Technische Universität München

Ingenieurfakultät Bau Geo Umwelt

Lehrstuhl für Computergestützte Modellierung und Simulation

# Computer Representation of Building Codes using Natural Language Processing (NLP) Techniques

Interdisciplinary Project

for Master of Science Department of Informatics

| | |
|---|---|
| Author: | Andaç Kürün |
| Supervisor: | Prof. Dr.-Ing. André Borrmann |
| | Jimmy Abualdenien, M.Sc |
| Submission Date: | 20th January 2021 |

# Contents

# 1 Introduction

We are living in buildings that are designed and constructed based on a set of rules to make sure of solidity and safety. Buildings should comply to the set of rules throughout its lifecycle like design, construction and demolition. These set of rules are complex and often changing regulations in other words codes and guidelines. In order to guarantee the safety of users, structural stability and solidity, code compliance check is a vital process during the design and construction phases of the building [1]. Nowadays the checking process is performed manually by consultants and experts from the field and also building permission authorities. Because it also involves several disciplines to go through this process, it often leads to problems, delays, inconsistencies. It makes a cumbersome and error-prone process.

In order to reduce these problems, there has been lots of research ongoing in automation of code compliance check. With the help of the enormous progress in Building Information Modeling and technical development, researchers can offer various approaches and applications. One of these is “Automated Code Compliance Check Based on Visual Language” [2]. The other one is “Extracting Implementable Methods for Automated Code Compliance Check” [3]. In that project, researchers wanted to improve rule based approach with high-level methods.

In this project, we aimed to present a new approach to automate code compliance check with Natural Language Processing (NLP). For this purpose, we want to show the advantages and disadvantages of the Natural Language Processing Approach in comparison to Rule Based Approach. In the second part, we investigated several techniques in NLP to find out the best method for this project. In the third part, we applied the NLP technique to extract “components” from the code compliance guideline.

# 2 Researched Approaches and Techniques

There are two approaches I researched in this project in order to analyze and extract information from building code regulations. First one is Rule Based Approach which is the conventional method. The other is Natural Language Processing Approach which is an advanced method especially in the last decade.

## 2.1 Researched Approaches

### 2.1.1 Rule Based Approach

Rule based approaches are the traditional and the oldest approaches in the field of information extraction. Due to their high accuracy and proven-well history, they are still in use for some of the specific tasks. The system relies on rules which are based on linguistic structures. It stores all the rules in the system. When one sentence is queried, it does parsing according to rules it memorized. It can extract the information by regular expression patterns or context free grammar rules. It can offer a lot of insight about the text by finding which words are nouns or parsing with patterns.

For example:

Sentence = “Jul 29 is the 210th day of the year”

Pattern = r'((Jan|Feb|Mar|Apr|May|Jun|Jul|Aug|Sep|Oct|Nov|Dec) [0-9]{2})'

Result = [('Jul 29', 'Jul')]

Here we can understand that the sentence is about time/date to deduct from the regular expression pattern.

### 2.1.2 Natural Language Processing (NLP)

Natural Language Processing is the field of Machine Learning to understand, process and derive meaning from human language. It gives an ability to machines to analyze human language. Because understanding human language is a comprehensive task, NLP is considered a difficult task in computer science. Thanks to advances in machine learning and technical development, NLP research has been rapidly growing in recent years. Natural Language Processing analyzes human language mainly in two ways. First one is syntactic analysis. Syntactic analysis is referring to the arrangement of the words in the sentence. It analyzes the sentences such that they make grammatically correct sentences. The second one is semantic analysis. Semantic analysis is referring to the meaning of the words in the sentence. It involves understanding the words and interpretation of these words. This task is one of the difficult aspects in NLP that has not been fully resolved. NLP applications in the real world seem to increase day by day and can help you with lots of tasks in the daily routine. Some of them are:

- Information Extraction
- Translation
- Recommendation
- Question & Answering
- Sentiment Analysis
- Speech Recognition

How does NLP work? Natural Language Processing requires to apply algorithms to identify and extract the natural language rules with probabilistic relation. When the text is given, the model will utilize the algorithm to extract the meaning associated with every sentence or words. Based on the probabilistic relation it does, it comes out the best possible result for that task.

### 2.1.3 Comparison of Rule Based Approach and NLP

Both of them get used to extract key information from text widely. For our specific project, I compared them in multiple ways to which one is suitable for our task.

1. In order to extract components and properties, we need to generate every rule in regulation sets for rule based approach. It needs to be done by some experts in the field of construction. Also the regulations contain thousands of pages in order to review which will take quite time. On the other hand NLP needs some of the data from regulations. You don’t have to label every sentence in the regulations.
2. For a rule based approach, we need structured data to generate patterns or regular expressions to identify components. Regulation sentences contain complex sentences or exceptions. Sometimes these sentences’ rules can contradict with another rule. But NLP needs labeled data for every sentence. Sentences do not depend on each other. Each sentences’ component can be labeled based on the meaning in that sentence.
3. For every new rule in regulations, we need to improve the knowledge base for rule based approach. At every time, it needs some development. When we look

at NLP approach, it doesn't need to improve again and again for future regulations. Because the model has a learning ability and uses the probabilistic relation between words, it infers the components in the new rules based on the past data.

4. Both approaches need massive data for better results. For a rule based approach, it needs to extract every rule from all regulations. For an NLP approach, it needs a big training dataset for better accuracy.

Because of these reasons, we decided to use NLP techniques for our project. I researched several techniques, namely Random Forest [4], Named Entity Recognition [5] and Question & Answering [6], to identify 'components'. Because Random Forest is built on rule based approach, we eliminated it. From now on we will focus on Named Entity Recognition and Question & Answering.

## 2.2 NLP Techniques

### 2.2.1 Question & Answering (QA)Technique

Question & Answering is focusing on building systems that automatically answer questions posed by humans in natural language. In other words, QA systems can be described as a method that provides the right short answer to a question. Let me explain how QA systems work.

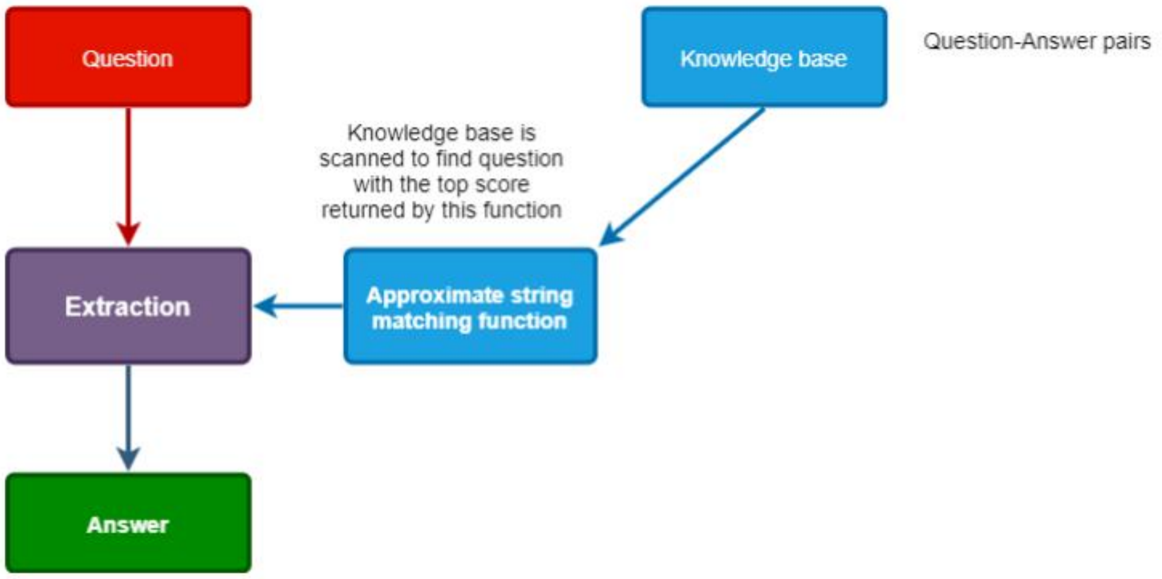
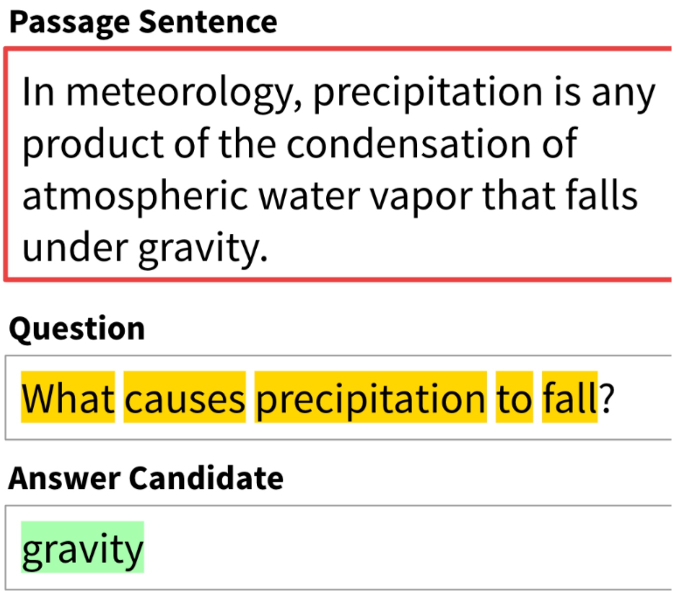

Question
Knowledge base
Question-Answer pairs
Knowledge base is scanned to find question with the top score returned by this function
Extraction
Approximate string matching function
Answer

Figure – 1 Q&A system [7]

As can be seen in Figure – 1, the QA system trained by several question-answer pairs with their texts. And then when the model gets a question, it uses the probabilistic relation to find the best possible answer. It relates the words of the question with the knowledge base and then provides us one best answer.

Example:



**Passage Sentence**

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity.

**Question**

What causes precipitation to fall?

**Answer Candidate**

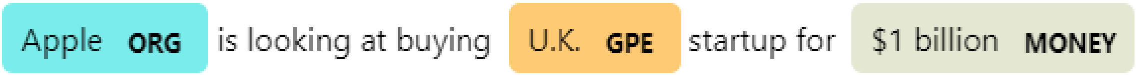
gravity

Figure – 2 Q&A Example [8]

### 2.2.2 Named Entity Recognition (NER) Technique

Named Entity Recognition is the task of identifying and categorizing key information in the text. Basically it classifies into predefined categories, such as place, time, organization, quantity etc. Let me explain the NER system. The NER system is designed to build a relation between words and categories/entities. The model is trained by labeled data which is words in sentence and corresponding predefined category pairs. Then when the model gets one sentence, it uses the probabilistic relation to find out which word or word groups belong to predefined categories.

Example:



Apple ORG is looking at buying U.K. GPE startup for $1 billion MONEY

Figure – 3 NER Example [9]

As can be seen in Figure – 3, words in the sentence are categorized according to their meaning. “Apple” is recognized as an organization. Extracting these key information in the text helps us to understand what the text is about.

### 2.2.3 Comparison of NER and QA

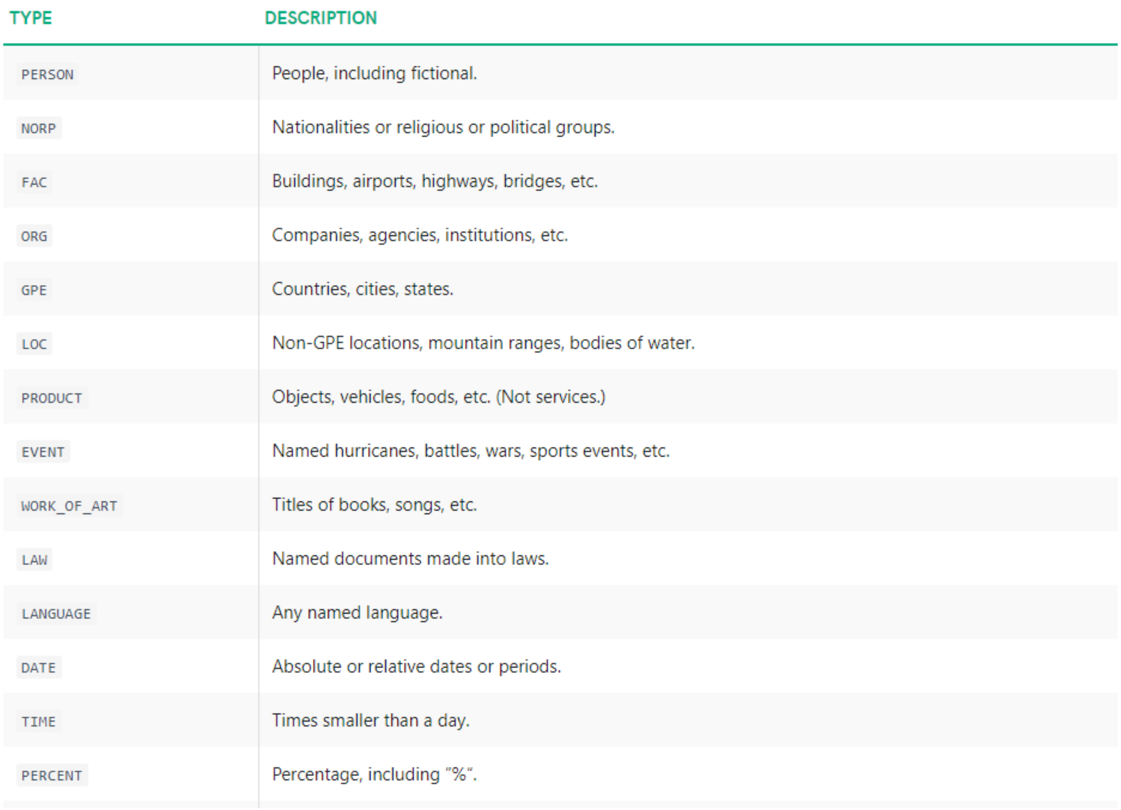
I compared them in different ways to find out which one is suitable for our project.

- You can train and build a model with both of them even if you have a small sized dataset. But accuracy and the applicability of your model will not be good in QA technique compared to NER. When I tried both of them with a small subset of dataset, QA accuracy was so low. It needs a massive dataset to get a good result. On the other hand, even with the small sized dataset, I get a quite good prediction result with NER technique.
- When we compared both of them in terms of development and training time, NER technique would be faster than QA technique. QA models have a more complex architecture than NER.
- When we start this project, we need to generate a dataset for training. Generating a dataset for NER and QA is a totally different process from each other. NER needs a labeling for each sentence for entities. QA needs a question-answer pair for each text. When we compared these two processes, we decided that NER would be faster than QA technique.

Because of these reasons, NER technique would be more suitable than QA technique. Thus, we decided to apply NER technique to extract “component” entities from the sentences in the regulations.

## 2.3 Framework for NER – SpaCY

Spacy is a free, open-source library for advanced NLP tasks in Python. It is designed specifically for production use to build real-world applications for businesses to release to the market fast. It supports multi-language models within their system and each of the languages has more than one pre-trained model to choose for your specific task. SpaCY provides us several built-in features in its pipeline such as tagging, parsing, named entity recognition etc. It helps us to develop our project quickly. When we look at the NER pipeline, SpaCY recognizes more than 20 entities in the model. It has a proven-well accuracy. You can see some of the entities in Figure – 4.



| TYPE | DESCRIPTION |
|---|---|
| PERSON | People, including fictional. |
| NORP | Nationalities or religious or political groups. |
| FAC | Buildings, airports, highways, bridges, etc. |
| ORG | Companies, agencies, institutions, etc. |
| GPE | Countries, cities, states. |
| LOC | Non-GPE locations, mountain ranges, bodies of water. |
| PRODUCT | Objects, vehicles, foods, etc. (Not services.) |
| EVENT | Named hurricanes, battles, wars, sports events, etc. |
| WORK_OF_ART | Titles of books, songs, etc. |
| LAW | Named documents made into laws. |
| LANGUAGE | Any named language. |
| DATE | Absolute or relative dates or periods. |
| TIME | Times smaller than a day. |
| PERCENT | Percentage, including "%". |

Figure – 4 NER Entities [9]

For our specific project, we need to extract “components” and it is not supported by the built-in SpaCY model. But SpaCY provides us a feature to categorize custom entities by re-training its model. What we called to this feature is Custom Named Entity Recognition (Custom NER). The method which I used by re-training is called Transfer Learning. Transfer Learning is a method where a model developed for a task is reused as the starting point for a model on a second task. Instead of starting from scratch every time for similar tasks, we reuse the pre-trained model to adapt it to our specific task.

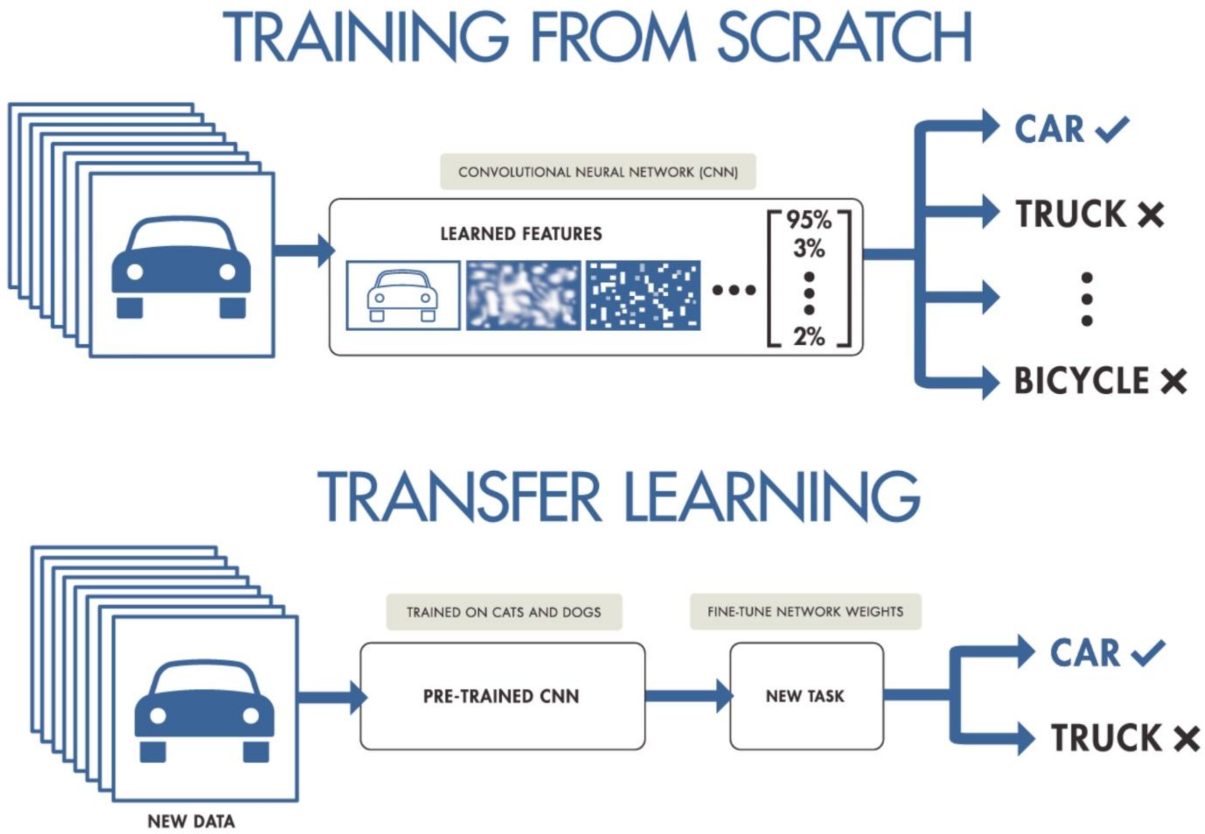


Figure – 5 Transfer Learning [10]

As it can be seen in Figure – 5, the upper model has been designed to classify several vehicle types. If we want to classify only cars and trucks, then we can re-train the upper model with our new data and make the fine-tune settings. Thus we can get a better accuracy result for the specific task. I used the same methodology with SpaCY. I re-trained the NER pipeline with our regulation dataset to classify the “components”.

# 3 Training Model

Training model consists of three parts which are preprocessing, training and evaluation of the model.

## 3.1 Part I – Data Generation & Preprocessing

Real world data tends to be noisy and inconsistent. This can lead to a poor quality for the model. In order to avoid this, data preprocessing comes forward. In natural language processing tasks, data preprocessing is consisting of several steps to clean text for better results. It is one of the key points in NLP for better models.

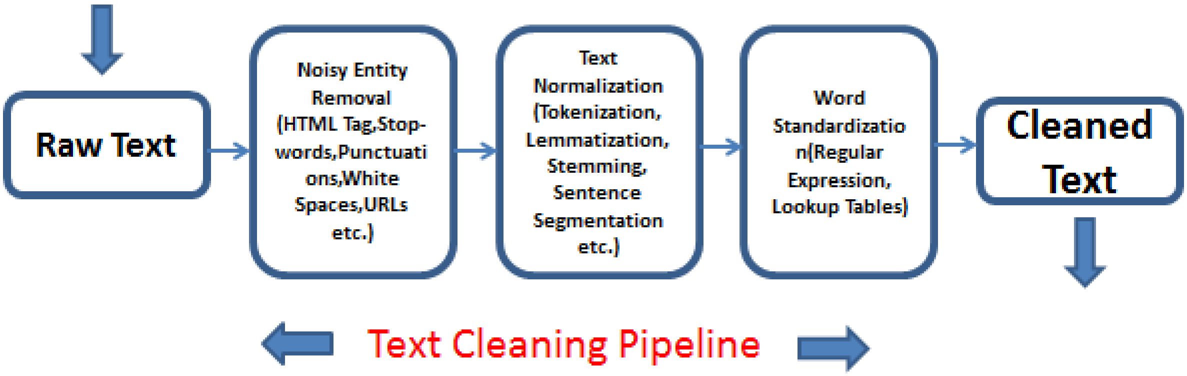
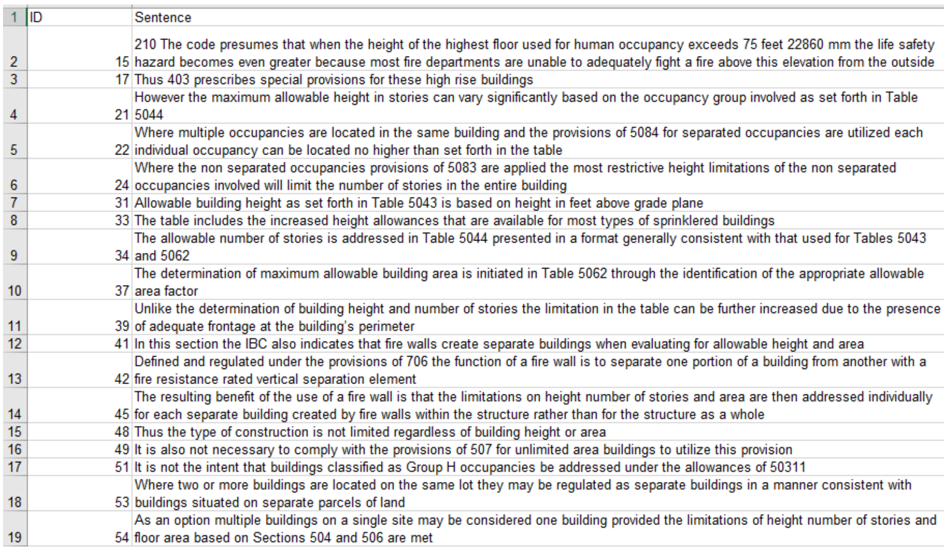

Raw Text
Noisy Entity Removal (HTML Tag,Stop-words,Punctuations,White Spaces,URLs etc.)
Text Normalization (Tokenization, Lemmatization, Stemming, Sentence Segmentation etc.)
Word Standardization(Regular Expression, Lookup Tables)
Cleaned Text
Text Cleaning Pipeline

Figure – 6 Preprocessing Pipeline [11]

As it can be seen in Figure – 6, raw text can go through the pipeline for cleaning. In NLP tasks, every step does not have to be considered to be done. Each step in preprocessing should be decided according to project or problem. In our project, I decided to use the first two steps to clean text. Our steps from pdf to labeled data:

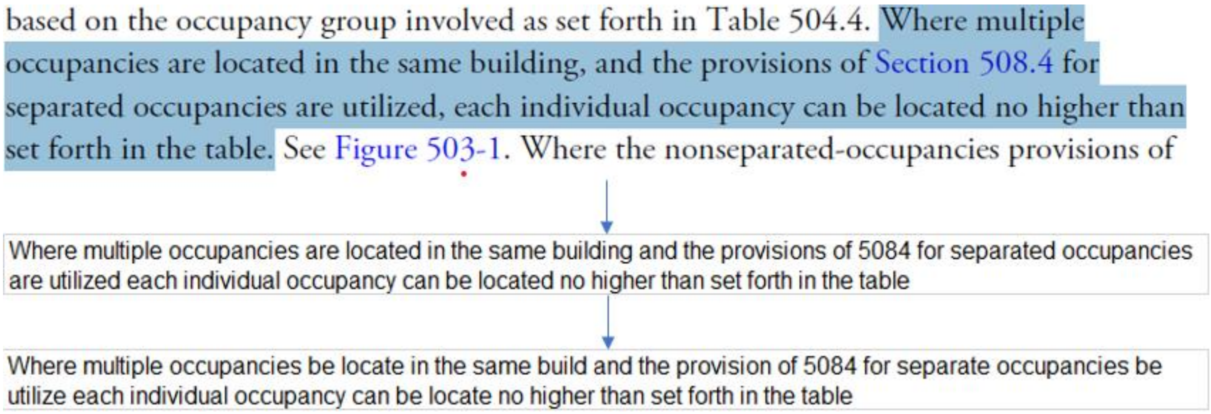
1. Each sentence in IBC regulation is extracted to one excel file.



| 1 | ID | Sentence |
|---|---|---|
| 2 | 15 | 210 The code presumes that when the height of the highest floor used for human occupancy exceeds 75 feet 22860 mm the life safety hazard becomes even greater because most fire departments are unable to adequately fight a fire above this elevation from the outside |
| 3 | 17 | Thus 403 prescribes special provisions for these high rise buildings |
| 4 | 21 | However the maximum allowable height in stories can vary significantly based on the occupancy group involved as set forth in Table 5044 |
| 5 | 22 | Where multiple occupancies are located in the same building and the provisions of 5084 for separated occupancies are utilized each individual occupancy can be located no higher than set forth in the table |
| 6 | 24 | Where the non separated occupancies provisions of 5083 are applied the most restrictive height limitations of the non separated occupancies involved will limit the number of stories in the entire building |
| 7 | 31 | Allowable building height as set forth in Table 5043 is based on height in feet above grade plane |
| 8 | 33 | The table includes the increased height allowances that are available for most types of sprinklered buildings |
| 9 | 34 | The allowable number of stories is addressed in Table 5044 presented in a format generally consistent with that used for Tables 5043 and 5062 |
| 10 | 37 | The determination of maximum allowable building area is initiated in Table 5062 through the identification of the appropriate allowable area factor |
| 11 | 39 | Unlike the determination of building height and number of stories the limitation in the table can be further increased due to the presence of adequate frontage at the building's perimeter |
| 12 | 41 | In this section the IBC also indicates that fire walls create separate buildings when evaluating for allowable height and area |
| 13 | 42 | Defined and regulated under the provisions of 706 the function of a fire wall is to separate one portion of a building from another with a fire resistance rated vertical separation element |
| 14 | 45 | The resulting benefit of the use of a fire wall is that the limitations on height number of stories and area are then addressed individually for each separate building created by fire walls within the structure rather than for the structure as a whole |
| 15 | 48 | Thus the type of construction is not limited regardless of building height or area |
| 16 | 49 | It is also not necessary to comply with the provisions of 507 for unlimited area buildings to utilize this provision |
| 17 | 51 | It is not the intent that buildings classified as Group H occupancies be addressed under the allowances of 50311 |
| 18 | 53 | Where two or more buildings are located on the same lot they may be regulated as separate buildings in a manner consistent with buildings situated on separate parcels of land |
| 19 | 54 | As an option multiple buildings on a single site may be considered one building provided the limitations of height number of stories and floor area based on Sections 504 and 506 are met |

Figure – 7 Extracted Excel File

2. Each sentence in the excel file is gone through the preprocessing pipeline to clean text. First sub step is the removal of punctuation, white spaces and noisy parts. Second sub step is lemmatization and tokenization. We noticed that lemmatization changes our components. Thus we decided to remove the lemmatization technique in the second sub step. We just did tokenization which gives one token to each word.



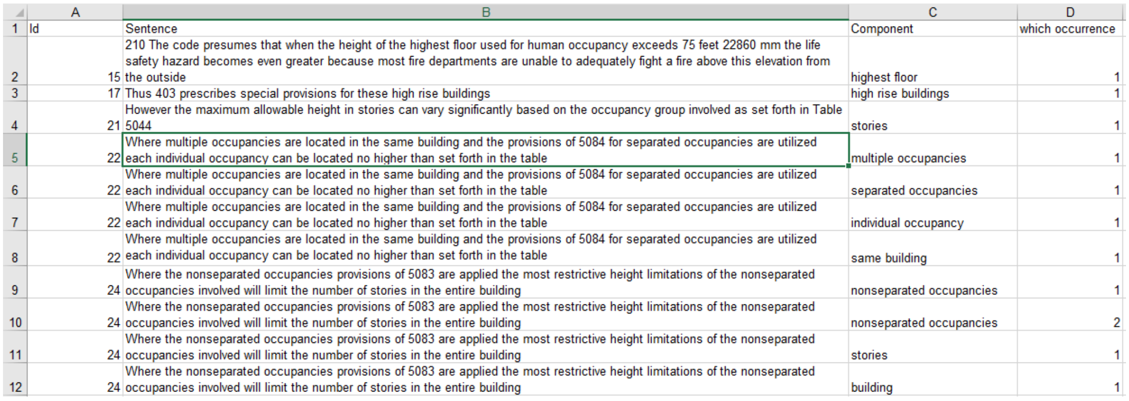
based on the occupancy group involved as set forth in Table 504.4. Where multiple occupancies are located in the same building, and the provisions of Section 508.4 for separated occupancies are utilized, each individual occupancy can be located no higher than set forth in the table. See Figure 503-1. Where the nonseparated-occupancies provisions of

Where multiple occupancies are located in the same building and the provisions of 5084 for separated occupancies are utilized each individual occupancy can be located no higher than set forth in the table

Where multiple occupancies be locate in the same build and the provision of 5084 for separate occupancies be utilize each individual occupancy can be locate no higher than set forth in the table

Figure – 8 Preprocessing Steps [12]

3. I identified every component one by one manually. The labeled data and its position in the sentence is needed for the model. That's why I also wrote a script to get an index position for every component in the sentence.



| | A | B | C | D |
|---|---|---|---|---|
| 1 | Id | Sentence | Component | which occurrence |
| 2 | 15 | 210 The code presumes that when the height of the highest floor used for human occupancy exceeds 75 feet 22860 mm the life safety hazard becomes even greater because most fire departments are unable to adequately fight a fire above this elevation from the outside | highest floor | 1 |
| 3 | 17 | Thus 403 prescribes special provisions for these high rise buildings | high rise buildings | 1 |
| 4 | 21 | However the maximum allowable height in stories can vary significantly based on the occupancy group involved as set forth in Table 5044 | stories | 1 |
| 5 | 22 | Where multiple occupancies are located in the same building and the provisions of 5084 for separated occupancies are utilized each individual occupancy can be located no higher than set forth in the table | multiple occupancies | 1 |
| 6 | 22 | Where multiple occupancies are located in the same building and the provisions of 5084 for separated occupancies are utilized each individual occupancy can be located no higher than set forth in the table | separated occupancies | 1 |
| 7 | 22 | Where multiple occupancies are located in the same building and the provisions of 5084 for separated occupancies are utilized each individual occupancy can be located no higher than set forth in the table | individual occupancy | 1 |
| 8 | 22 | Where multiple occupancies are located in the same building and the provisions of 5084 for separated occupancies are utilized each individual occupancy can be located no higher than set forth in the table | same building | 1 |
| 9 | 24 | Where the nonseparated occupancies provisions of 5083 are applied the most restrictive height limitations of the nonseparated occupancies involved will limit the number of stories in the entire building | nonseparated occupancies | 1 |
| 10 | 24 | Where the nonseparated occupancies provisions of 5083 are applied the most restrictive height limitations of the nonseparated occupancies involved will limit the number of stories in the entire building | nonseparated occupancies | 2 |
| 11 | 24 | Where the nonseparated occupancies provisions of 5083 are applied the most restrictive height limitations of the nonseparated occupancies involved will limit the number of stories in the entire building | stories | 1 |
| 12 | 24 | Where the nonseparated occupancies provisions of 5083 are applied the most restrictive height limitations of the nonseparated occupancies involved will limit the number of stories in the entire building | building | 1 |

Figure – 9 Labeled Data Excel File

**Sentence:** “Where multiple occupancies are located in the same building and the provisions of 5084 for separated occupancies are utilized each individual occupancy can be located no higher than set forth in the table”

**Component:** “multiple occupancies”

**Index:** 6 – 26

4. All labeled data is saved to “.tsv” or “.csv” file in order to be used by the model for the training part.

## 3.2 Part II – Training Model

As I mentioned before, I used the transfer learning methodology for our project. I retrained the powerful Spacy model for our specific task. Training a model part consists of several steps to acquire a good model.

1. In the previous part, I generated a clean dataset for the model. Before starting to train, we need to divide this dataset into two parts. One should be a training dataset which  is 80% of the dataset and the other one is a test dataset which is 20% of the dataset. The test dataset is never touched until the evaluation phase because it represents the unseen future data for us. This separation of the dataset should be balanced so that when we evaluate the model with the test dataset, it will perform a better result. The model uses the training dataset to adjust the weights.

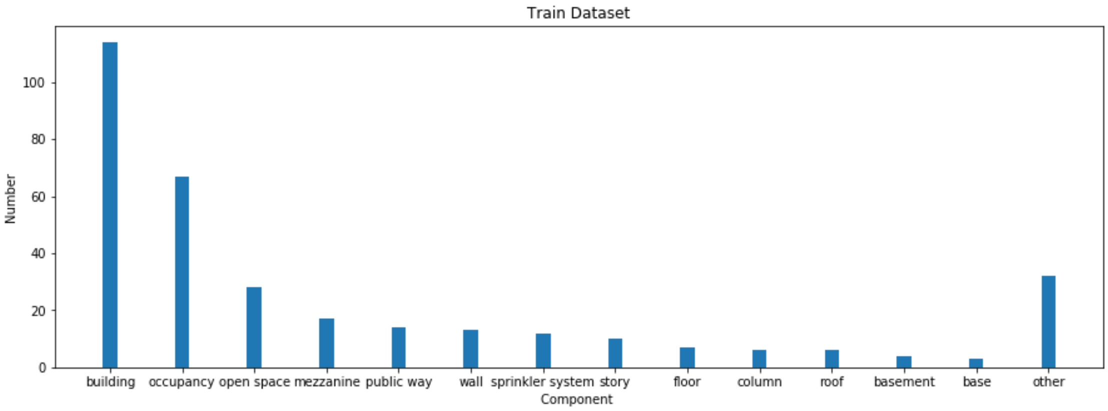


Figure – 10 Training Dataset Distribution

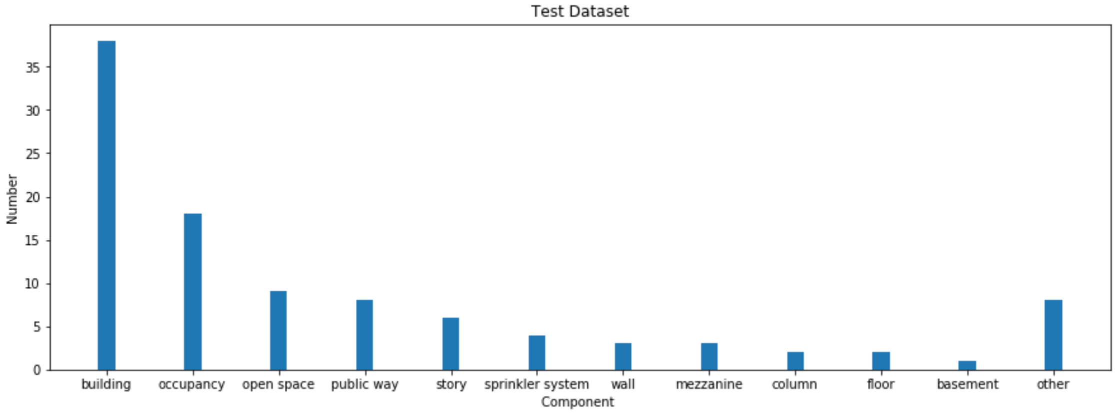
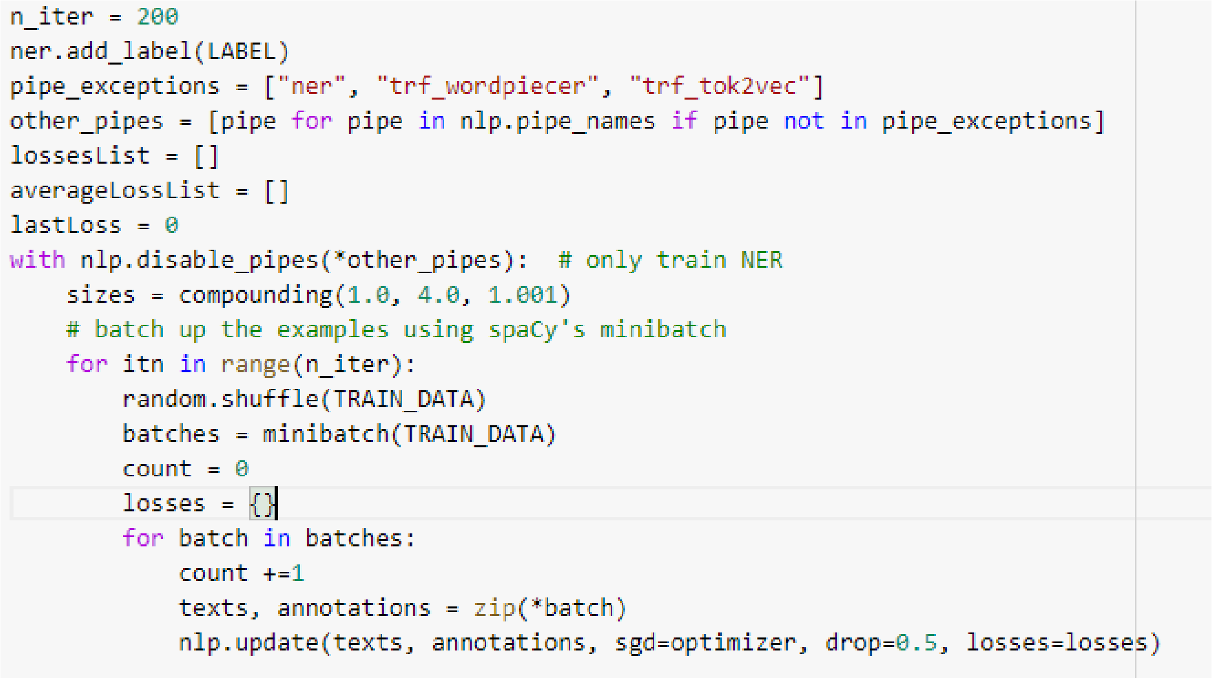


Figure – 11 Test Dataset Distribution

As it can be seen in Figure – 10 and Figure – 11, the components in the sentences are distributed in two datasets with protecting the ratio. That is also one of the keys for the model that affects the performance.

2. In Spacy, there are three different models, small, medium and large, for the English language. We decided to use the large one, called “en_core_web_lg”, because it has the most extensive and best NER F-1 score. It has 685k keys and unique vectors. Its F-1 score for NER task is 86.4.
3. After that I started the training phase with our training dataset. I wrote a small script to re-train the model for our task.



```
n_iter = 200
ner.add_label(LABEL)
pipe_exceptions = ["ner", "trf_wordpiecer", "trf_tok2vec"]
other_pipes = [pipe for pipe in nlp.pipe_names if pipe not in pipe_exceptions]
lossesList = []
averageLossList = []
lastLoss = 0
with nlp.disable_pipes(*other_pipes):  # only train NER
    sizes = compounding(1.0, 4.0, 1.001)
    # batch up the examples using spaCy's minibatch
    for itn in range(n_iter):
        random.shuffle(TRAIN_DATA)
        batches = minibatch(TRAIN_DATA)
        count = 0
        losses = {}
        for batch in batches:
            count +=1
            texts, annotations = zip(*batch)
            nlp.update(texts, annotations, sgd=optimizer, drop=0.5, losses=losses)
```

Figure – 12 Training script

As it can be seen in Figure – 12, the training dataset is divided into minibatches and then the model is updated with each batch and every iteration. In here, we are only updating the “ner” pipeline in the model because our task is related with NER.

4. After the training is finished, we need to validate the model to see how good it is. In order to do this, we can check the average losses through the iterations and the loss chart to decide the model. If you are not satisfied with the result, you can change some hyperparameters and then start the training process again. What are these hyperparameters:
    - Dropout value: can be 0, 0.25, 0.5 …
    - Iteration value: can be 100, 200, 500 …
    - Batch size: can be 8, 16, 32 …
    - Dataset separation ratio: can be 80-20%, 70-30%

My final settings for my model:

| | |
|---|---|
| Dropout | 0.05 |
| Batch size | 8 |
| Iteration | 200 |

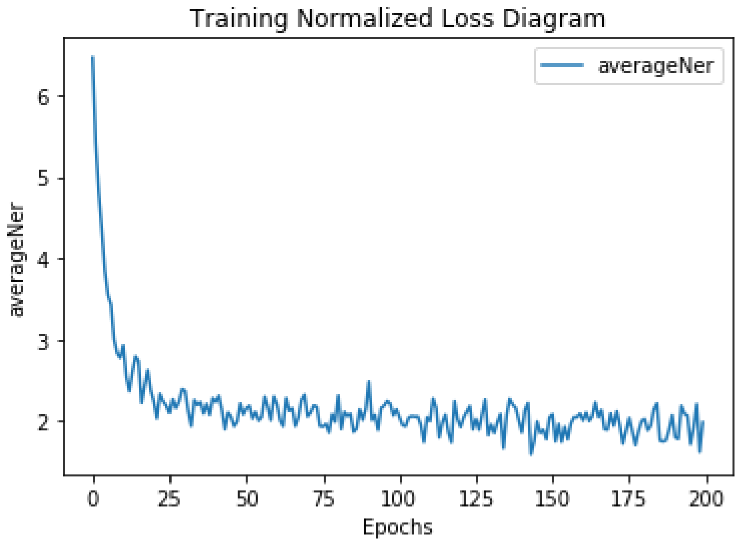
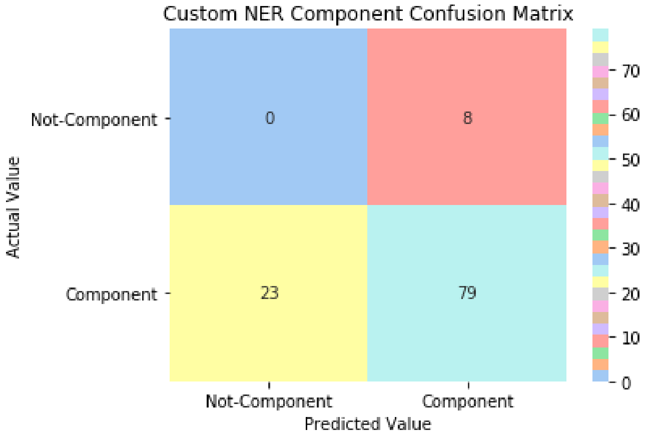


Figure – 13 Training Loss Diagram

As it can be seen in Figure – 13, my final model loss is below 1.5 and the loss diagram seems promising for evaluation. When the model result is good enough, it is saved to the disk to be reused again.

## 3.3 Part III – Evaluation

After the training part is over, we need to evaluate our model with the test dataset to see how good it goes with unseen data. In order to do that, I wrote a script that checks the model predictions with the ground truth value and then identifies the true-positive, false-positive and false-negative results.



| | |
|---|---|
| Precision | 90.80 |
| Recall | 77.45 |
| F-1 Score | 83.60 |

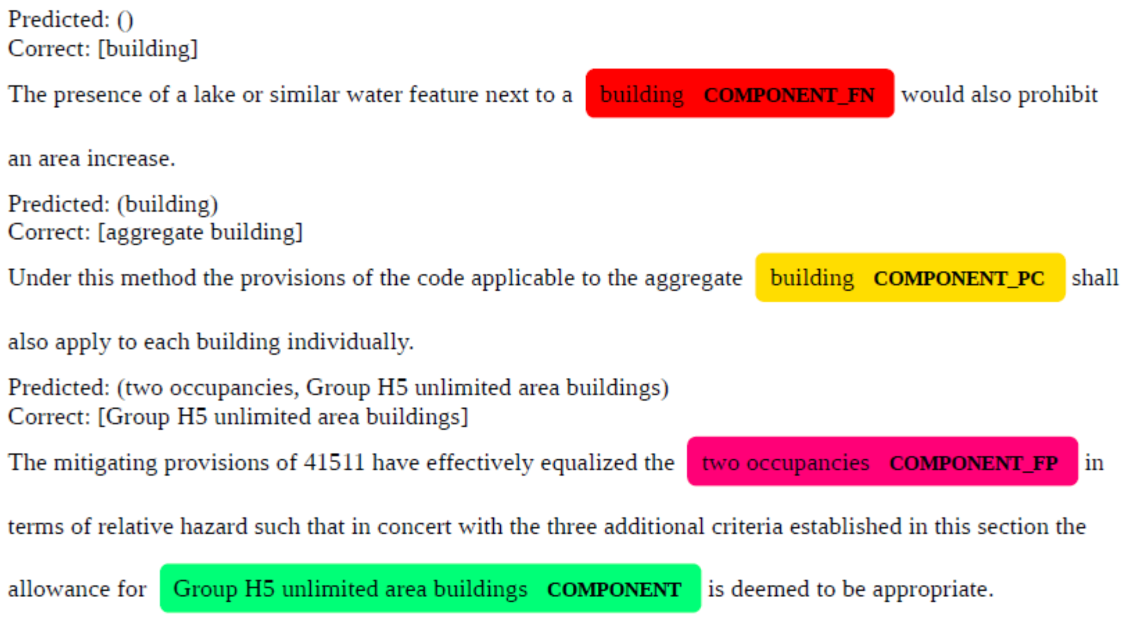
Figure – 14 Confusion Matrix

As it can be seen in Figure – 14, my model predicts 79 components correctly. 23 components couldn't be recognized by the model and those are false-negatives. 8

components are classified as a component but actually they are not components and those are false-positives.

When we look at the scores, the model’s precision is good. 90% precision shows us that most of the predicted components are correct. The recall which defines how many actual components are classified correctly, is a bit low and it can be improvable. Overall, the F-1 score which is the harmonic mean of Precision and Recall is also good enough to use the model in the real world.

My other visualizer script helps us to investigate the test dataset to get better insight about the result.



Predicted: ()
Correct: [building]

The presence of a lake or similar water feature next to a building **COMPONENT_FN** would also prohibit an area increase.

Predicted: (building)
Correct: [aggregate building]

Under this method the provisions of the code applicable to the aggregate building **COMPONENT_PC** shall also apply to each building individually.

Predicted: (two occupancies, Group H5 unlimited area buildings)
Correct: [Group H5 unlimited area buildings]

The mitigating provisions of 41511 have effectively equalized the two occupancies **COMPONENT_FP** in terms of relative hazard such that in concert with the three additional criteria established in this section the allowance for Group H5 unlimited area buildings **COMPONENT** is deemed to be appropriate.

Figure – 15 Visualizer of the Test Dataset

As it can be seen in Figure – 15, the visualizer shows us the components’ status of each sentence. The red label´s meaning is that the model couldn’t recognize the component and also is labeled as a false-negative. The yellow label´s meaning is that the model partly recognized the component. The pink label´s meaning is that the model recognized the word group as a component but actually it is not a component. That represents false-positives. The green label´s meaning is that the model predicts the component correctly. With this visualizer, we can understand the model’s prediction mistakes and what we can do to overcome those mistakes.

### 3.3.1 Investigation of False-Negative Entities

When I look into the false negative entities detailedly, I created a chart to see which components couldn't be recognized by the model.

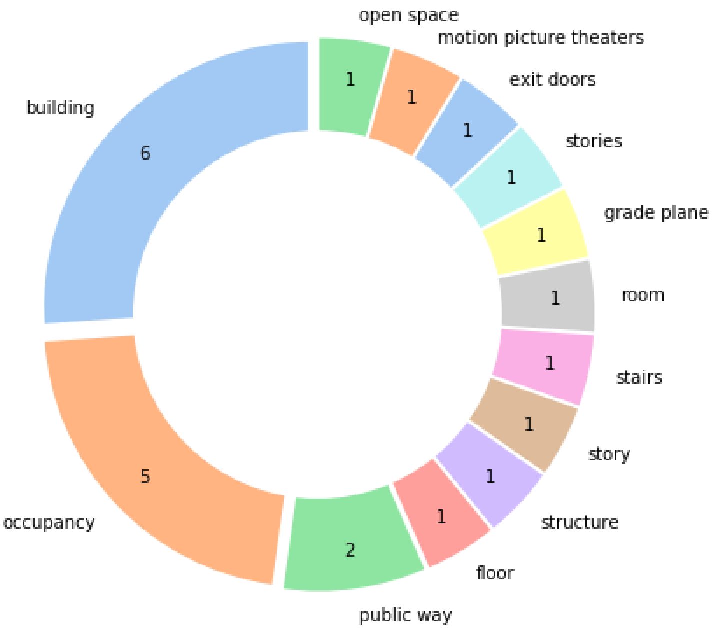


Figure – 16 False-Negative Components

As it can be seen in Figure – 16, half of the false-negative components' count is 1 and most of them didn't exist in the training dataset. This could be one of the reasons that they couldn't be classified. Regarding the other half, when we looked at the sentences, some of the components involved the "apostrophe (')" character. To overcome this issue, I can suggest that you can add removal of apostrophe in the preprocessing phase.

### 3.3.2 How to Improve the Model

As I mentioned in the previous section, we need to analyze the test results in detail to improve the model. From my observations, this is what can be done to improve the model for further progress:

- The dataset size would be increased by adding more sentences from regulation.
- The dataset consists mostly of "building, open space, wall" components. Adding more from other components to make a balanced dataset would affect directly the model prediction result.
- Preprocessing phase can be reviewed to add the removal of apostrophe characters.

# 4 Further Progress and Conclusion

In order to achieve the automated code compliance checks, we need to extract properties of the components and the rules belonging to them. Further progress should be to extract the properties using the same methodology and then combine these two to provide a rough representation of the natural text rule in a computer readable way.

In summary, we have shown a new approach and made initial integration with NLP to automate code compliance checks. NLP usage in construction is a new and wide area. In recent years, both technical development and accessing massive data leads to a major improvement in NLP. Likewise with the help of BIM, the construction field is going through a technological transformation. This area has still much research ongoing and it proves that it needs many more ways to make a fully automated compliance check. However our research shows us NLP could be a key factor to do this. Our project result seems promising and it can be one of the applications in the future.

## References

[1] [Abualdenien et al., 2019] Abualdenien, J., & Borrmann, A. (2019). A meta-model approach for formal specification and consistent management of multi-LOD building models. *Advanced Engineering Informatics*, 40, 135–153. https://doi.org/10.1016/j.aei.2019.04.003

[2] [Cornelius et al., 2015] Preidel C. and Borrmann A. (2015) "Automated Code Compliance Checking Based on a Visual Language and Building Information Modeling". *2015 Proceedings of the 32nd ISARC, Finland, ISBN 978-951-758-597-2.*

[3] [Lee et al., 2020] Lee J.K., Kim J., Lee G., Choi J., Kim I. and Lee Y. (2020) "Extracting Implementable Methods from the Predicates in Korean Building Act Sentences for Automated Code Compliance Checking". *Journal of Computational Design and Engineering - JCDE-2020-227*.

[4] [Ren et al., 2017] Ren Q, Cheng H, Han H. (2017) Research on machine learning framework based on random forest algorithm, AIP Conference Proceedings, AIP Publishing, https://doi.org/10.1063/1.4977376

[5] [Wang et al., 2020] Y. Wang, Y. Sun, Z. Ma, L. Gao, Y. Xu and T. Sun, "Application of Pre-training Models in Named Entity Recognition," *2020 12th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, Hangzhou, China, 2020, pp. 23-26, doi: 10.1109/IHMSC49165.2020.00013

[6] [Stroh et al., 2016] Stroh, E., & Mathur, P. (2016), Tuestion Answering Using Deep Learning, accessed 12.01.2021, <https://cs224d.stanford.edu/reports/Stroh-Mathur.pdf/>

[7] [Zola A, 2020] Andrew Zola, April 2020, Simple Question Answering (QA) Systems That Use Text Similarity Detection in Python, digital image, accessed 12.01.2021, <https://www.kdnuggets.com/2020/04/simple-question-answering-systems-text-similarity-python.html>

[8] [Rajpurkar P., 2017] Pranav Rajpurkar, April 2017, "The Stanford Question Answering Dataset Background, Challenges, Progress", digital image, accessed 12.01.2021, <https://rajpurkar.github.io/mlx/qa-and-squad/>

[9] Spacy IO "Get Started - Named Entities", digital image, accessed 12.01.2021, <https://spacy.io/usage/spacy-101>

[10] [Gomez C., 2020] Cristian Gomez Sep 24, 2020, Transfer Learning, digital image, accessed 12.01.2021, <https://medium.com/@1154_75881/transfer-learning-628e83df5c8a>

[11] [Saluja A.] Anmol Saluja, Text Preprocessing in Python using spaCy library, digital image, accessed 12.01.2021, <https://iq.opengenus.org/text-preprocessing-in-spacy/>

[12] [Douglas et al., 2018] Thornburg D. W., Kimball C. and Bracken W. C. "2018 International Building Code Illustrated Handbook". pages 210-211. *ISBN: 978-1-26-013229-8.*

## Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich den vorliegenden IDP-Report selbstständig angefertigt habe. Es wurden nur die in der Arbeit ausdrücklich benannten Quellen und Hilfsmittel benutzt. Wörtlich oder sinngemäß übernommenes Gedankengut habe ich als solches kenntlich gemacht.

Ich versichere außerdem, dass die vorliegende Arbeit noch nicht einem anderen Prüfungsverfahren zugrunde gelegen hat..

München, 16. February 2021

Andaç Kürün

Andaç Kürün