NATURE: A TOOL RESULTING FROM THE UNION OF ARTIFICIAL INTELLIGENCE AND NATURAL LANGUAGE PROCESSING FOR SEARCHING RESEARCH PROJECTS IN COLOMBIA

Felipe Cujar-Rosero, David Santiago Pinchao Ortiz,
Silvio Ricardo Timarán Pereira and Jimmy Mateo Guerrero Restrepo

Systems Department, University of Nariño, Pasto, Colombia

ABSTRACT

This paper presents the final results of the research project that aimed for the construction of a tool which is aided by Artificial Intelligence through an Ontology with a model trained with Machine Learning, and is aided by Natural Language Processing to support the semantic search of research projects of the Research System of the University of Nariño. For the construction of NATURE, as this tool is called, a methodology was used that includes the following stages: appropriation of knowledge, installation and configuration of tools, libraries and technologies, collection, extraction and preparation of research projects, design and development of the tool. The main results of the work were three: a) the complete construction of the Ontology with classes, object properties (predicates), data properties (attributes) and individuals (instances) in Protegé, SPARQL queries with Apache Jena Fuseki and the respective coding with Owlready2 using Jupyter Notebook with Python within the virtual environment of anaconda; b) the successful training of the model for which Machine Learning algorithms were used and specifically Natural Language Processing algorithms such as: SpaCy, NLTK, Word2vec and Doc2vec, this was also performed in Jupyter Notebook with Python within the virtual environment of anaconda and with Elasticsearch; and c) the creation of NATURE by managing and unifying the queries for the Ontology and for the Machine Learning model. The tests showed that NATURE was successful in all the searches that were performed as its results were satisfactory.

KEYWORDS

Artificial Intelligence, Natural Language Processing, Ontology, Machine Learning, Search Engine, Semantic Web.

1. INTRODUCTION

The Internet was conceived by Tim Berners-Lee as a project to manage and share knowledge and information among a select group of scientists. With the pass of the time and with the advances in the development of hardware that made possible the communication around the world, the necessary applications were developed to meet the needs of users. The large volume of content available online makes searching and processing difficult, the need to devise new ways to optimize the treatment given to such content has been vital; for the information available on the Web to be interpreted by computers without the need for human intervention, the Semantic Web is required. It is said that in Internet computers are not only capable of presenting the information contained in web pages, else they should also "understand" such information [1].

DOI: 10.5121/ijaia.2021.12401

According to Berners Lee and Hendler, on the Semantic Web, information is offered with a well-defined meaning, allowing computers and people to work cooperatively. The idea behind the Semantic Web is to have data on the Web defined and linked so these can be used more effectively for discovery, automatization, integration and reuse between different applications. The challenge of the Semantic Web is to offer the language that expresses data and rules to reason about many data and also allows the rules on any knowledge representation system to be exported to the Web, providing a significant degree of flexibility and "freshness" to traditional centralized knowledge representation systems, which become extremely overwhelming, and its growing in size is unmanageable. Different web systems can use different identifiers for the same concept; thus, a program that wants to compare or combine information between such systems has to know which terms mean the same thing; ideally the program should have a way of discovering the common meanings of whatever database it encounters. One solution to this problem is to add a new element to the Semantic Web; collections of information called Ontologies [2].

In the same way, it is known that the large amount of textual information available on the WEB with the increase in demand by users, makes necessary to have systems that allow access to that interest information in an efficient and effective way for saving time in the search and consultation. Among the existing techniques to achieve this efficiency and effectiveness, and in turn to provide access or facilitate the management of text document information are Machine Learning techniques, using them is highly convenient, this can be evidenced in a large number of applications in different areas [3].

This is because the factors that have generated the success of the Internet have also caused problems such as: information overload, heterogeneity of sources and consequent problems of interoperability. The Semantic Web helps to solve these problems by allowing users to delegate tasks to software tools. By incorporating semantics in the Web, the software is capable of processing content, reasoning with it, combining it and making logical deductions to solve problems automatically. Automatic ability is the result of the application of artificial intelligence techniques, which require the participation of intelligent agents that improve searches, adding values for reasoning and making decisions to web services that store high content [4].

According to Kappel, it is pertinent to make use of semantics, which is reflected in the responses that a user receives to their requests in search engines, since these go beyond the state in which users simply asked a question and received a set sorted by web page priority. Users want targeted answers to their questions without superfluous information. Answers should contain information from authorized sources, terms with the same meaning as those used in the question, relevant links, etc. So, the Semantic Web tries to provide a semantic structure to the significant contents of the Web, creating an environment in which software agents navigate through the pages performing complex tasks for users [1].

It is assumed that this Web has the ability to build a knowledge base on the preferences of users and that, through a combination of its ability to understand patterns and the information available on the Internet, it is able to meet exactly the information demands from users, for example: restaurant reservation, flight scheduling, medical consultations, purchase of books, etc. Thus, the user would obtain exact results on a search, without major complications because the Semantic Web provides a way to reason on the Web as it is an infrastructure based on metadata (highly structured data describing information), thus extending its capabilities. That is, it is not a magic artificial intelligence that allows web servers to understand the words of the users, it is only the construction of a skill arranged in a machine, in order to solve well-defined problems, through similar operations well defined to be carried out on existing data [4].

In the systematic review of the literature, a search engine is defined as an application and / or computer resource that allows information to be located on the servers of a certain website, resulting in a list that is consistent with the files or materials stored on the corresponding servers and responding to the needs of the user. Search engines make easy to locate the information that is scattered around the Web, but it is crucial to know the way in which the search is being carried out [5]. Syntactic search engines make use of keywords, where the search result depends on an indexing process, which is the one that will allow organizing searches with these keywords or through the use of hierarchical trees categorized by a certain topic. Despite the power shown by syntactic search engines, they are still far from being able to provide to the user adequate results for the queries made, since the number of results can be too many and therefore it will be quite tedious to find the desired result or else not getting any results, with the addition that much of the responsibility for the search can fall into the hands of the user, who would have to filter and categorize their search to get a clear and concise answer [6].

In this way, it can be observed that these problems can be solved with the use of semantic search engines which, on the other hand, facilitate the user's work, are efficient in the search since they find results based on the context, thus providing information more exact about what is sought, offering a more biased number of results, facilitating the work of filtering the results by the user. In this way that these search engines interpret user searches by making use of algorithms that symbolize comprehension or understanding, offering precise results quickly and thus recognizing the correct context for the search words or sentences. It is nothing more than a semantic search engine, one that performs the search by looking at the meaning of the group of words that are written [7].

ERCIM digital library [8], NDLTD [9], Wolfram Alpha [10] use semantics to find results based on context. The last one is capable of directly answering the questions asked by the user instead of providing a list of documents or web pages that could contain the answer, as Google does. Once the question is asked, the tool calculates different answers by selectively choosing the information from the Web to end up giving a precise answer. Swotti is another search engine that uses Semantic Web technologies to extract the opinions made by users in blogs and forums about companies or products. It is able to identify the adjectives and verbs that define what people are looking for, and therefore allows people to deduce if the comment is positive or negative. When people make a search in Swotti they get not only results, else a qualitative assessment [11]. Swoogle is a document search engine for the Semantic Web, a Google for the Semantic Web although it is not aimed at the end user yet, it has been created at the University of Maryland, it is not intended for the common user, but for the crawling of semantic web documents whose formats are OWL, RDF or DAML. Swoogle is a search engine that detects, analyzes and indexes the knowledge encoded as Semantic Web documents. Swoogle understands by Semantic Web documents those that are written with some of the languages oriented to the construction of Ontologies (RDF, OWL, DAML, N3, etc). It retrieves both documents written entirely in these languages (which for Swoogle are strict Semantic Web documents) and documents partially written with some of them. It also provides an algorithm also inspired by Google's Page Rank algorithm, which for Swooglehas been called Ontology Rank. The Ontology Rank algorithm has been adapted to the semantics and usage patterns found in the Semantic Web documents. Swoogle currently has around 1.5M Semantic Web documents indexed. This information is available through an internal link to statistical data related to their status [12]. Other works such as that of Camacho Rodríguez in her undergraduate work to obtain the degree in Telematics Engineering propose incorporating a semantic search engine in the LdShake platform for the selection of educational patterns. This work was developed at the PompeuFabra-UPF University of Barcelona, Spain in 2013. This work analyzes the efficiency of using Ontologies to considerably improve the results and at the same time gain speed in the search [13]. Amaral presents a semantic search engine for the Portuguese language where it makes use of Natural

Language Processing tools and a multilingual lexical corpus where the user's queries are evaluated, for the disambiguation of polysemic words, it uses pivots shown on the screen with the different meanings of the word where the user chooses the meaning with which he wants to make the query [14]. Aucapiña and Plaza in their thesis for obtaining the Degree in Systems Engineering propose a semantic search engine for the University of Cuenca in Cuenca, Ecuador in 2018, where they describe in detail the use of SPARQL as a query language and the various stages carried out to achieve the prototype of the semantic search engine following proven methodologies and in certain cases those are supported by automated processes [15].Umpiérrez Rodríguez in his final graduate project in Computer Engineering called "SPARQL Interpreter" at the University Of Las Palmas Of Gran Canaria, developed in 2014, where he explains how SPARQL Interpreter addresses the problem of communication between a query language and a database of specific data [16]. Baculima and Cajamarca in their graduate thesis in Systems Engineering developed a "Design and Implementation of an Ecuadorian Repository of Linked Geospatial Data" at the University of Cuenca Ecuador, in 2014, they work on the solution for generation, publication and visualization of data Geospatial Links, for which they rely on web search engines, this since the Web focuses on the publication of this type of data, allowing them to be structured in such a way that they can be interconnected between different sources. This work is supported by SPARQL and GEOSPARQL to be able to carry out queries, insert modification and elimination of data [17]. Iglesias, developed his project at the Simón Bolívar University of Barranquilla, his objective was to build an ontological search engine that allows semantic searches to be carried out online for master's and doctorate training works, where people can find this kind of work or topics that can serve as a guide for new research to emerge, thus improving searches when selecting research topics for undergraduate projects [18]. Bustos Quiroga in the thesis in the Master's Degree in Computer and Systems Engineering develops a "Prototype of a system for integrating scientific resources, designed to function in the space of linked open data to improve collaboration, efficiency and promote innovation in Colombia" in 2015 at the National University of Colombia. In this work he used the Semantic Web in linked data to improve integration in timelessness between applications and facilitate access to information through unified models and shared data formats [19]. Moreno and Sánchez in their undergraduate work to obtain the title of Systems and Computing Engineer propose a prototype of semantic search engines applied to the search for books on Systems Engineering and Computing in the Jorge RoaMartínez library of the Technological University of Pereira. This work was developed in 2012. This prototype was developed based on the existing theoretical foundations and the analysis that was carried out on the technologies involved, such as intelligent software agents, Ontologies that are implemented in languages such as RDF and XML, and other development tools [20]. Likewise, at the University of Nariño, Benavides and Guerrero developed the undergraduate work project to obtain the title of Systems Engineer, in 2013, called "UMAYUX: a knowledge management software model supported by a coupled-weakly dynamic Ontology with a database manager for the University of Nariño" whose objective was to convert the knowledge that was tacit, in the academic and administrative processes of the University of Nariño, into explicit knowledge that allows to collect, structure, store information and transform through the use of domain-specific Ontologies, in a way that each academic unit or administrative unit can build and couple to the model. The UMAYUX model was implemented through the construction of MASKANA, a knowledge management tool supported by a dynamic Ontology on graduate works of undergraduate students of the Systems Engineering program of the Systems department of the Faculty of Engineering, weakly coupled with the PostgreSQL DBMS (Data Base Management System) [21].

Currently in the Research System of the University of Nariño in the VIIS, the engineer in charge, at the time of searching for these types of research: graduate projects, student projects and teaching projects, comments that there is a delay in the processes, he says that these processes are

not optimal, sometimes it is difficult to find what he wants, in many occasions he has not been able to find what he needs. This indicates that there is neither efficiency nor

Effectiveness guaranteed at the time of conducting research searches, since the search is being performed manually. That is to say that by having a manual information search system that does not even have the qualification of a syntactic search engine, it is deduced that the information does not have a clear structure to be presented and that the processes are inefficient in the searches. This leads to the conclusion that it is absolutely necessary to build the intelligent semantic search engine, since if the problem persists, as the information increases, the searches will be more tedious and wasteful, additionally with the intelligent semantic search engine the amount of user population searching on a certain topic will be greater and will be satisfied by finding the desired results.

Thus, this work provides a tool that allows teachers, students and other researchers to search and consult about the researches that have been carried out at the Universidad de Nariño. A semantic search engine has been built using semantics through the SPARQL query language and RDF language with the management of Ontologies, all this integrated with a Machine Learning model that uses algorithms and libraries such as: NLTK, SpaCy, Word2Vec, Doc2Vec, TF-IDF (Term Frequency- Inverse Document Frequency), BOW (Bag Of Words), among others. In this way we can facilitate the work and allow researchers and the community in general to retrieve and find the requested information efficiently from the research projects that are digitized in the Research System of the Universidad de Nariño. The work was developed by student researchers of the GRIAS research group of the Systems Engineering program of the Faculty of Engineering of the Universidad de Nariño.

2. METHODOLOGY

The methodology used for the work comprises the following stages: appropriation of knowledge; installation and configuration of tools, libraries and technologies; collection, extraction and preparation of research projects; design and development of NATURE.

3. RESULTS

3.1. Appropriation of knowledge

It is highlighted the result of the acquired knowledge of all the topics covered by the project, as well as the various tools and languages used. The learning of topics such as: Semantics, Semantic Web, Ontologies, Search Engines, Machine Learning, Natural Language Processing, Artificial Intelligence and Methontology was obtained. In the same way, the learning in languages such as Python, XML, RDF, OWL and SPARQL was known and reinforced.

3.2. Installation and Configuration of Tools, Libraries and Technologies

It is highlighted the result of the installation and configuration of: Jupyter notebook, Protégé, Owlready2, Apache Jena Fuseki, Elasticsearch, Visual Studio Code, Anaconda, Gensim with Word2Vec and Doc2Vec, Pandas, Numpy, NLTK, SpaCy, etc.

3.3. Collection and Extraction of Research Projects

It is highlighted the result of collecting and extracting information from the research projects of teaching projects, student projects and graduate works that are stored in the research system of the University of Nariño.

It is clarified that currently the difference between student projects and graduate works is that student projects are registered from the first semesters of the university career (from first to eighth) while graduate projects are registered from the last semesters of the university career (seventh onwards) until the moment of appearing as a graduate (if it is the case).

3.4. Preparation of Research Projects

The result of preparing the research projects is highlighted, in such a way that this allowed for navigating through the following stages, anticipating and avoiding inconveniences, errors or problems with respect to the quality of the data.

In this order of ideas, the following phases (from the stage of preparation of research projects) are highlighted:

3.4.1. Data Organization Phase

In this phase, algorithms (created by the authors of this work) were applied to the research projects, this because the projects in the collection and extraction phase were untidy and in conditions not suitable to be treated, managed and worked. Jupyter Notebook was used with Python and Pandas scripts to facilitate the handling of data in series and data frames.

3.4.2. Corpus Creation Phase

In this phase, the corpus for the research projects was created, which was the most powerful input of semantics, as can be seen in the later stages. This corpus resulted from unifying all the data from the research projects (already organized in the previous phase), which were: title and summary of the research; keyword 1, keyword 2, keyword 3, keyword 4, keyword 5; names, surnames, program, faculty, department, research group and line of research for each of the authors and advisers. In this phase, like the previous one, Jupyter Notebook, Python and Pandas were also used to facilitate the handling of data in series and data frames.

3.4.3. Data Pre-processing Phase

In this phase, the NLTK and SpaCy libraries were used to preprocess the data obtained in the previous phase. For this, the following subphases (from the data-preprocessing phase) were used:

Figure. 1 shows an example of a project fragment dealing with physics, to which it will be shown how the different subphases were applied.

En este trabajo se aplica una técnica novedosa conocida como Deep Learning que forma parte de las herramientas de inteligencia computacional, particularmente del área de las Redes Neuronales Artificiales para clasificar dos tipos de sismos volcánicos: volcano-tectónicos o VT y largo periodo o LP. Se introduce al sistema ejemplos que contienen las series de tiempo correspondientes a los registros de sismos de los tipos LP y VT para realizar el 'entrenamiento' del sistema, éste extrae características generales que diferencian un tipo del otro; una vez el sistema ha 'aprendido', es capaz de reconocer, en un registro que nunca ha conocido, las características que distinguen un tipo de otro y catalogarlo de manera correcta.

Figure 1. Example of a project fragment on physics

3.4.3.1. Data Tokenization Subphase

In this subphase, algorithms from the NLTK library were executed to separate all the words and to be able to work with them individually.

Figure 2 shows the data tokenization subphase for Figure 1.

```
['En', 'este', 'trabajo', 'se', 'aplica', 'una', 'técnica', 'novedosa', 'conocida', 'como', 'Deep',
'Learning', 'que', 'forma', 'parte', 'de', 'las', 'herramientas', 'de', 'inteligencia',
'computacional', ',', 'particularmente', 'del', 'área', 'de', 'las', 'Redes', 'Neuronales',
'Artificiales', 'para', 'clasificar', 'dos', 'tipos', 'de', 'sismos', 'volcánicos', ':', 'volcano',
'tectónicos', 'o', 'VT', 'y', 'largo', 'periodo', 'o', 'LP', '.', 'Se', 'introduce', 'al', 'sistema',
'ejemplos', 'que', 'contienen', 'las', 'series', 'de', 'tiempo', 'correspondientes', 'a', 'los',
'registros', 'de', 'sismos', 'de', 'los', 'tipos', 'LP', 'y', 'VT', 'para', 'realizar', 'el',
"'entrenamiento", "'", 'del', 'sistema', ',', 'éste', 'extrae', 'características', 'generales', 'que',
'diferencian', 'un', 'tipo', 'del', 'otro', ';', 'una', 'vez', 'el', 'sistema', 'ha', "'aprendido", "'", ',',
'es', 'capaz', 'de', 'reconocer', ',', 'en', 'un', 'registro', 'que', 'nunca', 'ha', 'conocido', ',', 'las',
'características', 'que', 'distinguen', 'un', 'tipo', 'de', 'otro', 'y', 'catalogarlo', 'de', 'manera',
'correcta', '.']
```

Figure 2. Data tokenization subphase in all the words of Figure 1

3.4.3.2. Data Normalization Subphase

For this subphase, many algorithms were applied so that all the data were under the same standard.

Figure 3 shows the data normalization subphase for Figure 2

```
en este trabajo se aplica una técnica novedosa conocida como deep learning que forma
parte de las herramientas de inteligencia computacional , particularmente del área de las
redes neuronales artificiales para clasificar dos tipos de sismos volcánicos : volcano
tectónicos o vt y largo periodo o lp . se introduce al sistema ejemplos que contienen las
series de tiempo correspondientes a los registros de sismos de los tipos lp y vt para
realizar el 'entrenamiento ' del sistema , éste extrae características generales que
diferencian un tipo del otro ; una vez el sistema ha 'aprendido ' , es capaz de reconocer , en
un registro que nunca ha conocido , las características que distinguen un tipo de otro y
catalogarlo de manera correcta .
```

Figure 3. Data normalization subphase in all the words of Figure 2

3.4.3.3. Data Cleaning Subphase

In this subphase, NLTK and SpaCy algorithms were applied together with regular expressions so that the data is totally clean, this with the elimination of null data, punctuation marks, "non-ascii" characters and stopwords.

Figure 4 shows the data cleaning subphase for Figure 3.

trabajo aplica tecnica novedosa conocida deep learning forma parte herramientas inteligencia computacional particularmente area redes neuronales artificiales clasificar dos tipos sismos volcanicos volcano tectonicos largo periodo introduce sistema ejemplos contienen series tiempo correspondientes registros sismos tipos realizar entrenamiento sistema extrae caracteristicas generales diferencian tipo vez sistema aprendido capaz reconocer registro nunca conocido caracteristicas distinguen tipo catalogarlo manera correcta

Figure 4. Data cleaning subphase in all the words of Figure 3

3.4.3.4. Data Lemmatization Subphase

Finally, in this subphase, the data resulting from the cleaning stage were lemmatized.

Figure 5 shows the data lemmatization subphase for Figure 4.

trabajar aplicar tecnica novedoso conocido deep learning formar partir herramienta inteligencia computacional particularmente area redar neuronal artificial clasificar do tipo sismo volcanicos volcano tectonicos largar periodo introducir sistema ejemplo contener serie tiempo correspondiente registro sismo tipo realizar entrenamiento sistema extraer caracteristicas general diferenciar tipo vez sistema aprender capaz reconocer registrar nunca conocer caracteristicas distinguir tipo catalogarlo manera correcto

Figure 5. Data lemmatization subphase in all the words of Figure 4

3.5. Design of NATURE

Once the previous stage of preparation of the research projects was completed, NATURE was designed. This design was carried out taking into account the conceptualization phase of the Methontology methodology, where the following results stand out:

3.5.1. Conceptualization Phase

Within the conceptualization phase, eleven specific tasks were developed that allowed to successfully conceptualize: classes, attributes, relationships and instances of Ontology. The most important of these are the following:

Task 1. Build the glossary of terms:

This task listed all the important terms selected after analyzing the previous specification phase with its knowledge acquisition process, also this task presented a brief description of each term as shown in Table 1.

Table 1. Glossary of terms of Ontology

Term	Description
Universidad	The education-oriented entity that contains faculties.
Facultad	The entity that contains academic departments.
VIIS	Vice-Chancellor of Research and Social Interaction, is the entity in charge of the research aspect throughout the University, is the one who manages the economic resources for research projects.
Departamento	The entity that contains academic programs.
Convocatoria	This term refers to the convocatory by the VIIS for researchers to come to this convocatory and submit projects in order for them to be financed.
Programa	It is the academic program that is conformed by teachers and students.
Grupo de investigación	It is the group conformed by teachers and/or research students in order to submit projects to the VIIS convocatory.
Docente	Is a researcher who belongs to the University, who carries out projects of teaching type.
Estudiante	Is a researcher who belongs to the University, who carries out projects of student type and/or graduate works.
Investigadorexterno	Is a researcher who is external to the University but who
	presents for the convocatory for VIIS.
Línea de investigación	It is a branch that the research group manages, focused on aspecific area of knowledge.
Investigador	Is the one who develops research projects and submits them tothe VIIS convocatory. This researcher may be a teacher, student or external researcher.
Proyecto de investigación	It is perhaps the most important entity within the researchdomain that contains everything related to a research project.
Palabra	This entity refers to each of the words that conforms the research project, these were used for building the thesaurusand generating a big part of the semantic.

Task 2. Build concept taxonomies:

This task defined the taxonomy or hierarchy of ontology concepts or classes that were obtained from the glossary of terms in task 1, this taxonomy is shown in Figure. 6.

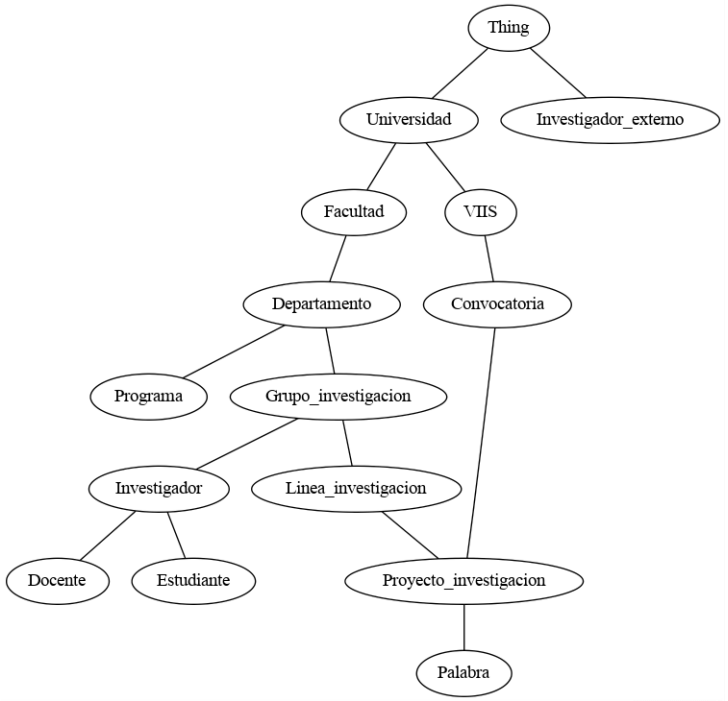

Figure 6. Taxonomy of ontology concepts Task 3. Build binary relation diagrams:

In this task the binary relations diagram that contains the predicates of Ontology was elaborated.

The relations of the most important class of Ontology are visualized in Figure 7, which is "Proyecto de investigación".

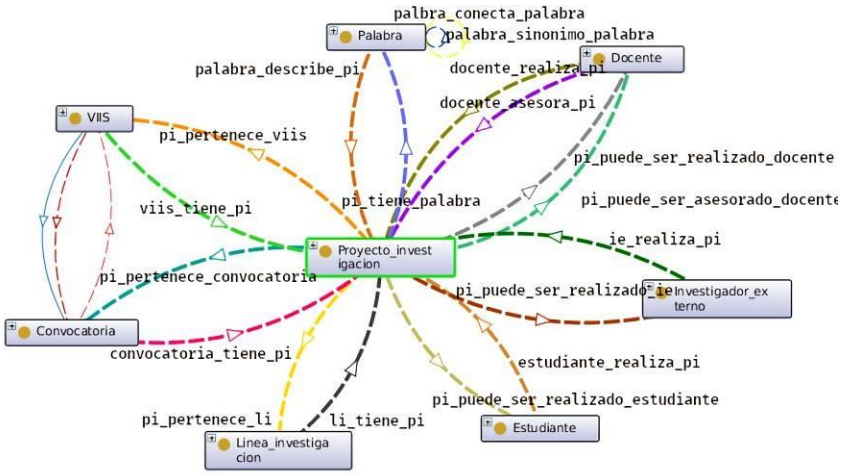

Figure 7. Binary relations diagram for class: "Proyecto de Investigación"

3.6. Development of NATURE

NATURE was developed based on three phases in which the following results are highlighted:

3.6.1. Development with Methontology Phase

For this phase the three subphases of Methontology were applied which are: formalization, implementation and evaluation.

3.6.1.1. Formalization Subphase

This phase highlights the results obtained after using the Protégé tool for the construction of Ontology in semi-computable terms.

3.6.1.2. Implementation Subphase

This phase highlights the results of using the Owlready2 library to encode a computable version of Ontology. Scripts were created and encoding was performed for the handling of Ontology with Python where an entire process of instantiating objects of all classes was performed:

Owlready2 "DataProperties" that correspond to ontology attributes, along with Owlready2 "ObjectProperties" that correspond to Ontology relations were also encoded within the scripts; for each of these elements mentioned, the domain and range were determined. It should be said that Owlready2 reverse relationships are executed in the background, so it was only enough execute the direct relationship.

In synthesis, all classes, attributes, and relations were instantiated within Ontology.

3.6.1.3. Evaluation Subphase

This phase highlights results after having performed functional tests locally and having successfully retrieved the data and other components of ontology with the use of SPARQL and Apache Jena Fuseki server by handling triples of RDF (subject predicate object).

3.6.2. Development with Machine Learning Phase

This phase highlights the results of training with the Machine Learning algorithm with Natural Language Processing such as Word2Vec, which helped to find the context that a word has, in addition a model was trained with the Doc2Vec algorithm, which relies on Word2Vec to find documents that relate to each other, these models make use of neural networks. In this case, the model was trained with the algorithms previously mentioned based on the Skip-Gram model, which attempts to predict words or documents in context given a word or set of base words to search for.

It should be clarified that the output returned by Word2Vec was the input for the process performed with Doc2Vec, this is possible since both algorithms work hand in hand to achieve discover semantic relationships and retrieve information semantically effectively.

To perform the search for similarity between words or documents, of a set of given words, the Gensim library was used, which makes use of the normalization of the vectors obtained from the words to be searched and the calculation of the product point between the normalized vector and each of the vectors corresponding to each word or document trained.

The model was created with data from the preparation stage of research projects, the respective hyper parameters were assigned, the model was trained, the results were evaluated and the hyper parameters were re-fed to satisfactory results, as it is evidenced in Table 2.

Table 2. Hyper parameters for word2vec and doc2vec models

Name	Value	Description
vector_size	300	Dimension of the vector of each of the words in the corpus.
window	5	Refers to the context where the distance between predicted words is chosen.
min_count	1	Minimum words to look for.
dm	0	) indicates that Doc2Vec PV-DBOW is used which is analogous to the Skip-Gram model used in Word2vec. indicates that Doc2Vec PV-DM is used which is analogous to the CBOW model used in Word2Vec.
dbow_words	1	0 indicates that it will train with Doc2Vec. 1 indicates that it will train with Doc2Vec taking Word2Vec input.
hs	0	It is the value with which the neuron will be punished in case the task done is not correct.
negative	20	Number of irrelevant words for negative sampling.
ns_exponent	-0.5	Indicates that frequencies will be sampled equally.
alpha	0.015	Neural network learning rate
min_alpha	0.0001	Rate to be reduced during training.
seed	25	Seed to generate hash for words.
sample	5	Reduction number for high frequency words
epochs	150	Epochs, number of iterations for training.

In Figure 8, Figure 9 and Figure 10 are presented the results of executing the order to find 10 more similar and related words (according to the cosine similarity of the algorithm ordered in percentage terms from highest to lowest) to another word that is specified within of the entire research corpus with a method of the Word2Vec algorithm.

Figure 8 indicates the 10 words most similar and related to the word “cultivos”.

```
modelo_cargado.wv.most_similar(
    positive=['cultivos'], topn=10)
```

```
[('andinos', 0.5301966667175293),
 ('sustituir', 0.5119062662124634),
 ('agrotecnologias', 0.4994411766529083),
 ('totipotencia', 0.49643680453300476),
 ('cebada', 0.49597498774528503),
 ('invitro', 0.49116745591163635),
 ('transitorios', 0.4897231459617615),
 ('hechas', 0.48805299401283264),
 ('potencializador', 0.4799562096595764),
 ('recesivo', 0.47675687074661255)]
```

Figure 8. Result of method with Word2vec for word “cultivos”

Figure 9 indicates the 10 words most similar and related to the word “fresa”.

```
modelo_cargado.wv.most_similar(
    positive=['fresa'], topn=10)
```

```
[('cereza', 0.9345278739929199),
 ('citricos', 0.9311402440071106),
 ('ciruela', 0.9139297604560852),
 ('pera', 0.8887712359428406),
 ('yogurt', 0.7925349473953247),
 ('banano', 0.7829478979110718),
 ('manzana', 0.7650518417358398),
 ('subtropico', 0.6739339828491211),
 ('canada', 0.6399896144866943),
 ('usado', 0.6375434994697571)]
```

Figure 9. Result of method with Word2vec for word “fresa”

Figure 10 indicates the 10 words most similar and related to the word “historia”.

```
modelo_cargado.wv.most_similar(
    positive=['historia'], topn=10)
```

```
[('guerreras', 0.5881358981132507),
 ('feminista', 0.5563334822654724),
 ('musicologia', 0.5516680479049683),
 ('diversion', 0.5437859296798706),
 ('empoderarse', 0.5405762195587158),
 ('juventud', 0.5386302471160889),
 ('constatado', 0.5351422429084778),
 ('enhem', 0.5351074934005737),
 ('amor', 0.5341321229934692),
 ('resignificacion', 0.53216552734375)]
```

Figure 10. Result of method with Word2vec for word “historia”

Figure 11 shows possible semantic relationships between the words:

- entidades - instituciones - academicas - ministerio - promover
- impacto - condiciones
- empresas - comercio - turismo – emprendimiento

Figure 12 shows possible semantic relationships between the words:

- adolescentes - escolares - escolar - bullying
- nivel - mundial
- aportar - contribuir
- relaciones - económicos
- social – sociales

Figure 13 shows possible semantic relationships between the words:

- lado - humano
- programa - diseño - industrial - artes - tecnologicos

- investigacion - curriculo
- desarrollo - proyecto - estudios - docente - grado
- universidad – Nariño
- desarrollar – propuesta

Figure 11. First plot for words semantically related

Figure 12. Second plot for words semantically related

Figure 13. Third plot for words semantically related

3.6.3. Integration of Ontology and Machine Learning Phase

For this phase, Ontology and Machine learning are integrated, providing potency, effectiveness and semantic power to optimize times, resources, and to have greater chances of finding

successful and satisfactory results to certain searches in NATURE, the results are observed in: Figure 14, Figure 15, Figure 16 and Figure 17.

This was achieved by bringing the vectors that Doc2Vec generated to Elasticsearch; Elastic helped in the ranking stage by having speed, scalability and being a distributed analysis engine that favors the search and indexing of research projects.

Afterwards, scripts were created to manage the queries of the research projects for the Ontology with SPARQL, which relies on the trained Word2Vec model to add additional words to the search that are related to those requested and thus find research related to a certain query. In the same way, with Doc2Vec it was possible to infer vectors from a set of supplied words, then as a partial result, the investigations that are related to inferred vectors are presented. Finally, the results obtained in the SPARQL query and the Doc2Vec algorithm are joined, so the final ranking of a search will show consistent, coherent, successful and satisfactory results as requested with the additional ability to recommend documents that may be useful and interest to the user.

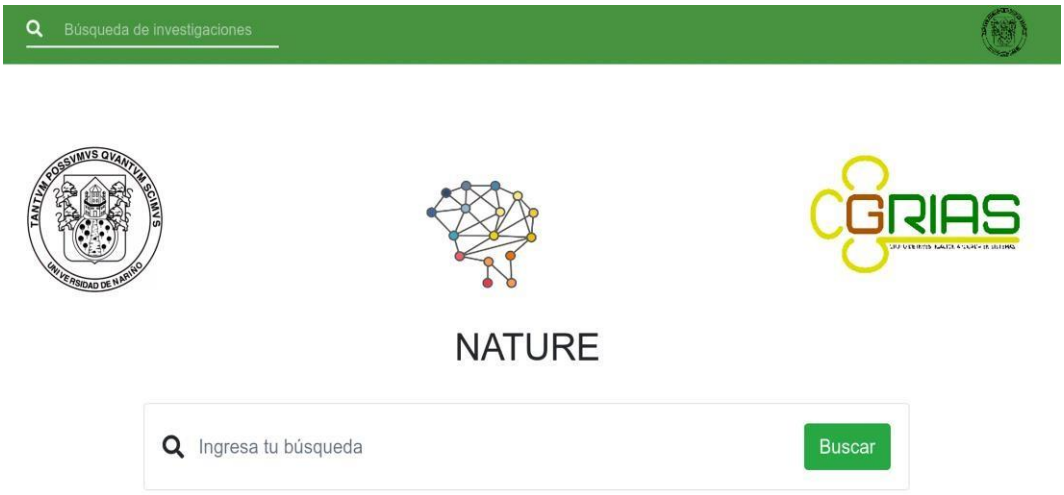

Figure 14. NATURE tool interface

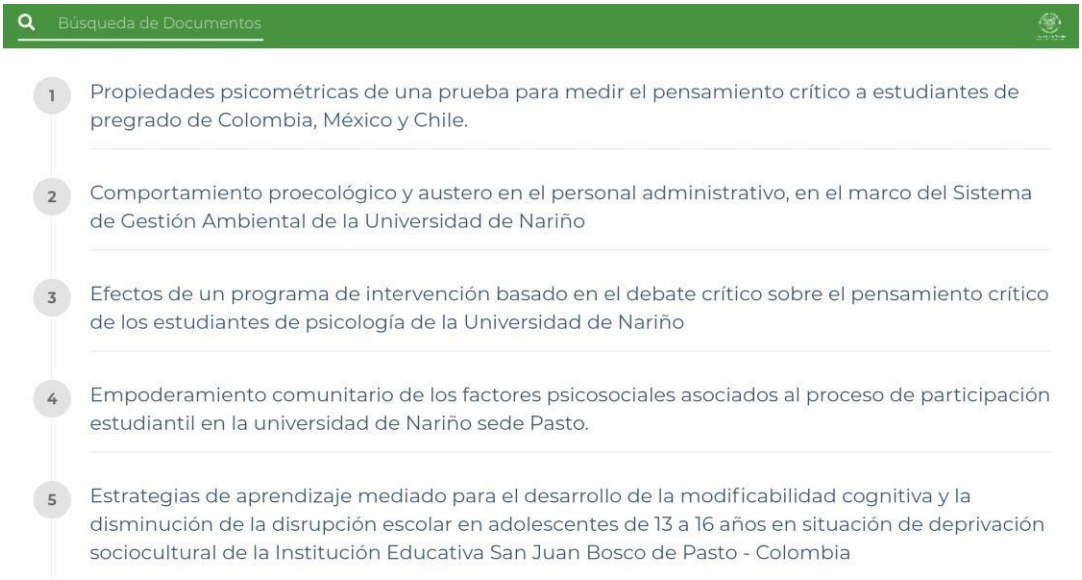

Figure 15. First results for the search: “investigaciones de psicología”

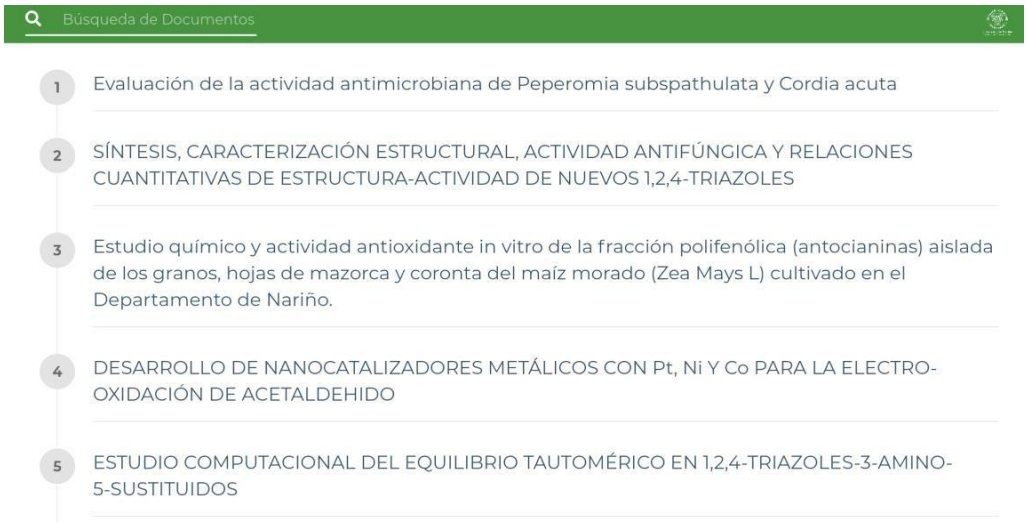

Figure 16. First results for the search: "investigacionessobrequímica"

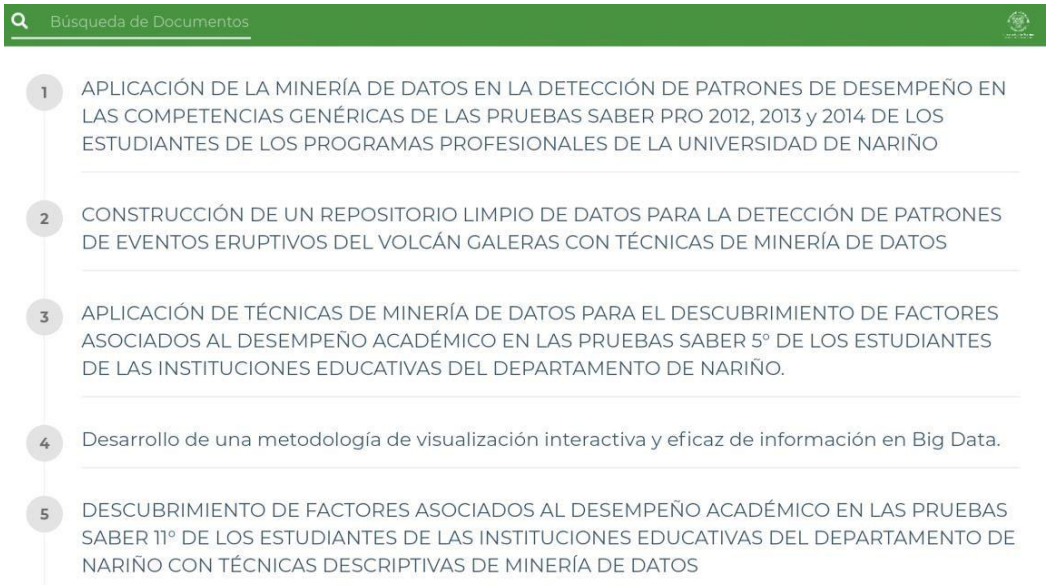

Figure 17. First results for the search: "investigacionesacerca de minería de datos"

4. DISCUSSION

- NATURE provides a degree of optimization, originality and innovation compared to other search engines and knowledge databases such as WordNet, Freebase, DLBP (Digital Bibliography and Library Project), ERCIM digital library, Swoogle, NDLTD, Wolfram Alpha (and others mentioned above) because ontologies are being integrated with Machine Learning with the aforementioned scripts where the ontology is well set up and the algorithms are well trained. In addition to this, the vectors of the words are being managed with Elasticsearch, which save significant memory consumption. Searches are also done with Elasticsearch which is another reason because the search engine is so fast and accurate.

- It is recommended to carry out tests with more data to see how NATURE behaves in the face of an expandable size in the information. This is because the data used were all the research projects that were in the VIIS Research System, but all the projects at the University level are not within that system, but 10%.

- It is proposed to do the coupling of NATURE in other universities and in various non-academic environments, determining the structure of the Ontology and Machine Learning models with their possible variants.

- It is suggested to carry out an analysis of what users are looking for, analyzing the records of searches, downloads, storing everything in the database, then applying data mining with all the information to possibly determine aspects such as: "What semesters do belong people who make queries about astronomy ?" or "What ages do belong people who make queries about psychology?". Machine Learning could be used for this future work perfectly.

- It is also proposed to incorporate in the search engine page a view with its respective database that allows to rate and comment on the search engine in order to observe and analyze how users are rating NATURE, as well as to realize their opinions and whether they are satisfied or not, thus determining the usability of NATURE.

5. CONCLUSIONS

- With the culmination of this research work, NATURE is obtained: A tool resulting from the union of Artificial Intelligence and Natural Language Processing for searching research projects in Colombia. Through the successful development of the project stages, the formulated problem is solved, the objectives set are fullfiled and satisfactory results are obtained. In this way, this tool facilitates the successful search for research projects for teaching projects, student projects and graduate projects at the University of Nariño.

- In the stages of appropriation of knowledge and installation and configuration of the tools, a domain of the various topics was acquired and this contributed to the development of the work and led to the personal training of the researchers as well as made outstanding contributions to the group of GRIAS research (Grupo de InvestigaciónAplicadoenSistemas) and for the University of Nariño in general.

- The stages of collection, extraction and preparation of research projects were extremely important stages that acted as preliminary and prelude stages as input for NATURE. In this vein, it is correct to affirm that without these stages a good development of NATURE could not have been achieved.

- Methontology was a methodology that was perfectly coupled to the project and allowed to build the Ontology following specific phases and tasks with an order, comprehension and accuracy in the processes.

- The Ontology integrated with Machine Learning demonstrated great potency, semantic power and effectiveness in the processes to obtain concrete results according to the searches carried out. This is because Machine Learning algorithms, specifically Natural Language Processing algorithms such as Word2vec and Doc2vec work with neural networks, which were trained with the words from the research project corpus, adapting them to the context and finding the various semantic relationships between them. Likewise,

Ontology acted as a great semantic network whose instances, hand in hand with classes, relations and attributes, interacted under the triple scheme handled by RDF and consulted by SPARQL to extract all the knowledge from the domain of the research projects.

ACKNOWLEDGMENT

To the University of Nariño, to the VIIS (Vicerrectoría de Investigación e Interacción Social) for financing this project and to the Research Community in general for supporting the successful completion of this work.

REFERENCES

[1] VELÁSQUEZ, Torcoroma, PUENTES, Andrés & GUZMÁN, Jaime. Ontologías: unatécnica de representación de conocimiento. En: AvancesenSistemas e Informática. Vol. 8. No. 2. (Julio, 2011), p. 211-216. [Enlínea]. Disponibleen: https://revistas.unal.edu.co/index.php/avances/article/view/26750

[2] GARCÍA, Francisco. Web Semántica y Ontologías. [Enlínea]. Disponibleen: https://www.researchgate.net/publication/267222548_Web_Semantica_y_Ontologias

[3] MOURIÑO, M. Clasificaciónmultilingüe de documentosutilizando machine learning y la wikipedia. [Enlínea]. Disponibleen: https://dialnet.unirioja.es/servlet/tesis?codigo=150295

[4] EFIGENIA, Ana & CANTOR, Sandoval. USO DE ONTOLOGÍAS Y WEB SEMÁNTICA PARA APOYAR LA GESTIÓN DEL CONOCIMIENTO. En: Ciencia e IngenieríaNeogranadina. Vol. 17 No.2. (Diciembre, 2007), p.111-129. [En línea]. Disponible en: https://dialnet.unirioja.es/descarga/articulo/2512191.pdf

[5] GALLO, Manuel, FABRE, Ernesto & GALLO, Manuel. ¿Quéesunbuscador? [Enlínea]. Disponibleen: http://media.axon.es/pdf/98234_1.pdf

[6] FAZZINGA, Bettina, GIANFORME, Giorgio, GOTTLOB, Georg & LUKASIEWICZ, Thomas. Semantic Web Search Based On Ontological Conjunctive Queries. En: SSRN Electronic Journal. [Enlínea]. Disponible en: https://www.researchgate.net/publication/326473981_Semantic_Web_Search_Based_on_Ontological_Co njunctive_Queries

[7] DE PEDRO, A. BuscadoresSemánticos, para quésirven. Ususen la AAPP. [Enlínea]. Disponibleen: http://www.alejandrodepedro.es/buscadores-semanticos-el-paso-al-30

[8] ANDREONI, Antonella, BALDACCI Maria, BIAGONI, Stefania, CARLESI, Carlo, CASTELLI, Donatella, PAGANO, Pasquale, PETERS, Carol & PISANI, Serena. The ERCIM Technical Reference Digital Library. En: D-Lib Magazine. Vol. 5. No. 12. (Diciembre, 1999). [Enlínea]. Disponibleen: http://www.dlib.org/dlib/december99/peters/12peters.html

[9] NDLTD. Networked Digital Library of Theses and Dissertations. [Enlínea]. Disponibleen: http://www.ndltd.org

[10] WolframAlpha Computational Intelligence. [Enlínea]. Disponibleen: https://www.wolframalpha.com

[11] MARTÍN, Javier. Swottibuscador de opiniones. [Enlínea]. Disponibleen: https://loogic.com/swotti-buscador-de-opiniones

[12] BARBERÁ, Consuelo, MILLET, Mercé& TORRES, Emiliano. Estudio del buscadorsemánticoSwoogle. [Enlínea]. Disponibleen: https://www.uv.es/etomar/trabajos/swoogle/swoogle.pdf

[13] CAMACHO, María. Incorporación de unbuscadorsemánticoen la plataformaLdShake para la selección de patroneseducativos. Barcelona, 2013, 76p. Trabajo de grado. Universidad PompeuFabra. Escuela Superior Politécnica UPF. Ingeniería de Telemática. [Enlínea]. Disponibleen: https://repositori.upf.edu/handle/10230/22172

[14] AMARAL, Carlos, LAURENT, Dominique, MARTINS, André, MENDES, Alfonso & PINTO, Cláudia. Design and Implementation of a Semantic Search Engine for Portuguese. [Enlínea]. Disponibleen: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.129.4090&rep=rep1&type=pdf

[15] AUCAPIÑA, Yolanda & PLAZA, C. Buscadorsemánticouniversitario: Caso de estudio Universidad de Cuenca. Cuenca, 2018, 200p. Trabajo de grado (Tesisprevia a la obtencióndelTítulo de IngenieroenSistemas). Universidad de Cuenca. Facultad de Ingeniería. Ingeniería de Sistemas. [Enlínea]. Disponibleen: http://dspace.ucuenca.edu.ec/handle/123456789/30291

[16] UMPIÉRREZ, Francisco. SPARQL Interpreter. Las Palmas de Gran Canaria, 2014, 65p. Trabajo de grado (Trabajo Final de GradoenIngenieríaInformática). Universidad de Las Palmas de Gran Canaria. EscuelaIngenieríaInformática. IngenieríaInformática. [Enlínea]. Disponibleen: https://nanopdf.com/download/0701044000000000pdf_pdf

[17] BACULIMA, Jhon& CAJAMARCA, Marcelo. Diseño e Implementación de unRepositorioEcuatoriano de DatosEnlazadosGeoespaciales. Cuenca, 2014, 131p. Trabajo de grado (Tesis de Gradoprevia a la obtencióndelTítulo: Ingeniero de Sistemas). Universidad de Cuenca. Facultad de Ingeniería. Ingeniería de sistemas. [Enlínea]. Disponibleen: http://dspace.ucuenca.edu.ec/handle/123456789/19876

[18] IGLESIAS, Daniela, MEJÍA, Omar, NIETO, Julio, SÁNCHEZ, Steven & MORENO, Silvia. Construcción de unbuscadorontológico para búsquedassemánticas de proyectos de maestría y doctorado. En: Investigación y Desarrolloen TIC. Vol. 7. No. 1. (Mayo, 2017), p. 7-13. [Enlínea].Disponibleen: https://revistas.unisimon.edu.co/index.php/identic/article/view/2501

[19] BUSTOS, Gabriel. Prototipo de unsistema de integración de recursoscientíficos, diseñado para sufuncionamientoen el espacio de losdatosabiertosenlazados para mejorar la colaboración, la eficiencia y promover la innovaciónen Colombia. Bogotá, 2018. Tesis de Maestría. Universidad Nacional de Colombia. Facultad de Ingeniería. Ingeniería de Sistemas e Industrial. [Enlínea]. Disponibleen: https://repositorio.unal.edu.co/handle/unal/55245

[20] MORENO, Carlos & SÁNCHEZ, Yakeline. Prototipo de buscadorsemánticoaplicado a la búsqueda de libros de Ingeniería de Sistemas y Computaciónen la biblioteca Jorge RoaMartínez de la Universidad Tecnológica de Pereira. Pereira, 2012, 66p. Trabajo de grado. Universidad Tecnológica de Pereira. Facultad de Ingenierías: Eléctrica, Electrónica, Física y Ciencias de la Computación. Ingeniería De Sistemas y Computación. [En línea]. Disponible en:http://repositorio.utp.edu.co/dspace/bitstream/11059/2671/1/0057565M843.pdf

[21] BENAVIDES, Mauricio & GUERRERO, Jimmy. UMAYUX: unmodelo de software de gestión de conocimientosoportadoenunaontologíadinámicadébilmenteacoplado con un gestor de base de datos. San Juan de Pasto, 2014, 145p. Trabajo de grado (Trabajo de gradopresentadocomorequisitoparcial para optar al título de Ingeniero de Sistemas). Universidad de Nariño. Facultad de Ingeniería. Ingeniería de Sistemas. [Enlínea]. Disponibleen: http://sired.udenar.edu.co/2030

[22] Apache Software Foundation. Apache Jena Fuseki. [Enlínea]. Disponibleen: https://jena.apache.org/documentation/fuseki2

[23] ARAUJO, Joaquín. ¿Quées Docker? ¿Qué son loscontenedores? y ¿Porqué no usar VMs? [Enlínea]. Disponibleen: https://platzi.com/tutoriales/1432-docker/1484-guia-del-curso-de-docker

[24] BUDHIRAJA, Amar. A simple explanation of document embeddings generated using Doc2Vec. [Enlínea]. Disponibleen: https://medium.com/@amarbudhiraja/understanding-document-embeddings-of- doc2vec-bfe7237a26da

[25] CHALLENGER, Ivet, DÍAZ, Yanet& BECERRA, Roberto. El lenguaje de programación Python. En: Ciencias Holguín. Vol. XX. No. 2. (Junio, 2014), p. 1-13. [Enlínea]. Disponibleen: www.redalyc.org/articulo.oa?id=181531232001

[26] CHECA, Diego & ROJAS, Oscar. ONTOLOGÍA PARA LOS SISTEMAS HOLÓNICOS DE MANUFACTURA BASADOS EN LA UNIDAD DE PRODUCCIÓN. En: RevistaColombiana de Tecnologías de Avanzada. Vol. 1. No. 23. (Noviembre, 2013), p. 134-141. [Enlínea]. Disponibleen: http://revistas.unipamplona.edu.co/ojs_viceinves/index.php/RCTA/article/view/2334

[27] CLASSORA. Sacandoprovecho a la Web Semántica: SPARQL. [Enlínea]. Disponibleen: http://blog.classora.com/2012/11/05/sacando-provecho-a-la-web-semantica-sparql

[28] CODINA, Lluís& CRISTÒFOL, Rovira. La Web Semántica. En: JesúsTramullas (coord.). Tendenciasendocumentación digital. Guijón: Trea, 2006. p. 9-54. [Enlínea]. Disponibleen: http://eprints.rclis.org/8899

[29] EDWARDS, Gavin. Machine Learning An Introduction. [Enlínea]. Disponibleen: https://towardsdatascience.com/machine-learning-an-introduction-23b84d51e6d0ElasticsearchB.V.Elasticsearch. [En línea]. Disponible en:https://www.elastic.co/es/what- is/elasticsearch

[30] FLORES, Pedro & PORTILLO, Julio. ELABORACIÓN DE PROPUESTA DE GUÍA DE IMPLEMENTACIÓN DE SCRUM PARA EMPRESA SALVADOREÑA, UN CASO DE ESTUDIO. Antiguo Cuscatlán, 2017, 117p. Trabajo de grado (MAESTRO EN ARQUITECTURA DE SOFTWARE). Universidad Don Bosco. Arquitectura de Software. [Enlínea]. Disponibleen: http://rd.udb.edu.sv:8080/jspui/bitstream/11715/1264/1/documento.pdf

[31] FLÓREZ, Héctor. Construcción de ontologías OWL. En: VÍNCULOS. Vol. 4. No. 1. (Diciembre, 2007), p. 19-34. [En línea]. Disponible en: https://revistas.udistrital.edu.co/index.php/vinculos/article/view/4112

[32] Kit de herramientas de lenguaje natural. [Enlínea]. Disponibleen: https://www.nltk.org

[33] LAMY, Jean. Owlready: Ontology-oriented programming in Python with automatic classification and high level constructs for biomedical ontologies. En: Artificial Intelligence in Medicine. Vol. 80. (Agosto, 2017), p. 11-28. [En línea]. Disponible en:https://www.sciencedirect.com/science/article/pii/S0933365717300271

[34] LINCOLN, Matthew. Uso de SPARQL para acceder a datosabiertosenlazados. [Enlínea].Disponibleen: https://programminghistorian.org/es/lecciones/sparql-datos-abiertos-enlazados

[35] LOZANO, Adolfo. Ontologíasen la Web Semántica. [Enlínea]. Disponibleen: http://eolo.cps.unizar.es/docencia/MasterUPV/Articulos/Ontologias%20en%20la%20Web%20Semantica. pdf

[36] MUÑOZ, José. Introducción a flask. [Enlínea]. Disponibleen: https://plataforma.josedomingo.org/pledin/cursos/flask/curso/u05/

[37] PEDRAZA, Rafael, CODINA, Lluís& CRISTÒFOL, Rovira. Web semántica y ontologíasen el procesamiento de la información documental. En: El profesional de la información. Vol. 16. No. 6. (Noviembre, 2007), p. 569-579. [Enlínea]. Disponibleen: https://repositori.upf.edu/handle/10230/13141

AUTHORS

FELIPE CUJAR ROSERO: Research student of the GRIAS group of the University of Nariño with publication of papers, presentations, poster exhibition and certifications in the areas of database knowledge, artificial intelligence and web development. Link:https://www.linkedin.com/in/felipe-cujar/.https://scienti.minciencias.gov.co/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0001853544.https://scholar.google.com/citations?user=dX12cEAAAAAJ. https://github.com/fcujar

DAVID SANTIAGO PINCHAO ORTIZ: Research student of the GRIAS group of the University of Nariño with publication of papers, presentations, poster exhibition and certifications in the areas of database knowledge, artificial intelligence and web development. Link: https://co.linkedin.com/in/sangeeky. https://github.com/SanGeeky

SILVIO RICARDO TIMARÁN PEREIRA: Doctor of Engineering. Director of Research Group GRIAS. Professor in the Systems Department of the University of Nariño. Link: Researcher:http://scienti.colciencias.gov.co:8081/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0000250988

MATEO GUERRERO RESTREPO: Master in Engineering. GRIAS Group Researcher. Professor Hora chair of Systems Department of the University of Nariño. Link: Researcher: http://scienti.minciencias.gov.co:8081/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0001489230