Dynamic Extraction of Holiday Data for Use in a Predictive Model for Workplace Accidents

Danilo Magone Duarte Martins[1], Felipe G. Silva[1], Inês Sena[1], Laires A. Lima[1], Florbela P. Fernandes[1], Maria F. Pacheco[1], Clara Vaz[1], José Lima[1], and Ana I. Pereira[1]

Research Center in Digitalization and Intelligent Robotics (CeDRI), Instituto Politécnico de Bragança, Campus de Santa Apolónia, 5300-253 Bragança, Portugal {m309041}@alunos.ipb.pt,{gimenez, ines.sena, laires.lima, fflor, jllima, pacheco, clvaz, apereira}@ipb.pt

Abstract. Workplace accidents are a concern for companies nowadays and can occur due to internal and external factors of the company. Thereby, several strategies are developed to predict and minimize the hazards in this environment. Companies resort to intelligent solutions, such as predictive analytics, aiming to increase productivity while ensuring safety in the work environment. In terms of accident prediction analysis, different input data are needed to ensure the accuracy of a predictive model. Therefore, this study aims to automatic collect and pre-process data from holidays for subsequent implementation in an accident-oriented predictive model to demonstrate its relevance in predicting accidents in the workplace.

Keywords: Data Extraction · Web Service · Data mining · ElementTree

1 Introduction

Accidents in the workplace have been extensively studied by developing theories that try to explain, prevent and reduce them. The most used approaches are the accidents research and the implementation of preventive actions. However, digital technology solutions, such as predictive analytics, have started to be studied and implemented, contributing to the improvement of safety in the workplace [3]. Several areas, such as energy and construction, use accident prediction through predictive models with different input parameters referring to the accident and workplace [1,10]. In addition to factors relating to the company and employees, external factors influence the occurrence of accidents as they are part of the employees' daily routine, and can directly affect their performance [2,7]. Some authors have developed occupational accident prediction models considering weather data as predictors [9,11].

Recent literature has demonstrated the relationship between psychological aspects and occupational accidents. Studies mention that employees who endure high pressure for productivity are more likely to develop high levels of anger, sleep disturbances, health problems, and an increased risk of accidents and/or injuries [12]. Other studies reinforce the link between occupational stress and workplace accidents [8]. A model that intends to predict accidents can be directly influenced by factors extrinsic to the workplace.

In this context, this work aims to develop a Python script that can automatically and dynamically collect, from a Web Service, dates (and respective days of the week) of Portuguese national and municipal holidays between 2018 and 2022, considering that some holidays alter their date periodically, such as Easter. In addition, it is not easy

to find a single calendar that gathers information about the municipal holidays of the various Portuguese cities, which usually do not occur on the same date and weekly day.

By obtaining these holidays' data, we intend to prepare a dataset and use it to study the impact that national and municipal holidays have on the retail sector and, consequently, on employees. It is mandatory to find input parameters for the accident prediction model¸ because several factors influence the occurrence of accidents. After an extensive review, none studies were found involving this type of data nor on the influence that holidays have on stores in the retail sector. In other words, holiday data can be explored in conjunction with accident information to increase the performance of predictive models and minimize risk. The data needed to feed a predictive system can be collected from different sources: sensors, forms, databases (online and offline), and directly from the internet, among others. One of the options for obtaining data collected directly from the internet is Web Services, which are applications without a graphical interface designed to receive and respond to requests from other applications. This communication with other applications is done using the HTTP protocol, and the XML specification for the data [15].

HTTP request libraries are easy to find in modern programming languages [14]; some of the most used Python HTTP request libraries are httplib2, urllib2, and requests [4]. The response of HTTP requests is divided into three parts: the status code, in a set of headers, and the entity-body. The entity-body is the most crucial part as it is where the web service client will get the information it needs when parsing an XML document [14]. Works such as those by [5,6,13] use the ElementTree library to analyze XML documents with the tree style strategy because, with this library, it is possible to parse an entire XML document with just one line of code.

This paper is organized into four sections. The Section 1 describes the study's structure, the motivation in the development, and a brief literature review on the relevance of holiday data in accident prediction and the usual procedure used to extract this type of data. The procedure adopted to collect the data is described in Section 2. In Section 3 the results obtained are analyzed and discussed. Finally, the study is concluded, and future work is presented in Section 4.

2 Methodology

Fig. 1 shows the steps used to obtain the dataset that contains Portuguese national and municipal holidays. As it is possible to observe, six steps were taken into account to obtain the desired data. A description of each step follows.

- Data Collection (Step 1) – To find or identify possible sources for the desired data, among web pages, databases, and Web Services, for example.

- Literature Review (Step 2) – After choosing the selected data source, there will be a literature review about the existing methods to extract the data and about the pre-processing methods.

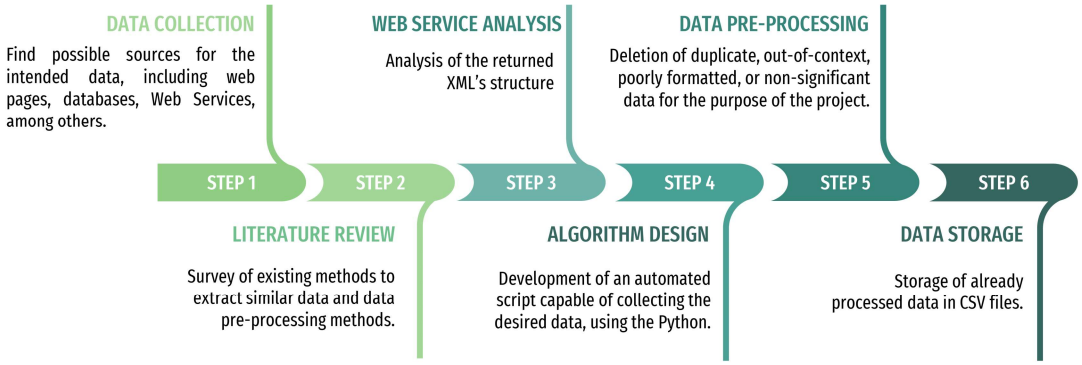

Fig. 1: Methodology used.

– Web Service Analysis (Step 3) – The structures of the XMLs returned from the Web Service were analyzed to prepare the algorithm's elaboration.

– Algorithm Design (Step 4) – Taking into account all the collected information, the algorithm will be designed through the development of an automated script capable of managing the desired data, using the Python programming language to create a client that receives documents XML from a Web Service.

– Data Pre-processing (Step 5) – After extracting the data through the algorithm developed in the previous task, pre-processing methods will be used to exclude duplicate, out-of-context, poorly formatted, or non-significant data for the project, as well as generating the day of the week when the holidays occurred, since the day can have some influence on accidents.

– Data Storage (Step 6) – The processed data will be stored in CSV files and used for further studies.

3 Results and Discussion

After searching for possible sources for the intended data, two Web Services were selected to return XML documents with information about municipal and national holidays. This choice is justified since this source had well-standardized data, facilitating the collection.

For Web Service, the year of interest must be a requesting data. In the case of municipal holidays, the ID number of the municipality is required, following a pattern provided by another service of the same Web Service. It was decided to collect data on all cities available in the Web Service and only then filter them according to the need. This approach was chosen since it is easier to autonomously filter the data once it is stored in a data frame of the Pandas library. This strategy could be problematic in large amounts of data, but this is not the case.

Before creating the script, the XML structures provided by the Web Services were analysed. Fig. 2 illustrates the hierarchy found for extracting data referring to municipal holidays.

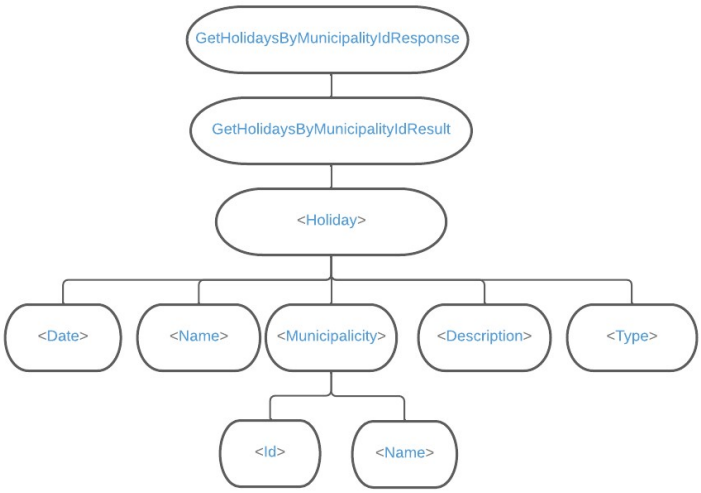

Fig. 2: XML Structure of Municipal Holidays

Fig. 3 presents the hierarchy found for collecting data on national holidays.

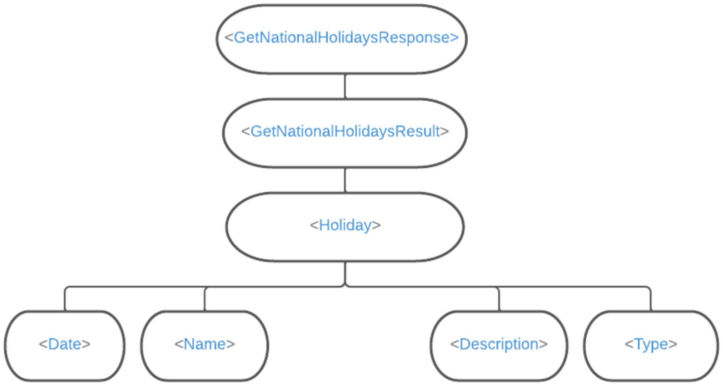

Fig. 3: National Holidays XML Structure.

The codes created make GET method requests for each year and municipality to the service in question, and the answer to these requests is stored in a variable of string type that has its content increased after each request made to the Web Service. This approach was chosen to facilitate the pre-processing data procedure.

With the content of all requests in a single variable, a pre-processing of the resulting XML is performed, that is, a single XML with the data of all the XMLs returned. The pre-processing consists of removing the root element, meaning, the tag at the top of the hierarchies shown in Figures 2 and 3 to allow the subsequent interpretation of the file by the libraries used.

Finally, the "fromString" function is used since string variables from the ElementTree library are used to obtain an Element object from the XML documents, one node from the tree. The node can be iterated to get access to the child nodes. The nodes have a "text" attribute with which it is possible to extract the text associated with each node and then obtain the desired data. Once the dates of each holiday are collected, the script get the corresponding day of the week using the weekday method of Python's datetime package.

The collected data are placed in data frames of the Pandas library in order to use the methods of this library to create *.csv* files as the final result of the scripts. In the next step of the research yet to be carried out, this library will also contribute to facilitate the filtering of the data.

The extracted dataset referring to municipal holidays contains about 914 data, showing the municipality's name, the holiday's representative date, and the day of the week that they occurred in each year selected, as can be seen in the sample presented in Table 1.

Table 1: Sample of the holidays data extracted in relation to municipal holidays

Municipality	**Date**	**Weekday**
Alcobaça	2018-08-20 00:00:00	Monday
Alvaiázare	2018-06-13 00:00:00	Tuesday
Ansião	2018-05-10 00:00:00	Thursday
...		
Elvas	2019-01-14 00:00:00	Monday
Fronteira	2019-04-06 00:00:00	Saturday
...		
Constância	2020-04-13 00:00:00	Monday
Coruche	2020-08-17 00:00:00	Monday
...		
Castro Daire	2021-06-29 00:00:00	Tuesday
Cinfães	2021-06-24 00:00:00	Thursday
...		
Calheta	2022-10-25 00:00:00	Tuesday
Velas	2022-04-23 00:00:00	Saturday

The dataset of national holidays extracted comprises 69 data, divided into the name of the holiday, the dates, and the day of the week that they occurred in each year selected, as shown in the sample shown in Table 2.

As shown in the tables, the data collected for each type of holiday were stored together, regardless of the year, to facilitate later use.

Table 2: Sample of the data extracted about national holidays.

Holiday	Date	Weekday
Ano Novo	2018-01-01 00:00:00	Monday
Carnaval	2018-02-13 00:00:00	Tuesday
Sexta-Feira Santa	2018-03-30 00:00:00	Friday
...		
Páscoa	2019-04-21 00:00:00	Sunday
Dia da Liberdade	2019-04-25 00:00:00	Thursday
...		
Restauração da Independência	2020-12-01 00:00:00	Tuesday
Imaculada Conceição	2020-12-08 00:00:00	Tuesday
...		
Dia de Camões, de Portugal e das Comunidades Portuguesas	2021-06-10 00:00:00	Thursday
Assunção de Nossa Senhora	2021-08-15 00:00:00	Sunday
...		
Dia da Liberdade	2022-04-25 00:00:00	Monday
Dia do Trabalhador	2022-05-01 00:00:00	Sunday

4 Conclusion and Future Work

The literature review made exposed that studies concerning the influence of holidays in workplace accidents occurrence are scarce. In this context, this research aimed to collect data to do such correlation. In this work scripts were created that can make requests to Web Services and interpret their response, and to store the collected information in *.csv* files.

To do so, extensive computational resources were not needed due to the use of the existing Python modules. The approach used may not be the most suitable for Web Services that can return extensive data, considering that the filtering was done only after data collection. However, the large computational cost involved in this process is not as significant for smaller applications as presented in this work.

After being processed and filtered by cities of interest, the obtained dataset will be used to study the impact that holidays may have on the occurrence of accidents. A new dataset will be prepared with the accidents between 2018 and 2022, already collected from a retail company in Portugal and associated to the occurrence of the holidays, the time spent from the last one and the proximity of the accident. These elements referring to holidays will be correlated with the cause of the accident to understand the influence of which a holiday may have on the behavior of the store and employees.

Acknowledgement

This work has been supported by FCT—Fundação para a Ciência e Tecnologia within the Project Scope UIDB/05757/2020 and NORTE-01-0247-FEDER-072598 iSafety: Intelligent system for occupational safety and well-being in the retail sector.

References

1. Ajayi, A., Oyedele, L., Delgado, J.M.D., Akanbi, L., Bilal, M., Akinade, O., Olawale, O.: Big data platform for health and safety accident prediction. World Journal of Science, Technology and Sustainable Development **16**, 2–21 (1 2019). https://doi.org/10.1108/WJSTSD-05-2018-0042
2. Alruqi, W.M., Hallowell, M.R., Techera, U.: Safety climate dimensions and their relationship to construction safety performance: A meta-analytic review. Safety Science **109**, 165–173 (11 2018). https://doi.org/10.1016/j.ssci.2018.05.019
3. Blanchard, D.: Ehs today: A smarter way to safety (2019), `https://www.ehstoday.com/safety-technology/article/21920103/a-smarter-way-to-safety`, last accessed 5 May 2022
4. Chandra, R.V., Varanasi, B.S.: Python Requests Essentials. Packt Publishing Ltd, 1 edn. (2015)
5. Dysarz, T.: Application of python scripting techniques for control and automation of hec-ras simulations. Water **10**, 1382 (10 2018). https://doi.org/10.3390/w10101382
6. Garabík, R.: Processing xml text with python and elementtree – a practical experience. p. 160 (2006)
7. Jiang, L., Lavaysse, L.M., Probst, T.M.: Safety climate and safety outcomes: A meta-analytic comparison of universal vs. industry-specific safety climate predictive validity. Work & Stress **33**, 41–57 (1 2019). https://doi.org/10.1080/02678373.2018.1457737
8. Kim, K.W., Park, S.J., Lim, H.S., Cho, H.H.: Safety climate and occupational stress according to occupational accidents experience and employment type in shipbuilding industry of korea. Safety and Health at Work **8**, 290–295 (9 2017). https://doi.org/10.1016/j.shaw.2017.08.002
9. Ksenhuck, B.C., Vieira, M.E.M., Lechuga, T.A., Bezerra, F.O., Corrêa, P.L.P.: Desenvolvimento e análise de algoritmos aplicados na predição de acidentes em ambiente fabril. In: Anais do Congresso Latino-Americano de Software Livre e Tecnologias Abertas (Latinoware 2019). pp. 22–31. Sociedade Brasileira de Computação - SBC (11 2019). https://doi.org/10.5753/latinoware.2019.10329
10. Liu, M., Tang, P., Liao, P.C., Xu, L.: Propagation mechanics from workplace hazards to human errors with dissipative structure theory. Safety Science **126**, 104661 (6 2020). https://doi.org/10.1016/j.ssci.2020.104661
11. da Luz Pola, C.: Aplicação de processo de classificação e técnica de bayes na base de dados de acidentes ocupacionais de uma empresa metalúrgica. Ph.D. thesis, Universidade de Caxias do Sul (2018)
12. Petitta, L., Probst, T.M., Ghezzi, V., Barbaranelli, C.: The impact of emotional contagion on workplace safety: Investigating the roles of sleep, health, and production pressure. Current Psychology (3 2021). https://doi.org/10.1007/s12144-021-01616-8
13. Pushpa, C.N., Shankar, R., Thriveni, J., Venugopal, K.R., Patnaik, L.M.: Emet: Extracting metadata using elementtree to recommend tags for web contents. International Journal of Computer Applications **5**, 8170–8177 (2014)
14. Richardson, L., Ruby, S.: RESTful web services. O'Reilly Media, Inc., 1 edn. (2007)
15. Zanetti, A.R., Camargo, V.V.: Web service no ambiente escolar: um estudo de caso. Tecnologias, Infraestrutura e Software **1**, 35–53 (2012)