Operational text-mining methods for enhancing building maintenance management

Marco Marocco and Ilaria Garofolo

Department of Civil Engineering and Architecture, University of Trieste, Trieste, Italy

ABSTRACT

Facility managers can significantly benefit from operational data, such as maintenance requests, stored in computerized maintenance management systems (CMMSs). This data is a valuable means to assess building performance and gain insights for preventive maintenance actions. However, databases are not always organized in such a way that allow undertaking analytics, therefore resulting in troubles when trying to generate useful information from raw data. This paper presents two methods based on a text-mining approach to extract valuable information from textual maintenance requests. The first method aims to extract the room identifier (ID) numbers where faults mainly occur, while the second one aims to identify the most problematic building elements and systems. The text-mining-based methods were tested by using a data set which contains 12,655 maintenance requests derived from a cluster of 33 buildings managed by the local administration of the Municipality of Trieste (Italy).

KEYWORDS

Text mining; maintenance requests; facilities management (FM); computerized maintenance management system (CMMS)

Introduction

The facilities management (FM) accounts for more than 60% of the building lifecycle for both cost and time (Wang et al., 2013). During this phase, operational data is widely regarded as crucial for diminishing FM expenses and providing value-added services, such as comfort and safety, remote monitoring and control of facility systems (Akcamete et al., 2010; Becerik-Gerber et al., 2012; Ignatov & Nørkj, 2019). In particular, detailed data on installed components and equipment stored in archives, along with inspections can support FM decision-making processes (Volk et al., 2014). In situ inspections allow FM personnel to diagnose and assess the condition of building elements at an exact point in time, but cannot be valuable for developing an exhaustive view of what the principal issues of buildings are and what buildings are mainly affected by failings over long time periods. In contrast, FM systems can integrate data and support analyses in order to present a comprehensive picture of building conditions. For instance, these tools can show what assets exist and keep trace of all actions taken, such as fixings, upgrades and replacements (Bortolini et al., 2016; Kensek, 2015).

A maintenance request is generated when there is a perception of a building element malfunction or when building components have a deficient performance (Bortolini & Forcada, 2020; Teicholz, 2004). During the operation and maintenance (O&M) phase, this data is usually stored in FM systems and can generate a general knowledge of building fault trends. Although the data extracted from CMMS data sets can be used to provide insights into building performance, FM systems are often organized in an unstructured manner, which leads to difficulties in analysing data (Gunay et al., 2019). Maintenance requests usually contain the textual descriptions of issues, which can include irrelevant details, but also miss important ones. However, this textual data can be helpful and worthwhile to glean insights, but generating invaluable and appropriate information from textual data is a challenging task (Bortolini & Forcada, 2020).

With the increasing interest in boosting the management of buildings, there is a great need of analysing data stored in FM databases to generate potential benefits for organizations. Thus, the research question of this paper is the following:

- How can facility owners improve the operational management of building maintenance if FM systems do not have a structured database that can properly store data?

This paper aims to provide two methods to support FM by generating useful insights from maintenance requests stored in a CMMS database, examine current limitations and propose future research directions.

CONTACT Marco Marocco marco.marocco2@phd.units.it

The proposed methods focus on investigating textual maintenance requests to conduct operational actions for preventive maintenance by extrapolating the room ID numbers where faults mainly occur and the most problematic building elements/systems. These methods are based on the development of text mining algorithms that allow the extraction of this useful information from data sets. Eventually, these methods were applied to the case study of the local administration of the Municipality of Trieste, where 12,655 maintenance requests derived from 33 buildings were analysed. The paper is structured as follows: the next section provides a literature review of building maintenance management, the third section focuses on the research methodology, the fourth section presents the case study and identifies limits and future directions and finally the conclusion is presented.

Literature review

Building maintenance management requires continuous information updates to maintain a high level of efficiency during the operational phase (Allen, 1993; Becerik-Gerber et al., 2012). Since building condition is based on the quality of maintenance management, comprehensive procedures and methods for investigating potential failings of building components and systems are needed (Lateef, 2009). Building information used to be stored in sheets of paper and/or file notes, but manual data management was not able to generate a proper method for information management (Lin et al., 2012), leading to irregularity for data management storage (Aziz et al., 2016; Corneli et al., 2019; Gursel et al., 2009; Kelly et al., 2013). To this end, several kinds of FM systems have been developed to help manage different data (Bortolini et al., 2016; Labib, 2004; Majerník et al., 2016).

Some techniques for examining building condition which often exploit data stored in CMMSs have been used by previous studies. Failure mode and effect analysis (FMEA) is an analysis technique which enhances risk management decisions by investigating failure modes and their effects (Schmittner et al., 2014). This method is based on a two-step process. The former involves a session of systematic brainstorming to identify all possible potential failure modes of systems, while the latter concerns a critical analysis of these failure modes, which considers some risk factors, such as occurrence, severity and detection, to prioritize the failure modes of systems (Liu et al., 2013; Vilarinho et al., 2017). Several studies depending on the failure mode prioritization method have been carried out in different applications (Chin et al., 2009; Franceschini & Galetto, 2001; Gargama & Chaturvedi, 2011; Kutlu & Ekmekçioğlu, 2012). Another technique for enhancing the processes of detecting failings and diagnosing their causes concerns the automated fault detection and diagnosis (AFDD). This process refers to the ongoing monitoring of system operations in order to prevent abnormal conditions and loss of service and guide planned maintenance decisions (Katipamula & Brambley, 2005; Zhang & Hong, 2017). Research studies pertaining to AFDD methods are categorized into three main groups including process history-based, qualitative model-based and quantitative model-based (Bruton et al., 2014; Fan et al., 2010; Kim & Katipamula, 2018; Provan, 2011). However, limitations in knowledge capture and data analysis are recognized when dealing with FM systems (Pärn et al., 2017; Vilarinho et al., 2017).

Recently, text mining-based methods have been developed to enhance maintenance management by analysing textual maintenance requests contained in FM systems (Bortolini & Forcada, 2020; Gunay et al., 2019). Text-mining concerns the discovery of unknown information by analysing large unstructured documents (Allahyari et al., 2017; Witten et al., 2011). Maintenance requests stored in CMMS encompass relevant information including dates, fault locations, fault types, fault symptoms, problem descriptions, actions taken to fix the problems, and causes (Mmelesi & Nwaigwe, 2020; Yang et al., 2018). The first application was carried out by Gunay et al. (2019), who focused on exploiting unstructured data contained in CMMSs in order to identify HVAC faults (Gunay et al., 2019). The used method allowed clustering sections of CMMS database which contain work orders about failures and avoid routine maintenance requests. As a result, the coexistence tendencies among terms of interested clusters, such as HVAC systems and terms used for describing modes of failure, were defined. Another text-mining approach was conducted by Bortolini and Forcada (2020), who analysed the conditions of the different systems of a building and the correlations between building characteristics and faults (Bortolini & Forcada, 2020). This analysis took into consideration how data including gross floor area, year of construction, type of building use and building property are linked to maintenance requests.

Nevertheless, in literature there are no publications that take into consideration the necessity to extract information from CMMSs to pinpoint building areas where failures more often occur and identify precisely flawed building elements and systems, leading to acquiring a comprehensive knowledge of building faults to conduct operational actions for preventive maintenance.

Methodology

The research methodology employs two methods to acquire a comprehensive knowledge of maintenance work orders (WOs). Firstly, a method to detect the identifier number of rooms where faults mainly occur was proposed. Secondly, a method to identify the most problematic building elements and systems was suggested. For each of the proposed methods, a text-mining approach was implemented to obtain information that can be useful to generate insights for building management from unstructured textual data contained in the CMMS. The process of identifying and extracting parts of texts is subjected to conditions that are dictated by the language rules of the analysed texts. This means that parts of texts are not always organized in the same way in different languages. On the other hand, ways of writing specific pieces of sentences are not influenced by the syntactic structure of single languages, but are shared by most of them. For instance, 'there is a fault in room 44' in English, 'c'è un guasto nella stanza 44' in Italian, 'il y a un défaut dans la chambre 44' in French and 'hay una avería en la habitación 44' in Spanish collocate the word 'room' next to the number of the room. Databases are not always highly organized, but owing to previous decisions that organizations make when they develop or buy their maintenance management platform, they are structured in a way that data cannot generate useful information when integrated. To this end, organizations that are in possession of such unorganized data have to find other ways to efficiently analyse their databases. For instance, as shown in Figure 1, room and element ID numbers might be omitted, leading to difficulties in extracting this data from the description of maintenance requests.

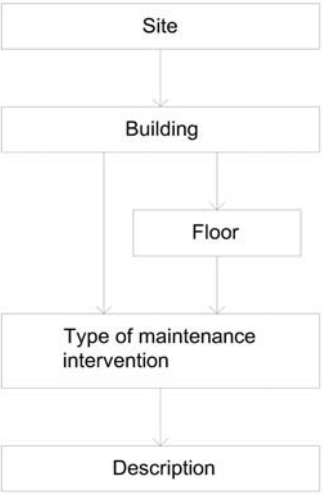

Figure 1. Example of a database structure.

To this end, text-mining algorithms can be a solution. This technique can analyse textual data which is contained in CMMS databases to obtain information about specific warnings of maintenance requests. The process of analysing these warnings begins with exporting the maintenance requests from the database and generating a data set with an extension, such as.xlsx or.csv, which can be read by advanced tools of programming, such as Jupyter Notebook. These tools exploit programming language scripts to modify and exclude unnecessary parts of text, as shown in Figure 2. Firstly, a list of work order descriptions is identified and selected from the database. Secondly, for each description, the text is broken into single words and unnecessary punctuation marks and spaces are removed. Then, all the words are converted into lower case and added to a list of words. This process is repeated until the

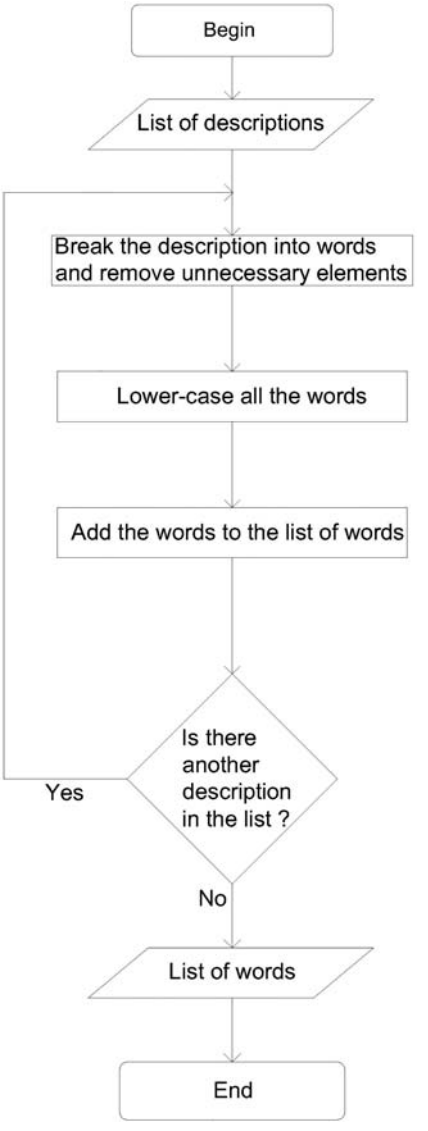

Figure 2. The process of generating the list of words.

last description of the list is analysed. The list of words can contain both words and numerals and preserves the original order and sequence of words as they were in the sentences. At the end of the process, an extensive list of words is created and ready for subsequent elaborations.

Identifying room ID numbers

The proposed method, which is shown in Figure 3, aims to identify room ID numbers from textual data in order to allow FM personnel to enhance maintenance management by shrinking and limiting the area where faults usually occur. Before starting the process, an empty list which will store room ID numbers is created and a set of secondary variables which are crucial to generate pieces of sentence by keeping string values are defined. This method is composed of four phases including adjusting numeric words, defining variables, extracting room ID numbers and setting new rules for variables. This method is repeated in a loop cycle as many times as the quantity of words that composes the list of words previously generated.

The first phase standardizes the formulation of words, as people in charge of reporting issues in buildings use different expressions. In particular, this step takes into account each word and checks whether it starts with a zero. If the Boolean validation produces a positive result, the initial letter of the word is removed, or else the word is not modified. This allows standardizing the ways representatives[1] write the room ID numbers.

The second phase concerns the statement of a series of principal variables that are used for building chunks of sentences. The number of variables depends on the number of different logic structures and vocabulary compositions identified for describing the room ID numbers. In the former case, the position of specific words is different from one sentence to another, while in the latter case, words that carry a particular meaning can be written in different forms. Five representative cases that can describe the logical sequences were identified. For each of these cases, a principal string variable composed of two or more secondary string variables was defined. As a result, complex and composite pieces of phrases can be created. As well as these main variables, a subsidiary one that represents whether the analysed word is or is composed of a number was defined, too.

The third phase exploits the previous steps to add one of the primary variables above described to the list of room ID numbers after processes of validation. Thus, five processes of validation, which take into account different syntax structures and several synonyms to write 'room', were applied by following an exact order. These processes were developed in such a way that one differs from another unequivocally, consequently enabling unique choices in precise situations. This means that two validation processes cannot be involved at the same time, therefore avoiding multiple extractions of the same information.

The fourth phase is a crucial part of the algorithm, which allows retaining specific words as values of the secondary variables for successive extractions. Five conditional processes of verification were employed to alter the values of the secondary variables, which are used in the consecutive loop cycle that restarts from the beginning of the four phases of the method. Each of these five conditional processes checks whether a precise variable possesses or not an exact value. Since the word 'room' can be written in various manners in a text, not only were synonyms considered, but also possible abbreviations. As a result, the considered secondary variable acquires a new value or maintains its value according to a specific verification code. Similar to the second phase of the method, also this set of conditional verification processes is univocal, which leads to assigning a certain value to an exact variable at a specific point of the entire loop cycle.

In order to strengthen the code, potential human mistakes in writing texts were considered. For instance, while it is less probable that orthographic mistakes are overlooked, when typing, it is possible that workers neglect typing errors, such as double spaces among words. In addition, due to a lack of attention and unusual ways of writing maintenance requests, a threshold of word frequency occurrence was defined as a limit for acceptance of results. The whole process keeps on extracting pieces of phrases until the last word is analysed. In the end, a list of room ID numbers is extracted from raw data.

Identifying the most problematic building elements

Although identifying room ID numbers is a useful step to better analyse maintenance problems of buildings, there is also a great need to discover what building components and systems are the most problematic. However, since the database is not organized in order to store element ID numbers of failing elements, workers in charge of warning malfunctions and faults are not used to referring to a specific ID number, but they use simple sentences to describe these deteriorated elements. To this end, the proposed method, as shown in Figure 4, aims to identify the most frequent words from textual data in order to pinpoint the primary elements that present problems.

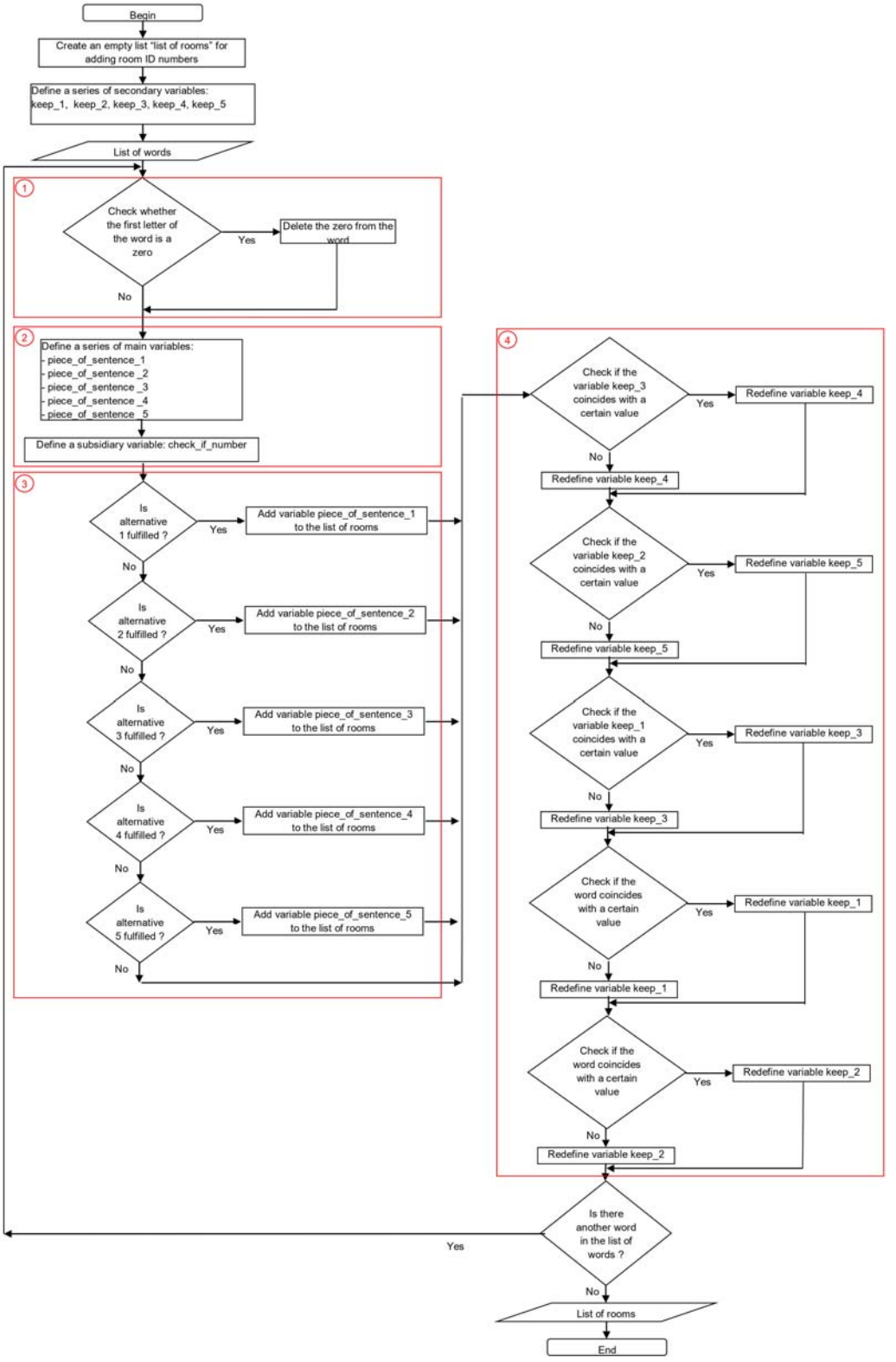

Figure 3. The process of identifying room ID numbers.

This method is composed of a few steps including creating a Bag of words, comparing the Bag of words with a list of stop words[2] and then updating the Bag of words. The method begins with adding each word of the list of words previously generated at the end of the process shown in Figure 2 to a Bag of words. In this process, the Bag of words represents a data frame composed of words which is organized in descending

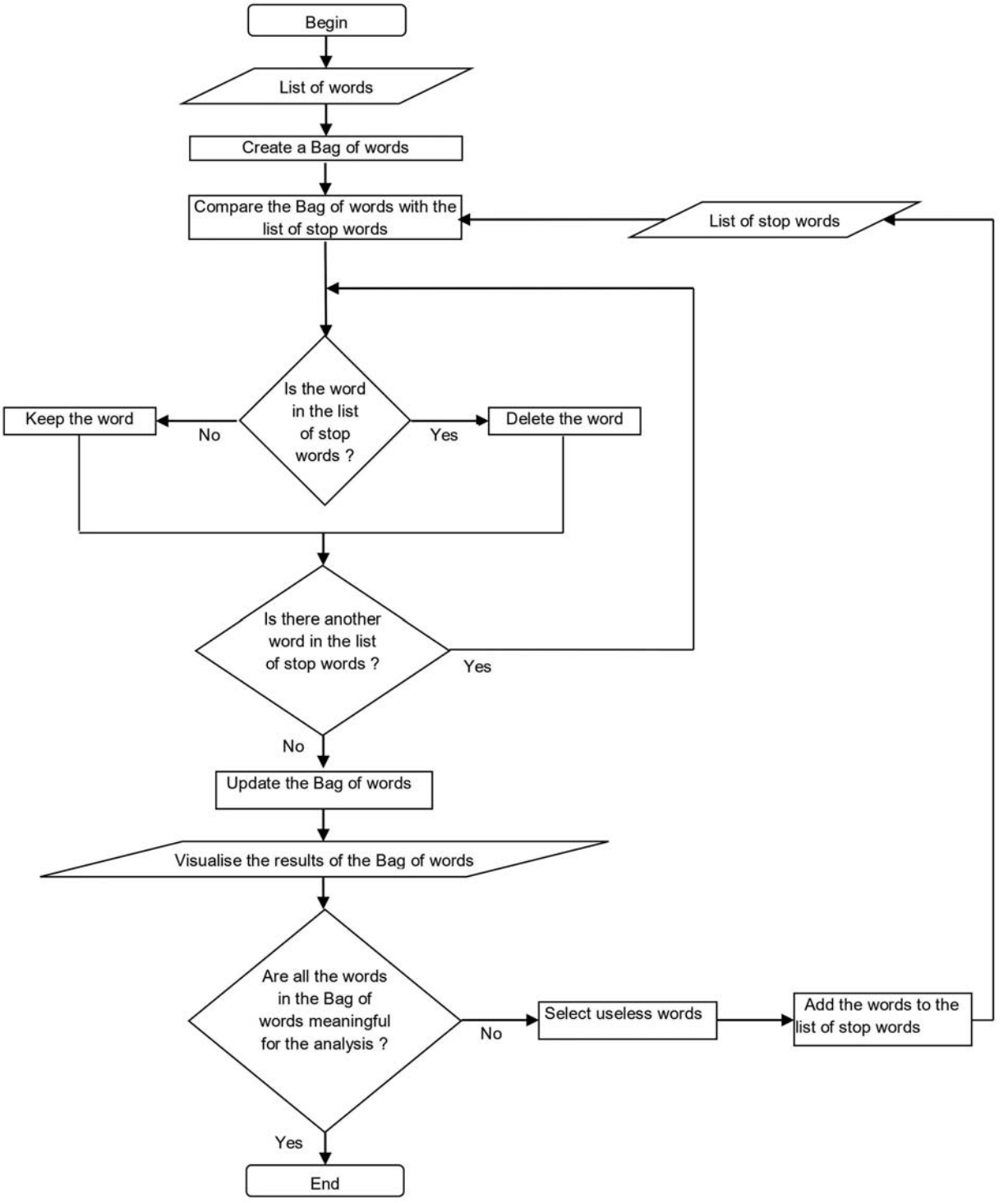

Figure 4. The process of identifying the most repeated words.

order where each word has an absolute value that derives from counting how many times it appears in the initial list of words generated at the end of the process shown in Figure 2. In the second step, the Bag of words is compared with the list of stop words. Before starting this step, there is a need to create the list of stop words, namely a list of words that do not refer to building elements and systems. Next, any word of the Bag of words that is contained in the list of stop words is removed from the Bag of words because it is not useful to identify issues in buildings, while the others are preserved. Afterwards, the components of the Bag of words are updated and then sorted again in descending order. Eventually, the results allow

classifying words to give major attention to those which have the highest frequency.

Since creating on the first try a list of stop words with all the unnecessary words, which must be removed from the Bag of words, is quite challenging, this is an iterative process. Thus, when the Bag of words is sorted, the most repeated words are assessed and those which are not considered interesting for the analysis are selected and added to the list of stop words. This means updating the list of stop words to allow restarting a new cycle of the method with a major number of common words regarded as noisy data removed.

As well as identifying single words that can specify issues, pairs of words can often describe better what the problem is. This is why a pattern to extract pairs of words that can have a more meaningful value for stakeholders was developed. Similar to the process for room ID number identification, a process to gather meaningful single words to create pairs of words is shown in Figure 5.

This method exploits an iterative process that begins with selecting one word from the list of the most repeated words, namely the Bag of words. This selected word is regarded as the reference for creating a valuable pair of meaningful words, which means that at the end of the cycle another word is added to the referenced one. Similar to the method of identifying room ID numbers, an empty list for adding the pairs of meaningful words is created and secondary variables are defined.

Next, for each of the selected words of the Bag of words, a sub-loop cycle is created, where the other word which will generate the pair of meaningful words is extracted from the initial list of words generated at the end of the process shown in Figure 2. Inside the sub-loop cycle, the process consists of three phases, including defining principal variables, extracting pairs of meaningful words and setting new rules for variables. The first phase states the main variables, which are composed of two or more secondary string variables, used for generating the pairs of words. The number of variables depends on the number of different logic structures and vocabulary compositions identified for describing the pairs of meaningful words specifically. A representative case that can describe the logical sequence was identified. The second phase exploits the previous step to add the primary variable to the list of pairs of meaningful words after a process of validation. The process of validation, which takes into account a specific syntax structure, was developed in such a way that pairs of words can be retrieved only in a particular situation, avoiding erroneous extractions. The third step alters the value of the secondary variable by retaining specific words as value for successive extractions.

This process is repeated in the sub-loop cycle as many times as the number of words that composes the initial list of words generated at the end of the process shown in Figure 2. As a result, a list of pairs of words is extracted from raw data. Eventually, the whole loop cycle is reiterated as many times as the number of words in the Bag of words.

Case study

The case study examines the maintenance management of the facilities that are managed by the local administration of the Municipality of Trieste (Italy), as shown in Figures 6 and 7. The buildings that are under this local administration control are deployed around the city and surrounding areas and include schools, churches, residential houses, public toilets, nursing homes, libraries, museums and offices. Similar to other public administrations, the strategy of the local administration of the Municipality of Trieste is based on a mix of outsourcing and in-house processes. With reference to the maintenance service, this service is outsourced, namely the management is entrusted to a third party, which provides the platform to manage the maintenance of buildings and the document management of WOs and conducts repair works on site. Data regarding maintenance requests is stored in different databases according to different types of service contracts. In this particular situation, the number of cases, namely the maintenance requests, derives from 33 buildings including buildings for assistance, museums and offices, as they are part of a unique contract. This set of buildings was chosen as test subject because the Municipality of Trieste regarded data contained in this database as high priority data. As shown in Table 1, both museums and offices account for nearly 40% of the analysed buildings, while the remaining 20% are assistance buildings.

The data set consists of 12,655 maintenance requests gathered from the local administration's CMMS for which building maintenance works were conducted. The period of analysis includes requests submitted between 2013 and 2020 by end users and the FM team through the maintenance platform linked to the CMMS database. The information gathered is restricted to the fields of input that the maintenance platform provides. These fields include request code, date, requester (name, phone number and e-mail), the location of the problem by specifying site, building and floor, the category of problem type by using predefined labels, the description of the problem, the priority of intervention, supplier for the service and a field for additional notes.

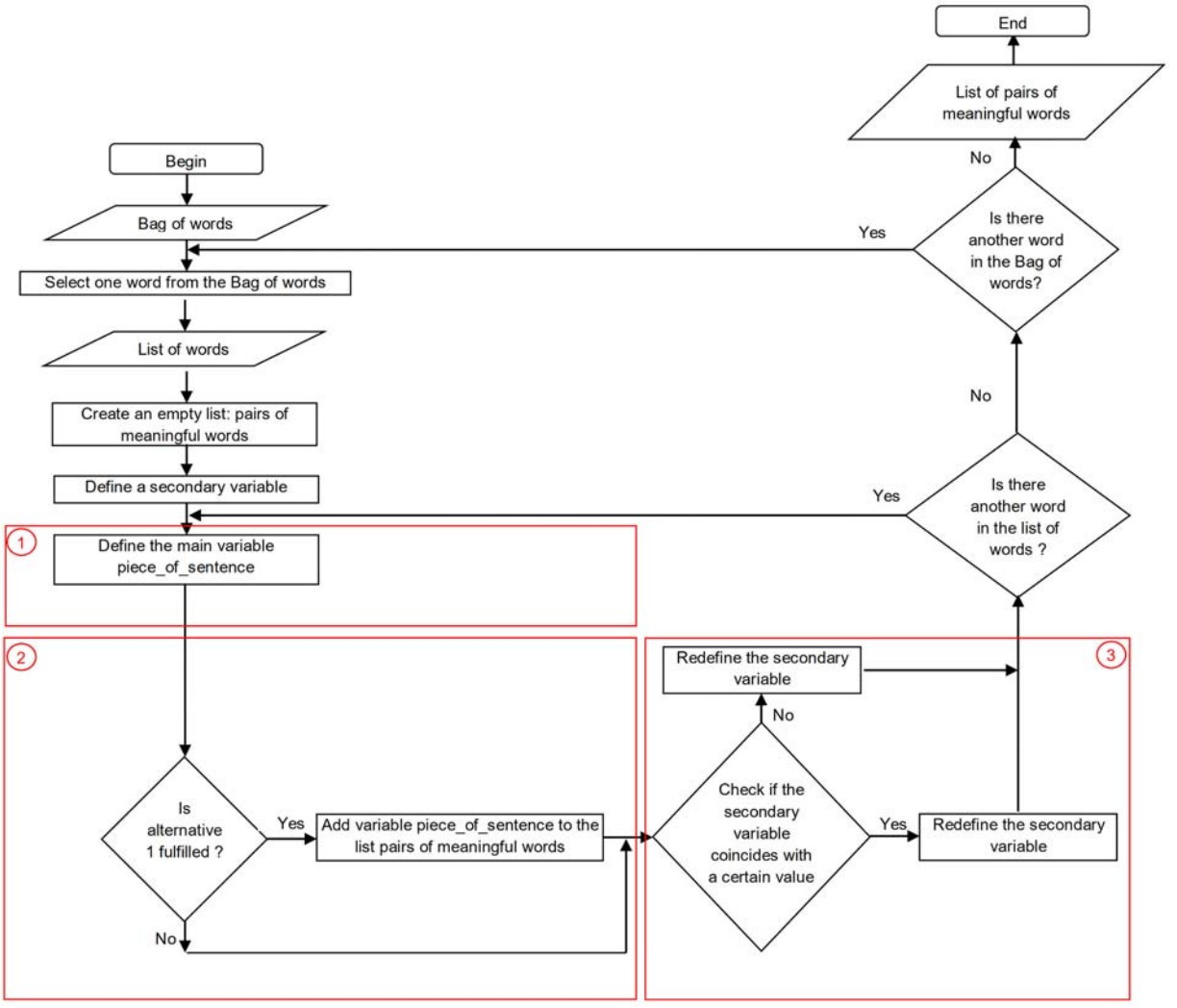

Figure 5. The process of identifying pairs of meaningful words.

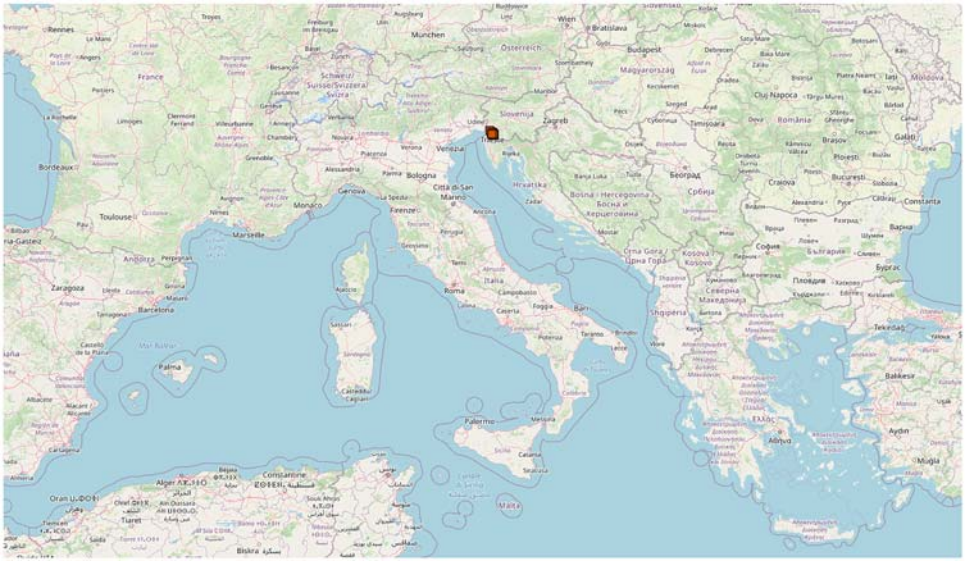

Figure 6. Pinpointing Trieste in the map of Italy.

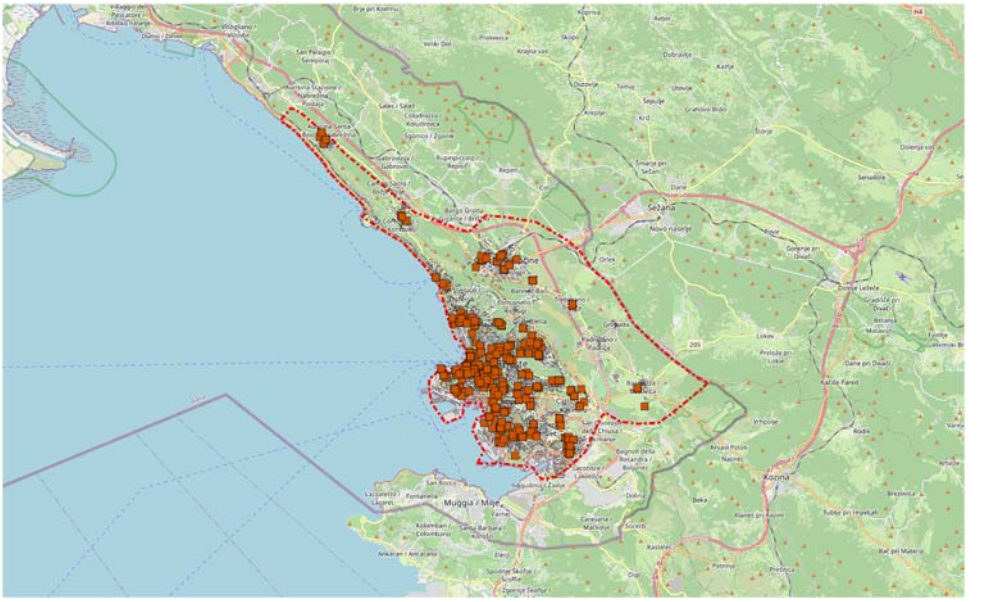

Figure 7. The buildings managed by the local administration of the Municipality of Trieste.

In order to compare buildings according to their typology, the distribution of building maintenance requests was normalized with respect to their volume, as shown in Figure 8. The chart shows a significant quantity of maintenance requests for buildings for assistance, while a less remarkable importance for offices and museums. It appears that since buildings for assistance aim to help and accommodate several people, this may cause repeated WOs in buildings.

Table 1. The main characteristics of the analysed buildings.

Building	Typology of building	Volume (m^3)	Number of cases
0165	assistance	27,275	1737
0260	offices	40,733	1171
0733	assistance	11,892	683
0261	offices	23,885	603
1077	museums	19,194	577
0277_1	offices	20,878	508
0387	assistance	19,295	492
0710	museums	19,716	452
1119	offices	29,050	447
0281	museums	40,306	429
0723	offices	66,254	409
0048	offices	16,771	352
0262	offices	9450	332
0622	museums	8030	319
0282	museums	39,540	298
0545_1	museums	5483	278
0289_0484	museums	7970	278
0545_2	museums	38,910	267
0439_1	assistance	8562	255
0732	offices	8520	253
0729_1	museums	2740	244
1105	offices	1662	221
0726	offices	6136	218
0591	museums	15,870	212
0439_2	assistance	12,329	203
0574	museums	28,398	193
0236	offices	35,900	187
1055_1	museums	3037	184
0717	assistance	2130	182
0751	assistance	6352	174
0249	offices	1500	171
0718	assistance	2130	163

Based on an internal organization of problem types adopted by the local administration of the Municipality of Trieste, the categories of problem types that were analysed include: electrical maintenance, metalwork/carpentry maintenance, water/sanitary maintenance, special installations maintenance, construction maintenance, external areas maintenance, elevators maintenance, sewer maintenance and fire extinguisher/hydrant maintenance. The three most prevalent typologies of intervention, namely electrical, metalwork/carpentry and water/sanitary, account for nearly 80% of the total of maintenance interventions, as shown in Table 2.

Eventually, analysing the priorities of intervention, it is noticeable that routine maintenance requests account for 60% of the total maintenance requests, urgent maintenance requests account for 29% and extreme urgent maintenance requests account for 11%. The distribution of maintenance request priorities of the analysed buildings is shown in Figure 9. For each building shown along the x-axis, the total number of maintenance requests, which are divided in routine, urgent and extreme urgent, is shown using a stacked bar chart. It is worth noticing that buildings 0260 and 0165 have received nearly double and triple maintenance requests compared to the third most problematic building, respectively.

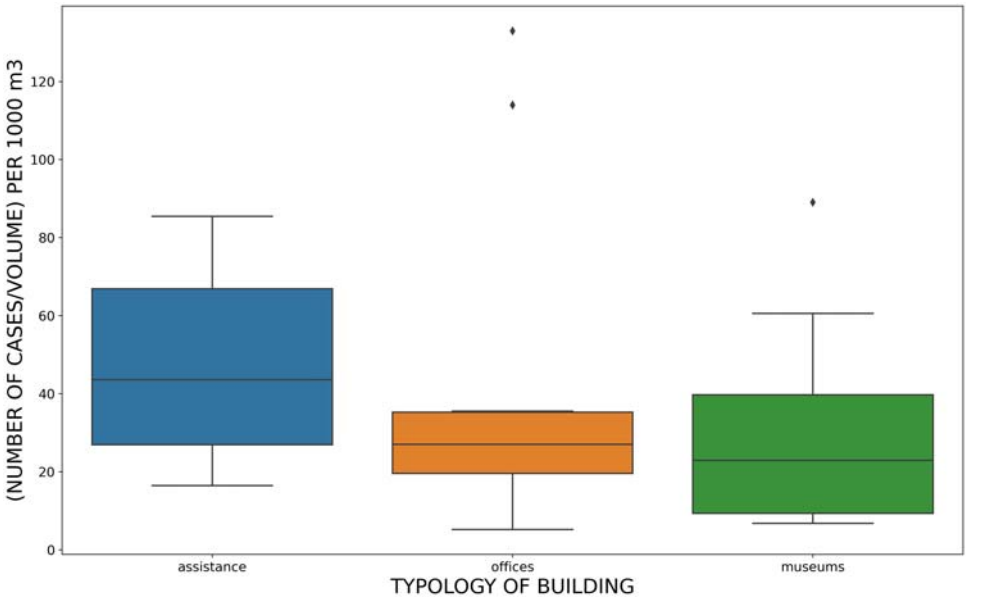

Figure 8. The distribution of maintenance requests per 1000 m^3 with respect to building typologies.

Results for room ID number identification

The first part of the research focused on developing a method that can extract the ID number of rooms where faults mainly occur. Before beginning the analysis, all the maintenance requests pertaining to the types of intervention 'External Areas' and 'Elevators' were excluded, as they cannot contain room ID numbers in their maintenance request descriptions. Thus, the analysed data frame for this specific investigation contained 12,233 maintenance requests. Three kinds of analysis were conducted: a general analysis, an analysis of the single types of intervention and a detailed analysis taking into account a specific building.

The first analysis was conducted to have a comprehensive view of the total amount of identified rooms. As a result, 3939 room ID numbers were identified. Comparing this number with the overall number of maintenance requests, the method identified room ID numbers for 32.2% of the total number of descriptions.

The second investigation analysed maintenance requests of all buildings taking into account the specific types of intervention. Results showed that the percentages of room ID numbers found for the type of intervention compared to the total number of maintenance requests for the specific type of intervention were: 43.8% for the water/sanitary maintenance, 31.2% for the electrical maintenance, 35.3% for the metalwork/carpentry maintenance, 33% for the construction maintenance, 9.1% for the special installations maintenance, 13.6% for the sewer maintenance and 10.5% for the fire extinguisher/hydrant maintenance, as shown in Figure 10.

The third analysis took into consideration the maintenance requests of a specific building so as to assess the method precisely and whether incoherent results occur. The building chosen was the 0165, a five-storey building, where rooms are numerated by hundreds, namely room ID numbers of the ground floor are within a range between 1 and 99, room ID numbers of the first floor are within a range between 100 and 199 and the other storeys have the same numeration pattern. Thus, checking whether some results are inconsistent can be conducted in two ways.

The first way checked the room ID numbers of the overall building. In this case, room ID numbers can be within a range of value between 1 and 499. Results showed that none of the room ID numbers extracted had an incoherent value compared to what expected, as shown in Figure 11.

The second way checked the room ID numbers for a specific floor. The fourth floor was chosen as a reference,

Table 2. Problem type categories.

Type of intervention	Number of intervention	% of the total	% Routine	% Urgency	% Max urgency
Electrical	4473	35.3	70.6	22.2	7.2
Metalwork/ Carpentry	2931	23.2	57.6	30.1	12.3
Water/Sanitary	2487	19.7	60.8	31.2	8.0
Special Installations	1241	9.8	39.1	36.8	24.1
Construction	837	6.6	56.0	32.7	11.3
External Areas	268	2.1	46.3	37.3	16.4
Elevators	154	1.2	47.4	35.7	16.9
Sewer	140	1.1	25.0	42.9	32.1
Fire Extinguisher/ Hydrant	124	1.0	44.4	45.1	10.5
Total	12,655	100			

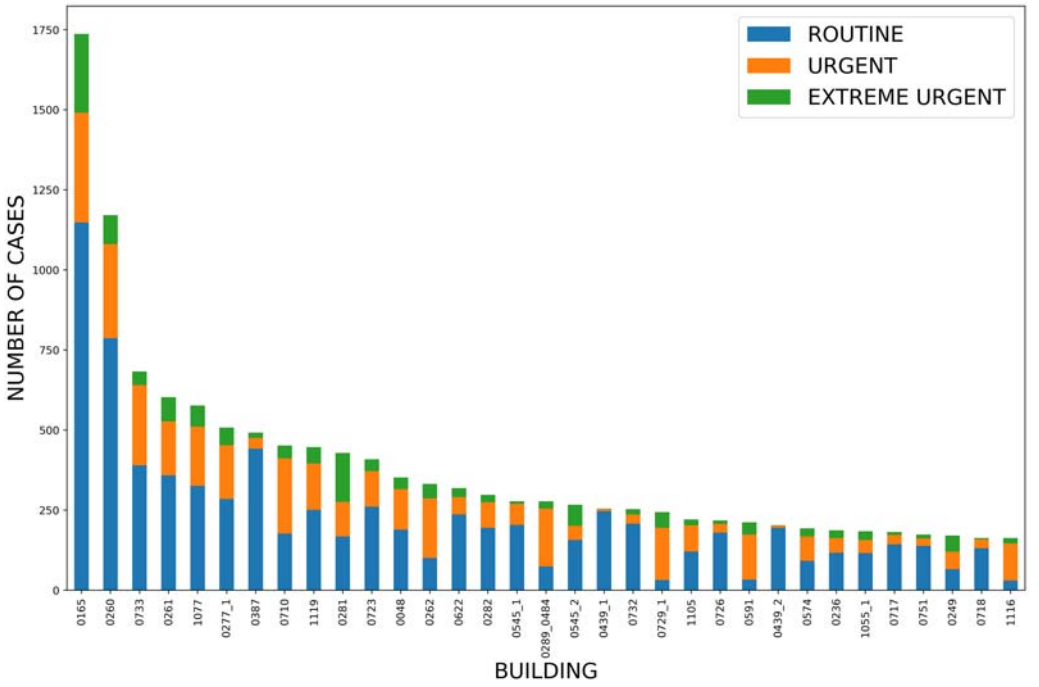

Figure 9. The analysis of intervention priorities.

as any kind of numbers in descriptions that do not concern with the room ID numbers, for instance the quantity of windows broken, are quite difficult to reach a value between 400 and 499, therefore allowing a valuable test. Results demonstrated that all room ID numbers extracted had a value included in the predefined range, as shown in Figure 12.

Results for the most problematic element identification

The second part of the research focused on generating a method that can acquire the most repeated words in the maintenance requests for discovering what building components and systems are the most problematic. Since the analysis included all the potential faults, the analysed data frame for this specific research contained the whole data set, namely 12,655 maintenance requests. To better analyse the results, Figure 13 displays the occurrence value for each of the most repeated words shown along the x-axis. The results showed that the two most repeated words were terms related to electrical issues, such as *neon light*, which was cited 3312 times, and *light bulb*, which was cited 1948 times. The third most quoted word referred to water/sanitary problems, such as *wc*, which was cited 1659 times, namely less than half of the word *neon light*. The other words that were considered relevant were part of a range of frequency from slightly below 1000 times to 250 times.

After identifying single words that can specify the issues, the analysis focused on the pairs of meaningful words that can describe better what the problems are by exploiting the words found in the previous analysis. The considered words included *wc*, *window* and *door*, namely the three most repeated words with general meaning so that they could be combined with other words to generate terms with specific significance. The results of the meaningful pairs of words are shown in Figures 14–16, where only the results above a threshold of fixed frequency, namely 10% of the total amount of extracted pairs of words for each analysis, were considered in order to avoid meaningless outcomes derived from misleading ways of writing. The results revealed that the most problematic fault related to water/sanitary problems and linked to the word *wc* was toilet seats, which occurred 307 times, namely more than twice of the second most frequent issue. The other failings included basins, flushes, drains, neon lights, light bulbs and sinks. With reference to the words linked to *window*, glass and handles problems were the most significant ones, while the other words encompassed issues connected with blinds and frames. Eventually, problems related to *door* consisted of locks, closers, handles, bells and glass.

Discussion

Theoretical implications

Similar to other research on this topic (Bortolini & Forcada, 2020; Gunay et al., 2019), this study presents results

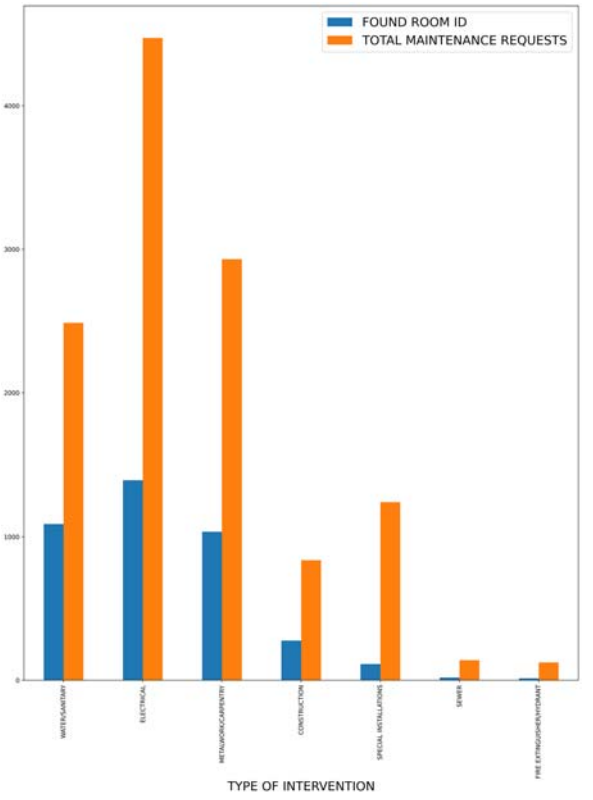

Figure 10. Room ID number percentages according to the type of intervention.

from a real case study and contributes to the body of knowledge regarding the extraction of valuable information from CMMSs by using text-mining algorithms.

Aligned with previous studies found in literature, this research overcomes issues identified by (Pärn et al., 2017), which pointed out a lack of practical use regarding most of the data collected from computerized systems. To this end, this study proposes two methods to allow exploiting unused data, such as text records, in order to provide an exhaustive view of what the principal issues of buildings are and what buildings are mainly affected by failings. The proposed methods also support the problem of insufficient level of data usability identified by (Tretten & Karim, 2014). According to Tretten and Karim (2014), information stored in the FM system needs to be easily analysed to plan and conduct maintenance tasks. Along these lines, the two proposed methods support effective building maintenance management by generating useful information from raw textual data. Furthermore, this study also addressed issues such as improperly standardized and organized maintenance requests pointed out by several research (Becerik-Gerber et al., 2012; Federspiel et al., 2003; Vilarinho et al., 2017), as this study focuses on developing methods to precisely identify operational faults in buildings by using unstructured data contained in CMMSs. Considering data as fundamental for supporting the decision-making processes for maintenance management (Lateef, 2009), this work also aligns with Lateef (2009) because it enhances the performance of building maintenance management by using a data-driven approach instead of relying on hypothetical and experience-based systems.

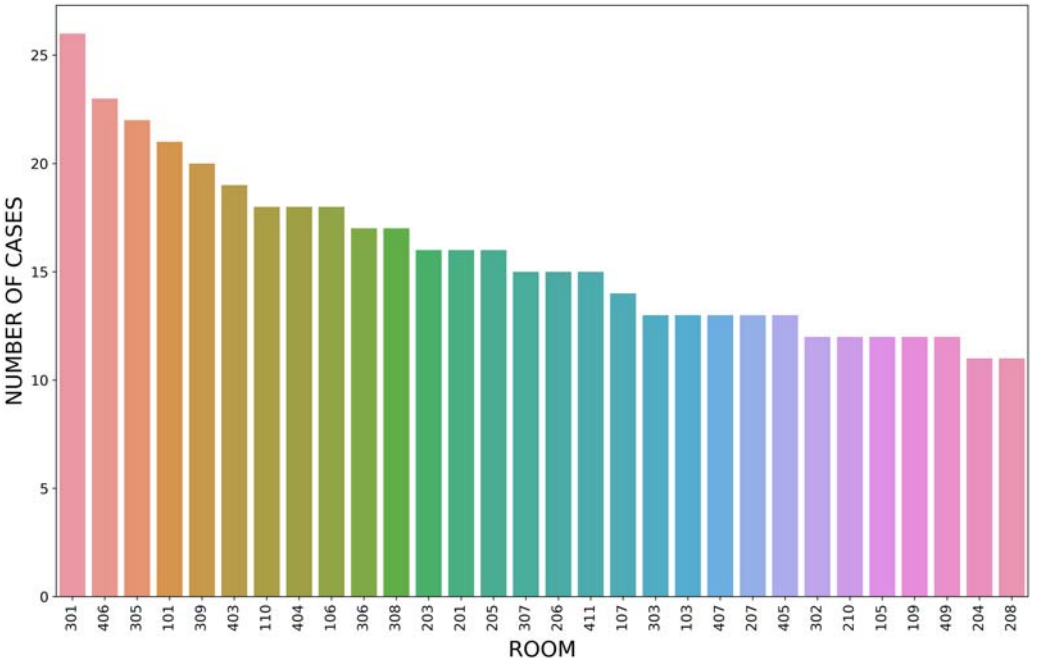

Figure 11. Room ID numbers of the building 0165.

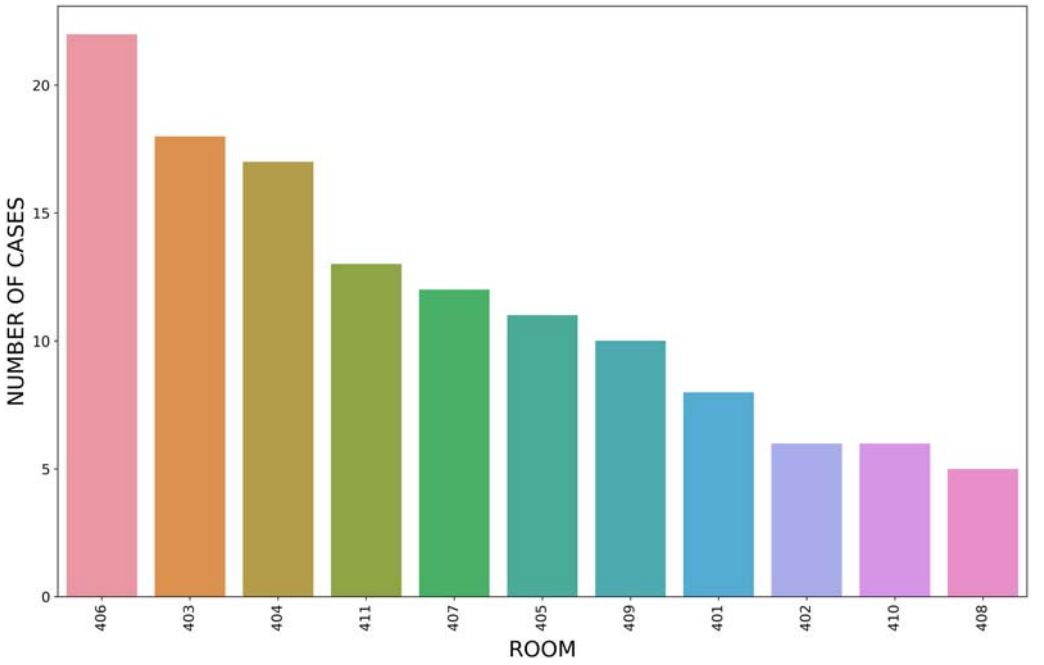

Figure 12. Room ID numbers of the floor P4 of the building 0165.

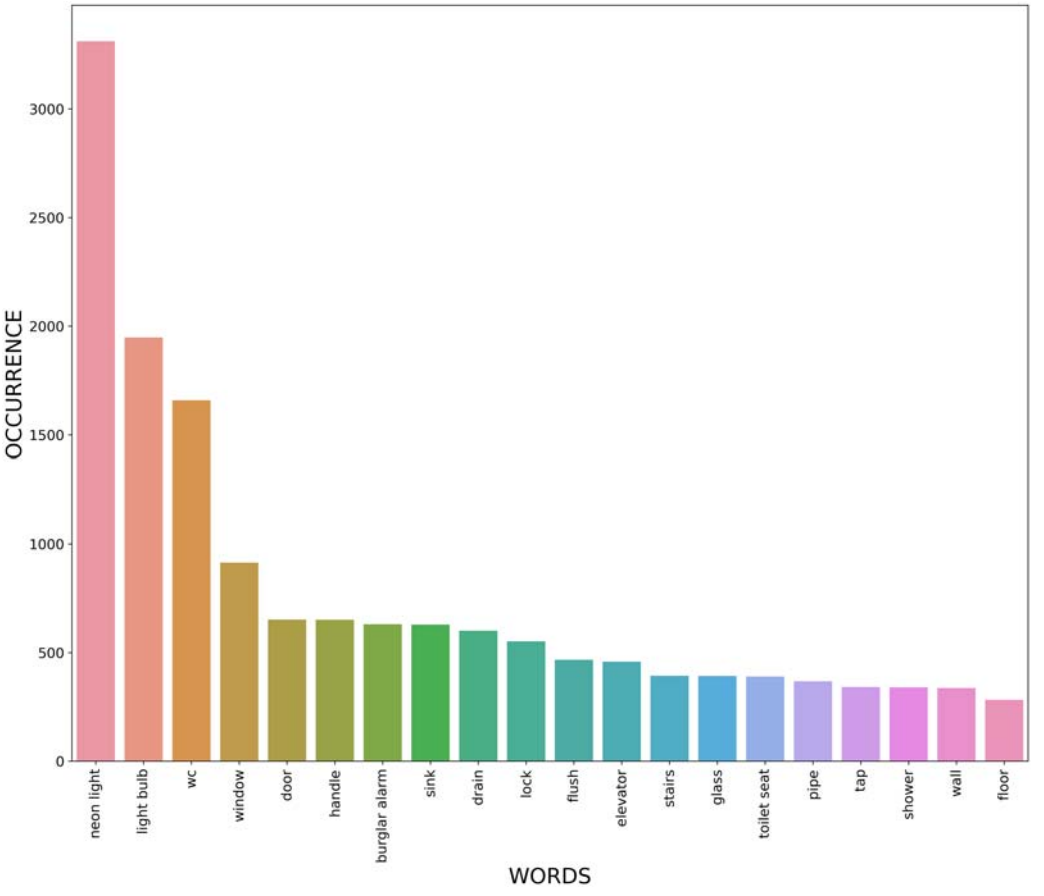

Figure 13. The most repeated words.

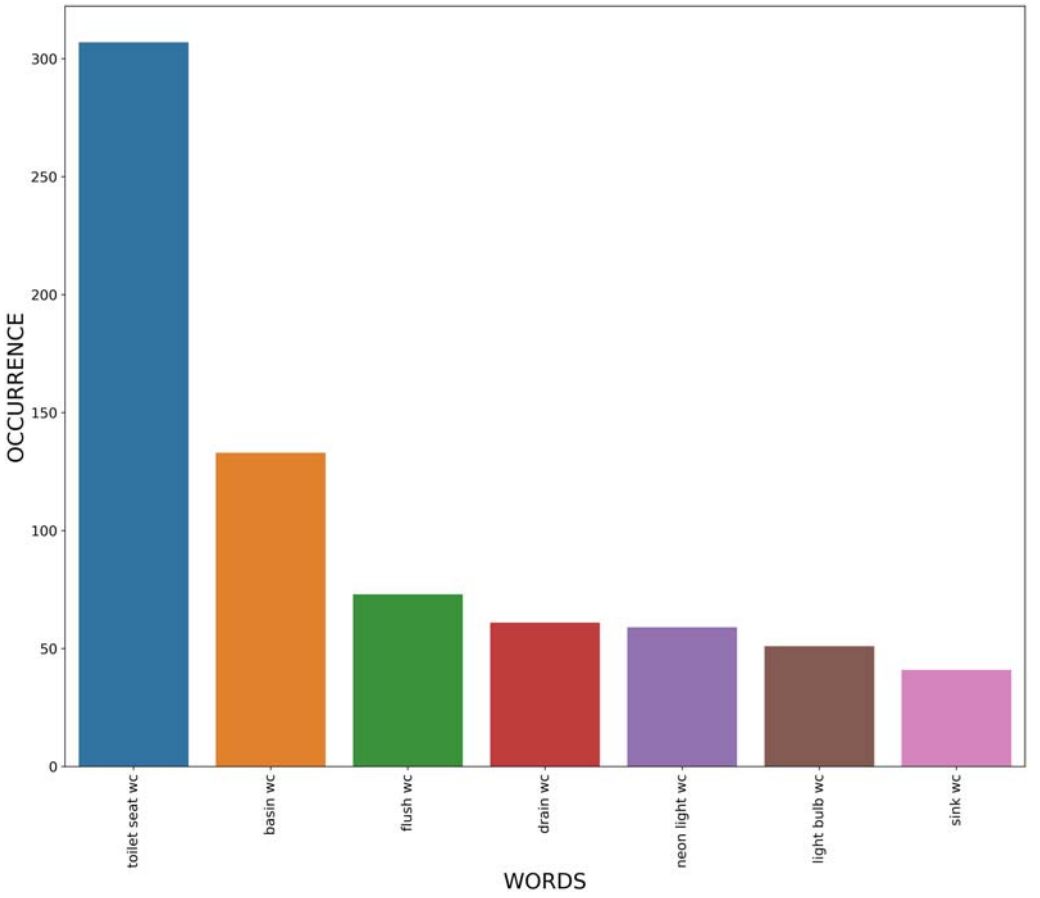

Figure 14. The pairs of meaningful words connected with the word 'wc'.

With reference to the adoption of text-mining for extracting useful information from CMMSs, the work of Gunay et al. (2019) showed how to cluster specific work orders limited to HVAC issues, avoiding superfluous and misleading data, such as routine maintenance requests, whereas the research carried out by Bortolini and Forcada (2020) identified the correlations between building characteristics and faults by considering data including gross floor area, year of construction, type of building use and building property. Compared to the works of Bortolini and Forcada (2020), and Gunay et al. (2019), this research proposed two methods which focus on pinpointing building areas where failures more often occur and identifying precisely all flawed building elements and systems. This means that not only principal failures are recognized, but also a specific component of the flawed element is identified, leading to a major accuracy of what sub-component is damaged.

Practical implications

The first part of the research focused on developing a method that can extract the ID number of rooms where faults mainly occur. Three types of analysis were conducted in order to assess whether the methodology works or needs adjustments. The general analysis concluded that the method identified the room ID numbers for 32.2% of the total number of maintenance requests. With reference to the analysis of the specific types of intervention, the results demonstrated to be consistent with the previous outcome, as for most of the types of intervention, the percentages of found room ID numbers were close to a third of the overall number of maintenance requests per the same type of intervention, as shown in Figure 10. On the other hand, some limitations of the proposed method were identified. Firstly, some of the types of intervention showed a lower percentage of found room ID numbers due to their limited pertinence with identifying room ID numbers. For instance, the category 'Fire extinguisher/hydrant' is rarely applicable to a single room, as fire extinguishers are usually located in building corridors. Secondly, the results are influenced by the maintenance representatives' accuracy and completeness ways of writing. Indeed, most of the WOs do not contain a room ID number, but rooms are described and pinpointed by using another name, such as 'bathroom' instead of 'room 12', or describing how to reach them.

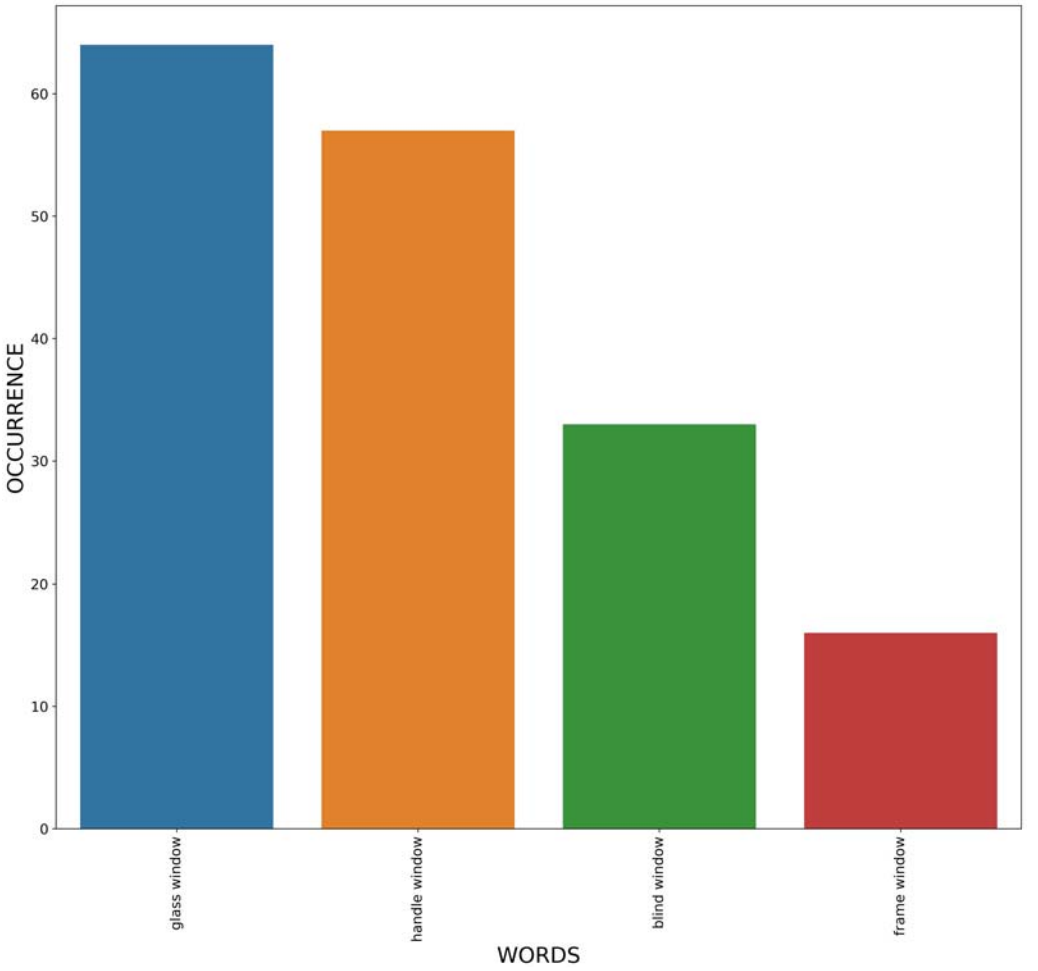

Figure 15. The pairs of meaningful words connected with the word 'window'.

Furthermore, some manners of writing room ID numbers can be extremely unusual, such as 'room ID number 112-3', which means 'room ID number 112 and 113'. Another issue concerns the limits to explain precisely the whole problem, which can affect other spaces and elements close to damaged areas. For instance, a leaking problem may involve two floors, but rigid forms for stating maintenance requests do not allow inputting two information in the same field, leading to difficulties in identifying and localizing an issue properly. Finally, this method can be extended to other maintenance scenarios, such as industrial and commercial scenarios, to reach similar results due to the fact that each of these scenarios usually have a digital maintenance management system. In these cases, instead of searching for room ID numbers, the validation processes would need to be adapted to identify the ID numbers of areas, such as laboratories or stores, on the basis of the type of portfolio analysed, On the other hand, the identified limitations might occur, as well. Indeed, problems related to limited pertinence with identifying area ID numbers are troublesome if these areas cannot be defined properly, while issues related to the maintenance representatives' accuracy and completeness ways of writing mainly depends on the instructions and education received.

The second part of the research focused on generating a method that can acquire the most repeated words in the maintenance requests for discovering what building components and systems are the most problematic. Since the data set is derived from an Italian database, the translation of some words into English words needs more than one word, such as light bulb and neon light. The analysis demonstrated to be coherent with the information shown in Table 2, as most of the maintenance requests concerned electrical, metalwork/carpentry and water/sanitary problems. One part of the results can be considered as stand-alone words, which can provide useful insights into what needs more attention. For instance, the most significant result was the word *neon light*, which was cited more than twice than the third most quoted word. On the other hand, the other part of the results cannot supply meaningful

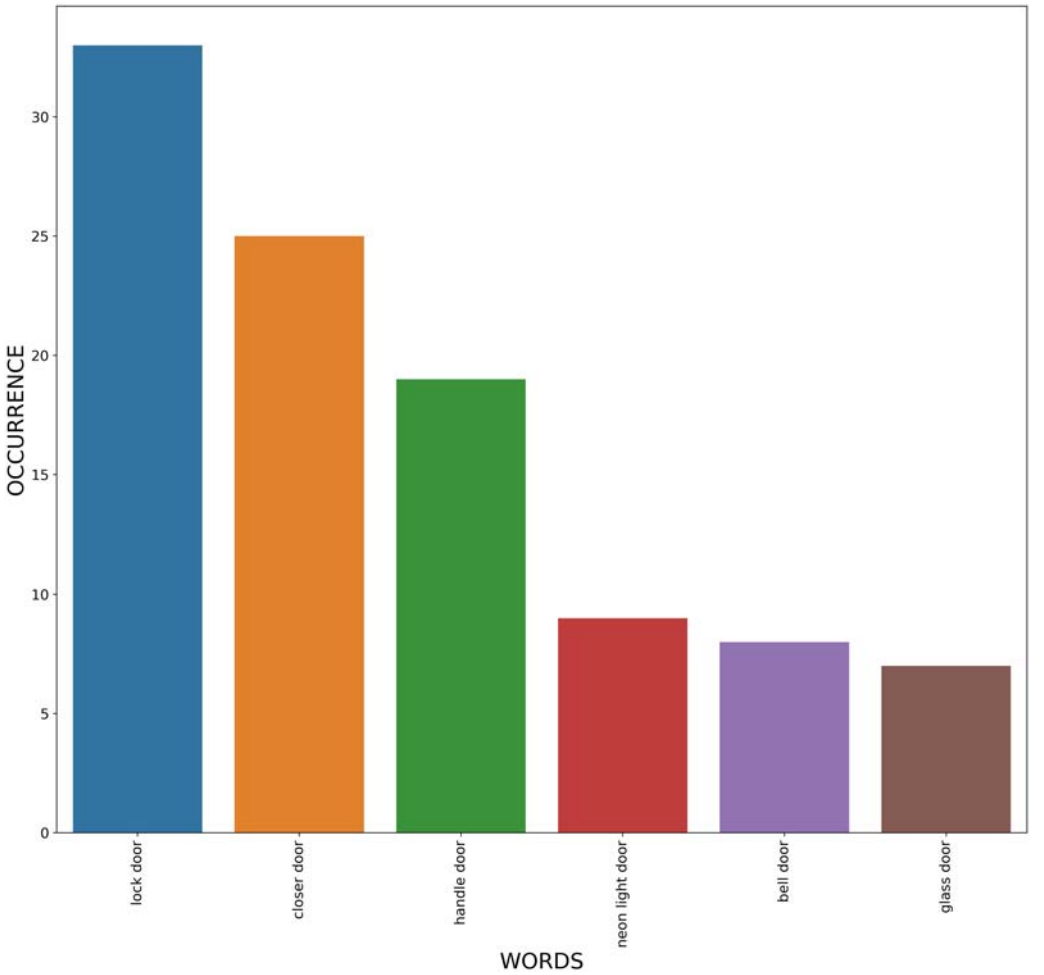

Figure 16. The pairs of meaningful words connected with the word 'door'.

outcomes, but needs another word to describe better what the problems are. For instance, results revealed that many water/sanitary problems related to the word *wc* were toilet seats, basins, flushes, drains, neon lights, light bulbs and sinks. Generalizing these results with other maintenance scenarios, it can be noticed that this method might be also valuable for identifying more specific and technical issues related to other scenarios, such as hospital and factory scenarios. In particular, the proposed method might be helpful to identify the groups of or individual most problematic specialized machines and systems. To adapt the algorithm, there would be a need of grouping similar ways of naming equipment at the beginning of the extraction process to identify unequivocal and peculiar machinery ID numbers.

Another limitation of the proposed methods concerns the fact that WOs are usually subjected to misleading information due to space name abbreviations, no space names or incorrect space names, which leads to time-consuming operations of identifying the correct place or component. Although commercial systems, such as CMMS, can support keeping trace of building maintenance requests, they are often badly organized and cannot always constrain storing all the necessary information. One solution to this problem which avoids complex activities, such as modifying the structure of the CMMS database, concerns improving the internal processes of stating maintenance requests. Representatives of maintenance work orders should be instructed on how to write a request in a standard way. A default form which expresses maintenance requests in order to get the right data at the right point should be developed. Descriptions should contain the position of where the problem occurs, a summary of the problem which states what element or system has a malfunction and then a brief explanation, as shown in Table 3. This form might be able to allow FM personnel to easily extract and exploit data due its well-organized pattern.

Although this solution can be useful to generate insights to manage buildings, it cannot provide the

Table 3. The proposed default form for maintenance requests.

Position of the problem:	*e.g. 'Room 19'*
Summary of the problem:	*e.g. 'Toilet seat broken'*
Brief explanation:	*e.g. '… text … '*

exact identification of elements. For instance, a building corridor can have several neon lights, therefore without specifying which one is the problem, it is not possible to carry out an in-depth analysis. This leads to a need to improve the structure of databases, where each element is linked to a specific ID number in the database so as to identify it whenever necessary. Well-structured databases can be organized according to a classification system of building components as proposed in Figure 17. The classification system has a tree structure, which starts from the site of where buildings are located up to specific element ID numbers. Information, such as work order ID number, facility ID number, location, the description of previous works, but also documents required to perform maintenance, has to be attached to building elements during FM processes.

However, this solution cannot provide comprehensive spatial and topological information that is needed to identify and fix issues. Thus, these platforms should be expanded by adding additional modules that can achieve this accuracy. Digital twin platforms can be the technology which enables an effective management of buildings in the operational phase. As well as dynamic data retrieved by smart devices, such as sensors, deployed around buildings, the other root component of these platforms is the use of BIM models to represent buildings. One of the basic properties of a BIM object is its Globally Unique Identifier (GUID), which is unique for each element in the model. Due to this characteristic, these systems allow detecting elements, simplify access to data and automate the process of linking elements with WOs.

Conclusion

Maintenance requests stored in CMMSs are directly linked to operational faults and building deteriorations. However, due to unstructured CMMSs, this data is not valuable to plan effective maintenance strategies until integrated and transformed into a piece of valuable information. To this end, this paper proposes two text mining-based methods to extract relevant information from maintenance requests. The results of the first method show a significant efficacy to detect room ID numbers where this approach is applicable, while the results of the second method reveal the typology and overall significance of specific building element/system faults. Based on these results, the proposed methods can be extended to other maintenance situations, such as industrial and commercial scenarios. Therefore, future studies could focus on applying these methods to other scenarios in order to identify major faults pertaining to different typologies of assets.

Although the proposed methods can be useful to generate insights to manage buildings, WOs are influenced by the maintenance representatives' accuracy and completeness ways of writing. Results demonstrate to be subjected to misleading information due to space name abbreviations, no space names or incorrect

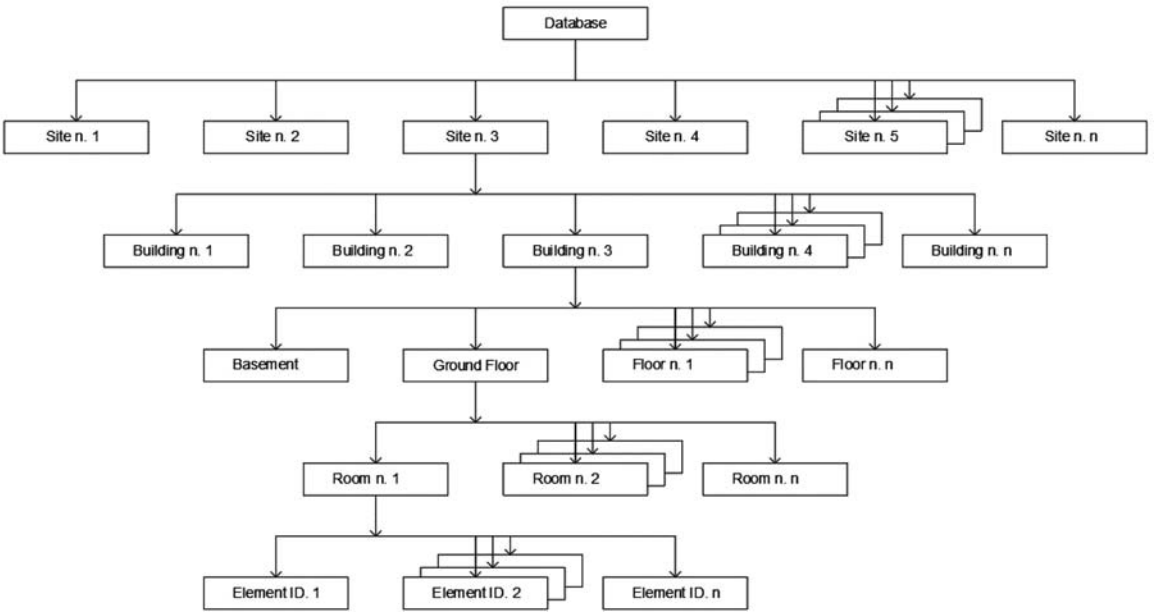

Figure 17. An example of a structured database.

space names. Based on the evidence, scholars and practitioners are recommended to identify and put in place well-organized patterns and internal processes of requesting maintenance works to improve the extraction of information from CMMS databases. Furthermore, results point out that issues concerning limits to explain precisely and completely the problem might lead to difficulties in identifying and localizing failures properly. Thus, additional investigations are needed in order to detect elements easily, simplify access to information and automate the process of linking elements with WOs. Future research will focus on employing Digital Twin platforms for enabling appropriate preventive strategies based on spatial and topological information, but also predicted maintenance by performing condition monitoring and assessment and analysing the trends of building system deteriorations.

Notes

1. Representatives: people in charge of reporting issues in buildings.
2. Stop word: a word which is excluded from a text because it is useless for the scope of analyses.

Acknowledgements

The authors thank the local administration of the Municipality of Trieste for sharing their maintenance data records.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Data availability statement

The data that supports the findings of this study is available from the local administration of the Municipality of Trieste. Restrictions apply to the availability of these data, which were used under license for this study. Data is available from the authors with the permission of the local administration of the Municipality of Trieste.

ORCID

Marco Marocco http://orcid.org/0000-0002-2055-6896

References

Akcamete, A., Akinci, B., & Garrett, J. H. (2010, June 30–July 2). Potential utilization of building information models for planning maintenance activities. *Proceedings of the International Conference on Computing in Civil and Building Engineering* (pp.151–157). https://www.researchgate.net/publication/260056325_Potential_utilization_of_building_information_models_for_planning_maintenance_activities.

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. *ArXiv:1707.02919 [Cs]*. http://arxiv.org/abs/1707.02919.

Allen, D. (1993). What is building maintenance? *Facilities*, *11* (3), 7–12. https://doi.org/10.1108/EUM0000000002230

Aziz, N. D., Nawawi, A. H., & Ariff, N. R. M. (2016). Building information modelling (BIM) in facilities management: Opportunities to be considered by facility managers. *Procedia – Social and Behavioral Sciences*, *234*, 353–362. https://doi.org/10.1016/j.sbspro.2016.10.252

Becerik-Gerber, B., Jazizadeh, F., Li, N., & Calis, G. (2012). Application areas and data requirements for BIM-enabled facilities management. *Journal of Construction Engineering and Management*, *138*(3), 431–442. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000433

Bortolini, R., & Forcada, N. (2020). Analysis of building maintenance requests using a text mining approach: Building services evaluation. *Building Research & Information*, *48*(2), 207–217. https://doi.org/10.1080/09613218.2019.1609291

Bortolini, R., Forcada, N., & Macarulla, M. (2016, September 7–9). BIM for the integration of building maintenance management: A case study of a university campus. *11th European Conference on Product & Process Modelling, Cyprus*, 9. https://www.researchgate.net/publication/308020370_BIM_for_the_integration_of_Building_Maintenance_Management_A_case_study_of_a_university_campus

Bruton, K., Raftery, P., O'Donovan, P., Aughney, N., Keane, M. M., & O'Sullivan, D. T. J. (2014). Development and alpha testing of a cloud based automated fault detection and diagnosis tool for air handling units. *Automation in Construction*, *39*, 70–83. https://doi.org/10.1016/j.autcon.2013.12.006

Chin, K.-S., Wang, Y.-M., Ka Kwai Poon, G., & Yang, J.-B. (2009). Failure mode and effects analysis using a group-based evidential reasoning approach. *Computers & Operations Research*, *36*(6), 1768–1779. https://doi.org/10.1016/j.cor.2008.05.002

Corneli, A., Naticchia, B., Cabonari, A., & Bosché, F. (2019, May 24). Augmented reality and deep learning towards the management of secondary building assets. *36th International Symposium on Automation and Robotics in Construction*. https://doi.org/10.22260/ISARC2019/0045.

Fan, B., Du, Z., Jin, X., Yang, X., & Guo, Y. (2010). A hybrid FDD strategy for local system of AHU based on artificial neural network and wavelet analysis. *Building and Environment*, *45*(12), 2698–2708. https://doi.org/10.1016/j.buildenv.2010.05.031

Federspiel, C., Martin, R., & Yan, H. (2003). *Thermal comfort models and complaint frequencies*. Center for the Built Environment, University of California, Berkeley, 34.

Franceschini, F., & Galetto, M. (2001). A new approach for evaluation of risk priorities of failure modes in FMEA. *International Journal of Production Research*, *39*(13), 2991–3002. https://doi.org/10.1080/00207540110056162

Gargama, H., & Chaturvedi, S. K. (2011). Criticality assessment models for failure mode effects and criticality analysis using fuzzy logic. *IEEE Transactions on Reliability*, *60*(1), 102–110. https://doi.org/10.1109/TR.2010.2103672

Gunay, H. B., Shen, W., & Yang, C. (2019). Text-mining building maintenance work orders for component fault frequency. *Building Research & Information*, *47*(5), 518–533. https://doi.org/10.1080/09613218.2018.1459004

Gursel, I., Sariyildiz, S., Akin, Ö, & Stouffs, R. (2009). Modeling and visualization of lifecycle building performance assessment. *Advanced Engineering Informatics*, *23*(4), 396–417. https://doi.org/10.1016/j.aei.2009.06.010

Ignatov, I. I., & Nørkj, P. (2019, November 13–15). Data formating and visualization of BIM and sensor data in building management systems. *19th International Conference on Construction Applications of Virtual Reality*, 12. https://www.researchgate.net/publication/337274994_Data_formating_and_visualization_of_BIM_and_sensor_data_in_building_management_systems

Katipamula, S., & Brambley, M. (2005). Review article: Methods for fault detection, diagnostics, and prognostics for building systems – a review, part I. *HVAC&R Research*, *11*(1), 3–25. https://doi.org/10.1080/10789669.2005.10391123

Kelly, G., Serginson, M., Lockley, S. R., Dawood, N., & Kassem, M. (2013, October 30–31). BIM for facility management: A review and a case study investigating the value and challenges. *Proceedings of the 13th International Conference on Construction Applications of Virtual Reality*. https://www.researchgate.net/publication/312469604_BIM_for_facility_management_a_review_and_a_case_study_investigating_the_value_and_challenges

Kensek, K. (2015). BIM guidelines inform facilities management databases: A case study over time. *Buildings*, *5*(3), 899–916. https://doi.org/10.3390/buildings5030899

Kim, W., & Katipamula, S. (2018). A review of fault detection and diagnostics methods for building systems. *Science and Technology for the Built Environment*, *24*(1), 3–21. https://doi.org/10.1080/23744731.2017.1318008

Kutlu, A. C., & Ekmekçioğlu, M. (2012). Fuzzy failure modes and effects analysis by using fuzzy TOPSIS-based fuzzy AHP. *Expert Systems with Applications*, *39*(1), 61–67. https://doi.org/10.1016/j.eswa.2011.06.044

Labib, A. W. (2004). A decision analysis model for maintenance policy selection using a CMMS. *Journal of Quality in Maintenance Engineering*, *10*(3), 191–202. https://doi.org/10.1108/13552510410553244

Lateef, O. A. (2009). Building maintenance management in Malaysia. *Journal of Building Appraisal*, *4*(3), 207–214. https://doi.org/10.1057/jba.2008.27

Lin, Y.-C., Su, Y.-C., & Chen, Y.-P. (2012, July 23–24). Mobile 2D Barcode/BIM-based facilities maintaining management system. *2nd International Conference on Strategy Management and Research*, Singapore, 5. https://www.semanticscholar.org/paper/Mobile-2-D-Barcode-%2F-BIM-based-Facilities-System-Lin-Su/34050f89edffbea993f41e0f91c9a2b0462d479d.

Liu, H.-C., Liu, L., & Liu, N. (2013). Risk evaluation approaches in failure mode and effects analysis: A literature review. *Expert Systems with Applications*, *40*(2), 828–838. https://doi.org/10.1016/j.eswa.2012.08.010

Majerník, M., Daneshjo, N., & Bosák, M. (2016). *Production management and engineering sciences*. http://www.crcnetbase.com/isbn/9781315673790.

Mmelesi, T., & Nwaigwe, K. N. (2020). A computerised maintenance management system as a teaching aid. *World Transactions on Engineering and Technology Education*, *18*(3), 6.

Pärn, E. A., Edwards, D. J., & Sing, M. C. P. (2017). The building information modelling trajectory in facilities management: A review. *Automation in Construction*, *75*, 45–55. https://doi.org/10.1016/j.autcon.2016.12.003

Provan, G. (2011, October 4–7). Generating reduced-order diagnosis models for HVAC systems. *22nd International Workshop on Principles of Diagnosis*, 9.

Schmittner, C., Gruber, T., Puschner, P., & Schoitsch, E. (2014). Security application of failure mode and effect analysis (FMEA). In A. Bondavalli & F. Di Giandomenico (Eds.), *Computer safety, reliability, and security* (Vol. 8666, pp. 310–325). Springer International Publishing. https://doi.org/10.1007/978-3-319-10506-2_21.

Teicholz, E. (2004). Bridging the AEC/FM technology gap. *Journal of Facilities Management*, *8*, https://docplayer.net/64590079-Bridging-the-aec-fm-technology-gap-eric-teicholz-ifma-fellow.html .

Tretten, P., & Karim, R. (2014). Enhancing the usability of maintenance data management systems. *Journal of Quality in Maintenance Engineering*, *20*(3), 290–303. https://doi.org/10.1108/JQME-05-2014-0032

Vilarinho, S., Lopes, I., & Oliveira, J. A. (2017). Preventive maintenance decisions through maintenance optimization models: A case study. *Procedia Manufacturing*, *11*, 1170–1177. https://doi.org/10.1016/j.promfg.2017.07.241

Volk, R., Stengel, J., & Schultmann, F. (2014). Building information modeling (BIM) for existing buildings – literature review and future needs. *Automation in Construction*, *38*, 109–127. https://doi.org/10.1016/j.autcon.2013.10.023

Wang, Y., Wang, X., Wang, J., Yung, P., & Jun, G. (2013). Engagement of facilities management in design stage through BIM: Framework and a case study. *Advances in Civil Engineering*, *2013*, 1–8. https://doi.org/10.1155/2013/189105

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2011). *Data mining: Practical machine learning tools and techniques*. Elsevier. https://doi.org/10.1016/C2009-0-19715-5.

Yang, C., Shen, W., Chen, Q., & Gunay, B. (2018). A practical solution for HVAC prognostics: Failure mode and effects analysis in building maintenance. *Journal of Building Engineering*, *15*, 26–32. https://doi.org/10.1016/j.jobe.2017.10.013

Zhang, R., & Hong, T. (2017). Modeling of HVAC operational faults in building performance simulation. *Applied Energy*, *202*, 178–188. https://doi.org/10.1016/j.apenergy.2017.05.153