FENIX: A SEMANTIC SEARCH ENGINE BASED ON AN ONTOLOGY AND A MODEL TRAINED WITH MACHINE LEARNING TO SUPPORT RESEARCH

Felipe Cujar-Rosero, David Santiago Pinchao Ortiz, Silvio Ricardo Timaran 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 to build a Semantic Search Engine that uses an Ontology and a model trained with Machine Learning to support the semantic search of research projects of the System of Research from the University of Nariño. For the construction of FENIX, as this Engine is called, it was used a methodology that includes the stages: appropriation of knowledge, installation and configuration of tools, libraries and technologies, collection, extraction and preparation of research projects, design and development of the Semantic Search Engine. 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 and specifically Natural Language Processing algorithms were used such as: SpaCy, NLTK, Word2vec and Doc2vec, this was also done in Jupyter Notebook with Python within the virtual environment of anaconda and with Elasticsearch; and c) the creation of FENIX managing and unifying the queries for the Ontology and for the Machine Learning model. The tests showed that FENIX was successful in all the searches that were carried out because its results were satisfactory.

KEYWORDS

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

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].

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

David C. Wyld et al. (Eds): CCSIT, SIPP, PDCTA, AISC, NLPCL, BIGML, NCWMC - 2021
pp. 97-115, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.110709

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 Swoogle has 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 Pompeu Fabra-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 degree 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 degree 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 Roa Martínez 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 degree 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, the Research System of the University of Nariño does not have a tool that allows teachers, students and other researchers to carry out effective searches and queries about the research projects that have been carried out in that University. For this reason, in order to solve this problem, it was proposed to build a search engine making use of semantics through the SPARQL query language, the RDF language with the management of Ontologies and Machine Learning with a specific area called Natural Language Processing. In this way, the work can be

facilitated and the researchers and the community in general can recover and find the information requested, successfully, from the research projects that are digitized in the Research System of the University of Nariño. 85% of research projects are in Spanish language.

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 the semantic search engine.

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 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 degree works that are stored in the research system of the University of Nariño.

It is clarified that currently the difference between student projects and degree works is that student projects are registered from the first semesters of the university career (from first to eighth) while degree 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:

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.

3.4.3.2. Data Normalization Subphase

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

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.

3.4.3.4. Data Lemmatization Subphase

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

3.5. Design of Semantic Search Engine

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

3.5.1. Specification Phase

Within this phase, the reasons that that allowed to make the Ontology were identified, it was also described the end users who make use of the Semantic Search Engine. Also a knowledge acquisition process was carried out; this process of acquiring knowledge differs from the appropriation of the general knowledge of the project, since this acquisition is mainly focused on the design of the Search Engine.

Thus, it is highlighted that the Ontology was created to give a robust semantic component to the Search Engine interacting with the research projects and their components. The end users will be all those researchers who wish to access the research project resources through various searches.

3.5.2. Conceptualization Phase

Within the conceptualization phase, eleven specific tasks were developed that allowed to successfully conceptualize: classes, attributes, relationships and instances of Ontology. These tasks were:

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	He/She is a researcher who belongs to the University, who carries out projects of teaching type.
Estudiante	He/She is a researcher who belongs to the University, who carries out projects of student type and/or degree works.
Investigador externo	He/She 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 a specific area of knowledge.
Investigador	He/She is the one who develops research projects and submits them to the VIIS convocatory. This researcher may be a teacher, student or external researcher.
Proyecto de investigación	It is perhaps the most important entity within the research domain 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 thesaurus and 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. 1.

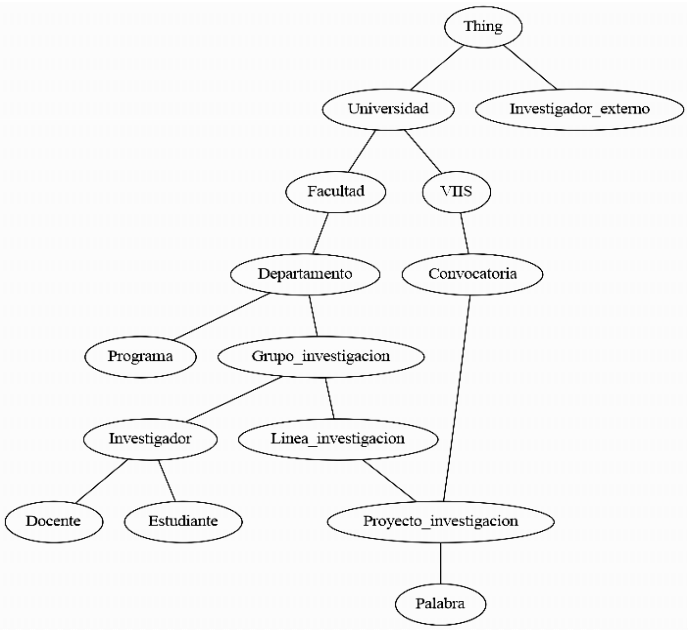

Figure. 1 Taxonomy of ontology concepts

Task 3. Build ad hoc 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. 2, which is "Proyecto de investigación".

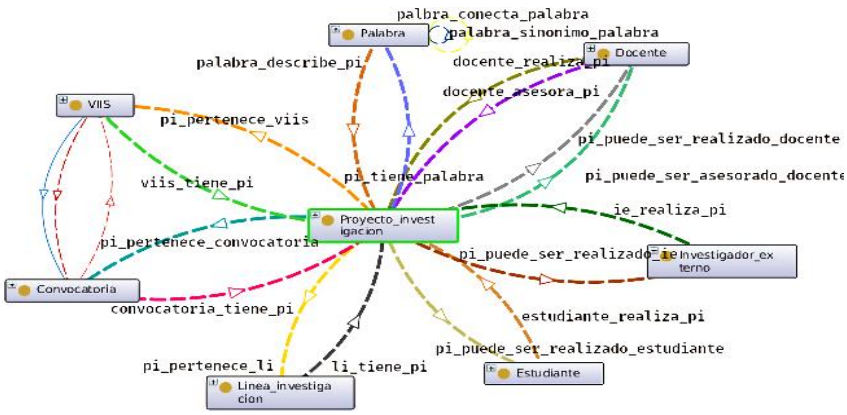

Figure. 2 Binary relations diagram for class: "Proyecto de Investigación"

Task 4. Build concepts dictionary:

This task detailed the most important or relevant concepts or classes within the research domain of Ontology, highlighting its attributes, relations and instances. "Proyecto de investigación" was

chosen because this class have all the raw material with many words for discovering the semantic power. It is shown in Table 2.

Table 2. Concept dictionary for "Proyecto de Investigación"

Class Proyecto de Investigación	
Attributes	id_proyecto_investigacion titulo_proyecto_investigacion resumen_proyecto_investigacion palabras_clave estado_proyecto_investigacion tipo_proyecto_investigacion
Relations	pi_tiene_palabra pi_pertenece_li pi_pertenece_convocatoria pi_puede_ser_realizado_estudiante pi_puede_ser_realizado_docente pi_puede_ser_realizado_ie pi_puede_ser_asesorado_docente pi_pertenece_viis
Instances	Proyecto investigación

Task 5. Describe ad hoc binary relations:

A total of 55 binary relationships were obtained, of which "pi_tiene_palabra" is highlighted because each project relates to its words in such "pi_tiene_palabra" relation facilitates searches.

It is shown in Table 3.

For each relationship its origin class (domain), destination class (range), inverse relation and cardinality were obtained.

Table 3. Binary relation (pi_tiene_palabra) in detail

Relation pi_tiene_palabra	
Origin Class (Domain)	Proyecto de investigación
Destination Class (Range)	Palabra
Inverse Relation	palabra_describe_pi
Cardinality	1:N

Task 6. Describe instance attributes:

A total of 45 attributes were obtained, of which the most representative class ("Proyecto de investigación") is shown. Instance attributes of this class can be seen in Table 4.

For each class, the instance attributes, the class name (domain), the data type (range) and its cardinality are displayed.

Table 4. Instance attributes for class: “Proyecto de Investigación”

Attribute	Data Type (Range)	Cardinality
id_proyecto_investigacion	Int	1
titulo_proyecto_investigacion	String	1
resumen_proyecto_investigacion	String	1
palabras_clave	String	0:5
estado_proyecto_investigacion	String	1
tipo_proyecto_investigacion	String	1

Task 7. Describe class attributtes:

This task defined class attributes that serve as cardinal constraints for each class. These can be observed in Table 5 for classes “Investigador”, “Docente” and “Proyecto de Investigación”.

Table 5. Class attributes for: “Investigador”, “Docente” and “Proyecto de Investigación”

Class	Attribute
Investigador	Maximum 4 researchers per research project.
Docente	Maximum 2 teachers can advise research projects.
Proyecto de Investigación	Maximum 2 years and minimum 6 months duration of the research project.

Task 8. Describe constants:

The need to use constants for this Ontology was not contemplated.

Task 9. Describe formal axioms:

There was no need to use axioms that are predicates (relationships) that are always fulfilled, i.e. they are always affirmative.

Task 10. Describe rules:

Because this Ontology did not introduce formal axioms, no rules were necessary.

Task 11. Describe instances:

Instances were obtained for each of the classes contemplated in ontology: “Universidad, Facultad, VIIS, Departamento, Convocatoria, Programa, Grupo de investigación, Docente, Estudiante, Investigador externo, Línea de investigación, Investigador, Proyecto de investigación y Palabra”. This was achieved with the previous stage of preparation of research projects.

3.6. Development of the Semantic Search Engine

FENIX was developed based on three phases (from the stage of Development of the Semantic Search Engine) 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 6.

Table 6. Hyperparameters 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	0 indicates that Doc2Vec PV-DBOW is used which is analogous to the Skip-Gram model used in Word2vec. 1 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. 3, Figure. 4 and Figure. 5 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. 3 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. 3 Result of method with Word2vec for word “cultivos”

Figure. 4 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. 4 Result of method with Word2vec for word “fresa”

Figure. 5 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. 5 Result of method with Word2vec for word “historia”

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 FENIX, the results are observed in: Figure. 6, Figure. 7 and Figure. 8.

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.

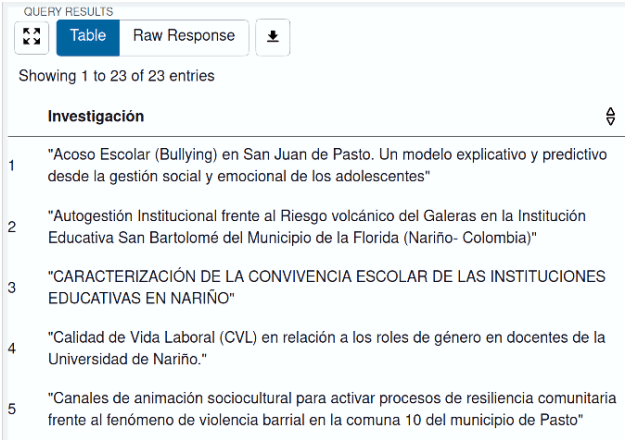
QUERY RESULTS

Table | Raw Response

Showing 1 to 23 of 23 entries

	Investigación
1	"Acoso Escolar (Bullying) en San Juan de Pasto. Un modelo explicativo y predictivo desde la gestión social y emocional de los adolescentes"
2	"Autogestión Institucional frente al Riesgo volcánico del Galeras en la Institución Educativa San Bartolomé del Municipio de la Florida (Nariño- Colombia)"
3	"CARACTERIZACIÓN DE LA CONVIVENCIA ESCOLAR DE LAS INSTITUCIONES EDUCATIVAS EN NARIÑO"
4	"Calidad de Vida Laboral (CVL) en relación a los roles de género en docentes de la Universidad de Nariño."
5	"Canales de animación sociocultural para activar procesos de resiliencia comunitaria frente al fenómeno de violencia barrial en la comuna 10 del municipio de Pasto"

Figure. 6 First results for the search: “investigaciones de psicología”

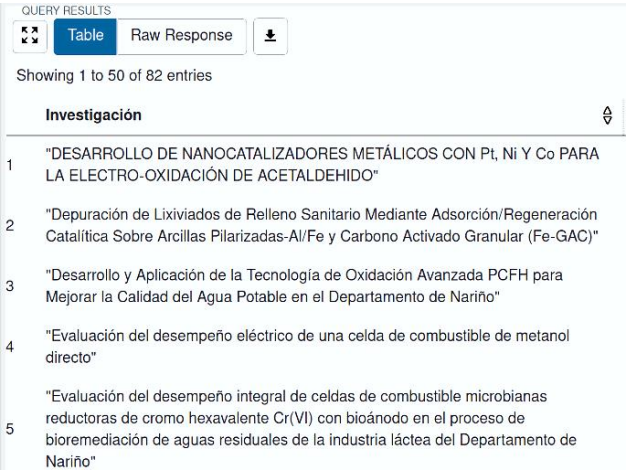
QUERY RESULTS

Table | Raw Response

Showing 1 to 50 of 82 entries

	Investigación
1	"DESARROLLO DE NANOCATALIZADORES METÁLICOS CON Pt, Ni Y Co PARA LA ELECTRO-OXIDACIÓN DE ACETALDEHIDO"
2	"Depuración de Lixiviados de Relleno Sanitario Mediante Adsorción/Regeneración Catalítica Sobre Arcillas Pilarizadas-Al/Fe y Carbono Activado Granular (Fe-GAC)"
3	"Desarrollo y Aplicación de la Tecnología de Oxidación Avanzada PCFH para Mejorar la Calidad del Agua Potable en el Departamento de Nariño"
4	"Evaluación del desempeño eléctrico de una celda de combustible de metanol directo"
5	"Evaluación del desempeño integral de celdas de combustible microbianas reductoras de cromo hexavalente Cr(VI) con bioánodo en el proceso de bioremediación de aguas residuales de la industria láctea del Departamento de Nariño"

Figure. 7 First results for the search: “investigaciones sobre química”

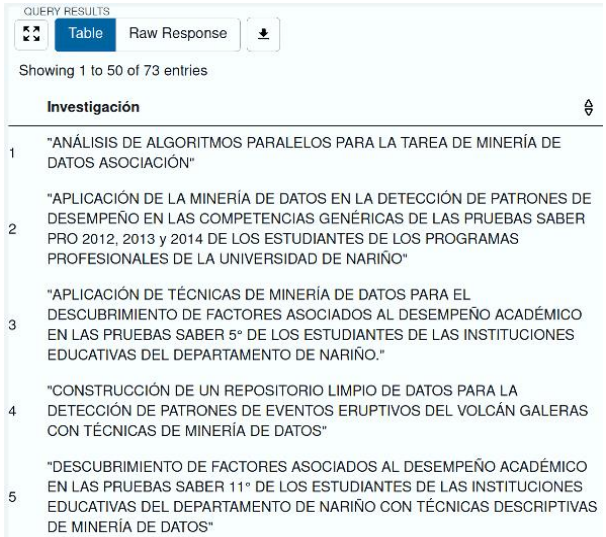
QUERY RESULTS

Table | Raw Response

Showing 1 to 50 of 73 entries

	Investigación
1	"ANÁLISIS DE ALGORITMOS PARALELOS PARA LA TAREA DE MINERÍA DE DATOS ASOCIACIÓN"
2	"APLICACIÓN DE LA MINERÍA DE DATOS EN LA DETECCIÓN DE PATRONES DE DESEMPEÑO EN LAS COMPETENCIAS GENÉRICAS DE LAS PRUEBAS SABER PRO 2012, 2013 y 2014 DE LOS ESTUDIANTES DE LOS PROGRAMAS PROFESIONALES DE LA UNIVERSIDAD DE NARIÑO"
3	"APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA EL DESCUBRIMIENTO DE FACTORES ASOCIADOS AL DESEMPEÑO ACADÉMICO EN LAS PRUEBAS SABER 5° DE LOS ESTUDIANTES DE LAS INSTITUCIONES EDUCATIVAS DEL DEPARTAMENTO DE NARIÑO."
4	"CONSTRUCCIÓN DE UN REPOSITORIO LIMPIO DE DATOS PARA LA DETECCIÓN DE PATRONES DE EVENTOS ERUPTIVOS DEL VOLCÁN GALERAS CON TÉCNICAS DE MINERÍA DE DATOS"
5	"DESCUBRIMIENTO DE FACTORES ASOCIADOS AL DESEMPEÑO ACADÉMICO EN LAS PRUEBAS SABER 11° DE LOS ESTUDIANTES DE LAS INSTITUCIONES EDUCATIVAS DEL DEPARTAMENTO DE NARIÑO CON TÉCNICAS DESCRIPTIVAS DE MINERÍA DE DATOS"

Figure. 8 First results for the search: “investigaciones acerca de minería de datos”

4. DISCUSSION

√ FENIX 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 FENIX 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 FENIX 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 FENIX, as well as to realize their opinions and whether they are satisfied or not, thus determining the usability of FENIX.

5. CONCLUSIONS

√ With the culmination of this research work, FENIX is obtained: A Semantic Search Engine based on an Ontology and a Machine Learning model for research projects at the University of Nariño. 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 degree 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 Investigación Aplicado en Sistemas) 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 FENIX. In this vein, it is correct to affirm that without these stages a good development of FENIX 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: una técnica de representación de conocimiento. En: Avances en Sistemas e Informática. Vol. 8. No. 2. (Julio, 2011), p. 211-216. [En línea]. Disponible en: https://revistas.unal.edu.co/index.php/avances/article/view/26750

[2] GARCÍA, Francisco. Web Semántica y Ontologías. [En línea]. Disponible en: https://www.researchgate.net/publication/267222548_Web_Semantica_y_Ontologias

[3] MOURIÑO, M. Clasificación multilingüe de documentos utilizando machine learning y la wikipedia. [En línea]. Disponible en: 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 Ingeniería Neogranadina. 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. ¿Qué es un buscador? [En línea]. Disponible en: 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. [En línea]. Disponible en: https://www.researchgate.net/publication/326473981_Semantic_Web_Search_Based_on_Ontological_Conjunctive_Queries

[7] DE PEDRO, A. Buscadores Semánticos, para qué sirven. Usos en la AAPP. [En línea]. Disponible en: 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). [En línea]. Disponible en: http://www.dlib.org/dlib/december99/peters/12peters.html

[9] NDLTD. Networked Digital Library of Theses and Dissertations. [En línea]. Disponible en: http://www.ndltd.org

[10] WolframAlpha Computational Intelligence. [En línea]. Disponible en: https://www.wolframalpha.com

[11] MARTÍN, Javier. Swotti buscador de opiniones. [En línea]. Disponible en: https://loogic.com/swotti-buscador-de-opiniones

[12] BARBERÁ, Consuelo, MILLET, Mercé & TORRES, Emiliano. Estudio del buscador semántico Swoogle. [En línea]. Disponible en: https://www.uv.es/etomar/trabajos/swoogle/swoogle.pdf

[13] CAMACHO, María. Incorporación de un buscador semántico en la plataforma LdShake para la selección de patrones educativos. Barcelona, 2013, 76p. Trabajo de grado. Universidad Pompeu Fabra. Escuela Superior Politécnica UPF. Ingeniería de Telemática. [En línea]. Disponible en: 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. [En línea]. Disponible en: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.129.4090&rep=rep1&type=pdf

[15] AUCAPIÑA, Yolanda & PLAZA, C. Buscador semántico universitario: Caso de estudio Universidad de Cuenca. Cuenca, 2018, 200p. Trabajo de grado (Tesis previa a la obtención del Título de Ingeniero en Sistemas). Universidad de Cuenca. Facultad de Ingeniería. Ingeniería de Sistemas. [En línea]. Disponible en: 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 Grado en Ingeniería Informática). Universidad de Las Palmas de Gran Canaria. Escuela Ingeniería Informática. Ingeniería Informática. [En línea]. Disponible en: https://nanopdf.com/download/0701044000000000pdf_pdf

[17] BACULIMA, Jhon & CAJAMARCA, Marcelo. Diseño e Implementación de un Repositorio Ecuatoriano de Datos Enlazados Geoespaciales. Cuenca, 2014, 131p. Trabajo de grado (Tesis de Grado previa a la obtención del Título: Ingeniero de Sistemas). Universidad de Cuenca. Facultad de Ingeniería. Ingeniería de sistemas. [En línea]. Disponible en: http://dspace.ucuenca.edu.ec/handle/123456789/19876

[18] IGLESIAS, Daniela, MEJÍA, Omar, NIETO, Julio, SÁNCHEZ, Steven & MORENO, Silvia. Construcción de un buscador ontológico para búsquedas semánticas de proyectos de maestría y doctorado. En: Investigación y Desarrollo en TIC. Vol. 7. No. 1. (Mayo, 2017), p. 7-13. [En línea]. Disponible en: https://revistas.unisimon.edu.co/index.php/identic/article/view/2501

[19] BUSTOS, Gabriel. Prototipo de un sistema de integración de recursos científicos, diseñado para su funcionamiento en el espacio de los datos abiertos enlazados para mejorar la colaboración, la eficiencia y promover la innovación en Colombia. Bogotá, 2018. Tesis de Maestría. Universidad Nacional de Colombia. Facultad de Ingeniería. Ingeniería de Sistemas e Industrial. [En línea]. Disponible en: https://repositorio.unal.edu.co/handle/unal/55245

[20] MORENO, Carlos & SÁNCHEZ, Yakeline. Prototipo de buscador semántico aplicado a la búsqueda de libros de Ingeniería de Sistemas y Computación en la biblioteca Jorge Roa Martínez 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: un modelo de software de gestión de conocimiento soportado en una ontología dinámica débilmente acoplado con un gestor de base de datos. San Juan de Pasto, 2014, 145p. Trabajo de grado (Trabajo de grado presentado como requisito parcial para optar al título de Ingeniero de Sistemas). Universidad de Nariño. Facultad de Ingeniería. Ingeniería de Sistemas. [En línea]. Disponible en: http://sired.udenar.edu.co/2030

[22] Apache Software Foundation. Apache Jena Fuseki. [En línea]. Disponible en: https://jena.apache.org/documentation/fuseki2

[23] ARAUJO, Joaquín. ¿Qué es Docker? ¿Qué son los contenedores? y ¿Por qué no usar VMs? [En línea]. Disponible en: https://platzi.com/tutoriales/1432-docker/1484-guia-del-curso-de-docker

[24] BUDHIRAJA, Amar. A simple explanation of document embeddings generated using Doc2Vec. [En línea]. Disponible en: 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. [En línea]. Disponible en: 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: Revista Colombiana de Tecnologías de Avanzada. Vol. 1. No. 23. (Noviembre, 2013), p. 134-141. [En línea]. Disponible en: http://revistas.unipamplona.edu.co/ojs_viceinves/index.php/RCTA/article/view/2334

[27] CLASSORA. Sacando provecho a la Web Semántica: SPARQL. [En línea]. Disponible en: http://blog.classora.com/2012/11/05/sacando-provecho-a-la-web-semantica-sparql
[28] CODINA, Lluís & CRISTÒFOL, Rovira. La Web Semántica. En: Jesús Tramullas (coord.). Tendencias en documentación digital. Guijón: Trea, 2006. p. 9-54. [En línea]. Disponible en: http://eprints.rclis.org/8899
[29] EDWARDS, Gavin. Machine Learning An Introduction. [En línea]. Disponible en: https://towardsdatascience.com/machine-learning-an-introduction-23b84d51e6d0
Elasticsearch B.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. [En línea]. Disponible en: 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. [En línea]. Disponible en: 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 datos abiertos enlazados. [En línea]. Disponible en: https://programminghistorian.org/es/lecciones/sparql-datos-abiertos-enlazados
[35] LOZANO, Adolfo. Ontologías en la Web Semántica. [En línea]. Disponible en: http://eolo.cps.unizar.es/docencia/MasterUPV/Articulos/Ontologias%20en%20la%20Web%20Semantica.pdf
[36] MUÑOZ, José. Introducción a flask. [En línea]. Disponible en: https://plataforma.josedomingo.org/pledin/cursos/flask/curso/u05/
[37] PEDRAZA, Rafael, CODINA, Lluís & CRISTÒFOL, Rovira. Web semántica y ontologías en el procesamiento de la información documental. En: El profesional de la información. Vol. 16. No. 6. (Noviembre, 2007), p. 569-579. [En línea]. Disponible en: https://repositori.upf.edu/handle/10230/13141
[38] PEREZ, Fernando & GRANGER, Brian. Project Jupyter. [En línea]. Disponible en: https://jupyter.org PostgreSQL. [En línea]. Disponible en https://www.postgresql.org/about/
[39] ROCCA, Joseph. A simple introduction to Machine Learning. [En línea]. Disponible en: https://towardsdatascience.com/introduction-to-machine-learning-f41aabc55264
[40] SHETTY, Badreesh. Natural Language Processing (NLP) for Machine Learning. [En línea]. Disponible en: https://towardsdatascience.com/natural-language-processing-nlp-for-machine-learning-d44498845d5b
[41] SHPERBER, Gidi. A gentle introduction to Doc2Vec. [En línea]. Disponible en: https://medium.com/wisio/a-gentle-introduction-to-doc2vec-db3e8c0cce5e
[42] Sistema de Información de Investigaciones. [En línea]. Disponible en: http://sisinfoviis.udenar.edu.co
[43] SpaCy 101: todo lo que necesita saber. [En línea]. Disponible en: https://spacy.io/usage/spacy-101
[44] TABARES, John & JIMÉNEZ, Jovani. Ontología para el proceso evaluativo en la educación superior. En: Revista Virtual Universidad Católica del Norte. Vol. 1. No. 42. (Agosto, 2014), p. 68-79. [En línea]. Disponible en: https://revistavirtual.ucn.edu.co/index.php/RevistaUCN/article/view/495

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://scholar.google.com/citations?user=dX12cEAAAAAJ
https://scienti.minciencias.gov.co/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0001853544

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

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