Semantic Text Matching of Company Policies and Regulatory Documents using Text Similarity Measures

Christoph Erl, February 26 2018, Master's Thesis Final Presentation

Chair of Software Engineering for Business Information Systems (sebis)
Faculty of Informatics
Technische Universität München
wwwmatthes.in.tum.de

Outline

Motivation
- References between Regulatory Documents and Company Policies
- Industry Partner Alyne GmbH
- Problem Statement
- Proposed Solution

Research Approach
- Iterative Approach
- Research Questions

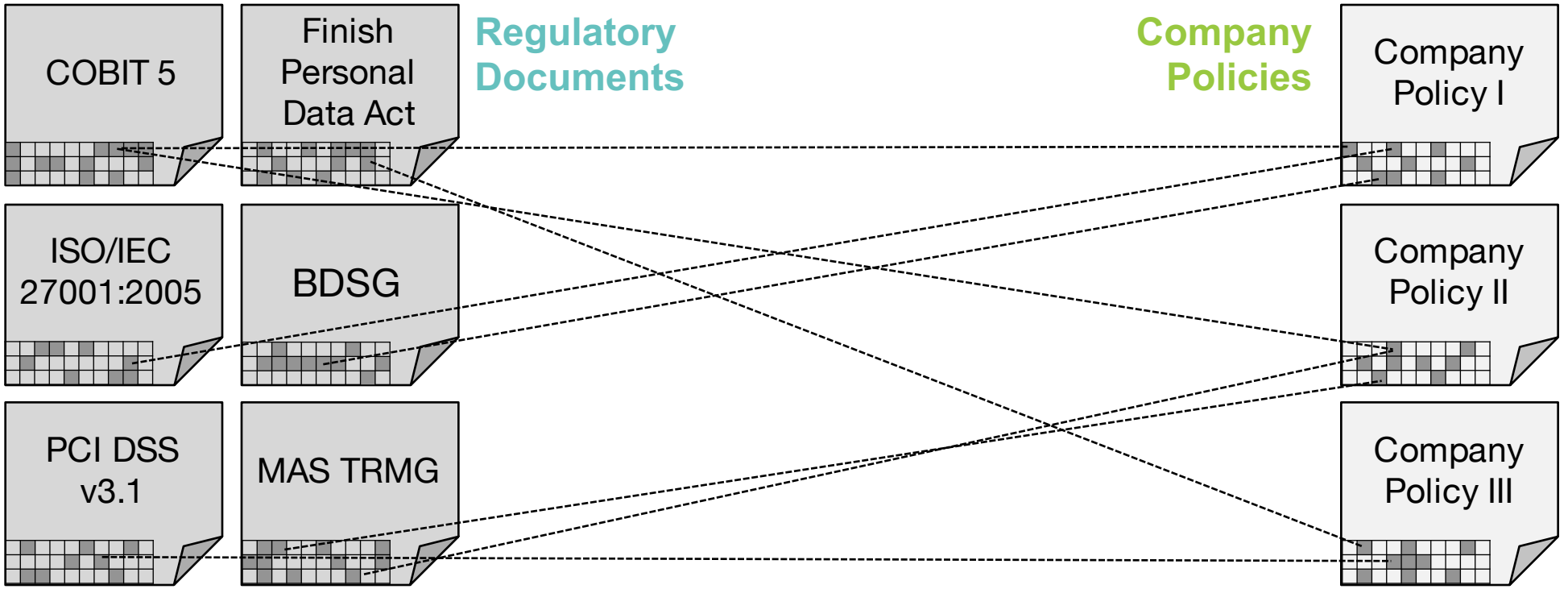
Implementation
- Recommender System
- Management & Analysis GUI

Evaluations & Results

Motivation

References between Company Policies and Regulatory Documents

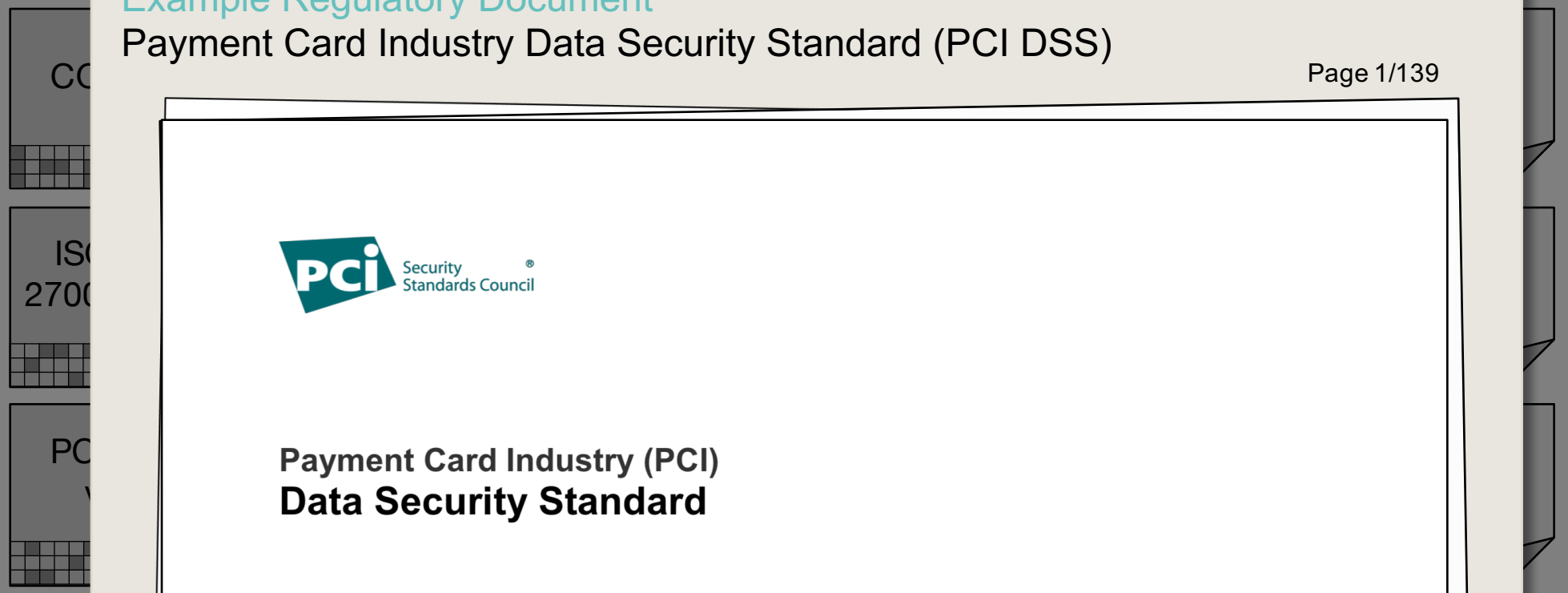
- Deriving company policies from regulatory documents is a legal requirement for companies
  → **Implicit reference**



- **Explicit references** are required for future audits or updates

Motivation
References between Company Policies and Regulatory Documents

- Deriving company policies from regulatory documents is a legal requirement for companies
  → **Implicit**



Example Regulatory Document
Payment Card Industry Data Security Standard (PCI DSS)

Page 1/139

PCI Security Standards Council

Payment Card Industry (PCI)
Data Security Standard

Requirements and Security Assessment Procedures

Version 3.2
April 2016

- **Explicit ref**

Motivation

References between Company Policies and Regulatory Documents

- Deriving company policies from regulatory documents is a legal requirement for companies
  → **Implicit**

CC

ISO 270

PC

- **Explicit ref**

Example Regulatory Document: PCI DSS

Payment Card Industry Data Security Standard

Page 69/139

PCI Security Standards Council

Requirement 8: Identify and authenticate access to system components

Assigning a unique identification (ID) to each person with access ensures that each individual is uniquely accountable for their actions. When such accountability is in place, actions taken on critical data and systems are performed by, and can be traced to, known and authorized users and processes.

The effectiveness of a password is largely determined by the design and implementation of the authentication system—particularly, how frequently password attempts can be made by an attacker, and the security methods to protect user passwords at the point of entry, during transmission, and while in storage.

Note: *These requirements are applicable for all accounts, including point-of-sale accounts, with administrative capabilities and all accounts used to view or access cardholder data or to access systems with cardholder data. This includes accounts used by vendors and other third parties (for example, for support or maintenance). These requirements do not apply to accounts used by consumers (e.g., cardholders).*

However, Requirements 8.1.1, 8.2, 8.5, 8.2.3 through 8.2.5, and 8.1.6 through 8.1.8 are not intended to apply to user accounts within a point-of-sale payment application that only have access to one card number at a time in order to facilitate a single transaction (such as cashier accounts).

PCI DSS Requirements	Testing Procedures	Guidance
8.1 Define and implement policies and procedures to ensure proper user identification management for non-consumer users and administrators on all system components as follows:	**8.1.a** Review procedures and confirm they define processes for each of the items below at 8.1.1 through 8.1.8	By ensuring each user is uniquely identified—instead of using one ID for several employees—an organization can maintain individual responsibility for actions and an effective audit trail per employee. This will help speed issue resolution and containment when misuse or malicious intent occurs.
	8.1.b Verify that procedures are implemented for user identification management, by performing the following:	
8.1.1 Assign all users a unique ID before allowing them to access system components or cardholder data.	**8.1.1** Interview administrative personnel to confirm that all users are assigned a unique ID for access to system components or cardholder data.	
8.1.2 Control addition, deletion, and modification of user IDs, credentials, and other identifier objects.	**8.1.2** For a sample of privileged user IDs and general user IDs, examine associated authorizations and observe system settings to verify each user ID and privileged user ID has been implemented with only the privileges specified on the documented approval.	To ensure that user accounts granted access to systems are all valid and recognized users, strong processes must manage all changes to user IDs and other authentication credentials, including adding new ones and modifying or deleting existing ones.

Motivation

References between Company Policies and Regulatory Documents

- Deriving company policies from regulatory documents is a legal requirement for companies
  → **Implicit**

CC

ISO 270

PC

- **Explicit ref**

Example Regulatory Document: PCI DSS

Payment Card Industry Data Security Standard

Page 70/139

PCI Security Standards Council

PCI DSS Requirements	Testing Procedures	Guidance
8.2.5 Do not allow an individual to submit a new password/passphrase that is the same as any of the last four passwords/passphrases he or she has used.	**8.2.5.a** For a sample of system components, obtain and inspect system configuration settings to verify that password parameters are set to require that new passwords/passphrases cannot be the same as the four previously used passwords/passphrases.	If password history isn't maintained, the effectiveness of changing passwords is reduced, as previous passwords can be reused over and over. Requiring that passwords cannot be reused for a period of time reduces the likelihood that passwords that have been guessed or brute-forced will be used in the future.
	8.2.5.b ***Additional testing procedure for service provider assessments only:*** Review internal processes and customer/user documentation to verify that new non-consumer customer user passwords/passphrase cannot be the same as the previous four passwords.	***Note:*** *Testing Procedure 8.2.5.b is an additional procedure that only applies if the entity being assessed is a service provider.*
8.2.6 Set passwords/passphrases for first-time use and upon reset to a unique value for each user, and change immediately after the first use.	**8.2.6** Examine password procedures and observe security personnel to verify that first-time passwords/passphrases for new users, and reset passwords/passphrases for existing users, are set to a unique value for each user and changed after first use.	If the same password is used for every new user, an internal user, former employee, or malicious individual may know or easily discover this password, and use it to gain access to accounts.
8.3 Secure all individual non-console administrative access and all remote access to the CDE using multi-factor authentication. ***Note:*** *Multi-factor authentication requires that a minimum of two of the three authentication methods (see Requirement 8.2 for descriptions of authentication methods) be used for authentication. Using one factor twice (for example, using two separate passwords) is not considered multi-factor authentication.*		Multi-factor authentication requires an individual to present a minimum of two separate forms of authentication (as described in Requirement 8.2), before access is granted. Multi-factor authentication provides additional assurance that the individual attempting to gain access is who they claim to be. With multi-factor authentication, an attacker would need to compromise at least two different authentication mechanisms, increasing the difficulty of compromise and thus reducing the risk. Multi-factor authentication is not required at both the system-level and application-level for a particular system component. Multi-factor authentication can

Motivation

References between Company Policies and Regulatory Documents

- Deriving company policies from regulatory documents is a legal requirement for companies
  → **Implicit**

- **Explicit ref**

Example Regulatory Document: PCI DSS

Payment Card Industry Data Security Standard

Page 70/139

PCI Security Standards Council

PCI DSS Requirements	Testing Procedures	Guidance
8.2.5 Do not allow an individual to submit a new password/passphrase that is the same as any of the last four passwords/passphrases he or she has used.	**8.2.5.a** For a sample of system components, obtain and inspect system configuration settings to verify that password parameters are set to require that new passwords/passphrases cannot be the same as the four previously used passwords/passphrases.	If password history isn't maintained, the effectiveness of changing passwords is reduced, as previous passwords can be reused over and over. Requiring that passwords cannot be reused for a period of time reduces the likelihood that passwords that have been guessed or brute-forced will be used in the future.
	8.2.5.b ***Additional testing procedure for service provider assessments only:*** Review internal processes and customer/user documentation to verify that new non-consumer customer user passwords/passphrase cannot be the same as the previous four passwords.	***Note:*** *Testing Procedure 8.2.5.b is an additional procedure that only applies if the entity being assessed is a service provider.*
8.2.6 Set passwords/passphrases for first-time use and upon reset to a unique value for each user, and change immediat	**8.2.6** Examine password procedures and observe security personnel to verify that first-time passwords/passphrases for new users, and reset passwords/passphrases for existing	If the same password is used for every new user, an internal user, former employee, or malicious ver this to accounts.
8.3 Secure all indi administrative acc access to the CDE authentication.		an individual to forms of uirement 8.2),
Note: *Multi-factor a that a minimum of t authentication meth Requirement 8.2 fo authentication meth authentication. Usin (for example, using two separate passwords) is not considered multi-factor authentication.*		additional ting to gain multi-factor ed to thentication of compromise and thus reducing the risk. Multi-factor authentication is not required at both the system-level and application-level for a particular system component. Multi-factor authentication can

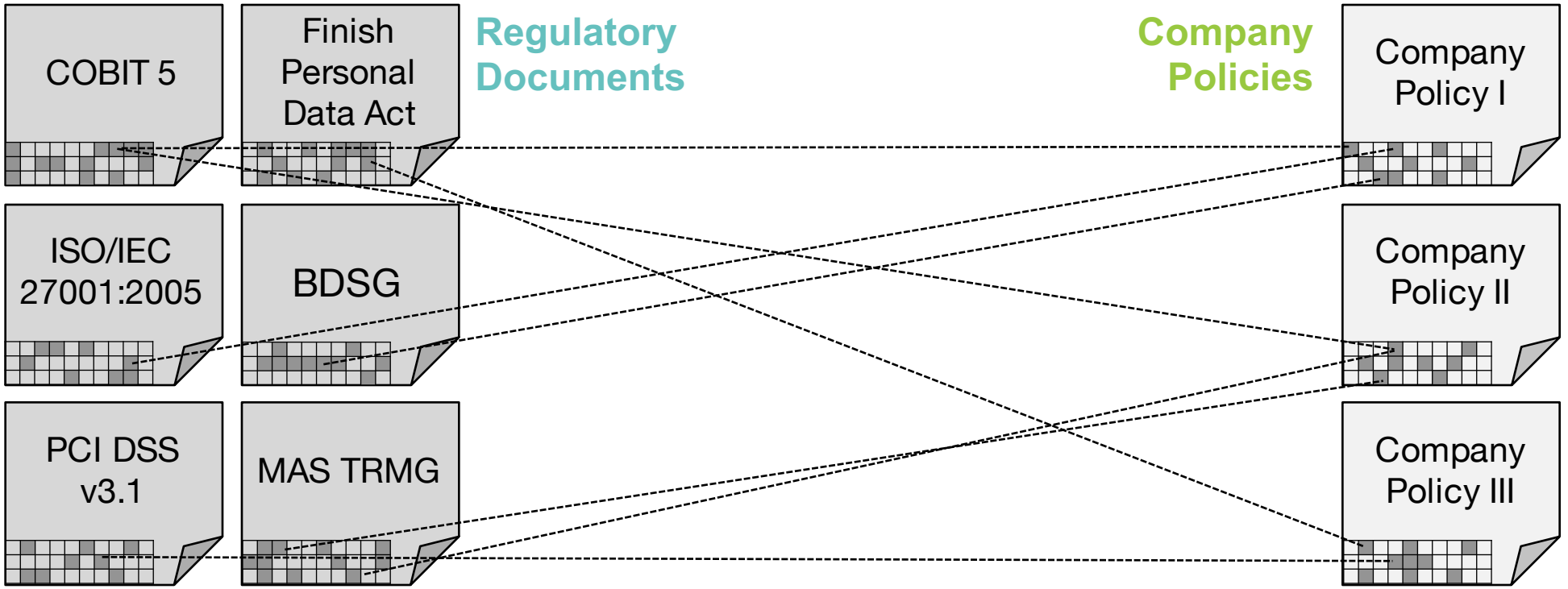
Example Paragraph: PCI DSS § 8.2.5

Do not allow an individual to submit a new password that is the same as any of the last four passwords he or she has used.

Motivation

References between Company Policies and Regulatory Documents

- Deriving company policies from regulatory documents is a legal requirement for companies
  → **Implicit reference**

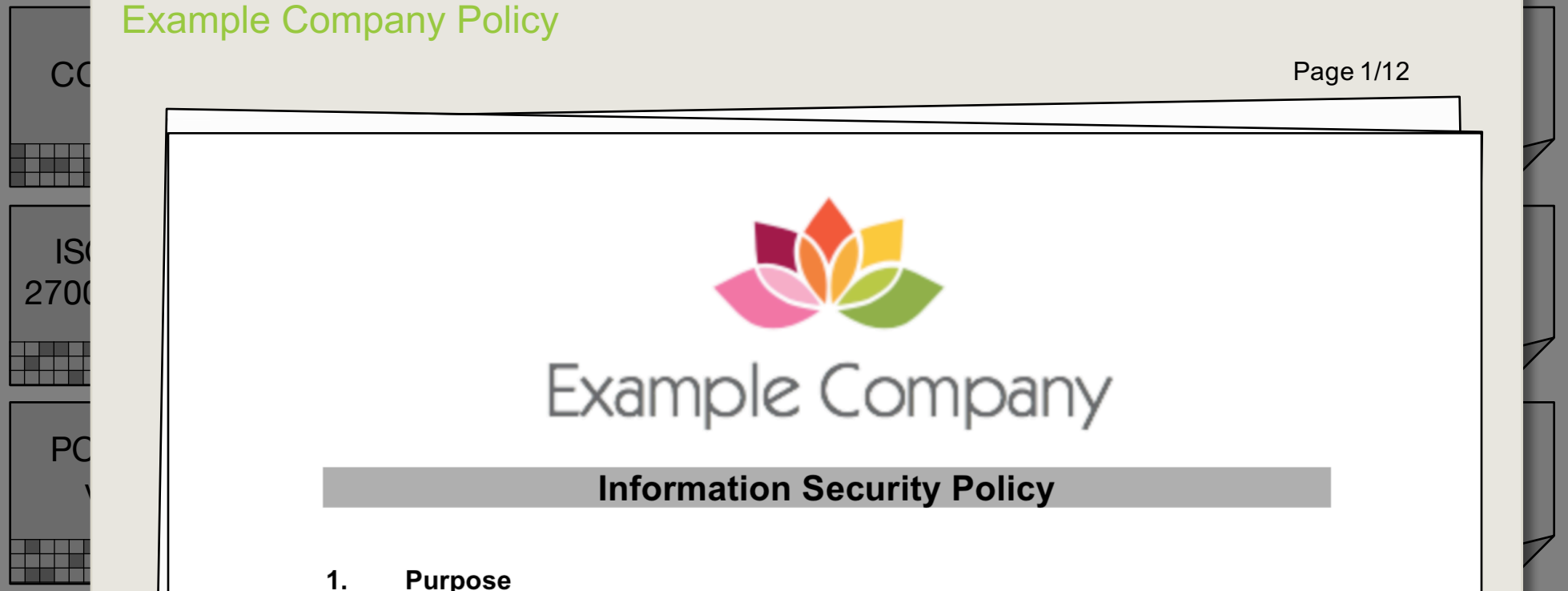


- **Explicit references** are required for future audits or updates

Motivation

References between Company Policies and Regulatory Documents

- Deriving company policies from regulatory documents is a legal requirement for companies
  → **Implicit**

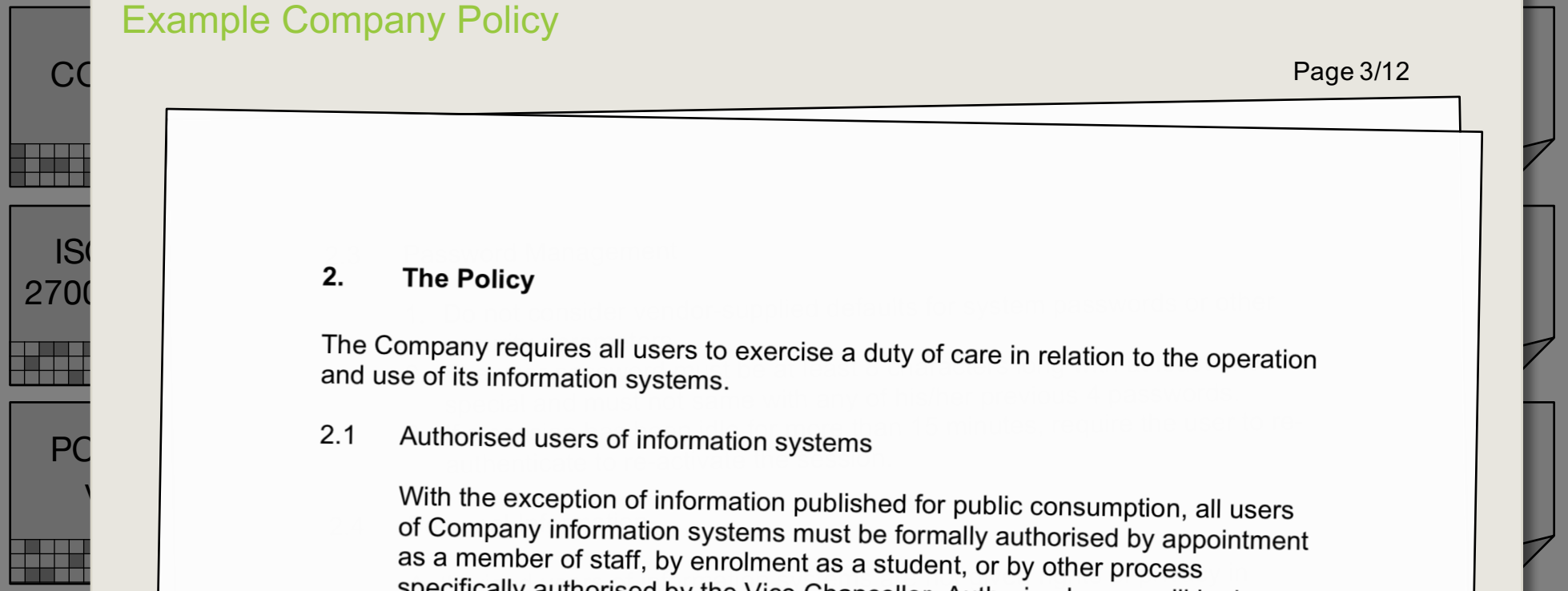


- **Explicit ref**

Motivation

References between Company Policies and Regulatory Documents

- Deriving company policies from regulatory documents is a legal requirement for companies
  → **Implicit**



- **Explicit ref**

Example Company Policy

Page 3/12

2. The Policy

The Company requires all users to exercise a duty of care in relation to the operation and use of its information systems.

2.1 Authorised users of information systems

With the exception of information published for public consumption, all users of Company information systems must be formally authorised by appointment as a member of staff, by enrolment as a student, or by other process specifically authorised by the Vice Chancellor. Authorised users will be in possession of a unique user identity. Any password associated with a user identity must not be disclosed to any other person. The "Network password policy" describes these principles in greater detail.

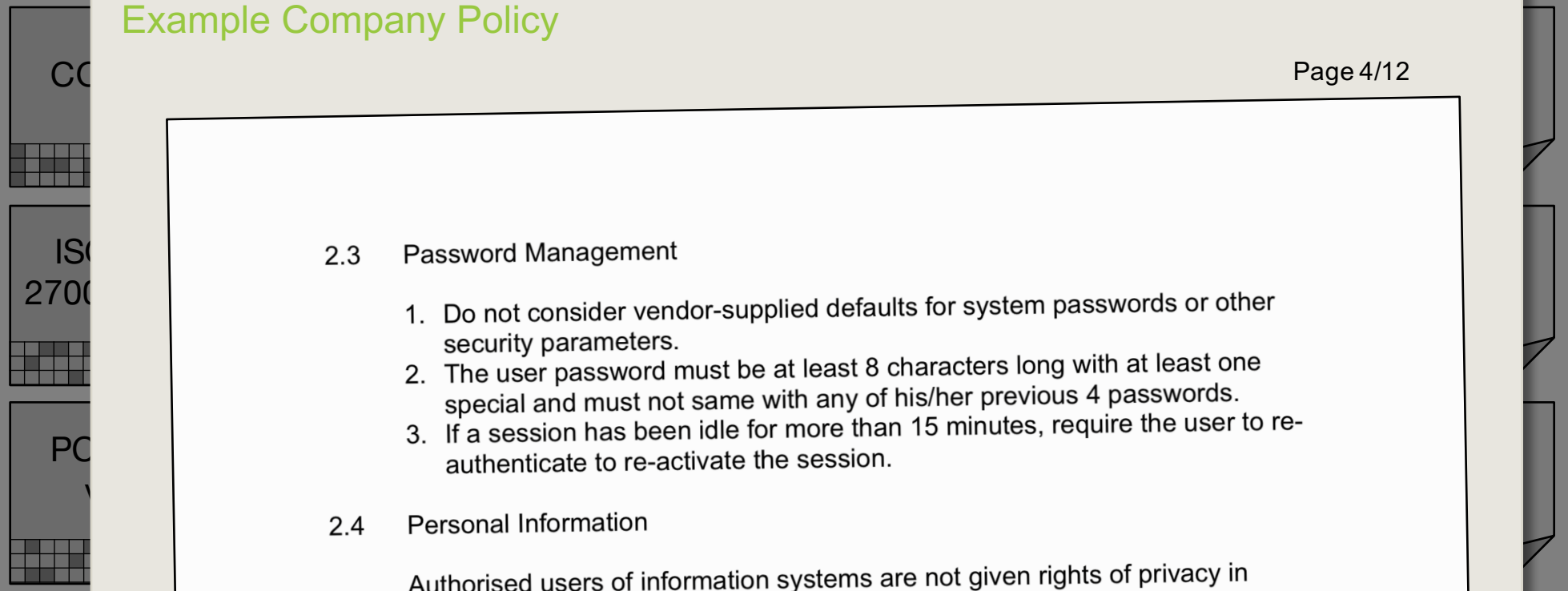
Authorised users will pay due care and attention to protect Company information in their personal possession. Confidential, personal or private information must not be copied or transported without consideration of:

- permission of the information owner

Motivation

References between Company Policies and Regulatory Documents

- Deriving company policies from regulatory documents is a legal requirement for companies
  → **Implicit**



- **Explicit ref**

Example Company Policy

Page 4/12

2.3 Password Management

1. Do not consider vendor-supplied defaults for system passwords or other security parameters.
2. The user password must be at least 8 characters long with at least one special and must not same with any of his/her previous 4 passwords.
3. If a session has been idle for more than 15 minutes, require the user to re-authenticate to re-activate the session.

2.4 Personal Information

Authorised users of information systems are not given rights of privacy in relation to their use of Company information systems. Duly authorised officers of the Company may access or monitor personal data contained in any Company information system (mailboxes, web access logs, file-store etc).

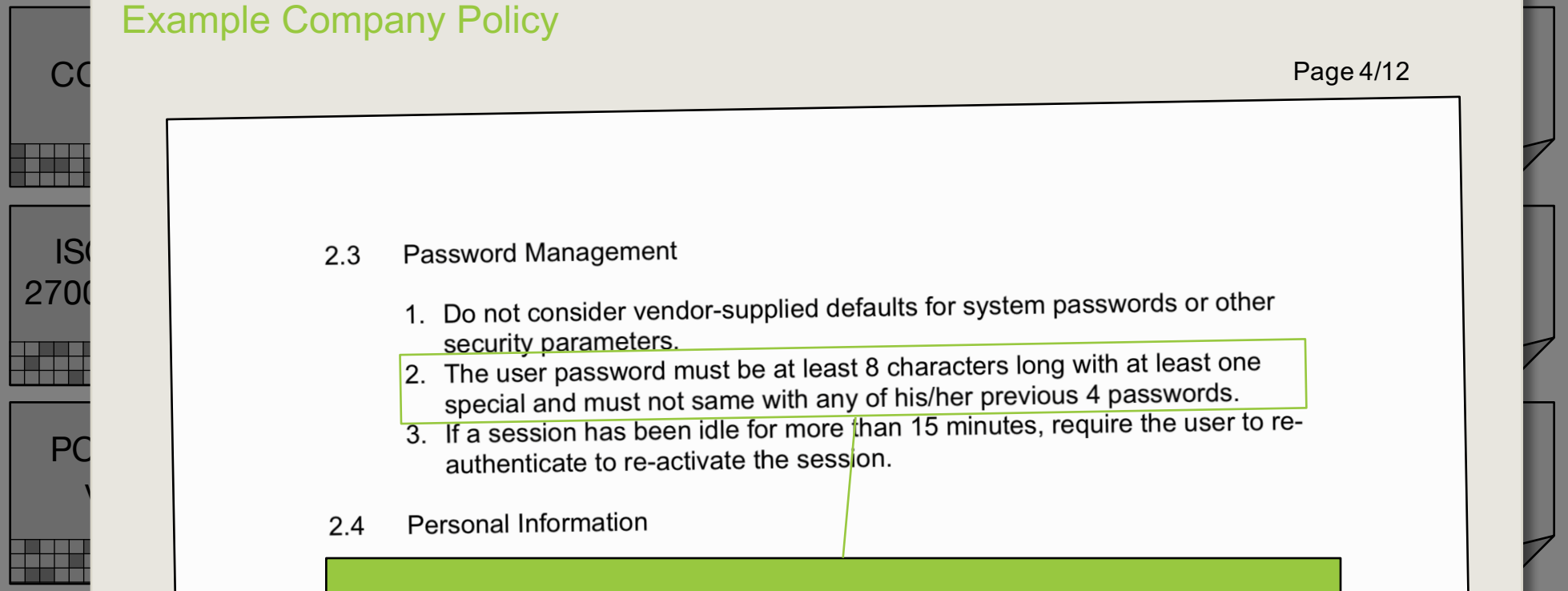
2.5 Individuals in breach of this policy are subject to disciplinary procedures (staff or student) at the instigation of the Dean/Director with responsibility for the relevant information system, including referral to the Police where appropriate.

The Company will take legal action to ensure that its information systems are

Motivation

References between Company Policies and Regulatory Documents

- Deriving company policies from regulatory documents is a legal requirement for companies
  → **Implicit**



- **Explicit ref**

Example Company Policy

Page 4/12

2.3 Password Management

1. Do not consider vendor-supplied defaults for system passwords or other security parameters.
2. The user password must be at least 8 characters long with at least one special and must not same with any of his/her previous 4 passwords.
3. If a session has been idle for more than 15 minutes, require the user to re-authenticate to re-activate the session.

2.4 Personal Information

Example Paragraph: § 2.3.2

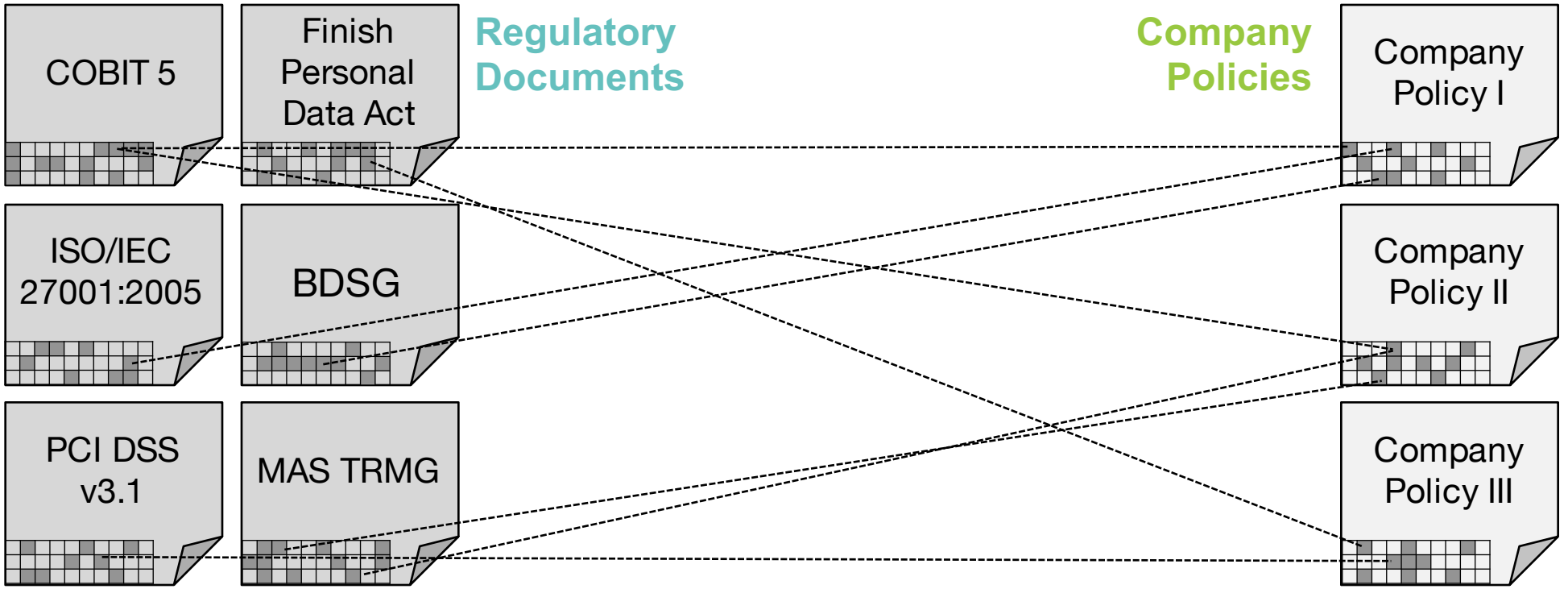
The user password must be at least 8 characters long with at least one special character and must not be same with any of his/her previous 4 passwords.

The Company will take legal action to ensure that its information systems are

Motivation

References between Company Policies and Regulatory Documents

- Deriving company policies from regulatory documents is a legal requirement for companies
  → **Implicit reference**

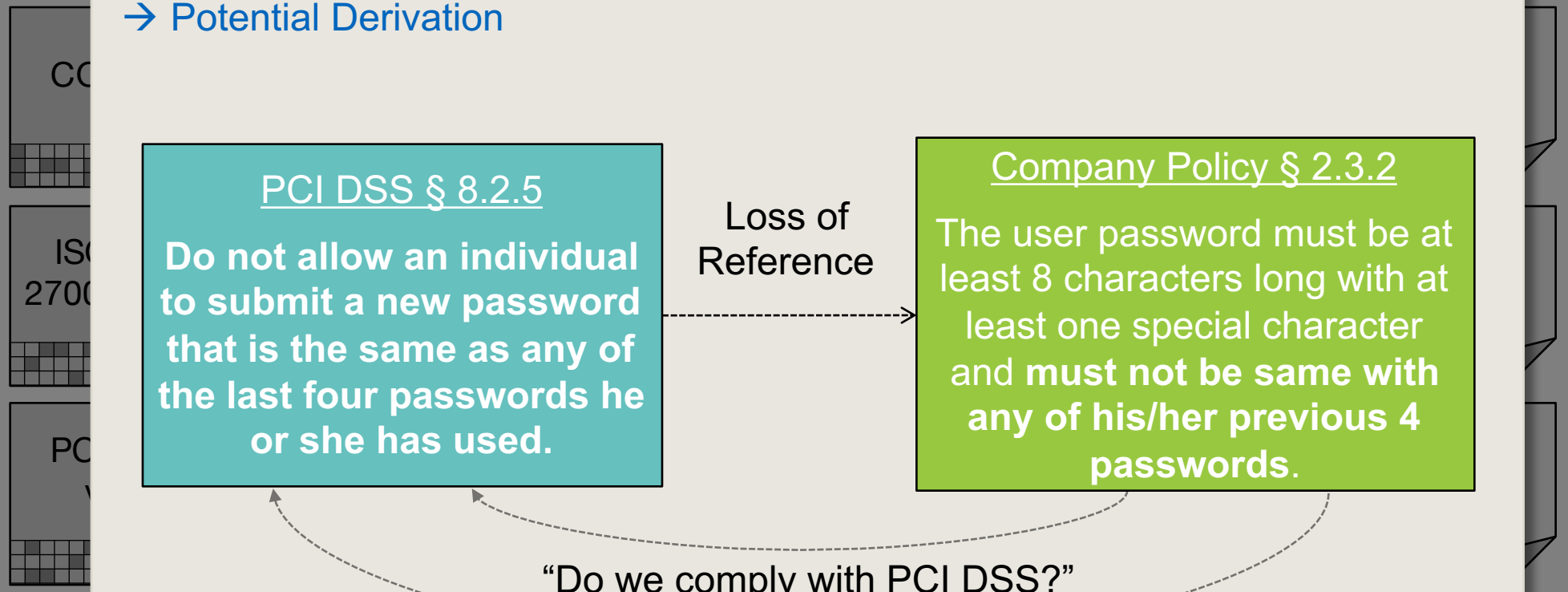


- **Explicit references** are required for future audits or updates

Motivation

References between Company Policies and Regulatory Documents

- Deriving company policies from regulatory documents is a legal requirement for companies
  → **Implicit**

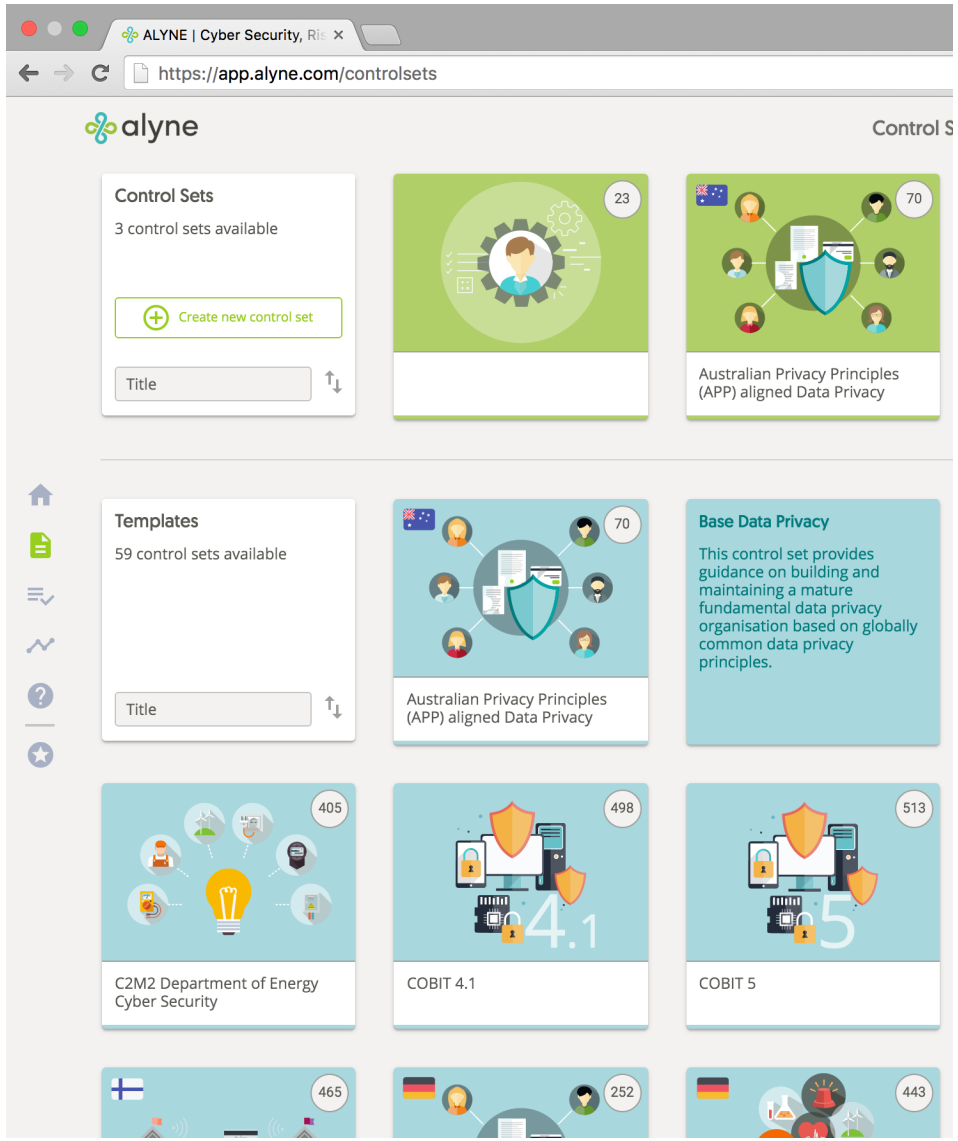


- **Explicit ref**

Motivation

Industry Partner Alyne GmbH

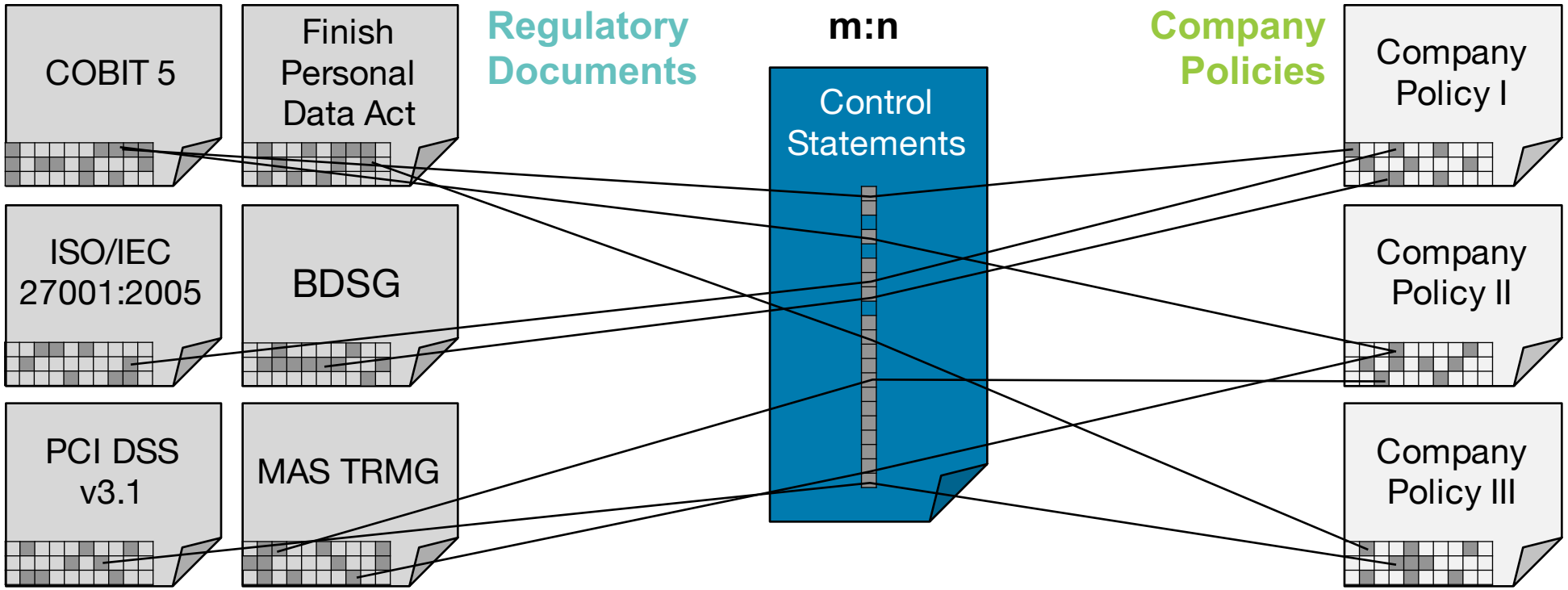
- RegTech Company
  - Launch in 2015
  - Software-as-a-Service (SAAS)
  - B2B
- Business Idea
  - Alyne delivers a software that supports organisations to manage their cyber security, risk management and compliance.
  - Generalized **control statements** serve as a glue between **company policies** and **regulatory documents**



Motivation

Industry Partner Alyne GmbH – Control Statements

- Paragraphs from **regulatory documents** and **company policies** are linked by generalized, well defined control statements
  → **Explicit References**



PCI DSS § 8.2.5
Do not allow an individual to submit a new password that is the same as any of the last four passwords […].

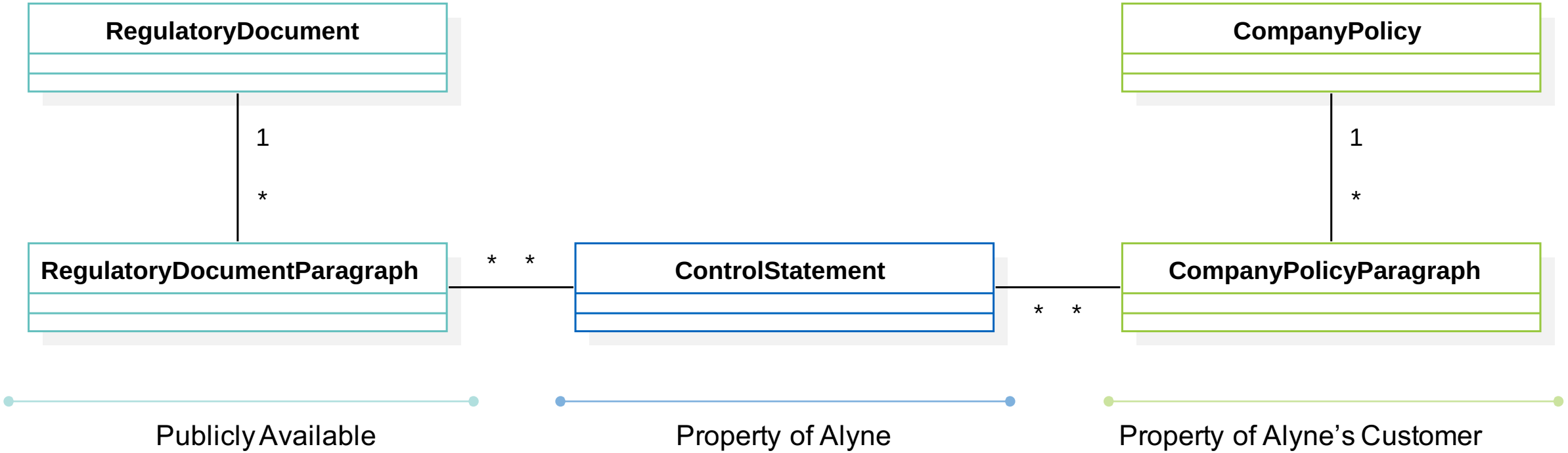
Control Statement
User passwords shall be prevented from being changed to any of the previous x passwords.

Company Policy § 2.3.2
The user password […] must not be same with any of his/her previous 4 passwords.

Motivation

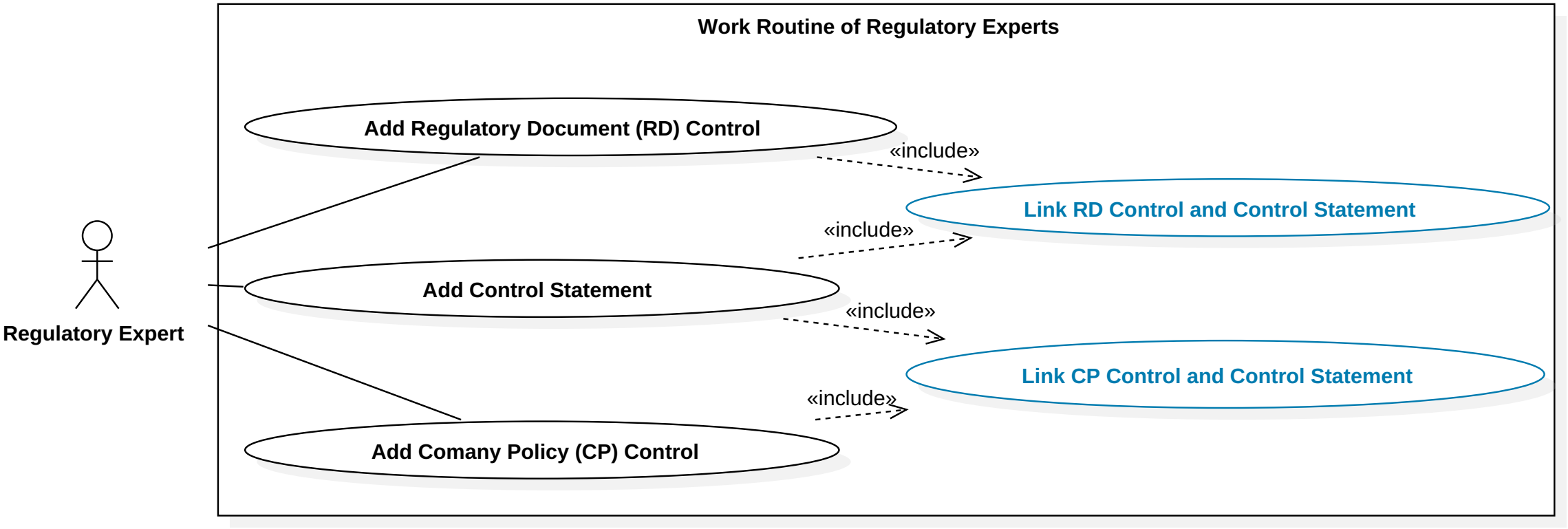
Industry Partner Alyne GmbH – Data Model

- Paragraphs from **regulatory documents** and **company policies** are linked by generalized, well defined control statements
  → **Explicit References**



Motivation
Problem Statement

- Linking Controls to Control Statements by hand is a **very labour intensive process for regulatory experts!**

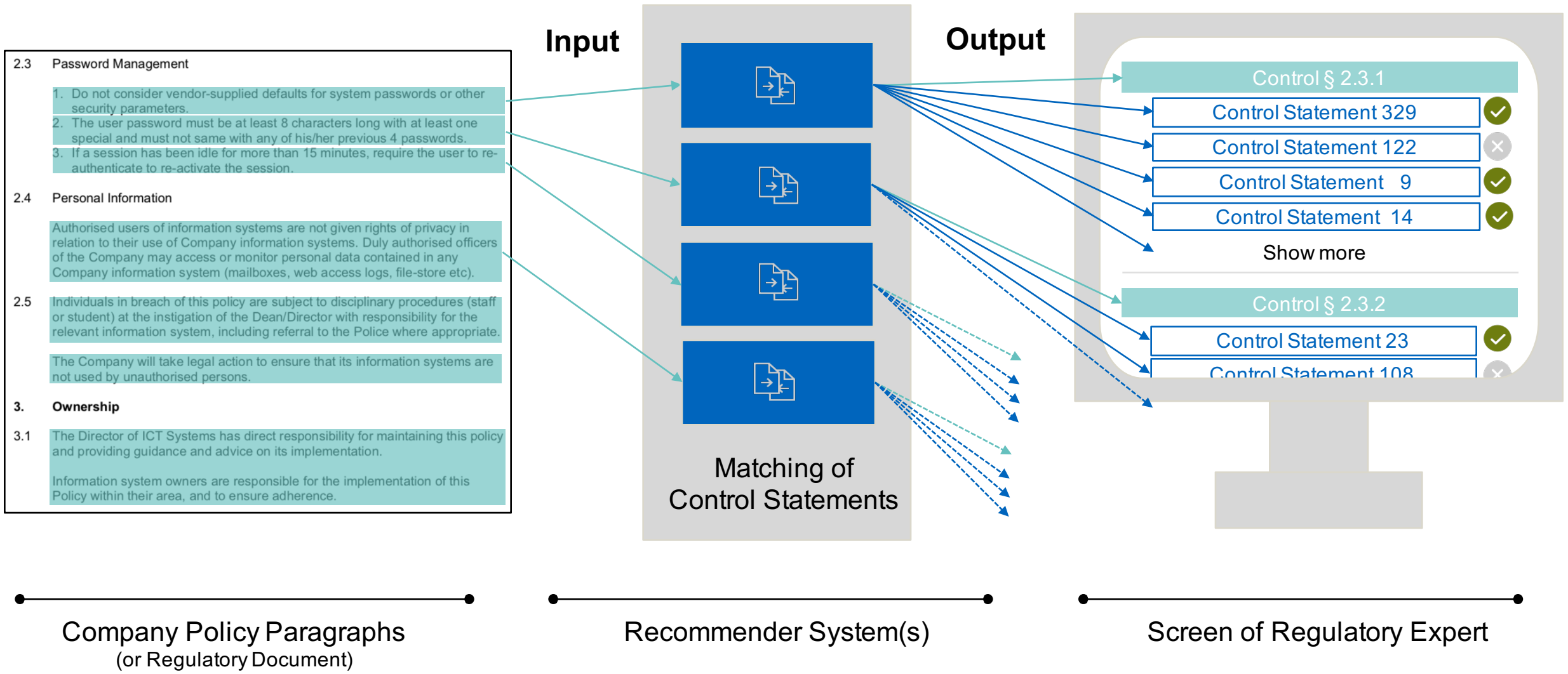


- **Problem Statement:**
  How can we support regulatory experts in their work routine, in particular, to link company policy controls and regulatory documents to control statements?

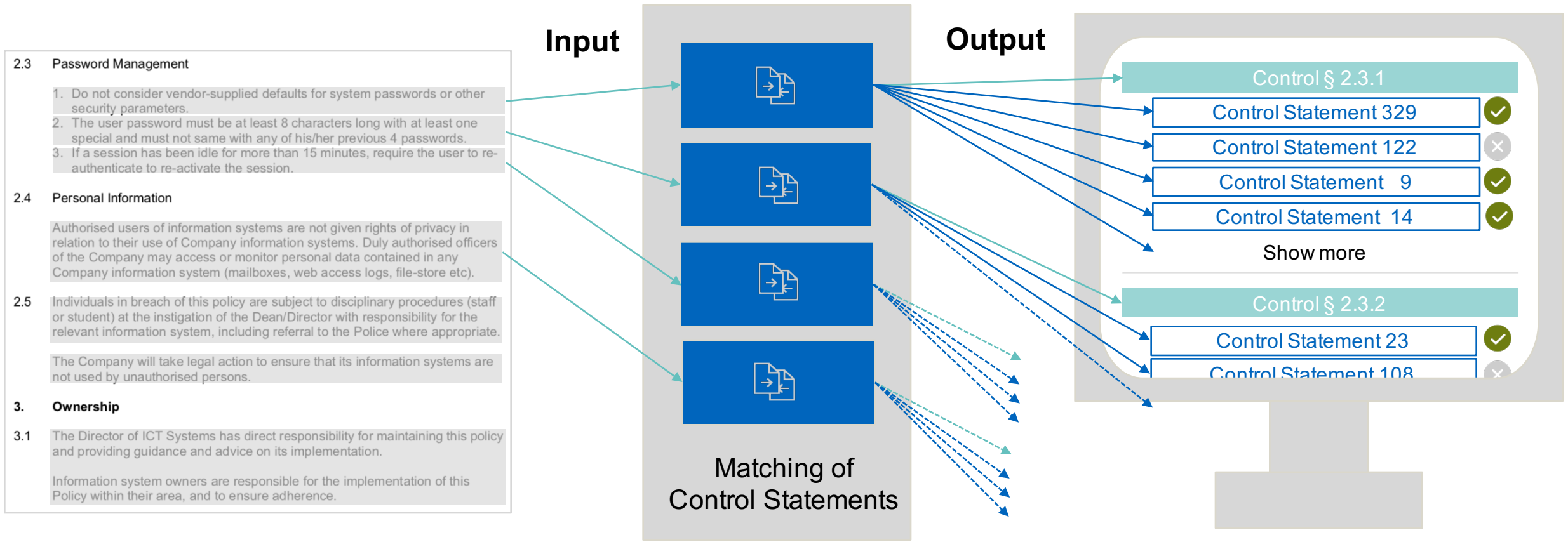
- **Proposed Solution:**
  - A **recommender system** suggests related control statements for a given input control **using text similarity approaches**.
  - Whenever an expert intends to link a control, the system sorts the collection of 879 control statements according to their semantic similarities to the control whereby the more similar control statements appear at the top.

- The intended effect is that experts
  - do not need to check all control statements but only the top results (**time savings**) and
  - find control statements which they might have forgotten (**improved data quality**).

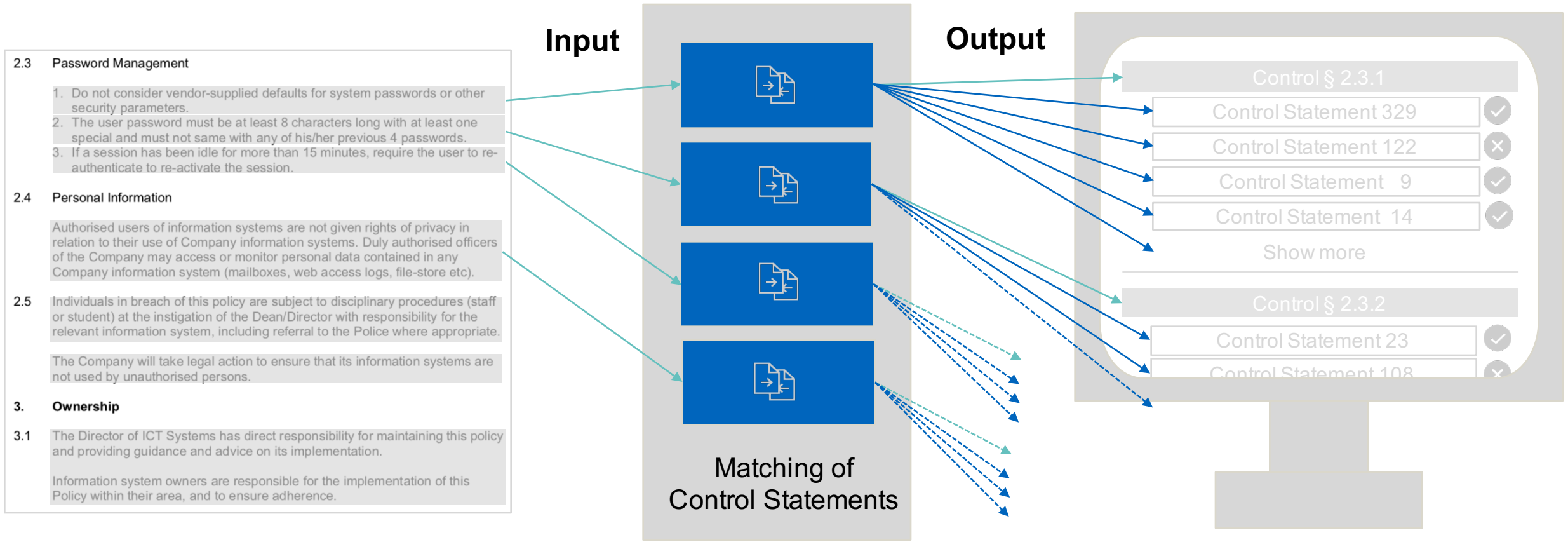
Motivation
Proposed Solution



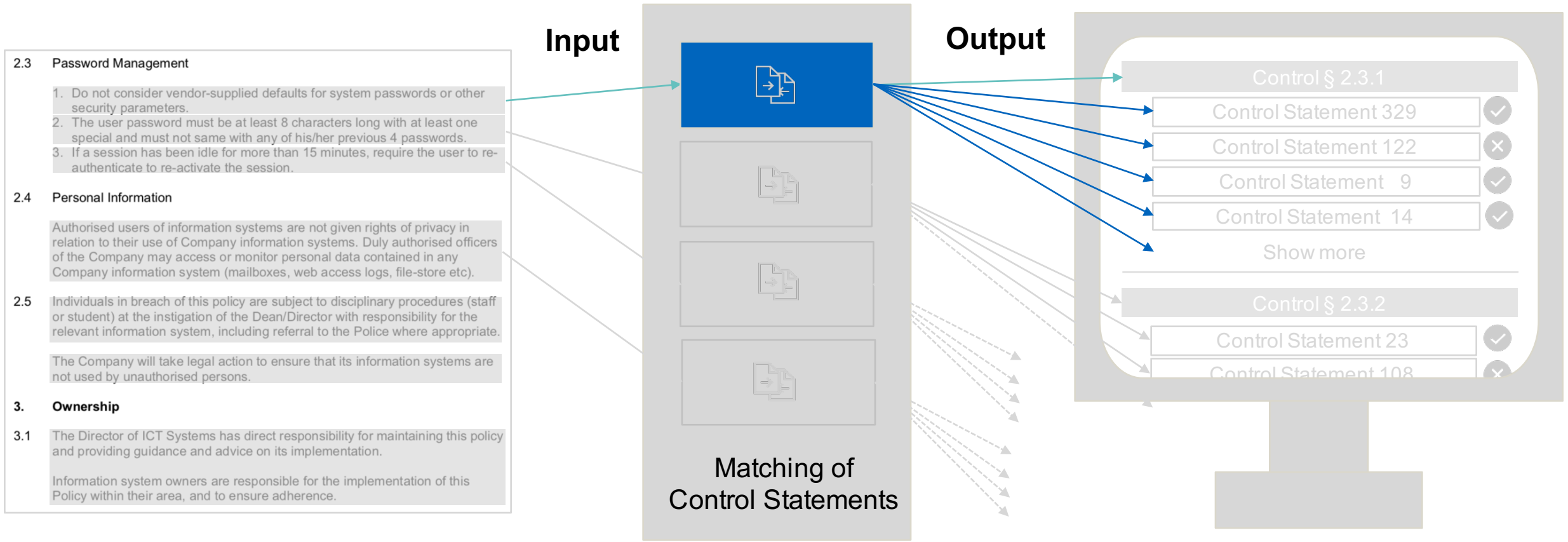
Motivation
Proposed Solution



Motivation
Proposed Solution

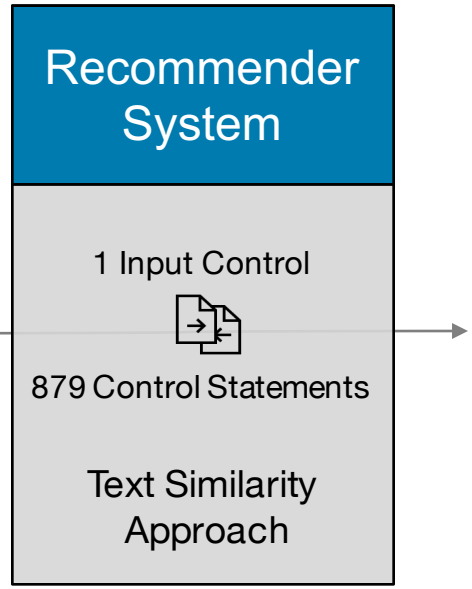


Motivation
Proposed Solution



Motivation
Proposed Solution

Do not allow an individual to submit a new password that is the same as any of the last four passwords he/she has used.



1. User passwords shall be prevented from being changed to any of the previous 10 passwords.
2. Privileged account passwords shall be prevented from being changed to any of the previous 15 passwords.

⋮

879. Recovery from backup media shall be tested at least every year.

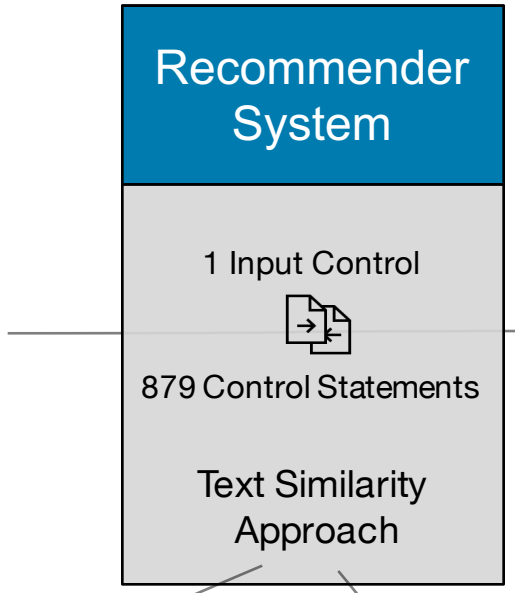
Input

Output

Motivation
Proposed Solution

Do not allow an individual to submit a new password that is the same as any of the last four passwords he/she has used.

[0.1, 0.5, 0.2, ..., 0.9]



Vector Representation
- TF-IDF **or**
- Word2Vec **or**
- Doc2Vec

(accurately matching terms only)

(synonyms to some degree)

Similarity Measure
- Cosine Similarity

1. User passwords shall be prevented from being changed to any of the previous 10 passwords.
   [0.2, 0.6, 0.3, ..., 0.7]
   → Similarity: 95%

2. Privileged account passwords shall be prevented from being changed to any of the previous 15 passwords.
   [0.4, 0.7, 0.2, ..., 0.8]
   → Similarity: 92%

⋮

879. Recovery from backup media shall be tested at least every year.
   [0.8, 0.1, 0.4, ..., 0.2]
   → Similarity: 40%

Input

Output

Outline

Motivation
- References between Regulatory Documents and Company Policies
- Industry Partner Alyne GmbH
- Problem Statement
- Proposed Solution

Research Approach
- Research Questions
- Iterative Approach

Implementation
- Recommender System
- Management & Analysis GUI

Evaluations & Results

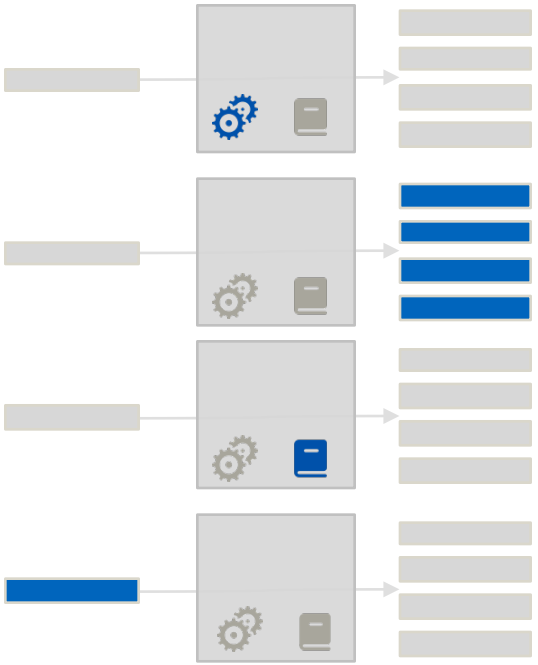
Research Approach
Research Questions

Overall Evaluation

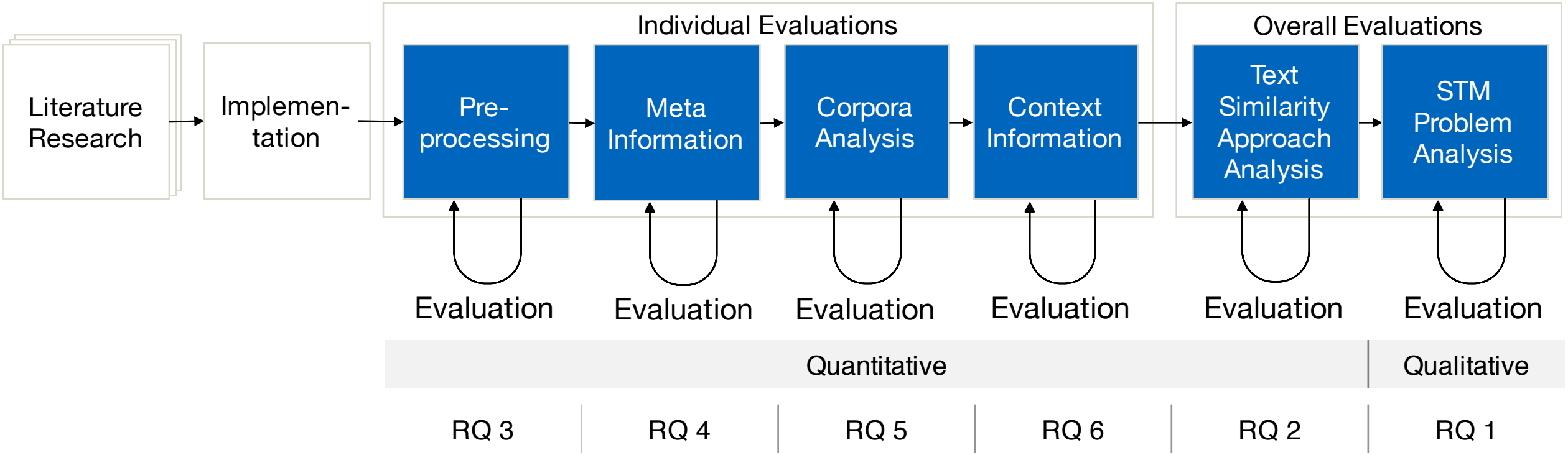
1. To which degree does the text similarity approach solve the semantic text matching problem?
2. What text similarity approach performs best - TF-IDF, Word2Vec or Doc2Vec?

Individual Evaluation

3. Which preprocessing technologies have a positive impact on the matching results?
4. Does the addition of meta information to control statements improve the results?
5. What are good corpora to train Word2Vec and Doc2Vec?
6. Can chapter or paragraph context help to improve the results?



Research Approach
Iterative Approach



Quantitative Evaluation: Rank-Position-Score (RPS, average rank)

- 1000 ground truth items from 10 regulatory documents

Qualitative Evaluation: Feedback sheet

RP	Control Statement Id
1 ✓	**Control_Statement_Id_00033**
2	Control_Statement_Id_00544
3 ✓	**Control_Statement_Id_00056**
4	Control_Statement_Id_00612
...	(and 875 more)

RPS = (1+3)/2 = 2

Outline

Motivation
- References between Regulatory Documents and Company Policies
- Industry Partner Alyne GmbH
- Problem Statement
- Proposed Solution

Research Approach
- Research Questions
- Iterative Approach

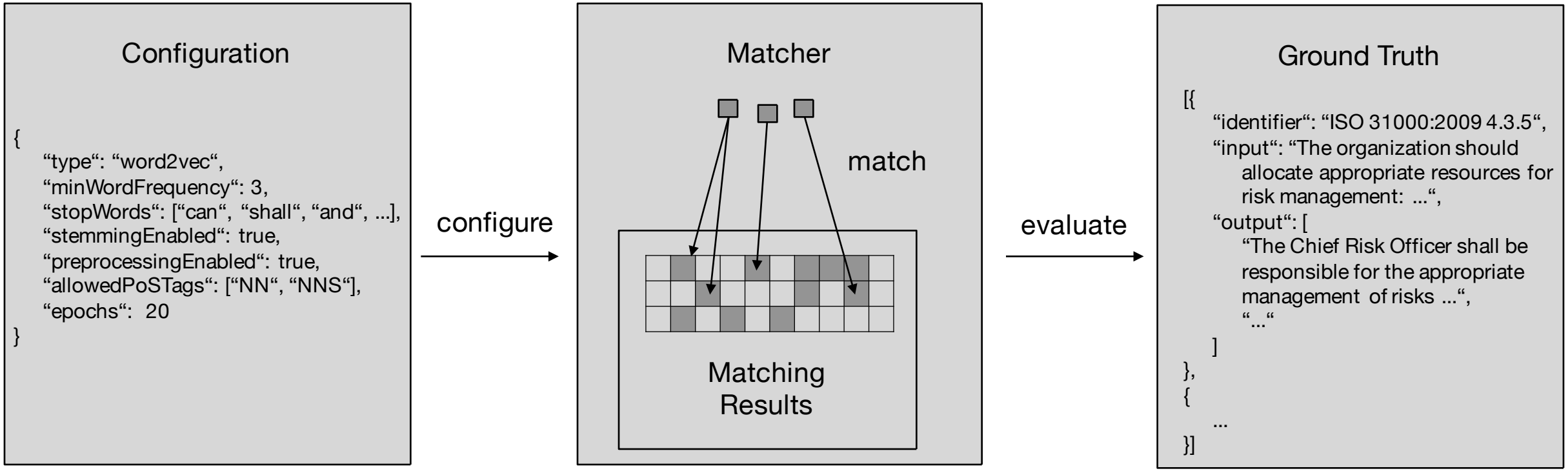
Implementation
- Recommender System
- Management & Analysis GUI

Evaluations & Results

Implementation
Recommender System

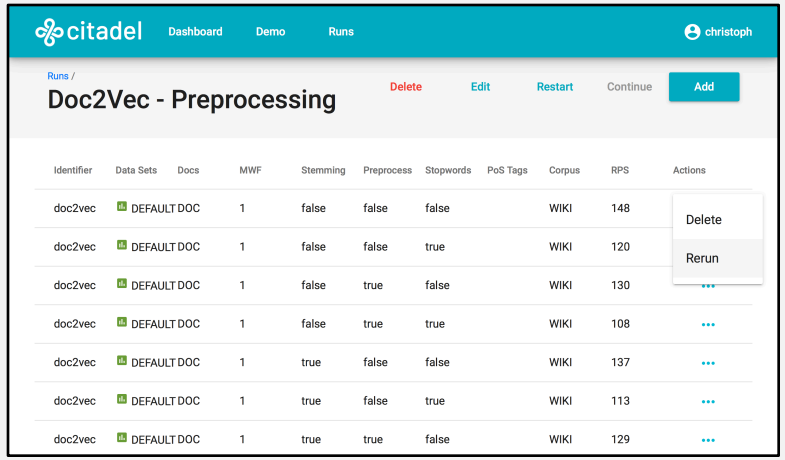
- **Core functionality**: Match a text passage against a collection of other text passages using text similarity approaches (TF-IDF, Word2Vec, Doc2Vec)

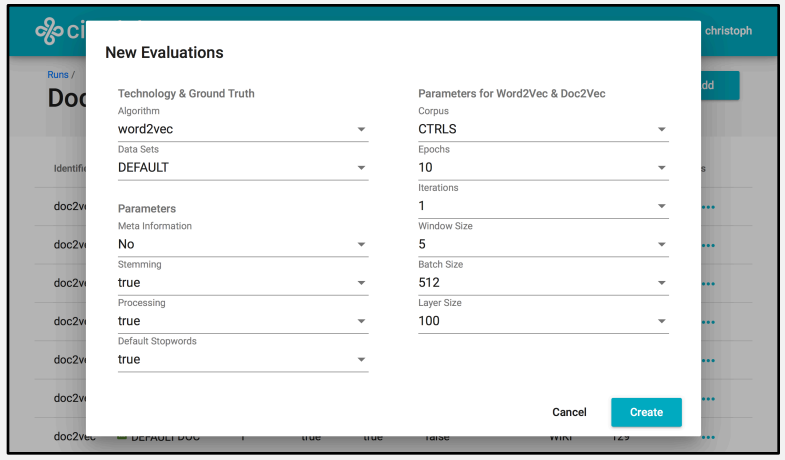
- Framework DeepLearning4Java (DL4J)



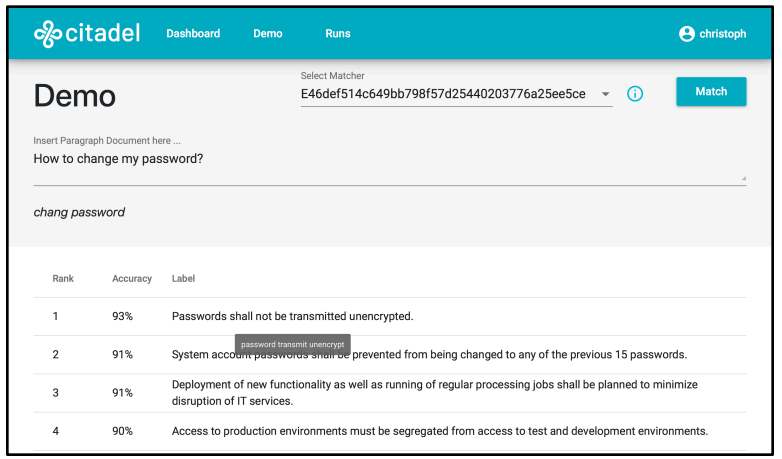
Implementation

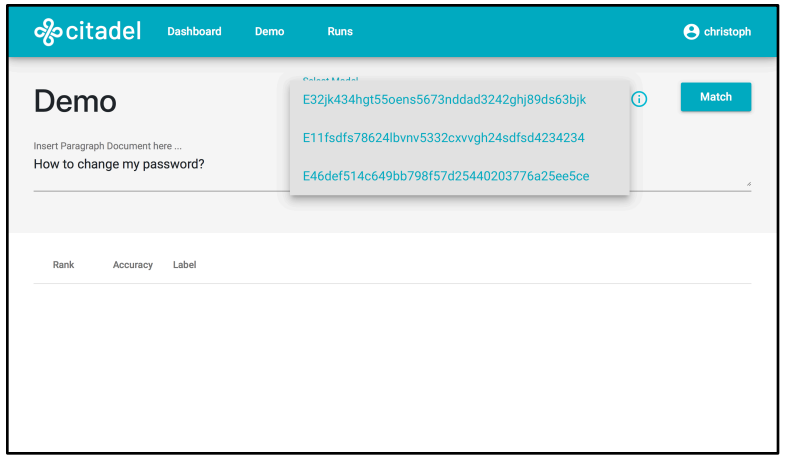
Management & Analysis GUI



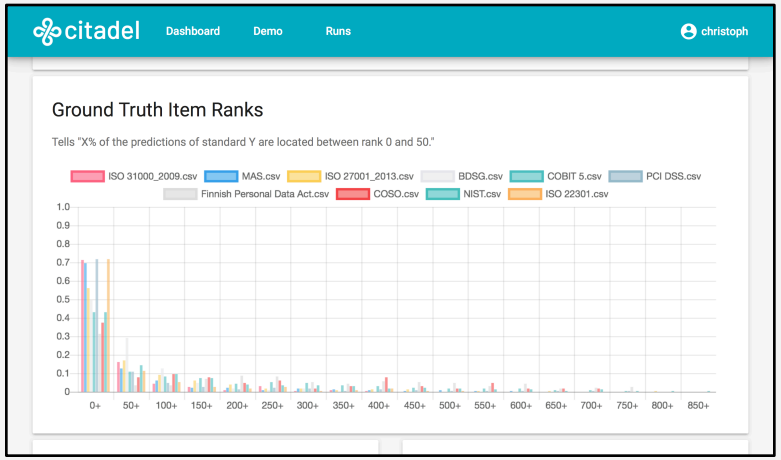


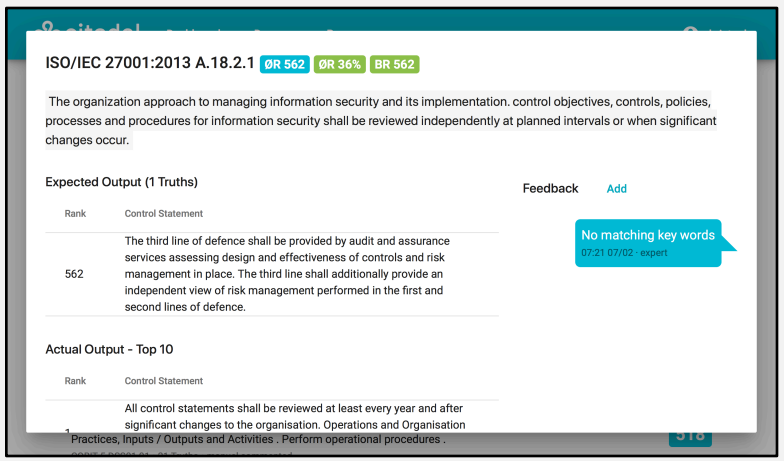
Evaluation Management





Evaluation Analysis – Demo





Evaluation Analysis – Evaluation Details & Feedback

Outline

Motivation
- References between Regulatory Documents and Company Policies
- Industry Partner Alyne GmbH
- Problem Statement
- Proposed Solution

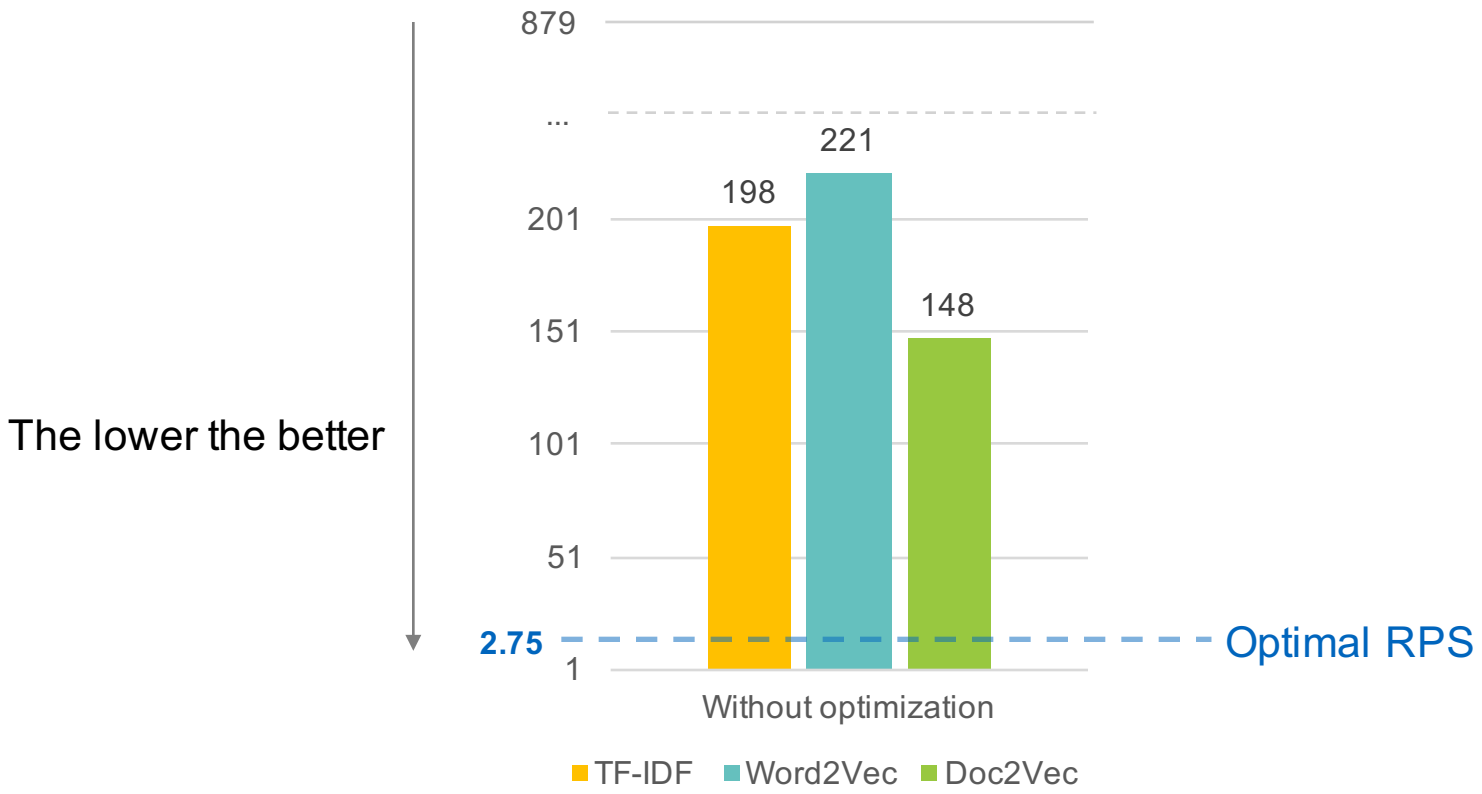
Research Approach
- Research Questions
- Iterative Approach

Implementation
- Recommender System
- Management & Analysis GUI

Evaluations & Results

Evaluations & Results
Initial Evaluation

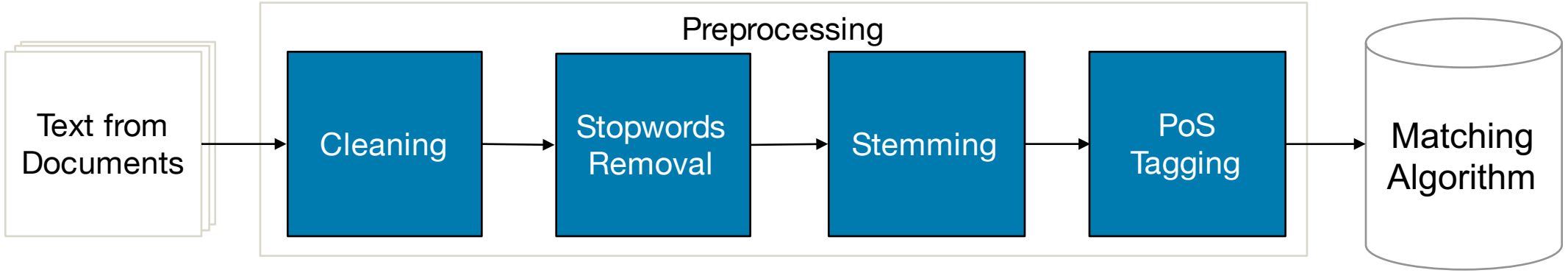
- Quantitative Evaluation without any optimizations
- Optimal Average Rank (RPS) is 2.75 → Allows fully automation, but unrealistic



Evaluations & Results

RQ 3: Impact of Preprocessing Technologies on Matching Results

- Examine of a selection of state-of-the-art preprocessing technologies



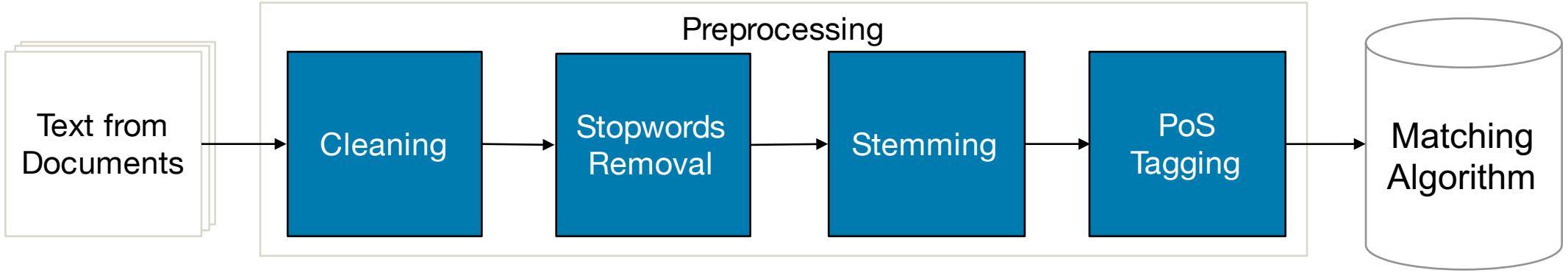
Example

	User passwords shall be prevented from being changed to any of the previous 10 passwords.
Cleaning	user passwords shall be prevented from being changed to any of the previous passwords
Stopwords R.	user passwords prevented changed previous passwords
Stemming	user password prevent chang previous password
PoS Tagging	user password password (only nouns)

Evaluations & Results

RQ 3: Impact of Preprocessing Technologies on Matching Results

- Examine of a selection of state-of-the-art preprocessing technologies



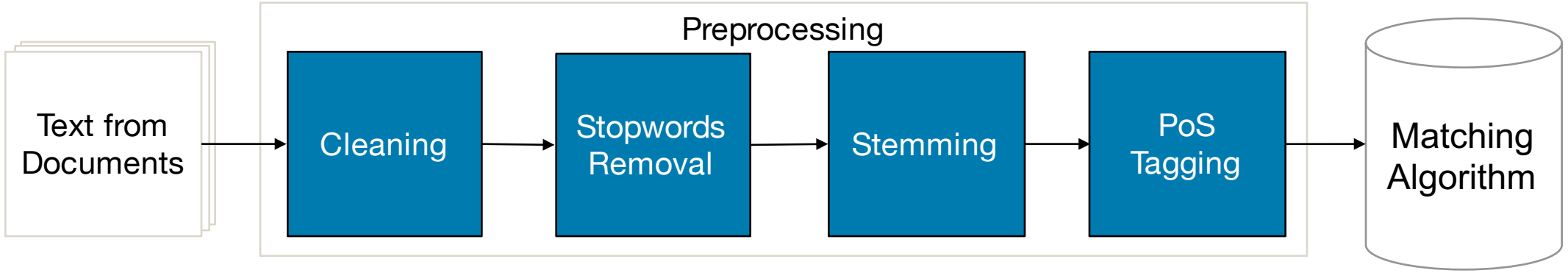
Example

	User passwords shall be prevented from being changed to any of the previous 10 passwords.
Cleaning	**user passwords shall be prevented from being changed to any of the previous passwords**
Stopwords R.	user passwords prevented changed previous passwords
Stemming	user password prevent chang previous password
PoS Tagging	user password password (only nouns)

Evaluations & Results

RQ 3: Impact of Preprocessing Technologies on Matching Results

- Examine of a selection of state-of-the-art preprocessing technologies



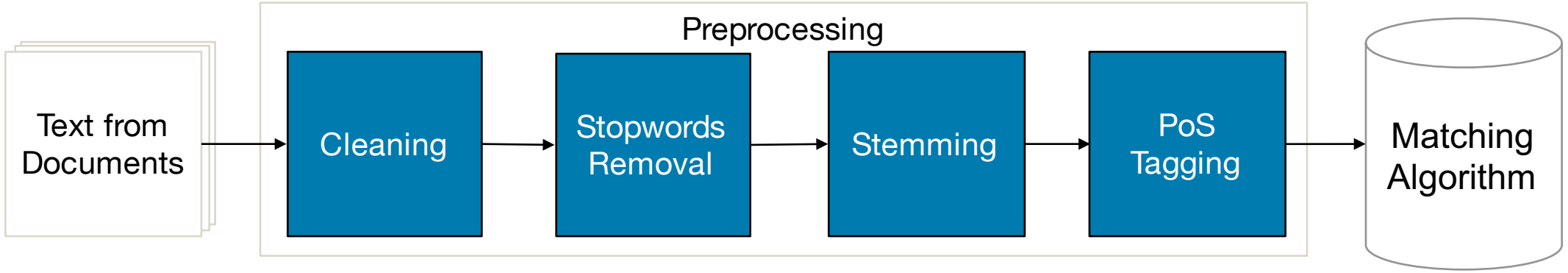
Example

	User passwords shall be prevented from being changed to any of the previous 10 passwords.
Cleaning	user passwords shall be prevented from being changed to any of the previous passwords
Stopwords R.	**user passwords prevented changed previous passwords**
Stemming	user password prevent chang previous password
PoS Tagging	user password password (only nouns)

Evaluations & Results

RQ 3: Impact of Preprocessing Technologies on Matching Results

- Examine of a selection of state-of-the-art preprocessing technologies



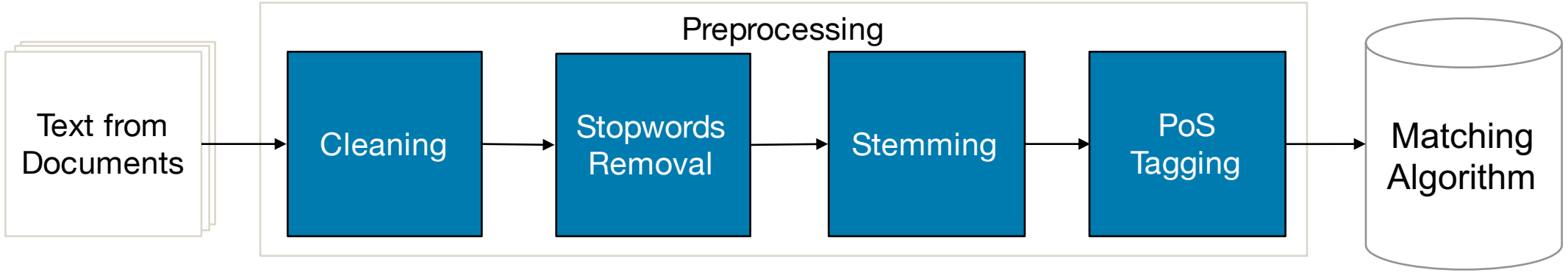
Example

	User passwords shall be prevented from being changed to any of the previous 10 passwords.
Cleaning	user passwords shall be prevented from being changed to any of the previous passwords
Stopwords R.	user passwords prevented changed previous passwords
Stemming	**user password prevent chang previous password**
PoS Tagging	user password password (only nouns)

Evaluations & Results

RQ 3: Impact of Preprocessing Technologies on Matching Results

- Examine of a selection of state-of-the-art preprocessing technologies



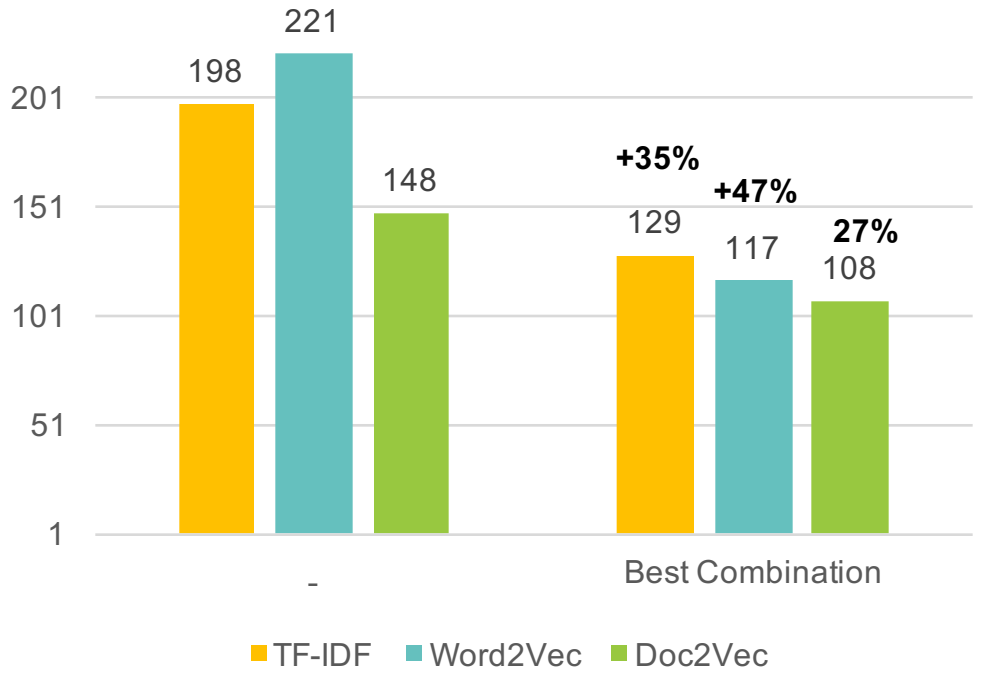
Example

	User passwords shall be prevented from being changed to any of the previous 10 passwords.
Cleaning	user passwords shall be prevented from being changed to any of the previous passwords
Stopwords R.	user passwords prevented changed previous passwords
Stemming	user password prevent chang previous password
PoS Tagging	**user password password** (only nouns)

RQ 3: Impact of Preprocessing Technologies on Matching Results

Results

- Best Combination
  - TF-IDF: Cleaning, stopwords removal, stemming
  - Word2Vec, Doc2Vec: Cleaning, stopwords removal, no stemming
- PoS Tagging had negative impact on results



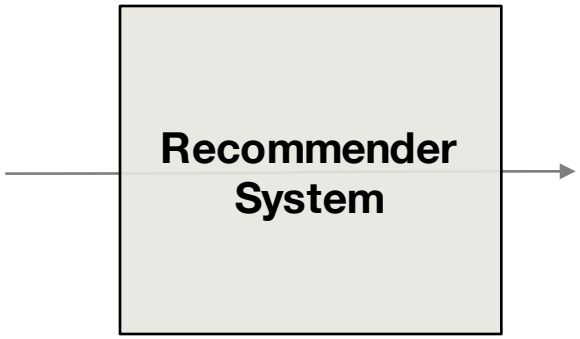
Evaluations & Results

RQ 4: Addition of Meta Information to Control Statements

- Control statements are linked with meta information such as **topic, subtopic, title** and **tags**.
  → Add information to control statements and examine the impact on matching results

Example

Do not allow an individual to submit a new password that is the same as any of the last four passwords he/she has used.



1. User passwords shall be prevented from being changed to any of the previous 10 passwords.
   Password History, Access Management, ...
2. Privileged account passwords shall be prevented from being changed to any of the previous 15 passwords.
   Password History, Access Management, ...

⋮

879. Recovery from backup media shall be tested at least every year. Data Lifecycle

Evaluations & Results

RQ 4: Addition of Meta Information to Control Statements

- Control statements are linked with meta information such as **topic, subtopic, title** and **tags**.
  → Add information to control statements and examine the impact on matching results

Example

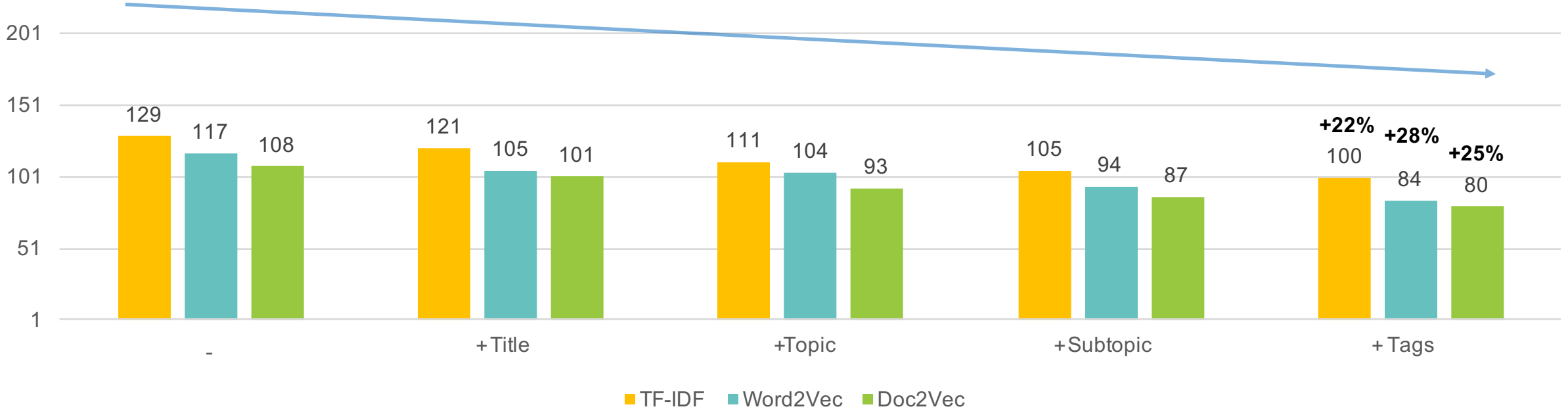
```
{
   "id":          "Control_Statement_Id_00033",
   "topic":       "Password Management",
   "subTopic":    "Password History",
   "title":       "User Password History Length",
   "statement":  "User passwords shall be prevented from being
      changed to any of the previous 10 passwords."
   "tags":        [ "Password History", "Access Management", ... ]
}
```

Evaluations & Results

RQ 4: Addition of Meta Information to Control Statements

Results

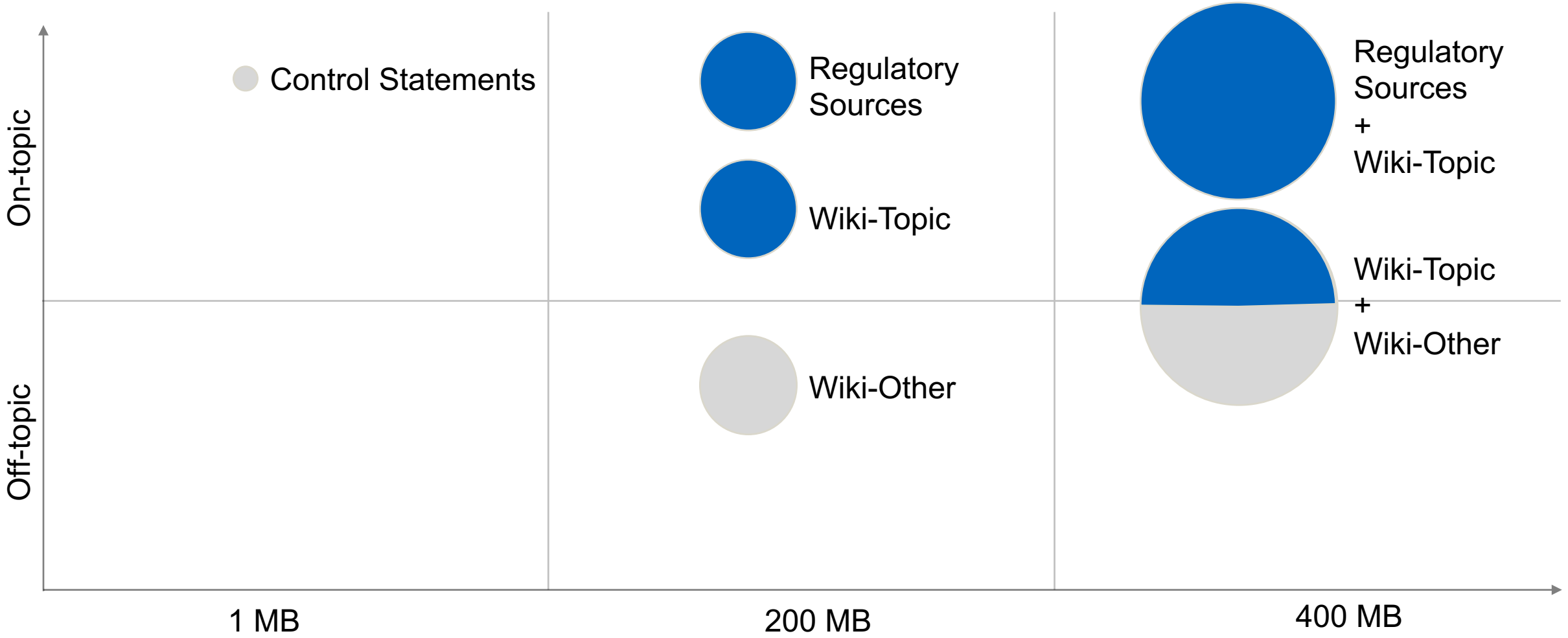
- “The more meta information, the better the results”
- A deterioration could not be stated with available meta information
  - General information that are not sufficiently disjunct to other control statements might worsen the RPS



Evaluations & Results

RQ 5: Corpora Analysis

- We investigate several corpora that differ in characteristics regarding content type (off-topic/on-topic) and size (small/big) to conclude the influence on matching results.

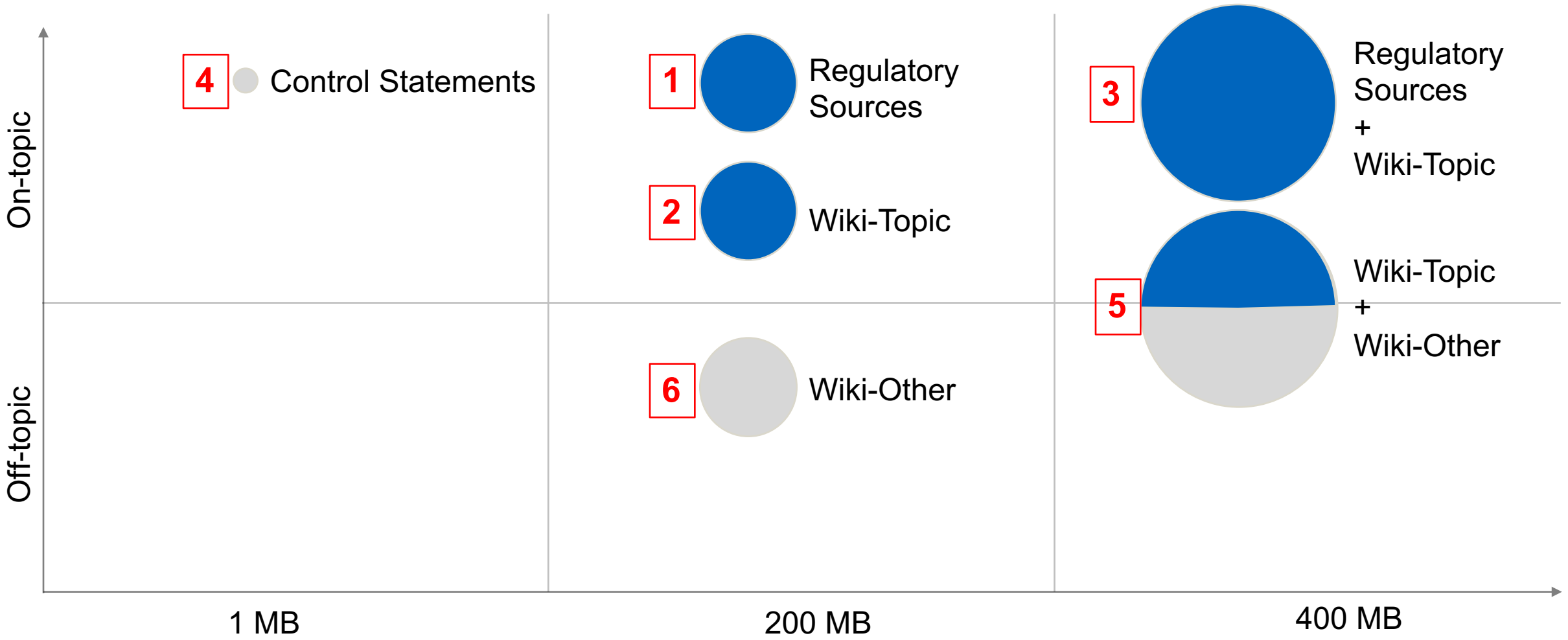


Evaluations & Results

RQ 5: Corpora Analysis

- We investigate several corpora that differ in characteristics regarding content type (off-topic/on-topic) and size (small/big) to conclude the influence on matching results.

1 good - 6 bad

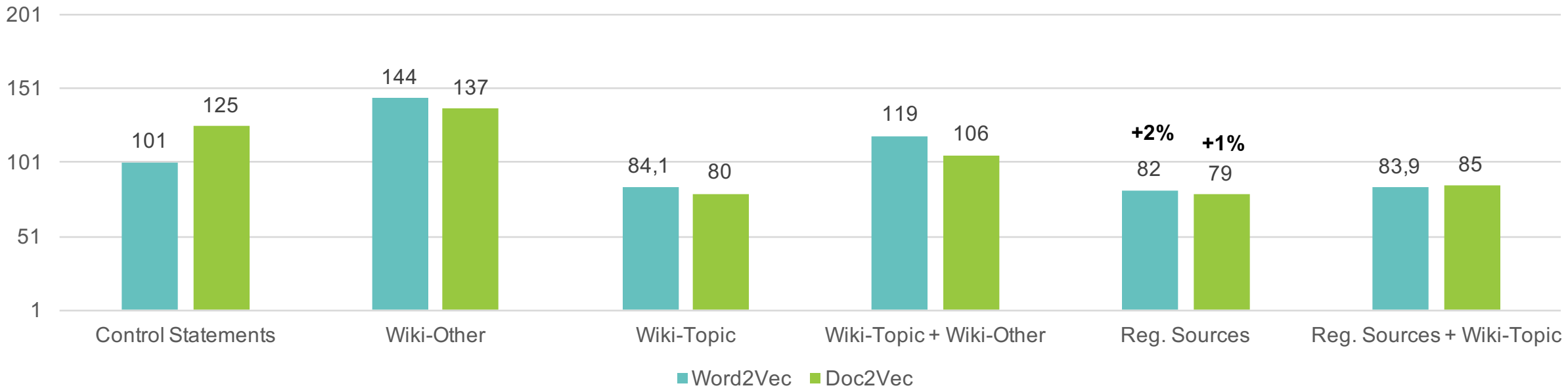


Evaluations & Results
RQ 5: Corpora Analysis

Results

- ~~On-topic corpus + off-topic corpus → slight improvements~~
  On-topic corpus + off-topic corpus → substantial deterioration

- Combination of on-topic corpus does not necessarily lead to better results

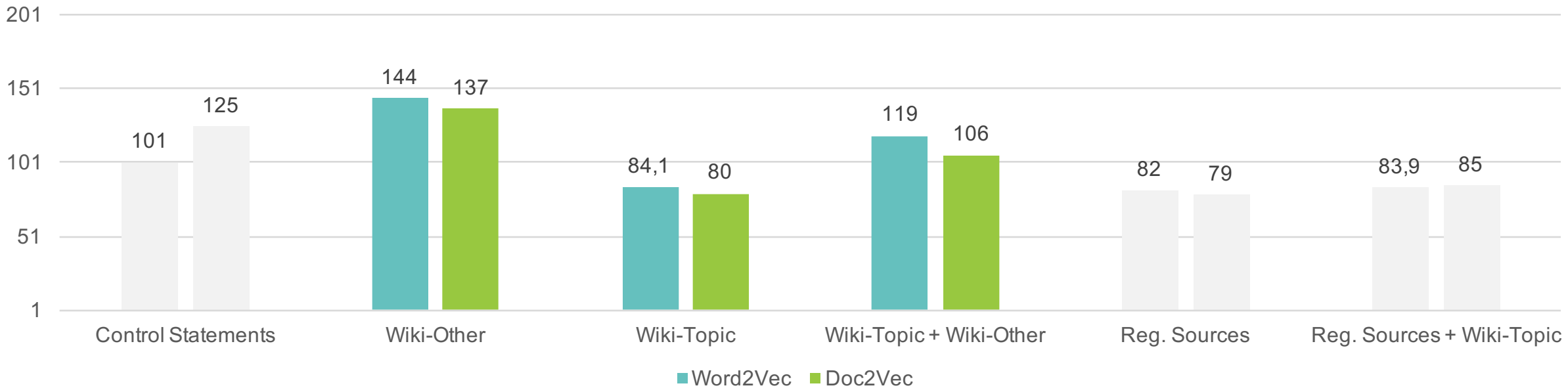


Evaluations & Results
RQ 5: Corpora Analysis

Results

- ~~**On-topic corpus + off-topic corpus → slight improvements**~~
  On-topic corpus + off-topic corpus → substantial deterioration

- Combination of on-topic corpus does not necessarily lead to better results

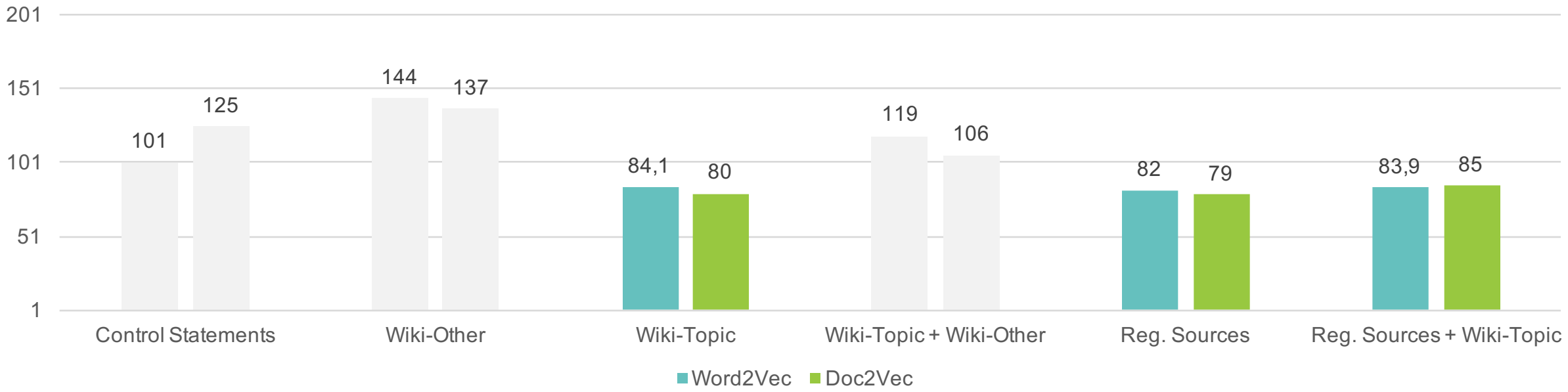


Evaluations & Results
RQ 5: Corpora Analysis

Results

- ~~On-topic corpus + off-topic corpus → slight improvements~~
  On-topic corpus + off-topic corpus → substantial deterioration

- **Combination of on-topic corpus does not necessarily lead to better results**



Evaluations & Results

RQ 6: Addition of Context Information to Input Controls

- Paragraph context of input controls, like section titles, often contains key words that might support the matching process.
  → Extract context information and examine its impact

Evaluation

Do not allow an individual to submit a new password that is the same as any of the last four passwords he/she has used.

Implement strong access control measures

1. User passwords shall be prevented from being changed to any of the previous 10 passwords.
   Password History, Access Management, ...
2. Privileged account passwords shall be prevented from being changed to any of the previous 15 passwords.
   Password History, Access Management, ...

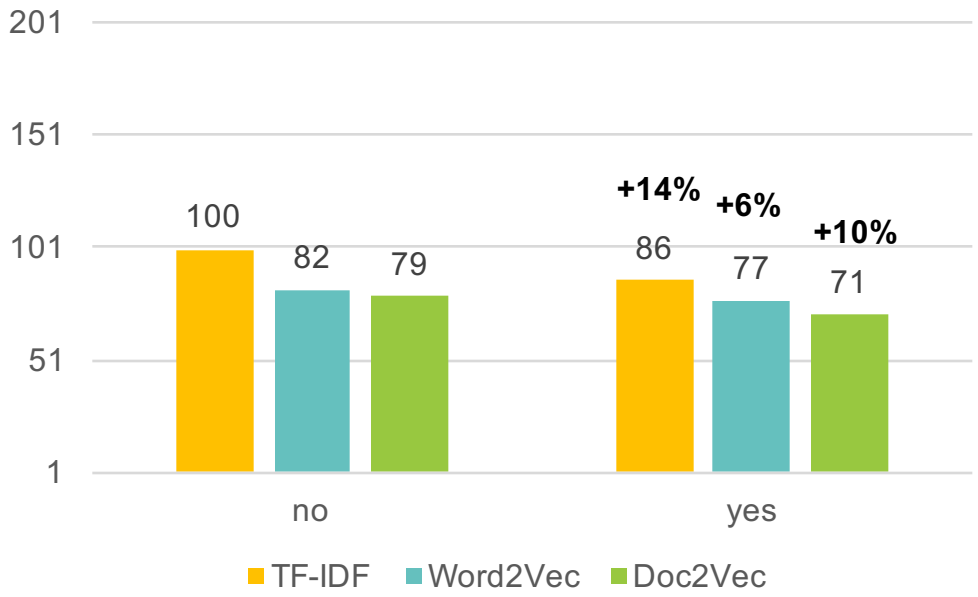
   ⋮

879. Recovery from backup media shall be tested at least every year. Data Lifecycle

RQ 6: Addition of Context Information to Input Controls

Results

- Improvements for all approaches

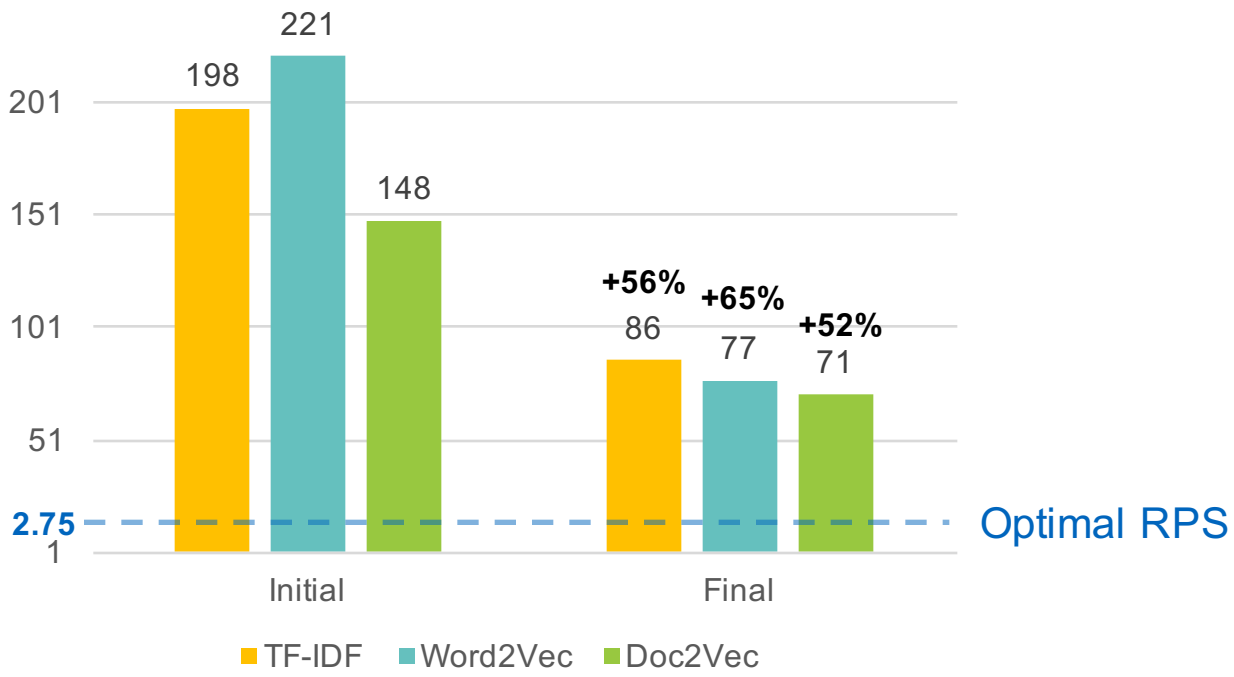


Evaluations & Results

RQ 2: What text similarity approach performs best?

Results

- Doc2Vec performed best with an average rank of 71
- All in all, good and balanced performance of all approaches



Evaluations & Results

RQ 1: To which degree does the text similarity approach solve the semantic text matching problem?

- Expert interview with 4 experts using feedback sheet
  - Inspection of matching results of best recommender system (Doc2Vec, average rank 71)
  - Focus on the 209 worst matchings (outliers) to see future improvement potential

- **Questions**
  - **For each of the 209 items: Are there reoccurring patterns?**
  - **All in all, is the recommender system ready for productive use?**

Evaluations & Results

RQ 1: To which degree does the text similarity approach solve the semantic text matching problem?

For each of the 209 matchings results: Are there reoccurring patterns?

- Abbreviations
  → Use the document's list of abbreviation, if available, or any online service

- Short input
  → Extract more context information of input controls
  - Special Case: Enumeration

- Missing synonyms
  → Enrich meta information of control statements

- Imbalance in depth of content between input control and control statement (general > specific)
  → Match Input with Controls that are already linked with control statements

Evaluations & Results

RQ 1: To which degree does the text similarity approach solve the semantic text matching problem?

Levels of maturity:

1. Expert works (almost) exclusively with recommender system
   Effect: Significant time savings
2. Expert works with recommender system (50%) but also traditional method (50%)
   Effect: Medium time savings
3. Expert uses recommender system only as cross-check and reassurance
   Effect: Improved data quality
4. No productive applicability yet

Evaluations & Results

RQ 1: To which degree does the text similarity approach solve the semantic text matching problem?

Levels of maturity:

1. Expert works (almost) exclusively with recommender system
   Effect: Significant time savings
2. Expert works with recommender system (50%) but also traditional method (50%)
   Effect: Medium time savings
3. Expert uses recommender system only as cross-check and reassurance
   Effect: Improved data quality
4. No productive applicability yet

Evaluations & Results
Summary

- Preprocessing
  - Balanced the performance of the text similarity approaches
  - PoS Tagging has a negative impact on results
  - Stemming great effect on TF-IDF, but can have negative effect on Word2Vec and Doc2Vec

- A larger corpus does not necessarily lead to better results
  → Corpus quality has a decisive influence on the matching quality

- Extending control statements or input controls by additional information (meta or context information) leads to great improvements, especially for shorter controls or statements

- Doc2Vec performs best, Word2Vec and TF-IDF also with good results

- **Already ready for productive use**:
  - cross-check and reassurance → improve data quality
    (already identified a number of cases where the manual initial mapping had errors)
  - But high number of outliers

B. Sc.

Christoph Erl

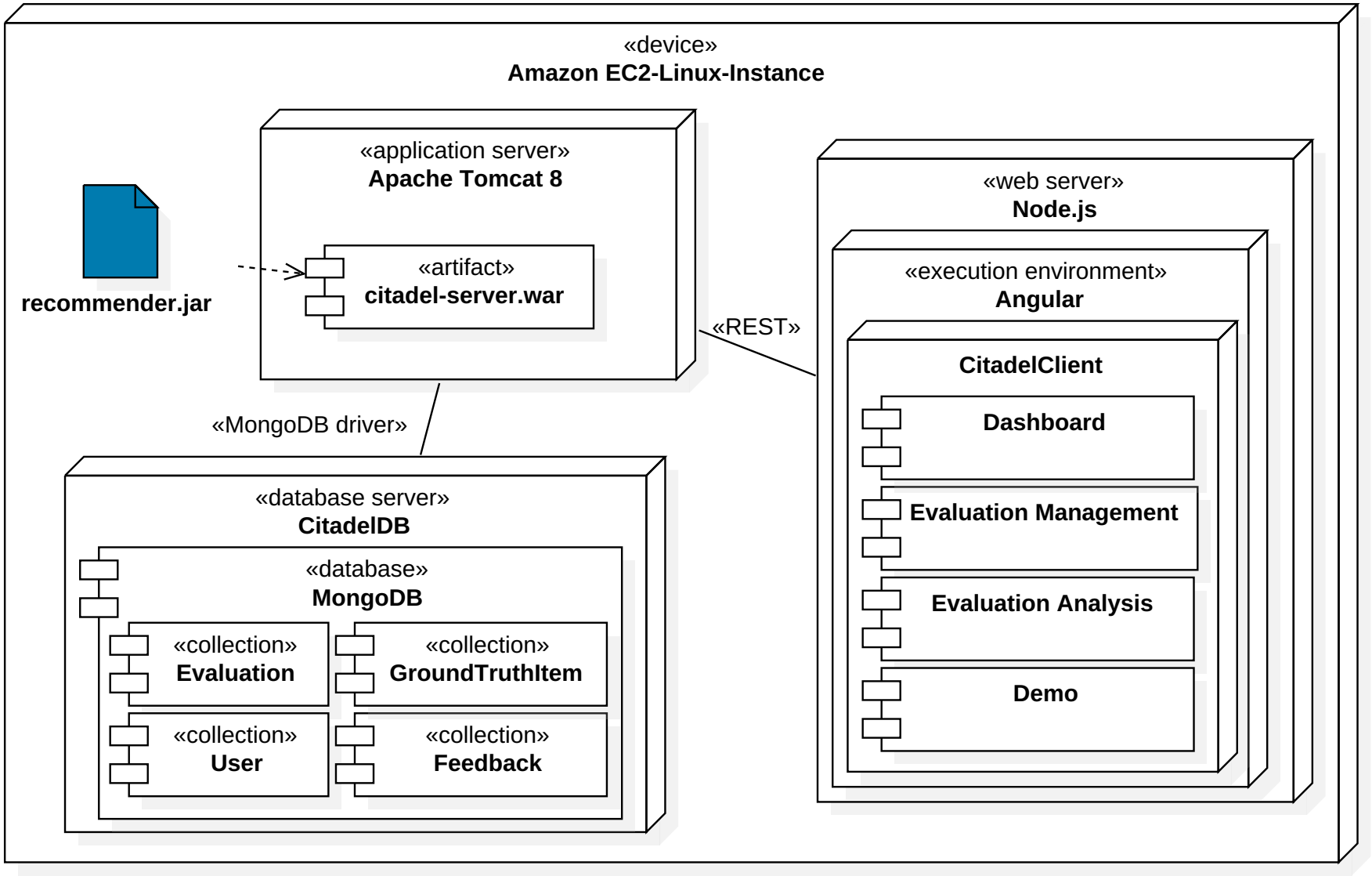
Technische Universität München
Faculty of Informatics
Chair of Software Engineering for Business Information Systems

Boltzmannstraße 3
85748 Garching bei München

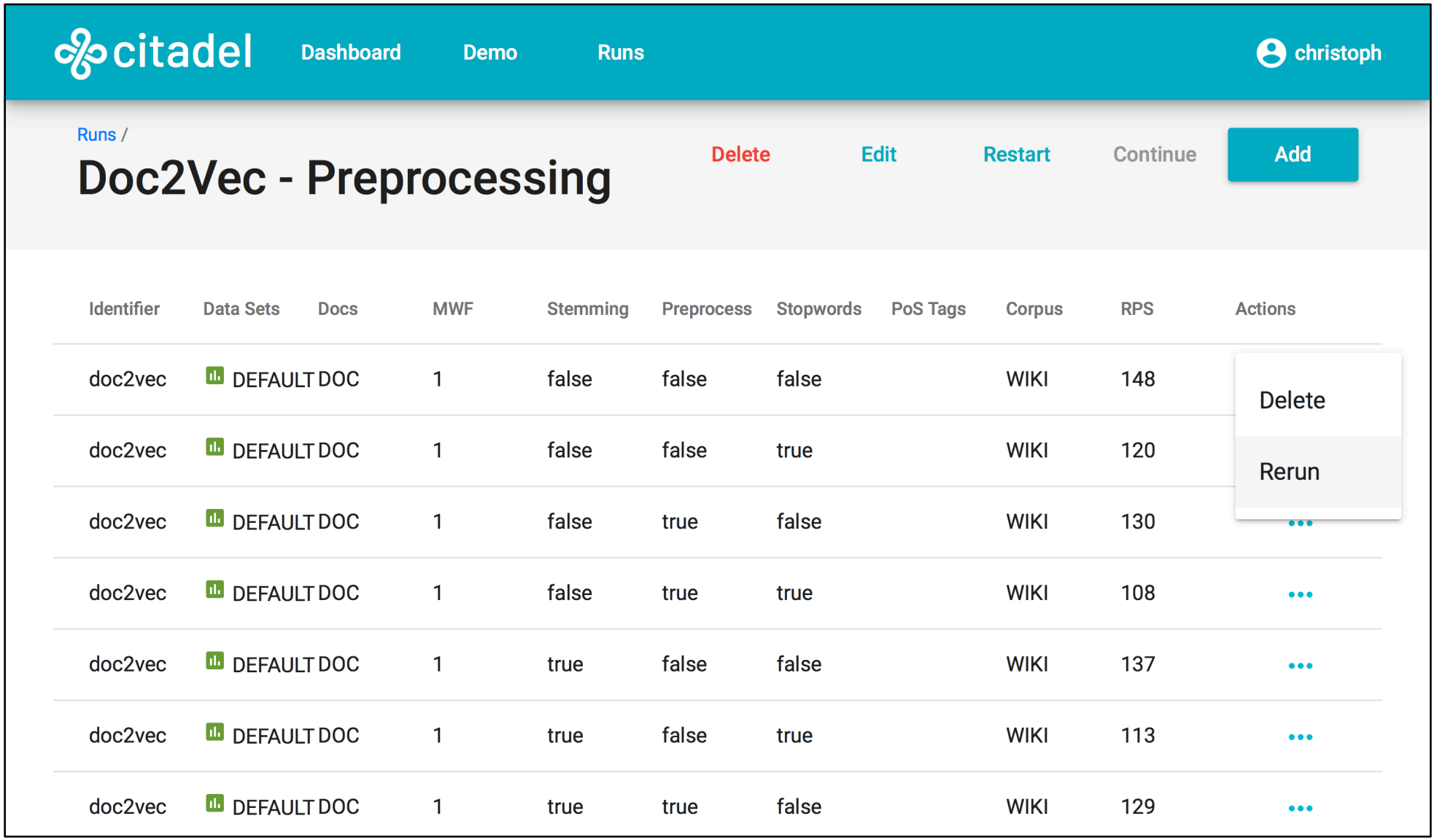
Christoph.erl@tum.de
wwwmatthes.in.tum.de

BACKUP SLIDES

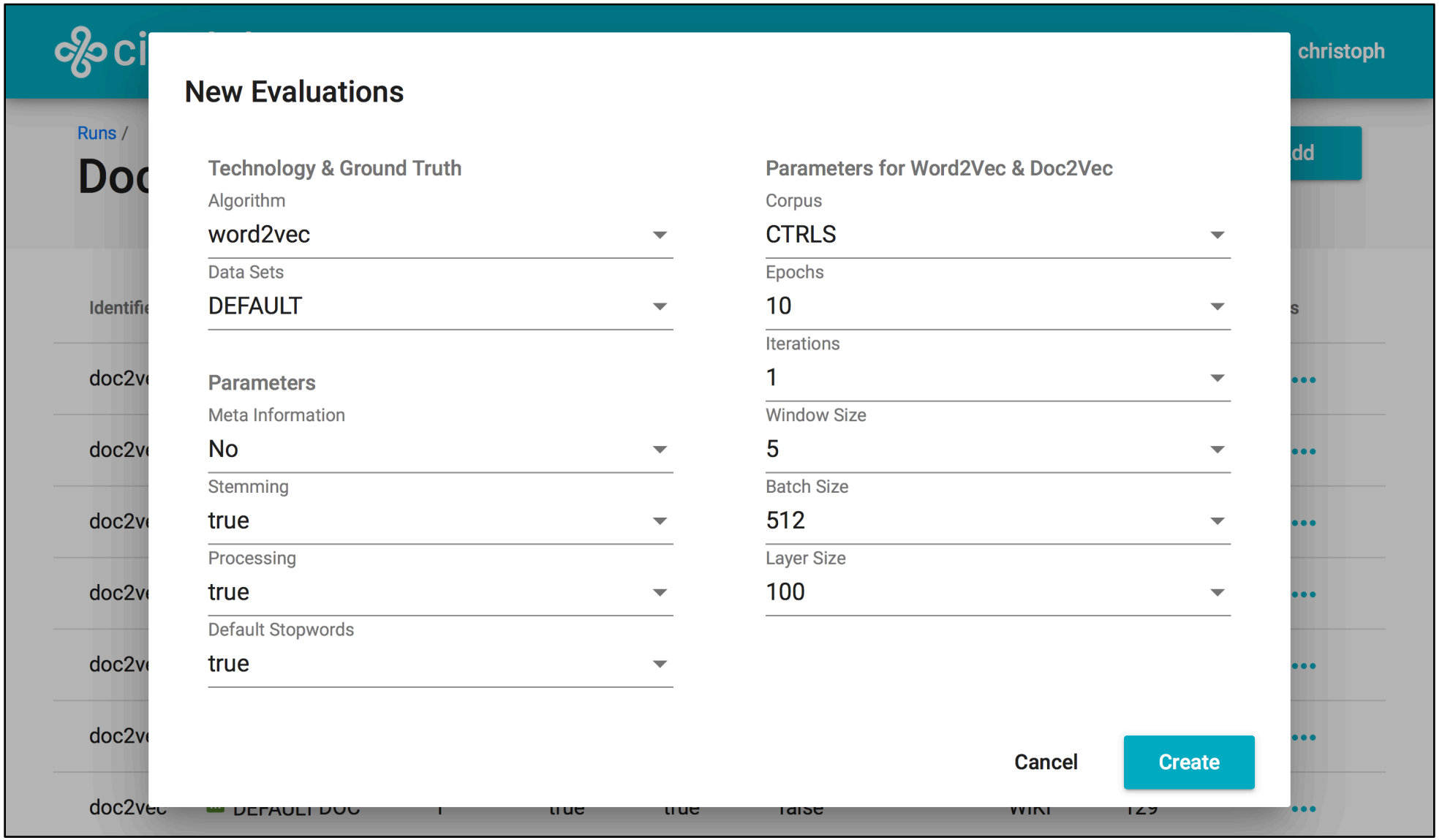
GUI: Deployment Diagram



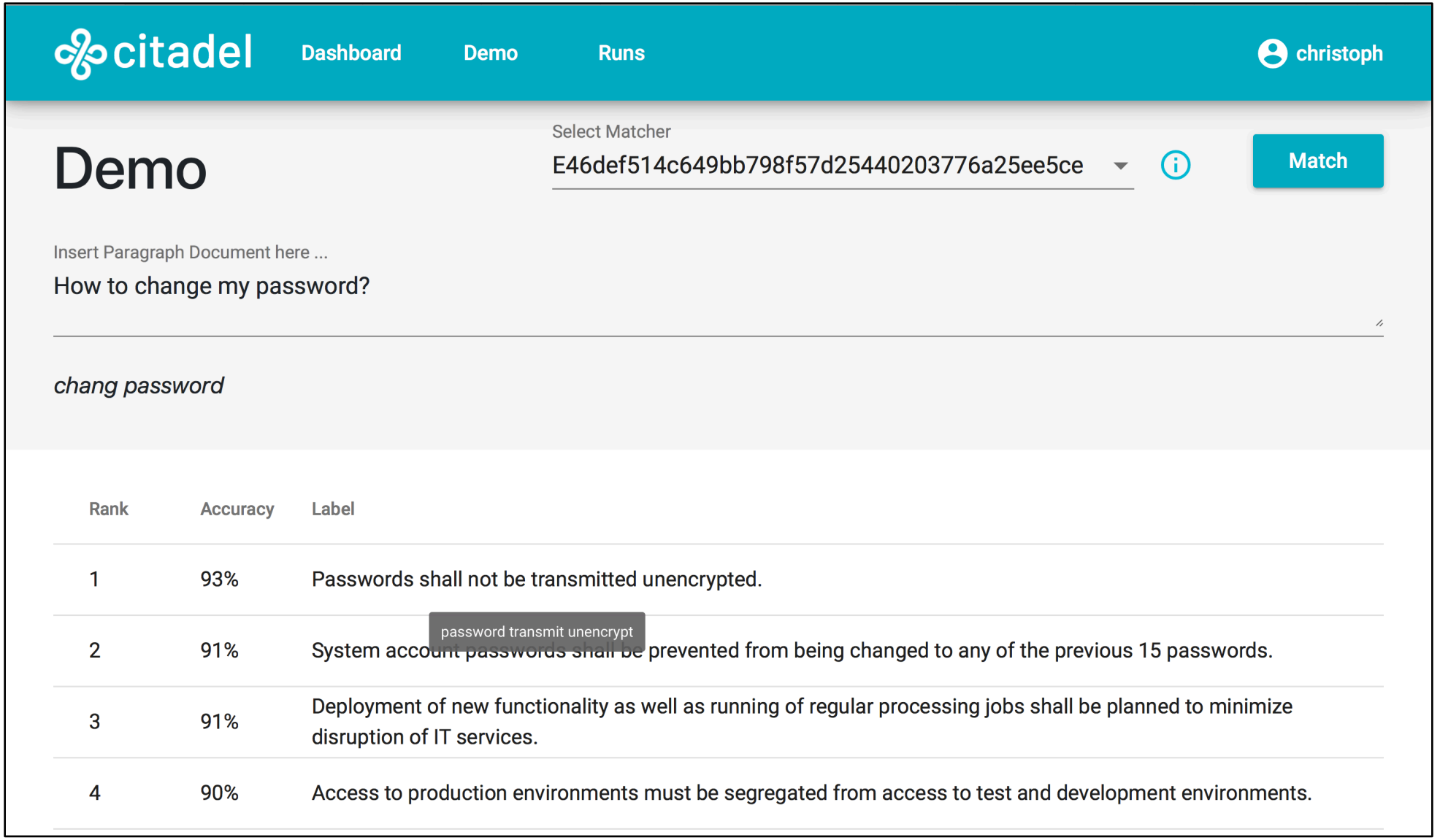
GUI: Screenshot – Evaluations



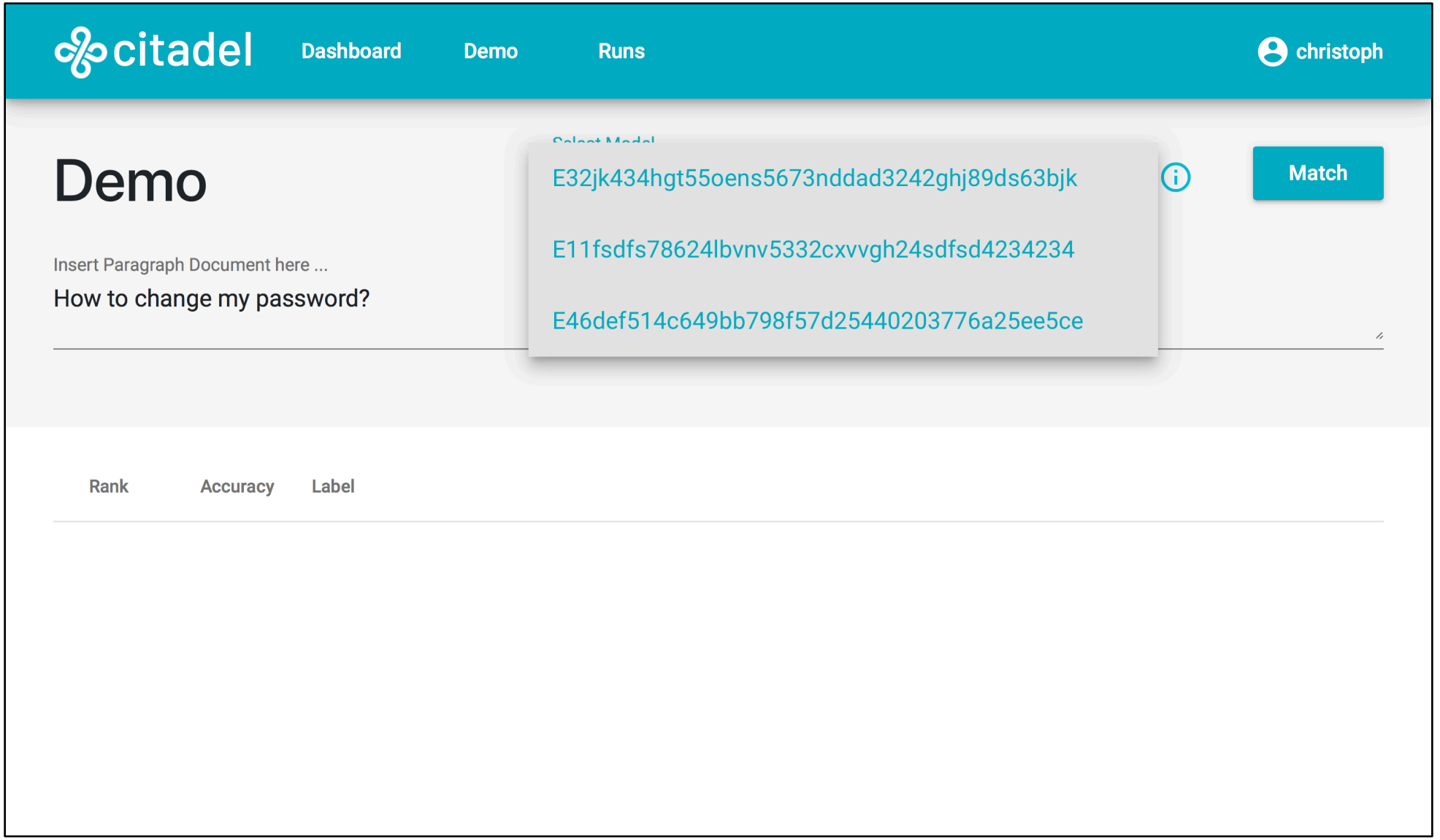
GUI: Screenshot – Create Evaluation



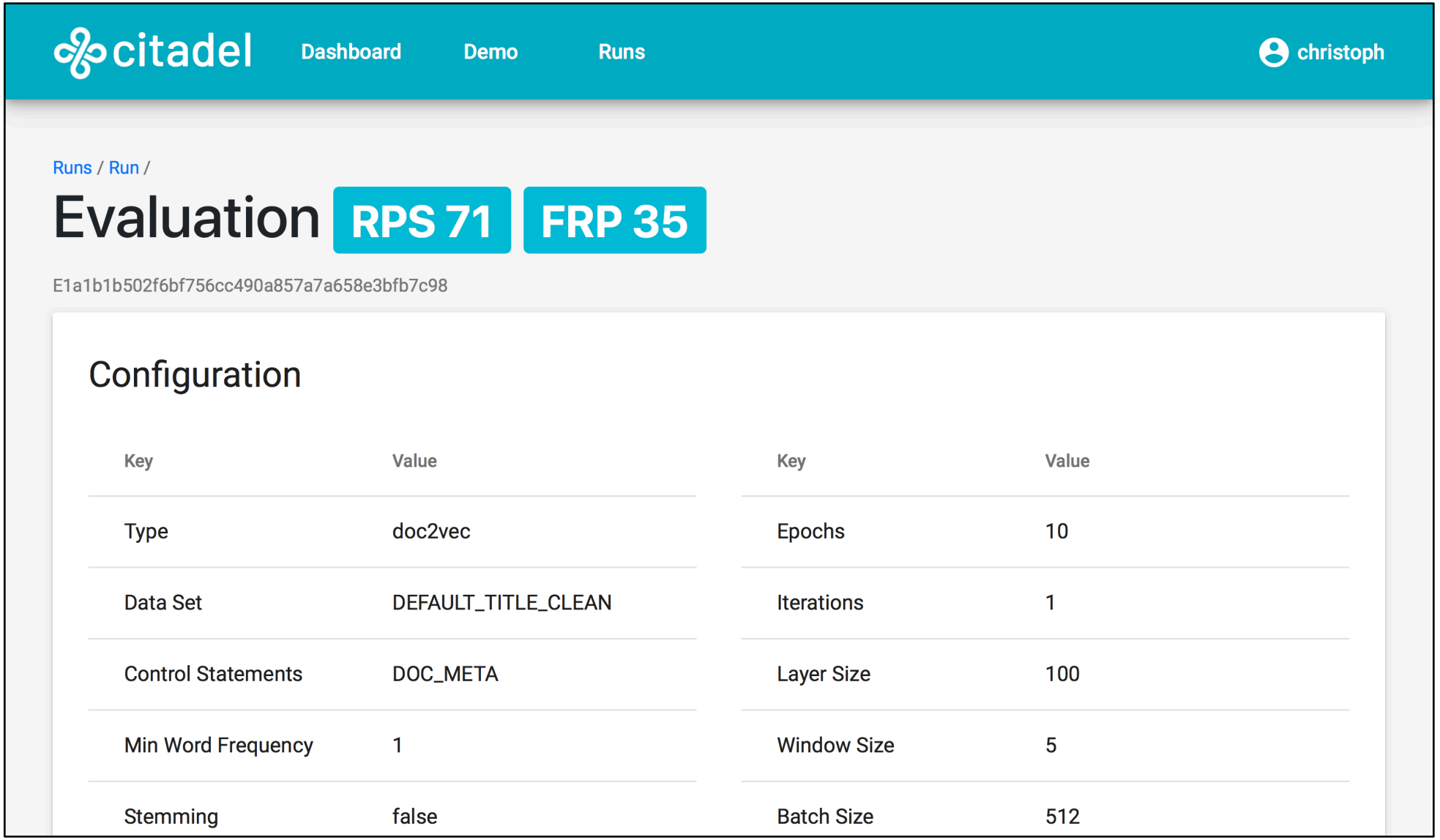
GUI: Screenshot – Demo



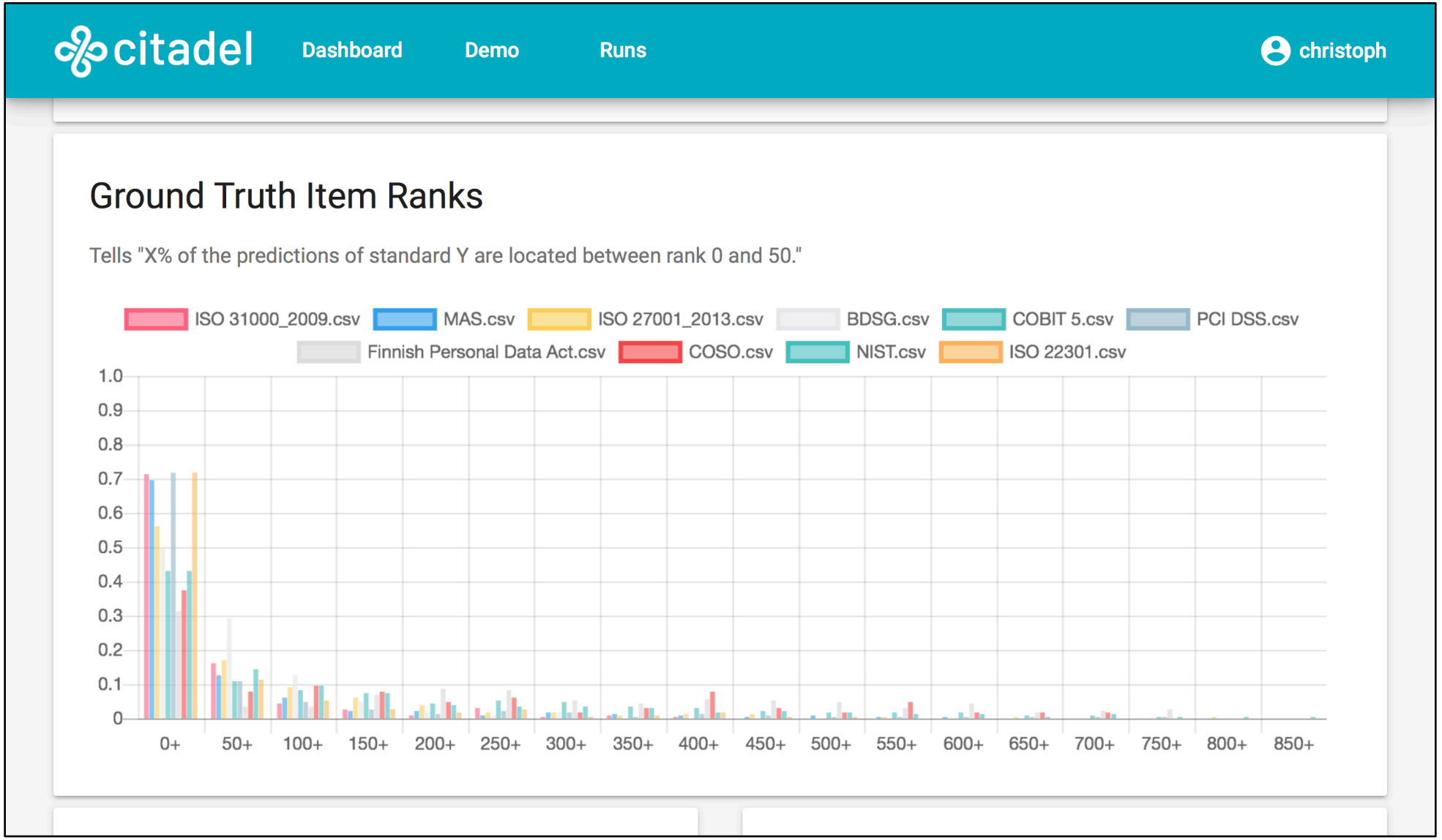
GUI: Screenshot – Demo – Select Matcher



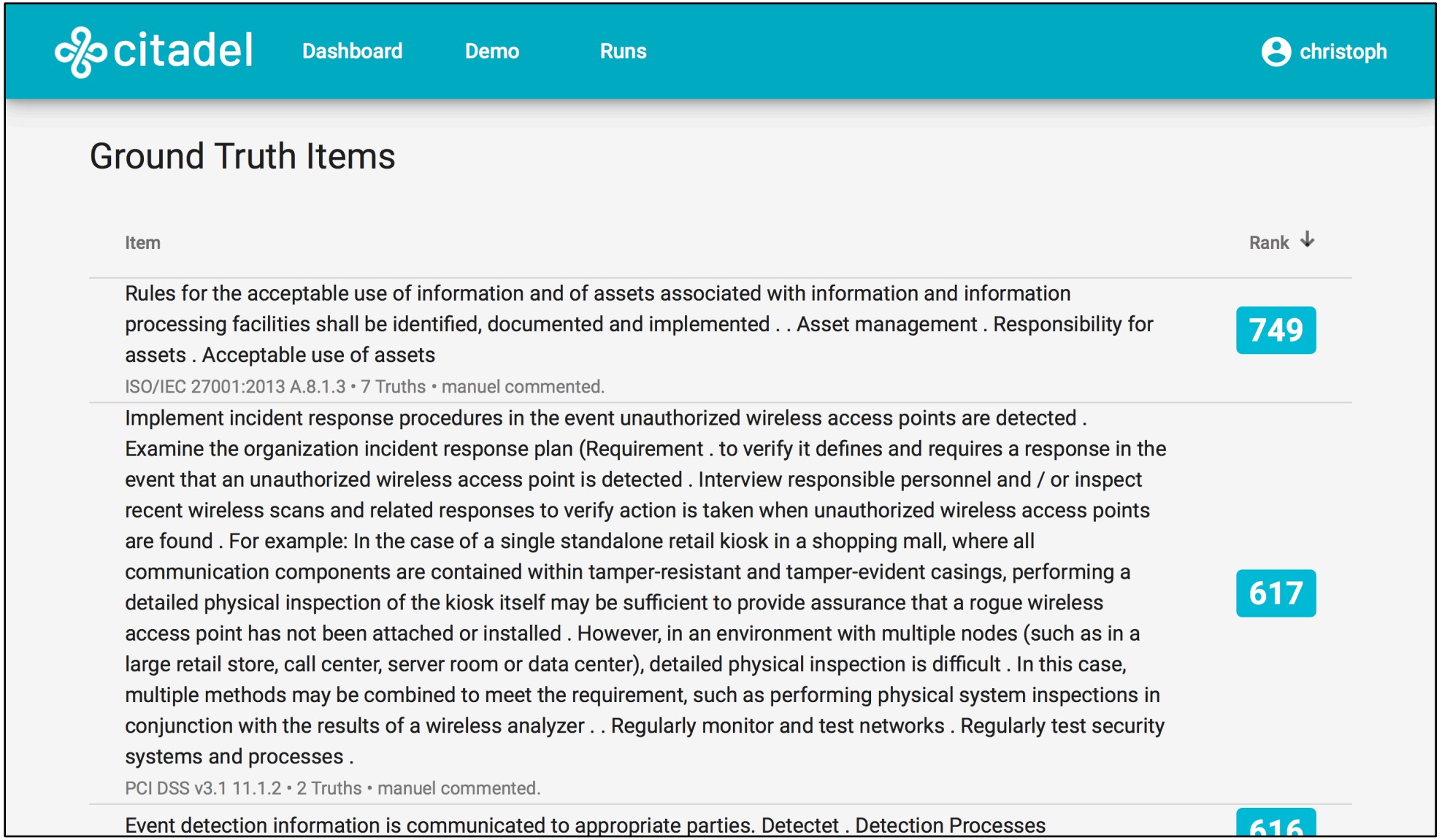
GUI: Screenshot – Evaluation Detail



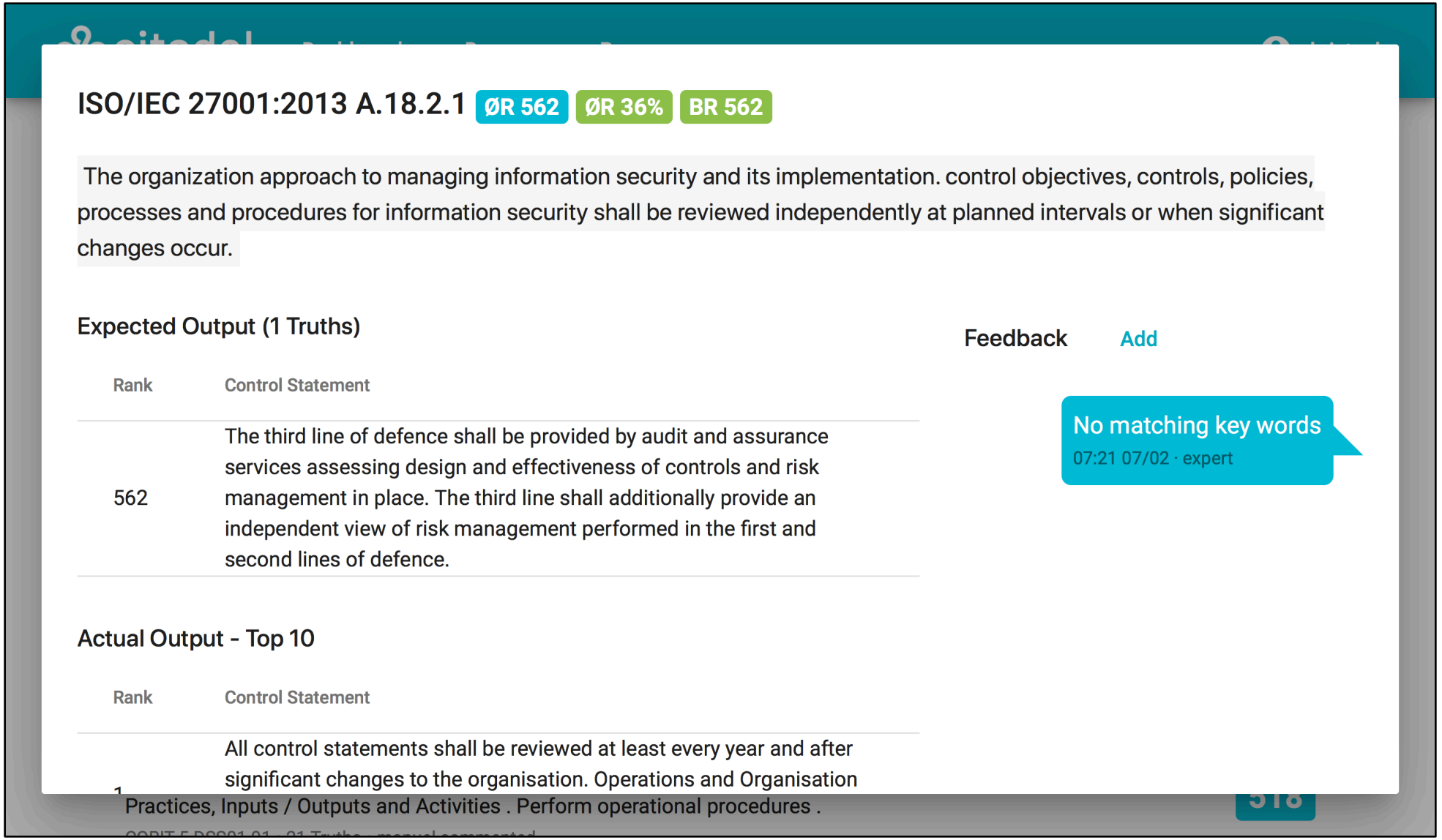
GUI: Screenshot – Evaluation Detail – Charts



GUI: Screenshot – Evaluation Detail – Matching Results



GUI: Screenshot – Evaluation Detail – Matching Result



ISO/IEC 27001:2013 A.18.2.1 ØR 562 ØR 36% BR 562

The organization approach to managing information security and its implementation. control objectives, controls, policies, processes and procedures for information security shall be reviewed independently at planned intervals or when significant changes occur.

Expected Output (1 Truths)

Rank	Control Statement
562	The third line of defence shall be provided by audit and assurance services assessing design and effectiveness of controls and risk management in place. The third line shall additionally provide an independent view of risk management performed in the first and second lines of defence.

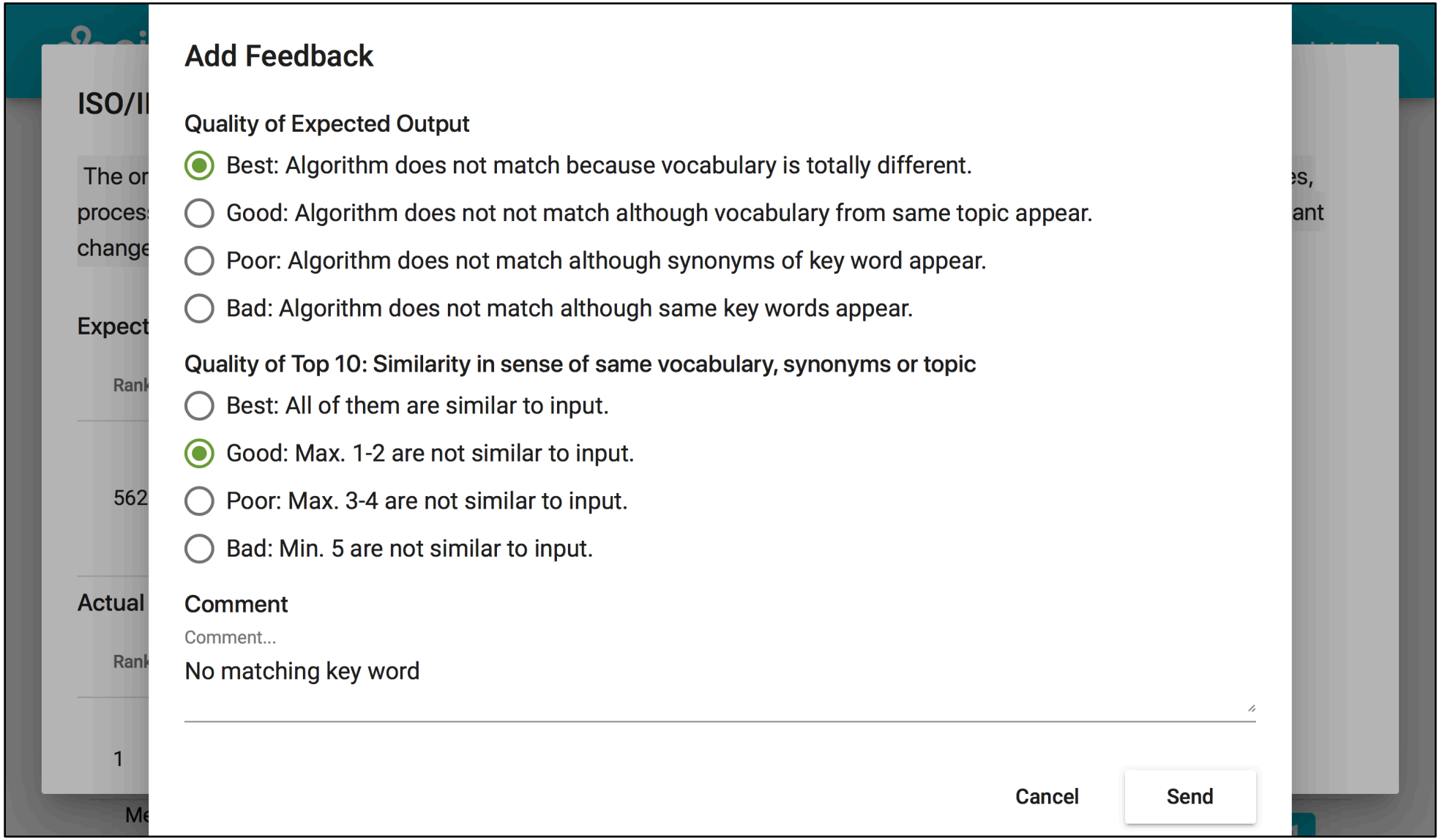
Actual Output - Top 10

Rank	Control Statement
1	All control statements shall be reviewed at least every year and after significant changes to the organisation. Operations and Organisation Practices, Inputs / Outputs and Activities . Perform operational procedures .

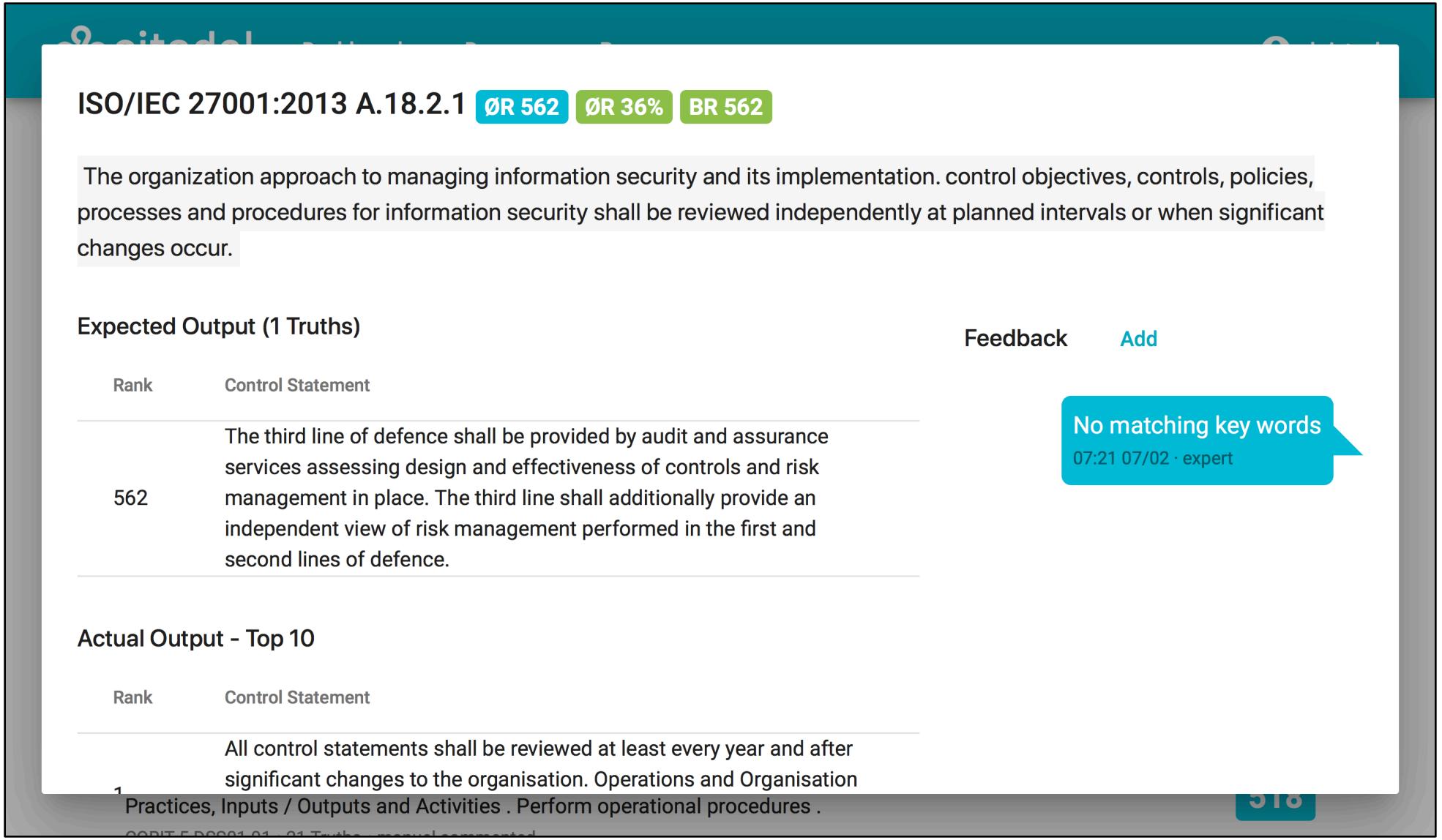
Feedback Add

No matching key words
07:21 07/02 · expert

GUI: Screenshot – Evaluation Detail – Matching Result Feedback



GUI: Screenshot – Evaluation Detail – Matching Result



ISO/IEC 27001:2013 A.18.2.1 ØR 562 ØR 36% BR 562

The organization approach to managing information security and its implementation. control objectives, controls, policies, processes and procedures for information security shall be reviewed independently at planned intervals or when significant changes occur.

Expected Output (1 Truths)

Rank	Control Statement
562	The third line of defence shall be provided by audit and assurance services assessing design and effectiveness of controls and risk management in place. The third line shall additionally provide an independent view of risk management performed in the first and second lines of defence.

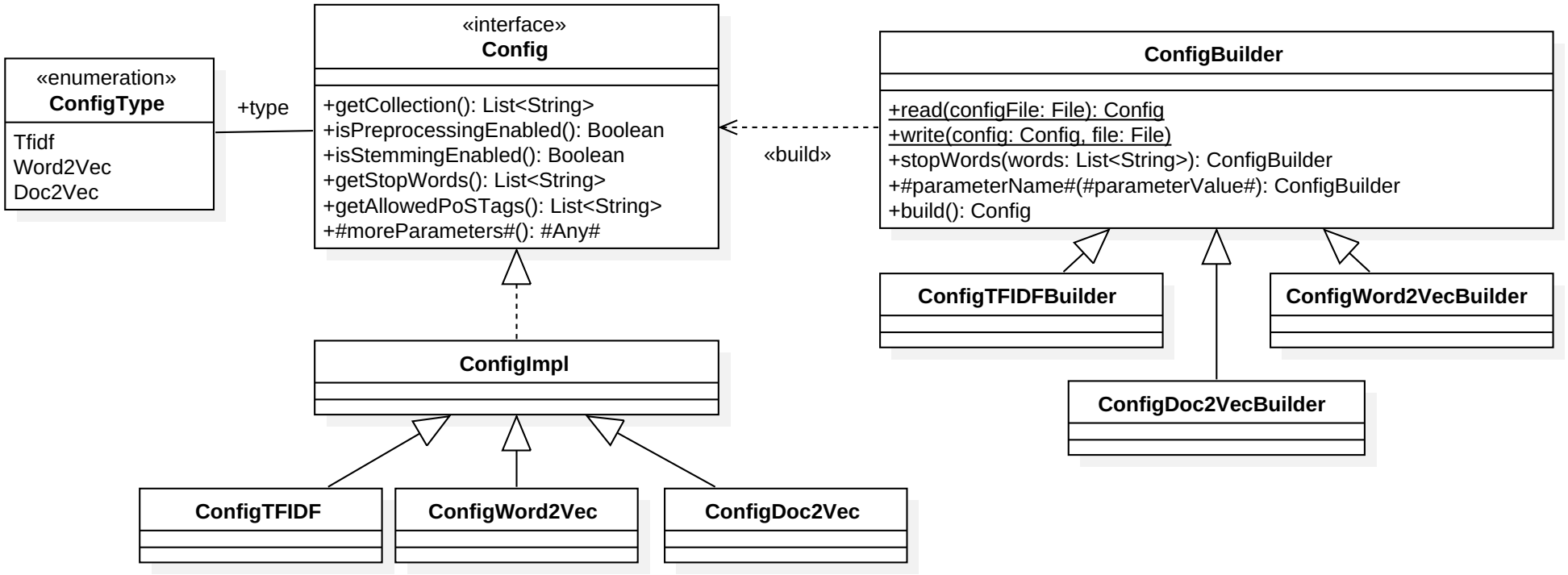
Actual Output - Top 10

Rank	Control Statement
1	All control statements shall be reviewed at least every year and after significant changes to the organisation. Operations and Organisation Practices, Inputs / Outputs and Activities . Perform operational procedures .

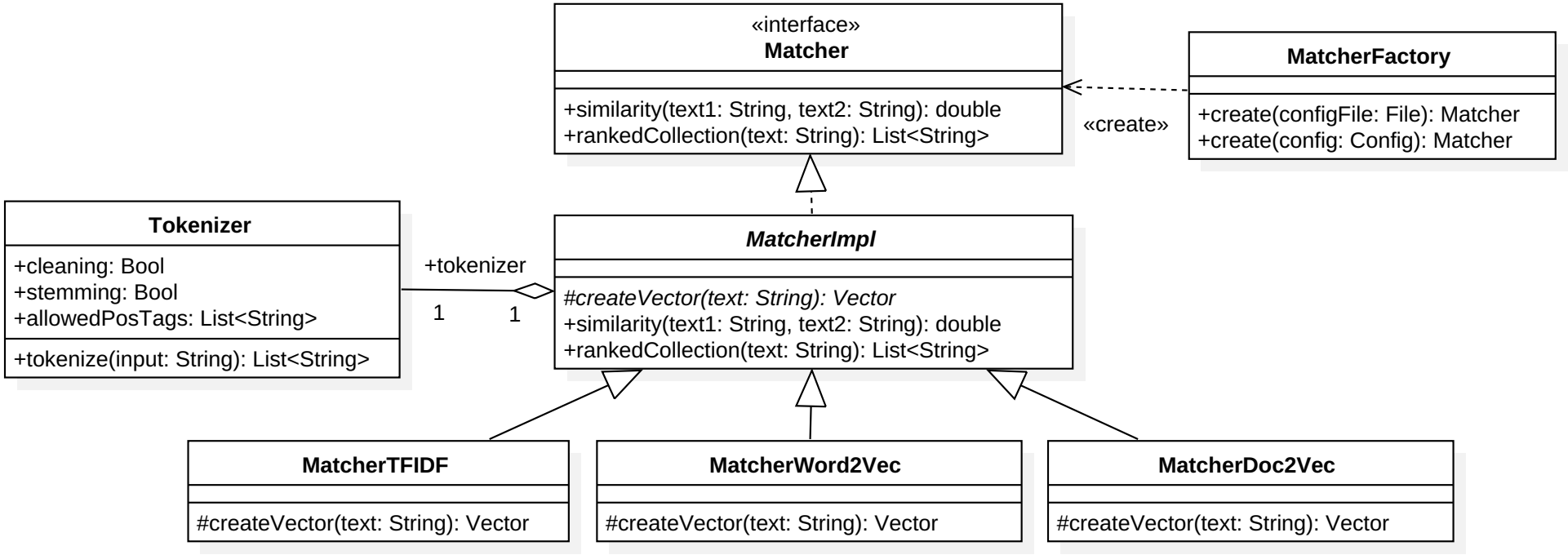
Feedback Add

No matching key words
07:21 07/02 · expert

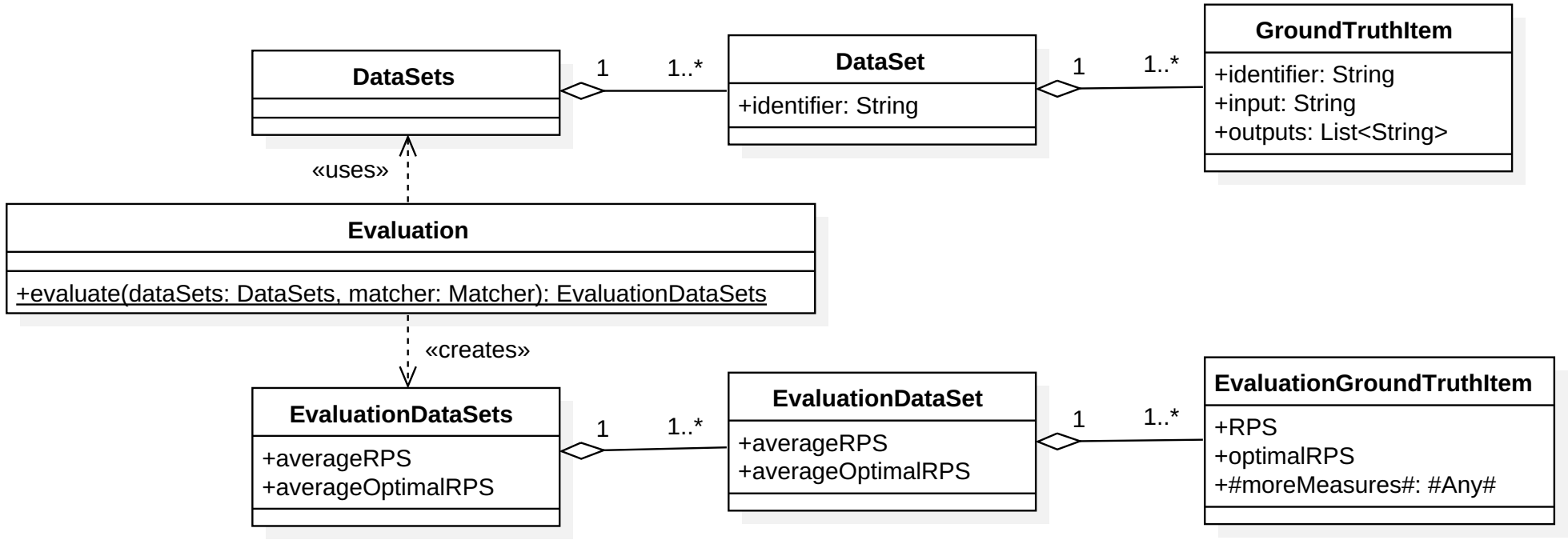
Recommender System: Configuration



Recommender System: Matcher

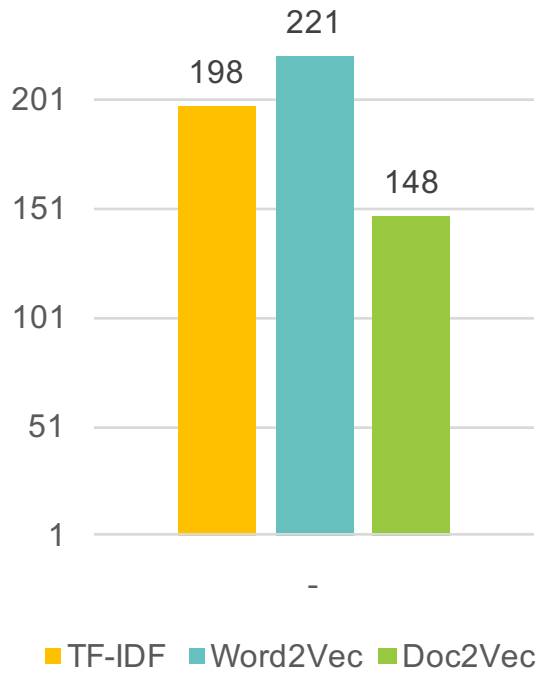


Recommender System: Evaluation



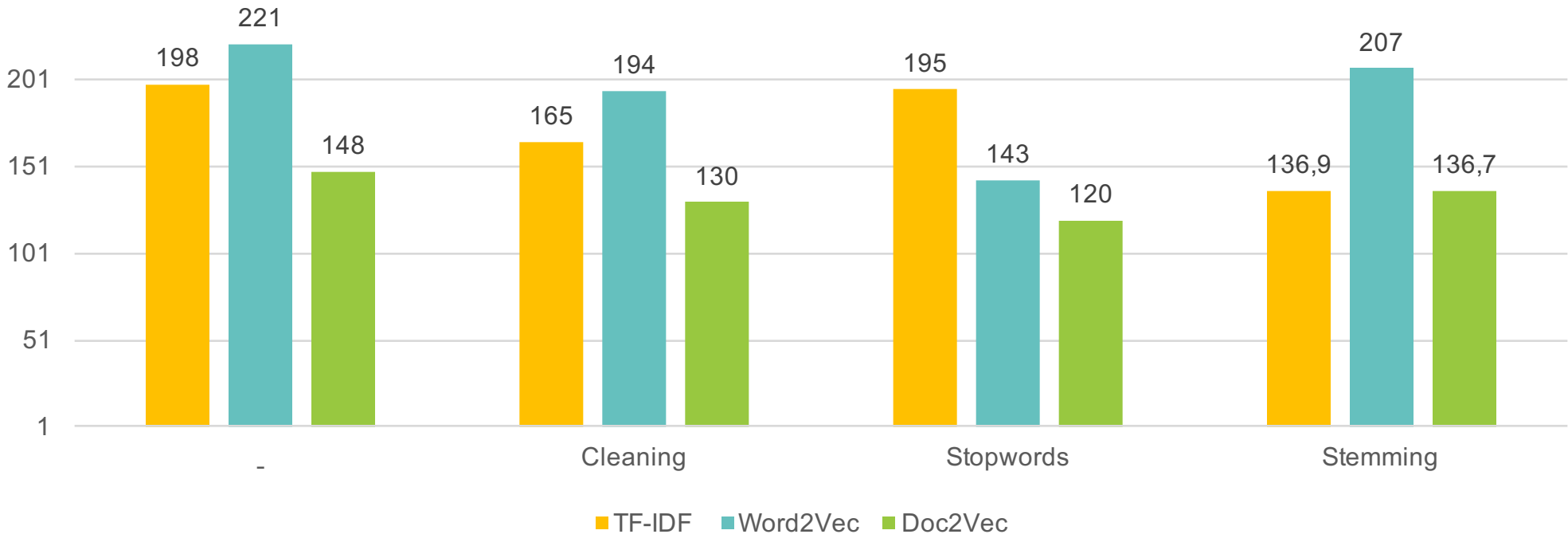
Evaluation & Results

Initial Evaluation



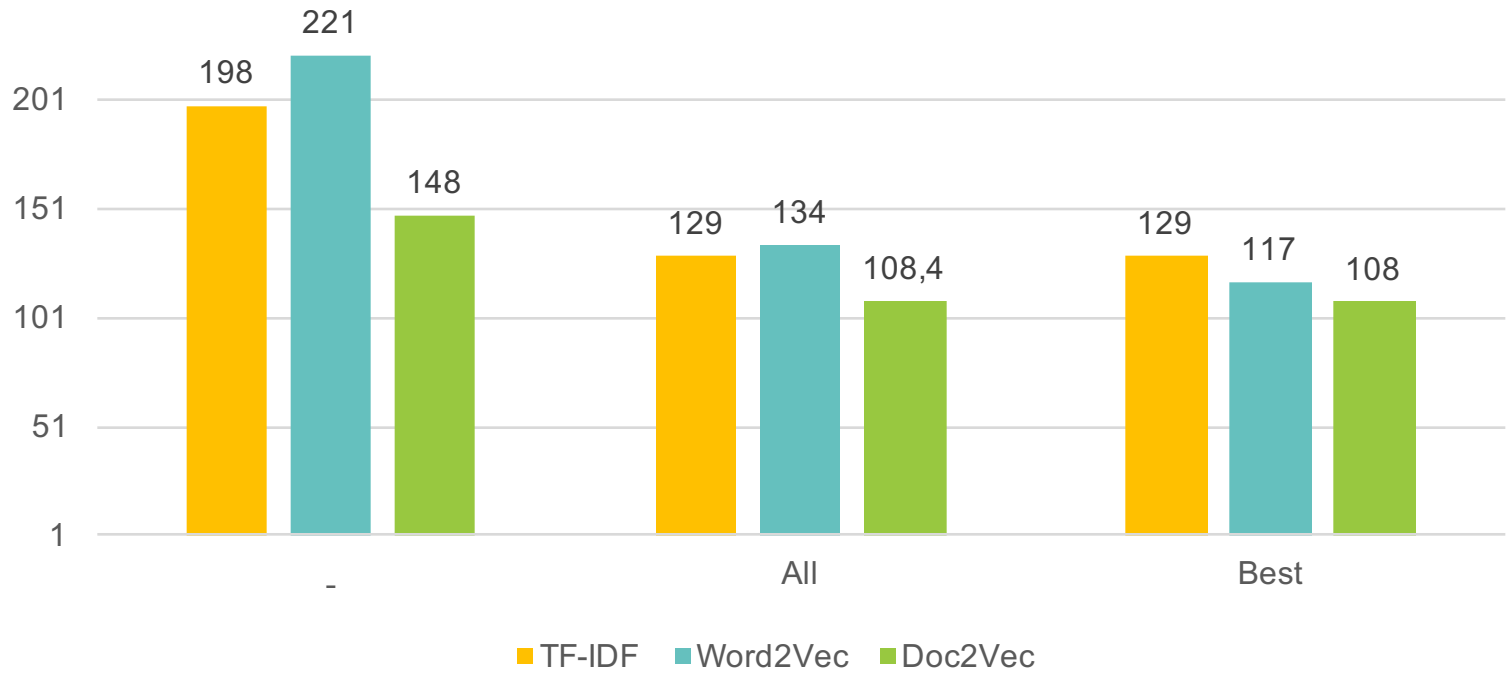
Evaluation & Results

Preprocessing – Separate Application



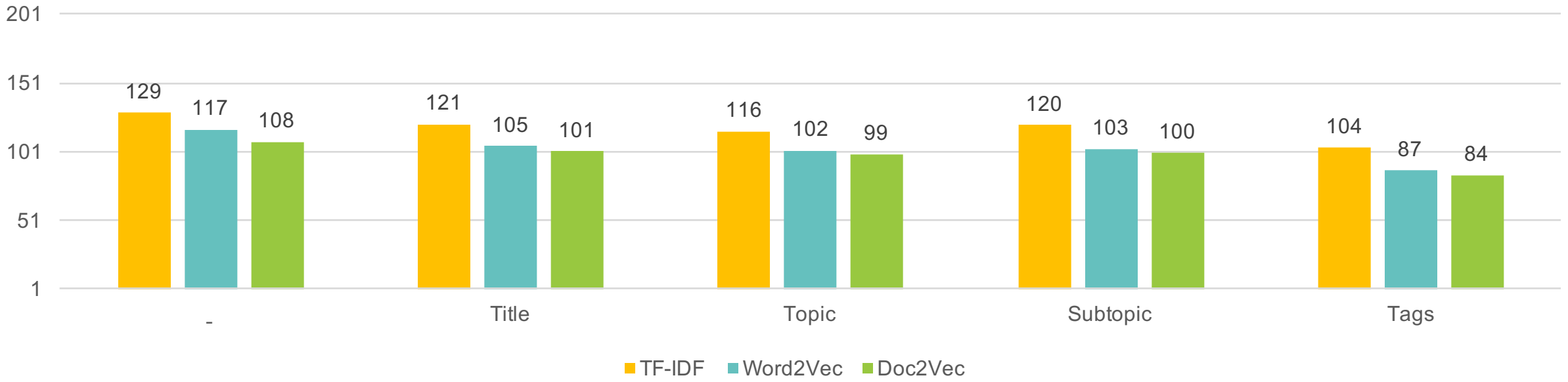
Evaluation & Results

Preprocessing – Combined Application



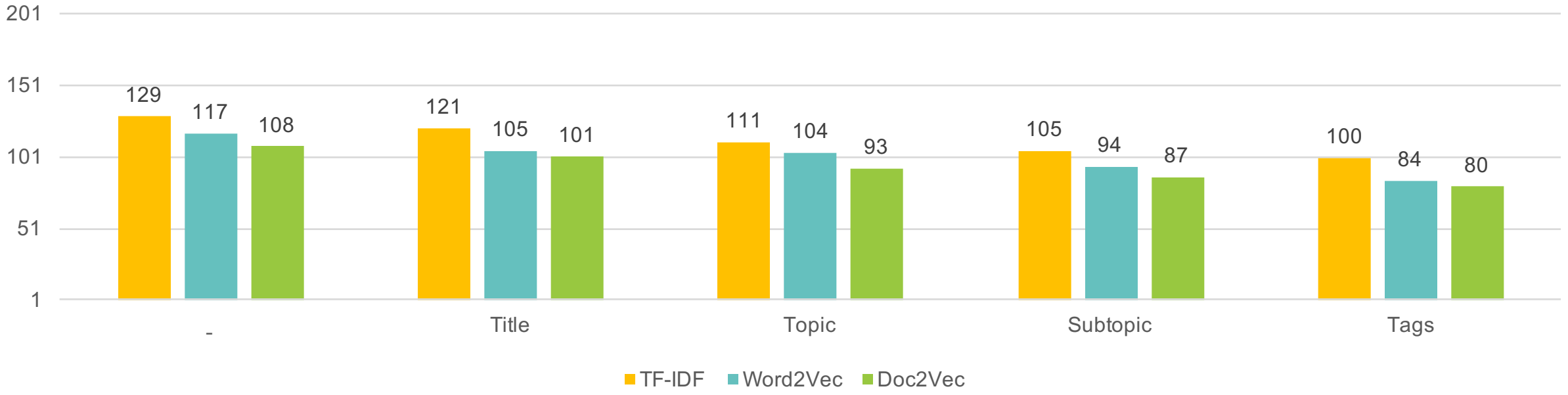
Evaluation & Results

Addition of Meta Information – Separate Addition



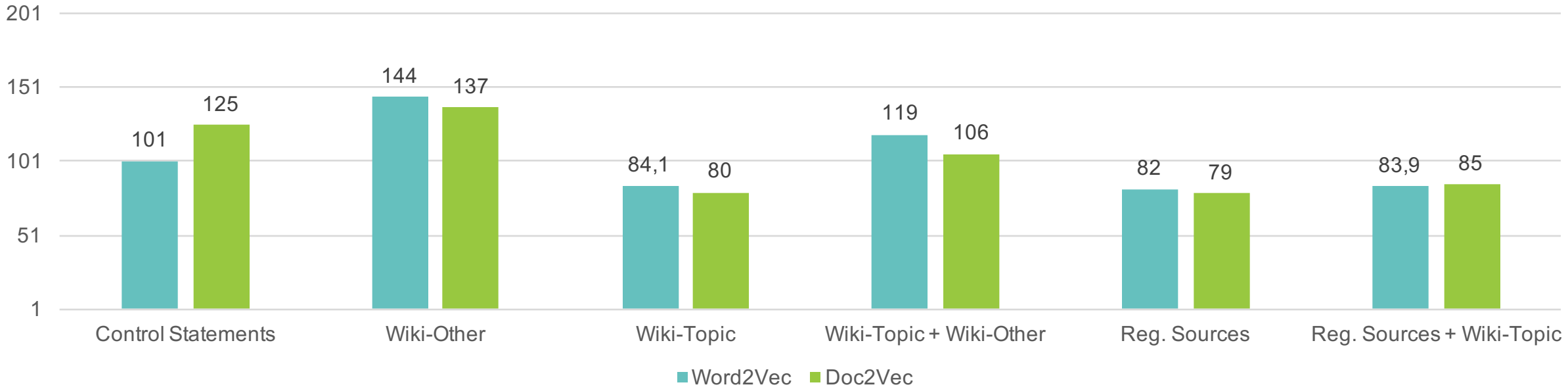
Evaluation & Results

Addition of Meta Information – Combined Addition



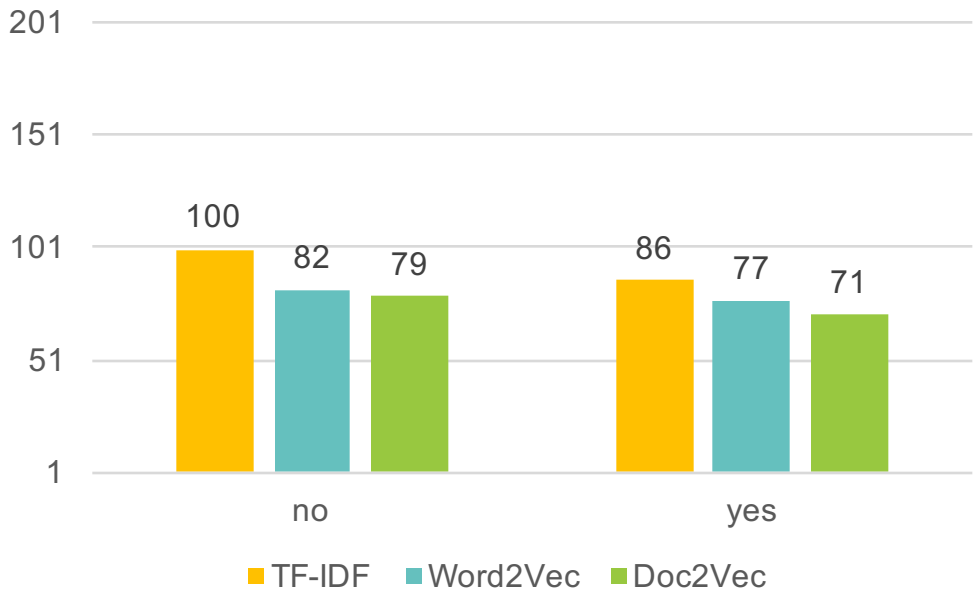
Evaluation & Results

Corpora Analysis

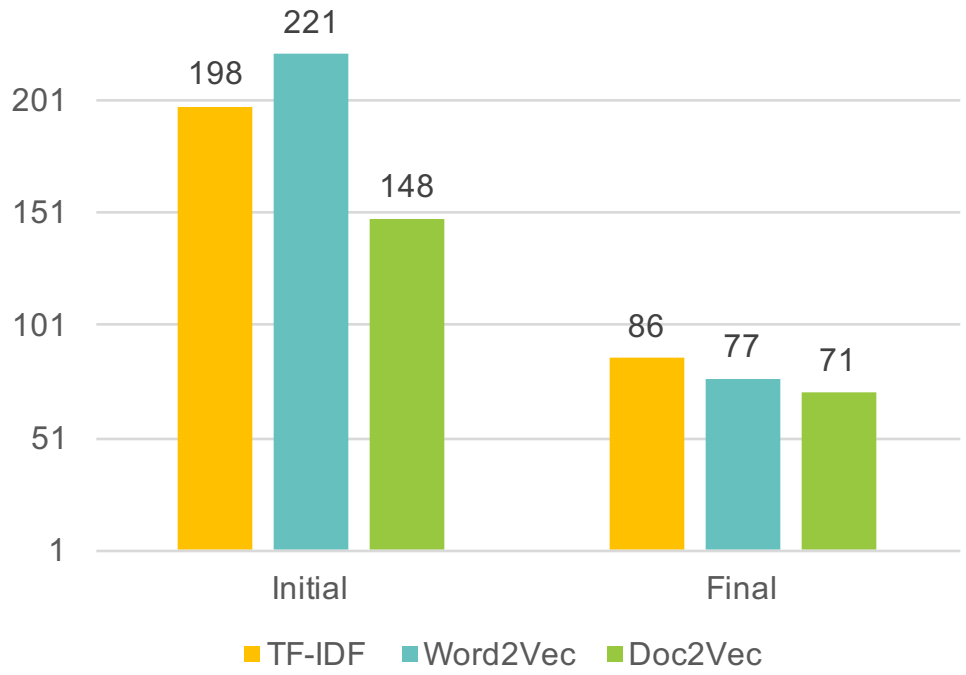


Evaluation & Results

Addition of Context Information

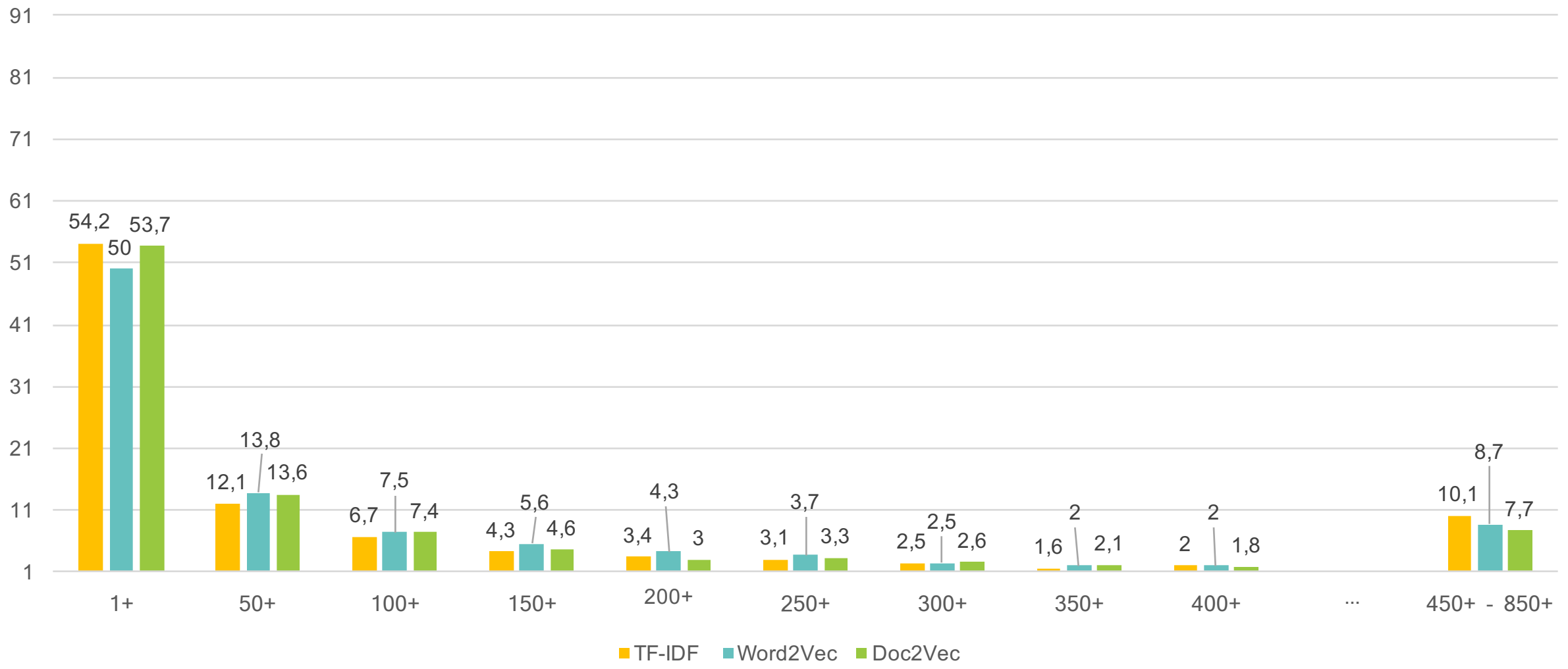


Evaluation & Results
Final Evaluation



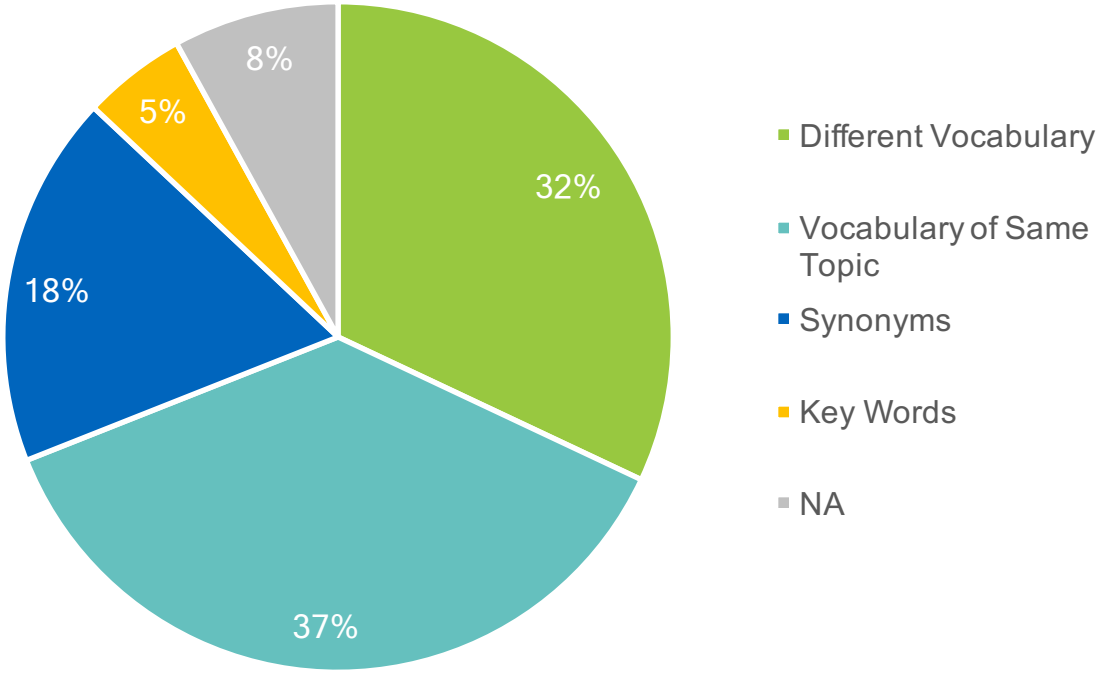
Evaluation & Results

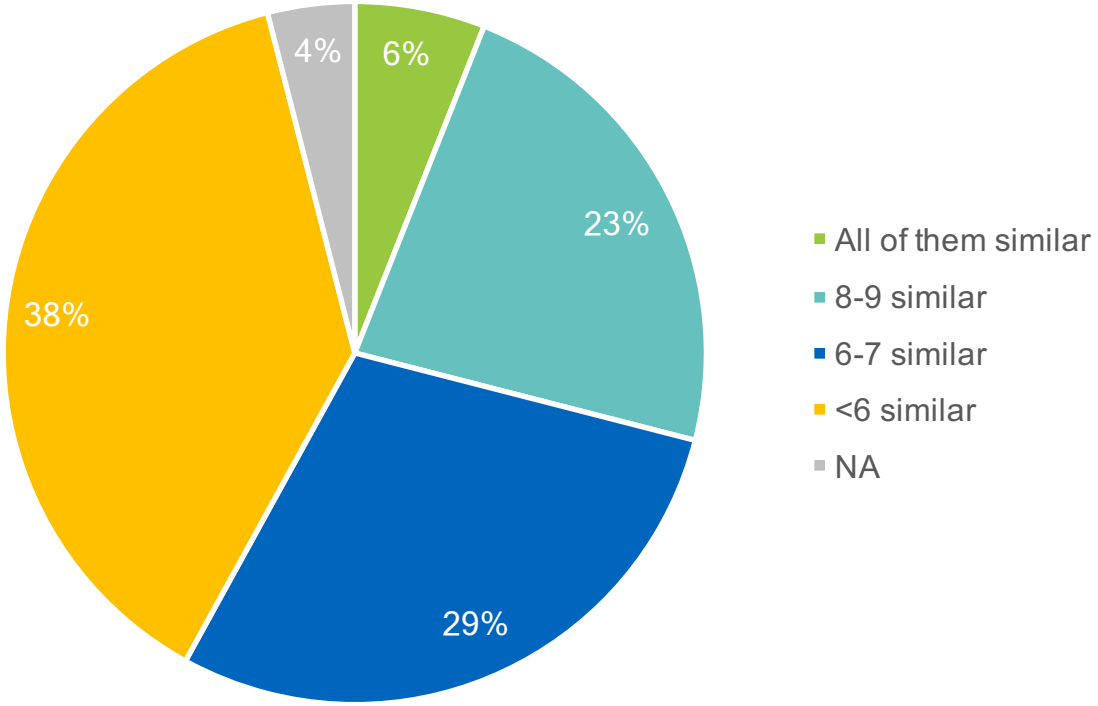
Final Evaluation



Evaluation & Results

Outlier Analysis





Code
Text To Vector

input : Text text, TFIDFVocabCache cache
output: Vector[] vector

```
1 words ← Tokenize(text)
2 vector ← InitialiseVector(Count(cache))
3 for word in words do
4 |   get cache index for word, calculate TF-IDF value and put value to vector
5 |   index ← IndexOf(word,cache)
6 |   vector[index] ← CalculateTFIDF(word)
7 end
```

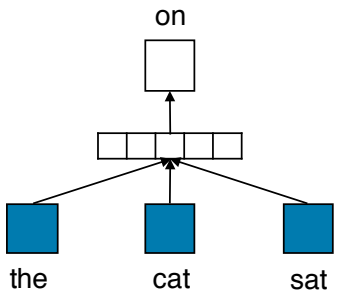
---

input : Text text, WordEmbeddingsVocabCache cache
output: Vector[] vector

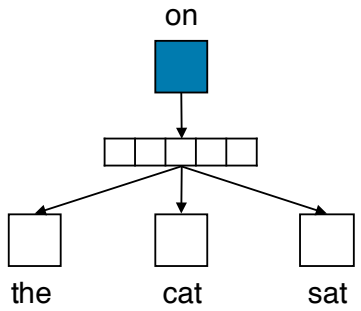
```
1 words ← Tokenize(text)
2 matrix ← InitialiseMatrix(Count(words), GetWordVectorSize (cache))
3 i ← 0
4 for word in words do
5 |   get word vector from cache and put to matrix
6 |   matrix[i] ← GetWordVector(word, cache)
7 |   i ++
8 end
9 average rows of matrix
10 vector ← Average(matrix)
```

Word2Vec, Doc2Vec

Word2Vec

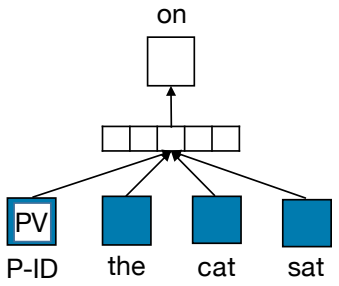


(a) CBOW

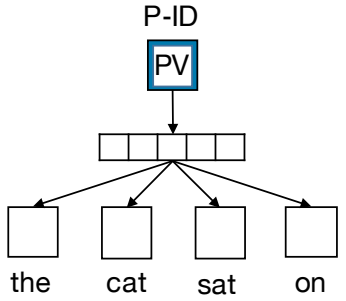


(b) Skip-gram

Doc2Vec



(a) PV-DM



(b) PV-DBOW

Ground Truth

Regulatoy Document	#Paragraphs	∅ Words/Paragraphs
BDSG[2]	24	362.88
COBIT[3] 5 2012	191	32.02
COSO[4] 2013	16	18.94
Finish Personal Data Act 523/1999	29	164.24
ISO 22301:2012	37	122.46
ISO 27001:2005	133	24.86
ISO 31000:2009	46	95.17
MAS-TRMG[5] 21	268	52.00
NIST[6] C2M2 Cyber Security Framework v1.1	95	9.02
PCI DSS[7] v3.1	161	170.16
All Ground Truths	1000	74.30
All Control Statements		21.55