Determining the semantic similarity of definitions by artificial intelligence for the needs of 3D Land Administration: a case of building units

Martin VANĚK, Karel JANEČKA, and Otakar ČERBA, Czech Republic

Key words: semantic similarity, artificial intelligence, NLP, cosine similarity, LADM

SUMMARY

This article focuses on comparing selected terms connected to the building units across different standards and on testing the possibility of using the calculation of semantic similarity by artificial intelligence in the field of 3D Land Administration.
In the first chapter, the reader is introduced to the issue itself. The second chapter is devoted to determining semantic similarity. First, it is mentioned that it is necessary to use NLP and the related text preprocessing steps. Subsequently, the methods of converting the text into numerical form (vectorization) and then the actual methods of calculation of the semantic similarity between two texts are described. In the third chapter, the reader will learn which standards and the corresponding terms and definitions have been selected. Subsequently, the results of the calculation of semantic similarity using artificial intelligences from OpenAI and Hugging Face are presented here. Finally, the results are compared with each other. In the last chapter, the entire text is summarized and some problems and possibilities for further research are discussed here.

Determining the semantic similarity of definitions by artificial intelligence for the needs of 3D Land Administration: a case of building units

Martin VANĚK, Karel JANEČKA, and Otakar ČERBA, Czech Republic

1. INTRODUCTION

In recent years, artificial intelligence has undergone significant development, and its influence is gradually beginning to manifest itself in an increasing number of different fields. The basic premise of artificial intelligence is that it can understand natural language correctly. In order for this to be possible, it is necessary to use the so-called Natural Language Processing (NLP). NLP is key to converting texts into numerical form and then determining semantic similarity, which is the calculation of a numerical value that determines how similar the texts are in meaning. With such a value, the computer can already work and it is important, in addition to artificial intelligence, for example, in various Internet search engines, in text generation and in a number of other diverse applications.

A problem that also occurs within the 3D Land Administration, for example, is that there are a number of different definitions for individual terms. This can be the cause of a number of misunderstandings, when anyone can understand something different by the given term. The need to talk in a unified language is especially crucial in the case of interconnecting two or more domains together, i.e. BIM/IFC, 3D Land Administration/ISO 19152 and GIS/CityGML. The problem can also occur in the case of various national legislations when each country could use different terms for the same thing/object.

In order to avoid this, it is also possible to use the calculation of semantic similarity, which will help to compare the agreement between individual definitions and allow to find the ones that are the most or, conversely, the least similar. Definitions that have a high semantic similarity are less likely to cause misunderstanding than those that are only minimally similar. Since every artificial intelligence must have some model implemented to calculate semantic similarity, it seems like a good option to use it to solve this problem. In order to calculate the semantic similarity, it is necessary to convert the text into numerical form, so that this is possible, it is necessary to use NLP first. Subsequently, it is already possible to proceed with the conversion of the text into numerical form (into the form of a vector). There are a number of different models that convert text to vector for this purpose. After the text is converted to a vector, the semantic similarity between the two texts (vectors) is calculated. For this purpose, it is possible to use the semantic similarity calculation model used by some of the artificial intelligences.

Artificial intelligence can be used to compare definitions with each other, which can have a significant positive effect on the field of 3D Land Administration as well. This article focuses on comparing selected terms connected to the building units across different standards. Which can help prevent future misunderstandings. There are several publications that focus on the issue of semantic similarity calculation with the use of artificial intelligence, but none of them have yet dealt with this issue for comparing terms and definitions in the field of land administration. The aim of this work is to test the possibility of using the calculation of semantic similarity by artificial intelligence in the field of 3D Land Administration.

2. DETERMINING SEMANTIC SIMILARITY

This chapter deals with the method of calculating the semantic similarity between two texts and the steps that precede the calculation itself. The basic idea is to convert text into a form that a computer can understand. It is necessary to preprocess the text based on NLP and then convert it into numerical (vector) form. In this chapter, the individual steps of preprocessing are described, followed by some models that are used for conversion into vector form and, finally, some options for calculating the textual similarity between two texts.

2.1 Natural Language Processing

To calculate semantic similarity, it is first necessary to convert the text into a form that a computer can understand. The so-called Natural Language Processing (NLP) is used for this purpose. It is a discipline that deals with the conversion of text into a computer-intelligible (usually numerical) form (Xieling et al., 2022). This is particularly advantageous for the automation of certain processes and essential for the correct functionality of artificial intelligence (AI) (Khurana et al., 2023).

In NLP, several basic steps are used during text preprocessing, which enable subsequent easier work with the text when converting it into numerical form (Kadhin, 2018). These are:

- Normalization: This step focuses on editing the text so that it is converted to a basic form. In the resulting form of the text, all letters should be lowercase, there should be no special characters, links, punctuation, numbers, etc. in the text (Pal, 2021).
- Tokenization: The goal of this step is to divide the text into certain continuous sequences of characters that have semantic meaning. This usually involves dividing the text into individual words, so-called tokens (Petrović and Stanković, 2019).
- Stop word removal: This step deals with the removal of so-called stop words. These are words that occur frequently in the text and do not add any new meaning to it. These are mainly conjunctions and prepositions. The word "and" can be an example of such a word (Denny and Spirling, 2018).
- Stemming: This step is aimed at finding the root of words by removing prefixes and suffixes (Hickman et al., 2022).
- Lemmatization: In this case, the corresponding lemma is searched for the word. Unlike stemming, meaning is also considered. Especially if it is a verb, a noun, etc. For example, the word "set" can be a verb in some cases and a noun in others (Chai, 2023).

2.2 Vectorization

Since computers understand numbers best, after preprocessing it is necessary to convert the text into numerical form, preferably into vector form. This process is called text vectorization or text embeddings (Rani et al., 2022). There are several different models used to vectorize text, only some of the most well-known models are described here:

- Bag of Words (BoG): This is the simplest method that is based on word count only. The resulting vector has a length corresponding to the number of different words in the text, and each value of the vector corresponds to the number of occurrences of the given word in the text. This approach does not consider the context in which

individual words are found, and therefore the subsequent calculation of semantic similarity can be quite inaccurate (Rani et al., 2022).

- Word2vec: This model already considers the context in which the words are found. It is based on a neural network using a training process. The resulting vectors of semantically similar words will have a strong relationship with each other and will be close to each other in the vector space. It follows that the vectors created by this method also contain semantic information (Yang et al., 2022).
- Global Vector for word representation (GloVe): It is a log-bilinear regression model and unsupervised learning model. Compared to Word2vec, it is faster and provides more accurate results (Pennington et al., 2014). This model is based on the creation of a matrix, the rows of which correspond to individual words and contain their vector representations, the columns of the matrix subsequently correspond to the different contexts in which the words can occur (Rani et al., 2022).

2.3 Computing semantic similarity

After the text is converted to a vector, the semantic similarity between the two texts (vectors) is calculated. This can be calculated in several possible ways. For example, the Euclidean distance calculation can be used, where the distance between two vectors is calculated. The formula for calculating the Euclidean distance is as follows:

$$d(\mathbf{p},\mathbf{q}) = d(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2},$$

where **p** and **q** are vectors of compared texts and n is the number of vector values (Pal, 2021). A more commonly used and more accurate method is cosine similarity. In this case, the cosine of the angle between two vectors is measured using the following formula:

$$cos\,\theta = \frac{\mathbf{p}\cdot\mathbf{q}}{\|\mathbf{p}\|\|\mathbf{q}\|} = \frac{\sum_{i=1}^{n} p_i q_i}{\sqrt{\sum_{i=1}^{n} p_i^2}\sqrt{\sum_{i=1}^{n} q_i^2}},$$

where **p** and **q** are vectors of compared texts and n is the number of vector values (Pal, 2021). The result of this calculation is the angle between the given vectors. If it is zero, the cosine similarity is equal to one, and such two vectors have the highest similarity (100%). Conversely, if the angle between the vectors is 90°, the cosine similarity is zero and such vectors have the lowest similarity (0%) (Pal, 2021).

3. METHODOLOGY

This chapter is devoted to the calculation of semantic similarity using AI on the example of terms and definitions from the field of 3D Land Administration. The results are also compared with each other.

3.1 Selection of terms and definitions

First it was necessary to determine the standards that will be used and then to select from them the terms and corresponding definitions from the area of 3D Land Administration related to building units. Land Administration Domain Model (LADM) was chosen as the

basic standard, the terms of which will be compared with the terms of other standards. Terms and definitions from the Industry Foundation Classes (IFC 4.3), Land and Infrastructure Conceptual Model Standard (LandInfra) and City Geography Markup Language (CityGML 3.0) standards were then selected for comparison.
Terms and definitions related to building units were selected from the above-mentioned standards and these were subsequently recorded in individual tables (see tables 1 – 4). The definitions were selected from part 3 of the IFC 4.3 standard, and from part 4 of the LADM, InfraGML and CityGML 3.0 standards.

Table 1 – terms and definition from LADM (ISO, 2012)

Land Administration Domain Model (LADM)		
1	basic administrative unit	Administrative entity, subject to registration (by law), or recordation [by informal right, or customary right, or another social tenure relationship], consisting of zero or more spatial units against which (one or more) unique and homogeneous rights [e.g. ownership right or land use right], responsibilities or restrictions are associated to the whole entity, as included in a land administration system.
2	boundary	Set that represents the limit of an entity.
3	boundary face	Face that is used in the 3-dimensional representation of a boundary of a spatial unit.
4	boundary face string	Boundary forming part of the outside of a spatial unit.
5	building unit	Component of building (the legal, recorded or informal space of the physical entity).
6	land	The surface of the Earth, the materials beneath, the air above and all things fixed to the soil.
7	spatial unit	Single area (or multiple areas) of land and/or water, or a single volume (or multiple volumes) of space.

Table 2 – terms and definitions from IFC 4.3 (ISO, 2024)

IFC 4.3		
1	building information modelling	Use of a shared digital representation of an asset to facilitate design, construction and operation processes to form a reliable basis for decisions.
2	element	Physical object with a stated function, form and position.
3	entity	Class of information defined by common properties.
4	facility	Physical structure, including the related site, serving one or more main purposes.
5	feature	Conceptualization of certain design or manufacturing functionality to implicitly alter the geometric form of an element to be computed at import.
6	model	Collection of entity data type instances.

7	object	Any part of the perceivable or conceivable world.
8	product	Thing or substance produced by a natural or artificial process.
9	property	Defined characteristic suitable for the description and differentiation of an object.
10	property set	Named set of properties grouped under some characteristics.
11	representation	Organized collection of associated data elements, collected together for one or more specific uses.
12	space	Limited three-dimensional extent defined physically or notionally.

Table 3 – terms and definitions from LandInfra (OGC, 2016)

Land and Infrastructure Conceptual Model Standard (LandInfra)		
1	administrative division	Division of state territory according to political, judicial, or executive points of view.
2	building	Construction works that has the provision of shelter for its occupants or contents as one of its main purposes, usually partially or totally enclosed and designed to stand permanently in one place.
3	building part	Floor-related part of a multi-storage building, subdivided according to management and use by a lawful process.
4	boundary	Set that represents the limit of an entity.
5	condominium	Concurrent ownership of real property that has been divided into private and common portions.
6	construction	Assembled or complete part of construction works that results from work on-site.
7	facility	Improvements of or on the land including buildings and civil engineering works and their associated siteworks.
8	feature	Abstraction of real world phenomena.
9	interest in land	Ownership or security towards real property.
10	land	Area of earth's surface, excluding the oceans, usually marked off by natural or political boundaries, or boundaries of ownership.
11		The surface of the Earth, the materials beneath, the air above and all things fixed to the soil.
12	land parcel	Contiguous part of the surface of the Earth (land and/or water) as specified through lawful process.
13	ownership (in land)	Includes the right to grant a lease, an easement, or a security interest and other lesser rights.
14	physical element	Any component defined within the spatial and functional context of a facility.
15	positioning element	Virtual element used to position, align, or organize physical elements.
16	product	Item manufactured or processed for incorporation in construction works.
17	retaining wall	Wall that provides lateral support to the ground or that resists

		pressure from a mass of other material.
18	site	Area of land or water where construction work or other development is undertaken.
19	spatial unit	Contiguous geometrical entity, which is delimited and located on or close to the surface of the Earth through the bounding elements of its boundary.
20	wall	Vertical construction that bounds or subdivides a space and usually fulfils a loadbearing or retaining function.

Table 4 – terms and definitions from CityGML 3.0 (OGC, 2023)

CityGML 3.0		
1	2D data	Geometry of features is represented in a two-dimensional space.
2	2.5D data	Geometry of features is represented in a three-dimensional space with the constraint that, for each (X,Y) position, there is only one Z.
3	3D data	Geometry of features is represented in a three-dimensional space.
4	city-object relation	Specific relation from the city object in which the relation is included to another city object.
5	feature	Abstraction of real world phenomena.
6	geometry	An ordered set of n-dimensional points in a given coordinate reference system; can be used to model the spatial extent or shape of a feature.
7	life-cycle information	Set of properties of a spatial object that describe the temporal characteristics of a version of a spatial object or the changes between versions.
8	space	Entity of volumetric extent in the real world.
9	space boundary	Entity with areal extent in the real world. Space boundaries are objects that bound a Space. They also realize the contact between adjacent spaces.
10	top-level feature	Feature that represents one of the main components of 3D city models; can be further semantically and spatially decomposed and substructured into parts.

3.2 Computing semantic similarity by artificial intelligence

Subsequently, it was already possible to proceed with the calculation of semantic similarities between the individual definitions of the given terms. The text-embedding-ada-002 model, which uses AI from OpenAI, was used for the calculation. To compare the approaches of different AIs, AI from Hugging Face was also used to calculate semantic similarity.

Individual definitions were passed to these artificial intelligences and the results were semantic similarity values between the given terms. The resulting values were recorded in tables, where each column was coloured separately with a colour transition from red to green, where red indicates the lowest value of semantic similarity in the given column, while green indicates the highest (see tables 5 – 7 for results from OpenAI AI and tables 8 – 10 for results from Hugging Face AI).

Table 5 – semantic similarity results from OpenAI AI between LADM and IFC 4.3 (results are in percentages)

IFC 4.3 \ Land Administration Domain Model (LADM)	1	2	3	4	5	6	7
1	76,98	70,95	76,28	74,42	79,34	72,99	77,41
2	77,91	74,00	81,03	80,45	81,16	79,88	80,84
3	77,92	73,32	73,99	74,63	78,18	72,84	76,09
4	79,86	70,83	79,30	81,43	85,29	78,98	83,41
5	73,76	70,19	76,83	76,50	76,62	71,97	74,41
6	74,57	75,70	71,09	71,81	72,88	69,22	72,61
7	77,89	72,57	79,43	81,62	82,86	81,65	81,79
8	75,45	69,25	74,96	75,07	77,44	78,79	78,05
9	76,64	73,57	77,38	76,24	75,50	73,27	76,86
10	78,78	75,36	76,78	77,57	77,45	76,13	79,85
11	80,71	72,27	75,66	76,96	79,06	74,12	81,15
12	80,21	78,40	84,51	83,64	80,84	79,49	85,41

Table 6 – semantic similarity results from OpenAI AI between LADM and LandInfra (results are in percentages)

LandInfra \ Land Administration Domain Model (LADM)	1	2	3	4	5	6	7
1	80,16	72,31	79,65	81,31	77,67	75,75	82,19
2	77,44	69,45	75,14	78,29	82,65	76,42	80,15
3	81,72	71,16	77,88	79,57	85,85	76,57	84,27
4	74,87	100,00	76,71	76,70	74,84	69,55	72,84
5	80,96	70,39	74,40	76,70	79,59	73,14	80,05
6	76,58	69,86	77,56	80,17	83,36	77,43	78,70
7	78,21	68,28	73,88	76,34	81,28	78,40	80,54
8	75,56	72,39	77,82	77,09	76,46	78,09	77,40
9	82,35	72,49	75,32	76,14	81,85	76,20	79,81
10	81,71	73,69	82,03	84,88	81,35	84,38	86,80
11	75,82	69,55	78,02	78,16	77,42	100,00	82,46
12	82,06	73,40	82,07	84,58	83,03	84,21	86,72
13	81,76	70,94	72,63	74,19	77,88	74,33	77,92
14	80,13	73,03	81,09	83,16	87,04	77,52	82,87
15	75,47	72,34	80,54	79,30	81,99	75,61	79,84
16	76,59	70,55	76,85	76,41	80,74	76,66	77,54
17	75,41	72,74	79,90	80,16	79,37	79,26	79,55
18	81,19	71,93	78,70	81,67	83,42	78,59	87,24
19	82,68	77,38	84,94	85,84	82,08	82,46	83,93
20	77,57	72,28	80,28	81,00	83,76	75,87	81,93

Table 7 – semantic similarity results from OpenAI AI between LADM and CityGML 3.0 (results are in percentages)

CityGML 3.0 \ Land Administration Domain Model (LADM)	1	2	3	4	5	6	7
1	74,96	71,61	84,21	82,67	78,29	78,37	82,58
2	76,6	73,87	84,65	80,20	77,27	78,95	82,48
3	75,01	71,23	86,74	81,97	78,21	78,83	81,88
4	71,78	73,11	73,07	75,51	73,93	70,06	71,69
5	75,56	72,39	77,82	77,09	76,46	78,09	77,36
6	79,06	74,68	84,84	84,53	78,54	76,08	82,53
7	79,83	74,89	78,98	80,22	80,38	75,04	79,92
8	80,58	78,94	84,40	82,34	81,74	78,89	84,80
9	81,21	77,36	83,17	85,15	83,01	78,96	84,02
10	78,38	72,83	85,52	82,88	82,64	75,16	80,53

Table 8 – semantic similarity results from Hugging Face AI between LADM and IFC 4.3 (results are in percentages)

IFC 4.3 \ Land Administration Domain Model (LADM)	1	2	3	4	5	6	7
1	22,40	12,30	19,40	13,80	43,10	1,80	18,30
2	24,60	23,90	39,90	33,00	41,30	24,90	22,80
3	29,60	38,20	12,80	10,40	31,30	12,10	22,20
4	22,60	19,70	19,80	27,20	57,50	36,40	38,50
5	9,00	4,80	32,30	19,70	34,00	14,40	13,30
6	39,60	47,00	7,10	6,60	26,90	6,20	17,60
7	11,40	22,20	18,30	25,10	21,40	22,60	32,20
8	16,40	13,30	9,00	17,40	35,70	25,30	7,50
9	30,40	31,10	35,80	21,70	34,70	12,30	16,20
10	29,20	45,90	23,60	23,80	40,10	19,30	27,90
11	27,40	26,10	13,40	8,30	35,50	14,10	25,90
12	21,00	27,70	54,40	35,50	31,90	15,10	39,20

Table 9 – semantic similarity results from Hugging Face AI between LADM and InfraGML (results are in percentages)

InfraGML/Lan \ Land Administration Domain Model (LADM)	1	2	3	4	5	6	7
1	28,30	19,70	28,30	38,00	31,90	14,80	43,00
2	19,40	11,90	20,40	28,60	55,80	24,10	24,40
3	33,90	15,70	21,10	28,50	51,80	19,30	39,40
4	43,70	100,00	20,10	18,00	29,20	8,70	23,10
5	35,10	25,70	11,10	26,20	38,90	8,00	26,60

	6	12,70	4,50	7,70	17,90	47,20	25,50	15,80
	7	22,40	12,40	10,00	14,10	36,50	35,40	18,90
	8	15,40	8,40	29,60	25,20	28,40	16,10	17,10
	9	39,60	23,10	14,70	20,00	38,90	10,30	16,10
	10	32,20	20,20	36,40	44,50	28,60	50,70	58,10
	11	16,30	8,70	15,20	24,20	25,00	100,00	44,10
	12	28,40	23,70	35,70	49,10	36,90	53,20	55,80
	13	35,20	15,30	13,90	18,10	28,00	16,30	20,80
	14	35,90	33,40	37,70	38,90	63,30	22,90	40,60
	15	23,00	19,00	29,80	26,10	50,90	27,50	30,50
	16	10,90	10,40	5,70	15,30	51,90	12,40	4,80
	17	7,10	11,30	29,80	34,80	37,80	39,00	26,50
	18	22,60	9,90	18,30	30,70	46,40	41,00	50,50
	19	35,10	34,60	44,30	54,20	38,20	42,70	48,80
	20	13,10	23,00	34,20	38,90	38,10	24,00	30,80

Table 10 – semantic similarity results from Hugging Face AI between LADM and CityGML 3.0 (results are in percentages)

		Land Administration Domain Model (LADM)						
		1	2	3	4	5	6	7
CityGML 3.0	1	19,80	16,20	51,80	35,70	30,70	9,30	34,00
	2	13,60	18,70	40,10	16,90	18,40	8,80	28,20
	3	16,00	15,90	59,50	28,40	29,40	13,00	33,50
	4	20,90	25,80	24,50	25,10	32,60	13,40	19,70
	5	15,40	8,40	29,60	25,20	28,40	16,10	17,10
	6	23,10	27,60	44,20	27,90	25,10	17,40	34,20
	7	36,80	34,00	34,90	31,30	43,70	21,20	29,20
	8	20,40	18,20	42,30	37,10	26,50	26,40	50,70
	9	41,20	36,60	42,50	53,90	39,50	27,80	48,80
	10	15,30	12,60	45,60	29,60	44,80	20,40	37,70

3.3 Comparison of results

Thanks to the coloring of the resulting values of semantic similarity, it is possible to estimate at first glance that the results are meaningful and that both artificial intelligences provide similar results. The difference in that the resulting values from the OpenAI AI are in the range of approximately 70 % – 100 % and the resulting values from the Hugging Face AI in the range of 0 % – 100 % is caused by the model that the given AI uses to calculate semantic similarity.

When comparing the numerical values, it turns out that both artificial intelligences determine the same two definitions for 15 out of 21 cases as the definitions with the greatest similarity. This is 71,4 % of cases. An example can be the definitions of the term spatial unit in LADM and space in IFC 4.3, when the similarity between these two definitions reaches 85.41 % in the case of calculation using OpenAI AI and 39.20 % in the case of calculation using Hugging

Face. In both cases, this is the greatest similarity between the individual definitions of terms from IFC 4.3 and the term spatial unit from LADM. Such results can be very interesting, especially when mapping the classes of individual standards to each other and during the subsequent transition from one standard to another. In this case, the result shows that spatial units from LADM, which can be for example rooms or property spaces, will be mapped to a class in IFC 4.3 that supports spaces (i.e. the IfcSpace class) (ISO, 2024).
Otherwise, when determining the definitions with the least similarity, the two AIs agree on only 4 out of 21 terms. This is 19 % of cases. An example can be the definition of boundary face from LADM and city-object relation from CityGML 3.0, where the similarity between these two definitions reaches 73.07 % in the case of calculation using OpenAI AI and 24.50 % in the case of calculation using Hugging Face AI. In both cases, this is the smallest similarity between the individual definitions from CityGML 3.0 and the boundary face definition from LADM. The differences in the resulting values are because the two AIs use a different model to calculate semantic similarity. Overall, when comparing the results of semantic similarity for a specific term, it can be stated that the results of both AIs are similar, since, for example, if the similarity between two specific definitions is low for the first AI it can be expected that the result of semantic similarity between the same definitions will reach a low value also in the case of the second AI.
In some cases, it may happen that even if two different standards define the same terms, it may turn out that their definitions do not have the highest semantic similarity. An example of this can be the term spatial unit, which occurs both in LADM and in InfraGML. The results show that even though both standards use the same term, their definitions do not have the highest similarity. Spatial unit from LADM has greater semantic similarity with terms site and land than with spatial unit from InfraGML.
Furthermore, for some terms, an own calculation of semantic similarity was performed using the BoG method (which is based on word counting and does not consider the context) and the subsequent calculation of cosine similarity. Since the given definitions do not contain many identical words, the results are quite imprecise. For example, for the terms spatial unit from LADM and space from CityGML 3.0, the semantic similarity from our own calculation was only 20.52 %, although according to the results from calculations by AI it should be high. The same case is the terms spatial unit from LADM and space from IFC 4.3, whose semantic similarity is high according to AI, but since the given definitions do not contain any of the same words, the semantic similarity from our own calculation came out to be 0 %. It follows that methods that do not take context into account are highly inaccurate.
By verifying that the semantic similarity calculation is going well, the semantic similarity result can be for the term boundary, which has the same definition in both LADM and LandInfra. As expected, the AI-calculated semantic similarity was 100 % in both cases.

4. DISCUSSION AND CONCLUSION

The method of determining semantic similarity was described within the text. First, it was mentioned that it is necessary to use NLP and the related text preprocessing steps. Subsequently, the methods of converting the text into numerical form (vectorization) and then the actual calculation of the semantic similarity between two texts were described. The next

chapter was devoted to the methodology of calculating semantic similarity using AI between selected terms from the field of 3D Land Administration. Finally, the resulting values were compared.

The results show that the calculation of the semantic similarity of the texts was successful in all cases. The differences between individual results are because each AI uses a different model to calculate semantic similarity. For example, even if the results of the AIs do not agree on which two definitions have the least similarity to each other, it can be expected that if one AI comes up with a low similarity value between two definitions, then the other AI will also come out with a low similarity value for these definitions. From the results, it can also be concluded that for determining the semantic similarity of definitions, it is more appropriate to use calculations that also consider the context. Otherwise, two definitions may have a high degree of similarity in meaning, but the resulting value will be zero due to the absence of common words. In conclusion, it is possible to state that artificial intelligence can help with comparing selected terms connected to the building units across different standards. Which can subsequently significantly help when trying to map different classes from different standards onto each other and prevent possible misunderstandings not only in the field of 3D Land Administration.

A problem that can occur when mapping between different standards is that it may happen that the same terms are not the closest in meaning in the two standards and then it is necessary to determine which two terms to use for mutual mapping. Further research would be needed for this. This work was understood as an introduction to the issue of comparing selected terms from the field of 3D Land Administration connected to the building units across different standards using AI. Furthermore, it would be advisable to carry out research that would verify the usability of the calculation of semantic similarity by artificial intelligence for the needs of class mapping between individual standards, which would be useful, for example, during the conversion between IFC 4.3 and CityGML 3.0.

REFERENCES

Chai, C. P. (2023). Comparison of text preprocessing methods. Natural Language Engineering. Volume 29. Number 3. Pages 509 – 553. Cambridge University Press.

Denny, M. J., Spirling, A. (2018). Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It. Political Analysis. Volume 26. Number 2. Pages 168 – 189. Cambridge University Press.

Hickman, L., Thapa, S., Tay, L., Cao, M., Srinivasan, P. (2022). Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations. Organizational Research Methods, Volume 25. Number 1. Pages 114 – 146.

ISO (2012). ISO 19152, Geographic information – Land Administration Domain Model (LADM). Geneva, Switzerland.

ISO (2024). ISO 16739-1, Industry Foundation Classes (IFC) for data sharing in the construction and facility management industries. buildingSMART.

Kadhin, A. I. (2018). An Evaluation of Preprocessing Techniques for Text Classification. International Journal of Computer Science and Information Security (IJCSIS). Volume 16. Number 6. Baghdad, Iraq.

Khurana, D., Koli, A., Khatter, K., Singh, S. (2023). Natural language processing: state of the art, current trends and challenges. Multimedia Tools and Applications. Volume 82. Pages 3713 – 3744.
OGC (2016). OGC Land and Infrastructure Conceptual Model Standard (LandInfra). Open Geospatial Consortium.
OGC (2023). OGC City Geography Markup Language (CityGML) Part 2: GML Encoding Standard. Open Geospatial Consortium.
Pal, S. (2021). What is Text Similarity and How to Implement it. Microsoft Learn Student Ambassador - KIIT Chapter.
Pennington, J., Socher, R., Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. Conference on Empirical Methods in Natural Language Processing (EMNLP). Pages 1532 – 1543. Doha, Qatar.
Petrović, Đ., Stanković, M. (2019). The influence of text preprocessing methods and tools on calculating text similarity. Facta Universitatis. Volume 34. Number 5. Pages 973 – 994. Niš, Serbia.
Rani, D., Kumar, R., Chauhan, N. (2022). Study and Comparision of Vectorization Techniques Used in Text Classification. 13th International Conference on Computing Communication and Networking Technologies (ICCCNT). Pages 1 – 6. Kharagpur, India.
Xieling, C., Haoran, X., Xiaohui, T. (2022). Vision, status, and research topics of Natural Language Processing. Natural Language Processing Journal. Volume 1.
Yang, X., Yang, K., Cui, T., Chen, M., He, L. (2022). A Study of Text Vectorization Method Combining Topic Model and Transfer Learning. Processes. Volume 10. Number 2. Page 350.

BIOGRAPHICAL NOTES

Martin VANĚK, Ing. in 2023 graduated in Geomatics at the Faculty of Applied Sciences of the University of West Bohemia in Pilsen (Czech Republic). He is studying a PhD program focusing on 3D cadastre and BIM.

Karel JANEČKA has a Ph.D. (2009) Geomatics, University of West Bohemia in Pilsen. He had been working as a database programmer at the Czech Office for Surveying, Mapping and Cadastre in Section of cadastral central database between 2006 and 2008. Since 2009 he is a researcher at University of West Bohemia, Department of Geomatics. His research activities are spatial data infrastructures (SDI), geographical information systems (GIS), spatial databases, spatial data mining, and 3D cadastre. He has experience with coordination of several EU projects and is also reviewer of several international scientific journals. Since 2012 he is the president of the Czech Association for Geoinformation and member of National Mirror Committee 122 Geographic information/Geomatics. Since 2021 he is Head of Department of Geomatics at the University of West Bohemia, Czech Republic.

Otakar ČERBA, Ph.D. works at the Department of Geomatics (Faculty of Applied Sciences, University of West Bohemia, Plzeň, Czech Republic). He is focused on web cartography, thematic cartography, Linked Data on the geographic domain and semantic issues of geographic data. He has been involved in many international projects such as Humboldt, SDI4Apps, SmartOpenData, Plan4all or ROSIE. Otakar Čerba is the members of the board of Czech Cartographic Association and the chair of the Commission on Maps and the the Internet of International Cartographic Association.

CONTACTS

Martin Vaněk
University of West Bohemia
Technická 8
Pilsen
CZECH REPUBLIC
Phone: + 420 773 967 168
E-mail: vanekma@kgm.zcu.cz

Karel Janečka
University of West Bohemia
Technická 8
Pilsen
CZECH REPUBLIC
Phone: + 420 377 639 200
E-mail: kjanecka@kgm.zcu.cz

Otakar Čerba
University of West Bohemia
Technická 8
Pilsen
CZECH REPUBLIC
Phone: + 420 377 639 206
E-mail: cerba@kgm.zcu.cz