Understanding semantics:
An natural-language-process-based intelligent extraction and classification engine for BIM names

CAO ZENGJIE
UNSW, Sydney, Australia

In Collaboration with:

Andrew Butler
COX Architecture, Sydney, Australia

Keywords. BIM; Natural language process; Intelligent engine

Abstract

Digital transformation is impacting the architecture industry most notably in relation to the uptake of Building information modelling (BIM) approaches in the production and delivery of architectural projects. BIM is an information model that is composed of completely sufficient information to support life cycle management. In BIM-based projects, multi-disciplinary actors can have a deeper and more efficient collaboration through their transparent information sharing. However, the lack of a unified standard in classification for BIM objects increases the difficulty of finding target information in BIM model and reduces the efficiency of both construction and management. In this case, natural language process (NLP) can be introduced into this finding process by building a auto-classification system for BIM objects.

NLP is a sub-field of artificial intelligence. NLP is about to program a computer to process and analyze natural language datasets. WordNet is a lexical database of semantic relations between words, which can select key words that are relevant to the certain theme. The combination of WordNet and NLP can offer an automated approach on semantic understanding BIM information. This research explores a natural-language-based intelligent extraction and classification engine to help users automatically find the relevant information by different target in BIM projects. By using WordNet with NLP, the semantic

relations of each BIM objects will be labeled. This enables the system to then classify each BIM objects by different purposes, which can save the BIM users' time of finding target information and allow them to work on more important design tasks' contributes to enhancing productivity in the AEC industry. NLP is a sub-field of artificial intelligence. NLP is about to program a computer to process and analyze natural language datasets. WordNet is a lexical database of semantic relations between words, which can select key words that are relevant to the certain theme. The combination of WordNet and NLP can offer an automated approach on semantic understanding BIM information. This research explores a natural-language-based intelligent extraction and classification engine to help users automatically find the relevant information by different target in BIM projects. By using WordNet with NLP, the semantic relations of each BIM objects will be labeled. This enables the system to then classify each BIM objects by different purposes, which can save the BIM users' time of finding target information and allow them to work on more important design tasks' contributes to enhancing productivity in the AEC industry.

1. Introduction: Research Aims and Motivations

With the development of digital technology, AI, big data, cloud computing, etc. continue to empower all walks of life, and the era of industrial digitalization is gradually coming. Facing this wave of digital technology, the digitalization of the architecture industry is lagging behind other industries. However, with the technologies such as Building information modelling (BIM), all actors of architecture industry can be coordinated on a “visualized and modelized” platform. BIM is a digital representation which allows all users share information based on open standards. (National Institute of Building Sciences, 2006). In the originally dispersed AEC (Architecture, Engineering & Construction) industry, by using BIM, its production lines, business lines and management lines can be effectively integrated. Project data and management information have also begun to be interconnected. Since the transparent sharing of data, BIM can improve project quality and reduce the risks and costs of construction projects.

However, a survey about the barriers to BIM implement, whose participants are from contractors constituting, client organizations and the BIM consultants' group, shows that one of barriers is poor collaboration among project participants (Chan et al., 2019). One of the factors leads to this barrier is the low efficiency of search technology. This is because there are variety ways of naming BIM room tag. For example, room “toilet” can be named as “lav”, “can”, “john”, etc. In the BIM software' own search engine, users can not type “john” to search all the toilets in the project. The “john” only can return the rooms which have the name tag “john”, which make project participants hard to find their target information, which increase the time in design and management process in BIM-based projects. This evidences the deficiencies and limitations of native search engines in BIM software. This research argues that it could be improved by developing an auto extraction and classification system of BIM room names to enhance the ability of search function in BIM software.

2. Research Objectives

The objectives of this research are to increase the efficiency of finding target information and expand the search methods for BIM users. Instead of matching input characters to find result, users can use natural language and computer can understand what are the information users wanted.

3. Research Questions

BIM can be seen as a collaboration tool for all AEC participants, which can increase the productivity by transparent data sharing. However, according to a UK BIM survey in 2019, it can be found there is a trend of falling first and then rising in the companies use BIM up to 25% projects and 50% projects [figure 1]. This means that they faced some difficulties and then solved them in BIM projects in after one year.

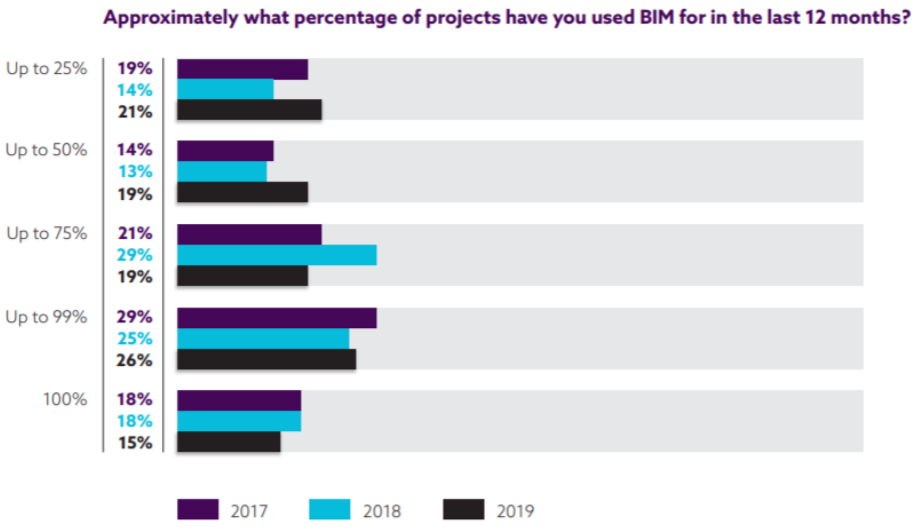

Figure 1. Approximately what percentage of projects have you used BIM for in the last 12 months? (Bain, 2019, p.18)

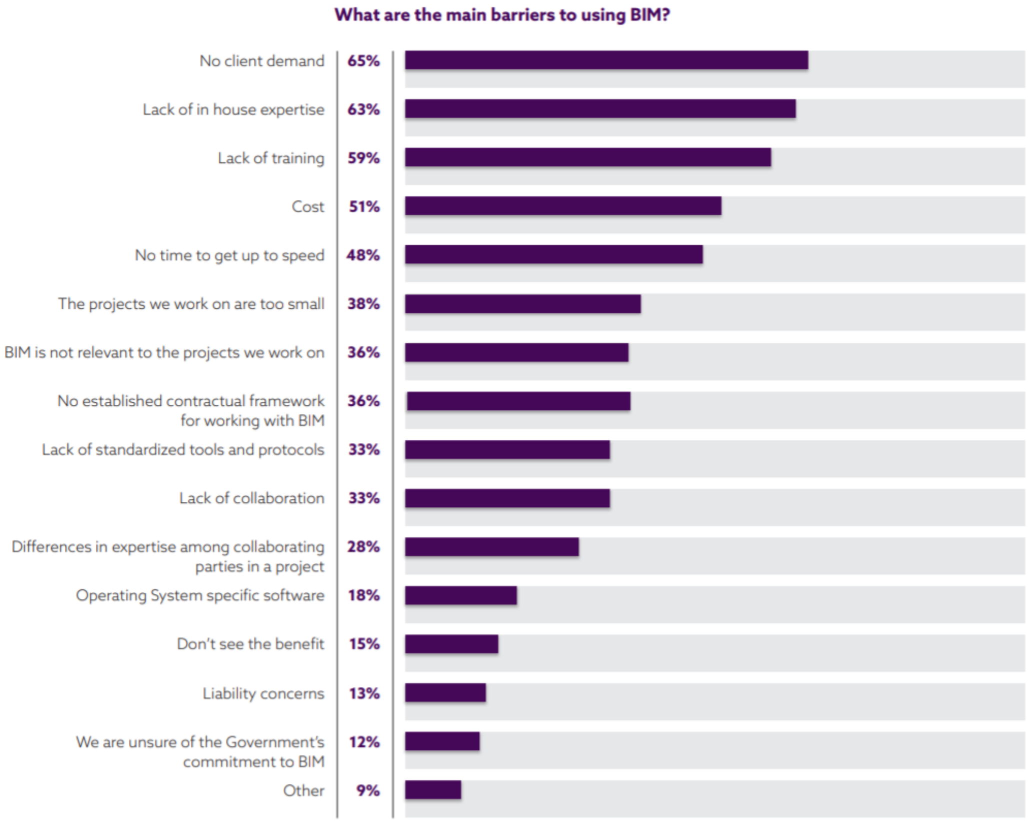

Figure 2. What are the main barriers to using BIM? (Bain, 2019, p.22)

From a comprehensive analysis with Figure 2, it is likely that AEC participants wanted to push using BIM but since the lack of expertise and training, it is hard for everyone to use it. After a year’s learning and training, more practitioners learned the skills of using BIM. However, the total amount of BIM-skilled practitioners is still not large enough, which can be seen from the overall drop in the companies using BIM more than three quarters [figure 1]. These data show to some extent that BIM is not easy to use.

A case study found that the most difficulties of using BIM is the inability to quickly and easily find faulty facilities and retrieve specific target information in operation and maintenance activities (Wang and Zhang, 2020).

BIM contains a large amount of data. If BIM software search engines could return users’ target information by their natural language, the difficulty of use BIM could decrease and the efficiency of information sharing among collaboration participants will increase, which may increase the adoption of BIM. In this case, my research question is

In what ways can natural language process be applied to enhance BIM software’s native retrieve engine to eliminate repetitive manual finding process?

This research explores the implement of natural language process to make the search conditions more flexible and closer to natural language, more semantic and easier for users to find their target information.

4. Methodology

The methodology of this research is that firstly I have a real-world BIM room names file created by COX Architecture. Then I am going to clean data, for example, I will delete the direction words, numbers and the words in brackets. This step is necessary for not confuse the system I will build later. Thirdly, I will try serval potential technologies, which can build the system, that other research mentions. Then, I am going to see which one will have a better performance.

5. Literature Review

Digital Transformation in Architecture and Cloud Co-Operation

New digital technologies often lead to the development of new solutions for existing problems. For example, social media technologies led to social media networks, which triggered the development of social media marketing concepts (Wiesböck and Hess, 2020). Digital transformation is impacting the architecture industry most notably in relation to the uptake of Building information modelling (BIM) approaches in the production and delivery of architectural projects. In BIM-based projects, multi-disciplinary actors can have a deeper collaboration through their transparent information sharing (Papadonikolaki et al., 2019).

In addition, (Abanda et al., 2018). Also, cloud BIM systems are parts of the workflow among designers. To increase the efficiency of the use of cloud BIM models, it is important that the BIM names of each parts are generalized.

Challenges of using BIM

The slow rate of BIM adoption in owners is diverse. In the studies of facility owners, it is observed that the complexity of exchange facility information is one of the most important issue (Cavka et al., 2017). Keeping the accuracy and relevance of information are the most critical variables in the process of exchanging information (Ghosh et al., 2015). For example, when designers create BIM models, different people will name a same room differently. Toilet can be called as 'W.C.', 'lavatory', 'lav', etc. To manage the lifecycle of all public toilet in the BIM models, if toilets are named differently, users need to rename all the room tags of toilets and it is possible to have omissions and it is time-consuming. To eliminate repetitive finding and renaming tasks and keep the information accuracy while exchange information, it is important that all the BIM names are generalized.

Machine Learning and Programming Language

Manually generalizing BIM names however is time-consuming, and mistakes and omissions are likely. The power of computation provides potentially one way to overcome this repetitive manual task. "It would be useful if computers could learn from experience and thus automatically improve the efficiency of their own programs during execution. (Michie et al., 1994)” Machine learning is often explained using metaphors and language of cognitive science, such as neural networks and connectionism, but it is more like a statistical correlation engine. By the inspiration of natural genetic algorithms, genetics-based machine learning was born (Goldberg and Holland, 1988). In addition, Python has been the most popular languages for data science (Millman and Aivazis, 2011). The combination of Python and machine learning allows easier use of data analysis (Pedregosa et al., 2011), which could have a positive impact on finding the pattern behind the auto-generalization of BIM names.

Potential Technologies Used in the Auto-Generalization of BIM Names System

Fuzzy theory is proposed by Prof. Zadeh in 1965 (Mukaidono, 2001). One of the contributions of fuzzy logic is that it allows intermediate degrees. For example, tall is a objective word in human cognition. For a set of 150cm people, 165cm is tall but for a set of 170cm people, 165cm is not tall. However, if we have the data sets of all human height, it is clear to define ‘tall’, which means a larger database will return a more accurate result (Zadeh, 2015). Analogy to the auto-generalization BIM rename system, the objects of computation are words instead of numbers and here is the weakness of fuzzy logic, which is that fuzzy logic do not have the ability to understand words. To add the ability of understanding words to the system, semantic relatedness ontology measures function is important.

YAGO is a large ontology driven by Wikipedia and WordNet (Suchanek et al., 2008). It uses Wikipedia redirect as one of its judgement. For example, if people type “Einstein” to Wikipedia search bar, the Wikipedia page will redirect to “Albert Einstein” so that the YAGO will know that “Einstein” links to “Albert Einstein”. Also, by extracting the information of attributes and relations in the infoboxes, YAGO can build category for each word (Suchanek et al., 2008). Analogy to the auto-generalization BIM rename system, for example, “toilet” can be same as “bog”, “lav”, “dunny”, etc. and it is under the category of “bathroom”. However, the lack of a set of self-contained and easily reproducible experiments to measure the accuracy of the system are the issue need to be solved. (Lastra-Díaz et al., 2017) .

In addition, another way to achieve semantic relatedness ontology measures ability is Natural Language Processing (NLP). NLP is about to program a computer to process and analyze natural language datasets. The combination of NLP and Machine Learning (ML) has been implemented to the area of semantic enrichment of BIM models by Tanya Bloch and Rafael Sacks (Bloch and Sacks, 2018). They used NLP and ML to extract the information from BIM objects’ properties and classified them based on different uses. For example, the classification of spaces is important for BIM spatial validation (Bloch and Sacks, 2018). It is a case for classification BIM objects and the concept of using NLP and ML can be also used for the auto-generalization BIM rename system to map ontology and other BIM objects.

This review aimed to reveal current difficulties in cloud BIM collaborations and find methods to increase the efficiency of using BIM models. It can be found that the ML has advantages in data analysis and self-learning. Fuzzy logic, WordNet and Wikipedia redirect together can build a more accurate ontology library, which can be useful for BIM objects semantic identifications. NLP shows its ability of BIM objects' different classification based on different uses and it can be critical in auto-generalization of BIM names system. It is proposed that future studies will further examine how NLP and ML can improve the efficiency of using BIM by help user extract their target information in BIM models accurately.

6. Methods

This research has investigated how to enhance BIM software's native retrieve engine by using NLP and an auto extraction and classification system.

6.1 Define standards for the auto extraction and classification system

With Cox Architecture, we decided to return 5 important categories for extracting BIM names' information.

1） Location

External Member’s Bar → #external

2） Size/Relation

Female Officials Area → #female

Main Storage Room → #main

Large Bathroom → #large

3） Sector

Catering Admin PWD Toilets → #Amenities

4） Ownership

Staff lift → #staff

5） Function

General Admission Lobby → #lobby

6.2 Datamuse Api

As this auto extraction and classification system is used in architecture area, we can do some pre-production and save them in local disk to reduce the responsive time, which will be the corpus for the system.

In this project, to implement NLP technology, I am using a word-finding engine called Datamuse Api, which can give developers programmatic access to words’ similarity and words’ relationship. By using this Api, I can use NLP in a highly customized and easier way.

6.3 Pre-production

The firstly step for this system is to find all the words related or with a similar meaning to room, space, facility, which will be the corpus for the function list. The result will be:

['bathroom', 'lounge', 'ballroom', 'bedroom', 'hall', 'bed', 'cubicle', 'hotel', 'house', 'bedrooms', 'cabin', 'auditorium', 'dorm', 'halls', …]

The second step is to find all the words which can describe the words in the function list and append them as the Size/Relation (S/R) list. The result will be:

['local', 'editorial', 'same', 'only', 'private', 'own', 'main', 'small', 'grand', 'large', 'own', 'small', 'great', 'large', 'own', 'same', 'small', 'little', …]

By analyzing a Revit sample file from COX Architecture, the ownership information of most BIM names is occupation. In this way, I decided to find occupation datasets and using Named-entity recognition (NER) by SpaCy, which is a sub area in NLP implement. By using SpaCy, it allows developers to create their own model to fit the tasks and provide ambiguous rules to be used in instances of ambiguous or specific ontology recognition tasks. The result will be:

[manager, senior, engineer, director, assistant, consultant, accountant, analyst, sales, advisor, supervisor, officer, programmer, specialist, …]

6.4 System Workflow

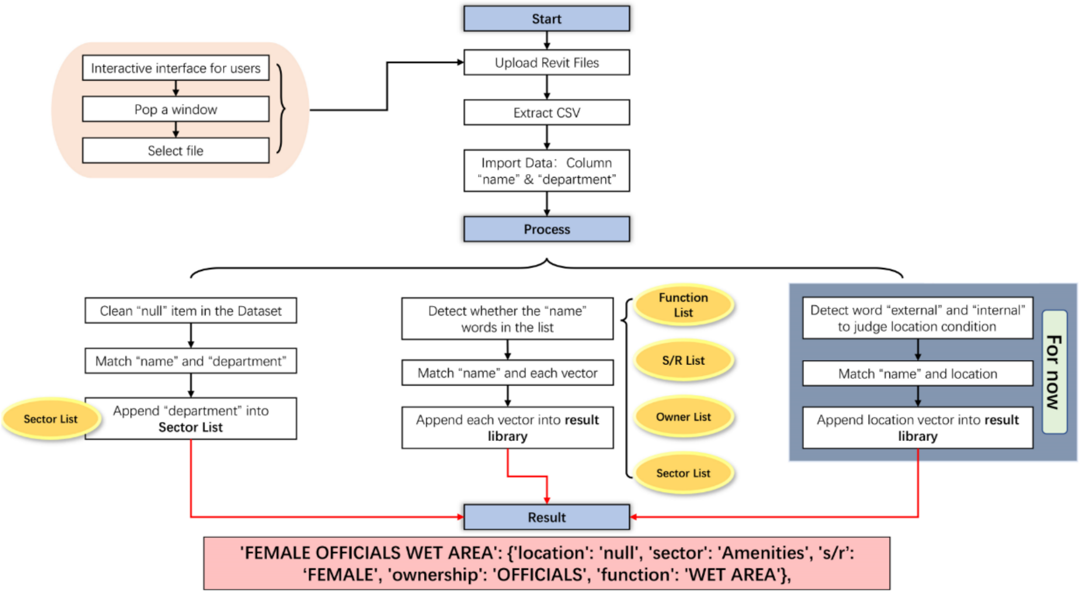
Start
Upload Revit Files
Extract CSV
Import Data: Column "name" & "department"
Process
Interactive interface for users
Pop a window
Select file
Clean "null" item in the Dataset
Match "name" and "department"
Append "department" into Sector List
Sector List
Detect whether the "name" words in the list
Match "name" and each vector
Append each vector into result library
Function List
S/R List
Owner List
Sector List
Detect word "external" and "internal" to judge location condition
Match "name" and location
Append location vector into result library
For now
Result
'FEMALE OFFICIALS WET AREA': {'location': 'null', 'sector': 'Amenities', 's/r': 'FEMALE', 'ownership': 'OFFICIALS', 'function': 'WET AREA'},

Figure 3. The System Workflow

6.5 Failed Attempt

1) Named Entity Recognition (NER)

Its concept is using word2vec to calculate the relationship between the entity and the words before entity. By training the models with a large dataset, NER model can be built. Since it is finding the word vectors' relationship between entity and the words before, it needs a context to do these works. Without context, the accuracy is about 1-2%, which I have tried.

2) Part of speech

NLP has a function which can identity each part of speech of phares. Then, I can use different strategy to classify information based on different part of speech.

However, the accuracy of tagging each word is low. The model I used is from SpaCy and the result is not like what I excepted, for example, a lot of name is using noun as attributives, which is hard to classify correctly.

7. Accuracy

To check the accuracy of this auto extraction and classification system, I used a real project from COX Architecture, which is University of Sydney ETP Stage 1.

7.1 Methodology

Step1. I export all the BIM names to a CSV file.

Step2. The extraction and classification system will process the CSV file.

Step3. I will randomly chose 100 results and if one of the 5 vectors is wrong, I marked the result is wrong.

Step4. Repeat the process 4 times and calculate the average accuracy.

7.2 Accuracy

In test 1, the accuracy is 81%. In test 2, the accuracy is 78%. In test 1, the accuracy is 87%. In test 1, the accuracy is 82%. In this way, the average accuracy is 82%.

8. Research Significance

Though several trying serval potential technology, it can be found that using Datamuse Api which can allow developers to build a high customized corpus by the implement of NLP. In addition, this research also explored the methods to enhanced native search engine, which is to build an auto extraction and classification system and expand the search key from only characters of BIM names to 5 area: Location, Size/Relation, Ownership, Sector, Function, which gives users more flexibility to search their target information.

In this way, the BIM software will be easier for novices and more efficient for experts to find their target information. This may further increase the number of BIM users.

9. Evaluation of research project

For the future evaluation, the whole system can be made as a new search engine plugin in Revit. Since the time limitation, I haven't finished this plugin but I have researched the architecture of building this plugin.

Firstly, I am using Revit python shell, which is an iron python console that runs inside of Revit and using pyRevit which allows user to create scripts (Iran-Nejad, 2020).

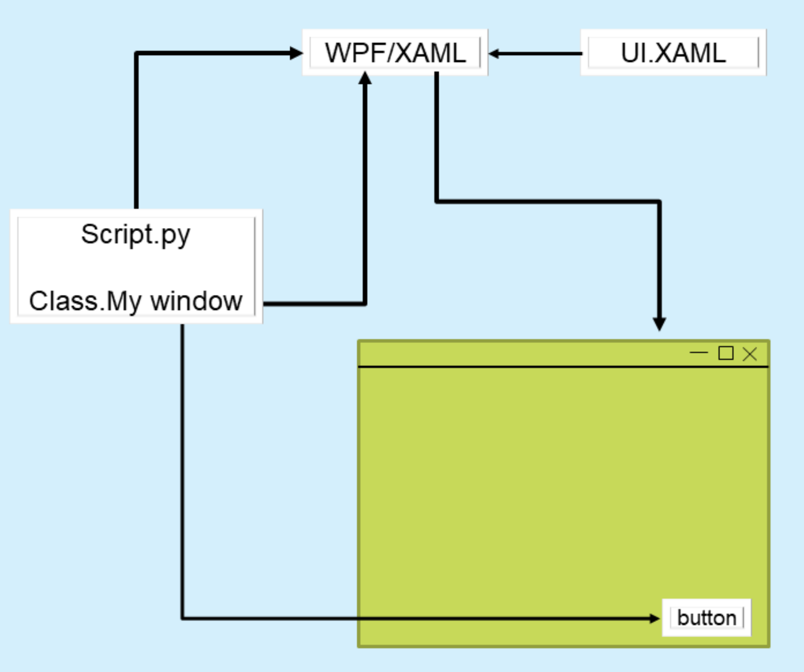

Figure 4. The Architecture of Building Plugin

WPF can be used for accessing the Script. Then it is necessary to deliver the path of "UI. XMAL" file to the XMAL or WPF engine. The engine will load it up and generate the components for the window and everything else. But before that it needs to have access to the "Class. my window", which also need be provide to the engine. we will create these components attach to the window to set the properties and function that needs to handle the click on the button. The class is like the code behind for the user interface.

In this way, by using this architecture, the function of the auto extraction and classification can be presented as a intelligent search engine in Revit as a plugin, to help users automatically find the relevant information by different target in BIM projects, which can save the BIM users' time of finding target information, reduce omissions, increase the accuracy of result and allow them to work on more important design tasks' contributes to enhancing productivity in the AEC industry.

10. Conclusion

BIM technology has been well-known in AEC industry. It combines and cooperates with all the information of all aspects of project delivery, namely architectural design, civil engineering design, structural design, mechanical design, construction, price estimation, schedule and project life cycle management. To put it simply, BIM enables the construction industry to achieve informatization and efficient production like ordinary industrial products. However, it is hard to find users' target information in a large amount of data.

This research focused on how to use NLP to bring intelligence to the Revit native search engine. By serval failed attempt, it is found that by using Datause Api to calculate the relationship between words, computer can have a clear definition of each word meaning. In this way, by building an auto extraction and classification system, the search methods can be expanded. In addition, if the whole system can be presented as a search engine plugin in the future, the difficulty of using BIM will decrease and the efficiency of information communication will increase.

This is a small step to make architecture filed more intelligent. In a society with rapid technological development, I believe there are other latest technology we can bring to the Architecture Field to achieve more improvements.

Acknowledgements

From the topic selection and data collection to the completion of this thesis, I have received many helps from both teachers and company. I would like to thank Nicole Gardner, Daniel and Hank Haeusler from UNSW, who have put forward a lot of valuable opinions on my research so that I can have a clear direction for this thesis. Secondly, I would like to thank Andrew Butler, who has taught me a lot of technical skills in last three years and as the representative of COX Architecture, he selected an interesting topic for me so that I can learn new things in a new filed for me. All of them help me finalize this thesis for obtaining my Bachelor of Computational Design.

Reference

Abanda, H., Mzyece, D., Oti, A. & Manjia, M. 2018. A Study of the Potential of Cloud/Mobile BIM for the Management of Construction Projects. *Applied System Innovation,* 1, 9.

Bloch, T. & Sacks, R. 2018. Comparing machine learning and rule-based inferencing for semantic enrichment of BIM models. *Automation in Construction,* 91, 256-272.

Cavka, H. B., Staub-French, S. & Poirier, E. A. 2017. Developing owner information requirements for BIM-enabled project delivery and asset management. *Automation in Construction,* 83, 169-183.

Chan, D. W. M., Olawumi, T. O. & Ho, A. M. L. 2019. Perceived benefits of and barriers to Building Information Modelling (BIM) implementation in construction: The case of Hong Kong. *Journal of Building Engineering,* 25, 100764.

Ghosh, A., Chasey, A. D. & Mergenschroer, M. 2015. Building Information Modeling for Facilities Management: Current Practices and Future Prospects. *Building Information Modeling.*

Goldberg, D. E. & Holland, J. H. 1988. Genetic algorithms and machine learning.

Iran-Nejad, E. 2020. *pyRevit* [Online]. Available: https://github.com/eirannejad/pyRevit [Accessed].

Lastra-Díaz, J. J., García-Serrano, A., Batet, M., Fernández, M. & Chirigati, F. 2017. HESML: A scalable ontology-based semantic similarity measures library with a set of reproducible experiments and a replication dataset. *Information Systems,* 66, 97-118.

Michie, D., Spiegelhalter, D. J. & Taylor, C. 1994. Machine learning. *Neural and Statistical Classification,* 13, 1-298.

Millman, K. J. & Aivazis, M. 2011. Python for Scientists and Engineers. *Computing in Science & Engineering,* 13, 9-12.

Mukaidono, M. 2001. *Fuzzy Logic for Beginners*, World Scientific.

National Institute of Building Sciences. 2006. *WHAT IS A BIM?* [Online]. Available: https://www.nationalbimstandard.org/faqs#faq1 [Accessed].

Papadonikolaki, E., van Oel, C. & Kagioglou, M. 2019. Organising and Managing boundaries: A structurational view of collaboration with Building Information Modelling (BIM). *International Journal of Project Management,* 37, 378-394.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R. & Dubourg, V. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research,* 12, 2825-2830.

Suchanek, F. M., Kasneci, G. & Weikum, G. 2008. YAGO: A Large Ontology from Wikipedia and WordNet. *Journal of Web Semantics,* 6, 203-217.

Wang, G. & Zhang, Z. 2020. BIM implementation in handover management for underground rail transit project: A case study approach. *Tunnelling and*

Underground Space Technology, 103684.

Wiesböck, F. & Hess, T. 2020. Digital innovations. *Electronic Markets,* 30**,** 75-86.

Zadeh, L. A. 2015. Fuzzy logic—a personal perspective. *Fuzzy Sets and Systems,* 281**,** 4-20.