An Application of the Disease Ontology (DO) for Clustering COVID-19 Hospitalizations in Rio de Janeiro

Lucas Maddalena [1] and Fernanda Baião [1]

[1] *Department of Industrial Engineering, Pontifícal Catholic University of Rio de Janeiro, Rua Marquês de São Vicente 225, Rio de Janeiro, 22451-900, Brazil*

Abstract

On the 21st century, the exponential growth of technology, led the world facing a myriad of information coming from multitudinous sources. Then, finding ways of storing knowledge committed to certain rules became imperious.

Ontologies have been playing an important role on connecting data to the semantics of the real world. Data, without such ontological commitment, could be interpreted as representations of different entities than the one it actually is, leading to biased analysis and inaccurate prediction on data-driven projects. Such kind of artifact formalizes shared knowledge regarding a domain of discourse.

Therefore, this study will, based on works showing the benefits of bringing ontologies to the scenario of Machine Learning techniques, enrich similarity metrics between instances of data. So, the Human Disease Ontology (DO) will be used. Instead of calculating pairwise similarities between two diseases (terms on DO), groups of diseases will be considered. Therefore, this work will rely on adapting a groupwise similarity metric

Data collection will be done considering the SIVEP-Gripe Dataset. Then, an analysis will be made on how better Machine Learning Algorithms can perform the analysis is made considering semantic rather than just numerical and categorical features.

Keywords

Disease Ontology, COVID-19, Clustering

Introduction

In December 2019, the first case of coronavirus disease (COVID-19), caused by the SARS- CoV-2 virus, was reported. It did not take long for the disease to get enormous proportions and become a worldwide concern, and on March 11th, 2020, the World Health Organization (WHO) declared the disease outbreak a global pandemic [1].

COVID-19 is affecting the four corners of the world, and data is coming from a thousand-and-one different providers. Therefore, data integration in the COVID-19 domain can be compromised and semantic commitments shall be considered when treating pandemic data. As an illustration, in China, from Jan 15 until March 2, 2020, there have been seven different versions of the COVID-19 case definition issued by the government, and [2] estimate that the lack of a temporal consensus on the definitions led China official pandemic tracking to increase up to 7.1 times (IC 95%, 4.8 – 10.9) from one definition to another.

One of the main purposes of ontologies is to make the real-world data semantics explicit [3]; consequently, many benefits can be extracted by this kind of artifact, including its use as a communication artifact among different stakeholders, as a common data model to mediate data

Proceedings of the 15th Seminar on Ontology Research in Brazil (ONTOBRAS) and 6th Doctoral and Masters Consortium on Ontologies (WTDO), December 22-25, 2022.

EMAIL: lucasgm@tecgraf.puc-rio.br (A. 1); fbaiao@puc-rio.br (A. 2)
ORCID: 0000-0001-6411-4259 (A. 1); 0000-0001-7932-7134 (A. 2)

integration and access, or even as a formal specification to enable reasoning on data. In the COVID-19 domain, several works already proposed ontologies and applications, such as [4,5,6,7].

Recently, the multiple benefits of ontologies (including foundational ontologies, conceptual models, and other semantically aware artifacts) to enhance data analysis and knowledge extraction have been increasingly advocated. In this context, [8] present how ontologies, and specifically foundational ontologies, can have multiple benefits on every step of the internal cycle of the Data Science Life Cycle, while [9] show the benefits of pairing conceptual models with ontologies to Machine Learning (ML) techniques.

The present work focuses on data regarding the comorbidities (i.e., diseases) of patients who have been diagnosed with COVID-19 and were hospitalized in the state of Rio de Janeiro. The main objective is to analyze the impact of a semantically aware approach when finding similar subsets of hospitalizations in the dataset.

To this end, we apply a partition-based clustering technique and compared its results in two scenarios. The first scenario (semantic unaware) represented each hospitalization as a binary vector of comorbidities and applied the conventional cosine similarity metric. The second (semantic aware) scenario was proposed as follows.

Disease Ontology (DO) [10] is an ontology which integrates disease and medical vocabularies through extensive cross mapping of DO terms to other medical ontologies, such as MeSH.

We matched each comorbidity found in the dataset with a corresponding concept in the Disease Ontology (DO). A total of 161 distinct diseases were linked to DO concepts, and we observed 465 different combinations of diseases, for all the patients in the dataset.

To compute similarities between individual comorbidities, we applied the measure proposed by [11], which addressed semantics to find similarities between data, and specifically proposed a similarity metric in the bio-ontologies domain using DO terms. However, since each hospitalized patient was characterized by a (possibly empty) set of comorbidities in the dataset, the similarity between distinct hospitalizations required a groupwise similarity metric, i.e., measuring the similarity between two different groups of diseases, which represents the diseases a COVID-19 hospitalized patient has. For instance, while the pairwise metric performs a comparison between two terms such as “diabetes” and “asthma”, the groupwise similarity metric compares two sets of terms, such as “Diabetes, gilbert’s syndrome and flu” and “Psoriasis and AIDS”. Therefore, we applied the metric proposed by [12] for calculating groupwise similarities between sets of DO terms. On [12], it is calculated groupwise similarities between terms on the SNOMED CT.

Hence, the semantic aware groupwise similarity between hospitalizations proposed in our work was computed by combining the groupwise metric of [12] with the pairwise similarity between DO terms of [11].

The impact of the proposed semantically aware approach when finding similar subsets of hospitalizations in the dataset is assessed in the Data Post-Processing step using metrics of cluster quality. An additional analysis was performed to show how well the resulting clusters from each scenario partitioned the subsets of diseases.

Disease Ontology (DO)

In this research, we make use of the Human Disease Ontology (DO), a domain ontology organized as a directed acyclic graph, representing the domain of ontologies and is mapped to uncountable others application ontologies.

DO makes the knowledge on the domain of human diseases explicit, by describing diseases through ontology properties, such as *is-a*, *has-material-basis-in* or *has-symptom*. For instance, DO states that:

bone disease *is-a* **connective tissue disease**
congenital megabladder *has-material-basis-in* **autosomal dominant inheritance allergic**
conjunctivitis *has-symptom* **allergic reaction**.

Also, as shown on Figure 1, a term on DO can be linked to other ontologies through relations such as *has-symptom* and *has-phenotype*.

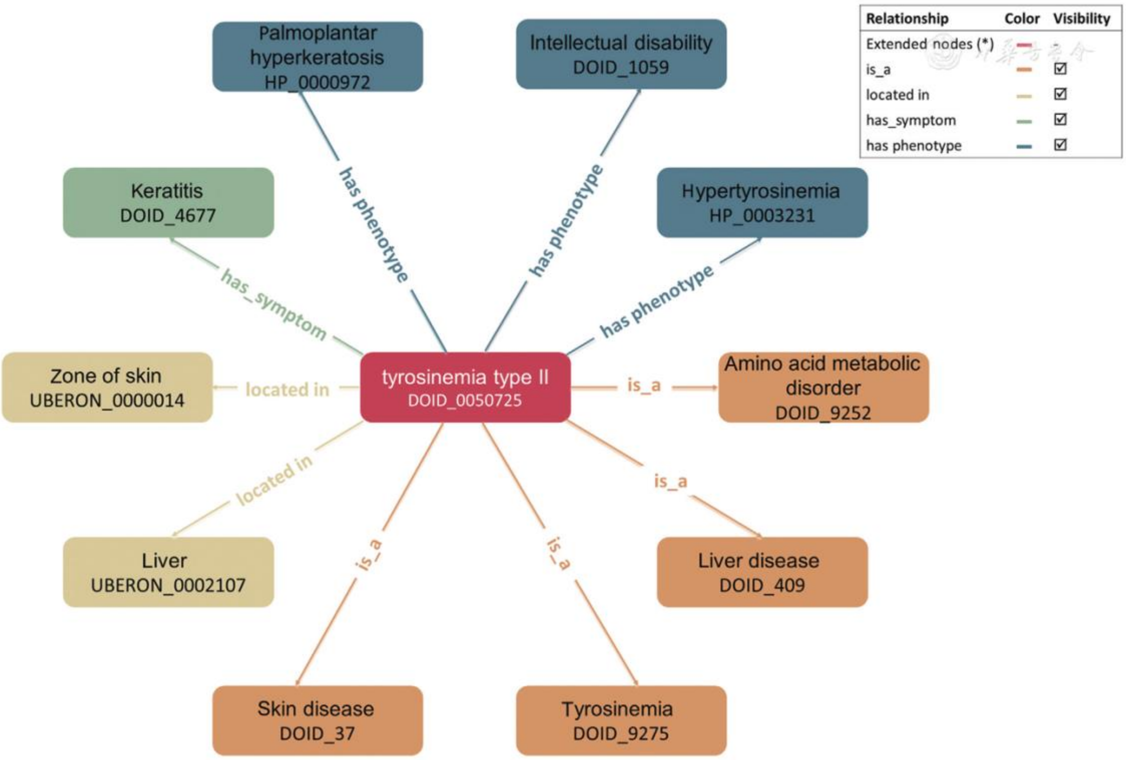

Figure 1: The representation of **tyrosinemia type II** in Disease Ontology (DO). Source: [13]

The Human Disease Ontology, in its last update on April 28th, 2022, comprises 17,840 classes and 45 properties [15] and is widely applied for several purposes in Academic and Industry contexts. In addition, it has been used by more than 50 other biomedical ontologies and there is a numerous list of software tools and other web resources that: (1) support the use of DO data, (2) have integrated or were built using DO data, or (3) provide data linkages to the DO website [16].

On the Benefits of Semantics, Ontologies and Conceptual Modeling in the Data Science Lifecycle

Managing data cannot be accomplished solely by humans with their limited cognitive capabilities [9]. Also, available data keeps growing and is becoming more important as a resource for decision-making. Thus, it is crucial to understand the domain which the data represents, to make a more precise usage of it.

Works [8,9] show that pairing conceptual modeling/ontologies artifacts with data science/machine learning techniques can not only enhance Data Science projects results but also support the development and evaluation of conceptual modelling approaches. However, this work will focus on the first mentioned kind of benefit, when semantical commitment helps on Machine Learning techniques.

In particular, [8] defend the benefits of using foundational and domain ontologies appears in each cycle of the Data Science Life Cycle, including Problem Understanding, Data pre- and post-processing, and Data Mining for different techniques (Classification and Clustering, for example). Such benefits are summarized on Table 1.

On the Data Pre-processing step, [8] defend ontologies could help on both on semantic interoperability and ontological commitment made explicit. These benefits refer to data integration which can be made not considering the ontological commitment of the sources providing the data and, therefore, joining data features which refers to different entities of the real world, leading to misinterpretations and false results on the DS project.

When clustering data, relying on foundational ontologies may lead to cluster results that better reflect real-world categorization. Moreover, calculating data similarity committed on ontological foundational can lead to similarities between data way more befitting to the domain where the treated data lays on.

Table 1

Multiple Benefits of Foundational Ontologies and Domain Ontologies on Data Science. Source: Adapted from [8]

DS Lifecycle Step	Benefit
Problem understanding	Semantic transparency Complexity management mechanisms for complex domains Data models are more uniform
Data pre-processing	Semantic interoperability Ontological commitments made explicit
Clustering	Higher probability of clusters that reflect genuine real-world categorizations Similarity calculation grounded on ontological foundations Easier to identify similarities that are not accidental Preventing unwarranted associations evaluation
Data post-processing	Improved understanding of the patterns discovered Systematic guidance in the validation of the patterns discovered grounded on ontological meta-properties

Traditional data mining methods and techniques treat data as merely “sums of attribute values”, and such approach can lead to biases and bad understanding of the patterns discovered [8]. Indeed, clustering techniques mostly relies on calculating similarities – a data pre-processing step – which does not consider semantical attributes and are basically mathematical operations to calculate Euclidian distance and other kind of metrics. However, there have been for the past few years many proposals of considering ontologies on the calculation of object similarities, such as [16,17]. Also, on the biomedical field, especially for Gene Ontology (GO) [19,20], there are several similarity metrics considering many different ontologies, such as Wang [11] and [21,22,23]. However, the metric proposed in [11] can also be extended for comparison between DO terms.

In this research scenario, ontologies will show up as a tool on data preparation step and, therefore, may enhance analysis results. The ontology terms (diseases) and taxonomic relations (*is-a*) will be considered when computing similarities between group of comorbidities, since each comorbidity is linked to a disease in the Disease Ontology. Similarities should be calculated following a groupwise approach, to enable a comparison between two groups of comorbidities. Pairwise similarities may be trivially computed by a simple application of a distance metric, either one of the four last mentioned metrics or any of the metrics available in HESML (Half-Edge Semantic Measures Library) [24].

Semantic aware groupwise metrics, however, are not that simple. According to [24], “A groupwise semantic similarity measure is used to compute the degree of similarity between two sets of concepts defined into an ontology. This type of measure is commonly used to compare sets of GO terms in genomics, although they could also be used to compare sets of WordNet synsets evoked by two words”. Section 6 details the approach used to calculate DO terms groupwise similarities.

Associating comorbidities to diseases in the Disease Ontology

We analyzed the dataset from SIVEP-Gripe (Sistema de Informação de Vigilância Epidemiológica da Gripe or Flu Epidemiological Vigilance Information System), a nationwide surveillance database used to monitor severe acute respiratory infections in Brazil. Each instance of such dataset represents a hospital admission due to COVID-19, characterized by several features regarding case evolution (Death or Recovery), patient previous COVID-19 vaccine administrations and others.

However, this dataset contains a lot of imprecise and missing data, specially on data referring to the patient comorbidities, which this work aims to tackle. Hence, data selection followed a semantic aware methodology, described as follows.

Data was selected by filtering the first three thousand hospitalization of 2021 in the State of Rio de Janeiro. However, since this work will rely mostly on analyzing each patient set of comorbidities, the filtering also considered instances of data with noisy, inaccurate and missing information regarding this feature. Also, since this study focuses on the pairing of ontologies to the Data Science Lifecycle, rather than discovering new patterns, we did not prioritize analyzing larger datasets.

Patient comorbidities which appeared in the dataset were then mapped to the ontology. Each comorbidity on the dataset was associated with a DO disease. This step was performed manually, by searching for DO classes whose names were syntactically similar to the comorbidity name appearing in the dataset. Some of these associations can be seen on Table 2.

For example, if a hospitalization entry on SIVEP-Gripe dataset has, for instance, the word “DPOC” (short for Doença Pulmonar Obstrutiva Crônica in Portuguese) in MORB_DESC column, we consider that the patient has “Chronic Obstructive Pulmonary Disease”, which has the ID DOID:3083 in the DO.

Table 2

Disease Matching between SIVEP-Gripe names with DO terms. Source: Authors

Name on SIVEP-Gripe Database	DO Match
ALCOOLISMO	alcohol use disorder
ALZHEIMER	Alzheimer’s disease
AMILOIDOSE	amyloidosis
ANEMIA	deficiency anemia

Calculating (dis)similarities between DO terms

There are several ways to calculate pairwise similarities between classes in an ontology. In this work, the proposed metric on [11] is applied to measure semantic similarity among DO terms. For computing such metric, Wang defines a term A in DO as $DAG = (A, T_A, E_A)$, where T_A is the set of all ancestors in DO graph and E_A is the set of edges connecting DO terms to A. The S-Value of DO term t related to term A is defined as the contribution of t to the semantics of A, such that, for any t in DAG_A, its S-value related to term A is defined on equation 1.

$$S_A(t) = \begin{cases} 1, if\ t = A \\ \max\{w_e \times S_A(t') | t' \in children\ of\ t\}, otherwise \end{cases} \quad (1)$$

However, w_e is a value representing the semantic contribution factor for edge $e \in E_A$ linking term t with its child t', thus for every e, a corresponding weight w_e may be predefined. Wang similarity measure for DO terms only considers is-a relationships, and the corresponding weight w_e is preset to be 0.7.

Also, for a given term A, the total semantic contribution of A, $SV(A)$ in DAG_A is given on equation 2.

$$Sim_{Wang}(A, B) = \frac{\sum_{t \in T_A \cup T_B} S_A(t) + S_B(t)}{SV(A) + SV(B)} \quad (2)$$

For computing such metrics, the R software package DOSE [24] was used, which is part of the open-source software for bioinformatics Bioconductor. Figure 3 shows a heatmap representing pairwise similarities among some DO terms. For instance, let A a vector of DO ID terms as follows on equation 3.

$$A = (8498,409,2841,850,2914,7148,8857) \quad (3)$$

The seven terms on vector A represent, respectively, the diseases in the following set: (**osteoarthritis**, **liver disease**, **asthma**, **lung disease**, **immune system disease**, **rheumatoid arthritis**, **lupus erythematosus**). We define a matrix S, such that the value on position $S_{Ai,Aj}$ represents the similarity $Sim_{Wang}(A_i, A_j)$, with the graphical representation on Figure 2.

Also, Figure 2 displays where in the ontology the terms on vector A are placed, with respect to their relationships and hierarchies between other terms. Moreover, the relationship *has-subclass* is equivalent to is-a in the way that, if A *is-a* B, then B has-subclass A.

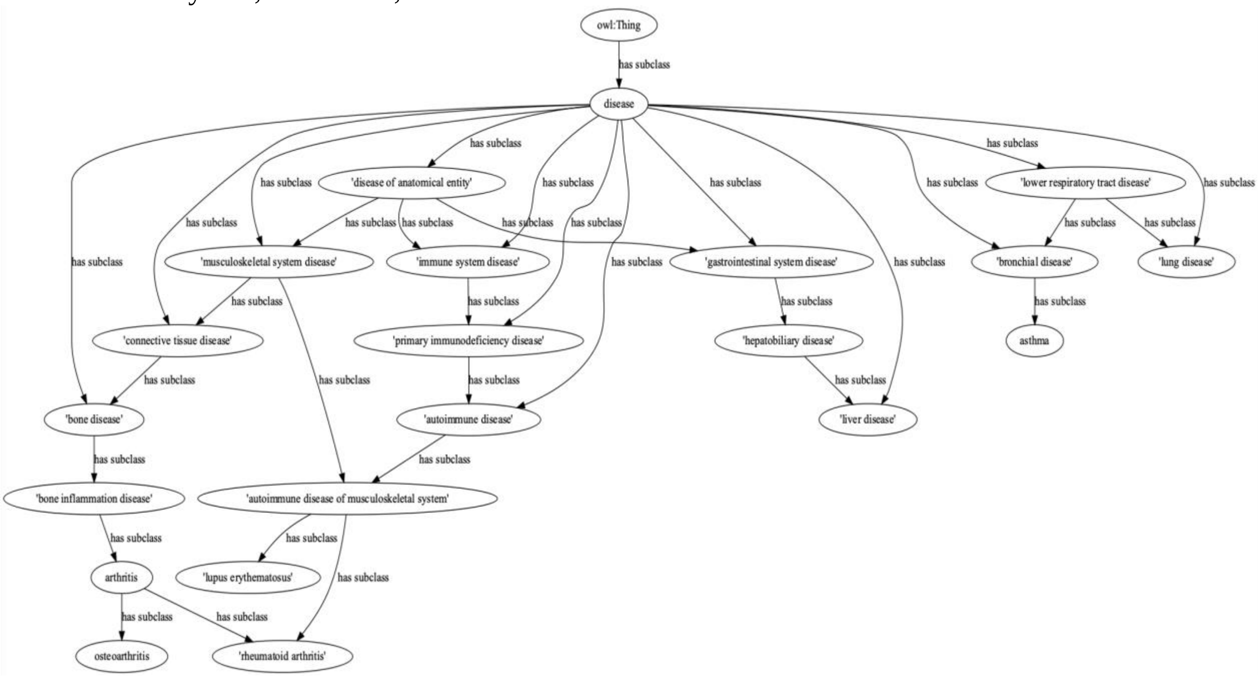
owl:Thing
has subclass
disease
'disease of anatomical entity'
'lower respiratory tract disease'
'musculoskeletal system disease'
'immune system disease'
'gastrointestinal system disease'
'bronchial disease'
'lung disease'
'connective tissue disease'
'primary immunodeficiency disease'
'hepatobiliary disease'
asthma
'bone disease'
'autoimmune disease'
'liver disease'
'bone inflammation disease'
'autoimmune disease of musculoskeletal system'
arthritis
'lupus erythematosus'
osteoarthritis
'rheumatoid arthritis'

Figure 2: Graph representing path-to-root concepts of six diseases in DO. Source: Author

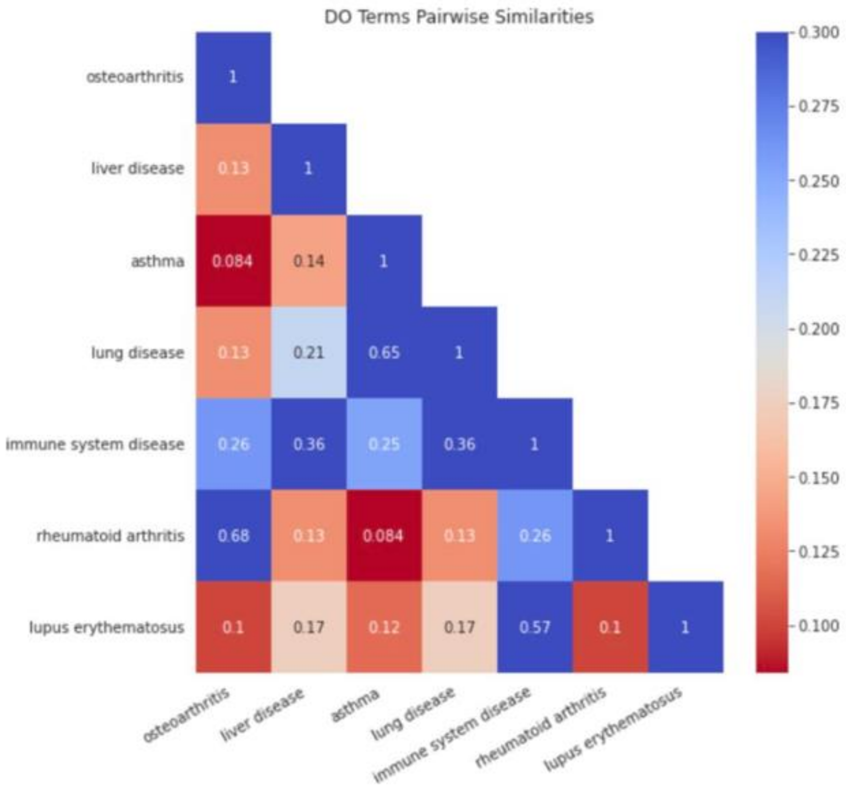

Figure 3: Pairwise similarities between DO terms. Source: Author

As can be seen on Figure 3, **rheumatoid arthritis** has a high similarity with **osteoarthritis** because both diseases have a relationship *is-a* with **arthritis**. Also, since **rheumatoid arthritis** *is-a* **autoimmune disease of musculoskeletal system** together with **lupus erythematosus**, such DO terms have higher pairwise similarity when comparing **lupus erythematosus** with **osteoarthritis**.

Calculating groupwise (dis)similarities

Each row in the hospitalization's dataset represents a hospital entry, which refers to a unique patient. As aforementioned, each entry contains data about the diseases a patient has. Hence, each instance on the dataset is characterized as a single group of DO terms. With the previous definitions, only pairwise similarity metrics between classes in the ontology can be computed. Then, for calculating similarities between set of diseases i.e., groupwise similarities, other approaches were required.

For instance, consider an ordered set C containing n terms from DO, and an example instantiation of C in which $n = 4$, as shown on equation 4.

$$\mathrm{C} = \{\boldsymbol{lupus\ erythemathosus}, \boldsymbol{rheumatoid\ arthritis}, \boldsymbol{liver\ disease}, \boldsymbol{asthma}\} \tag{4}$$

Also, let $D \subseteq C$ the subset representing the diseases a patient suffers, and an example instantiation of D, as on equation 5.

$$\mathrm{D} = \{\boldsymbol{rheumatoid\ arthritis}, \boldsymbol{asthma}\} \tag{5}$$

Any subset of diseases in C may be a represented as a document vector v, i.e., a n - dimensional binary vector, in which each coordinate represents if the concept of C is in D. Thus, in this case, $v_D^T = (0\ 1\ 0\ 1)$. This representation is useful and broadly used in Natural Language Processing models and some machine learning techniques that rely on similarity measures between instances of data.

Cosine (dis)similarity

Considering x, y vectors in the n-dimensional space, cosine similarity between these vectors is represented as on equation 6.

$$GSim_{cos}(x, y) = \frac{x \cdot y}{\|x\|\,\|y\|} \tag{6}$$

The operation $x \cdot y$ represents the usual $\mathbb{R}^n$ inner product and $\|x\|$ represents the Euclidian magnitude of a vector $x \in \mathbb{R}^n$.

Also, this similarity metric follows the property shown on equation 7.

$$\forall(x, y) \in \mathbb{R}^n \times \mathbb{R}^n : 0 \leq GSim_{cos}(x, y) \leq 1 \tag{7}$$

Therefore, cosine dissimilarity is defined on equation 8.

$$GDSim_{cos}(x, y) = 1 - GSim_{cos}(x, y) \tag{8}$$

Even though this metric represents, at some way, groupwise disease similarities, ontologies are not considered as semantical enrichment artifacts. Therefore, according to [8], data mining techniques relying in these metrics may lead to less genuine understanding of patterns discovered, due to the lack of semantics.

Hence, section 6.2 provides an ontologically well-founded (dis)similarity metric that may be considered as an extension of the original cosine similarity and is inspired on [12] work, which applies the metric on the domain of radiology.

Semantically aware cosine (dis)similarity

For introducing semantic similarity between document vectors, [12] first define (in their words, in a loosely way) the similarity between two concepts $C1, C2$ in an ontology as shown on equation 9.

$$Sim(C1, C2) = \frac{1}{d} \tag{9}$$

Where d is the number of nodes in the shortest path between concept nodes (inclusive of) $C1$ and $C2$. However, the authors clarify that other similarity measures can be used, as long as it preserves the basic property that increasing distance within the ontology is concomitant with a decrease in semantic similarity. Hence, the similarity measure defined by [11] for DO terms will be used, as displayed on equation 10.

$$Sim(C1, C2) = Sim_{Wang}(C1, C2) \tag{10}$$

Henceforward, each term of the domain ontology brought up by the dataset, together with all the other concepts in their paths-to-root (a.k.a. seed concepts), will represent each coordinate of the document vectors which will be further analyzed. However, Wang pairwise similarity measure already represents the weight of seed concepts in its formula. Hence, in this work, only the Disease Ontology terms presented on the explored dataset will be considered, and such group of diseases will be represented as a set C, called context set.

Finally, with the definitions above, the DO terms groupwise similarities, $GSim_{Wang}(A, B)$, with respect to a context can now be computed. Hence, let $C = \{C_1, C_2, \dots, C_n\}$ be a set of diseases representing the context set and let two group of disease terms, namely, A and B, which by definition, $A, B \subseteq C$. Then, groupwise similarity considering semantic is represented on equation 11.

$$GSim_{Wang}(A, B) = \frac{\sum_{c \in C \cap (A \cup B)} \max_{a \in A} Sim_{Wang}(a, c) \cdot \max_{b \in B} Sim_{Wang}(b, c)}{\sqrt{\sum_{c \in C \cap A} \left(\max_{a \in A} Sim_{Wang}(a, c)\right)^2} \cdot \sqrt{\sum_{c \in C \cap B} \left(\max_{b \in B} Sim_{Wang}(b, c)\right)^2}} \tag{11}$$

Also, this similarity metric ranges from 0 to 1, therefore, dissimilarity is derived as on equation 12.

$$GDSim_{Wang}(A, B) = 1 - GSim_{Wang}(A, B) \tag{12}$$

For instance, let's calculate the similarity between group of DO terms for context C, as in Table 3.

Table 3

Values for computing DO terms groupwise similarities. Source: Authors

	asthma	liver disease	lung disea se	immune system disease	rheumatoid arthritis
$A = \{asthma, liver\ disease\}$	1	1	0.65	0.36	0.13
$B = \{rheumatoid\ arthritis\}$	0.084	0.13	0.13	0.26	1

Similarity between groups of diseases A and B is then calculated as in equation 13 and 14.

$$GSim_{Wang}(A, B) = \frac{1 \cdot 0.084 + 1 \cdot 0.13 + 0.65 \cdot 0.13 + 0.36 \cdot 0.26 + 0.13 \cdot 1}{\sqrt{1^2 + 1^2 + 0.65^2 + 0.36^2 + 0.13^2} \cdot \sqrt{0.084^2 + 0.13^2 + 0.13^2 + 0.26^2 + 1}} \tag{13}$$

$$GSim_{Wang}(A, B) = 0.1824 \tag{14}$$

Therefore, in this work, both semantically aware and unaware groupwise dissimilarities were calculated. Figure 4 shows how smooth dissimilarity is when enriching data with semantics, while semantically unaware measures lead to false dissimilarities between data objects, which potentially may impact on further cluster analysis.

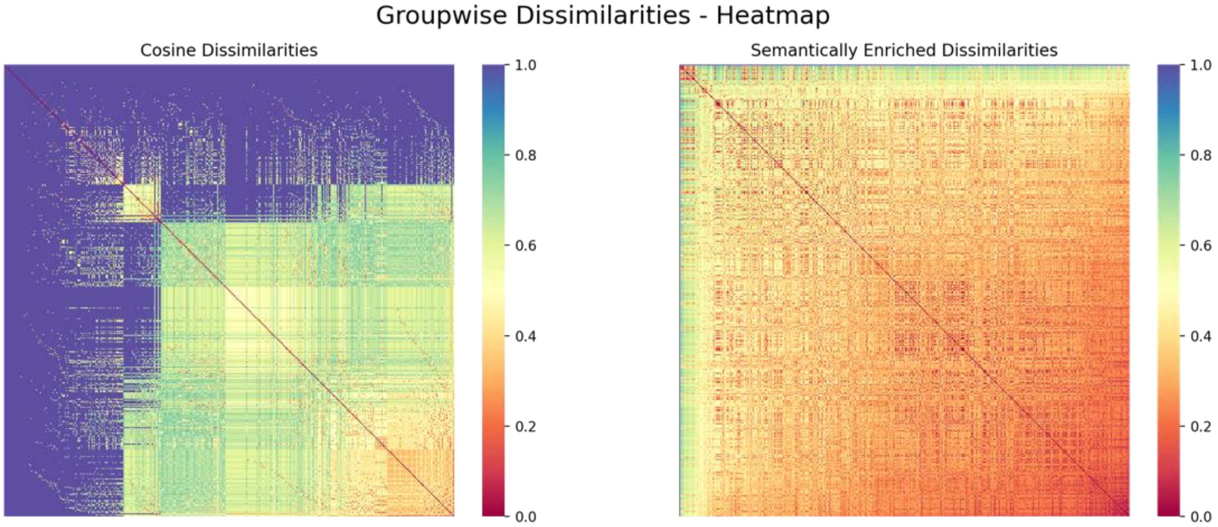

Figure 4: Heatmaps of groupwise dissimilarities using semantically unaware (left) and semantically aware (right) metrics. Source: Author

Hospitalizations cluster analysis

On [26], clustering is defined as the process of grouping a set of data objects into multiple groups or clusters so that the objects within a cluster have high similarity but are very dissimilar to objects in other clusters. Euclidian and Manhattan distance are often used as dissimilarity measure on clustering techniques. However, in this study, clustering analysis will rely on both cosine similarity and cosine similarity based on the prior mentioned Wang [11] measure.

This work focuses on the use of the K-medoids clustering technique [27], which is a Partitioning-Based clustering algorithm that is scalable and compatible to cluster objects upon precomputed dissimilarity metrics, which is the case of the data in this study.

Also, for choosing clustering algorithm parameters (such as the number of clusters) this work relies on the Silhouette Coefficient [28] as a metric which we want to maximize. Such metric, based on the intra-cluster and extra-cluster distances, provides information regarding the quality of the clusters.

For making use of such algorithms, Scikit-learn [29] implementation of K-Medoids and Silhouette Score on Python programming language [30] was used.

The average silhouette coefficient was then calculated for each instance of K-medoids application, on both semantically aware and unaware dissimilarity data and for different numbers k of clusters, ranging from 2 to 15. As seen on Figure 5, the optimal number of clusters $k = k^*$, which maximizes the average silhouette score was, on the semantic aware case was $k^*_{nosem} = 4$ and on the semantic unaware case was $k^*_{sem} = 5$, where each clustering obtained, respectively, scores of 0.277108 and 0.12143.

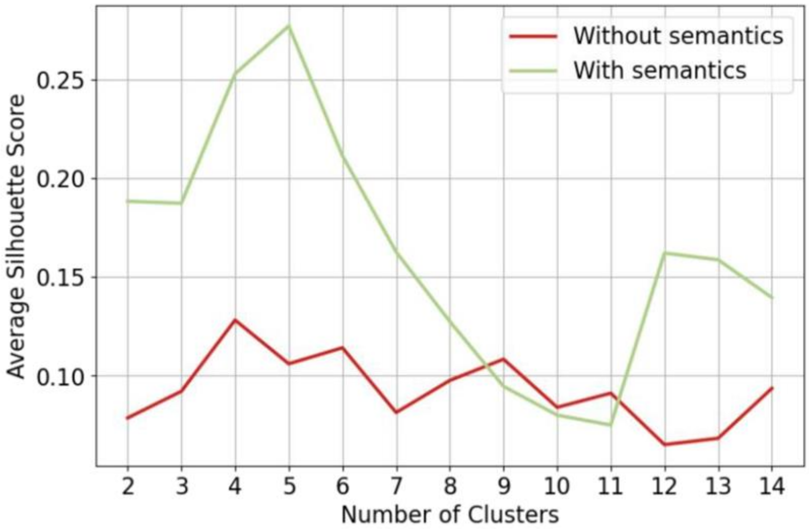

Figure 5: Comparing average silhouette score for different number of clusters. Source: Author

The obtained results regarding the quality of the clustering on both treated data are in fact encouraging. The bar plot displayed on figure 6 shows that not only the average silhouette is clearly higher, but the metric evaluated individually for each data point is clearly higher on the overall. Also, cluster 0 of the cosine dissimilarity clustering has mainly negative silhouette scores. Moreover, when clustering the semantically aware data, the average silhouette score was higher than 83% (386 out of 465) of the observations on the semantically unaware scenario.

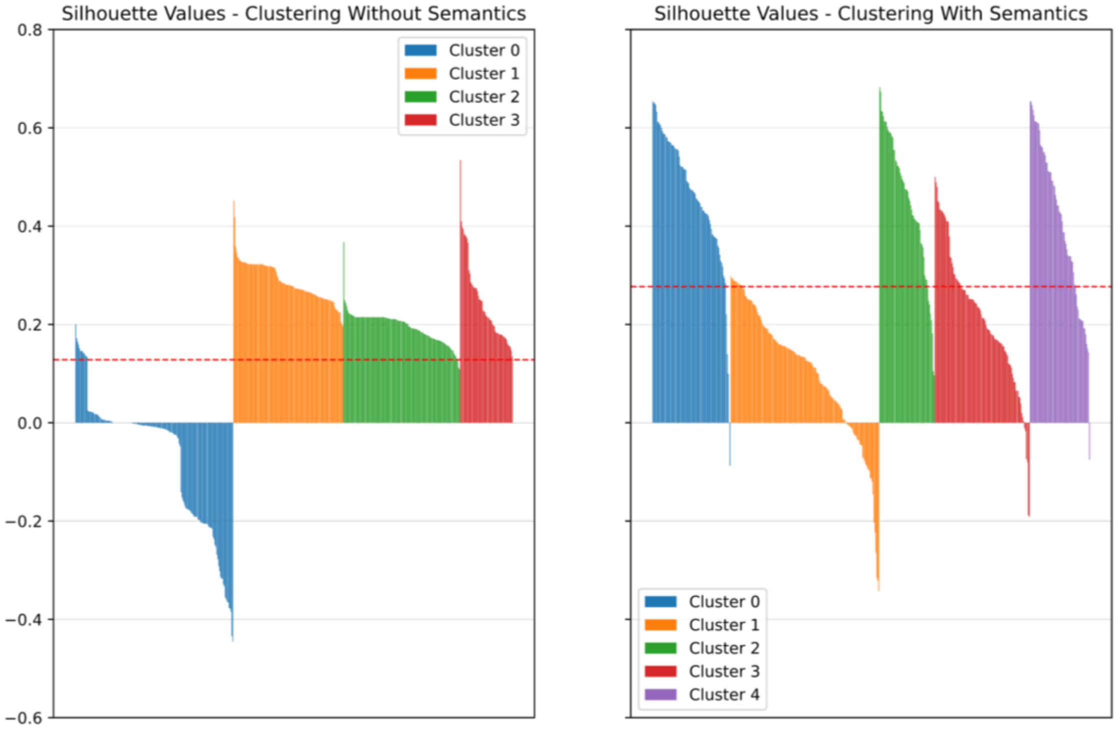

Figure 6: Bar plot with values of silhouette score for each data point. Source: Author

Dimensionality reduction

In this work, Multidimensional Scaling MDS [31], a dimensionality reduction technique was useful to transform groups of DO terms dissimilarities into points in the cartesian plane. Therefore, both charts displayed on Figure 7 were possible. Moreover, information regarding both the clustering results and the obtained silhouettes scores were represented, respectively, by introducing different colors and radius sizes for each point. Also, for every cluster, the medoid point was represented with a black cross, where it emerges a box displaying all DO terms presented by the highest 4 silhouette scored group of diseases of each cluster.

The results shown in Figure 7 are crucial to make explicit how clustering results are improved when adding semantics to data. While on the left chart clusters are overlapping (one more evidence to explain the low silhouette scores obtained), the one in the right, shows how the clusters were better separated, thus way closer to the main objective of this technique, which is to maximize intra-cluster similarities and maximize inter-cluster similarities. Lastly, our results evidenced the benefit of “Higher probability of clustering results that reflect real-world categorizations”, exactly as mentioned by [8]. When comparing both scenarios, the semantically unaware clusters grouped diseases which are, by common sense, dissimilar to each other; on the other hand, semantically aware clusters reflected real-world categorizations, i.e., diseases within the same cluster are clearly more similar to each other.

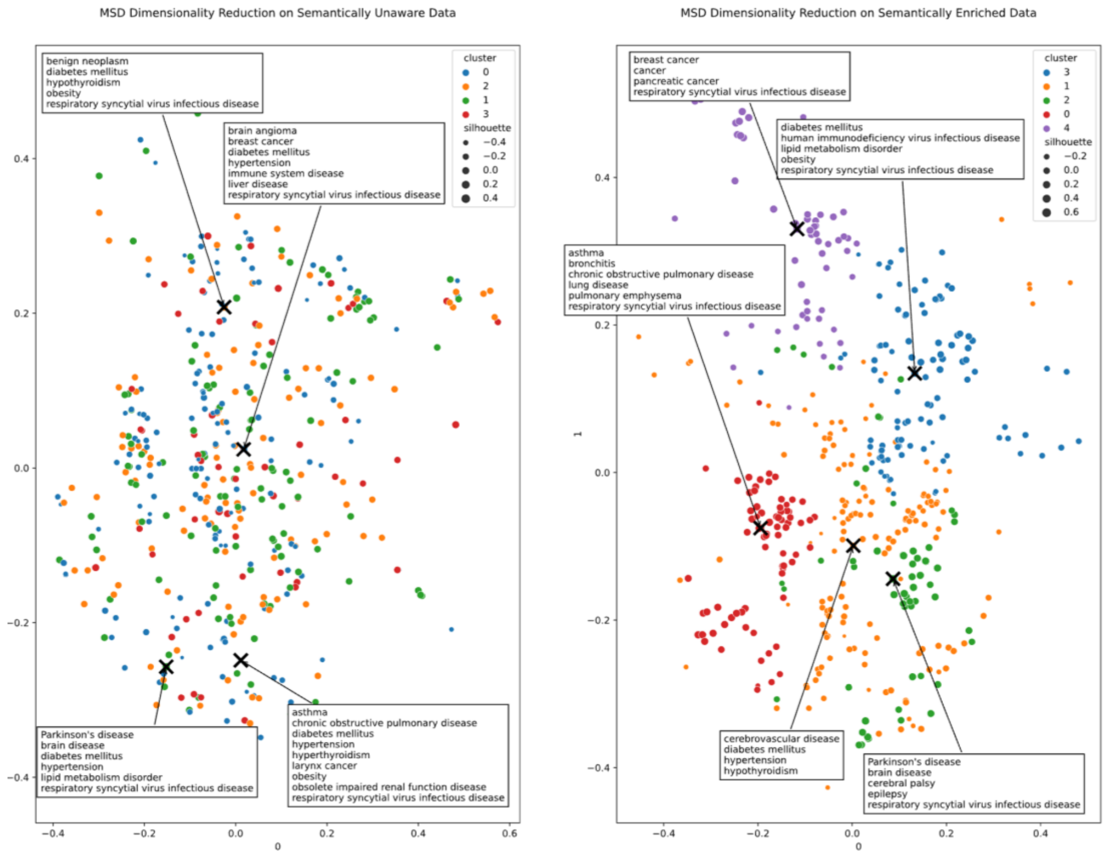

Figure 7: Graphical representation of clustering on semantically unaware (right) data and semantically aware (left) data. Source: Author

Conclusions

This work proposes a semantic awareness application of the Data Science Lifecycle on the COVID-19 domain and shows the benefits of considering ontologies and other semantic structures as tools for enhancing ML techniques.

Even though there are ontology terms groupwise metrics in the literature, they are not as present and accessible as the ones measuring pairwise similarities. So, in the context where groupwise distance between sets of objects are required, an adaption of the [12] proposal for calculating groupwise similarities was made so [11] was computed.

The benefits of enriching a disease dissimilarity metric with context information were evident. Firstly, when calculating groupwise similarities, Cosine Similarity, as shown on Figure 4 led to context inaccurate (dis)similarities between data objects and was pointed that could lead to bad results later, during cluster analysis. Figures 5 and 6 shows how the overall silhouettes score (i.e.) on an overall are considerably higher when enriching data with semantics.

Figure 7 aims at giving the reader a visualization of the most important results in a nutshell. It displays the overlapping clusters, which is a result of the semantically unaware similarity calculation. Also, such visual results agree with silhouette values found on the Data Pre-Processing step. On the other scatter plot, where data is semantically enriched, intra-clusters distances were minimized, and inter-cluster similarities were maximized. Finally, an analysis of the group of diseases grouping is visually represented on Figure 7.

Future Works

In this work, text treatment step on this work did not rely on modern Natural Language Processing (NLP) techniques. Leading, to manual tasks such as linking terms in the DO with data regarding comorbidities of the hospitalized patients. Therefore, as future work, such step can be automatized so more information can be considered.

Also, enriching similarity pairwise metric by not only considering *is-a* relationships, but many others an ontology can provide. Also, such as the work of [8], the use of foundational ontologies and their associated metaproperties can also be applied for a project using the Data Science Lifecycle. Hence, OntoCovid and OntoTB [32,33] are well-founded ontologies that may help when applying ML techniques.

An analysis on how clustering results are associated with mortality and to the use of mechanical ventilation will be made. Therefore, semantically enrichment of data could serve as an tool for better results on data-driven decision making.

Acknowledgements

This study was conducted in the context of the project "Effectiveness of COVID-19 Vaccination in Brazil Using Mobile Data" (EFFECT-BR), which is one out of ten, among 440 others worldwide, selected by the Grand Challenges ICODA COVID-19 Data Science, funded by the Bill & Melinda Gates foundation. Also, the Center for Healthcare Operations and Intelligence (NOIS[2]) which is part of PUC-Rio Industrial Engineering Department, together with Tecgraf institute, Fundação Oswaldo Cruz (FIOCRUZ) and Instituto D'Or (IDOR) gave the support needed so this study could be made. This work was was partially funded by projects FAPERJ E-26/010.002657/2019 and CNPq 422810/2021-5.

References

[1] World Health Organization, “WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020,” *World Health Organization*, Mar. 11, 2020. https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020 (accessed Apr. 22, 2022).

[2] T. K. Tsang, P. Wu, Y. Lin, E. H. Y. Lau, G. M. Leung, and B. J. Cowling, “Effect of changing case definitions for COVID-19 on the epidemic curve and transmission parameters in mainland China: a modelling study,” *The Lancet Public Health*, vol. 5, no. 5, Apr. 2020, doi: 10.1016/s2468-2667(20)30089-x.

[3] G. Guizzardi, “Ontology, Ontologies and the 'I' of FAIR,” *Data Intelligence*, vol. 2, no. 1–2, pp. 181–191, Jan. 2020, doi: 10.1162/dint_a_00040.

[2] http://www.nois.ind.puc-rio.br

[4] S. Babcock, J. Beverley, L. G. Cowell, and B. Smith, “The Infectious Disease Ontology in the age of COVID-19,” *Journal of Biomedical Semantics*, vol. 12, no. 1, Jul. 2021, doi: 10.1186/s13326-021-00245-1.

[5] L. Wan *et al.*, “Development of the International Classification of Diseases Ontology (ICDO) and its application for COVID19 diagnostic data analysis,” *BMC Bioinformatics*, vol. 22, Oct. 2021, doi: 10.1186/s12859021044022.

[6] H. Wu, Y. Zhong, Y. Tian, S. Jiang, and L. Luo, “Automatic diagnosis of COVID-19 infection based on ontology reasoning,” *BMC Medical Informatics and Decision Making*, vol. 21, no. S9, Nov. 2021, doi: 10.1186/s12911-021-01629-0.

[7] A. Sargsyan *et al.*, “The COVID-19 Ontology,” *Bioinformatics*, vol. 36, no. 24, pp. 5703–5705, Dec. 2020, doi: 10.1093/bioinformatics/btaa1057.

[8] G. Amaral, F. Baião, and G. Guizzardi, “Foundational ontologies, ontology-driven conceptual modeling, and their multiple benefits to data mining,” *WIREs Data Mining and Knowledge Discovery*, vol. 11, no. 4, Jul. 2021, doi: 10.1002/widm.1408.

[9] W. Maas and V. C. Storey, “Pairing conceptual modeling with machine learning,” *Data & Knowledge Engineering*, vol. 134, no. C, p. 101909, Jun. 2021, doi: 10.1016/j.datak.2021.101909.

[10] L. M. Schriml *et al.*, “Human Disease Ontology 2018 update: classification, content and workflow expansion,” *Nucleic Acids Research*, vol. 47, no. D1, pp. D955–D962, Jan. 2019, doi: 10.1093/nar/gky1032.

[11] J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, and C.-F. Chen, “A new method to measure the semantic similarity of GO terms,” *Bioinformatics*, vol. 23, no. 10, pp. 1274–1281, May 2007, doi: 10.1093/bioinformatics/btm087.

[12] T. Mabotuwana, M. C. Lee, and E. V. Cohen-Solal, “An ontology-based similarity measure for biomedical data – Application to radiology reports,” *Journal of Biomedical Informatics*, vol. 46, no. 5, pp. 857–868, Oct. 2013, doi: 10.1016/j.jbi.2013.06.013.

[13] J. Han, M. Kamber, and J. Pei, *Data mining : concepts and techniques*. Burlington, Ma: Elsevier, 2012.

[14] H. Pan *et al.*, “Biomedical ontologies and their development, management, and applications in and beyond China,” *Journal of Bio-X Research*, vol. 02, Art. no. 04, 2019, doi: 10.1097/JBR.0000000000000051.

[15] NCBO BioPortal, “Human Disease Ontology | NCBO BioPortal,” *bioportal.bioontology.org*. https://bioportal.bioontology.org/ontologies/DOID (accessed May 27, 2022).

[16] Disease Ontology, “Disease Ontology - Institute for Genome Sciences - Use Cases,” *disease-ontology.org*, 2022. https://disease-ontology.org/community/use-cases (accessed Jun. 08, 2022).

[17] K. Gibert, A. Valls, and M. Batet, “Introducing semantic variables in mixed distance measures: Impact on hierarchical clustering,” *Knowledge and Information Systems*, vol. 40, no. 3, pp. 559–593, Jun. 2013, doi: 10.1007/s10115-013-0663-5.

[18] W. Lee, N. Shah, K. Sundlass, and M. Musen, “Comparison of Ontology-based Semantic-Similarity Measures,” *AMIA ... Annual Symposium proceedings. AMIA Symposium*, vol. 2008, pp. 384–388, Nov. 2008, [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2655943/

[19] M. Ashburner *et al.*, “Gene Ontology: tool for the unification of biology,” *Nature Genetics*, vol. 25, no. 1, pp. 25–29, May 2000, doi: 10.1038/75556.

[20] Gene Ontology Consortium, “The Gene Ontology resource: enriching a GOld mine,” *Nucleic Acids Research*, vol. 49, no. D1, pp. D325–D334, Jan. 2021, doi: 10.1093/nar/gkaa1113.

[21] J. J. Jiang and D. W. Conrath, “Semantic similarity based on corpus statistics and lexical taxonomy,” in *The Association for Computational Linguistics and Chinese Language Processing (ACLCLP)*, Aug. 1997, vol. 10, pp. 19–33. [Online]. Available: https://aclanthology.org/O97-1002

[22] P. Resnik, “Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language,” *Journal of Artificial Intelligence Research*, vol. 11, pp. 95–130, 1999, doi: 10.1613/jair.514.

[23] D. Lin, “An information-theoretic definition of similarity,” in *IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology (CIBCB)*, 1998, pp. 296–304.

[24] J. J. Lastra-Díaz, A. García-Serrano, M. Batet, M. Fernández, and F. Chirigati, “HESML: A scalable ontology-based semantic similarity measures library with a set of reproducible

experiments and a replication dataset," *Information Systems*, vol. 66, pp. 97–118, Jun. 2017, doi: 10.1016/j.is.2017.02.002.

[25] G. Yu, L.-G. Wang, G.-R. Yan, and Q.-Y. He, "DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis," *Bioinformatics*, vol. 31, no. 4, pp. 608–609, Feb. 2015, doi: 10.1093/bioinformatics/btu684.

[26] J. Han, M. Kamber, and J. Pei, *Data mining : concepts and techniques*. Burlington, Ma: Elsevier, 2012.

[27] H.-S. Park and C.-H. Jun, "A simple and fast algorithm for K-medoids clustering," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3336–3341, Mar. 2009, doi: 10.1016/j.eswa.2008.01.039.

[28] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, Nov. 1987, doi: 10.1016/0377-0427(87)90125-7.

[29] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, Art. no. 85, 2011, [Online]. Available: http://jmlr.org/papers/v12/pedregosa11a.html

[30] G. van Rossum and F. L. Drake, *Python 3 : reference manual*. United States: Sohobooks, 2009.

[31] J. B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, vol. 29, no. 1, pp. 1–27, Mar. 1964, doi: 10.1007/bf02289565.

[32] L. Maddalena and F. Baião, "OntoCovid: Applying SABiO to conceptual modeling well grounded in the COVID-19 domain," in *CEUR Workshop Proceedings*, 2021, vol. 3050.

[33] T. Guarnier *et al.*, "Um Modelo Conceitual Baseado em Ontologia para Doenças Infecciosas com Ênfase em Tuberculose," 2020. [Online]. Available: http://ceur-ws.org/Vol-2728/short5.pdf