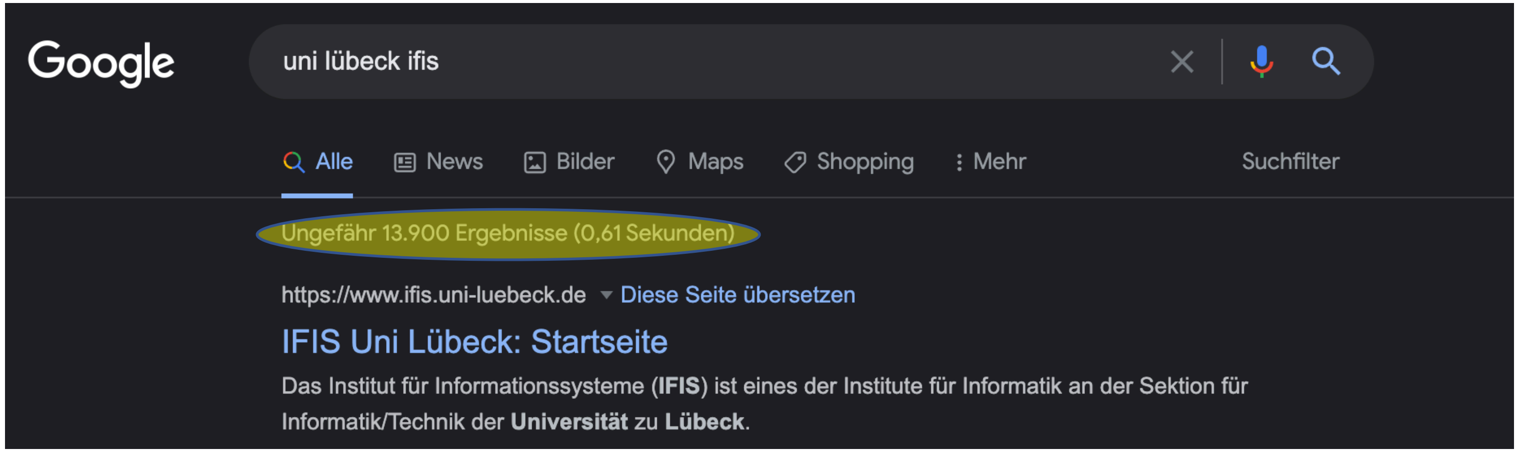
Das Institut für Informationssysteme (**IFIS**) ist eines der Institute für Informatik an der Sektion für Informatik/Technik der **Universität** zu **Lübeck**.

- Information retrieval (IR) is the task of finding documents that are relevant to a user's need for information
- Algorithms estimate relevance of displayed documents to searched queries:
  - Compare words in a query with content of documents

An information retrieval system can be characterized by:

- **Corpus** of documents
- **Queries** posed in a query language
- **Result** set of relevant documents

Are *documents* in the result set relevant to the information need of a user?



uni lübeck ifis

Ungefähr 13.900 Ergebnisse (0,61 Sekunden)

https://www.ifis.uni-luebeck.de ▾ Diese Seite übersetzen

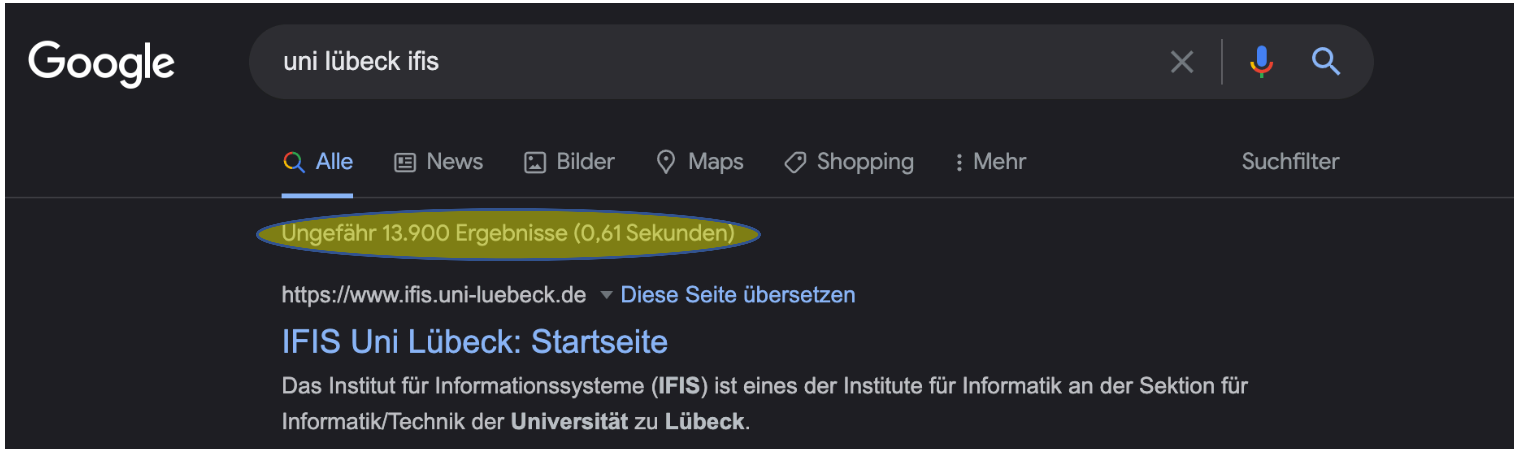
IFIS Uni Lübeck: Startseite

Das Institut für Informationssysteme (**IFIS**) ist eines der Institute für Informatik an der Sektion für Informatik/Technik der **Universität** zu **Lübeck**.

- Information retrieval (IR) is the task of finding documents that are relevant to a user's need for information
- Algorithms estimate relevance of displayed documents to searched queries:
  - Compare words in a query with content of documents

An information retrieval system can be characterized by:

- **Corpus** of documents
- **Queries** posed in a query language
- **Result** set of relevant documents
-



Ungefähr 13.900 Ergebnisse (0,61 Sekunden)

https://www.ifis.uni-luebeck.de ▾ Diese Seite übersetzen

IFIS Uni Lübeck: Startseite

Das Institut für Informationssysteme (**IFIS**) ist eines der Institute für Informatik an der Sektion für Informatik/Technik der **Universität** zu **Lübeck**.

Are *documents* in the result set relevant to the information need of a user?

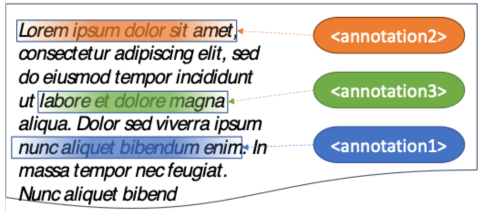
Annotation Systems

Annotation Systems

Manual Annotation Systems

⊕ Quality of annotations

User-centric annotations



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

Annotation Systems

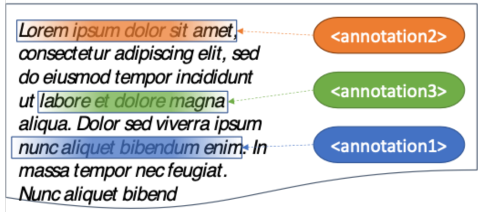
Manual Annotation Systems

⊕ Quality of annotations

User-centric annotations

⊖ Human annotation experts

High costs / time-consuming



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

Annotation Systems

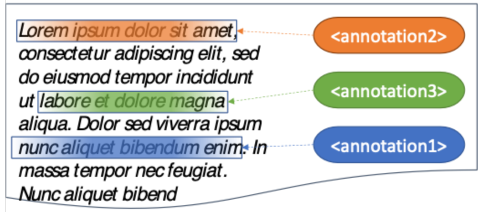
Manual Annotation Systems

⊕ Quality of annotations

User-centric annotations

⊖ Human annotation experts

High costs / time-consuming

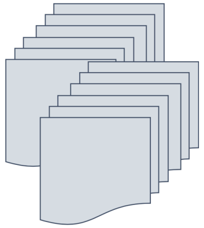


Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Dolor sed viverra ipsum nunc aliquet bibendum enim. In massa tempor nec feugiat. Nunc aliquet bibend

Automatic Annotation Systems

⊕ Low costs

Fast annotation process



Annotation Systems

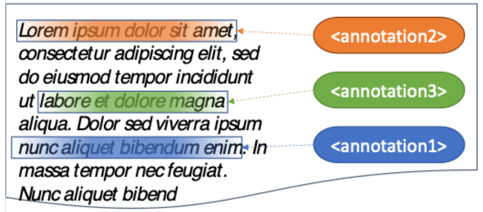
Manual Annotation Systems

⊕ Quality of annotations

User-centric annotations

⊖ Human annotation experts

High costs / time-consuming



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

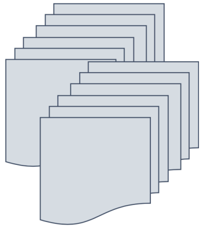
Automatic Annotation Systems

⊕ Low costs

Fast annotation process

⊖ Quality of annotations

Missing explainability



Annotation Systems

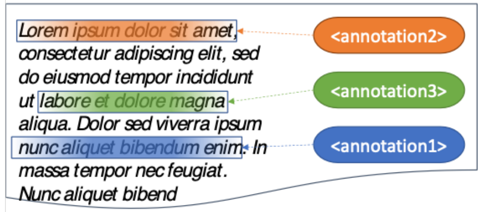
Manual Annotation Systems

⊕ Quality of annotations

User-centric annotations

⊖ Human annotation experts

High costs / time-consuming


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

⇆

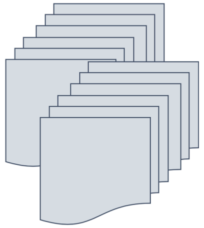
Automatic Annotation Systems

⊕ Low costs

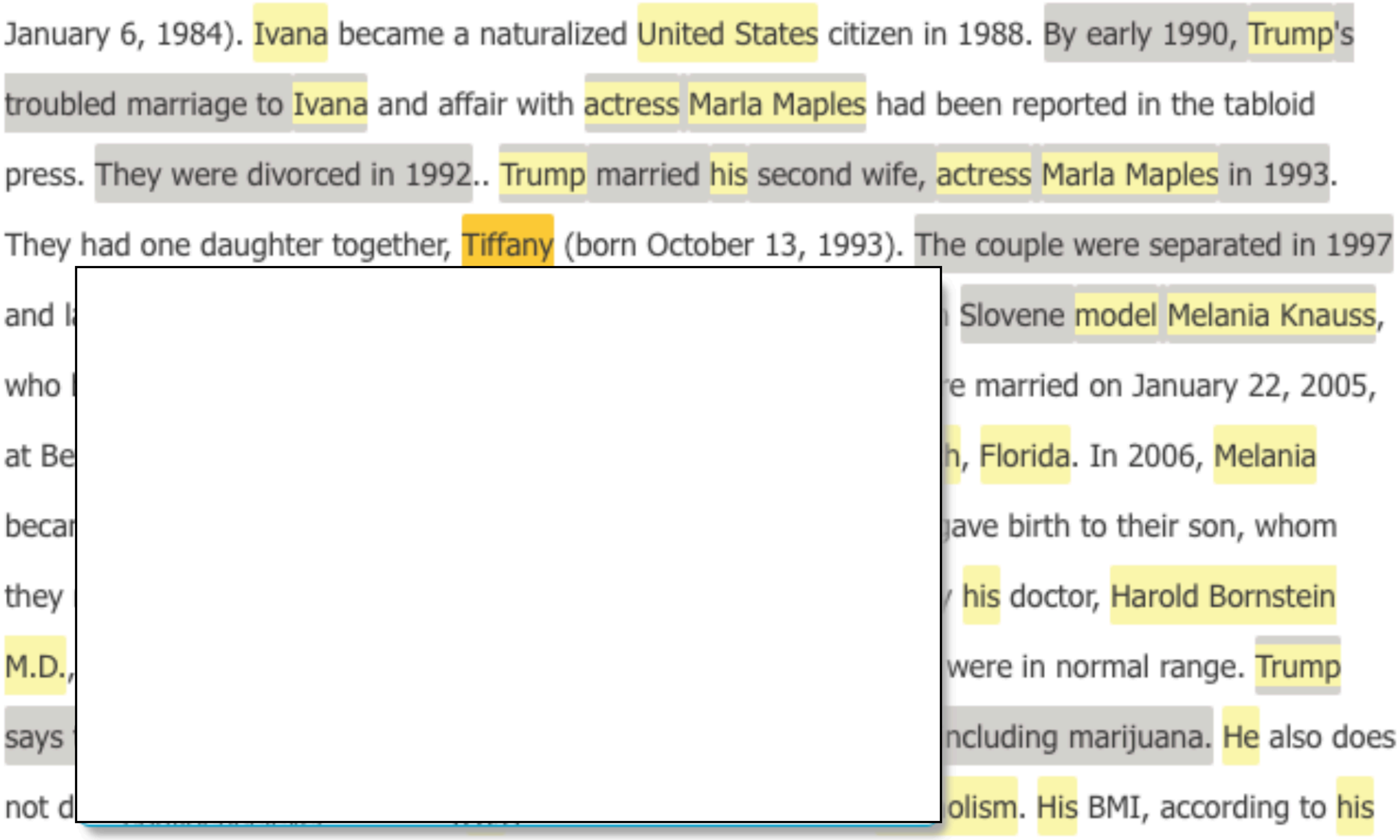
Fast annotation process

⊖ Quality of annotations

Missing explainability



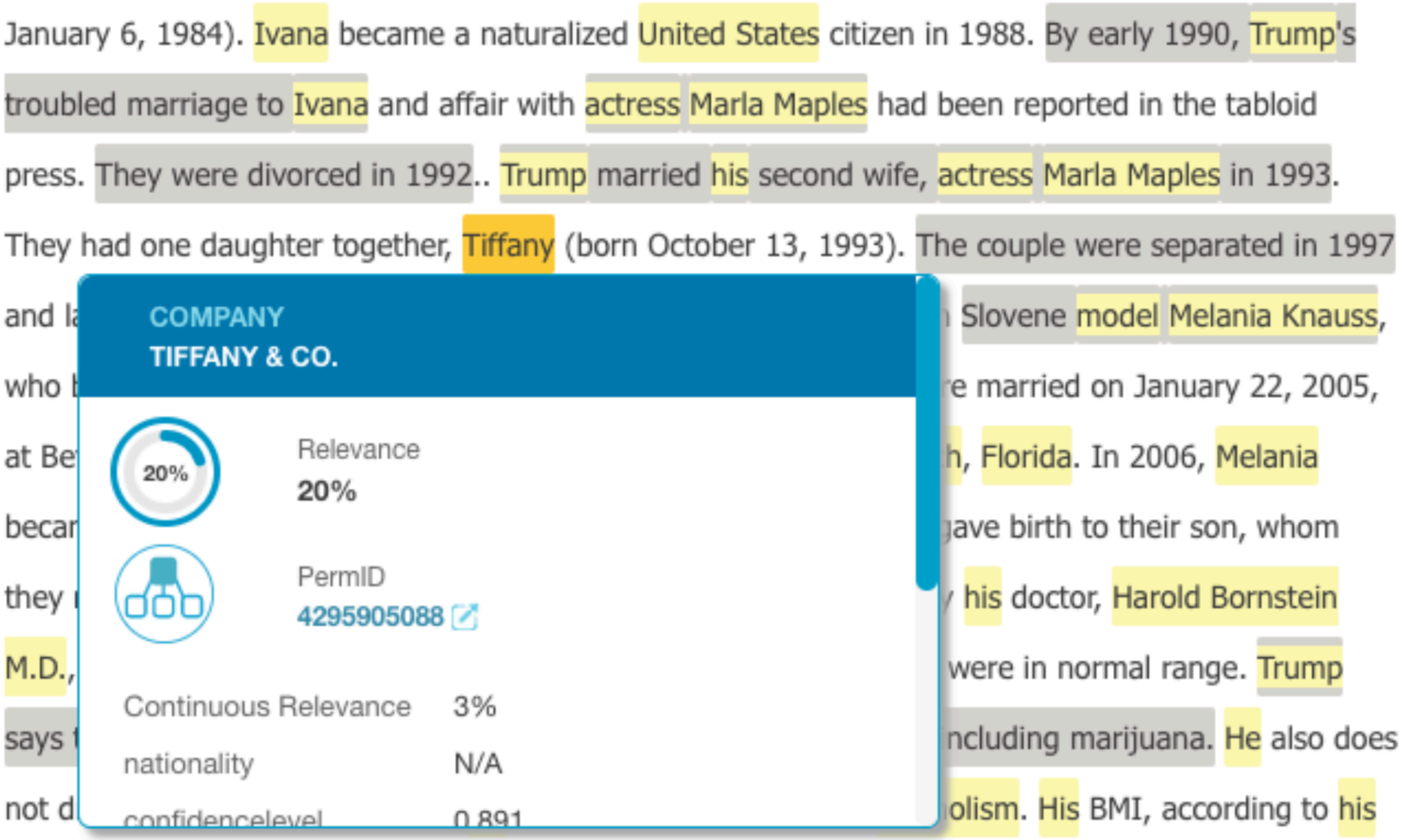
Context is the Key: Example (One Calais)*



January 6, 1984). Ivana became a naturalized United States citizen in 1988. By early 1990, Trump's
troubled marriage to Ivana and affair with actress Marla Maples had been reported in the tabloid
press. They were divorced in 1992.. Trump married his second wife, actress Marla Maples in 1993.
They had one daughter together, Tiffany (born October 13, 1993). The couple were separated in 1997
and l Slovene model Melania Knauss,
who l e married on January 22, 2005,
at Be h, Florida. In 2006, Melania
becar jave birth to their son, whom
they his doctor, Harold Bornstein
M.D., were in normal range. Trump
says ncluding marijuana. He also does
not d olism. His BMI, according to his

* https://permid.org/onecalaisViewer

Context is the Key: Example (One Calais)*



January 6, 1984). Ivana became a naturalized United States citizen in 1988. By early 1990, Trump's
troubled marriage to Ivana and affair with actress Marla Maples had been reported in the tabloid
press. They were divorced in 1992.. Trump married his second wife, actress Marla Maples in 1993.
They had one daughter together, Tiffany (born October 13, 1993). The couple were separated in 1997
and l Slovene model Melania Knauss,
who e married on January 22, 2005,
at Be h, Florida. In 2006, Melania
becar gave birth to their son, whom
they his doctor, Harold Bornstein
M.D., were in normal range. Trump
says ncluding marijuana. He also does
not d olism. His BMI, according to his

COMPANY
TIFFANY & CO.

20%

Relevance
20%

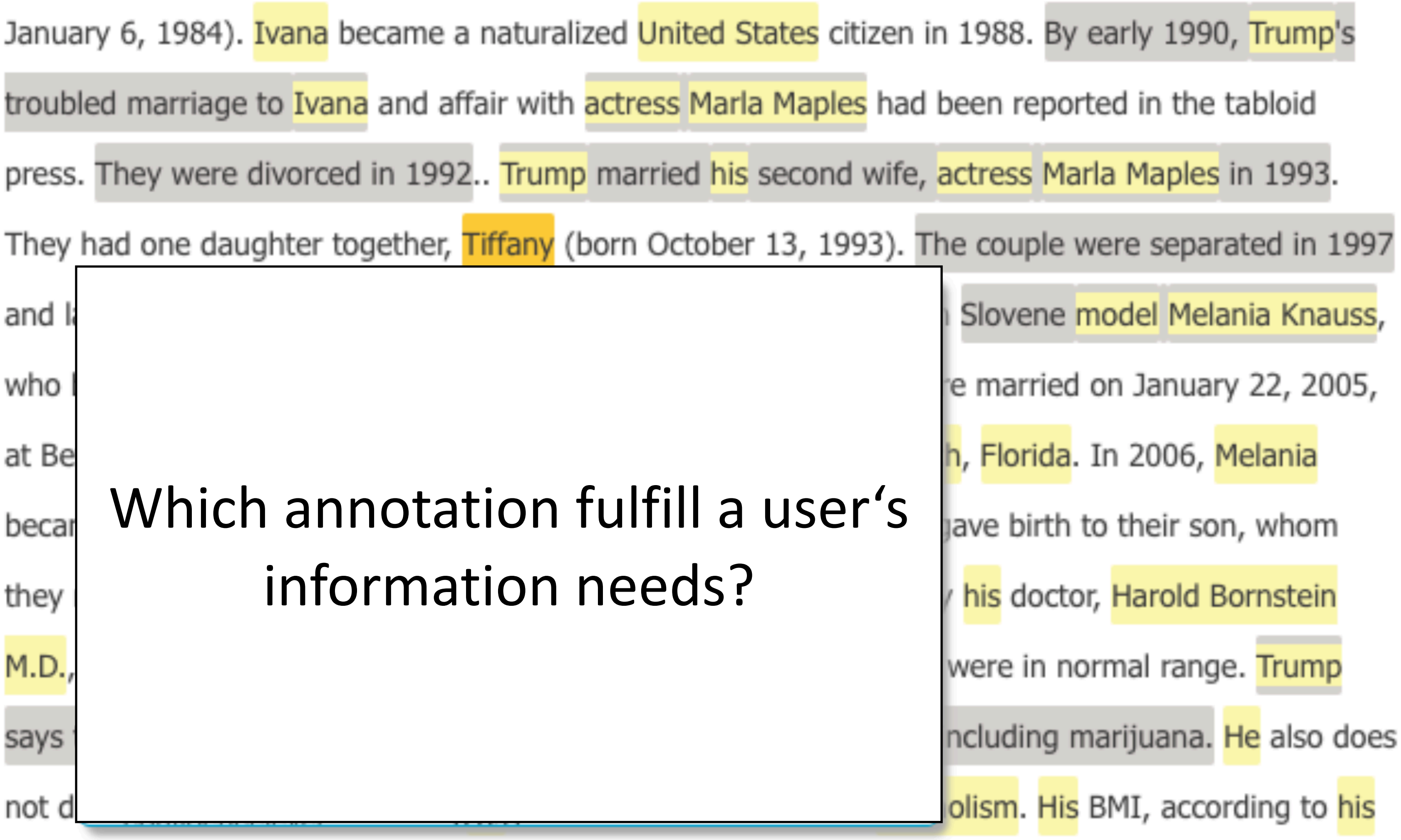
PermID
4295905088

Continuous Relevance	3%
nationality	N/A
confidencelevel	0.891

Add meta data to text by linking extractable entities to external data

* *https://permid.org/onecalaisViewer*

Context is the Key: Example (One Calais)*



January 6, 1984). Ivana became a naturalized United States citizen in 1988. By early 1990, Trump's
troubled marriage to Ivana and affair with actress Marla Maples had been reported in the tabloid
press. They were divorced in 1992.. Trump married his second wife, actress Marla Maples in 1993.
They had one daughter together, Tiffany (born October 13, 1993). The couple were separated in 1997
and l Slovene model Melania Knauss,
who e married on January 22, 2005,
at Be h, Florida. In 2006, Melania
becar gave birth to their son, whom
they his doctor, Harold Bornstein
M.D., were in normal range. Trump
says ncluding marijuana. He also does
not d olism. His BMI, according to his

Which annotation fulfill a user‘s information needs?

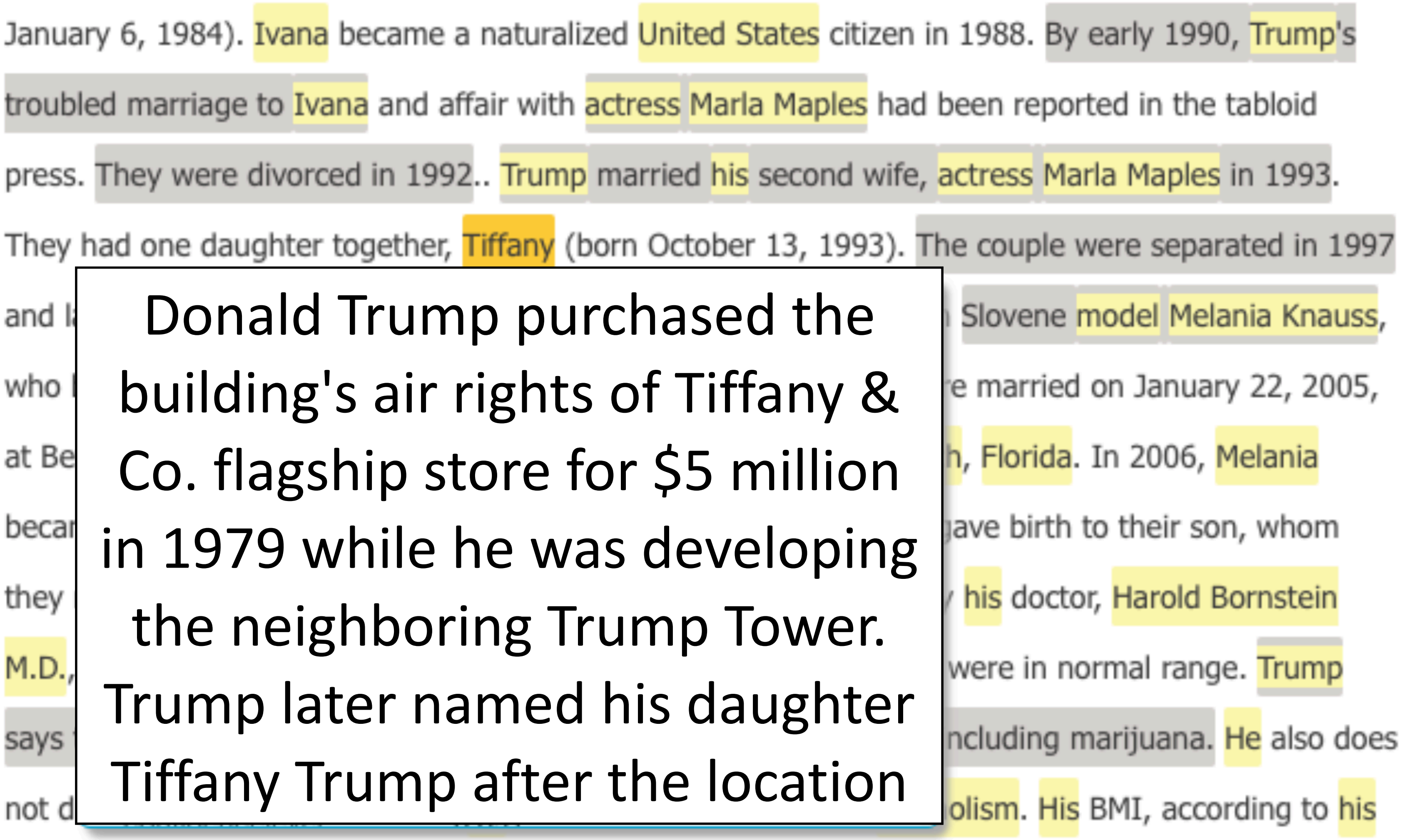
Add meta data to text by linking extractable entities to external data

Text related with Company Tiffany & Co.

Persons with names also used for companies

* https://permid.org/onecalaisViewer

Context is the Key: Example (One Calais)*



January 6, 1984). Ivana became a naturalized United States citizen in 1988. By early 1990, Trump's
troubled marriage to Ivana and affair with actress Marla Maples had been reported in the tabloid
press. They were divorced in 1992.. Trump married his second wife, actress Marla Maples in 1993.
They had one daughter together, Tiffany (born October 13, 1993). The couple were separated in 1997
and l
Slovene model Melania Knauss,
who
e married on January 22, 2005,
at Be
h, Florida. In 2006, Melania
becar
gave birth to their son, whom
they
his doctor, Harold Bornstein
M.D.,
were in normal range. Trump
says
ncluding marijuana. He also does
not d
olism. His BMI, according to his

Donald Trump purchased the building's air rights of Tiffany & Co. flagship store for $5 million in 1979 while he was developing the neighboring Trump Tower. Trump later named his daughter Tiffany Trump after the location

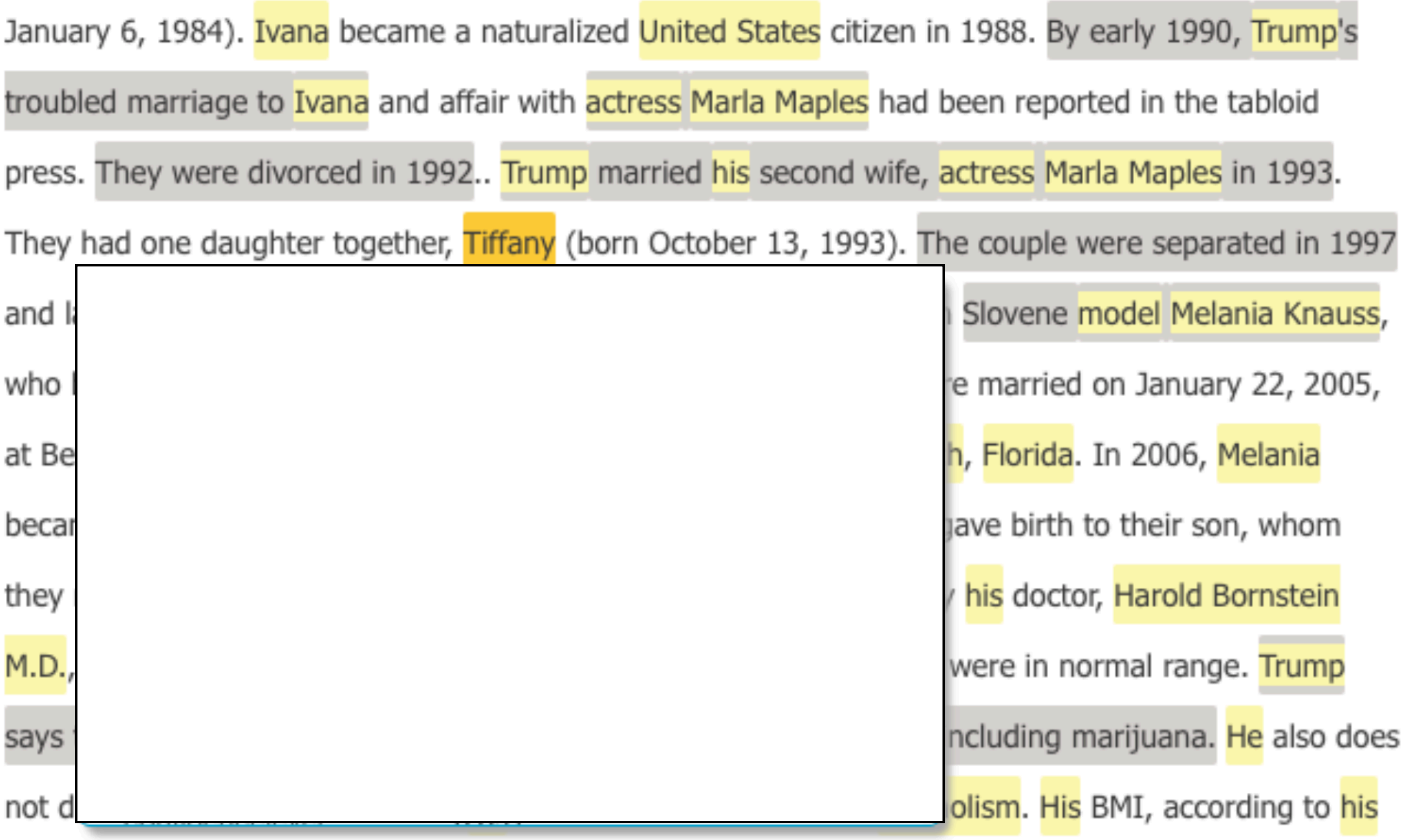
Add meta data to text by linking extractable entities to external data

Text related with Company Tiffany & Co.

Persons with names also used for companies

* *https://permid.org/onecalaisViewer*

Context is the Key: Example (One Calais)*



January 6, 1984). Ivana became a naturalized United States citizen in 1988. By early 1990, Trump's
troubled marriage to Ivana and affair with actress Marla Maples had been reported in the tabloid
press. They were divorced in 1992.. Trump married his second wife, actress Marla Maples in 1993.
They had one daughter together, Tiffany (born October 13, 1993). The couple were separated in 1997
and l Slovene model Melania Knauss,
who e married on January 22, 2005,
at Be h, Florida. In 2006, Melania
becar ave birth to their son, whom
they his doctor, Harold Bornstein
M.D., were in normal range. Trump
says ncluding marijuana. He also does
not d olism. His BMI, according to his

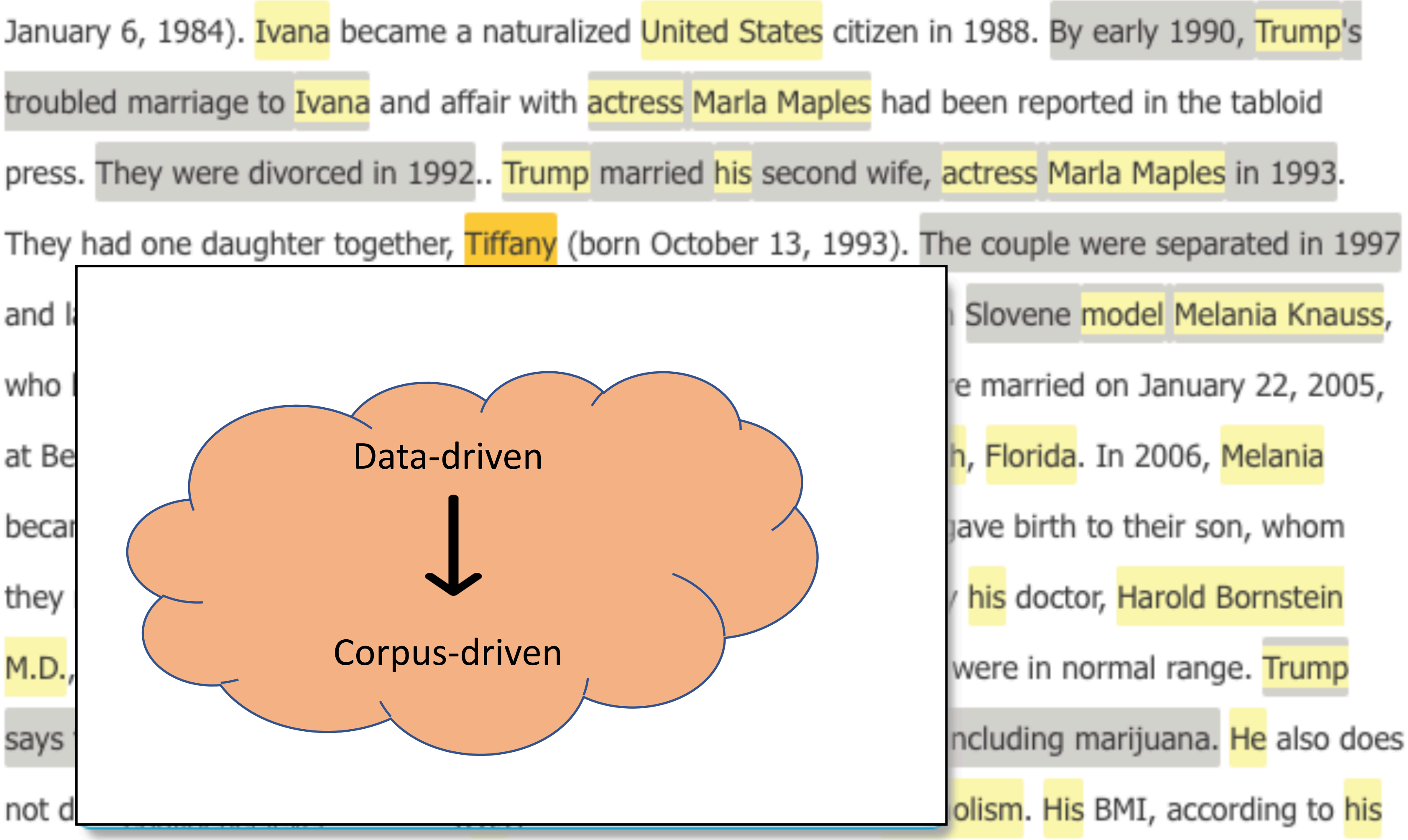
Add meta data to text by linking extractable entities to external data

Text related with Company Tiffany & Co.

Persons with names also used for companies

* https://permid.org/onecalaisViewer

Context is the Key: Example (One Calais)*



January 6, 1984). Ivana became a naturalized United States citizen in 1988. By early 1990, Trump's troubled marriage to Ivana and affair with actress Marla Maples had been reported in the tabloid press. They were divorced in 1992.. Trump married his second wife, actress Marla Maples in 1993. They had one daughter together, Tiffany (born October 13, 1993). The couple were separated in 1997

and l Slovene model Melania Knauss,

who e married on January 22, 2005,

at Be h, Florida. In 2006, Melania

becar gave birth to their son, whom

they his doctor, Harold Bornstein

M.D., were in normal range. Trump

says ncluding marijuana. He also does

not d olism. His BMI, according to his

Data-driven

↓

Corpus-driven

Add meta data to text by linking extractable entities to external data

Text related with Company Tiffany & Co.

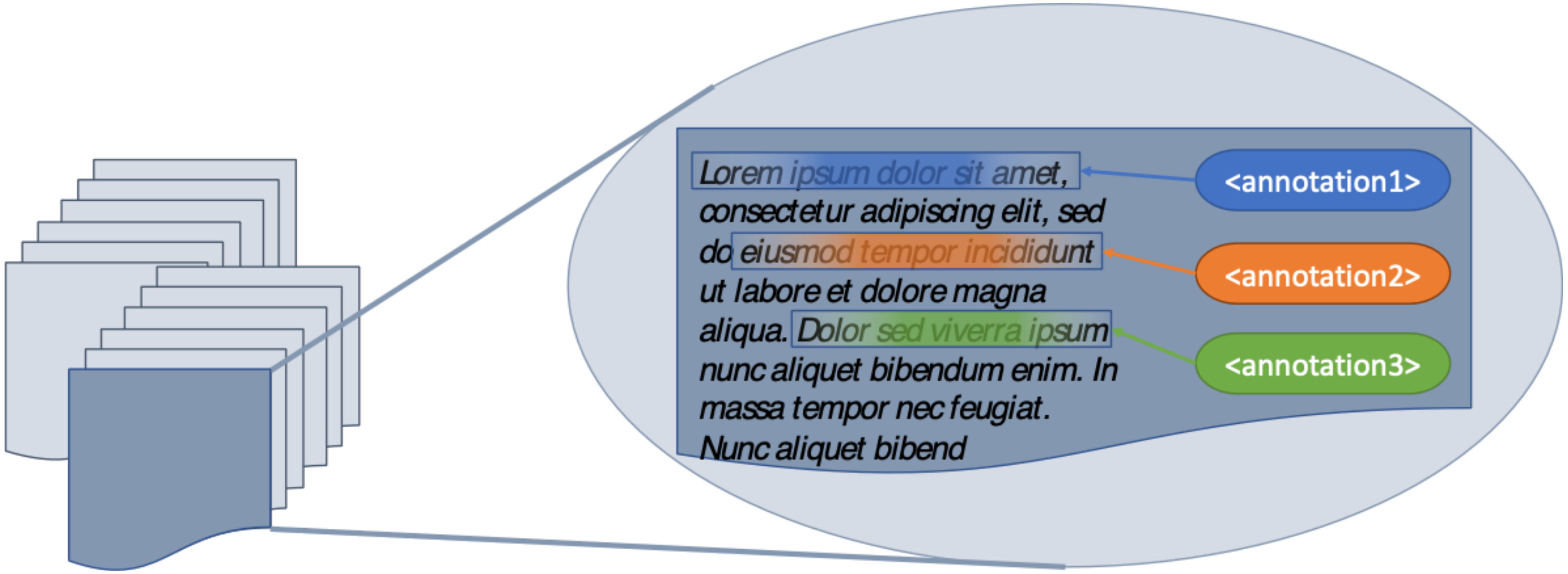
Persons with names also used for companies

* https://permid.org/onecalaisViewer

Subjective Content Descriptions (SCDs)

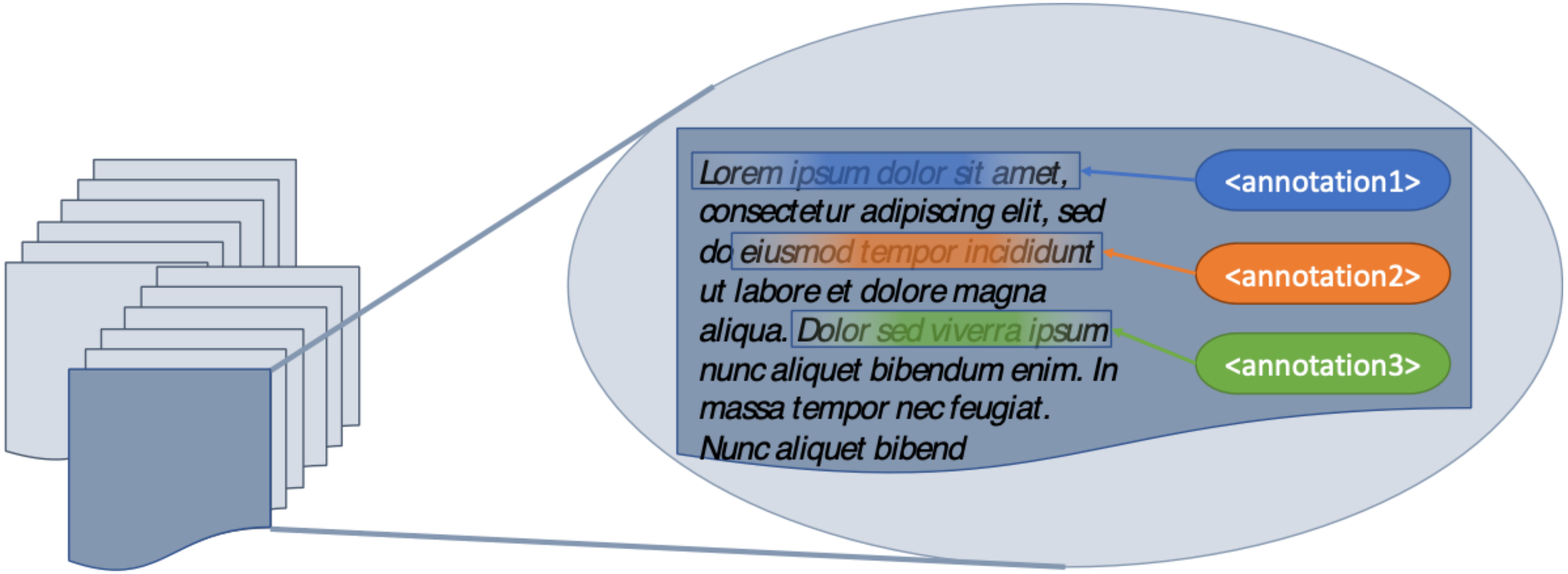
Subjective Content Descriptions (SCDs)

- … represent additional data for a document


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
<annotation1>
<annotation2>
<annotation3>

Subjective Content Descriptions (SCDs)

- … represent additional data for a document
- … are associated with a text span (window) in a document



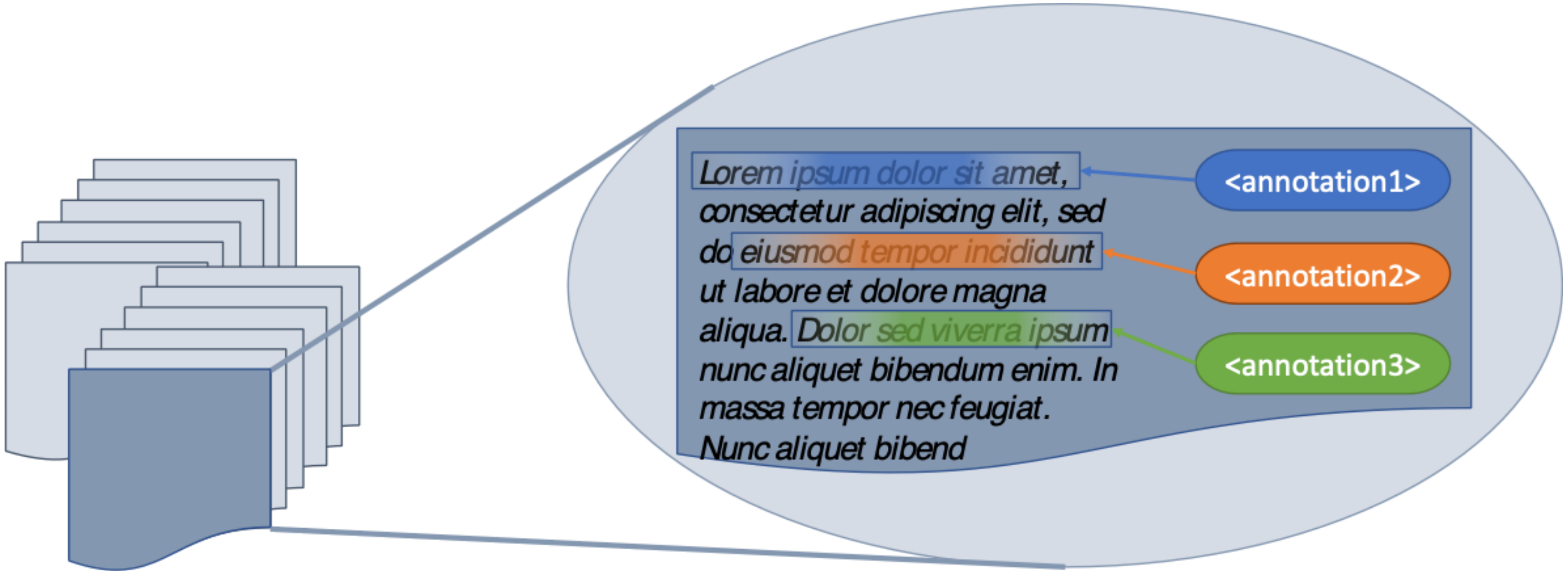
Lorem ipsum dolor sit amet,
consectetur adipiscing elit, sed
do eiusmod tempor incididunt
ut labore et dolore magna
aliqua. Dolor sed viverra ipsum
nunc aliquet bibendum enim. In
massa tempor nec feugiat.
Nunc aliquet bibend

Subjective Content Descriptions (SCDs)

- … represent additional data for a document
- … are associated with a text span (window) in a document

Different Typs of SCDs:

- Additional definitions

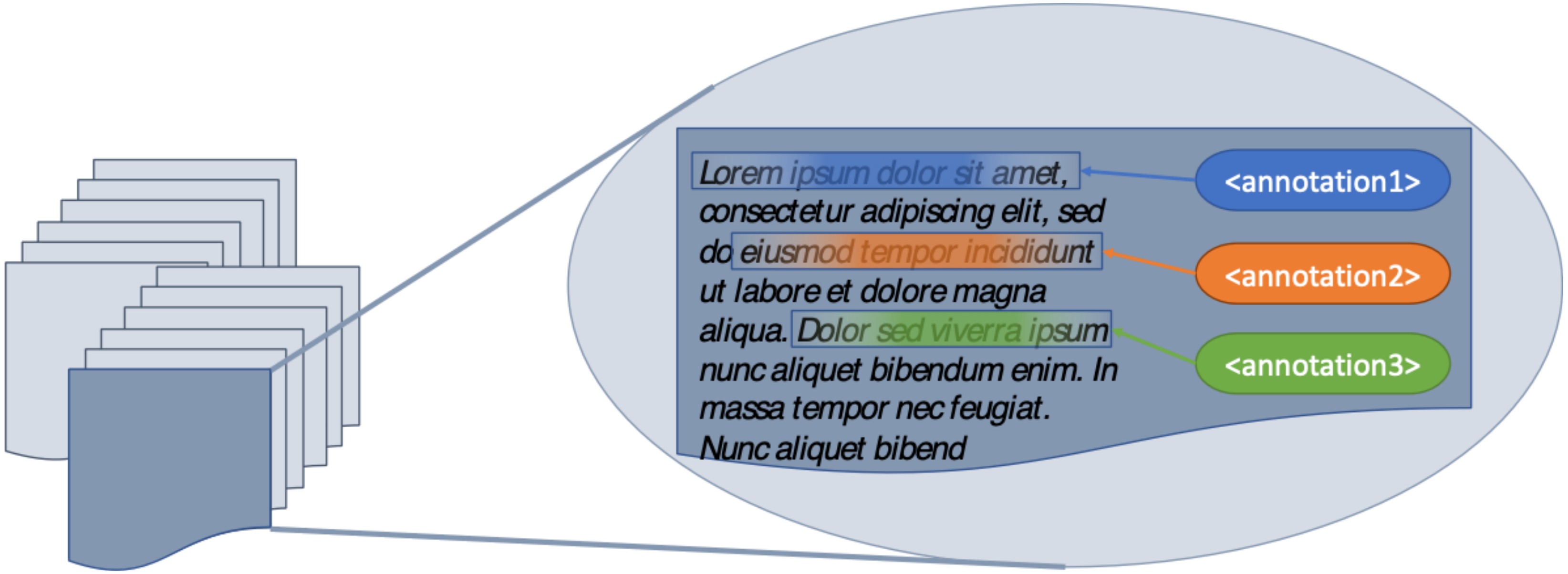

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Dolor sed viverra ipsum nunc aliquet bibendum enim. In massa tempor nec feugiat. Nunc aliquet bibend
<annotation1>
<annotation2>
<annotation3>

Subjective Content Descriptions (SCDs)

- … represent additional data for a document
- … are associated with a text span (window) in a document

Different Typs of SCDs:

- Additional definitions
- Links to external sources

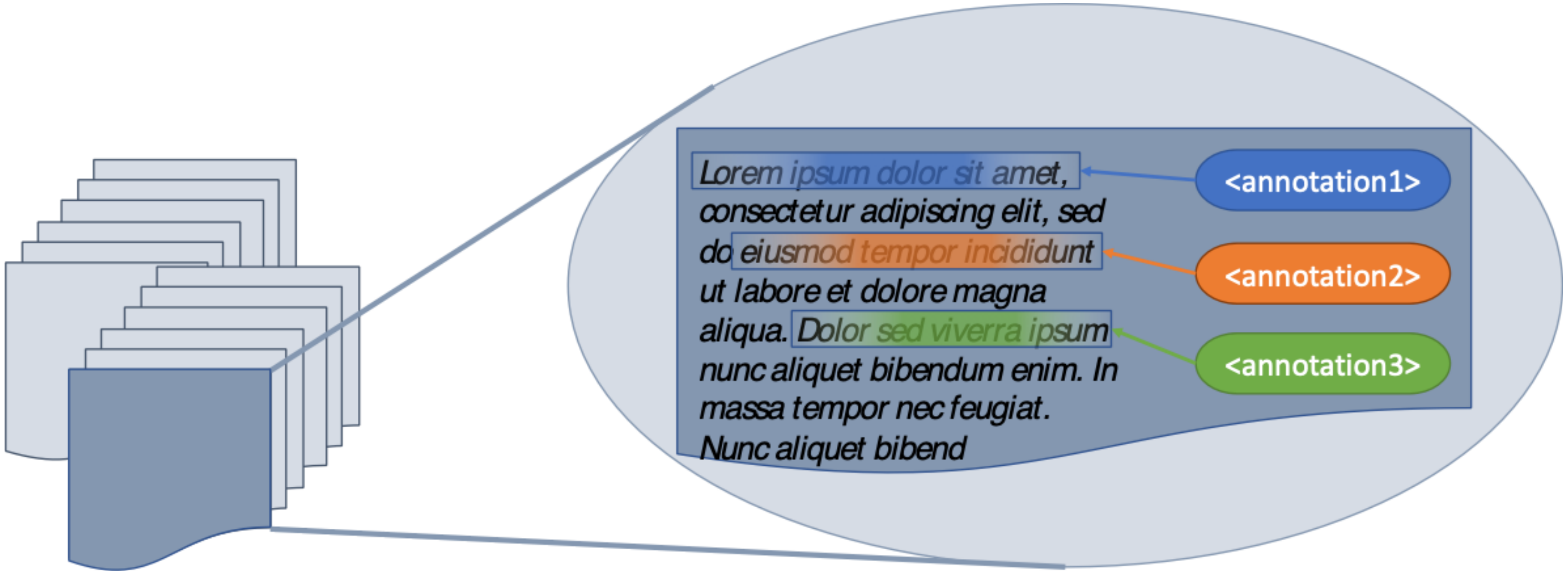

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Dolor sed viverra ipsum nunc aliquet bibendum enim. In massa tempor nec feugiat. Nunc aliquet bibend
<annotation1>
<annotation2>
<annotation3>

Subjective Content Descriptions (SCDs)

- … represent additional data for a document
- … are associated with a text span (window) in a document

Different Typs of SCDs:

- Additional definitions
- Links to external sources
- Relational data to clarify dependencies between entities

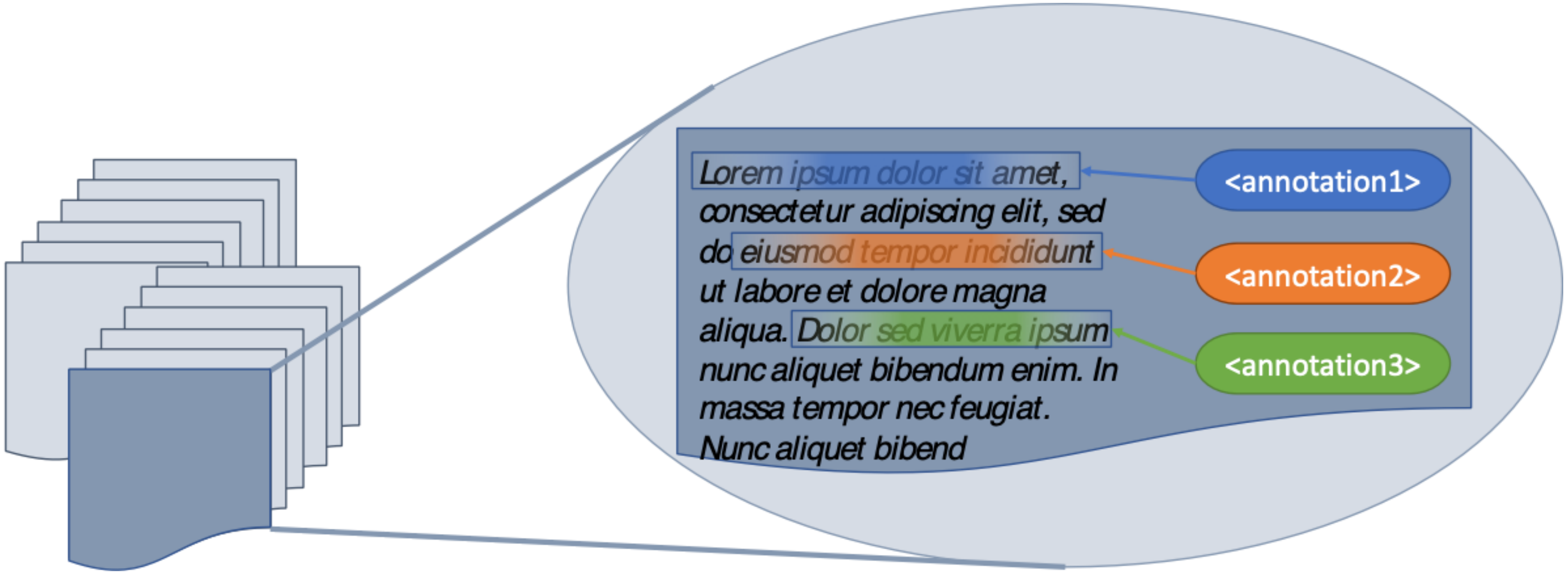

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Dolor sed viverra ipsum nunc aliquet bibendum enim. In massa tempor nec feugiat. Nunc aliquet bibend
<annotation1>
<annotation2>
<annotation3>

Subjective Content Descriptions (SCDs)

- … represent additional data for a document
- … are associated with a text span (window) in a document
- ... can take any form

Different Typs of SCDs:

- Additional definitions
- Links to external sources
- Relational data to clarify dependencies between entities

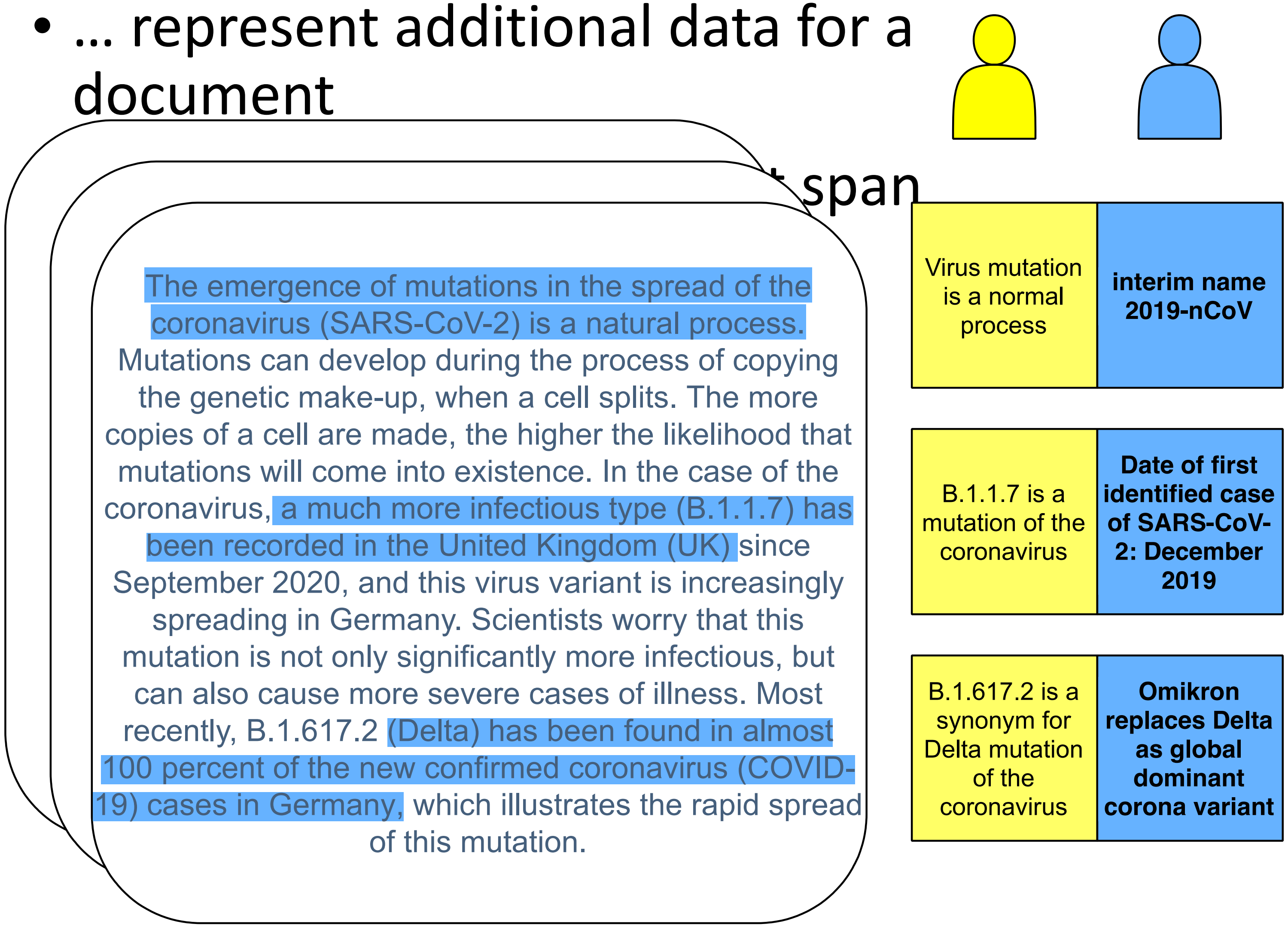


Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Dolor sed viverra ipsum nunc aliquet bibendum enim. In massa tempor nec feugiat. Nunc aliquet bibend

Subjective Content Descriptions (SCDs)

- … represent additional data for a document

- t span



The emergence of mutations in the spread of the coronavirus (SARS-CoV-2) is a natural process. Mutations can develop during the process of copying the genetic make-up, when a cell splits. The more copies of a cell are made, the higher the likelihood that mutations will come into existence. In the case of the coronavirus, a much more infectious type (B.1.1.7) has been recorded in the United Kingdom (UK) since September 2020, and this virus variant is increasingly spreading in Germany. Scientists worry that this mutation is not only significantly more infectious, but can also cause more severe cases of illness. Most recently, B.1.617.2 (Delta) has been found in almost 100 percent of the new confirmed coronavirus (COVID-19) cases in Germany, which illustrates the rapid spread of this mutation.

Virus mutation is a normal process	**interim name 2019-nCoV**
B.1.1.7 is a mutation of the coronavirus	**Date of first identified case of SARS-CoV-2: December 2019**
B.1.617.2 is a synonym for Delta mutation of the coronavirus	**Omikron replaces Delta as global dominant corona variant**

Different Typs of SCDs:

- Additional definitions
- Links to external sources
- Relational data to clarify dependencies between entities

Subjective Content Descriptions (SCDs)

- … represent additional data for a document
- … are associated with a text span (window) in a document
- ... can take any form
- Build SCD-word distribution matrix

→Use SCD-word distribution for additional tasks

Different Typs of SCDs:

- Additional definitions
- Links to external sources
- Relational data to clarify dependencies between entities

Subjective Content Descriptions (SCDs)

- … represent additional data for a document
- … are associated with a text span (window) in a document
- ... can take any form
- Build SCD-word distribution matrix

→Use SCD-word distribution for additional tasks

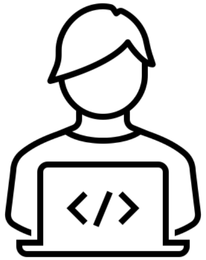
Different Typs of SCDs:

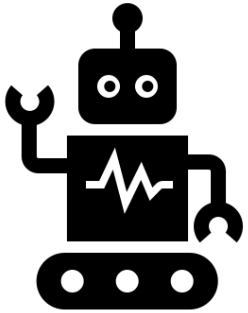
- Additional definitions
- Links to external sources
- Relational data to clarify dependencies between entities

Lead Scenario: Information Retrieval (IR) Agent

Lead Scenario: Information Retrieval (IR) Agent

- Agent‘s goal: Meet information need of a person

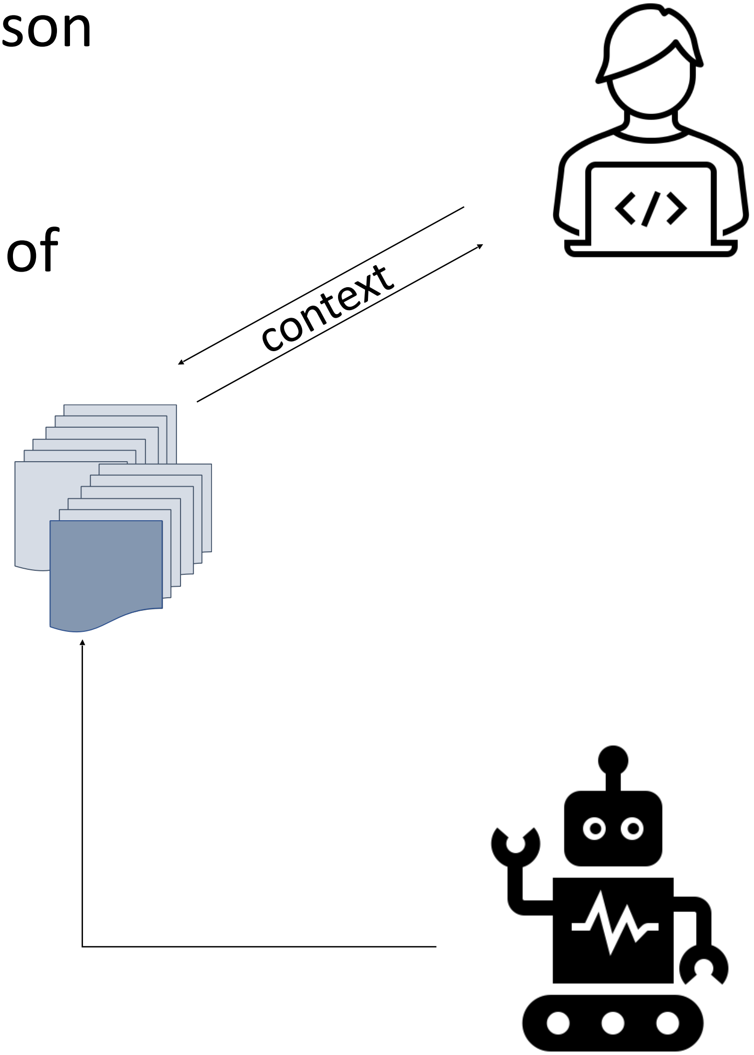




Lead Scenario: Information Retrieval (IR) Agent

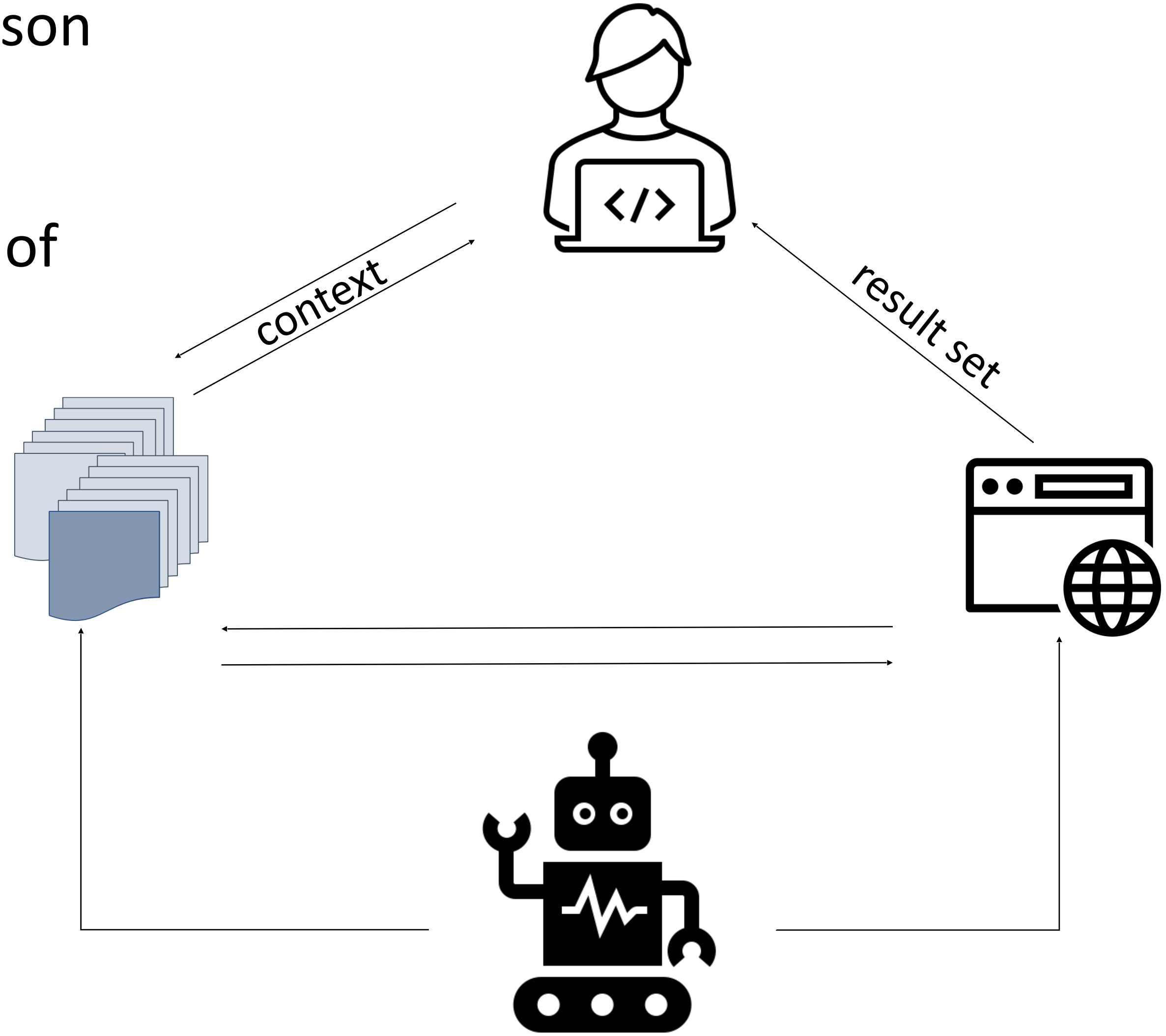
- Agent‘s goal: Meet information need of a person

- Agent is working on an IR-corpus that represents a model for the information need of a person



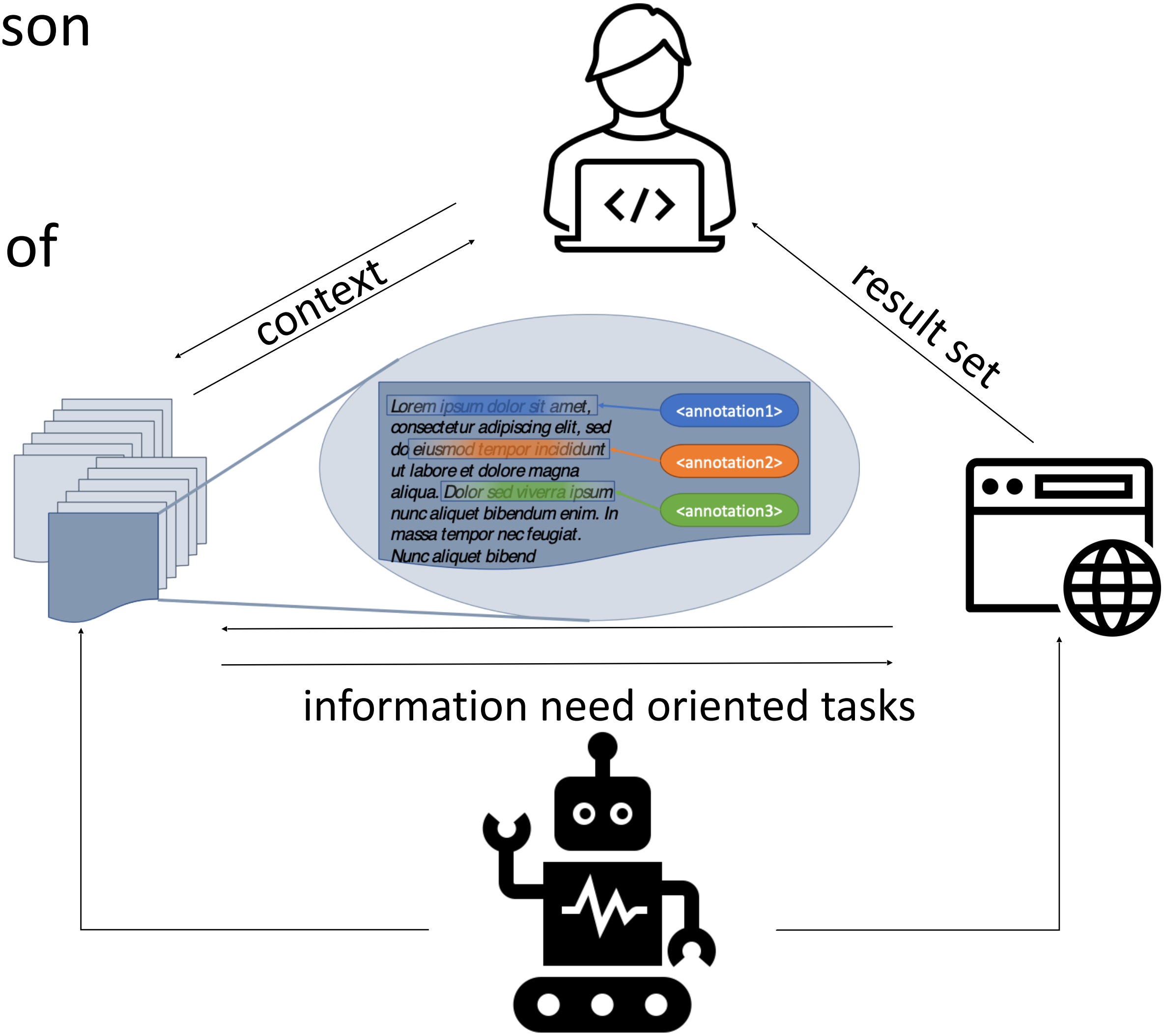
Lead Scenario: Information Retrieval (IR) Agent

- Agent‘s goal: Meet information need of a person
- Agent is working on an IR-corpus that represents a model for the information need of a person
- Agent optimizes the model to meet the information need



Lead Scenario: Information Retrieval (IR) Agent

- Agent‘s goal: Meet information need of a person
- Agent is working on an IR-corpus that represents a model for the information need of a person
- Agent optimizes the model to meet the information need
- Documents in the corpus are associated with annotations



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

Lead Scenario: Information Retrieval (IR) Agent

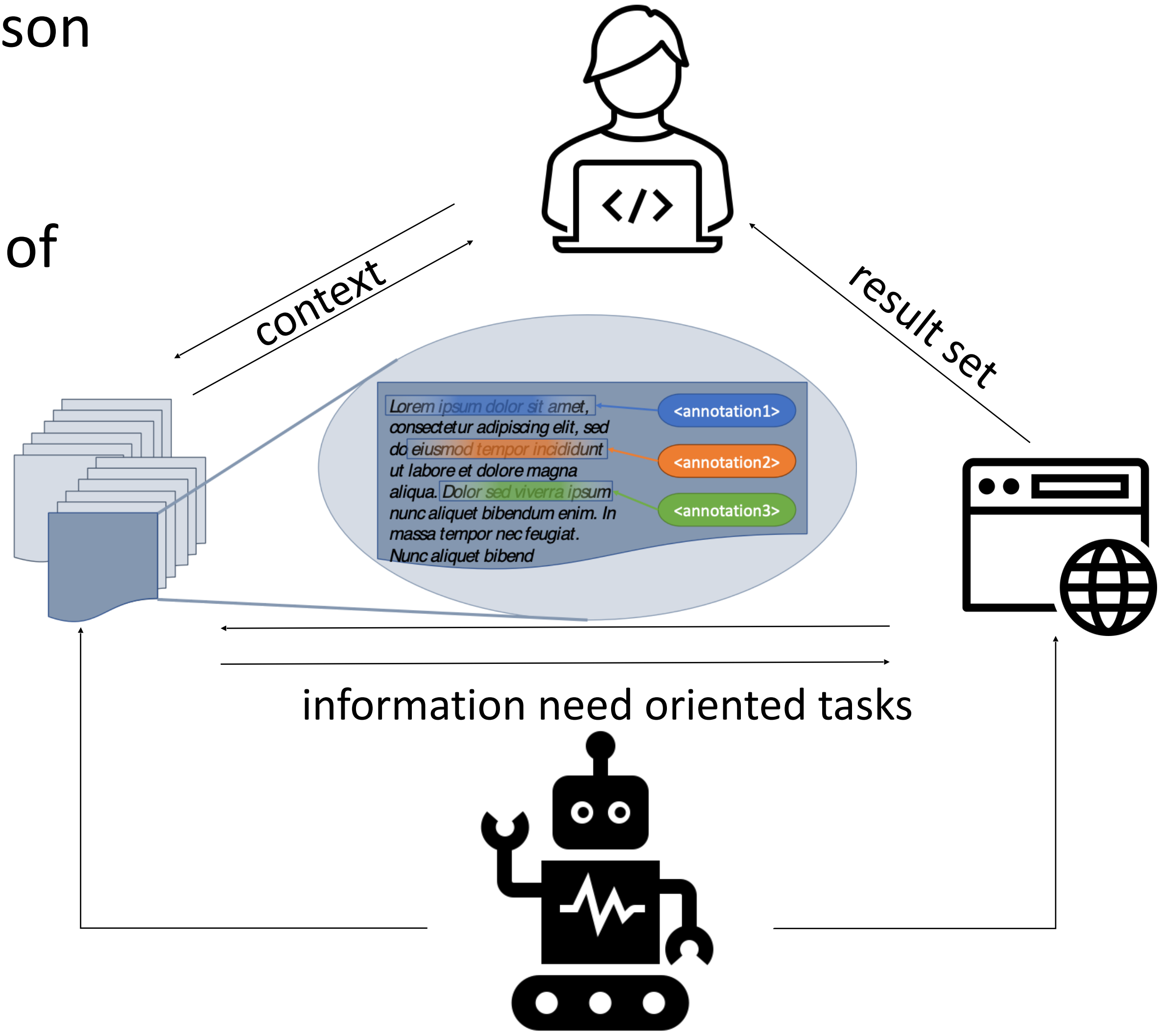
- Agent‘s goal: Meet information need of a person

- Agent is working on an IR-corpus that represents a model for the information need of a person

- Agent optimizes the model to meet the information need

- Documents in the corpus are associated with annotations

→ Subjective Content Descriptions (SCDs)



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

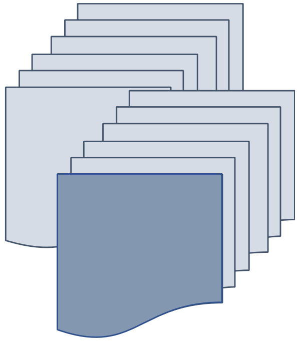
Foundations: SCD-Word Distribution [2]

- SCD-word distribution results from SCDs associated with *windows* in documents

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

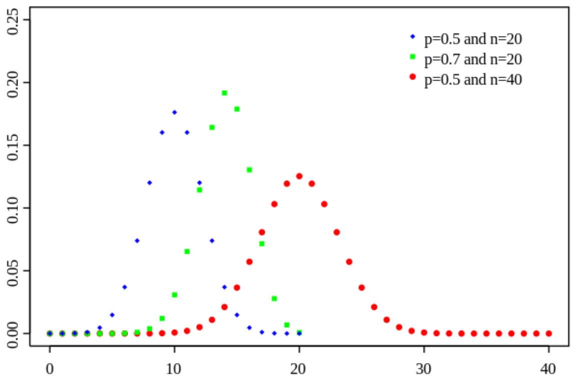
Foundations: SCD-Word Distribution [2]

- SCD-word distribution results from SCDs associated with *windows* in documents
- For each SCD estimate relative weighted frequency of words
- Binomial distribution to represent weights



Corpus $\mathscr{D}$

$$\delta(\mathscr{D}) = \begin{array}{c} \\ t_1 \\ t_2 \\ \vdots \\ t_m \end{array} \begin{array}{c} \begin{array}{ccccc} w_1 & w_2 & w_3 & \cdots & w_n \end{array} \\ \left[\begin{array}{ccccc} v_{1,1} & v_{1,2} & v_{1,3} & \cdots & v_{1,n} \\ v_{2,1} & v_{2,2} & v_{2,3} & \cdots & v_{2,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ v_{m,1} & v_{m,2} & v_{m,3} & \cdots & v_{m,n} \end{array}\right] \end{array}$$



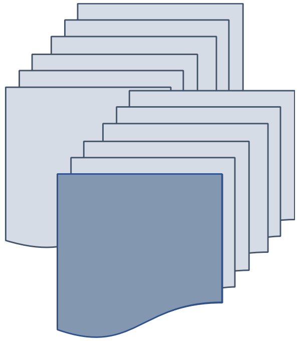
[2] Felix Kuhr, Tanya Braun, Magnus Bender, Ralf Möller: To Extend or not to Extend? Context-specific Corpus Enrichment. Proceedings of AI 2019: Advances in Artificial Intelligence, 2019

Foundations: SCD-Word Distribution [2]

- SCD-word distribution results from SCDs associated with *windows* in documents
- For each SCD estimate relative weighted frequency of words
- Binomial distribution to represent weights

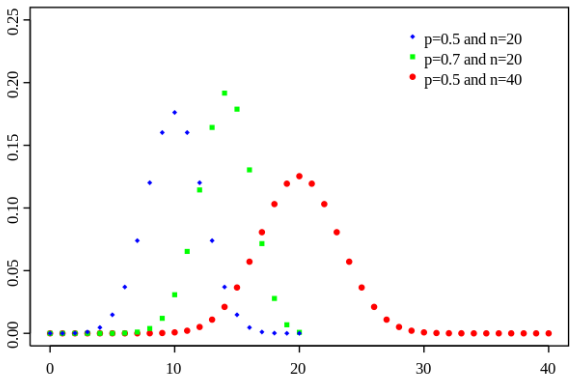
Algorithm 1 Forming SCD-word probability distribution matrix $\delta(\mathcal{D})$

```
function BUILDMATRIX(Corpus $\mathcal{D}$)
    Input: Corpus $\mathcal{D}$
    Output: SCD-word probability distribution matrix $\delta(\mathcal{D})$
    Initialize an $m \times V$ matrix $\delta(\mathcal{D})$ with zeros
    for each $d \in \mathcal{D}$ do
        for each $t \in T(d)$ do
            for $\rho$ of $t$ do
                for each $w \in win_{d,\rho}$ do
                    $\delta(\mathcal{D})[t][w] \mathrel{+}= I(w, win_{d,\rho})$
    Normalize $\delta(\mathcal{D})[t]$
    return $\delta(\mathcal{D})$
```



Corpus $\mathscr{D}$

$$\delta(\mathscr{D}) = \begin{array}{c|ccccc} & w_1 & w_2 & w_3 & \cdots & w_n \\ \hline t_1 & v_{1,1} & v_{1,2} & v_{1,3} & \cdots & v_{1,n} \\ t_2 & v_{2,1} & v_{2,2} & v_{2,3} & \cdots & v_{2,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ t_m & v_{m,1} & v_{m,2} & v_{m,3} & \cdots & v_{m,n} \end{array}$$



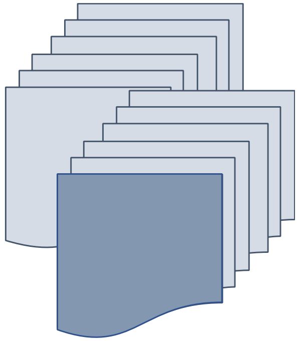
[2] Felix Kuhr, Tanya Braun, Magnus Bender, Ralf Möller: To Extend or not to Extend? Context-specific Corpus Enrichment. Proceedings of AI 2019: Advances in Artificial Intelligence, 2019

Foundations: SCD-Word Distribution [2]

- SCD-word distribution results from SCDs associated with *windows* in documents
- For each SCD estimate relative weighted frequency of words
- Binomial distribution to represent weights

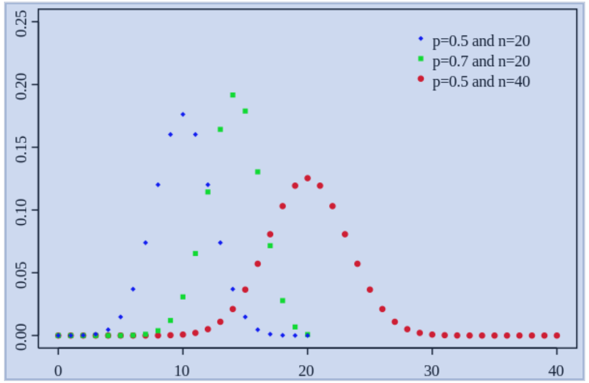
Algorithm 1 Forming SCD-word probability distribution matrix $\delta(\mathcal{D})$

```
function BUILDMATRIX(Corpus $\mathcal{D}$)
    Input: Corpus $\mathcal{D}$
    Output: SCD-word probability distribution matrix $\delta(\mathcal{D})$
    Initialize an $m \times V$ matrix $\delta(\mathcal{D})$ with zeros
    for each $d \in \mathcal{D}$ do
        for each $t \in T(d)$ do
            for $\rho$ of $t$ do
                for each $w \in win_{d,\rho}$ do
                    $\delta(\mathcal{D})[t][w] \mathrel{+}= I(w, win_{d,\rho})$
    Normalize $\delta(\mathcal{D})[t]$
    return $\delta(\mathcal{D})$
```



Corpus $\mathscr{D}$

$$\delta(\mathscr{D}) = \begin{array}{c|ccccc} & w_1 & w_2 & w_3 & \cdots & w_n \\ \hline t_1 & v_{1,1} & v_{1,2} & v_{1,3} & \cdots & v_{1,n} \\ t_2 & v_{2,1} & v_{2,2} & v_{2,3} & \cdots & v_{2,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ t_m & v_{m,1} & v_{m,2} & v_{m,3} & \cdots & v_{m,n} \end{array}$$



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

Most-Probably Suited SCDs (MPSCDs)[2]

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

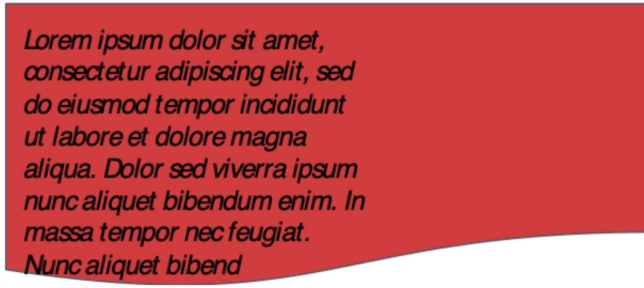
Most-Probably Suited SCDs (MPSCDs) [2]

- Given: Sequence of text windows in a (new) document

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

Most-Probably Suited SCDs (MPSCDs)[2]

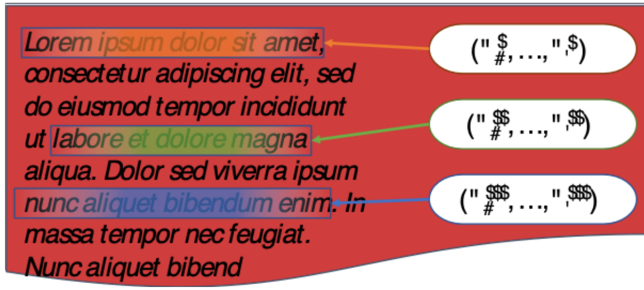
- Given: Sequence of text windows in a (new) document
- Words in window define word distribution (t')



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

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

Most-Probably Suited SCDs (MPSCDs) [2]

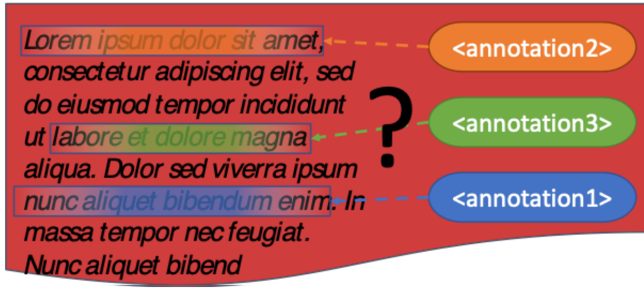

Lorem ipsum dolor sit amet,
consectetur adipiscing elit, sed
do eiusmod tempor incididunt
ut labore et dolore magna
aliqua. Dolor sed viverra ipsum
nunc aliquet bibendum enim. In
massa tempor nec feugiat.
Nunc aliquet bibend
(" $#, …, " $)
(" $$#, …, " $$)
(" $$$#, …, " $$$)

- <u>Given:</u> Sequence of text windows in a (new) document
- Words in window define word distribution (t')
- <u>Goal:</u> Estimate MPSCDs for windows in document

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

Most-Probably Suited SCDs (MPSCDs) [2]

- Given: Sequence of text windows in a (new) document
- Words in window define word distribution (t')
- Goal: Estimate MPSCDs for windows in document
- Each vector v_i defined in SCD-word matrix $\delta(D)$ defines an angle with window word vector t'

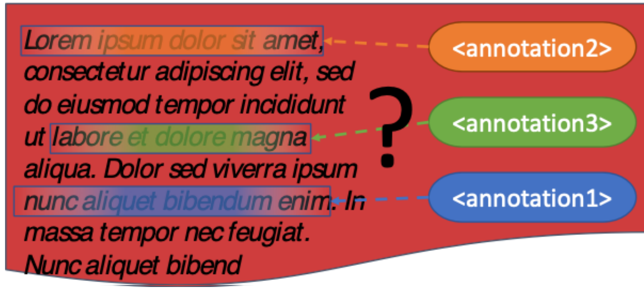


Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Dolor sed viverra ipsum nunc aliquet bibendum enim. In massa tempor nec feugiat. Nunc aliquet bibend

$$\begin{array}{c|ccccc} & w_1 & w_2 & w_3 & \cdots & w_n \\ \hline t_1 & v_{1,1} & v_{1,2} & v_{1,3} & \cdots & v_{1,n} \\ t_2 & v_{2,1} & v_{2,2} & v_{2,3} & \cdots & v_{2,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ t_m & v_{m,1} & v_{m,2} & v_{m,3} & \cdots & v_{m,n} \end{array}$$

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

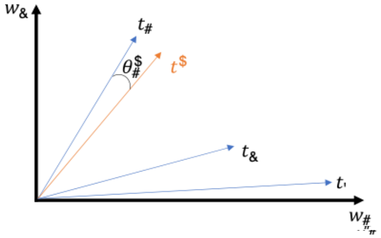
Most-Probably Suited SCDs (MPSCDs) [2]

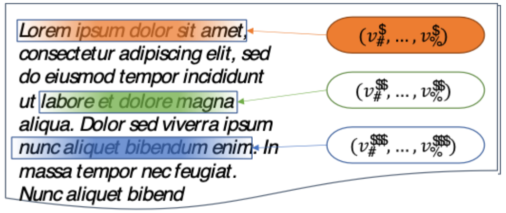


Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Dolor sed viverra ipsum nunc aliquet bibendum enim. In massa tempor nec feugiat. Nunc aliquet bibend

- Given: Sequence of text windows in a (new) document
- Words in window define word distribution (t')
- Goal: Estimate MPSCDs for windows in document
- Each vector v_i defined in SCD-word matrix $\delta(D)$ defines an angle with window word vector t'
- Define function $MPSCD(M,\ t')$
  - Provides SCD t_i that is associated with SCD-word matrix vector v_i with smallest angle to window word vector t'

$$
\begin{array}{c|ccccc|}
 & w_1 & w_2 & w_3 & \cdots & w_n \\
t_1 & v_{1,1} & v_{1,2} & v_{1,3} & \cdots & v_{1,n} \\
t_2 & v_{2,1} & v_{2,2} & v_{2,3} & \cdots & v_{2,n} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
t_m & v_{m,1} & v_{m,2} & v_{m,3} & \cdots & v_{m,n}
\end{array}
$$



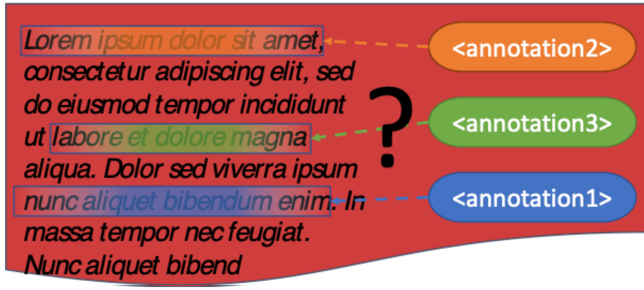


Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Dolor sed viverra ipsum nunc aliquet bibendum enim. In massa tempor nec feugiat. Nunc aliquet bibend

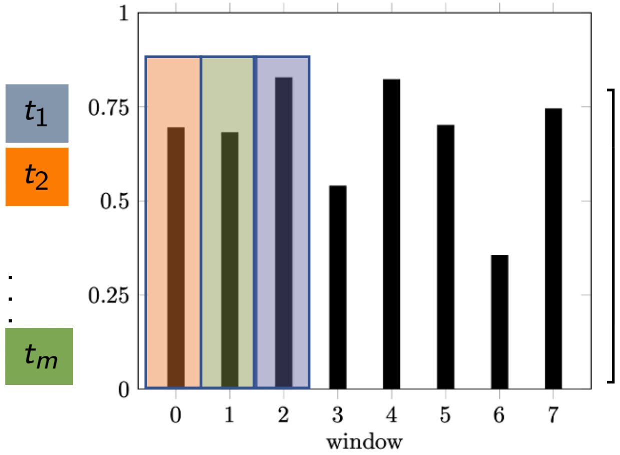
$(v_#^$, \ldots, v_%^$)$
$(v_#^{$$}, \ldots, v_%^{$$})$
$(v_#^{$$$}, \ldots, v_%^{$$$})$

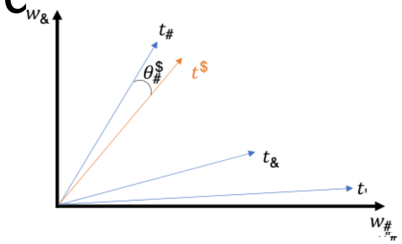
[2] Felix Kuhr, Tanya Braun, Magnus Bender, Ralf Möller: To Extend or not to Extend? Context-specific Corpus Enrichment. Proceedings of AI 2019: Advances in Artificial Intelligence, 2019

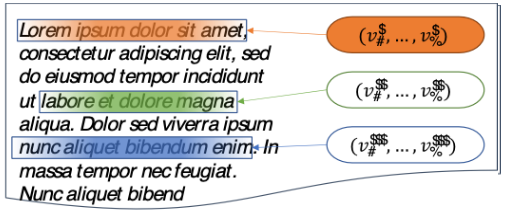
Most-Probably Suited SCDs (MPSCDs) [2]


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>

- <u>Given:</u> Sequence of text windows in a (new) document
- Words in window define word distribution (t')
- <u>Goal:</u> Estimate MPSCDs for windows in document
- Each vector v_i defined in SCD-word matrix $\delta(D)$ defines an angle with window word vector t'
- Define function $MPSCD(M,\ t')$
  - Provides SCD t_i that is associated with SCD-word matrix vector v_i with smallest angle to window word vector t'
- MPSCD with similarity measure is applied to each window of a document






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
$(v_\#^\$, \ldots, v_\%^\$)$
$(v_\#^{\$\$}, \ldots, v_\%^{\$\$})$
$(v_\#^{\$\$\$}, \ldots, v_\%^{\$\$\$})$

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

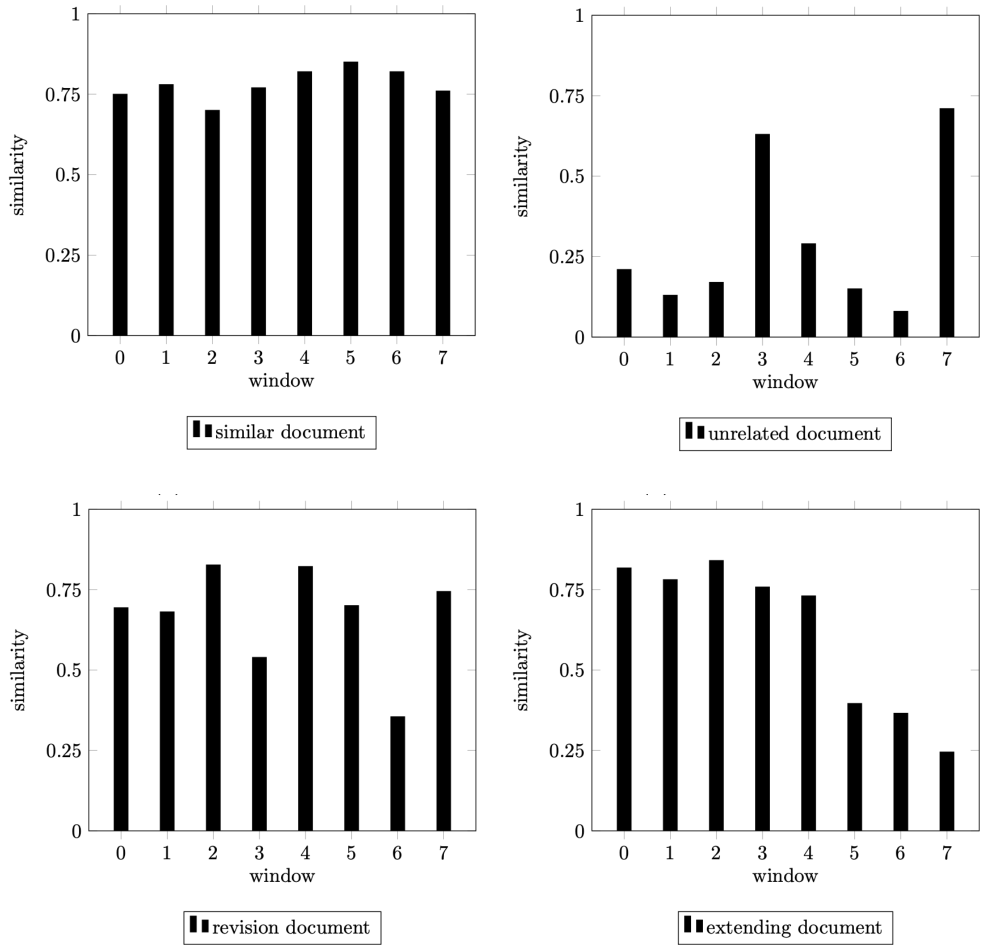
Context-specific Corpus Enrichment [3]

- Goal: Add new documents to IR corpus with an initial set of SCDs already associated with documents in the corpus

[3] Felix Kuhr, Tanya Braun, Ralf Möller: Augmenting and Automating Corpus Enrichment. Proceedings of the 14th IEEE International Conference on Semantic Computing (ICSC-20), 2020

Context-specific Corpus Enrichment [3]

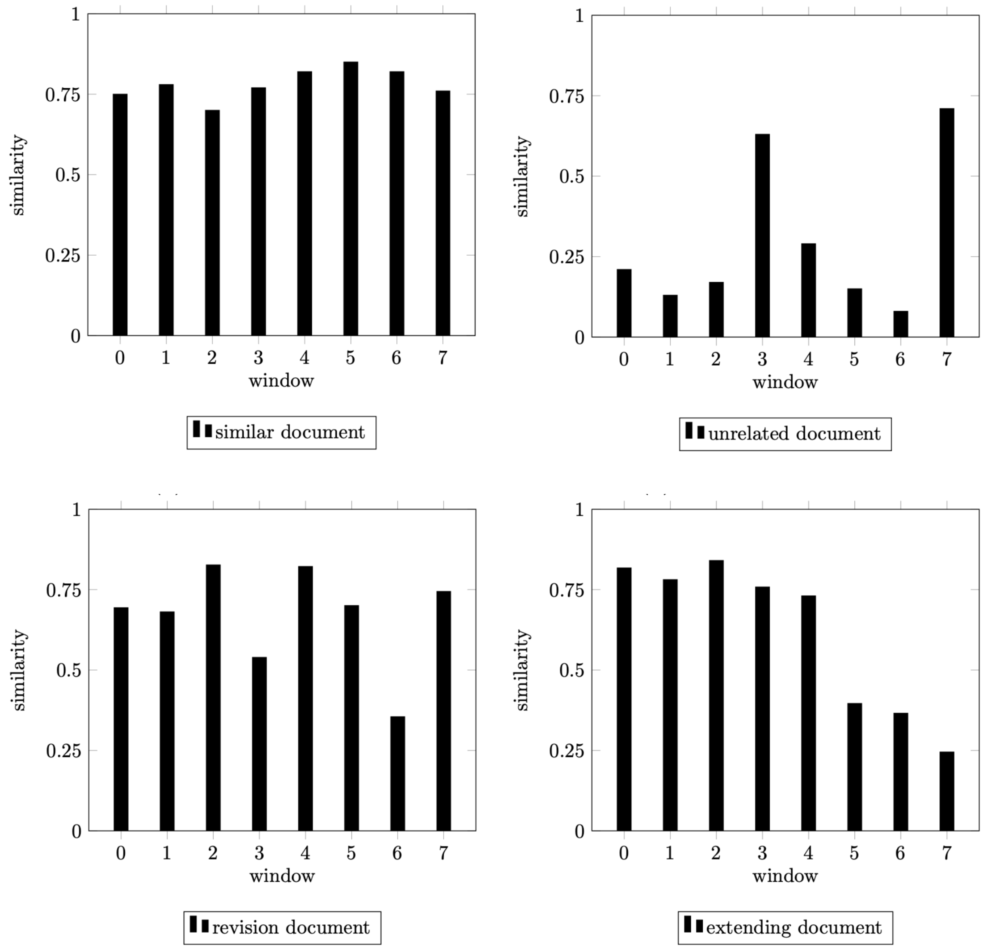
- Goal: Add new documents to IR corpus with an initial set of SCDs already associated with documents in the corpus
- Different Categories: sim, unrel, rev, ext



[3] Felix Kuhr, Tanya Braun, Ralf Möller: Augmenting and Automating Corpus Enrichment. Proceedings of the 14th IEEE International Conference on Semantic Computing (ICSC-20), 2020

Context-specific Corpus Enrichment [3]

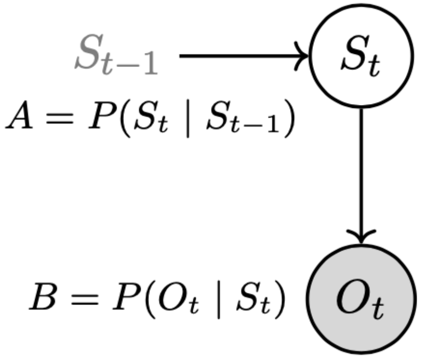
- Goal: Add new documents to IR corpus with an initial set of SCDs already associated with documents in the corpus
- Different Categories: sim, unrel, rev, ext
- Given: 4 category HMMs, each associated with a category label



[3] Felix Kuhr, Tanya Braun, Ralf Möller: Augmenting and Automating Corpus Enrichment. Proceedings of the 14th IEEE International Conference on Semantic Computing (ICSC-20), 2020

Context-specific Corpus Enrichment [3]

- Goal: Add new documents to IR corpus with an initial set of SCDs already associated with documents in the corpus
- Different Categories: sim, unrel, rev, ext
- Given: 4 category HMMs, each associated with a category label
- HMM Learning by using Baum-Welch Algorithm

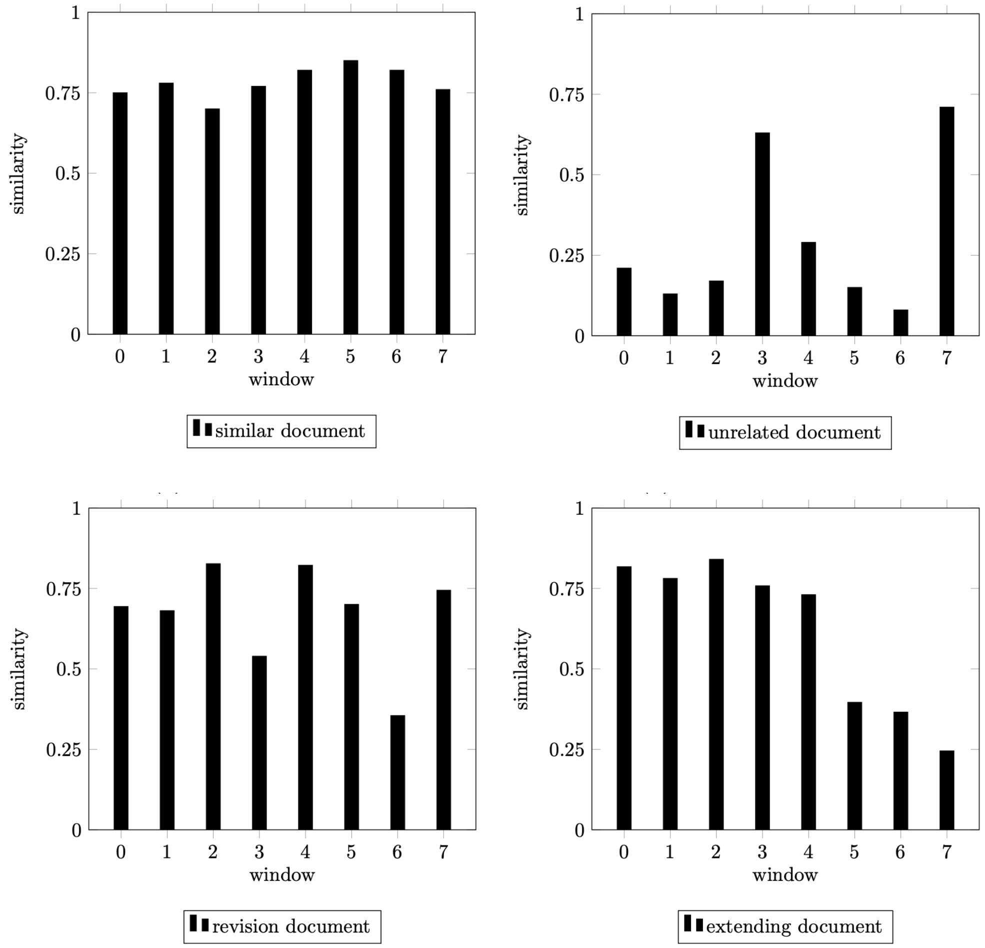


$A = P(S_t \mid S_{t-1})$

$B = P(O_t \mid S_t)$

Related content
or
Unrelated content

MPSCD similarity value
of the window (discretized)

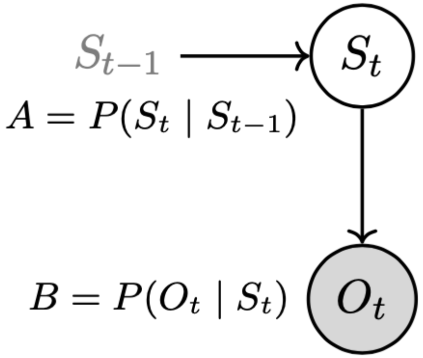


[3] Felix Kuhr, Tanya Braun, Ralf Möller: Augmenting and Automating Corpus Enrichment. Proceedings of the 14th IEEE International Conference on Semantic Computing (ICSC-20), 2020

Context-specific Corpus Enrichment [3]

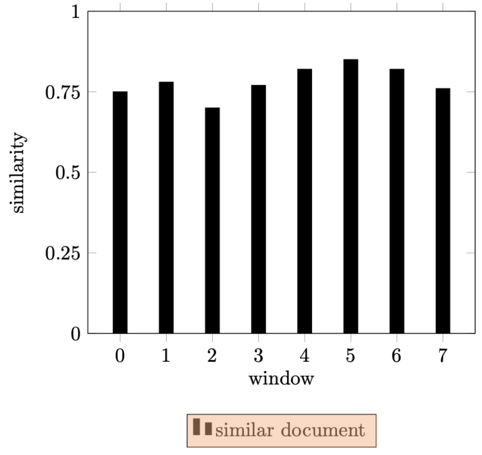
- Goal: Add new documents to IR corpus with an initial set of SCDs already associated with documents in the corpus
- Different Categories: sim, unrel, rev, ext
- Given: 4 category HMMs, each associated with a category label
- HMM Learning by using Baum-Welch Algorithm

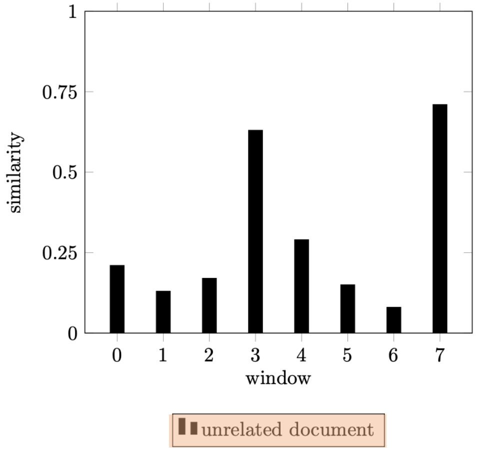
User input required

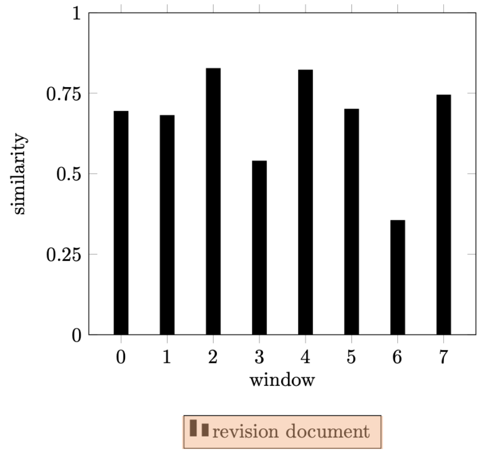


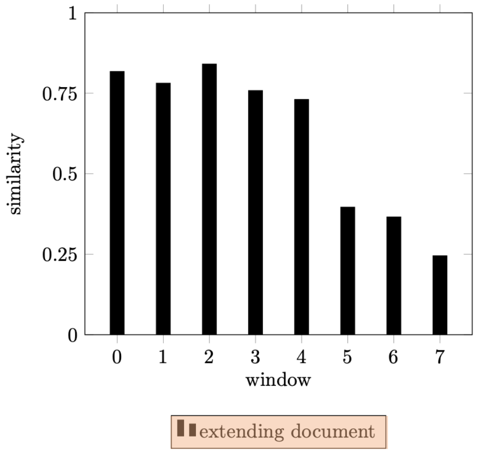
Related content
or
Unrelated content

MPSCD similarity value of the window (discretized)









[3] Felix Kuhr, Tanya Braun, Ralf Möller: Augmenting and Automating Corpus Enrichment. Proceedings of the 14th IEEE International Conference on Semantic Computing (ICSC-20), 2020

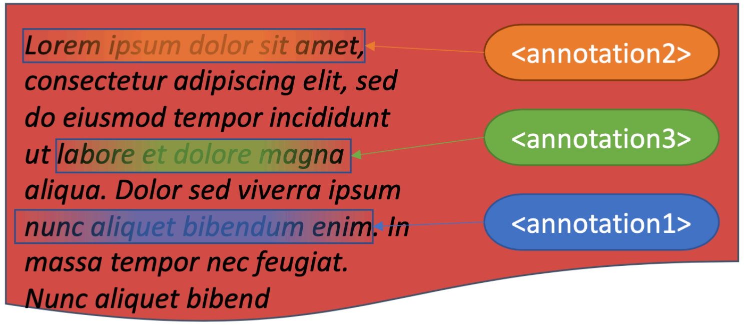
Context-specific Corpus Enrichment - Decision Making Process

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

Given: new document

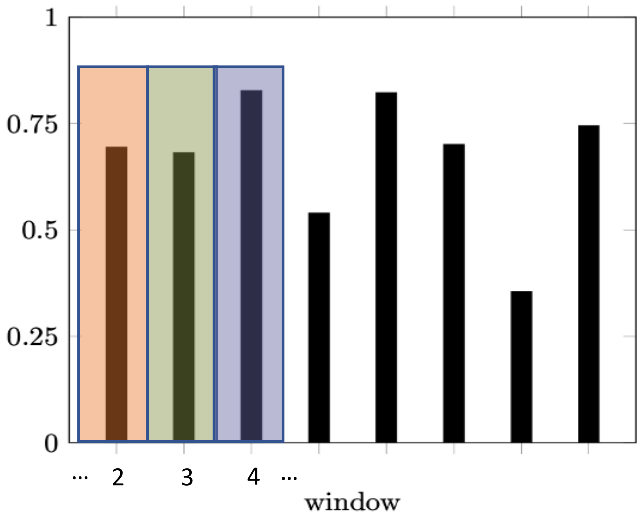
$$
\begin{array}{c|ccccc}
 & w_1 & w_2 & w_3 & \cdots & w_n \\
\hline
t_1 & v_{1,1} & v_{1,2} & v_{1,3} & \cdots & v_{1,n} \\
t_2 & v_{2,1} & v_{2,2} & v_{2,3} & \cdots & v_{2,n} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
t_m & v_{m,1} & v_{m,2} & v_{m,3} & \cdots & v_{m,n}
\end{array}
$$

Given: SCD-word distribution of IR corpus

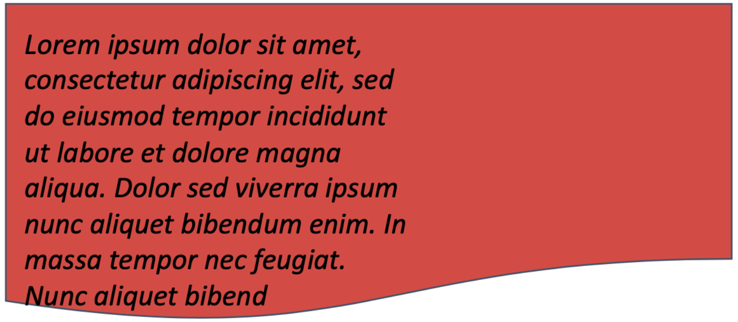


Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Dolor sed viverra ipsum nunc aliquet bibendum enim. In massa tempor nec feugiat. Nunc aliquet bibend

Determine the MPSCD sequence for the window sequence of the new document based on available SCD-word distribution



Context-specific Corpus Enrichment - Decision Making Process

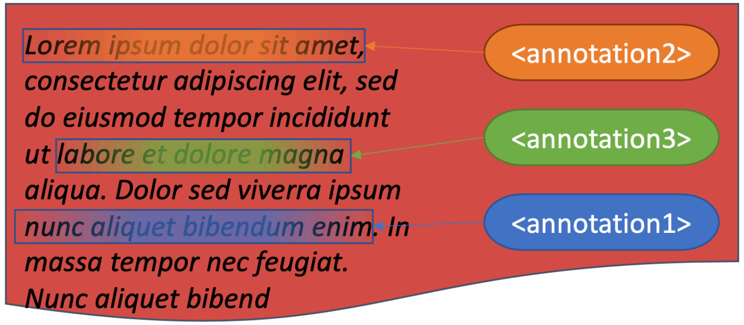


Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Dolor sed viverra ipsum nunc aliquet bibendum enim. In massa tempor nec feugiat. Nunc aliquet bibend

Given: new document

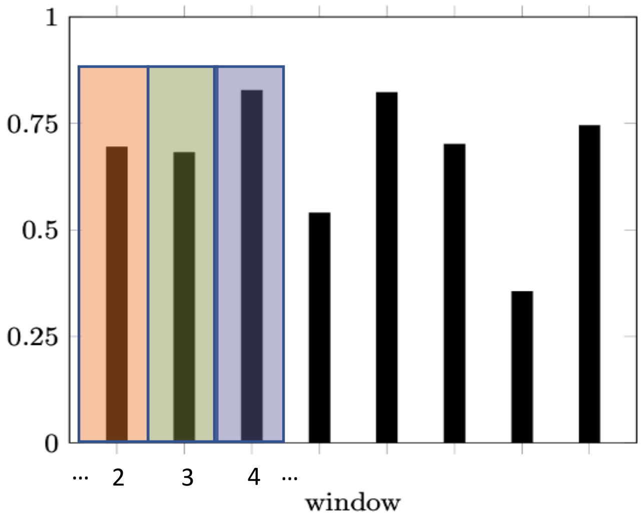
$$\begin{array}{c|ccccc} & w_1 & w_2 & w_3 & \cdots & w_n \\ \hline t_1 & v_{1,1} & v_{1,2} & v_{1,3} & \cdots & v_{1,n} \\ t_2 & v_{2,1} & v_{2,2} & v_{2,3} & \cdots & v_{2,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ t_m & v_{m,1} & v_{m,2} & v_{m,3} & \cdots & v_{m,n} \end{array}$$

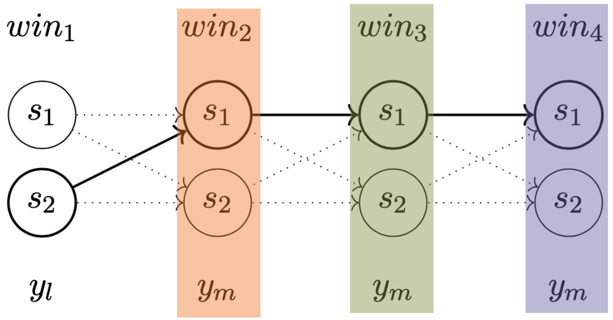
Given: SCD-word distribution of IR corpus



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

Determine the MPSCD sequence for the window sequence of the new document based on available SCD-word distribution

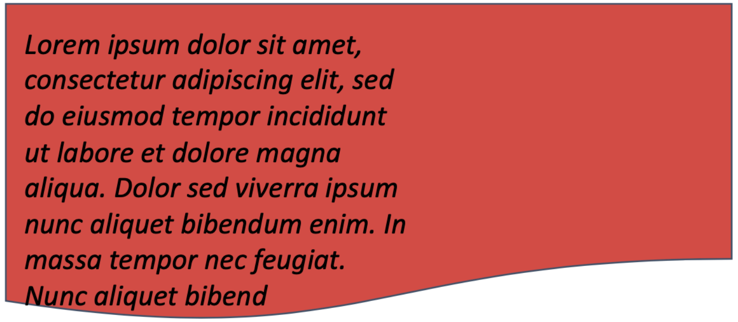




Discretize similarity values:

y_l: 0 – 0.3, y_m: 0.3 – 0.75, y_h: 0.75 – 1

Context-specific Corpus Enrichment - Decision Making Process

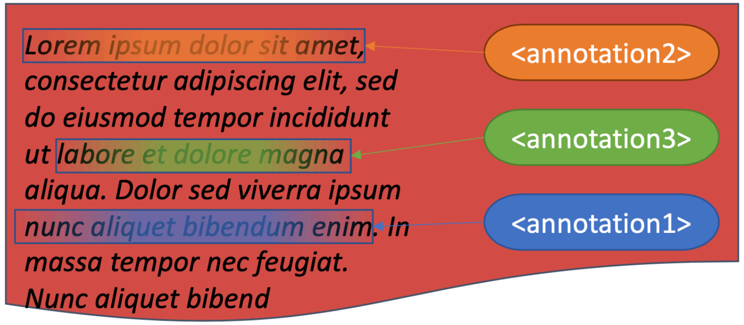


Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Dolor sed viverra ipsum nunc aliquet bibendum enim. In massa tempor nec feugiat. Nunc aliquet bibend

Given: new document

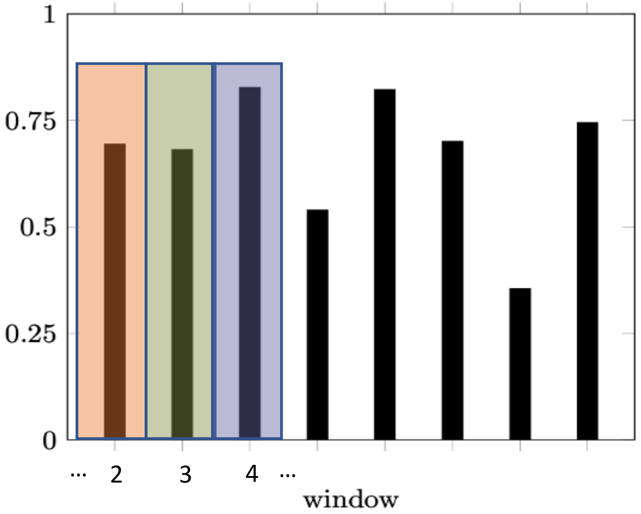
$$\begin{array}{c|ccccc} & w_1 & w_2 & w_3 & \cdots & w_n \\ t_1 & v_{1,1} & v_{1,2} & v_{1,3} & \cdots & v_{1,n} \\ t_2 & v_{2,1} & v_{2,2} & v_{2,3} & \cdots & v_{2,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ t_m & v_{m,1} & v_{m,2} & v_{m,3} & \cdots & v_{m,n} \end{array}$$

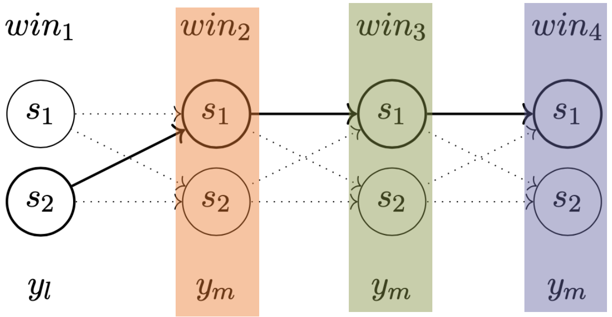
Given: SCD-word distribution of IR corpus



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

Determine the MPSCD sequence for the window sequence of the new document based on available SCD-word distribution



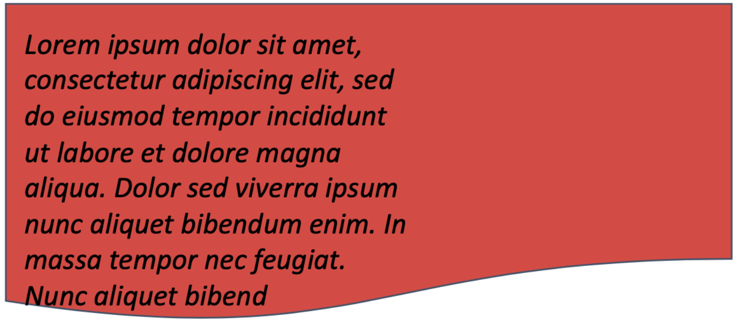


Discretize similarity values:

y_l: 0 – 0.3, y_m: 0.3 – 0.75, y_h: 0.75 – 1

Determine MPE sequence w.r.t. each category HMM on sequence of MPSCD similarity values.

Context-specific Corpus Enrichment - Decision Making Process

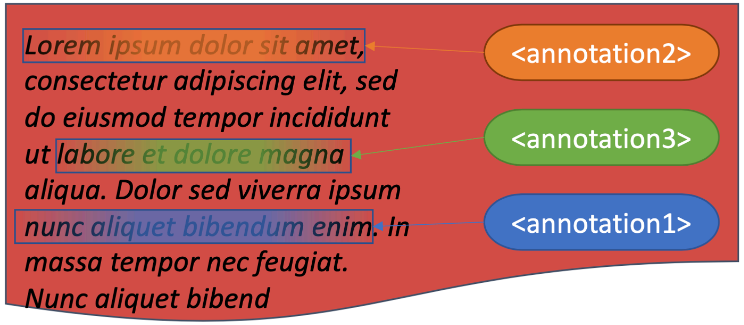


Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Dolor sed viverra ipsum nunc aliquet bibendum enim. In massa tempor nec feugiat. Nunc aliquet bibend

Given: new document

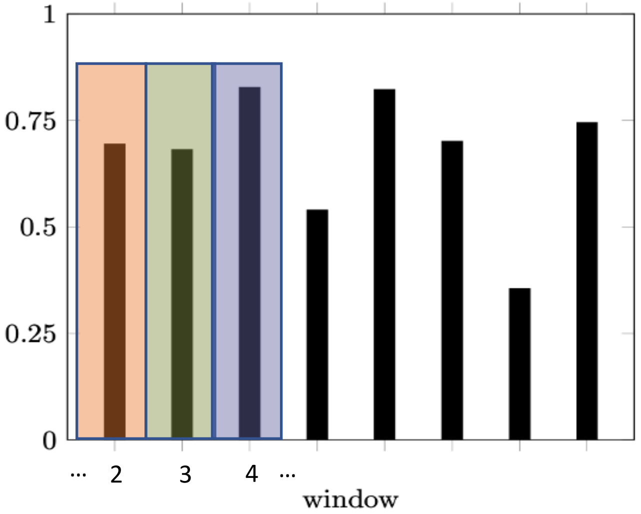
$$\begin{array}{cccccc} & w_1 & w_2 & w_3 & \cdots & w_n \\ t_1 & v_{1,1} & v_{1,2} & v_{1,3} & \cdots & v_{1,n} \\ t_2 & v_{2,1} & v_{2,2} & v_{2,3} & \cdots & v_{2,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ t_m & v_{m,1} & v_{m,2} & v_{m,3} & \cdots & v_{m,n} \end{array}$$

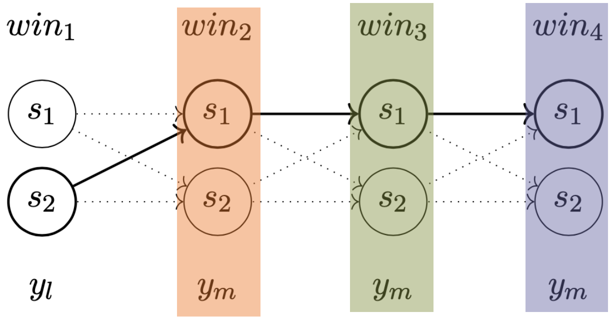
Given: SCD-word distribution of IR corpus



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

Determine the MPSCD sequence for the window sequence of the new document based on available SCD-word distribution





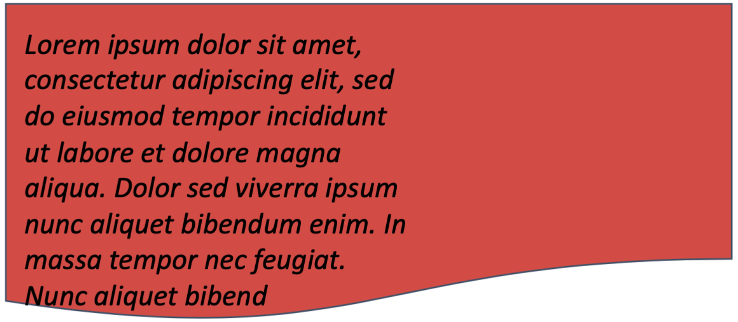
Discretize similarity values:

y_l: $0 - 0.3$, y_m: $0.3 - 0.75$, y_h: $0.75 - 1$

Determine MPE sequence w.r.t. each category HMM on sequence of MPSCD similarity values.

Take category of HMM with most-likely MPE sequence as classification

Context-specific Corpus Enrichment - Decision Making Process

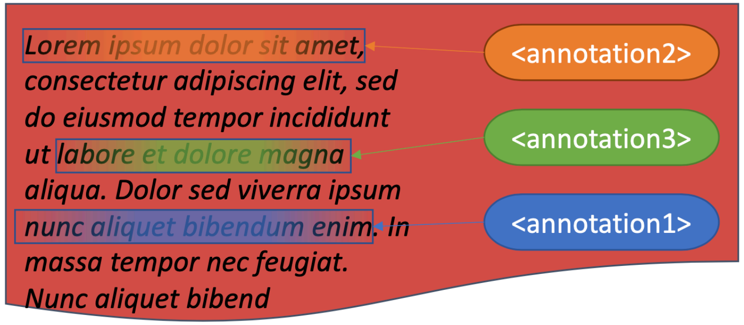


Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Dolor sed viverra ipsum nunc aliquet bibendum enim. In massa tempor nec feugiat. Nunc aliquet bibend

Given: new document

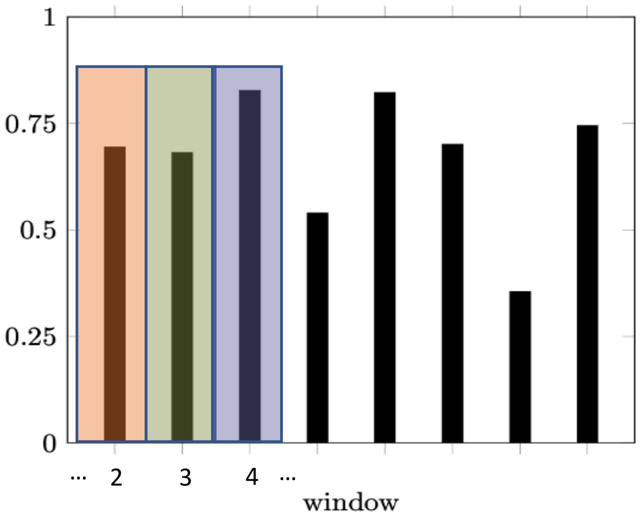
$$\begin{array}{c|ccccc} & w_1 & w_2 & w_3 & \cdots & w_n \\ \hline t_1 & v_{1,1} & v_{1,2} & v_{1,3} & \cdots & v_{1,n} \\ t_2 & v_{2,1} & v_{2,2} & v_{2,3} & \cdots & v_{2,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ t_m & v_{m,1} & v_{m,2} & v_{m,3} & \cdots & v_{m,n} \end{array}$$

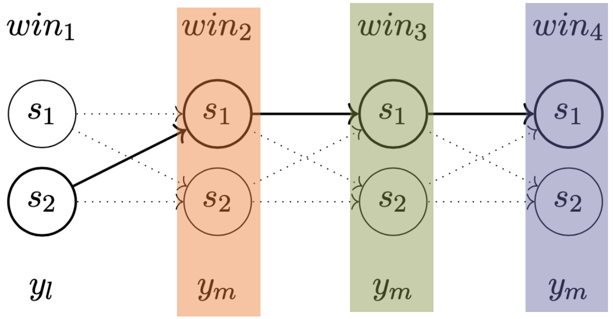
Given: SCD-word distribution of IR corpus



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

Determine the MPSCD sequence for the window sequence of the new document based on available SCD-word distribution





Discretize similarity values:

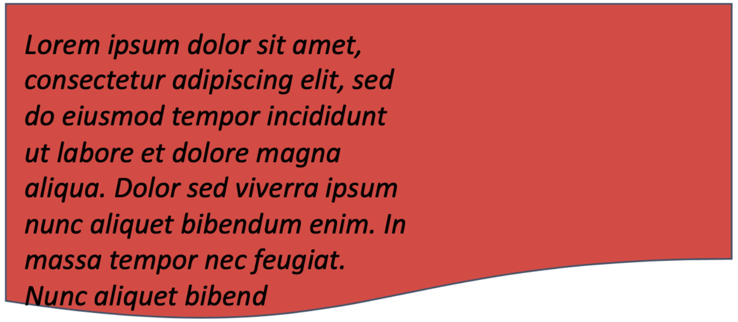
y_l: $0 - 0.3$, y_m: $0.3 - 0.75$, y_h: $0.75 - 1$

Determine MPE sequence w.r.t. each category HMM on sequence of MPSCD similarity values.

Take category of HMM with most-likely MPE sequence as classification

Extend corpus based on document category and transfer SCDs above a threshold

Context-specific Corpus Enrichment - Decision Making Process

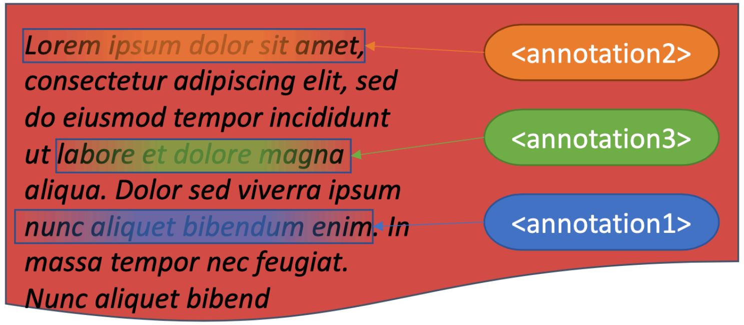


Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Dolor sed viverra ipsum nunc aliquet bibendum enim. In massa tempor nec feugiat. Nunc aliquet bibend

Given: new document

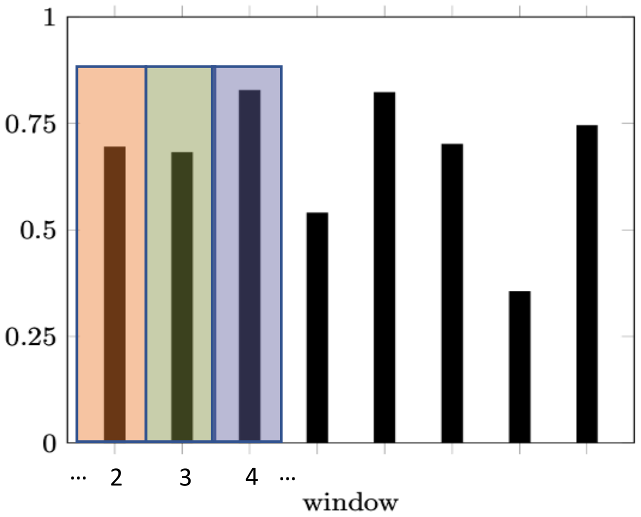
$$\begin{array}{c|ccccc} & w_1 & w_2 & w_3 & \cdots & w_n \\ t_1 & v_{1,1} & v_{1,2} & v_{1,3} & \cdots & v_{1,n} \\ t_2 & v_{2,1} & v_{2,2} & v_{2,3} & \cdots & v_{2,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ t_m & v_{m,1} & v_{m,2} & v_{m,3} & \cdots & v_{m,n} \end{array}$$

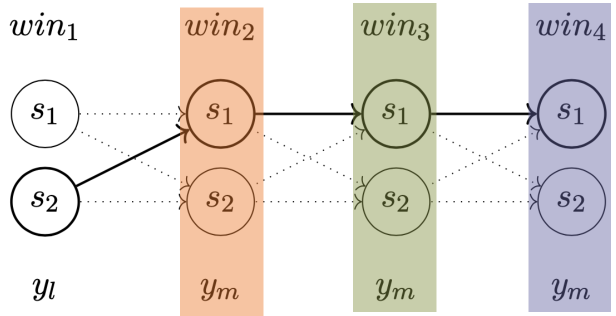
Given: SCD-word distribution of IR corpus



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

Determine the MPSCD sequence for the window sequence of the new document based on available SCD-word distribution





Focus on SCD similarity values

Discretize similarity values:

y_l: $0 - 0.3$, y_m: $0.3 - 0.75$, y_h: $0.75 - 1$

Determine MPE sequence w.r.t. each category HMM on sequence of MPSCD similarity values.

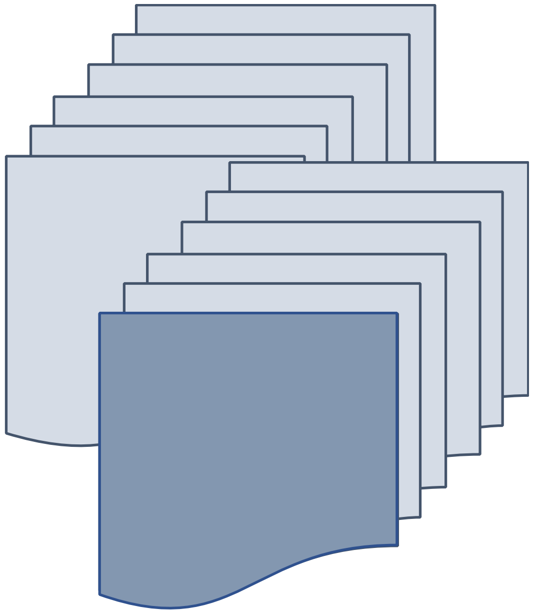
Take category of HMM with most-likely MPE sequence as classification

Focus on content of SCDs

Extend corpus based on document category and transfer SCDs above a threshold

[4]

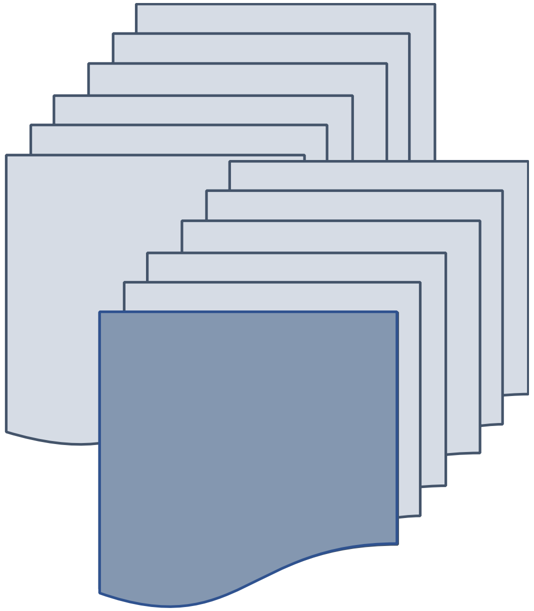
Corpus-Driven Document Enrichment using SCDs



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

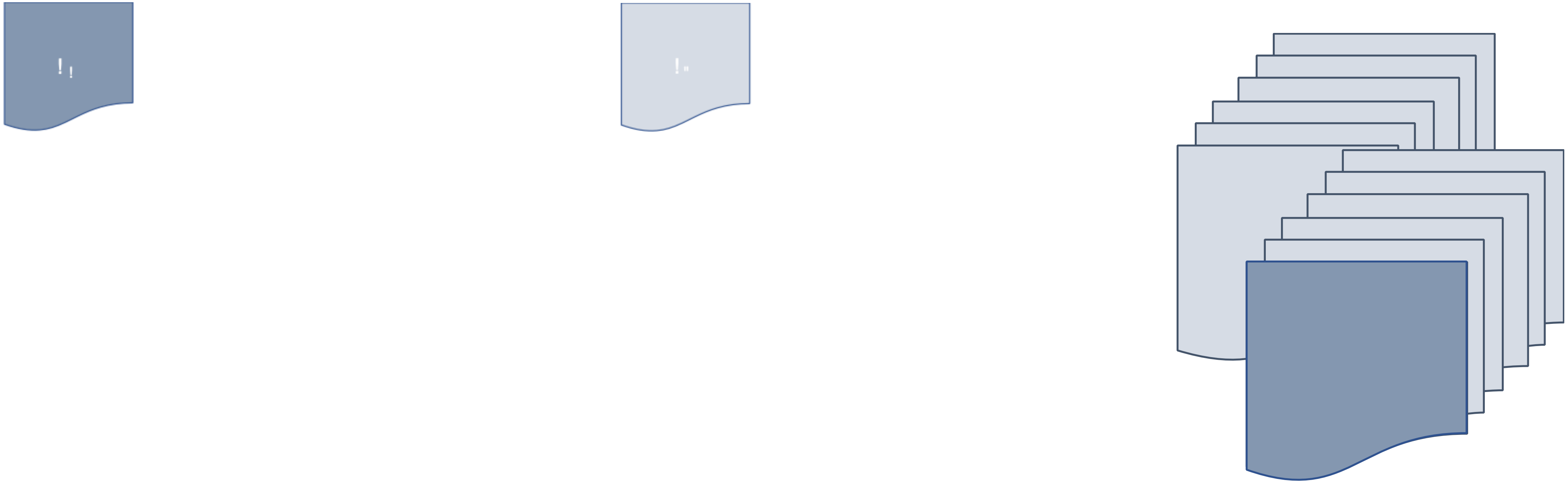
[4]

Corpus-Driven Document Enrichment using SCDs



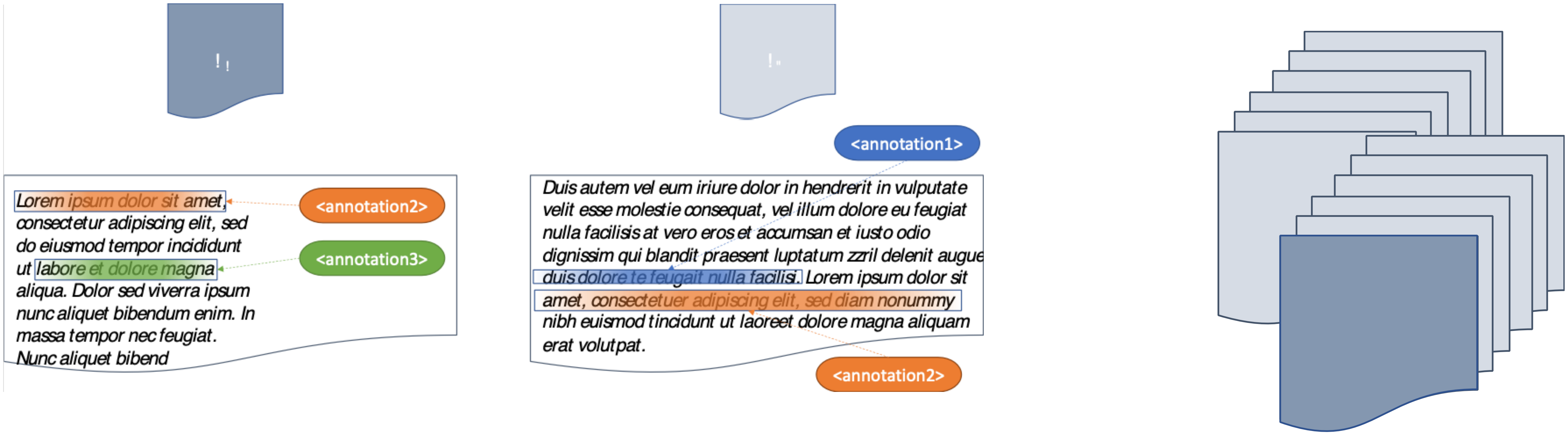
[4] Felix Kuhr, Bjarne Witten, Ralf Möller: Corpus-Driven Annotation Enrichment. Proceedings of the 13th IEEE International Conference on Semantic Computing (ICSC-19), 2019

Corpus-Driven Document Enrichment using SCDs



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

Corpus-Driven Document Enrichment using SCDs [4]



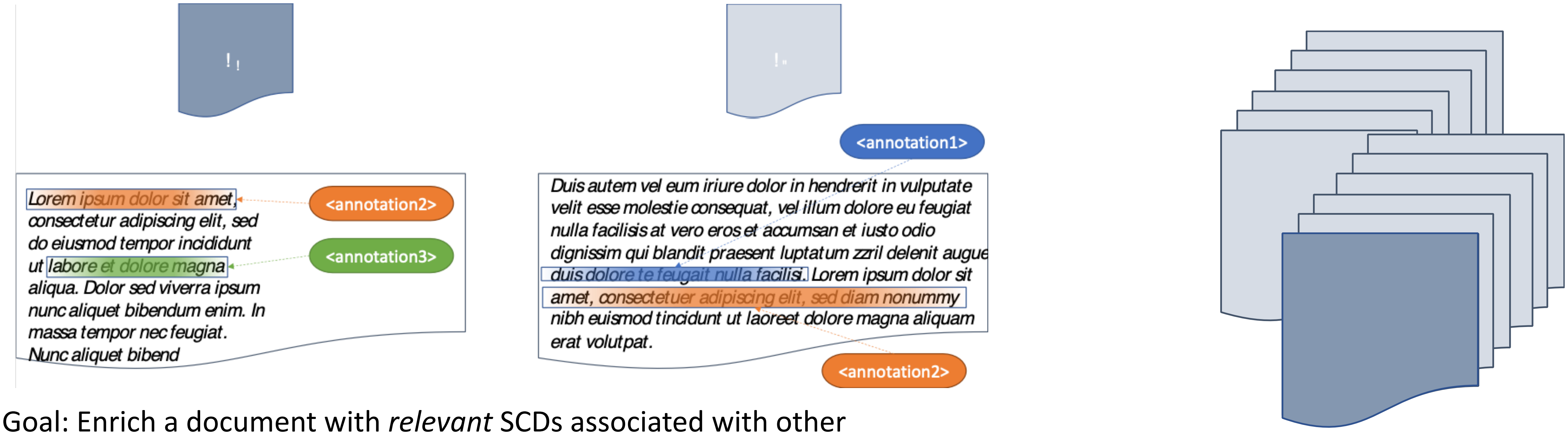
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Dolor sed viverra ipsum nunc aliquet bibendum enim. In massa tempor nec feugiat. Nunc aliquet bibend

Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.

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

[4]

Corpus-Driven Document Enrichment using SCDs



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

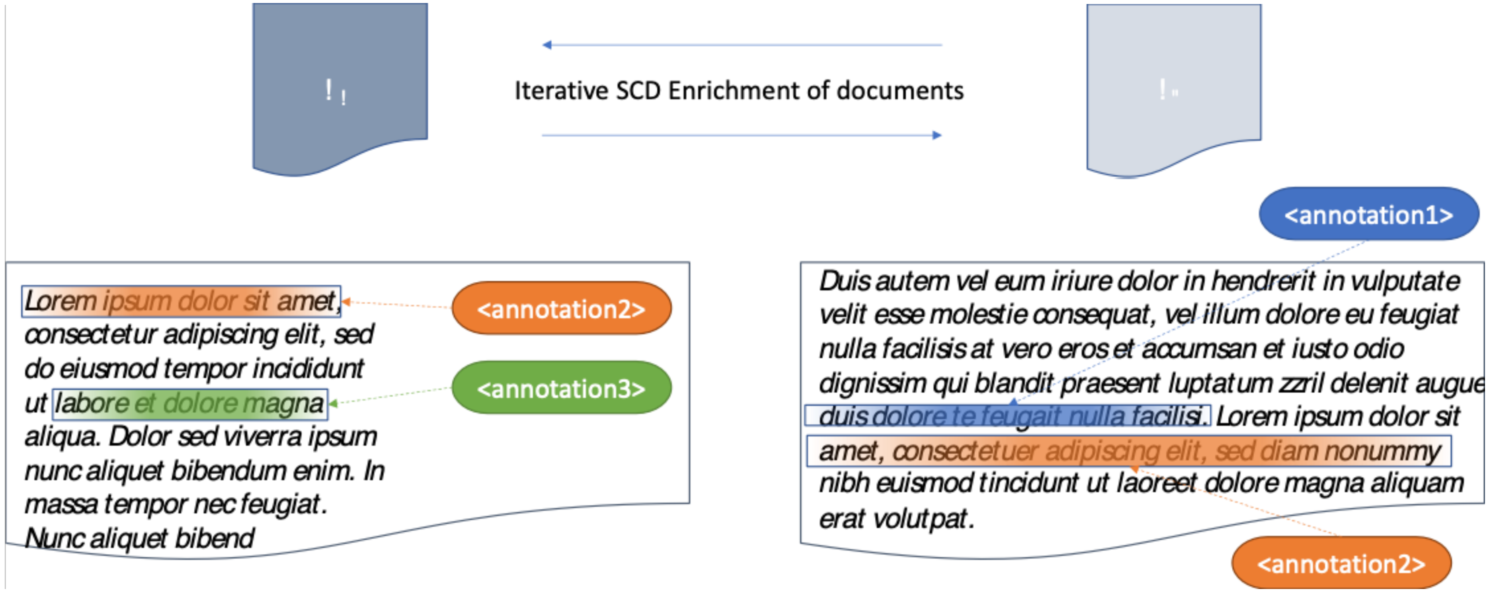
Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.

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

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

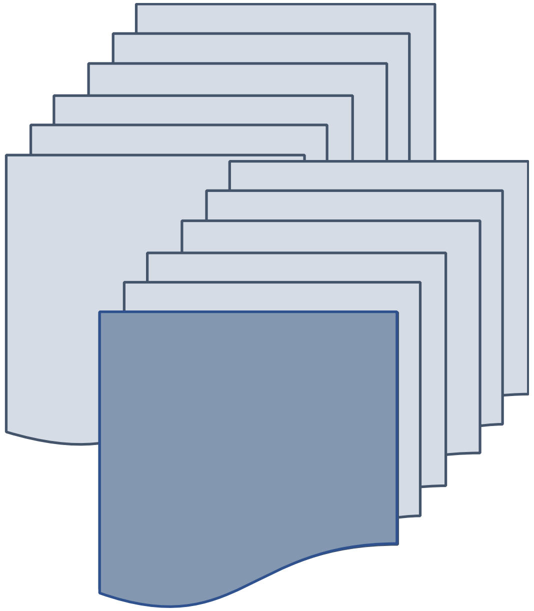
[4]

Corpus-Driven Document Enrichment using SCDs



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

Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.



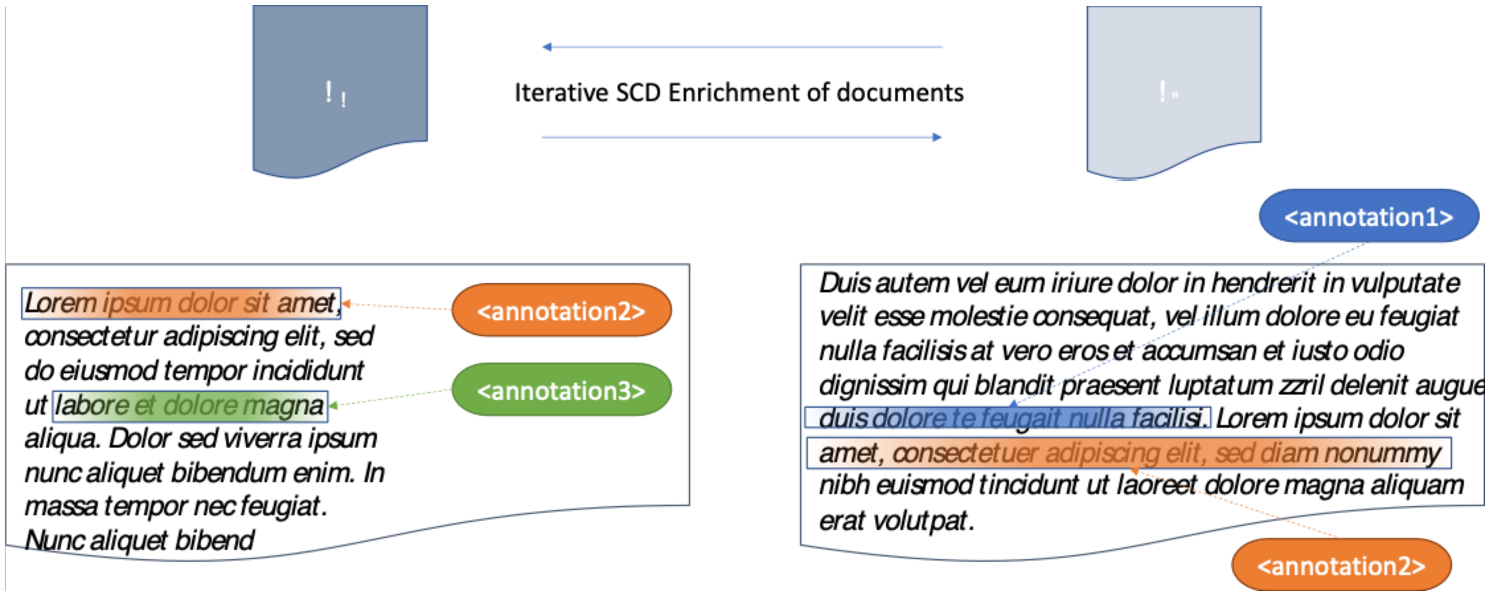
<u>Goal:</u> Enrich a document with *<u>relevant</u>* SCDs associated with other documents in an IR-corpus.

Fixed-point iteration procedure:

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

[4]

Corpus-Driven Document Enrichment using SCDs

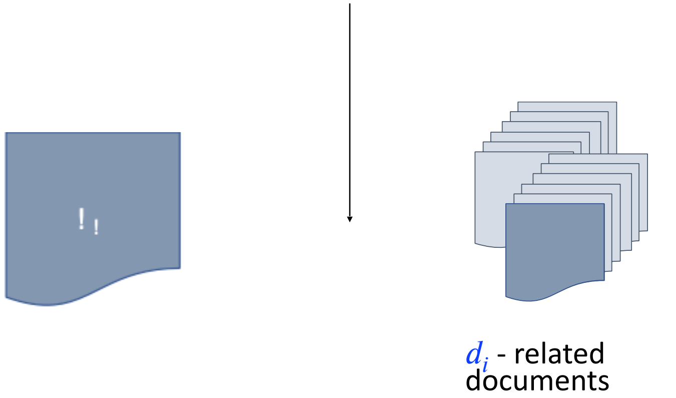


Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Dolor sed viverra ipsum nunc aliquet bibendum enim. In massa tempor nec feugiat. Nunc aliquet bibend

Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.

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

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



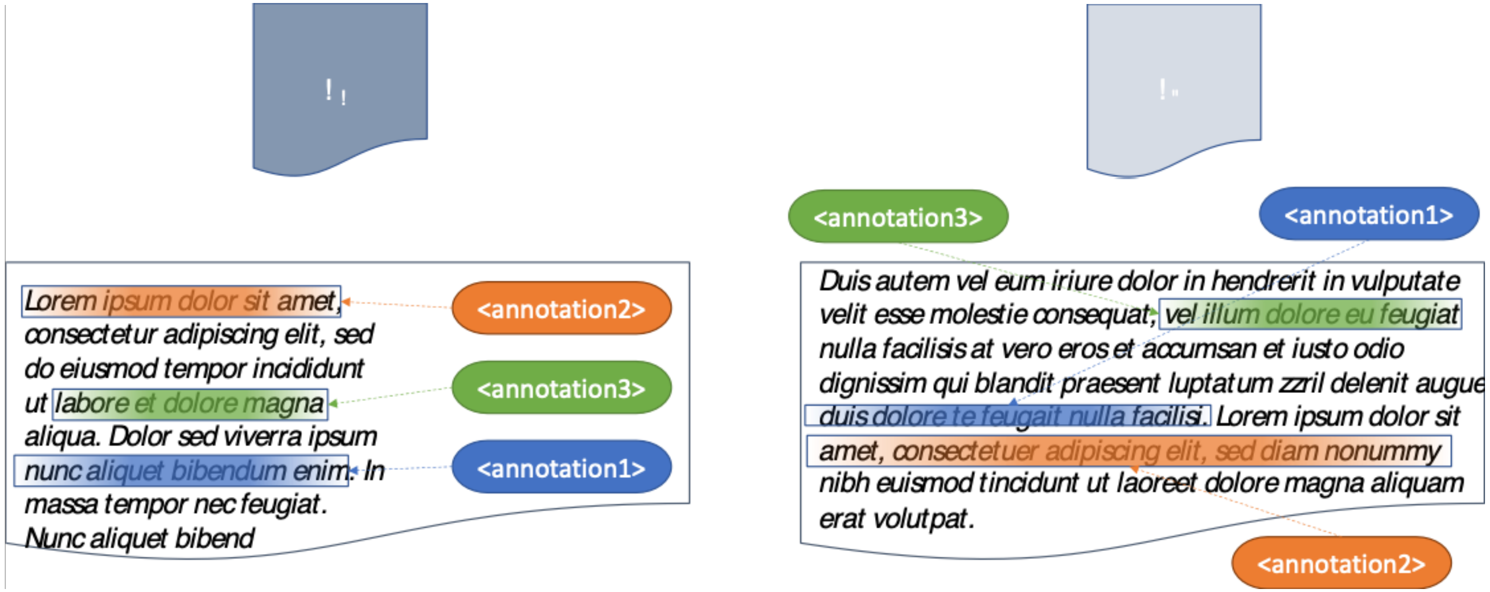
Goal: Enrich a document with *relevant* SCDs associated with other documents in an IR-corpus.

Fixed-point iteration procedure:

- determine the expected related documents in IR-corpus D,

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

Corpus-Driven Document Enrichment using SCDs [4]



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

Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.

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

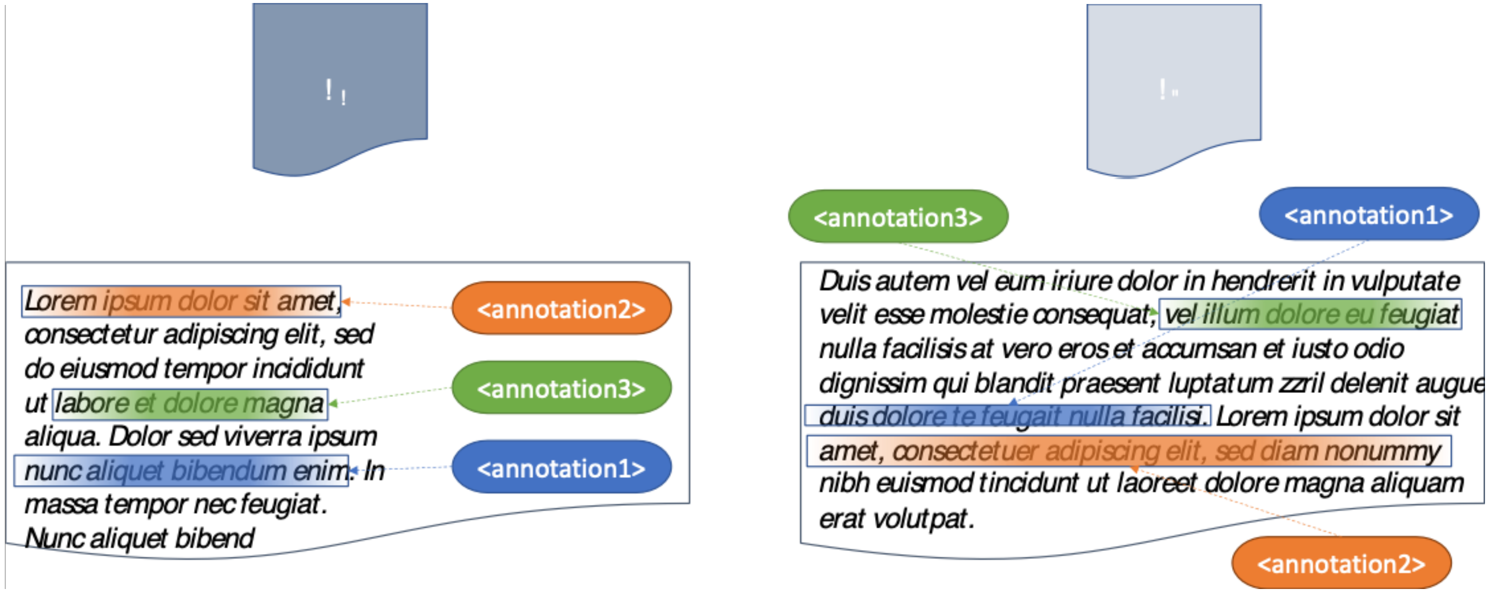
Fixed-point iteration procedure:

- determine the expected related documents in IR-corpus D,
- determine the set of SCDs T from D that are newly added to d, then

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

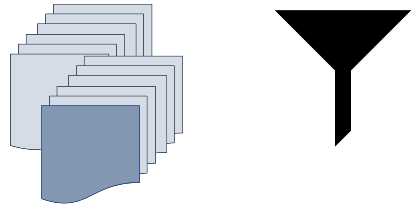
[4]

Corpus-Driven Document Enrichment using SCDs



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

Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.



Topic similarity
SCD similarity
Frequency

d_i - related documents

$expected\text{–}relevance(t, d_i)$

- 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

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

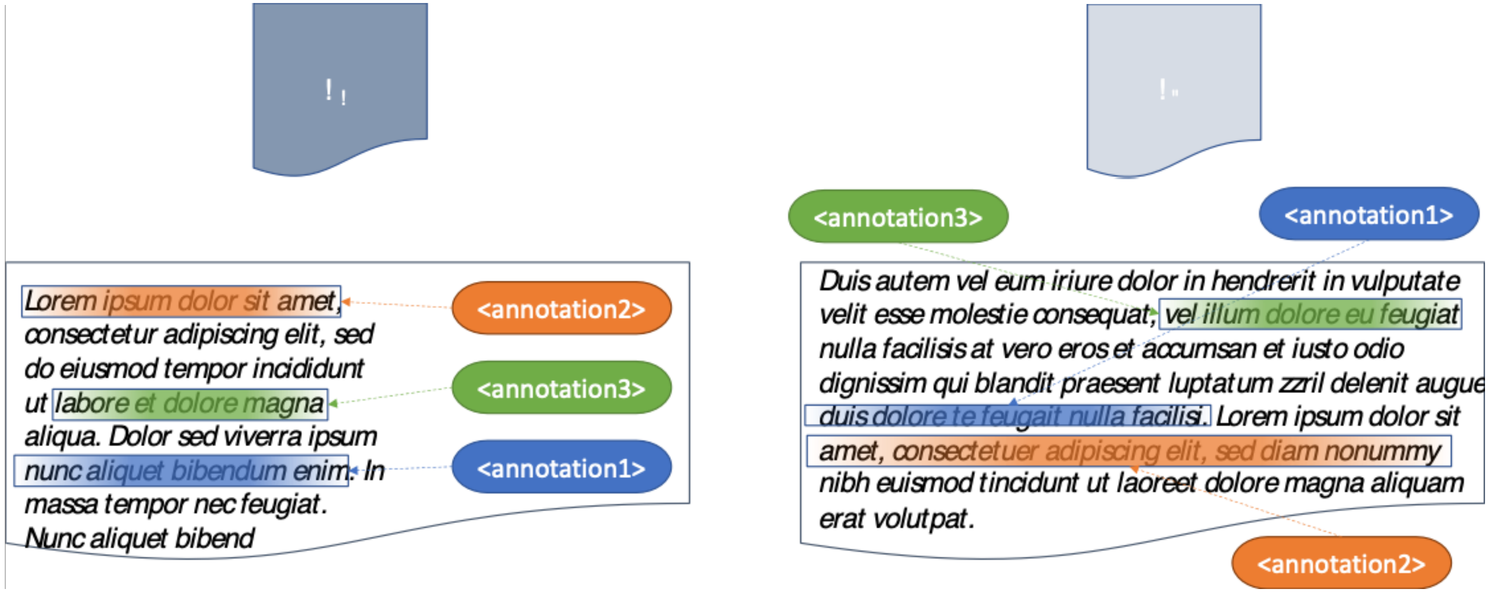
Fixed-point iteration procedure:

- determine the expected related documents in IR-corpus D,
- determine the set of SCDs T from D that are newly added to d, then

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

[4]

Corpus-Driven Document Enrichment using SCDs



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>
<annotation3>
<annotation1>
Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.
<annotation2>

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

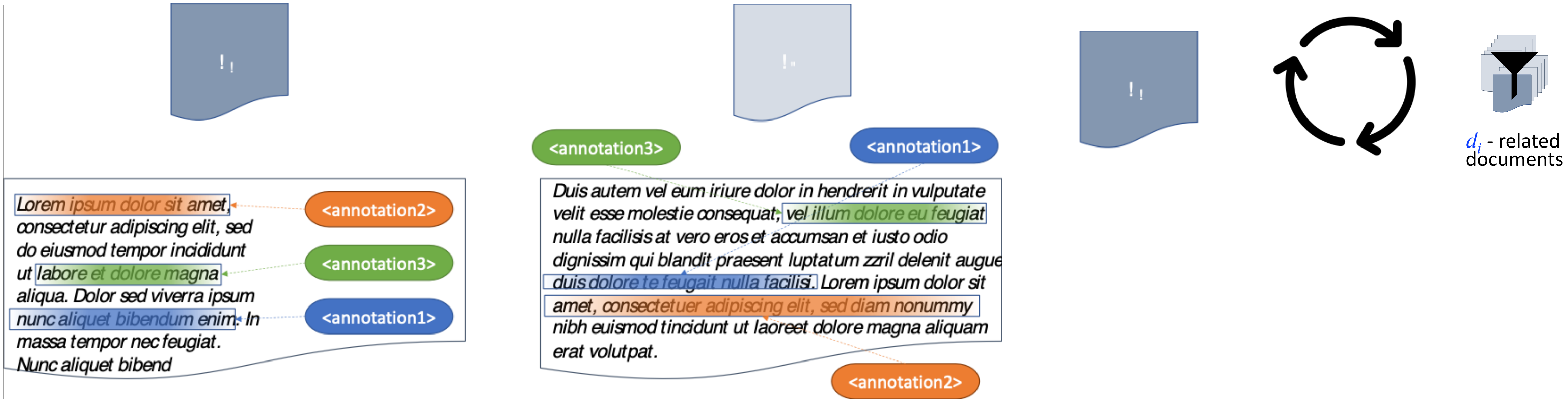
Fixed-point iteration procedure:

- determine the expected related documents in IR-corpus D,
- determine the set of SCDs T from D that are newly added to d, then

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

[4]

Corpus-Driven Document Enrichment using SCDs



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

Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.

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

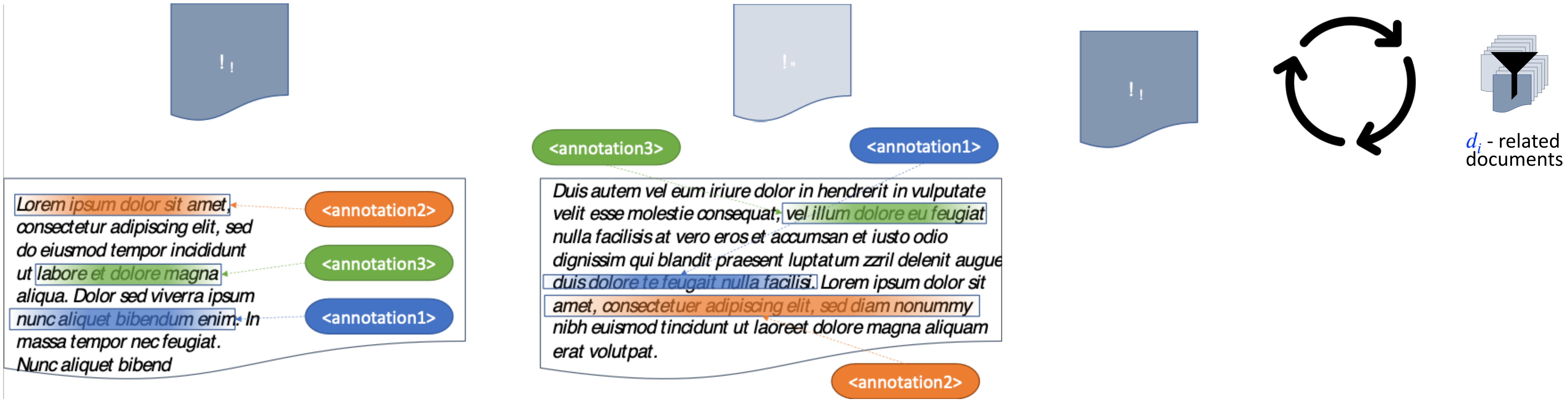
Fixed-point iteration procedure:

- determine the expected related documents in IR-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

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

[4]

Corpus-Driven Document Enrichment using SCDs



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

Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.

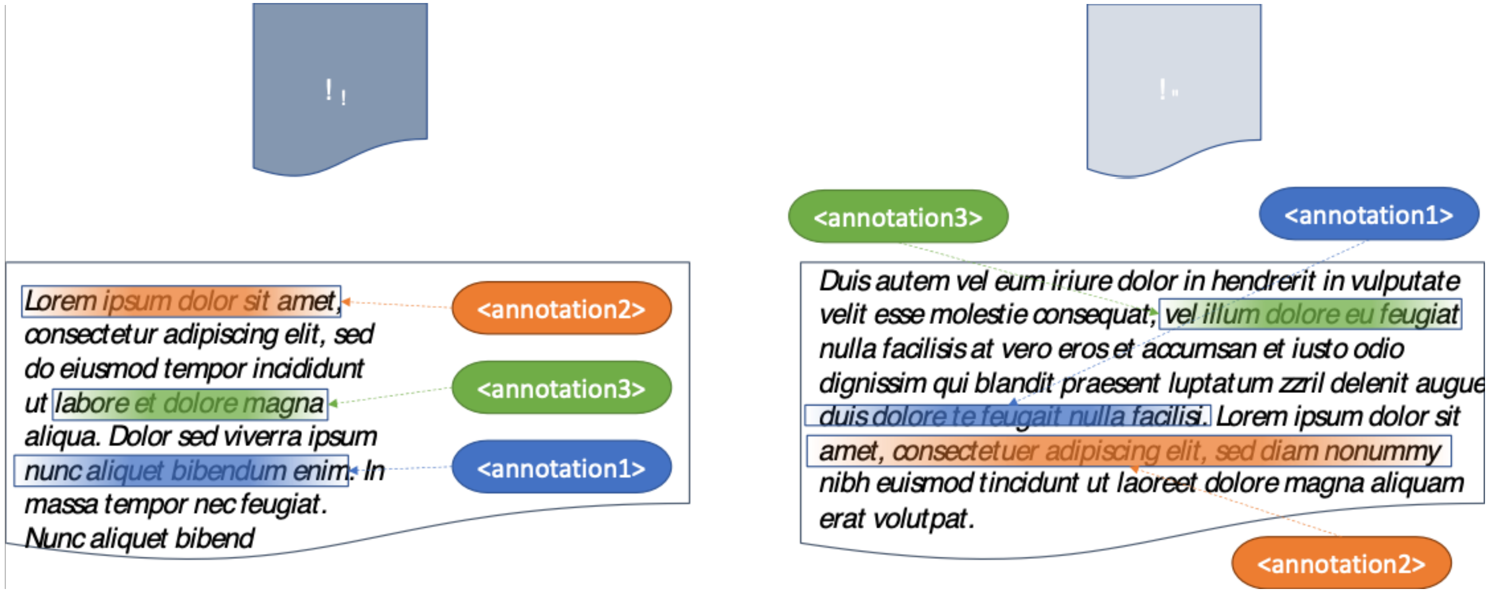
Goal: Enrich a document with ***relevant*** SCDs associated with other documents in an IR-corpus.

Fixed-point iteration procedure:

- determine the expected related documents in IR-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.

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

Corpus-Driven Document Enrichment using SCDs [4]



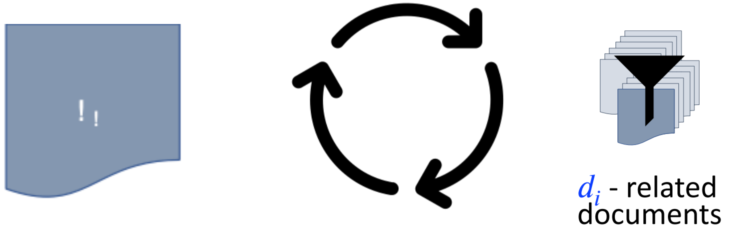
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Dolor sed viverra ipsum nunc aliquet bibendum enim. In massa tempor nec feugiat. Nunc aliquet bibend

Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.

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 IR-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.



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

- 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

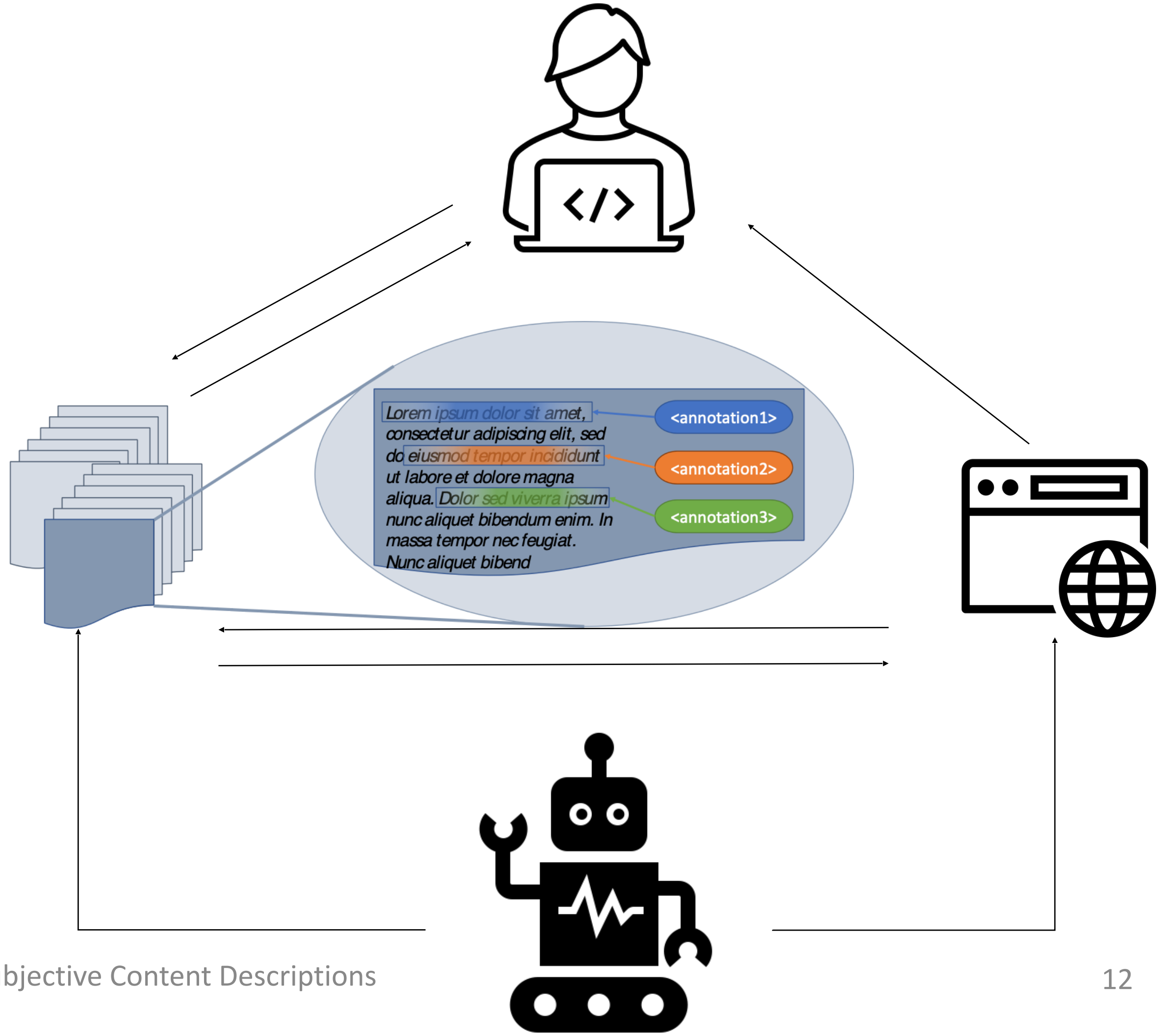
[4] Felix Kuhr, Bjarne Witten, Ralf Möller: Corpus-Driven Annotation Enrichment. Proceedings of the 13th IEEE International Conference on Semantic Computing (ICSC-19), 2019

Expected Relevance Value

Expected Relevance Value

- Given: document d_j from IR-corpus D

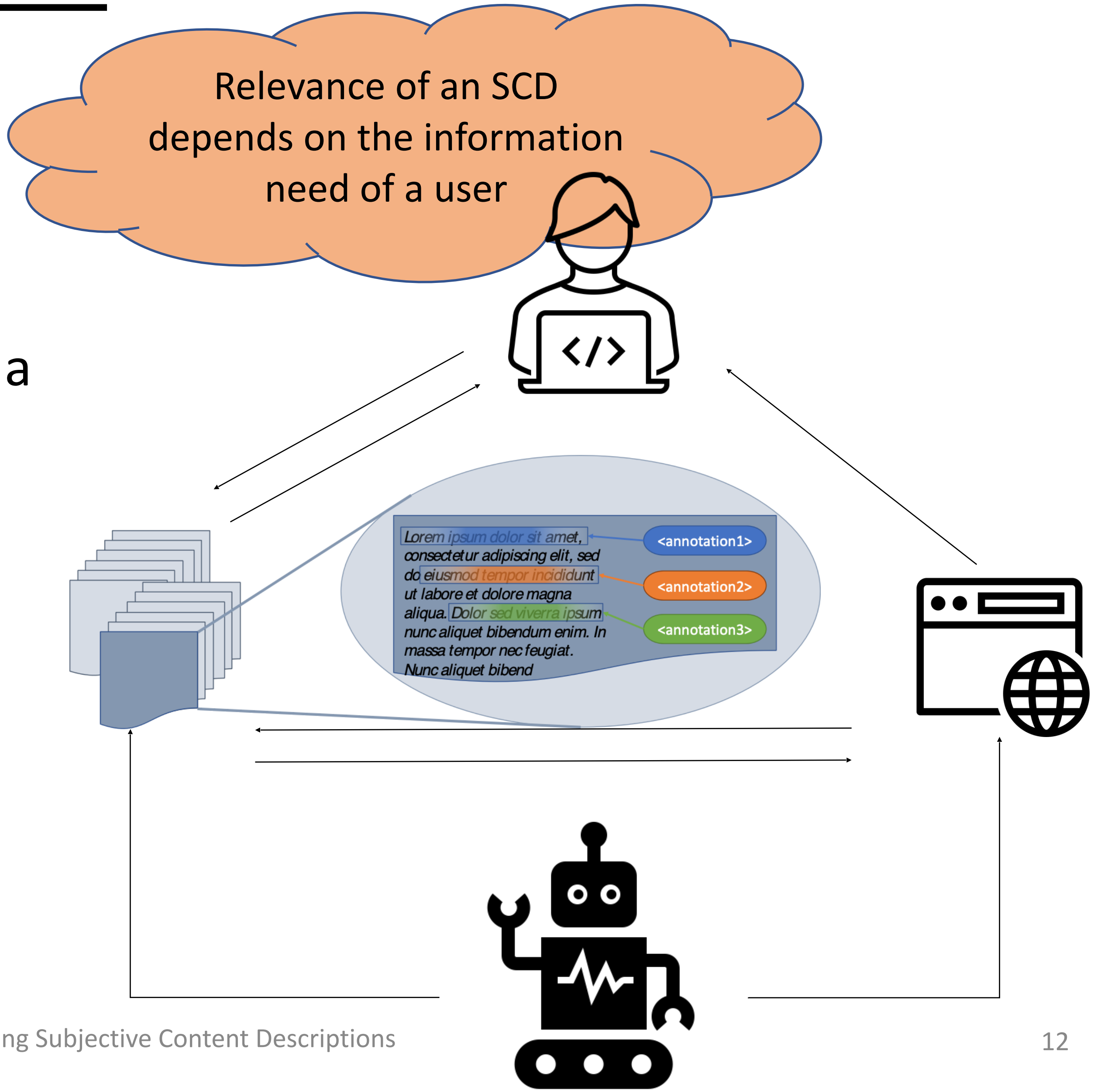
- Question: What is the expected relevance value of an SCD associated to a related document?



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
<annotation1>
<annotation2>
<annotation3>

Expected Relevance Value

- Given: document d_j from IR-corpus D
- Question: What is the expected relevance value of an SCD associated to a related document?



Relevance of an SCD depends on the information need of a user

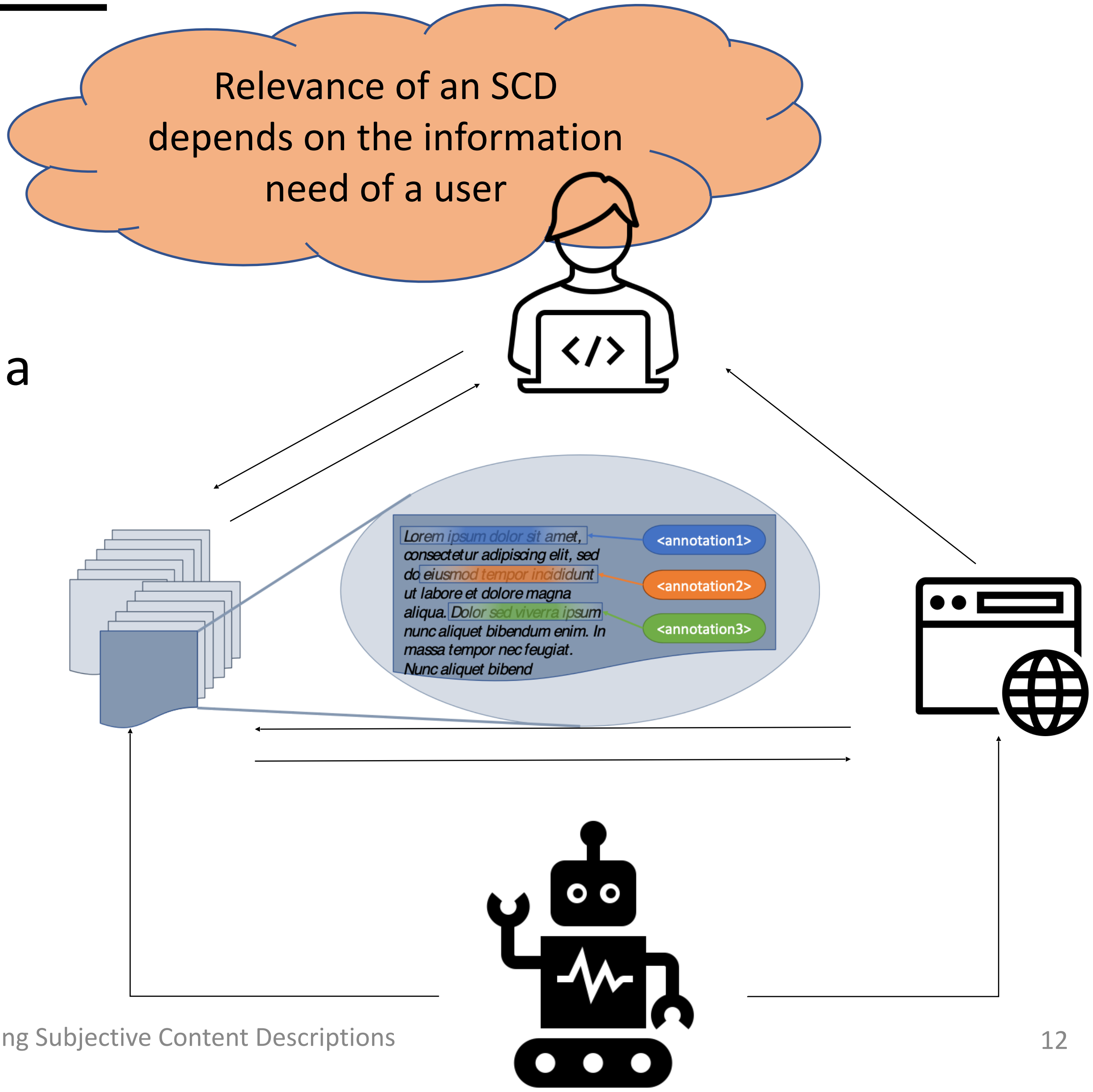
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Dolor sed viverra ipsum nunc aliquet bibendum enim. In massa tempor nec feugiat. Nunc aliquet bibend

Expected Relevance Value

- Given: document d_j from IR-corpus D

- Question: What is the expected relevance value of an SCD associated to a related document?

- Some ways to adjust performance:
  - Similarity between documents
  - Similarity between SCDs
  - Frequency of SCDs



Relevance of an SCD depends on the information need of a user

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

Bootstrap Approaches for SCDs

Inline SCDs[5]

[5] Felix Kuhr, Magnus Bender, Tanya Braun, Ralf Möller: Context-specific Adaptation of Subjective Content Descriptions. 15th IEEE International Conference on Semantic Computing, (ICSC 2021)

Bootstrap Approaches for SCDs

Inline SCDs[5]

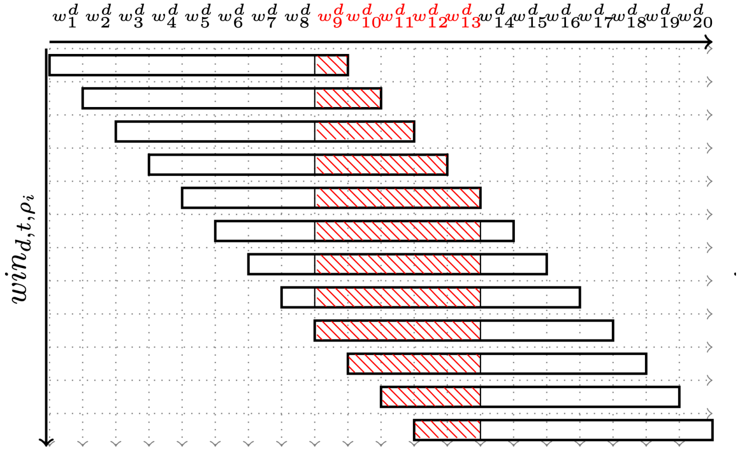
- Given: SCD word distribution, trained HMM to detect *inline* SCDs in text

[5] *Felix Kuhr, Magnus Bender, Tanya Braun, Ralf Möller: Context-specific Adaptation of Subjective Content Descriptions. 15th IEEE International Conference on Semantic Computing, (ICSC 2021)*

Bootstrap Approaches for SCDs

Inline SCDs [5]

- Given: SCD word distribution, trained HMM to detect *inline* SCDs in text
- Estimate MPSCDs and use trained HMM to analyse sequence of corresponding SCD similarity values
  - Small similarity values → different content → new inline-SCDs
  - Inline-SCD = Content of window
  - Inline-SCD represent new row in SCD word matrix
  - HMMs given as user input

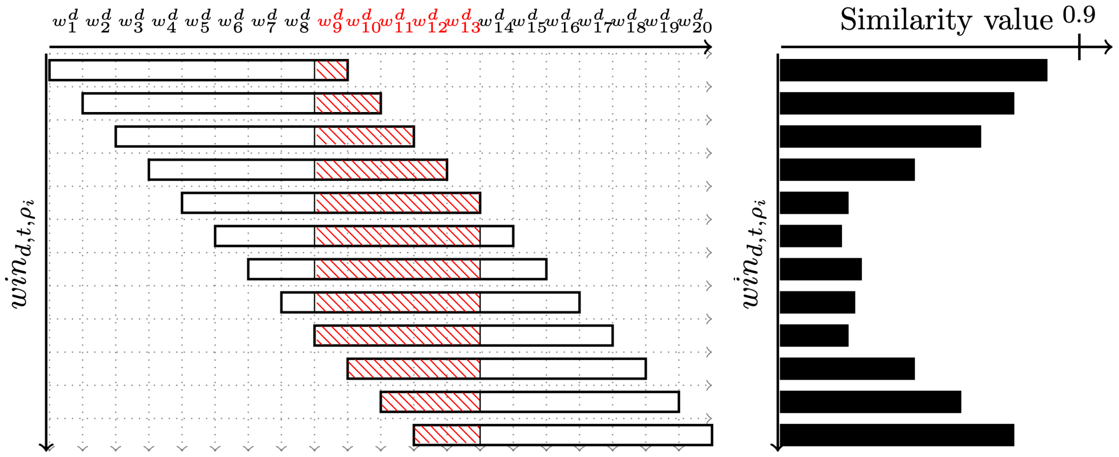

w_1^d w_2^d w_3^d w_4^d w_5^d w_6^d w_7^d w_8^d w_9^d w_{10}^d w_{11}^d w_{12}^d w_{13}^d w_{14}^d w_{15}^d w_{16}^d w_{17}^d w_{18}^d w_{19}^d w_{20}^d
win_{d,t,ρ_i}

[5] *Felix Kuhr, Magnus Bender, Tanya Braun, Ralf Möller: Context-specific Adaptation of Subjective Content Descriptions. 15th IEEE International Conference on Semantic Computing, (ICSC 2021)*

Bootstrap Approaches for SCDs

Inline SCDs[5]

- Given: SCD word distribution, trained HMM to detect *inline* SCDs in text
- Estimate MPSCDs and use trained HMM to analyse sequence of corresponding SCD similarity values
  - Small similarity values → different content → new inline-SCDs
  - Inline-SCD = Content of window
  - Inline-SCD represent new row in SCD word matrix
  - HMMs given as user input

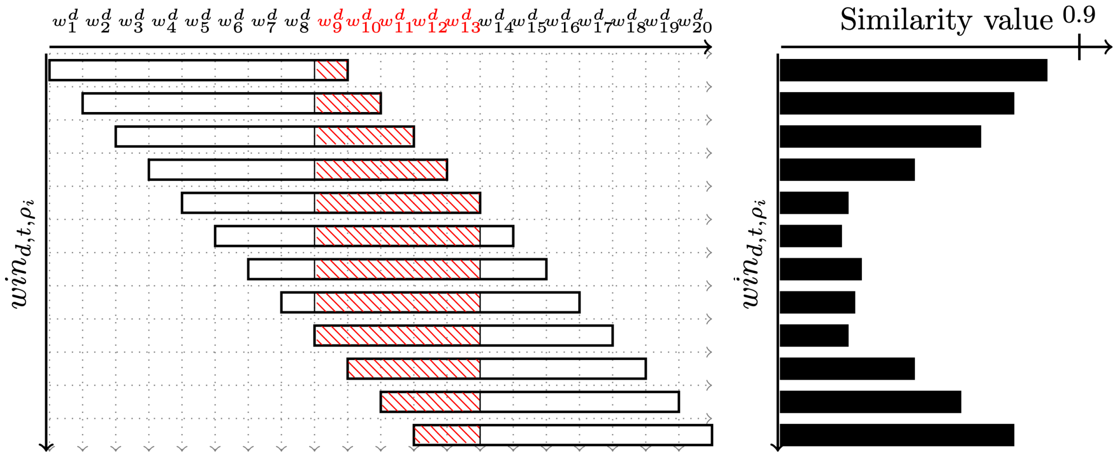

w_1^d w_2^d w_3^d w_4^d w_5^d w_6^d w_7^d w_8^d w_9^d w_{10}^d w_{11}^d w_{12}^d w_{13}^d w_{14}^d w_{15}^d w_{16}^d w_{17}^d w_{18}^d w_{19}^d w_{20}^d
win_{d,t,ρ_i}
Similarity value
0.9
win_{d,t,ρ_i}

[5] *Felix Kuhr, Magnus Bender, Tanya Braun, Ralf Möller: Context-specific Adaptation of Subjective Content Descriptions. 15th IEEE International Conference on Semantic Computing, (ICSC 2021)*

Bootstrap Approaches for SCDs

Inline SCDs[5]

- Given: SCD word distribution, trained HMM to detect *inline* SCDs in text
- Estimate MPSCDs and use trained HMM to analyse sequence of corresponding SCD similarity values
  - Small similarity values → different content → new inline-SCDs
  - Inline-SCD = Content of window
  - Inline-SCD represent new row in SCD word matrix
  - HMMs given as user input


w_1^d w_2^d w_3^d w_4^d w_5^d w_6^d w_7^d w_8^d w_9^d w_{10}^d w_{11}^d w_{12}^d w_{13}^d w_{14}^d w_{15}^d w_{16}^d w_{17}^d w_{18}^d w_{19}^d w_{20}^d
win_{d,t,ρ_i}
Similarity value
0.9
win_{d,t,ρ_i}

Adapting SCD word distribution from another IR-corpus[6]

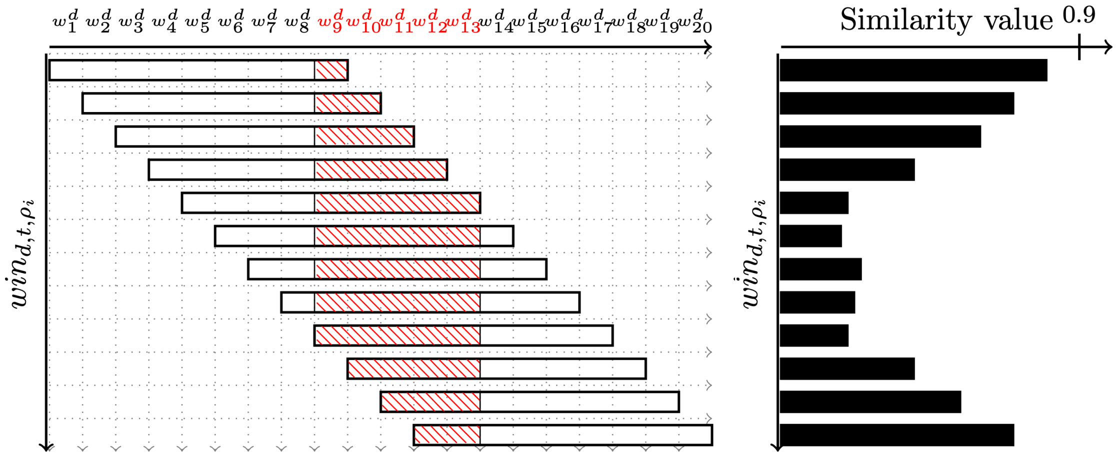
[5] *Felix Kuhr, Magnus Bender, Tanya Braun, Ralf Möller: Context-specific Adaptation of Subjective Content Descriptions. 15th IEEE International Conference on Semantic Computing, (ICSC 2021)*

[6] *Felix Kuhr, Magnus Bender, Tanya Braun, Ralf Möller: Context-specific Adaptation of Subjective Content Descriptions. 15th IEEE International Conference on Semantic Computing, (ICSC 2021)*

Bootstrap Approaches for SCDs

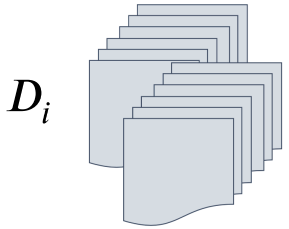
Inline SCDs [5]

- Given: SCD word distribution, trained HMM to detect *inline* SCDs in text
- Estimate MPSCDs and use trained HMM to analyse sequence of corresponding SCD similarity values
    - Small similarity values → different content → new inline-SCDs
    - Inline-SCD = Content of window
    - Inline-SCD represent new row in SCD word matrix
    - HMMs given as user input



w_1^d w_2^d w_3^d w_4^d w_5^d w_6^d w_7^d w_8^d w_9^d w_{10}^d w_{11}^d w_{12}^d w_{13}^d w_{14}^d w_{15}^d w_{16}^d w_{17}^d w_{18}^d w_{19}^d w_{20}^d
win_{d,t,ρ_i}
Similarity value
0.9
win_{d,t,ρ_i}

Adapting SCD word distribution from another IR-corpus [6]



$$\delta(D_i) = \begin{array}{c|ccccc} & w_1 & w_2 & w_3 & \cdots & w_n \\ \hline t_1 & v_{1,1} & v_{1,2} & v_{1,3} & \cdots & v_{1,n} \\ t_2 & v_{2,1} & v_{2,2} & v_{2,3} & \cdots & v_{2,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ t_m & v_{m,1} & v_{m,2} & v_{m,3} & \cdots & v_{m,n} \end{array}$$

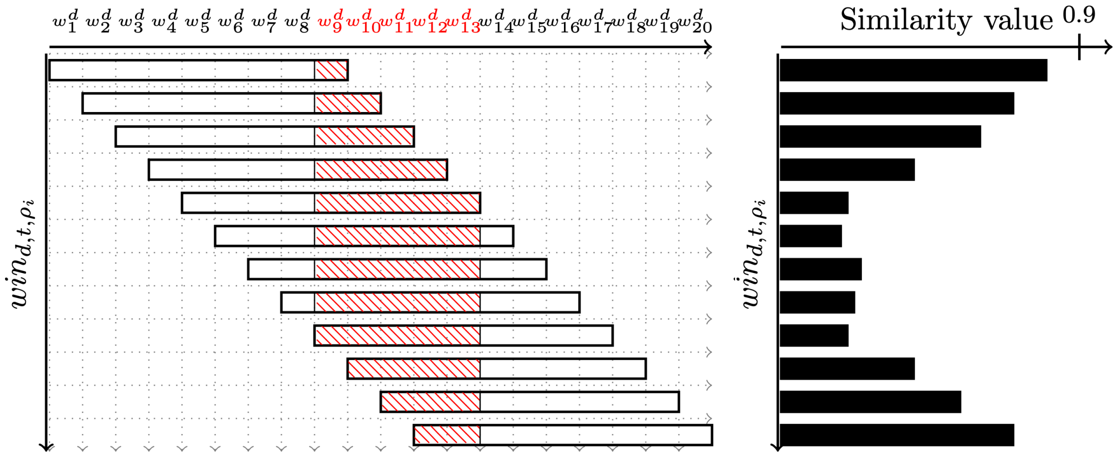
[5] Felix Kuhr, Magnus Bender, Tanya Braun, Ralf Möller: Context-specific Adaptation of Subjective Content Descriptions. 15th IEEE International Conference on Semantic Computing, (ICSC 2021)

[6] Felix Kuhr, Magnus Bender, Tanya Braun, Ralf Möller: Context-specific Adaptation of Subjective Content Descriptions. 15th IEEE International Conference on Semantic Computing, (ICSC 2021)

Bootstrap Approaches for SCDs

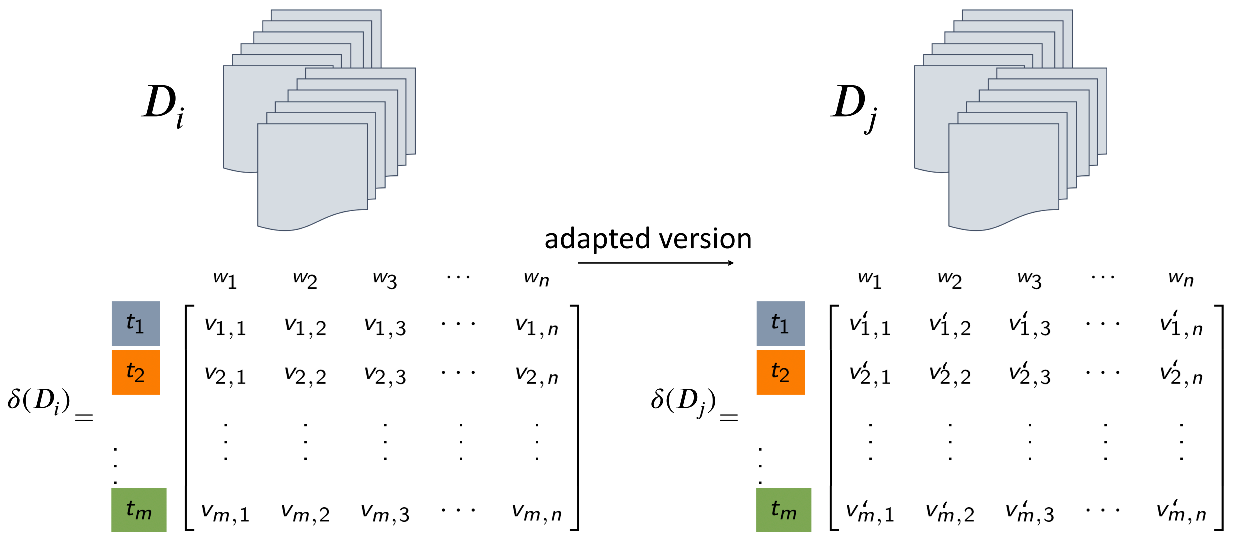
Inline SCDs[5]

- Given: SCD word distribution, trained HMM to detect *inline* SCDs in text
- Estimate MPSCDs and use trained HMM to analyse sequence of corresponding SCD similarity values
  - Small similarity values → different content → new inline-SCDs
  - Inline-SCD = Content of window
  - Inline-SCD represent new row in SCD word matrix
  - HMMs given as user input



[5] Felix Kuhr, Magnus Bender, Tanya Braun, Ralf Möller: Context-specific Adaptation of Subjective Content Descriptions. 15th IEEE International Conference on Semantic Computing, (ICSC 2021)

Adapting SCD word distribution from another IR-corpus[6]



$$\delta(D_i) = \begin{array}{c|ccccc} & w_1 & w_2 & w_3 & \cdots & w_n \\ \hline t_1 & v_{1,1} & v_{1,2} & v_{1,3} & \cdots & v_{1,n} \\ t_2 & v_{2,1} & v_{2,2} & v_{2,3} & \cdots & v_{2,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ t_m & v_{m,1} & v_{m,2} & v_{m,3} & \cdots & v_{m,n} \end{array}$$

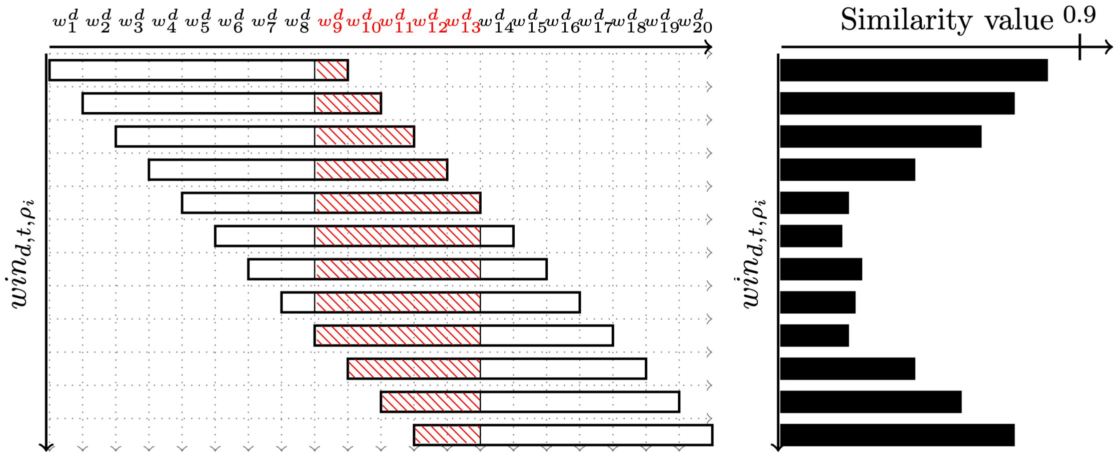
$$\delta(D_j) = \begin{array}{c|ccccc} & w_1 & w_2 & w_3 & \cdots & w_n \\ \hline t_1 & v'_{1,1} & v'_{1,2} & v'_{1,3} & \cdots & v'_{1,n} \\ t_2 & v'_{2,1} & v'_{2,2} & v'_{2,3} & \cdots & v'_{2,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ t_m & v'_{m,1} & v'_{m,2} & v'_{m,3} & \cdots & v'_{m,n} \end{array}$$

[6] Felix Kuhr, Magnus Bender, Tanya Braun, Ralf Möller: Context-specific Adaptation of Subjective Content Descriptions. 15th IEEE International Conference on Semantic Computing, (ICSC 2021)

Bootstrap Approaches for SCDs

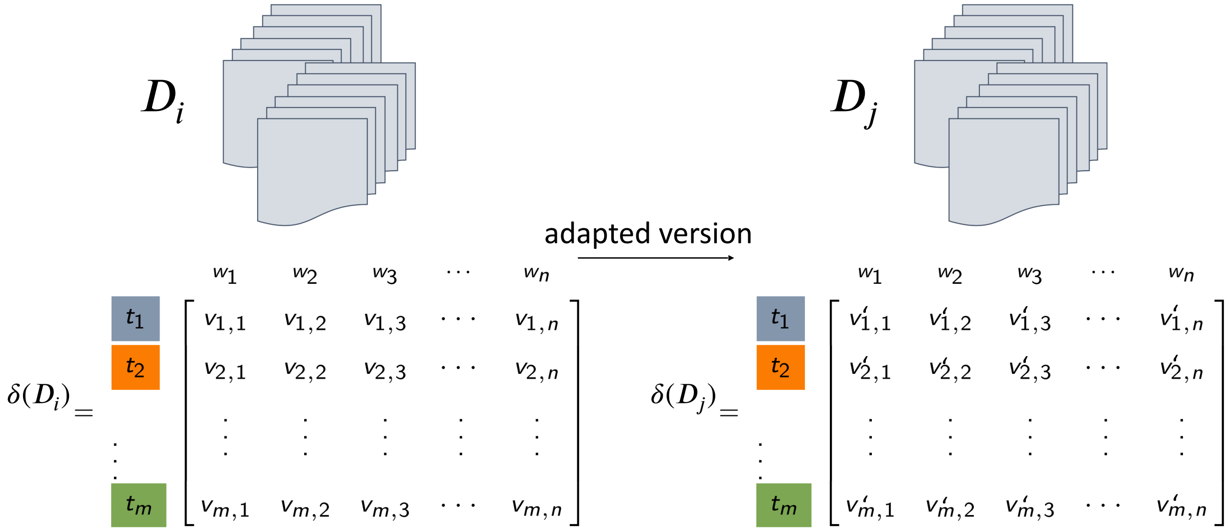
Inline SCDs [5]

- Given: SCD word distribution, trained HMM to detect *inline* SCDs in text
- Estimate MPSCDs and use trained HMM to analyse sequence of corresponding SCD similarity values
  - Small similarity values → different content → new inline-SCDs
  - Inline-SCD = Content of window
  - Inline-SCD represent new row in SCD word matrix
  - HMMs given as user input



w_1^d w_2^d w_3^d w_4^d w_5^d w_6^d w_7^d w_8^d w_9^d w_{10}^d w_{11}^d w_{12}^d w_{13}^d w_{14}^d w_{15}^d w_{16}^d w_{17}^d w_{18}^d w_{19}^d w_{20}^d
win_{d,t,ρ_i}
Similarity value
0.9
win_{d,t,ρ_i}

Adapting SCD word distribution from another IR-corpus [6]



$$\delta(D_i) = \begin{array}{c|ccccc} & w_1 & w_2 & w_3 & \cdots & w_n \\ \hline t_1 & v_{1,1} & v_{1,2} & v_{1,3} & \cdots & v_{1,n} \\ t_2 & v_{2,1} & v_{2,2} & v_{2,3} & \cdots & v_{2,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ t_m & v_{m,1} & v_{m,2} & v_{m,3} & \cdots & v_{m,n} \end{array} \xrightarrow{\text{adapted version}} \delta(D_j) = \begin{array}{c|ccccc} & w_1 & w_2 & w_3 & \cdots & w_n \\ \hline t_1 & v'_{1,1} & v'_{1,2} & v'_{1,3} & \cdots & v'_{1,n} \\ t_2 & v'_{2,1} & v'_{2,2} & v'_{2,3} & \cdots & v'_{2,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ t_m & v'_{m,1} & v'_{m,2} & v'_{m,3} & \cdots & v'_{m,n} \end{array}$$

- Adapt SCD word distribution from IR-corpus D_i to documents in D_j

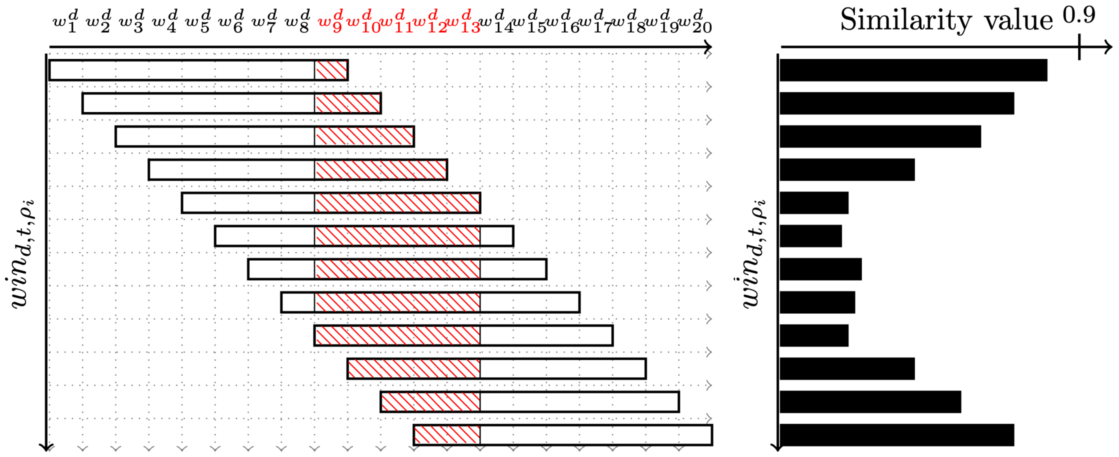
[5] *Felix Kuhr, Magnus Bender, Tanya Braun, Ralf Möller: Context-specific Adaptation of Subjective Content Descriptions. 15th IEEE International Conference on Semantic Computing, (ICSC 2021)*

[6] *Felix Kuhr, Magnus Bender, Tanya Braun, Ralf Möller: Context-specific Adaptation of Subjective Content Descriptions. 15th IEEE International Conference on Semantic Computing, (ICSC 2021)*

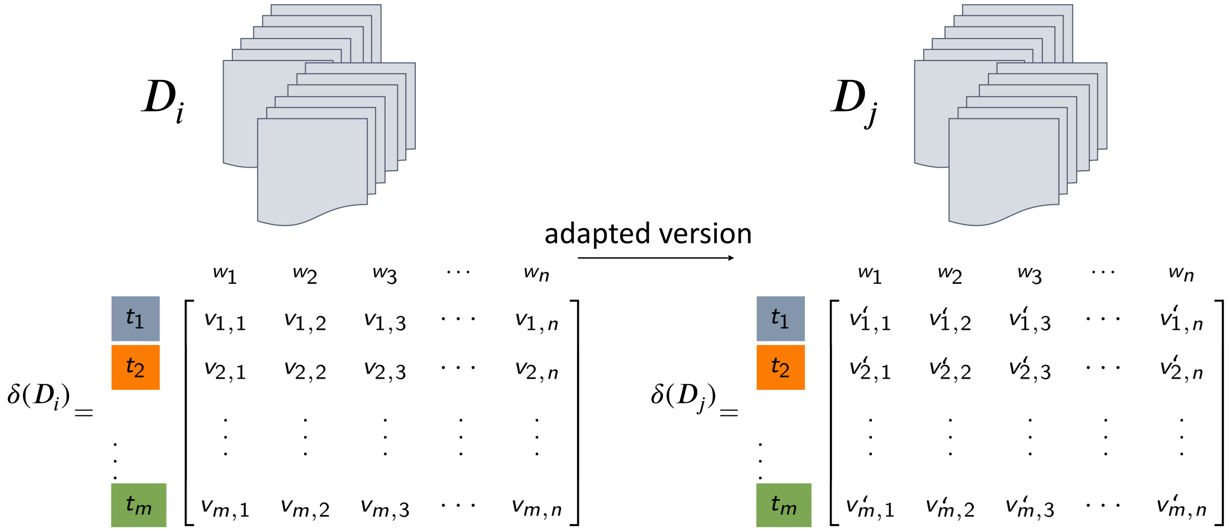
Bootstrap Approaches for SCDs

Inline SCDs[5]

- Given: SCD word distribution, trained HMM to detect *inline* SCDs in text
- Estimate MPSCDs and use trained HMM to analyse sequence of corresponding SCD similarity values
  - Small similarity values → different content → new inline-SCDs
  - Inline-SCD = Content of window
  - Inline-SCD represent new row in SCD word matrix
  - HMMs given as user input



Adapting SCD word distribution from another IR-corpus[6]



$$\delta(D_i) = \begin{array}{c|ccccc} & w_1 & w_2 & w_3 & \cdots & w_n \\ \hline t_1 & v_{1,1} & v_{1,2} & v_{1,3} & \cdots & v_{1,n} \\ t_2 & v_{2,1} & v_{2,2} & v_{2,3} & \cdots & v_{2,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ t_m & v_{m,1} & v_{m,2} & v_{m,3} & \cdots & v_{m,n} \end{array}$$

$$\delta(D_j) = \begin{array}{c|ccccc} & w_1 & w_2 & w_3 & \cdots & w_n \\ \hline t_1 & v'_{1,1} & v'_{1,2} & v'_{1,3} & \cdots & v'_{1,n} \\ t_2 & v'_{2,1} & v'_{2,2} & v'_{2,3} & \cdots & v'_{2,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ t_m & v'_{m,1} & v'_{m,2} & v'_{m,3} & \cdots & v'_{m,n} \end{array}$$

- Adapt SCD word distribution from IR-corpus D_i to documents in D_j
  - Analyze difference in word distributions of documents in corpus D_i and D_j

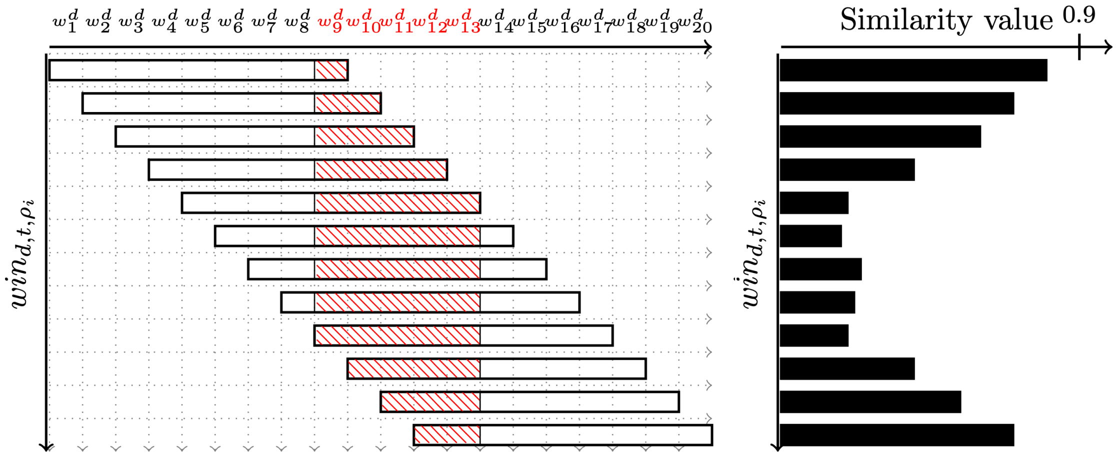
[5] *Felix Kuhr, Magnus Bender, Tanya Braun, Ralf Möller: Context-specific Adaptation of Subjective Content Descriptions. 15th IEEE International Conference on Semantic Computing, (ICSC 2021)*

[6] *Felix Kuhr, Magnus Bender, Tanya Braun, Ralf Möller: Context-specific Adaptation of Subjective Content Descriptions. 15th IEEE International Conference on Semantic Computing, (ICSC 2021)*

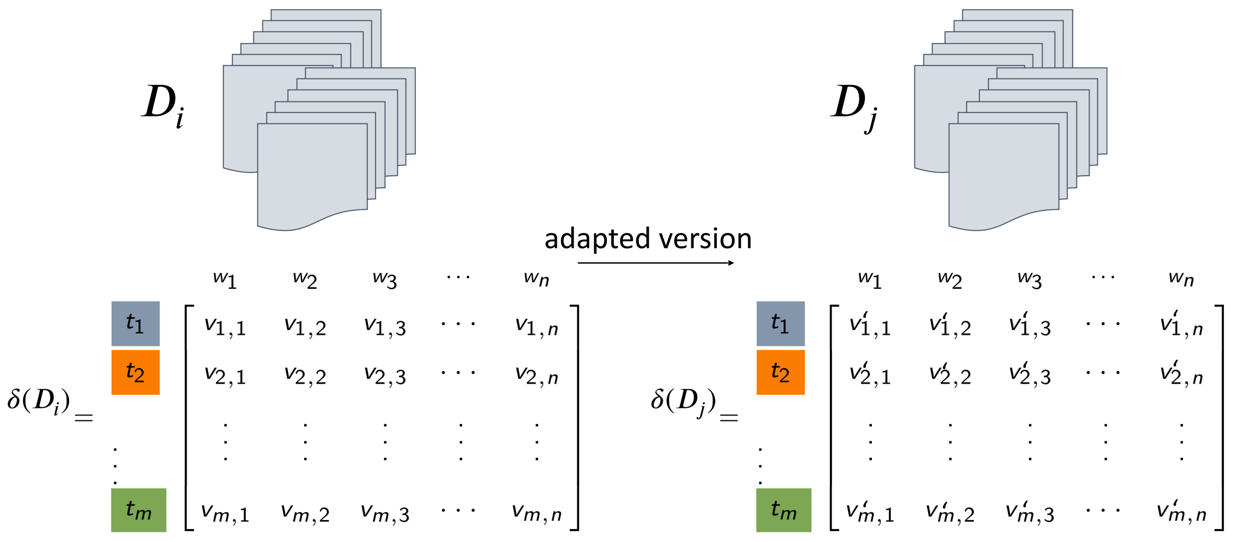
Bootstrap Approaches for SCDs

Inline SCDs [5]

- Given: SCD word distribution, trained HMM to detect *inline* SCDs in text
- Estimate MPSCDs and use trained HMM to analyse sequence of corresponding SCD similarity values
    - Small similarity values → different content → new inline-SCDs
    - Inline-SCD = Content of window
    - Inline-SCD represent new row in SCD word matrix
    - HMMs given as user input



Adapting SCD word distribution from another IR-corpus [6]



$$\delta(D_i) = \begin{array}{c} \\ t_1 \\ t_2 \\ \vdots \\ t_m \end{array} \begin{array}{c} \begin{array}{ccccc} w_1 & w_2 & w_3 & \cdots & w_n \end{array} \\ \begin{bmatrix} v_{1,1} & v_{1,2} & v_{1,3} & \cdots & v_{1,n} \\ v_{2,1} & v_{2,2} & v_{2,3} & \cdots & v_{2,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ v_{m,1} & v_{m,2} & v_{m,3} & \cdots & v_{m,n} \end{bmatrix} \end{array}$$

$$\delta(D_j) = \begin{array}{c} \\ t_1 \\ t_2 \\ \vdots \\ t_m \end{array} \begin{array}{c} \begin{array}{ccccc} w_1 & w_2 & w_3 & \cdots & w_n \end{array} \\ \begin{bmatrix} v'_{1,1} & v'_{1,2} & v'_{1,3} & \cdots & v'_{1,n} \\ v'_{2,1} & v'_{2,2} & v'_{2,3} & \cdots & v'_{2,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ v'_{m,1} & v'_{m,2} & v'_{m,3} & \cdots & v'_{m,n} \end{bmatrix} \end{array}$$

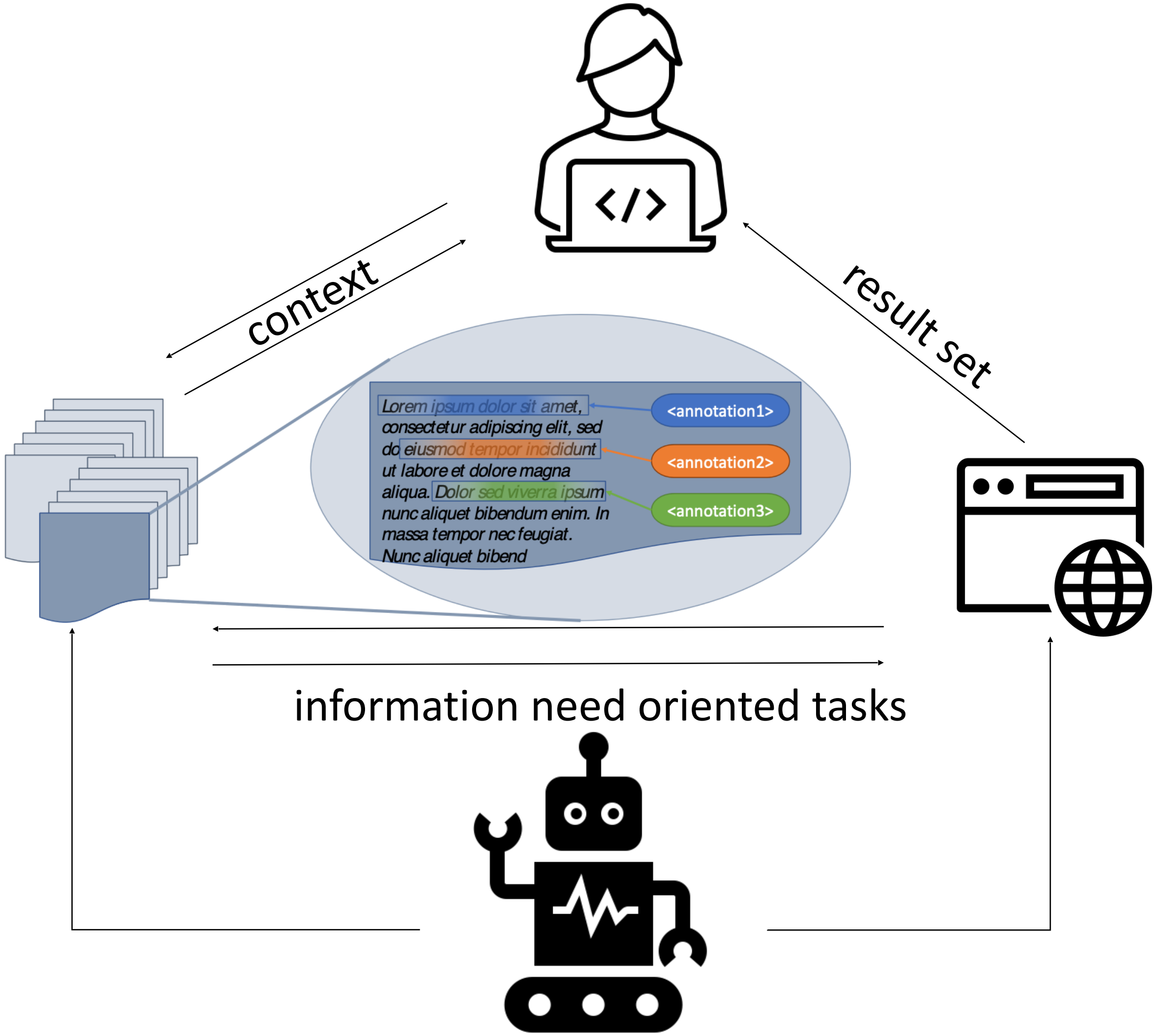
- Adapt SCD word distribution from IR-corpus D_i to documents in D_j
    - Analyze difference in word distributions of documents in corpus D_i and D_j
    - Reweight word distribution for each SCD in $\delta(D_i)$ s.t. distribution fits for D_j

[5] *Felix Kuhr, Magnus Bender, Tanya Braun, Ralf Möller: Context-specific Adaptation of Subjective Content Descriptions. 15th IEEE International Conference on Semantic Computing, (ICSC 2021)*

[6] *Felix Kuhr, Magnus Bender, Tanya Braun, Ralf Möller: Context-specific Adaptation of Subjective Content Descriptions. 15th IEEE International Conference on Semantic Computing, (ICSC 2021)*

Conclusion

- Human-aware information retrieval considering not only content of documents and queries
- Fully automated annotation approach considering the human information need represented by a corpus and SCDs
- Approach for the bootstrap problem considering inline-SCDs


context
result set
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
<annotation1>
<annotation2>
<annotation3>

information need oriented tasks

Conclusion

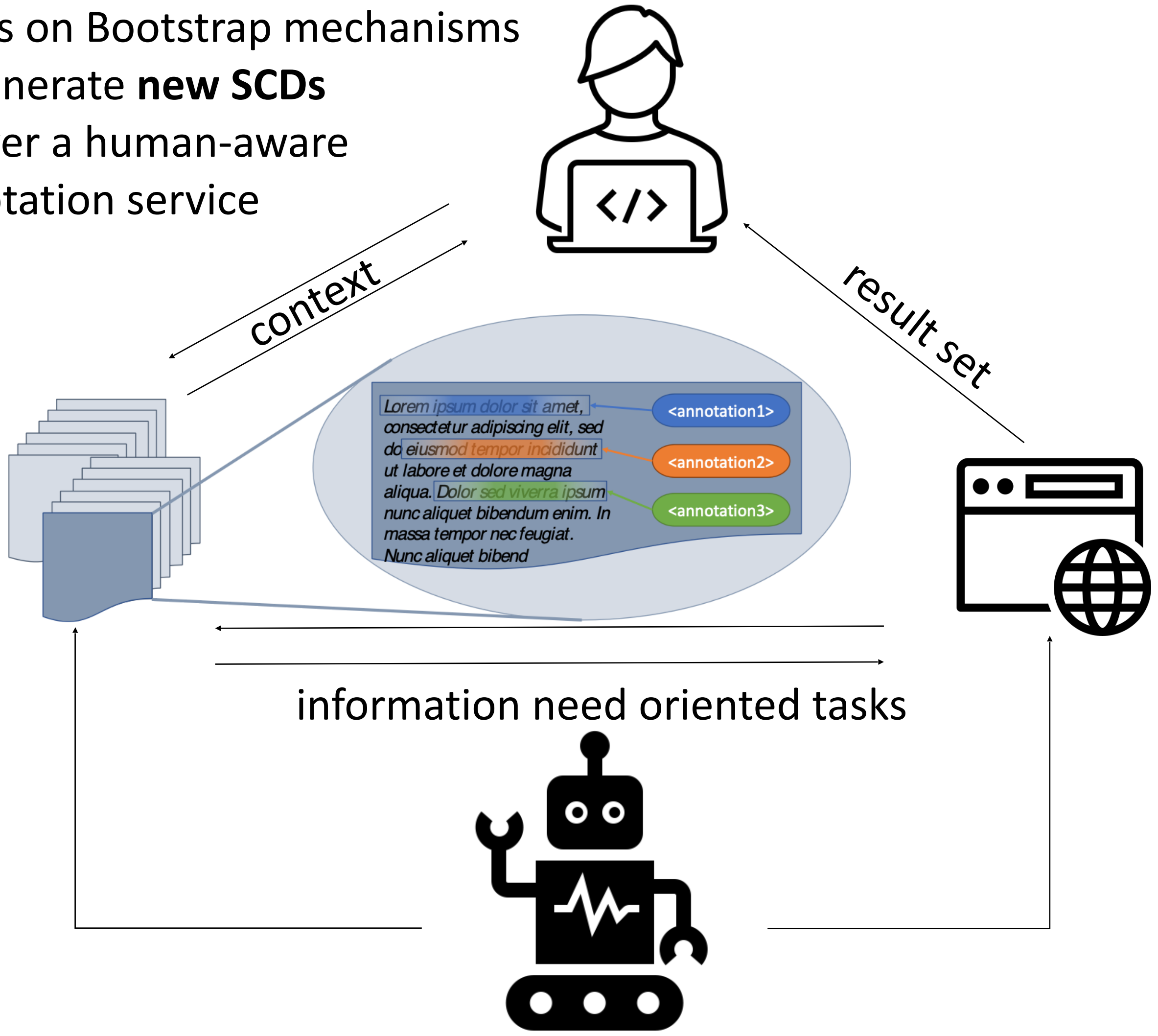
- Human-aware information retrieval considering not only content of documents and queries

- Fully automated annotation approach considering the human information need represented by a corpus and SCDs

- Approach for the bootstrap problem considering inline-SCDs

Future Work:

- Focus on Bootstrap mechanisms to generate **new SCDs**
- Deliver a human-aware annotation service


context
result set
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
<annotation1>
<annotation2>
<annotation3>
</>
information need oriented tasks

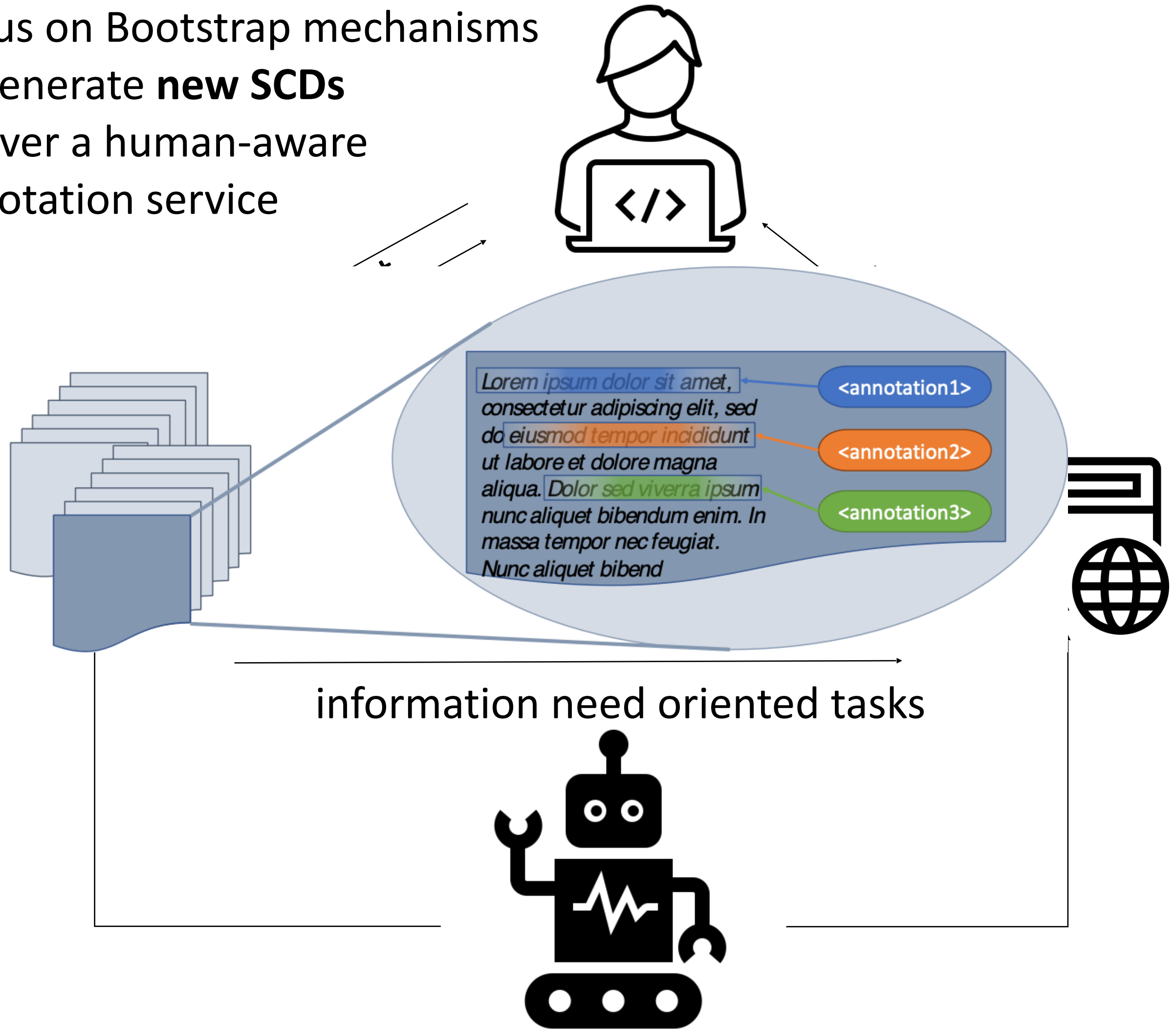
Conclusion

- Human-aware information retrieval considering not only content of documents and queries
- Fully automated annotation approach considering the human information need represented by a corpus and SCDs
- Approach for the bootstrap problem considering inline-SCDs

Focus on human-aware AI approaches:

Future Work:

- Focus on Bootstrap mechanisms to generate **new SCDs**
- Deliver a human-aware annotation service



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
<annotation1>
<annotation2>
<annotation3>
information need oriented tasks

Conclusion

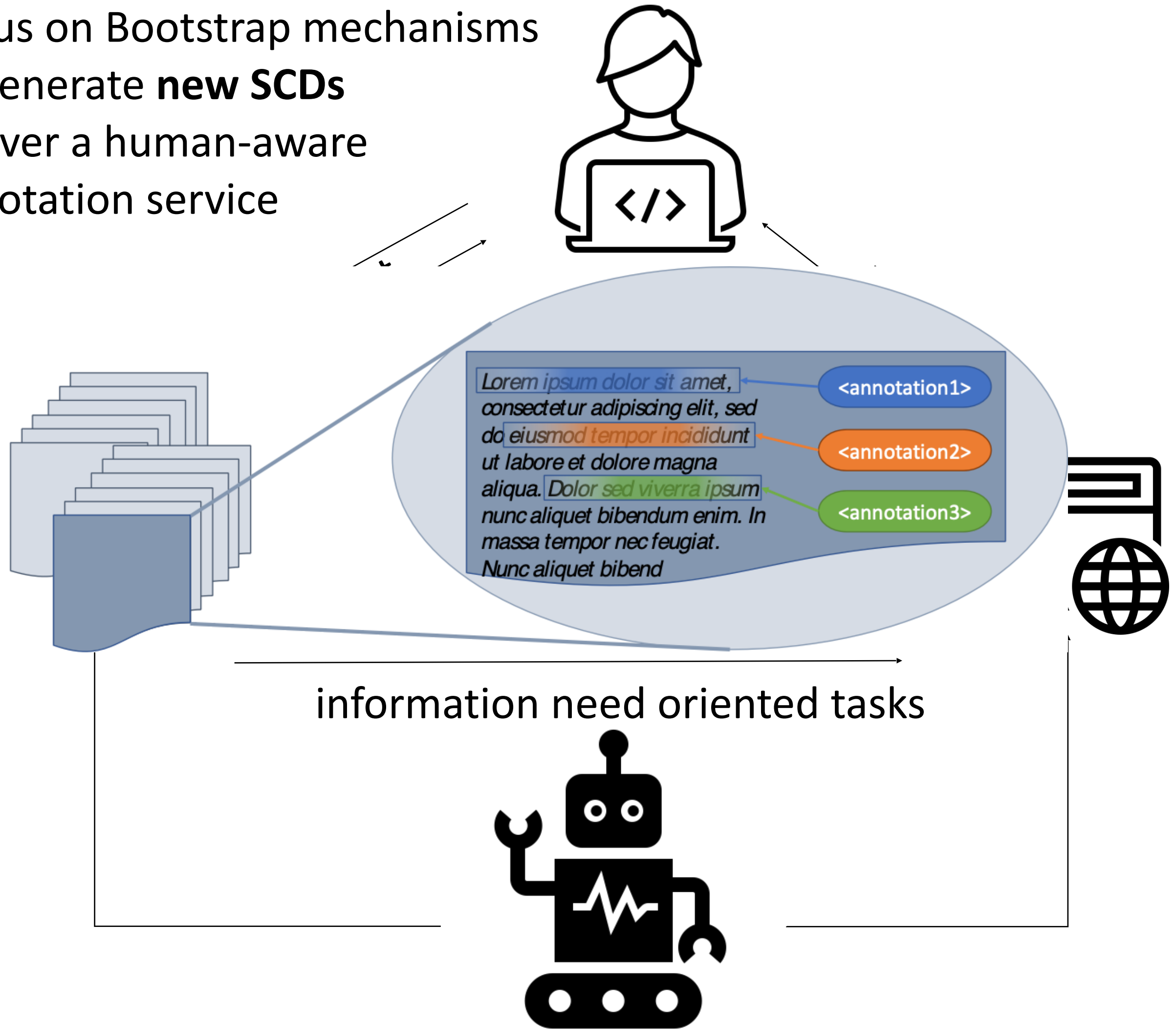
- Human-aware information retrieval considering not only content of documents and queries
- Fully automated annotation approach considering the human information need represented by a corpus and SCDs
- Approach for the bootstrap problem considering inline-SCDs

Focus on human-aware AI approaches:

→ Data linking **services** in a fashion that takes into aware **human expectations**

Future Work:

- Focus on Bootstrap mechanisms to generate **new SCDs**
- Deliver a human-aware annotation service



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
<annotation1>
<annotation2>
<annotation3>
information need oriented tasks

Referenzen

[1] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *Journal of machine Learning research* 3.Jan (2003): 993-1022.

[2] Felix Kuhr, Tanya Braun, Magnus Bender, Ralf Möller: To Extend or not to Extend? Context-specific Corpus Enrichment. *Proceedings of AI 2019: Advances in Artificial Intelligence, 2019, Springer, p.357-368*

[3] Felix Kuhr, Tanya Braun, Ralf Möller: Augmenting and Automating Corpus Enrichment. *Proceedings of the 14th IEEE International Conference on Semantic Computing (ICSC-20), 2020, <u>Best Student Paper Award</u>*

[4] Felix Kuhr, Bjarne Witten, Ralf Möller: Corpus-Driven Annotation Enrichment. *Proceedings of the 13th IEEE International Conference on Semantic Computing (ICSC-19), 2019, Jan, p.138-141*

Referenzen

[5] Magnus Bender, Tanya Braun, Marcel Gehrke, Felix Kuhr, Ralf Möller, Simon Schiff: Identifying Subjective Content Descriptions Among Texts. *15th IEEE International Conference on Semantic Computing, (ICSC 2021), Laguna Hills, CA, USA, January 27-29, 2021, IEEE, p.9-16*

[6] Felix Kuhr, Magnus Bender, Tanya Braun, Ralf Möller: Context-specific Adaptation of Subjective Content Descriptions. *15th IEEE International Conference on Semantic Computing, (ICSC 2021), Laguna Hills, CA, USA, January 27-29, 2021, IEEE, p.134-139*

[7] Felix Kuhr, Magnus Bender, Tanya Braun, Ralf Möller: Maintaining Topic Models for Growing Corpora. *Proceedings of the 14th IEEE International Conference on Semantic Computing (ICSC-20), 2020*