Ukrainian Catholic University

Bachelor Thesis

Optimizing the risk management process in outsourcing based on project management platform database

Author:
Diana Bondar

Supervisor:
Dmytro Irzhytkyi

A thesis submitted in fulfillment of the requirements
for the degree of Bachelor of Science

in the

Department of Computer Sciences
Faculty of Applied Sciences

УКРАЇНСЬКИЙ КАТОЛИЦЬКИЙ УНІВЕРСИТЕТ

APPLIED SCIENCES FACULTY

Lviv 2022

Declaration of Authorship

I, Diana BONDAR, declare that this thesis titled, “Optimizing the risk management process in outsourcing based on project management platform database” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

UKRAINIAN CATHOLIC UNIVERSITY

Faculty of Applied Sciences

Bachelor of Science

Optimizing the risk management process in outsourcing based on project management platform database

by Diana BONDAR

Abstract

This thesis focuses on optimizing aspects of the risk management process in an outsourcing IT organization. With application of the NLP algorithm, we assess the risk mitigation action items, entered by project teams into the risk register of a centralized risk management platform. As an outcome, we enable the system to provide optimal risk mitigation action items that can be used in the template risks, created by the system automatically, saving time for project leadership and improving the quality of the risk mitigation suggestions for project teams, based upon the automatically detected optimal best practice action items.

The suggested approach should contribute to decreased amount of the project risks in the long run, thus ensuring improvement of the overall delivery quality of organizations.

Keywords: *Natural Language Processing, Word Embedding, Risk Management, IT Outsourcing, Clustering, Unsupervised Learning.*

Acknowledgements

I want thank the Myroslava Garasym for providing me such an great opportunity to work on.

And I want to thank to my supervisor, Dmytro Irzhytkyi for his huge support and believing in me.

Contents

List of Figures

List of Tables

List of Abbreviations

NLP	**N**atural **L**anguage **P**rocessing
CSV	**C**omma-**s**eparated **V**alues
JSON	**J**avaScript **O**bject **N**otation
POS	**P**art **o**f **S**peech
W2V	**W**ord **T**o **V**ector
TF-IDF	**T**ext **F**requency - **I**nverse **D**ocument **F**requency
CFG	**C**ontext-**f**ree **G**rammar
CNF	**C**homsky **N**ormal **F**orm
DBI	**D**avies **B**ouldin **i**ndex
TSNE	**t**-distributed **S**tochastic **N**eighbor **E**mbedding
CBOW	**C**ontinuous **B**ag **o**f **W**ords
TF	**T**erm **F**requency
ITF	**I**nverse **T**erm **F**requency
PREX	**Pr**oject **Ex**cellence Platform
CRMS	**C**ollaborative **R**isk **M**anagement **S**ystem
PM	**P**roject **M**anager
PgM	**P**rogram **M**anager
PgD	**P**rogram **D**irector
ADO	**A**ccount **D**elivery **O**wner
DD	**D**elivery **D**irector
ERM	**E**nterprise **R**isk **M**anagement
ISO	**I**nternational **O**rganization for **S**tandartization
RPI	**R**iskiness **P**otential/Opportunities **I**ssues
SDLC	**S**oftware **D**evelopment **L**ife **C**ycle
CBR	**C**ase **B**ased **R**easoning

Non progredi est regredi

Chapter 1

Introduction

1.1 Background

Project Management Institute (PMI) identifies projects as temporary endeavors called to apply efforts in creating value through a unique product, service or result (Project Management Institute, 2022).

The goal of a product or service organization is to increase the rate of successful projects (or products) in order to meet the business goals or stakeholders' expectations. Regardless of efforts and focus from management of different scale organizations, the chances of the project to fail one of the constraints or totally, are quite high. One of the significant reasons of it is poor risk management, which contributes around 5 percent to it, based on research conducted in 2006-2009. (Ratsiepe and Yazdanifard, 2011)

Ability to identify, analyze the risk and prepare the proper actionable response allows organizations to be much more effective enabling success of commercial as well as internal projects. This brings effective risk management and mitigation in the spotlight of focus for executive teams of different organizations, quite often dedicating a lot of effort and resources on handling the risks.

1.2 Problem statement

From a project management standpoint, proper risk management significantly increases chances of project success, which makes it one of the critical activities for a project manager to focus on (Lukas, 2011). To make it an effective and consistent business process, organization may faces challenges that it needs to overcome to succeed:

- **Lack of time**. Effective ongoing risk management requires dedication of time not just from project management, but also from the project leadership team. This is a real challenge, considering the ongoing operational activities that can be quite time consuming in the specific periods of the SDLC. Quite often risk identification and refinement process is either being neglected, or attention to details is dramatically dropping, when risk management is being done just for the sake of the check mark;

- **Lack of experience**. Quite often project managers lack maturity to do the risk management at a proper level. Lacking the experience, they take it to rather theoretical level, quite often failing to identify significant risks, as well as preparing proper responses and mitigation plans;

- **Insufficient sharing of best practice around risk mitigation actions**. No collaborative access to information about activities that lead to successful risk mitigation results in teams failing to reuse the best practices in meeting similar risks. Considering the scale of some IT outsourcing organizations, the probability of similar risks appearing in different projects is quite significant, so using best practices in mitigating them is improving efficiency of delivery organizations;

- **Individual custom risks - quantity over quality**. All projects are required to track risks in the CRMS system. Each project leader has a different level of seniority and maturity in terms of risk management as well as constituency while tracking the risks. This results in creation of huge amounts of custom risks and mitigation action items. In its turn, this results in a greater amount of low quality risks, not being tracked or closed by project teams in the end.

The problems above have been validated with different level representatives of SoftServe delivery organization, who are doing risk management as a part of their routine activities. During the interviews, conducted with leaders of different levels: Project and Product Managers, Delivery Leaders and Vice Presidents, we've managed to confirm all the problems stated above. Experience is a determining factor for less skilled managers, who seek guidance with their direct managers, thus increasing overall time spent on risk management. Organizational best practices in terms of risk management, used in timely manner can increase the chances to mitigate the risks effectively. And following the process is sometimes quite complicated, when it comes to big amount of custom risks that need to be monitored and updated on regular basis.

1.3 Goal

In order to create business value and make organizations more efficient in terms of the risk management, we want to achieve the following goals:

- apply the NLP algorith to create a positive impact on the risk management process for an large-scale outsourcing IT organization;

- increase the efficiency of project/account leadership while planning the mitigation of project risks;

- decrease the amount of noise and improve the quality of action items on individual risks in the system.

1.4 Objectives

In order to achieve the goals of the work and deliver the desired business value, we focus on the following objectives:

1. Assess the collaborative approach towards risk management of SoftServe through centralized risk register of Collaborative Risk Management System (CRMS), used in this organization.

2. Define and shortlist the most prominent action items on individual risks' mitigation, using the NLP algorithm;

3. Improving the overall quality of risk mitigation action items by shortlisting the optimal and trimming the recurrent or low quality action items;

4. Provide optimal recommended action items for risk mitigation in system predefined risks.

1.5 Thesis structure

The thesis includes 8 main chapters. Introduction chapter covers background information as well as goals and objectives, we want to achieve with our work.

Chapter 2 introduces the Risk Management System as the primary tool of managing risks and uncovers some operational aspects of risk management. Starting from the third chapter, we share our approach towards the dataset, we've used for this work and data pre-processing activities and methods used (Chapter 4).

Chapter 5 provides an insight into the text embedding methods and models, we've considered for the implementation part of our work. The methods, flows and metrics, used to clusterize the data, are described in chapter 6. Chapter 7 outlines our selected model and presents the results and outcomes. The visuals for the outcomes part is presented in Appendix A.

The work is concluded by Chapter 8, presenting conclusions and findings.

Chapter 2

SoftServe's Collaborative Risk Management System

2.1 General information

In terms of being more effective in terms of risk assessment and mitigation, as well as adhering to ISO quality standards, SoftServe has developed a platform for centralized collaborative risk management - Collaborative Risk Management System (CRMS). CRMS allows working with the risks on the project and account organizational levels, as well as moving risks through the organizational levels pyramid (Figure 2.1) - from the project level under responsibility of the project manager, to account level, under responsibility of the Delivery Leader (Delivery Director or Associate Delivery Director).

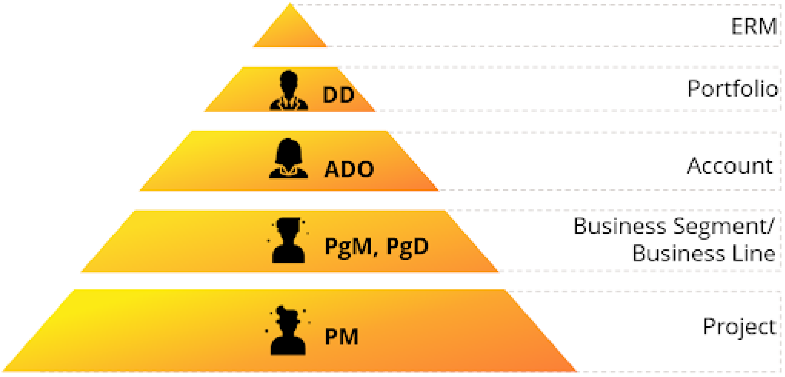


FIGURE 2.1: Collaborative risk management in the organizational pyramid

Collaborative Risk Management is a vital part of the Project Excellence Program. It drives the predictability of business and helps deliver success without surprises.

The Collaborative Risk Management System is a unified space for managing risks, which helps to:

- Drive success without surprises
- Communicate risks to appropriate stakeholders
- Predict and measure the outcomes proactively

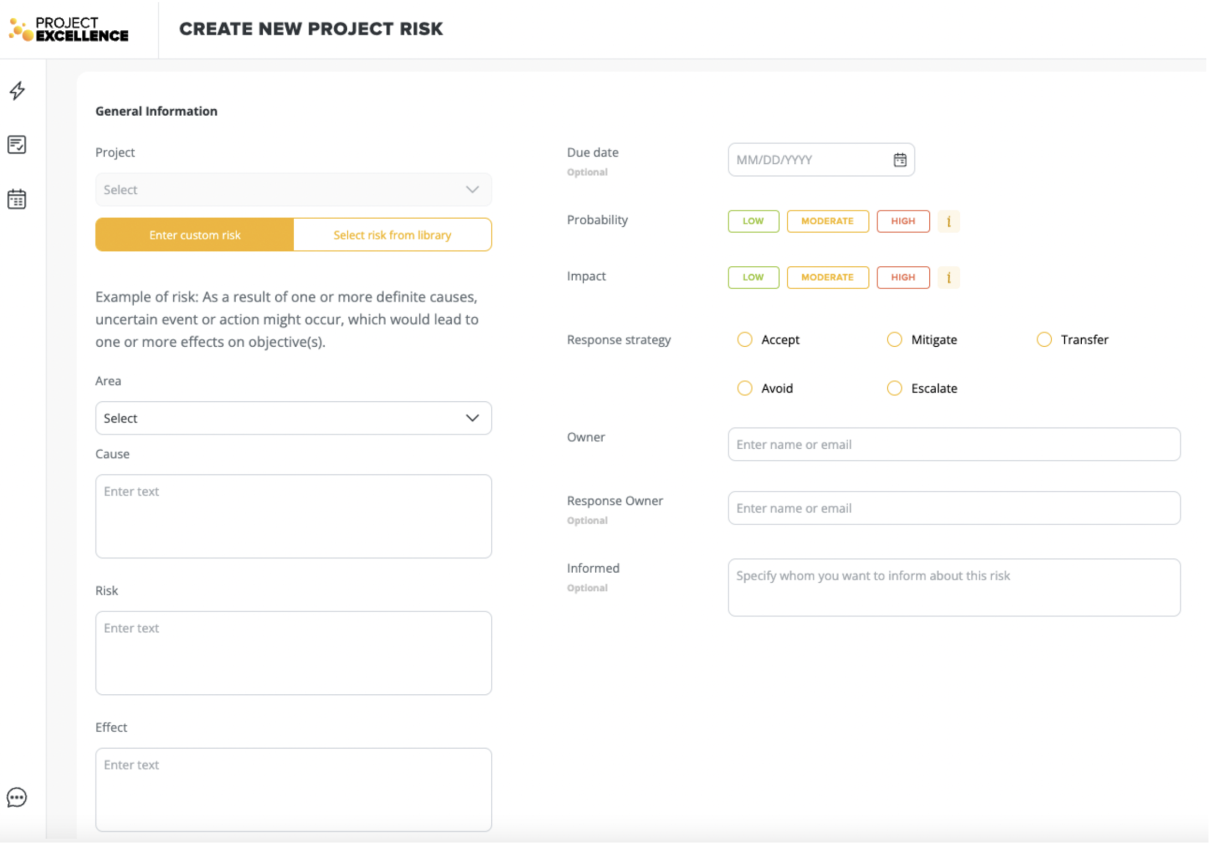
- Provide a unified approach for managing risks
- Enhance collaboration on Risk Management between the Project Managers and Delivery Leads

2.2 Risk tracking

The risks, tracked in the system are being addressed on 2 levels:

- **Individual risk**. These are the individual risk events within the project or account, that have positive or negative impact on project/account objectives. Focusing on such risks allows to enhance the chances of positive project outcome. Such risks are being tracked in a risk register of CRMS.
- **Overall risk**. Focuses on a risk on a level of exposure to project/portfolio, rather than on individual level. Is evaluated by RPI score (Riskiness Potential/Opportunities Issues).

In order to track individual risks, they are created in CRMS as new risks and are further updated on the regular basis by Project Leadership Team members. To create a new risk, you can either enter a custom risk or select risk from the library of predefined risks, broken down by risk areas.



PROJECT EXCELLENCE

CREATE NEW PROJECT RISK

General Information

Project: Select

Enter custom risk | Select risk from library

Example of risk: As a result of one or more definite causes, uncertain event or action might occur, which would lead to one or more effects on objective(s).

Area: Select

Cause: Enter text

Risk: Enter text

Effect: Enter text

Due date (Optional): MM/DD/YYYY

Probability: LOW | MODERATE | HIGH

Impact: LOW | MODERATE | HIGH

Response strategy: ○ Accept ○ Mitigate ○ Transfer ○ Avoid ○ Escalate

Owner: Enter name or email

Response Owner (Optional): Enter name or email

Informed (Optional): Specify whom you want to inform about this risk

FIGURE 2.2: Creating new individual risk in CRMS

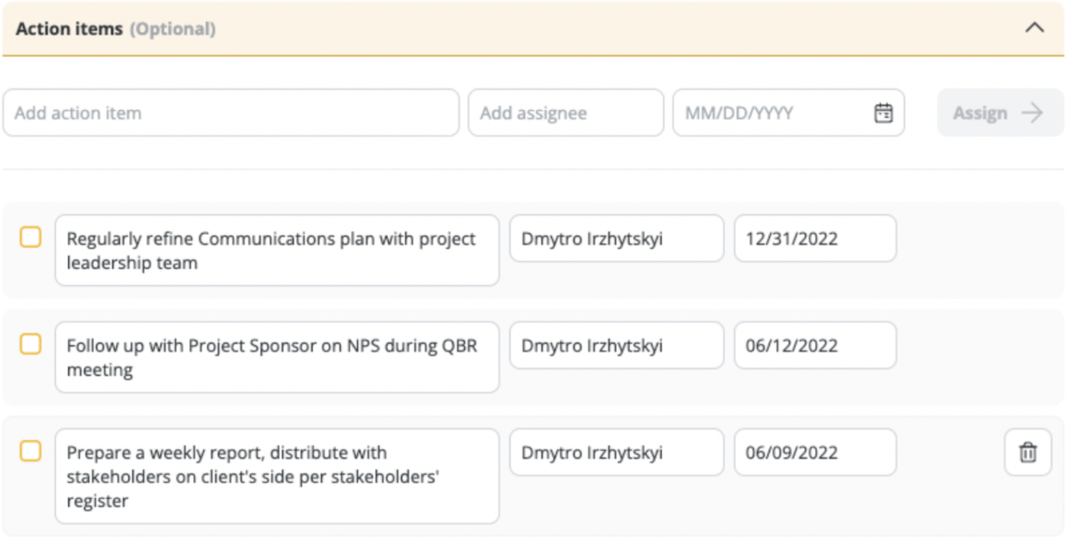
Each individual risk contains information about (Figure 2.2):

- *Risk area*. The category to which risk belongs to;
- *Cause*. What triggers a specific risk to happen;
- *Description*. Short description of a risk;
- *Effect*. The effect of the risk on the project in case it is triggered;
- *Action items*. List of action items for risk mitigation;
- *Comments*. Allows to leave any additional comments regarding a specific risk;
- *Probability*. Risk occurrence probability level - Low, Medium or High;
- *Impact*. The impact of the risk on the project/account - Low, Medium or High;
- *Response strategy*. Defines a risk management approach towards specific risk;
- *Owner*. A person who owns the risk.

Entering the individual risk into the risk register of CRMS and qualifying it with the probability and impact attributes helps to prioritize the risks and have a great visibility over risk management on a specific project/account.

2.3 Risk action items

Each individual risk provides the possibility to add a list of action items - actions that project team can take in order to prevent a risk from happening. The action items can be added while creating the risk or during the risk updating in CRMS (Figure 2.3)



□	Regularly refine Communications plan with project leadership team	Dmytro Irzhytskyi	12/31/2022
□	Follow up with Project Sponsor on NPS during QBR meeting	Dmytro Irzhytskyi	06/12/2022
□	Prepare a weekly report, distribute with stakeholders on client's side per stakeholders' register	Dmytro Irzhytskyi	06/09/2022

FIGURE 2.3: Example of risk mitigation action items

Each action item contains information about:

- *Action item description*. Short description of the action item.
- *Assignee*. A person, responsible for the specific action item.
- *Due date*. A date, when a specific action item becomes due.

The mitigation action items provide guidance to project teams, who know how to mitigate the risk from happening in a very pragmatic and measurable way. Having the owner for each item helps to avoid the shared responsibility and work on risk prevention in a focused way.

We will direct the efforts of our work on risk mitigation action items, to help the project teams be even more successful in risk management as a part of their daily operations.

Chapter 3

Literature Review

Case-Based Reasoning (CBR) is called a process of solving new problems based on the experience got from the previous past problems. (J.-P.Tixier, 2015). It is a popular approach in the Risk management of the Project as for continuous improvement it is important to take outputs from the previous experience. For any risks domain emphasis on similar cases may be extremely valuable to understanding how to deal with an upcoming issue.

> CBR is an important approach to safety risk management of construction projects. It emphasizes that previous knowledge and experience of accidents and risks are extremely valuable and could help to avoid similar risks in new situations.
>
> Bruno Daniotti, 2020

The implementation of CBR includes for main processes: Retrieve, Reuse, Revise, Retain.Plaza, 2001

- *Retrieve*. Stands for the gathering from database experiences clothes to the current problem.
- *Reuse*. Stands for applying the suggested solution for the risk and adapting them to the the current situation.
- *Revise*. Stands for evaluating the use of a solution in a new context.
- *Retain*. Stands for storing the problem-solving method in the database.

However, extracting the 'correct' cases and information from the database quickly and accurate may be a big challenge. Risks-related documents tend to be fulfilled with everyday language and miss the pre-categorized view. The poor case retrieval can significantly affect the user experience and it may lead to the overlooking of the previous experience. And it is considered the biggest problem and disadvantage of the CBR.

In article Zou, 2017 suggests an approach to improve the operation and efficiency of the risk retrieval process. For looking the similar cases it recommends using NLP techniques (document lemmatization and document embedding).

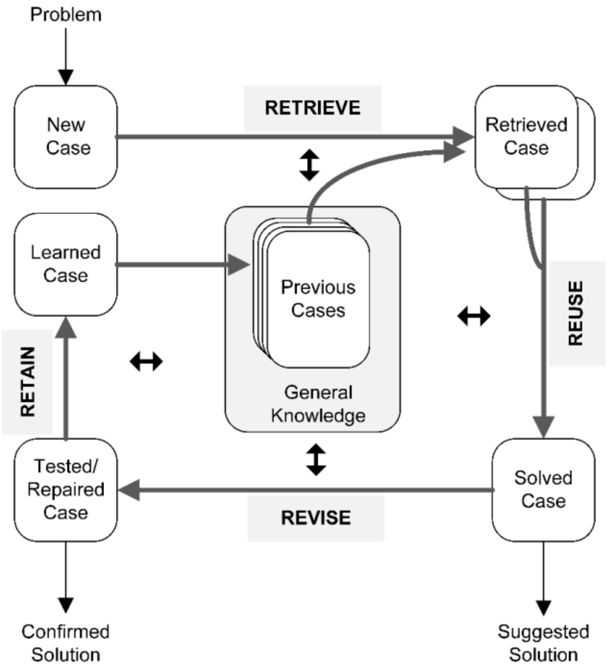


FIGURE 3.1: Classical model of a CBR system (Aamodt and Plaza, 1994)

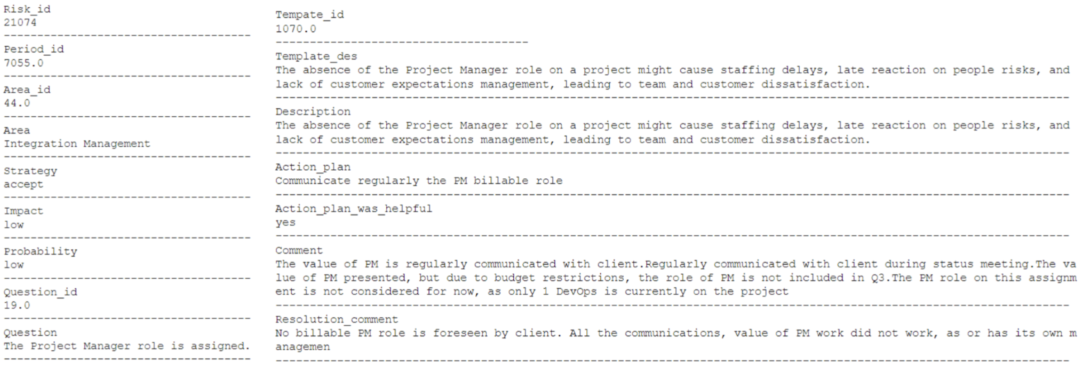
- Items collected from the risks database and data review.
- Textual information extraction from the collected items and processing of them into the corpus.
- Query operation process. Putting document context into the numerical representation.
- Recommendation for the most similar items as a Reuse.
- Adjust the numerical representation by putting additional weights on the numerical model.

Chapter 4

Datasets

4.1 Datasets from Collaborative Risk Management System (CRMS)

The Initial dataset we have used for our work was taken from CRMS system database dump. After completing data clearance procedure (removing empty records and duplicates), the nested fields were gathered from tables and where merged into a single CSV file. It contains the records, revealed among individual project risks as well as their properties. The table below provides the contents of CSV file with sample rows. 4.1



```
Risk_id
21074
-------------------------------------
Period_id
7055.0
-------------------------------------
Area_id
44.0
-------------------------------------
Area
Integration Management
-------------------------------------
Strategy
accept
-------------------------------------
Impact
low
-------------------------------------
Probability
low
-------------------------------------
Question_id
19.0
-------------------------------------
Question
The Project Manager role is assigned.
-------------------------------------
```

```
Tempate_id
1070.0
-------------------------------------
Template_des
The absence of the Project Manager role on a project might cause staffing delays, late reaction on people risks, and
lack of customer expectations management, leading to team and customer dissatisfaction.
----------------------------------------------------------------------------------------------------------------------
Description
The absence of the Project Manager role on a project might cause staffing delays, late reaction on people risks, and
lack of customer expectations management, leading to team and customer dissatisfaction.
----------------------------------------------------------------------------------------------------------------------
Action_plan
Communicate regularly the PM billable role
----------------------------------------------------------------------------------------------------------------------
Action_plan_was_helpful
yes
----------------------------------------------------------------------------------------------------------------------
Comment
The value of PM is regularly communicated with client.Regularly communicated with client during status meeting.The va
lue of PM presented, but due to budget restrictions, the role of PM is not included in Q3.The PM role on this assignm
ent is not considered for now, as only 1 DevOps is currently on the project
----------------------------------------------------------------------------------------------------------------------
Resolution_comment
No billable PM role is foreseen by client. All the communications, value of PM work did not work, as or has its own m
anagemen
----------------------------------------------------------------------------------------------------------------------
```

FIGURE 4.1: CSV file contents with sample rows

4.2 The database quantitative exploration

Dataset contains data gathered since 2020 year, since the CRMS rollout. The table below represents quantitative measurements of the text fields.

Field	**Records in database**
Risks (total)	37346
Risks (from template)	24306
Risks (from template unique)	1178
Risks (entered custom)	13040
Action plans (total)	44904
Action plans for (for risks from template)	22068
Action plans (for custom risks)	22836
Comments	5645
Resolution Comments	17982

TABLE 4.1: Quantitative measurements of the text fields in CRMS database

4.3 Datasets for the NLP

In order to improve NLP processing, the following datasets have been used:

1. **en_stop_words.json"** - JSON file, that contains English Stop-words;
2. **Words_dictionary.json** - JSON file, containing over 466k English Vocabulary words;
3. **Unigram_freq.csv** - CSV file, that contains the counts of the 333,333 most commonly-used single words on the English language web and their frequency, as derived from the Google Web Trillion Word Corpus;
4. **Domain_words.json** - custom file that contains Doman-specific words that are not English Vocabulary but should be taken into consideration (*Ex:* "pii", "dd", "cobol");
5. **glove-wiki-gigaword-50.model** - pretrained W2V embedding model. Model contains 50 dimensional vector representations of the words in Wikipedia articles.

Chapter 5

Data preprocessing

5.1 Tokenization

The first, preprocessing step for the NLP processing pipeline is a tokenization. Token is an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit.
Tokenization goals for the current case is to separate words from the punctuation from the text block.
Tokenization algorithm performs the following functions:

1. Splitting sentence by spacing and removing external spacing from split parts;
2. Detecting punctuation from the parts start/end and splitting punctuation as a different parts;
3. Splitting exceptions from the start/end of the split parts (*ex :* “‘s”, ‘n’t’).

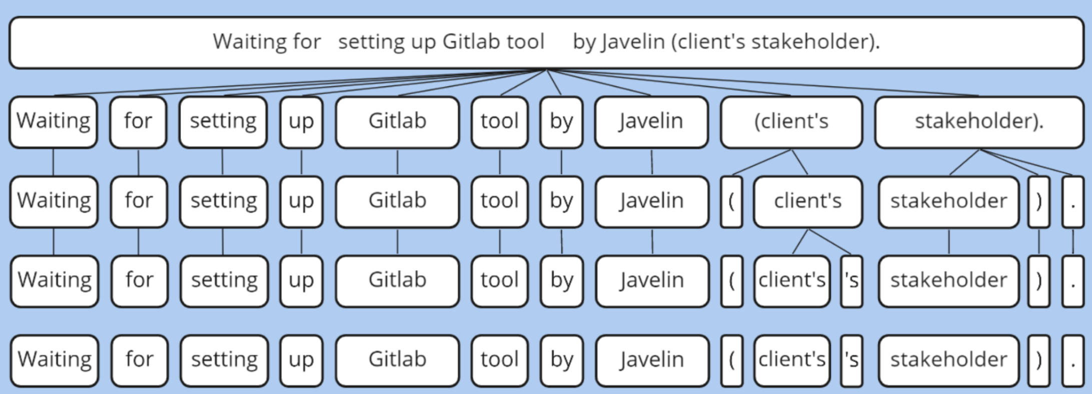

Waiting for setting up Gitlab tool by Javelin (client's stakeholder).
Waiting | for | setting | up | Gitlab | tool | by | Javelin | (client's | stakeholder).
Waiting | for | setting | up | Gitlab | tool | by | Javelin | (| client's | stakeholder |) | .
Waiting | for | setting | up | Gitlab | tool | by | Javelin | (| client's | 's | stakeholder |) | .
Waiting | for | setting | up | Gitlab | tool | by | Javelin | (| client's | 's | stakeholder |) | .

FIGURE 5.1: Tokenization algorithm

5.2 Text Normalization

Text normalization is a text transformation into a single canonical form. The main purpose behind text normalization is creation of common base form by reducing the forms (inflectional and derivative) of a single word. In this way, we can easily do word matching, that possess same meaning, provided by context.

5.2.1 Words filtering

The first part of the text normalization is removing tokens that are not words (consist of symbols different from [A-z]). So that we are getting rid of any punctuation, non-English words, hyperlinks and some misspelled words.

5.2.2 Lemmatization

In scope of the lemmatization the word is transferred into it's base form. For the lemmatization algorithm word's endings are checked with a corresponding endings and exceptions for its POS and are replaced by them.

For example, for adjectives endings 'er' , 'est' would be substituted with an ending 'e' or ''. (Wider -> Wide. Tallest -> Tall). Then, words are lowercased.

5.2.3 Stop-words filtering

Excluding stop words from the text set to avoid the additional noise at the model. Stop words are words that mostly don't bring a contextual value but occur in text frequent enough (articles, modal verbs, conjunctions). For our case all frequent words that don't bring any contextual/domain value to the text should be considered as stop-words. (mention JSON with stop_words was used to define and exclude them).

5.2.4 Removing words out of vocabulary

The final part of text normalization is to remove out of the vocabulary elements in order to eliminate the additional noise. So that First names, Last names, misspelled words and other non-common abbreviations are deleted. For the words, used to define and exclude out of vocabulary words.json' file was used enhanced with a domain.json set (a custom set with a frequently occuring words in a CRMS).

5.2.5 Result of sentence normalization

The table below provides an example of the sentence normalization:

Token Text	Token is Alpha	Lemmed Text	Token in Vocab	Token is Stop Word	Token is Domain Specific	Normalized Text
Waiting	True	wait	True	False	False	wait
for	True	for	True	True	False	
setting	True	set	True	False	False	set
up	True	up	True	True	False	
Gitlub	True	gitlub	False	False	True	gitlub
tool	True	tool	True	False	False	tool
by	True	by	True	True	False	
Javelin	True	javelin	True	False	False	
(	False					
client	True	client	True	False	False	client
‘s	False					
stakeholder	True	stakeholder	True	False	False	stakeholder
)	False					
.	False					

TABLE 5.1: Normalized sentence example

Chapter 6

Text embedding methods overview and implementation

6.1 Word2Vec model overview

Word2Vec is a technique of the Word embedding. It uses neural network to generate associations from a large corpus of text and produces a multidimensional vector space where each unique word is associated with its corresponding vector. Word vectors are positioned this way that words semantic similarity may be indicated with similarity of corresponding vector (Zhang et al., 2021).

There are two different methods within the Word2Vec algorithm. SkipGram and CBOW. In the CBOW surrounding words are combined to predict the word in the middle. When SkipGram uses one word to predict the context.

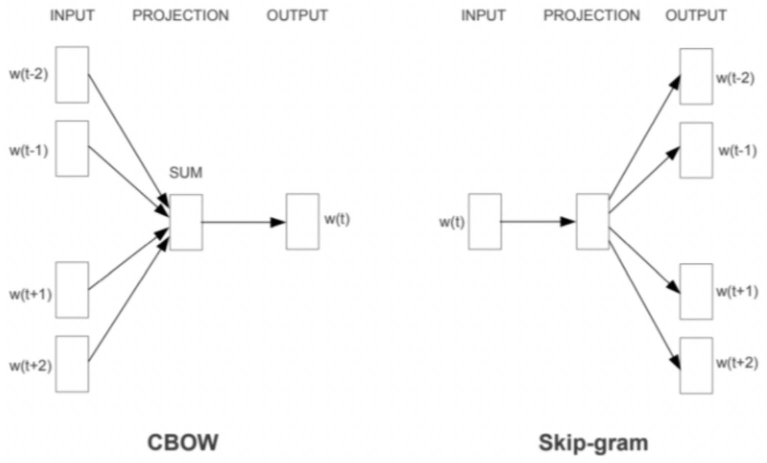


FIGURE 6.1: CBOW vs. SkipGram method

In our case, I used the SkipGram model as it will be needed to predict the general context of the whole sentence. Also, SkipGram is optimal for smaller datasets and can be much more meaningful in case we need to represent the words, that are met less frequently.

6.1.1 Algorithm

The Skip-Gram model assumes that a word can be used to predict its surrounding words in a text corpus. The main idea stands for this algorithm is that a center word (Vc) attempts to predict the conditional probability of it's neighboring words(Vw) and than, tries to maximize that probability of occurrence.

The number of neighboring words that will be calculated for the function depends on the selected the window size (m).

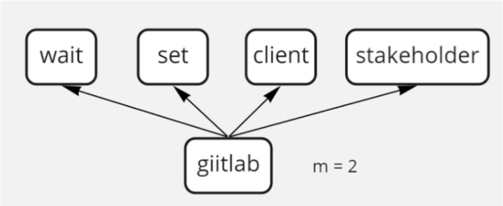


FIGURE 6.2: Window =2 selection for the Skip-Gram

By taking into account an assumption that all words are independently generated given the central word the condition probability can be:

$P(\text{‘}wait\text{”}|\text{'}gitlab\text{”}) * P(\text{‘}set\text{'}|\text{'}giitlab\text{'}) * P(\text{”}client\text{”}|\text{'}gitlab\text{”}) * P(\text{‘}stakeholder\text{”}|\text{”}gitlub\text{”})$

For Skip-Gram the softmax function is a probability of the output word given the input word. For the function, u and v are input and output vectors.

$$P(w_o \mid w_c) = \frac{\exp(\mathbf{u}_o^\top \mathbf{v}_c)}{\sum_{i\in\mathcal{V}} \exp(\mathbf{u}_i^\top \mathbf{v}_c)}$$

The likelihood function of the skip-gram model calculates the likelihood function of the probability of generating all context words given any center word:

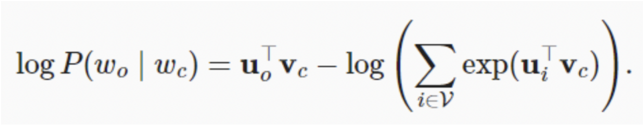
$$\prod_{t=1}^{T} \prod_{-m\leq j\leq m,\ j\neq 0} P(w^{(t+j)} \mid w^{(t)}),$$

Training w2v

For the word sequence w1,w2,w3. . . .wt on objective to maximize the probability of predicting similar words given a target word can be written as an average log probability function:

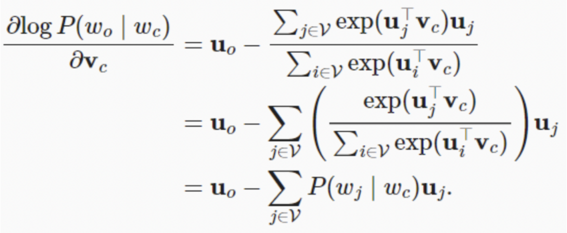
$$-\sum_{t=1}^{T} \sum_{-m\leq j\leq m,\ j\neq 0} \log P(w^{(t+j)} \mid w^{(t)}).$$

Than, gradient descent is used to gradually change the weights and increase the likelihood function,



$$\log P(w_o \mid w_c) = \mathbf{u}_o^\top \mathbf{v}_c - \log\left(\sum_{i\in\mathcal{V}} \exp(\mathbf{u}_i^\top \mathbf{v}_c)\right).$$

Afterwards, gradient with respect to the center word vector can be obtained through differentiation,



$$\begin{aligned}\frac{\partial \text{log}\, P(w_o \mid w_c)}{\partial \mathbf{v}_c} &= \mathbf{u}_o - \frac{\sum_{j\in\mathcal{V}} \exp(\mathbf{u}_j^\top \mathbf{v}_c)\mathbf{u}_j}{\sum_{i\in\mathcal{V}} \exp(\mathbf{u}_i^\top \mathbf{v}_c)}\\ &= \mathbf{u}_o - \sum_{j\in\mathcal{V}} \left(\frac{\exp(\mathbf{u}_j^\top \mathbf{v}_c)}{\sum_{i\in\mathcal{V}} \exp(\mathbf{u}_i^\top \mathbf{v}_c)}\right)\mathbf{u}_j\\ &= \mathbf{u}_o - \sum_{j\in\mathcal{V}} P(w_j \mid w_c)\mathbf{u}_j.\end{aligned}$$

As a result of the training, for all words in set the Vi (center word) and Ui (the context word) is updated. We will use a context word as a word representation.

6.2 Word2Vec model implementation

6.2.1 Parameters selection

For the model pretraining I used text normalized records from the following text data:

- Description (only Custom risk descriptions);
- Unique Risk Templates;
- Action plan;
- Risk Comment;
- Resolution Comment.

Model parameters were selected:

- Window = 2.
- Minimal word occurrence = 5.
- SkipGram method.

6.2.2 Model results

To evaluate model results we can test it by showing top results for the words to see whether it guesses context-similar words correctly. As you can see, our custom model provide much more narrow cottex and returns Domain specific-synonyms while PRETRAINED MODEL provides more general similarities.

The table below provides the results of the model trained on the CRMS text.

word	CRMS-trained TOP-7 closest words
escalate	rise notify vp highlight act resolve inform
promotion	annual career subordinate succession compensation pdp budgeting
sprint	pi iteration spring wbs splitting techdebt chunk
ukraine	war russia military relocation political invasion russian

TABLE 6.1: Results of the model trained on the CRMS text

The table below provides the results of the pretrained glove-wiki-gigaword-50.

word	glove-wiki-gigaword-50 TOP-7 clothest words
escalate	escalating escalation escalates aggravate provoke exacerbate bloodshed
promotion	promoted completion sponsorship promotions promote promoting participation
sprint	nextel competitor slalom race racing relay 200m
ukraine	belarus russia bulgaria romania moldova kazakhstan ukrainian

TABLE 6.2: Results of the pretrained glove-wiki-gigaword-50

The pictures below shows a TSNE representation of the CRMS w2v model words. The second picture is a zoomed example of the model.

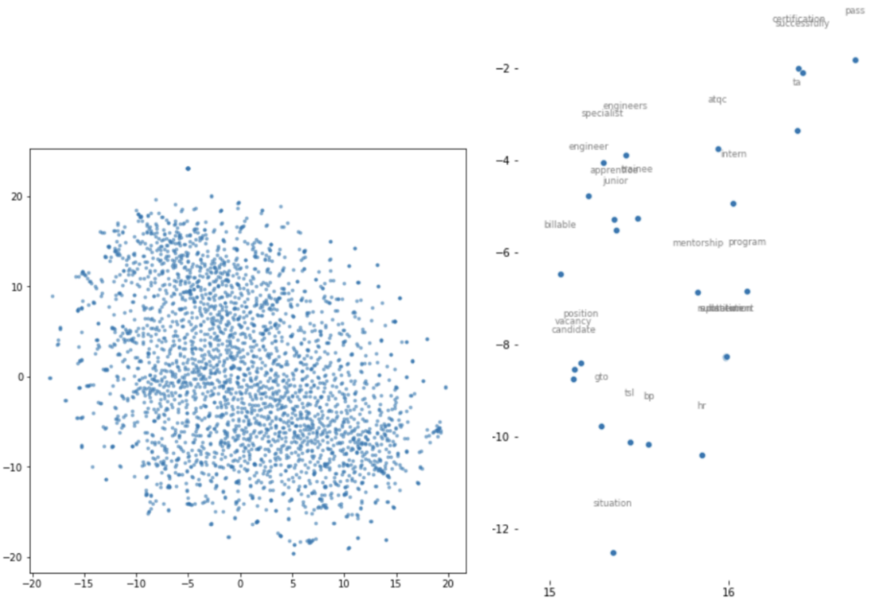


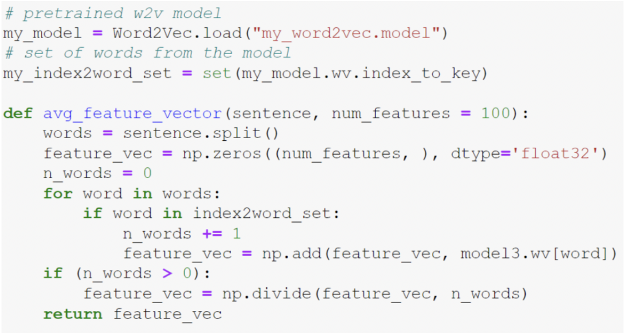
FIGURE 6.3: TSNE projection of the CRMS Word2Vec model word vectors

6.3 Sentence2Vec model

Sentence 2 Vec representation allows to represent the whole sentence at the Vector space by representing its' average vector based on their Word2Vec representations.

Word to Sentence approach is useful for the computation short sentences computations as vector direction is set preciacely enough. To calculate the Sentence embedding Averaged sum of the Vectors at the Sentence is calculated. The same as with Words Embedding, semantically cloth sentences are located next to each other into the multydimention space (Haw-Shiuan Chang, 2021).

For all words within the sentence Corresponding Vectors are taken from the model if the word is not at the model - it is skipped. The sentence vector is calculated as an average of the it's words vectors that are present in the model.



```
# pretrained w2v model
my_model = Word2Vec.load("my_word2vec.model")
# set of words from the model
my_index2word_set = set(my_model.wv.index_to_key)

def avg_feature_vector(sentence, num_features = 100):
    words = sentence.split()
    feature_vec = np.zeros((num_features, ), dtype='float32')
    n_words = 0
    for word in words:
        if word in index2word_set:
            n_words += 1
            feature_vec = np.add(feature_vec, model3.wv[word])
    if (n_words > 0):
        feature_vec = np.divide(feature_vec, n_words)
    return feature_vec
```

FIGURE 6.4: Sentence to Vector algorithm

6.4 Sentence2Vec weighting

There is an option to add weightings to the Sentence to vector model depending on the model goal. Applying a weighting depending on the goal of the model. It may be a POS tagging, to put more model attention on the exact part of speech. Or Entity tagging, to put more attention on the Entities in the sentence.

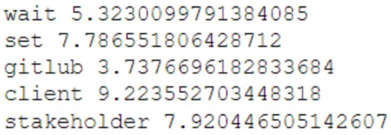
Algorithm

Sentence to Vec Weighted embedding implies multiplying each word in the sentence to its corresponding weighing.

$\sum_{k=1}^{i} idf(w_k) * v_k$

At this part, we attempted to implement the weighting dependent on the Term Frequency and Inverse Frequency. By applying these algorithms we expect sentences to simulate more by detecting the most valuable words.

Term Frequency (TF) approach implies using the weighting coefficient for the word proposed to it's frequency at the CRMS training corpus. Example: the result of weightings on words

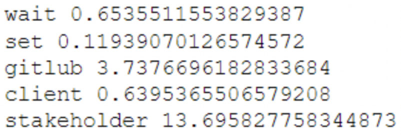


```
wait 5.3230099791384085
set 7.786551806428712
gitlub 3.7376696182833684
client 9.223552703448318
stakeholder 7.920446505142607
```

Inverse Terms Frequency approach is based on the TF-IDF method (in it term frequency is compared with a term frequency in all other documents) For calculation of inverse term frequency, each term frequency at the trained corpus is divided by term frequency stated at the **Unigram_freq.json**. The term frequency value is logged to avoid extremely scaling for outliers (frequent words). This approach should add more weight to the Domain-Specific words.

- The highest weight will get words that are rarely used in the common language but are often used at the CRMS;
- Words commonly used at the CRMS tool but also commonly used in a common language will have the same weights as the rarely used words at the CRMS and rarely used in the common language;
- Rarely used at the CRMS tool and commonly used in a common language will have the lowest value.

Example: result of weightings on words:



```
wait 0.6535511553829387
set 0.11939070126574572
gitlub 3.7376696182833684
client 0.6395365506579208
stakeholder 13.695827758344873
```

Chapter 7

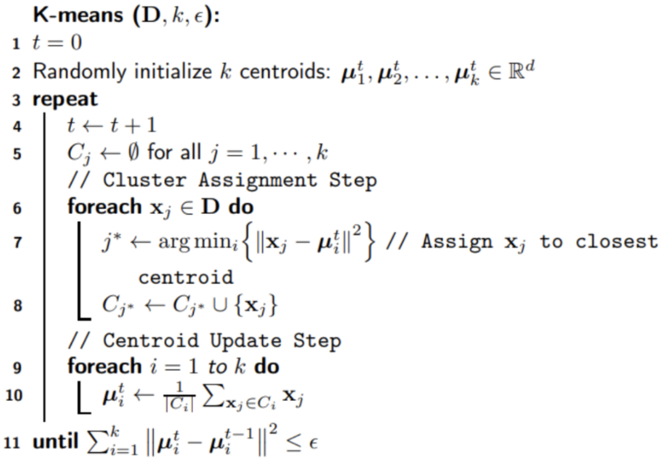
Text clusterization methods and performance measurements overview and implementation

7.1 K-means method

K-means method is is an iterative algorithm that attempts to partition dataset into predefined K groups by minimizing within-cluster variances and maximaxing distances within clusters, to ovoid cluster overlaps.

7.1.1 K-means clustering flow

K-means algorithm assigns datapoints of each dataset to the nearest (with a minimal euclidean distance) predefined cluster centers. Afterwards, it validates that for all clusters centroid id is the clothest datapoint to the actual cluster mean. If it doesn't - the closest datapoints are selected as a new cluster centroids and iteration reinializes.



```
K-means ($\mathbf{D}, k, \epsilon$):
$t = 0$
Randomly initialize $k$ centroids: $\boldsymbol{\mu}_1^t, \boldsymbol{\mu}_2^t, \ldots, \boldsymbol{\mu}_k^t \in \mathbb{R}^d$
repeat
    $t \leftarrow t + 1$
    $C_j \leftarrow \emptyset$ for all $j = 1, \cdots, k$
    // Cluster Assignment Step
    foreach $\mathbf{x}_j \in \mathbf{D}$ do
        $j^* \leftarrow \arg\min_i \left\{ \|\mathbf{x}_j - \boldsymbol{\mu}_i^t\|^2 \right\}$ // Assign $\mathbf{x}_j$ to closest
            centroid
        $C_{j^*} \leftarrow C_{j^*} \cup \{\mathbf{x}_j\}$
    // Centroid Update Step
    foreach $i = 1$ to $k$ do
        $\boldsymbol{\mu}_i^t \leftarrow \frac{1}{|C_i|} \sum_{\mathbf{x}_j \in C_i} \mathbf{x}_j$
until $\sum_{i=1}^k \left\|\boldsymbol{\mu}_i^t - \boldsymbol{\mu}_i^{t-1}\right\|^2 \leq \epsilon$
```

FIGURE 7.1: K-means clustering algorithm

7.1.2 K-means drawbacks

The main drawbacks of K-means method application are:

1. Clusters are sensitive to the initial centroid posinions, especially for the big K. Sensitivity to wrong the initial centroid selection can be easily bid with a Ensemble Clustering (Kuncheva and Vetrov, 2006). And repeating K-means algorithm several times.

2. Data are very sensitive to the outliers and to the data shape. Especially in the high dimension space outliers datapoints, that are significantly far away from the cluster can significantly change the cluster values distribution. (resolution is described below).

7.1.3 K-means outliers filtering

In this article (Hautamäki, 2005), we can find a suggested method for the mastering outliers problem at the K-means algorithm.

Algorithm description

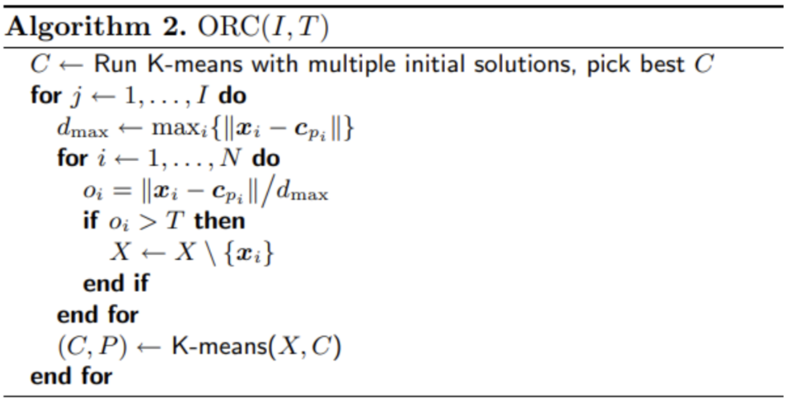
The algorithm consists of the repetitive consecutive stages. At the first stage K-means algorithm is run and silhouette index is calculated. For each cluster maximum distance from the centroid to the cluster datapoints is defined:

$$d_{max} = max_i\{\left\|x_i - c_{p_i}\right\|\}, i = 1, ..N$$

Then, for each datapoint in each cluster outliers factor is defined. It is a normalized value with a scale [0,1]. The greater the value is, the most likely datapoint to be an outlier of the cluster:

$$o_i = \frac{x_i - cp_i}{d_{max}}$$

At the second stage general T value is set. It is a threshold value to define datapoints which are located further away from the cluster center than expected. Then, each datapoint with an outlier factor less then T value will be removed from the dataset. The less T value, the more datapoints will be filtered from the dataset. It is recommended to use the removal threshold as 0.9 or 0.95..



```
Algorithm 2. ORC($I, T$)
  $C \leftarrow$ Run K-means with multiple initial solutions, pick best $C$
  for $j \leftarrow 1, \ldots, I$ do
    $d_{\max} \leftarrow \max_i\{\|\boldsymbol{x}_i - \boldsymbol{c}_{p_i}\|\}$
    for $i \leftarrow 1, \ldots, N$ do
      $o_i = \|\boldsymbol{x}_i - \boldsymbol{c}_{p_i}\| / d_{\max}$
      if $o_i > T$ then
        $X \leftarrow X \setminus \{\boldsymbol{x}_i\}$
      end if
    end for
    $(C, P) \leftarrow$ K-means($X, C$)
  end for
```

FIGURE 7.2: K-means outliers filtering algorithm

This method allows to filter outliers from the dataset and significantly improve the performance metrics (Siloutte score). However, there is no stated parameters for this approach (Thresholds value and number of iterations) and these numbers needed to validated depend on the dataset.

7.2 Performance measurement metrics

7.2.1 General information

Evaluating performance of the clustering algorithm is an important part of the clustering process. There are two methods, unsupervised (internal metrics are used for validation) and supervised (external metrics are used for validation).

> "In supervised learning,"the evaluation of the resulting classification model is an integral part of the process of developing a classification model and there are well-accepted evaluation measures and procedures" . pervised learning, because of its very nature, cluster evaluation, also known as cluster validation, is not as well-developed."
>
> Palacio-Nino and Berzal, 2019

When analyzing clustering results, several aspects must be taken into account for the validation of the algorithm results :

1. Defining whether the non-random structure is present within clusters;

2. Defining the correct number of clusters;

3. Assessing the quality of the clustering results with an internal information;

4. Comparing the results obtained with external information;

5. Comparing several sets of clusters to determine which one is better.

The first methods 1,2,3,5 can evaluate model with the internal information and the method 4 requires external information.

To measure the performance with an external way, a predefined knowledge of the real data distribution should be present. Due to the following factors external clusterization evaluation can not be taken into account and internal evaluation scheme is used as a primary one. (The correlation with a Risk properties is not so significant as to take it as a report point).

- Some risk parameters are not required and therefore are skipped;
- They can be quite new in the system and in most cases there is no data;
- Human factor. The measurement differs from manager to manager. Ex: by selecting only action plans such as “Escalate to DD”, risk strategies can be to escalate, and thus mitigate or accept the risk;
- Different action items may be generated for the same risk and vice versa one action item may cover several risks;
- Due to the fact that data is examined within the narrow domain specific it can not be evaluated with a similar datasets.

Thus, our case covers only the internal validation metrics.

Internal validation Metrics overview

In general, two metrics for the internal validation can be measured and combined **cohesion** and **separation**. Cohesion evaluates elements within the same cluster are cloth to each other. Separation stand for the level of separation between clusters.

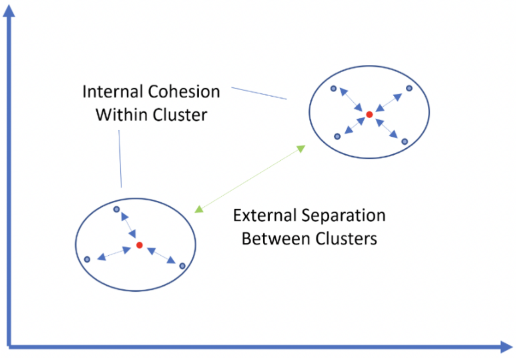


FIGURE 7.3: Cohesion and separation

7.2.2 Schillotte score

In order to measure the quality of a specific clustering technique is, Silhouette Score is quite commonly being used as unsupervised metrics for evaluation. As an outcome, we can get a short view on quality of presentation of objects’ classification, enables value by cohesion view within own cluster as well as separation view - comparison with other clusters. It's a common measure, including representation of both cohesion and separation metrics (Palacio-Nino and Berzal, 2019).

Schilotte score computation

Silhouette score is calculated for each datapoint at the cluster with a formula:

$$s(i) = \frac{b(i)-a(i)}{max\left\{a(i),b(i)\right\}}$$

- a(i) is an av distance between datapoint i and all other datapoints in cluster to which i belongs.

$$a(i) = \frac{1}{|C_a|} \sum_{j \in C_a, i \neq j} d(i,j)$$

b(i) is minimal distance from i to all other clusters cetroids to which datapoint i doesn't belong.

$$b(i) = \min_{C_b \neq C_a} \frac{1}{|C_b|} \sum_{j \in C_b} d(i,j)$$

Global Shilooette score for dataset is calculated as an average of add Shilooette scores.

$$S = \frac{1}{n} \sum_{i=1}^{n} s(i)$$

Silhouette score value ranges from -1 to 1. It may represent on the datapoint, cluster and dataset level how good clustering performance is. The higher value - the more evidence that clustering is implementing well. Cloth to 0 score means that there is not enough clustering evidence found.

Silhouette score is used for the clustering evaluation. If the dataset score is higher than 0.25 it may give us enough evidence that there clusters in this data are assigned 'good enough".

7.2.3 Davies–Bouldin index (DBI)

> "The Davies-Bouldin (DB) method also relates compactness to cluster separation. The score is defined as the average similarity measure of each cluster with its most similar cluster, where similarity is the ratio of within-cluster distances to between-cluster distances. This affirms the idea that no cluster has to be similar to another, and hence the best clustering scheme essentially minimizes the Davies–Bouldin index."

Baarsch and Celebi, 2012

For the cluster i index calculation each cluster is compared with a neigbourhood clusters by the following formula:

$$R_{ij} = (S_i + S_i)/M_{ij}$$

Where Si and Sj represents the sum of the average distances from each point in cluster i, j to the centroid of cluster i, j. And Mij is a distance between clusters i and j.

Afterwards, minimal value is selected from the cluster comparisons (formula below). The lowest value represent the closest, most similar cluster:

$$R_i = maxR_{ij}, i \neq j$$

Final result is an average of maximal results from for each cluster.

$$\frac{1}{k} \sum_{i=1}^{k} R(i)$$

Chapter 8

Model selection and final model results

In this chapter the models evaluation will be explored. Each model was run on a Action plan corpus, and based on the most efficient Siloutte score and DBI an appropriate clusterization method was selected. Then, each model is valalidated manually by exploring the recommendations for the Risks Templates.

The following models were examined and compared:

1. **CRMS, no weighting**. Using results for the Sen2Vec decomposition on a CRMP trained corpus, without additional actions;
2. **CRMS + Outliers, no weighting**. Using results for the Sen2Vec decomposition on a CRMP trained corpus. Having K-means outliers filtering algorithm applied;
3. **CRMS + TF**. Using results for the Sen2Vec decomposition on a CRPM trained corpus. Applying words weight based on the Term Frequency;
4. **CRMS + TF + Outliers**. Using results for the Sen2Vec decomposition on a CRPM trained corpus. Applying words weight based on the Inverse Term Frequency and having K-means outliers filtering algorithm applied;
5. **CRMS + ITF**. Using results for the Sen2Vec decomposition on a CRPM trained corpus. Applying words weight based on the Inverse Term Frequency;
6. **CRMS + ITF + Outliers**. Using results for the Sen2Vec decomposition on a CRPM trained corpus. Applying words weight based on the Inverse Term Frequency and having K-means outliers filtering algorithm applied.

8.0.1 Silhouette Score measurements

Results of the measuring Silhouette Score on models. The 'best' Silhouette Scores values are highlighted as bold.

K Clusters	CRMS w2v	CRMS + outliers filtering	CRMS + TF	CRMS + TF-IDF	CRMS+ TF +Outliers	CRMS + ITF + Outliers
20	0.021	0.175	0.045	0.64	0.08	**0.65**
35	0.035	0.24	0.053	**0.70**	0.1	0.6
50	0.041	**0.375**	0.06	0.55	0.14	0.62
70	0.076	0.35	0.055	0.56	0.16	0.5
100	0.059	0.33	0.058	0.45	0.19	0.2
150	**0.080**	0.335	**0.07**	0.4	0.2	0.21
200	0.092	0.356	0.068	0.33	**0.22**	0.25

TABLE 8.1: Measuring Silhouette Score on models

8.0.2 DBI index measurements

Results of the measuring DBI index on models. The 'best' DBI scores are highlighted as bold.

K Clusters	CRMS	CRMS + outliers filtering	CRMS + TF	CRMS + TF-IDF	CRMS+ TF +Outliers	CRMS + ITF + Outliers
20	3.21	2	2.85	1.3	2.2	0.6
35	2.83	1.6	2.83	1.15	2.25	0.53
50	**2.3**	**1.3**	2.65	1.15	1.9	0.4
70	2.34	1.4	2.74	**0.95**	**1.8**	0.41
100	2.1	1.6	2.55	1.05	1.85	0.2
150	2.6	1.65	2.50	0.98	1.8	0.15
200	2.1	1.72	**2.45**	0.8	1,79	**0.13**

TABLE 8.2: Measuring DBI index on models results

8.0.3 Visual representations

Visual representations are provided in Appendix A.

Each model was split by the K clusters, the most optimal number due to the DBI and Silhoutte score recommendations. For the visual representation each model Sentence vectors were projected to the 2D with a TSNE library. Representations are colored with respective clusters.

8.0.4 Summary

Despite the fact that frequency weightings models shows the best performance results they revealed to be not reliable one. Models that includes term frequency weightings allocates tends to picks clusters with not uniformly values amount distribution. (One/Several cluster captures most data while another clusters are very small). And this deviation scales even more after applying K-means Outliers filtering

Custom observation revealed that big clusters of frequency models contains sentences that with to be in a different sentences. (ex: "Hire junior BA" vs "Set up regular grooming sessions").

While little clusters are mostly connected to each other by having the same "Frequent Domain word".

Ex: Cluster 17 for the CRMS + ITF A.6 contains only the following records:

- 'We configured running tests in Docker using Selenium Grid'
- 'Ask DevOPs to set up Ngnx for Docker'
- 'Clarify the priority of Docker implementation'
- 'Figure out people in project who work closely with Docker'
- 'Assign tasks related with Docker to people who have intel based processor'
- 'The team is in a good spot migrating to Docker'

There are cases when frequency weighting may be useful, as an additional viewpoint. But it is not very good at the general splittings.

No weighting model works better with a clusterization approach. Clusters are split more eventually and Action Plans within them are mutually dependent. By manually exploration of cluster results was revealed that model works well for clusterization of the sentence general topics.

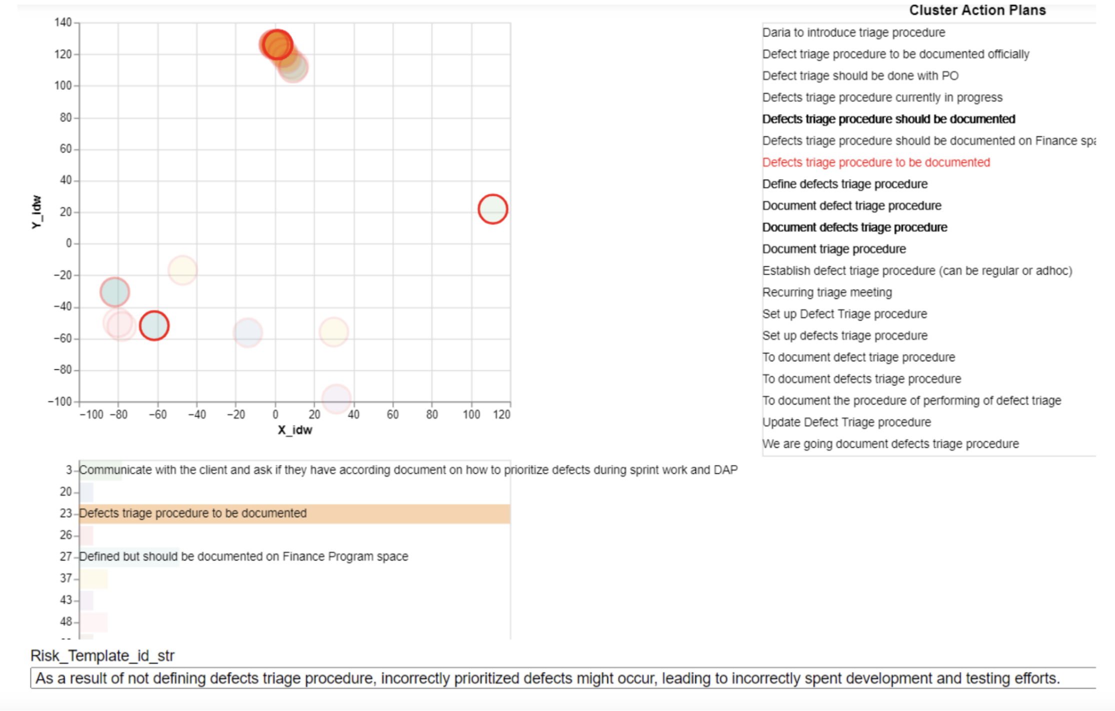
Outliers filtering approach lowers the interconnections between clusters and thus recommendations that appears at the model are more independent and intersects by the context similarity less.

So that, as a final model was selected **CRMS + Outliers** model A.2 .

8.0.5 Final Model Outcomes

After model was selected, its results are represented into the interactive way. The following properties may be explored from the chart :

- By selection the Risk Template automatic recommendations are shown with an ability to explore the details.
- Action items for the selected Risk are grouped by corresponding clusters and projected to the 2d with a TSNE decomposition.
- Final action items are shown only for the meaningful (valid) clusters. Cluster is considered as valid in case if at least 3 Action items from the Cluster were recorded.
- Final action items are set of action items with a closest euclidean distance to their cluster center (highlighted ones in Figure 8.1, Figure 8.2)
- By clicking on a datapoint, all action plans for the corresponding cluster are shown at the table. 8.1, Figure 8.2



Cluster Action Plans

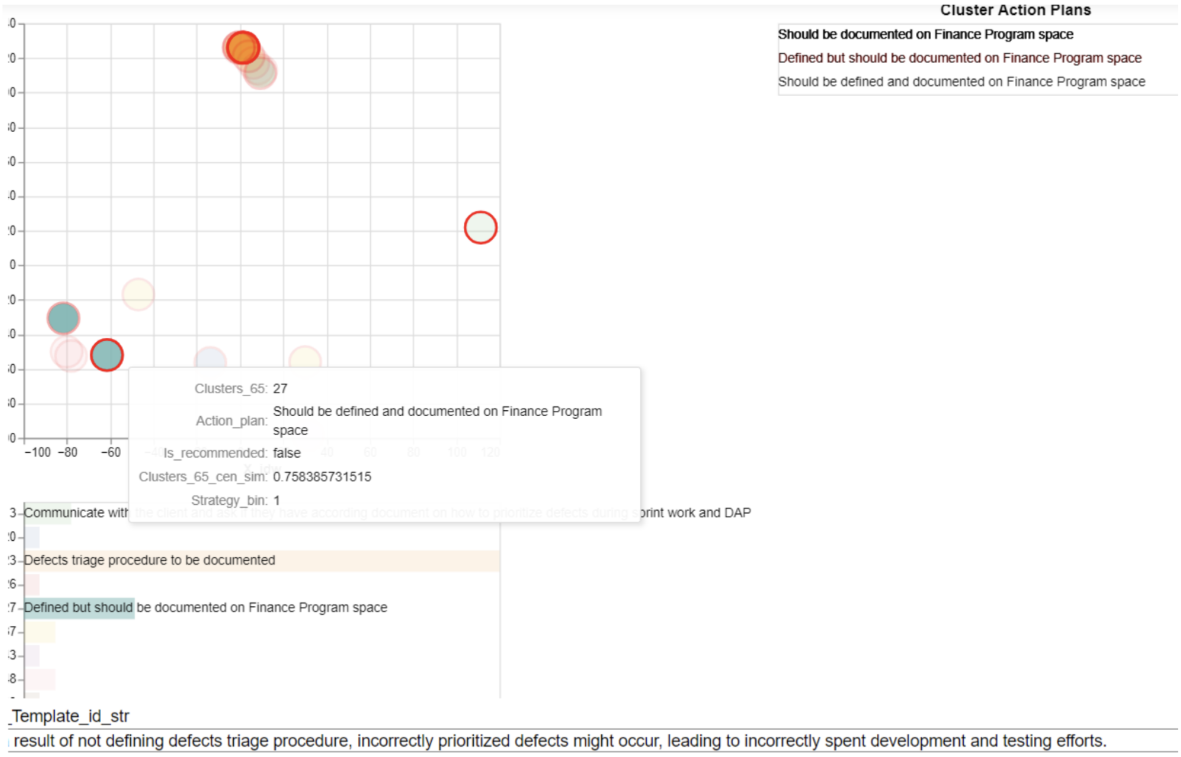
Daria to introduce triage procedure
Defect triage procedure to be documented officially
Defect triage should be done with PO
Defects triage procedure currently in progress
Defects triage procedure should be documented
Defects triage procedure should be documented on Finance spa
Defects triage procedure to be documented
Define defects triage procedure
Document defect triage procedure
Document defects triage procedure
Document triage procedure
Establish defect triage procedure (can be regular or adhoc)
Recurring triage meeting
Set up Defect Triage procedure
Set up defects triage procedure
To document defect triage procedure
To document defects triage procedure
To document the procedure of performing of defect triage
Update Defect Triage procedure
We are going document defects triage procedure

3 Communicate with the client and ask if they have according document on how to prioritize defects during sprint work and DAP
20
23 Defects triage procedure to be documented
26
27 Defined but should be documented on Finance Program space
37
43
48

Risk_Template_id_str

As a result of not defining defects triage procedure, incorrectly prioritized defects might occur, leading to incorrectly spent development and testing efforts.

FIGURE 8.1: Action items for the Cluster 23 are highlighted



Cluster Action Plans

Should be documented on Finance Program space

Defined but should be documented on Finance Program space

Should be defined and documented on Finance Program space

Clusters_65: 27

Action_plan: Should be defined and documented on Finance Program space

Is_recommended: false

Clusters_65_cen_sim: 0.758385731515

Strategy_bin: 1

Communicate wit[h] ... [s]print work and DAP

Defects triage procedure to be documented

Defined but should be documented on Finance Program space

_Template_id_str

result of not defining defects triage procedure, incorrectly prioritized defects might occur, leading to incorrectly spent development and testing efforts.

FIGURE 8.2: Action items for the Cluster 27 are highlighted

8.0.6 Further work for the model improvement

The following improvements can be suggested for the model optimization.

- Applying additional rules for the Recommended Action Plan selection. Based on the project specifics data and Risk details data.
- Applying the Constituency parsing to split the long sentences and avoid noise in clustering.
- Applying POS weighting for the Sen2Vec model with a highest weight assigned on the Word. So that Action Plans are additionally clustered based on the "Action Word".
- Revising the domain specifics words and their expansion. Substituting the unreadable abbreviations.

Chapter 9

Conclusions

Conducting corporate risk management is a complicated process, requiring a lot of focus of attention. As the human factor remains one of the biggest factors, jeopardizing the quality of the process and its outcomes, applying a data-driven approach greatly helps to achieve quality over quantity dynamics. We've seen that applying NLP algorithms can leave a positive impact on the business process, like risk management, and be a successful practice, maintained and developed further in outsourcing organizations.

We can say that goals and objectives are met and as result of the implementation of our model and the following results have been achieved:

- **Risks Templates coverage**. As a result, the 337 Risks Templates (30% of the total) are covered with Action Items recommendations and may be explored in the interactive chart;
- **Trimming controversial Action Items**. A number of Action Items that needed not to be reviewed per template decreased:
  - by 80% taking into consideration valid cluster exploration;
  - by 97% taking into account final action items.

  As an outcome, only the most important Action items are left and can be much easier analyzed;
- **Predicted increase of Action items quantity per risk**. Currently, the average number of Action items assigned per risk equals 1.23. By adding recommended predefined recommended Action items to the system this number should increase up to 0.7 points;

Not all the positive outcomes are visible immediately, some time will be required to rollout the changes and get project teams involved in trying out our suggestions.

To validate the performance of the model the following KPI should be measured in the future:

- **Time effectiveness** for the Action plan completion. Time difference between the time required to enter manual risks vs selecting predefined ones;

- **Satisfaction rate** of the usage of predefined Action Items as a separate feature. This KPI can be tracked in the feedback analysis on a project check;
- **Recommended Action Plan Usage %**. Measuring how often Predefined Action plans are used will let us know how relevant they are to the ongoing problems;
- Highlighting **Recommended Action Plan as useful %**. By measuring the percentage of selecting recommended suggested Action items as useful we can understand how relevant they were.

Appendix A

Visual representations

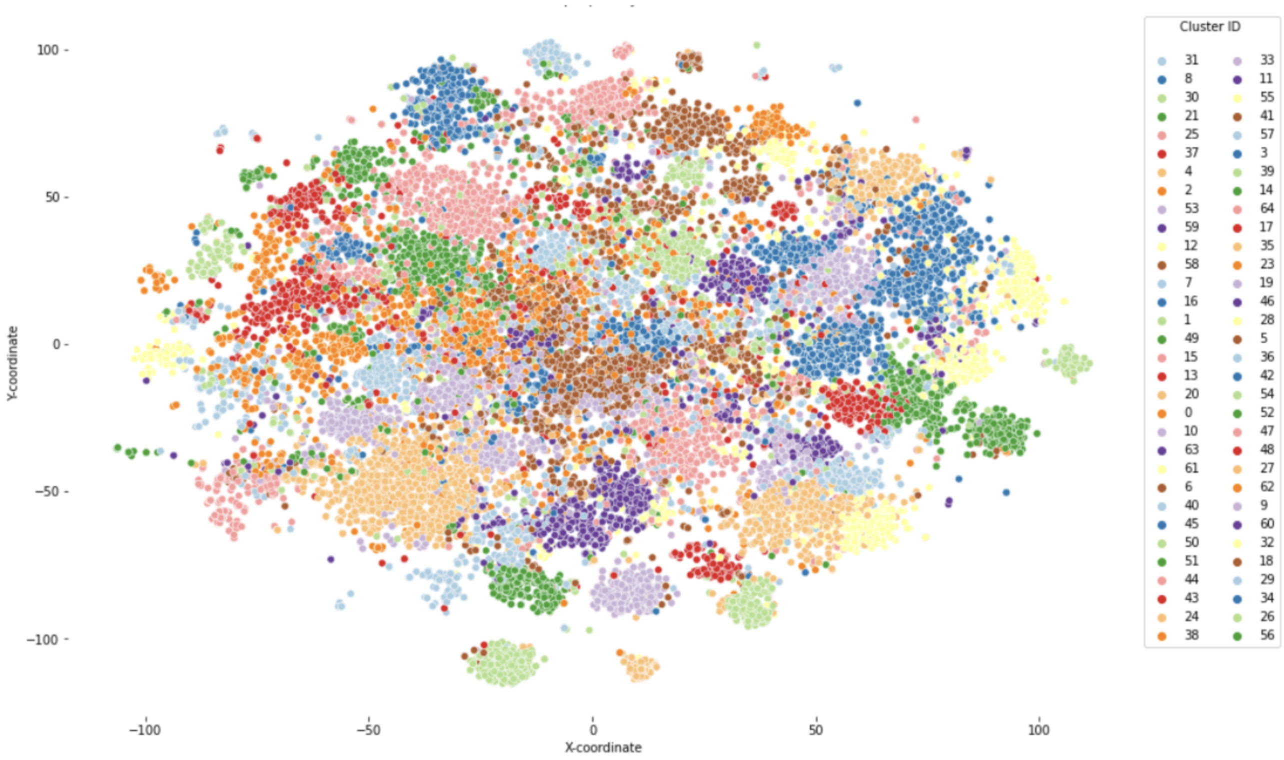


FIGURE A.1: CRMS, no weighting

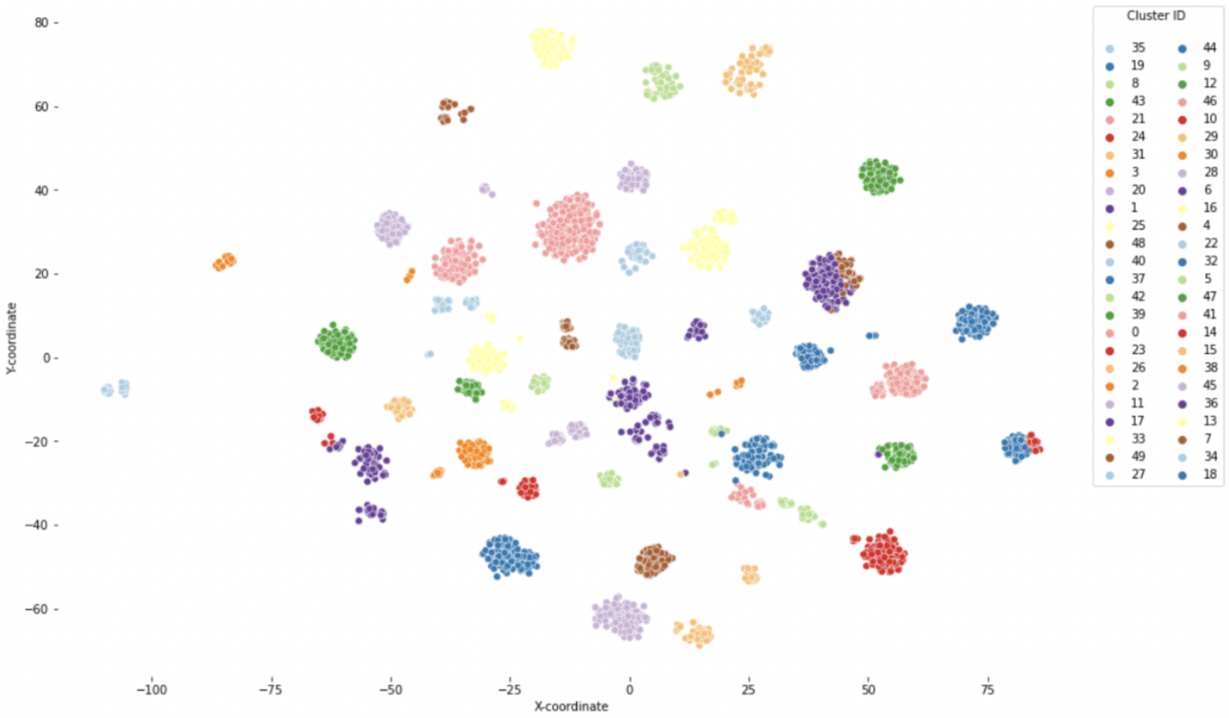


FIGURE A.2: CRMS + Outliers, no weighting

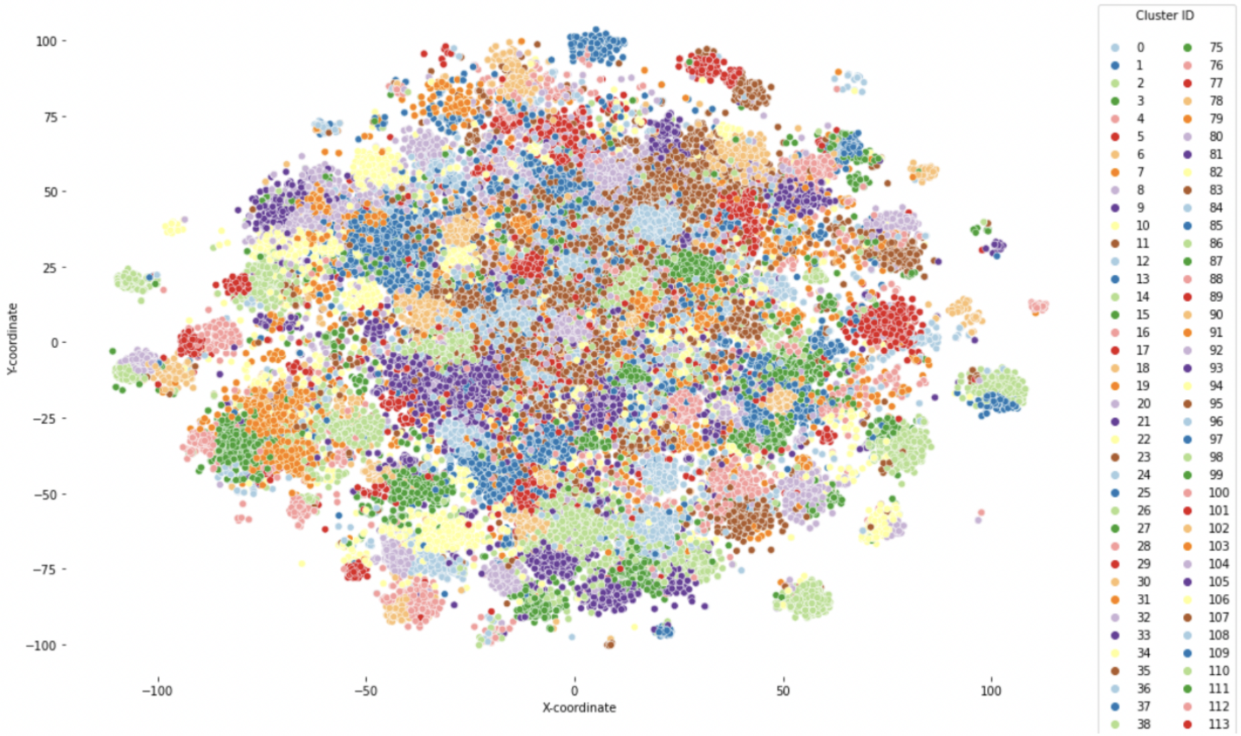


FIGURE A.3: CRMS + TF

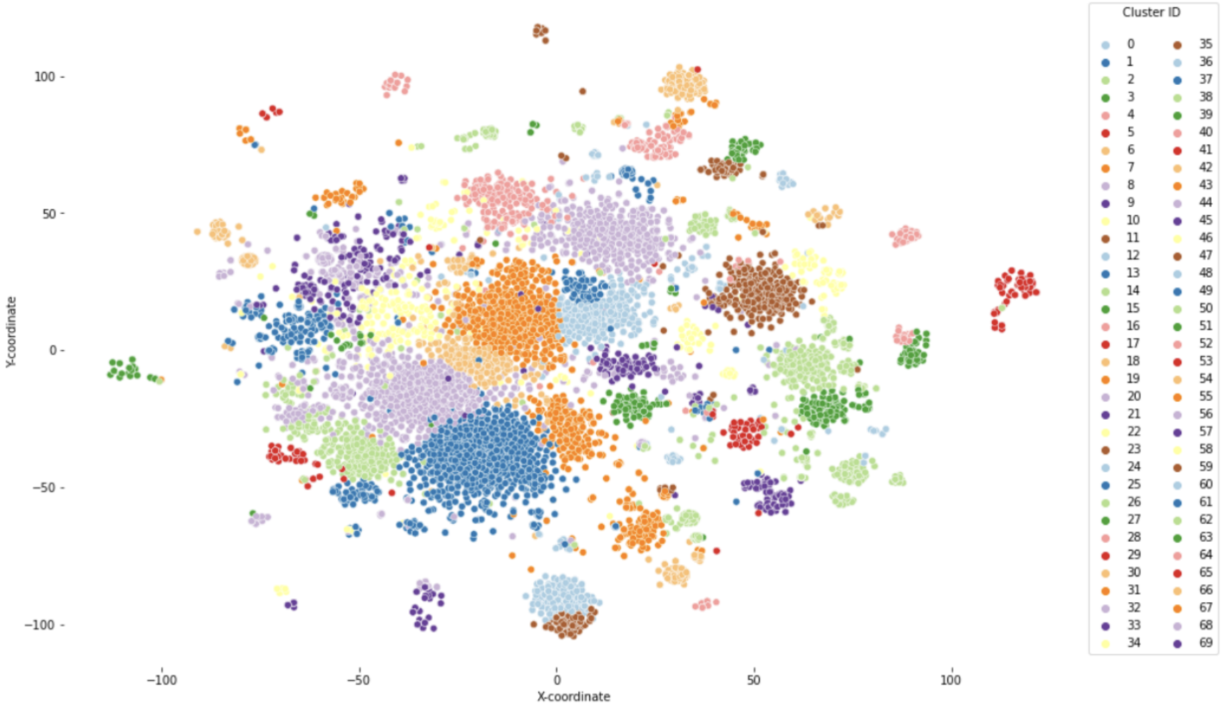


FIGURE A.4: CRMS + TF + Outliers

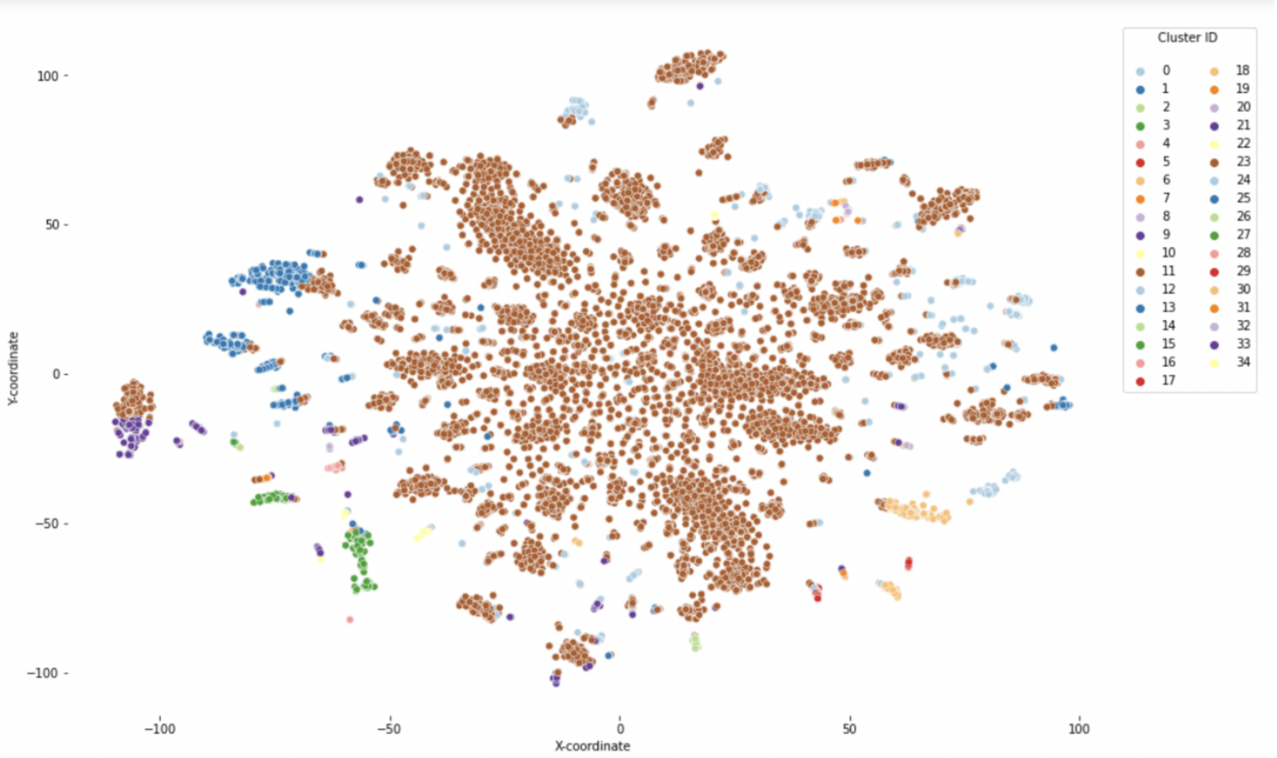


FIGURE A.5: CRMS + ITF

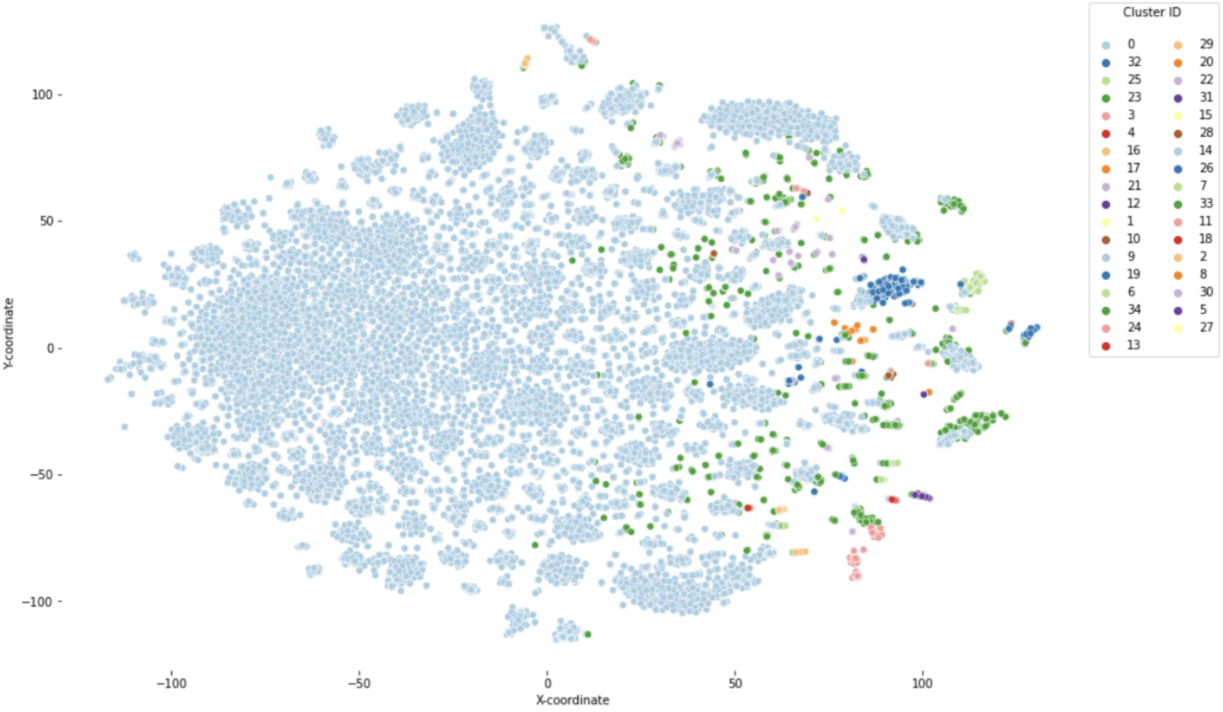


FIGURE A.6: CRMS + ITF + Outliers

Bibliography

Baarsch, Jonathan and M. Emre Celebi (2012). "Investigation of Internal Validity Measures for K-Means Clustering". In: ISSN: 2078-0966. URL: http://www.iaeng.org/publication/IMECS2012/IMECS2012_pp471-476.pdf.

Bruno Daniotti Marco Gianinetto, Stefano Della Torre (2020). *Digital Transformation of the Design, Construction and Management Processes of the Built Environment*. Milano: Fondazione Politecnice de Milano.

Hautamäki, Ville (2005). "Improving K-Means by Outlier Removal". In: DOI: 10.1007/11499145_99. URL: https://www.researchgate.net/publication/220809349_Improving_K-Means_by_Outlier_Removal.

Haw-Shiuan Chang Amol Agrawal, Andrew McCallum (2021). "Extending Multi-Sense Word Embedding to Phrases and Sentences for Unsupervised Semantic Applications". In: *arXiv preprint arXiv::2103.15330v2*.

J.-P.Tixier, Antoine (2015). "Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports". In: 2015.10, P10008. ISSN: 1742-5468. URL: https://www.sciencedirect.com/science/article/abs/pii/S0926580515002265?via%3Dihub.

Kuncheva, L.I. and D.P. Vetrov (2006). "Evaluation of Stability of k-Means Cluster Ensembles with Respect to Random Initialization". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28.11, pp. 1798 –1808. ISSN: 9181874. DOI: 10.1109/TPAMI.2006.226. URL: https://ieeexplore.ieee.org/abstract/document/1704835.

Lukas J. A. Clare, R. (2011). *Top 10 mistakes made in managing project risks. Paper presented at PMI® Global Congress 2011—North America, Dallas, TX. Newtown Square, PA: Project Management Institute.* URL: https://www.pmi.org/learning/library/mistakes-made-managing-project-risks-6239 (visited on 05/31/2022).

Palacio-Nino, Julio-Omar and Fernando Berzal (2019). "Evaluation Metrics for Unsupervised Learning Algorithms". In: DOI: 1905.05667v2. URL: https://arxiv.org/pdf/1905.05667.pdf.

Plaza, Enric (2001). "Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches". In: 2001.10, P10008. ISSN: 1742-5468. URL: https://www.researchgate.net/publication/225070522_Case-Based_Reasoning_Foundational_Issues_Methodological_Variations_and_System_Approaches.

Project Management Institute, website (2022). *What is Project Management?* URL: https://www.pmi.org/about/learn-about-pmi/what-is-project-management (visited on 05/31/2022).

Ratsiepe, Kaizer Boikanyo and Rashad Yazdanifard (2011). ""Poor Risk Management as One of the Major Reasons Causing Failure of Project Management." 2011 International Conference on Management and Service Science". In: DOI: 10.1109/ICMSS.2011.5999104. URL: https://doi.org/10.1109/ICMSS.2011.5999104.

Zhang, Aston et al. (2021). "Dive into Deep Learning". In: *arXiv preprint arXiv:2106.11342*.

Zou, Yang (2017). "Retrieving similar cases for construction project risk management using Natural Language Processing techniques". In: 2017.10, P10008. ISSN: 1742-5468. URL: https://www.sciencedirect.com/science/article/abs/pii/S0926580517303175?via%3Dihub.