Chang, M. et al. (Eds.) (2019). Proceedings of the 27th International Conference on Computers in Education. Taiwan: Asia-Pacific Society for Computers in Education

Multi-label search platform for open educational resources based on purposes learning

Julio Vera-Sancho[a], Gustavo Suero-Soto[b], Lushianna Tejada[c] , Sonia Castro-Cuba-Sayco [d] & Klinge Villalba-Condori[e]
[a, b, c, d & e]*Universidad Catolica de Santa Maria, Peru*

[a]jveras@ucsm.edu.pe
[b]gsuero@ucsm.edu.pe
[c]ltejada@ucsm.edu.pe
[c]scastroc@ucsm.edu.pe
[d]kvillalba@ucsm.edu.pe

Abstract: Open Educational Resources (OER) repositories stores a large amount and variety of data from multiple sources, and these presents relevant information to the educational learning process; but nowadays, thinking computing takes great importance, and that's why a search system of specialized educational resources in this area is of great need. This work presents the use of Data Mining in OER repositories, for the construction of a multi-label search platform. The process of extracting information is based on Web Scraping techniques, and the use of a multi-label classifier based on Multi Layer Perceptron. This work contributes to a search of OERs based on purposes learning, using as a case study the area of thinking computing.

Keywords: neural networks, classification, scraping, resources educational resources

1. Introduction

Currently, the internet has managed to connect all people and has generated a great amount of information, including in education, generating educational material to improve teaching in schools. All this educational material is available and stored in different repositories of open educational resources (OER) (Liñán and Perez, 2015) that are oriented to help teachers to prepare a better session to improve the learning process based on the characteristics and progress of the students. In this way, the teacher performs different searches to find content that suits what He needs and the class that he wants to teach. This work proposes a system of recommendation based on purpose, where the teacher at the time to performs the search for a particular topic in addition to choosing one or several labels, which will help to choose the purpose of the class to give. The proposed system enters to different repositories, to obtain the content related to the topic, and to identify the purpose of each metadata found, and to show the desired result by the teacher at the end.

2. Theorical Framework

2.1 *Purpose Learning*

When we talk about the purpose of learning, it is based on the pedagogical intentionality of the teacher, who is responsible of the design of your class. The Constructive Alignment is very pedagogically efficient and is based on the SOLO taxonomy (Structure of Observed Learning Outcome) (Levicoy, Obreque, Vásquez, y Salvatierra, 2017) which distinguishes between superficial learning and deep learning. As students learn, the results of their learning first show quantitative and then qualitative phases of increasingly complex structure (Villalba, Cuba, Deco, Bender, y García-Peñalvo, 2017) This

taxonomy is that we find levels of understanding and recurrent actions that are important for learning and each level of understanding is linked to actions as shown below in the table 1:

Table 1

Levels of understanding and recurrent action

Levels of understanding	Abstract enlarged
Relational	0
Multistructural	0
Unistructural	0
Recurrent actions	0

The SOLO taxonomy is an excellent means to classify the expected learning from the most concrete levels to the most abstract and complex levels. Which ensures us to have a learning that goes from the most superficial to the deepest (Villalba-Condori 2018).

2.2 *Open Educational Resources (OER)*

The OER may also be referred to as a digital object or object of digital learning, which serves to provide information and / or knowledge, as well as can help the generation of knowledge, skills and attitudes according to what the person needs(Aguila and Burgos, 2010). The use of TICs each time are more necessary in different fields, and one field where is taking strength is Education, to improve the process of teaching and learning. Open educational resources (OER) are found as shared open content in public repositories(Arias, Zermeño, and Chávez, 2015) which we can make use of, but going further we can say that the OER itself is to share the result of a design process (instructional), in which the knowledge and experience of a designer (teacher) is applied to produce a specific resource (Arias and cols., 2015), describing what the process is in a more exact way and how the content is shown in the REA, and in a certain way it is improving the content that is stored every day, thanks to the feedback that is generated (Espinoza-Suarez 2019).

2.3 *Web Scraping*

The World Wide Web contains a great amount of information that is valuable resource for tasks, such as machine learning, data mining or systems of recommendation (Smedt and Daelemans, 2012). Web Scraping is a software technique aimed at extracting information from websites (Vargiu and Urru, 2013). They generally simulate the manual exploration of the World Wide Web through the use of low-level hypertext extraction protocols or integration of web browsers. It is a technique of information retrieval from the Web through a bot that is generally called Spider.

This technique is used to extract information and be able to give life to search engines or recommendation systems. The extraction of information is mainly the extraction of text, an information retrieval task aimed at automatically discovering information from different text resources (Ristoski, Paulheim, Svátek, and Zeman, 2016). In extracting information, text mining is used to remove relevant information from text files by relying on linguistic and statistical algorithms *(Vargiu and Urru, 2013).*

2.4 *Scrapy*

Scrapy is an open source web crawling framework written in Python. Its main objective is to provide help and support for an efficient web crawling practice. Scrapy allows you to track websites and extract data that can be structured or not to be used in a wide range of applications of scientific interest (Kouzis-Loukas, 2016).

2.4.1 *Scrapy Architecture*

As shown in Figure 1, the architecture of the Scrapy library with its components that allows generating a data flow that takes place within the data extraction system. As a whole, each one of the components and the data flow allows obtaining more detailed and clear information of each website (Smedt and Daelemans, 2012). The Scrapy engine is its main component, whose main objective is to control the

flow of data among all the other components. The engine generates requests and manages events against an action. The Scheduler receives the requests sent by the engine and queues them (Wang and Guo, 2012). The aim of the downloader is to search all web pages and send them to the engine, which then sends web pages to spiders, which contains the logic that allows analyzing websites and extract data from each particular website. The element pipe processes the elements side by side after the spiders extract them (Myers and McGuffee, 2015).

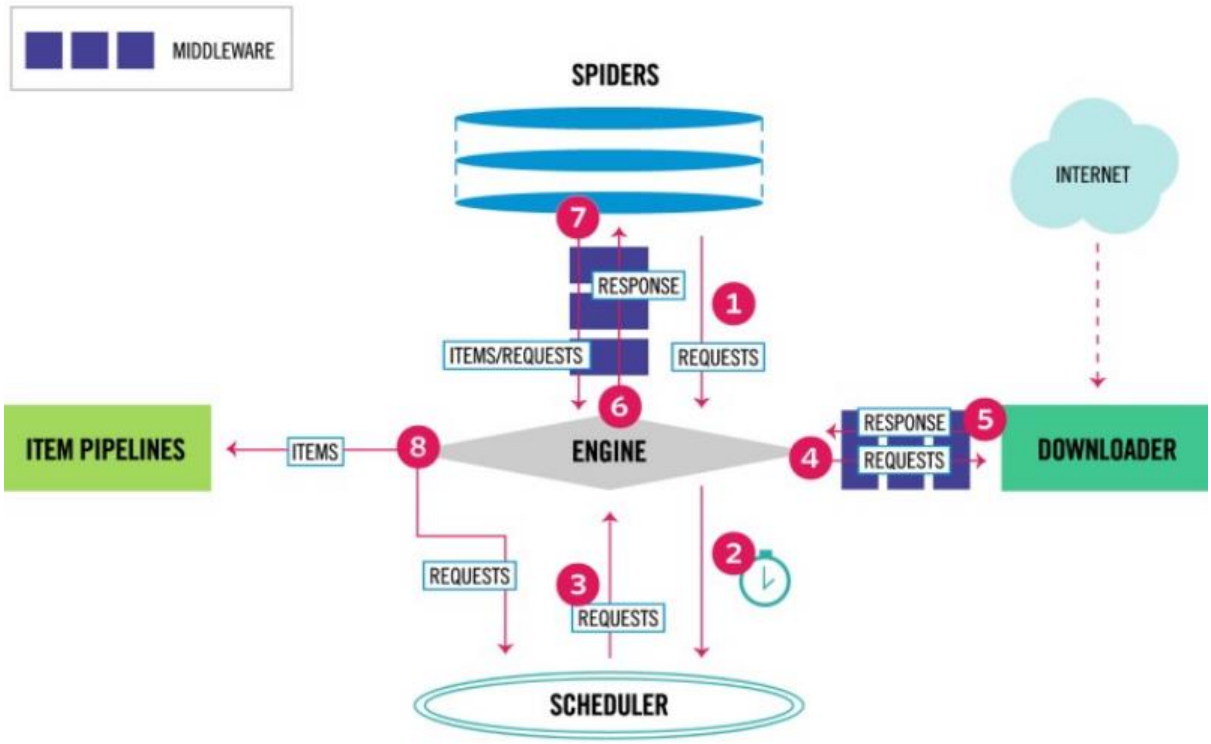

Figura 1: Scrapy Library architecture (Myers y McGuffee, 2015)

2.5 *TF=IDF*

It is an abbreviation of the Reverse Document Frequency formula, whose purpose is to define the importance of a keyword or phrase in a document (Dadgar, Araghi, and Farahani, 2016). In order to process the documents or texts, it is necessary to make a mathematical representation of the text, for which we use this algorithm.

The frequency of a term in a document is simply the number of times the term appears in that document. The value is usually normalized to prevent large documents from acquiring an unusual advantage. In this way, the importance of the term ti in the document dj is given by

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \qquad (1)$$

Where *n, j* is the number of occurrences of term considered in the document *dj* and the denominator is the number of occurrences of all the terms in the document *dj*. the inverse frequency of the document that is calculated with:

$$IDF_i = log\frac{|D|}{|\{d_j : t_i \in d_j\}|} \qquad (2)$$

Where we define the numerator as the total number of documents, and the denominator where the number of documents where the term *ti* appears (i.e., *nij/0*), and is where the calculation of TF-IDF for the term t i in the document *dj*

$$TF - IDF_{i,j} = TF_{i,j} x IDF_i \qquad (3)$$

2.6 *Multilayer Perceptron*

The multilayer perceptron (MLP) is one of the simplest neural networks, and it is based on a neural network that is the perceptron, and MLP architecture is characterized because it has its neurons grouped in layers of different levels and each one of the layers can be formed formed by a set of neurons and three different types of layers are distinguished which are: the input layer, the hidden layers and the output layer (Bishop, 2006). In the figure 1.6 an MLP model is shown, where the connections of the always are directed forward, that is, the neurons of one layer are connected with the neurons of the next layer, which means that it is a unidirectional network feedforward (Zhang, Towsey, Xie, Zhang, and Roe, 2016).

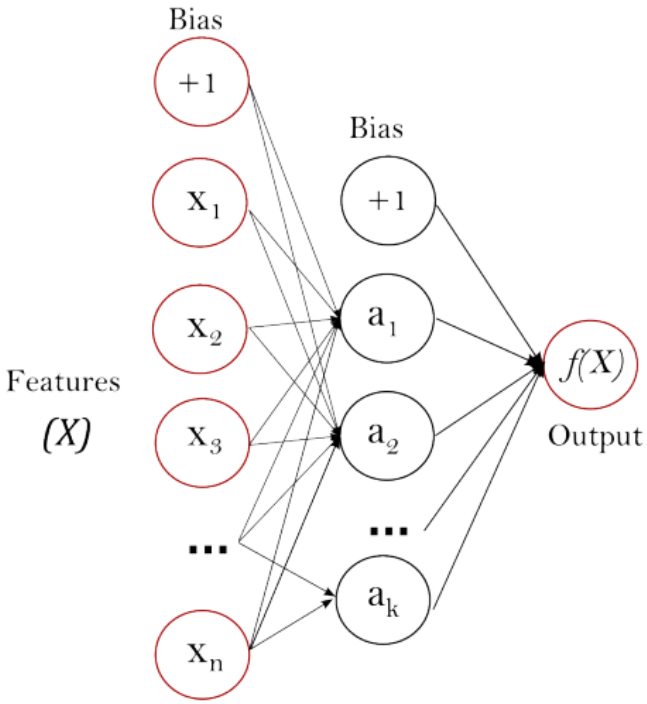

Figura 2:Architecture of a multilayer perceptron (Buitinck and cols., 2013)

The MLP is a supervised learning algorithm that learns a function to train a dataset, given a set of characteristics X=x1, x2, ..., xm where m is the number of dimensions of the input and these characteristics come to form the set of neurons {xi |x1, x2, ..., xm}, Where each neuron in the hidden layer is the value of the weight of each neuron by the value of the characteristic that is: w1x1 + w2x2 + ... + wmxm The output nodes are those that indicate if the new instance belongs to a class or not, according to the activation function these will take values from 0 to 1, or from −1 to +1.

3. Methodology

The present research project proposes, to create a search platform for open educational resources, using a multi-label classifier as a Machine Learning technique, which is a multilayer perceptron, for which the following architecture of the proposal is proposed:

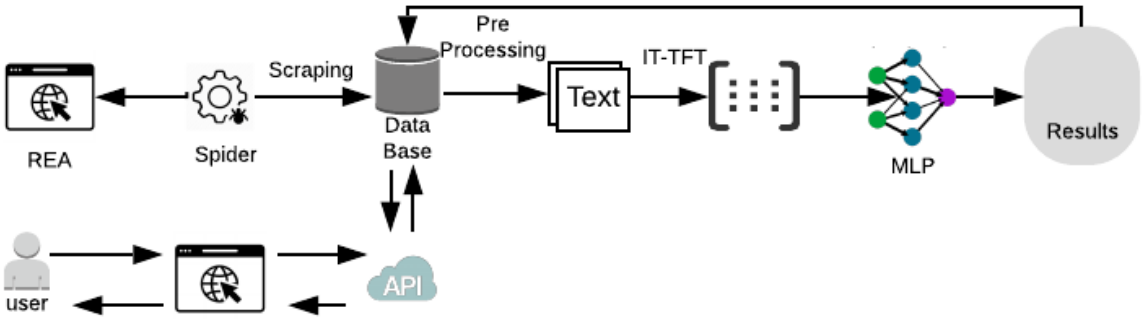

Figura 3: Architeture of model proposed (Buitinck y cols., 2013)

3.1 *Data Extraction*

In this stage of our proposal, scraping techniques are used to extract data from REA Repositories, such as: OERCOMMONS, Temoa, Procomun, for this we use the Scrapy library, which will allow us to create spiders, to extract the data more important to create a database with these resources, these are the fields extracted from these repositories: Title, Author, Description, Link, Subject, Material Type and also generate a field to perform the searches that we call “Labels"; field will serve as index of search of the contents.

A script (spider) was generated that helps to automate the searches on each page and get the information for each field that is being required. The script (spider) are divided into two parts:

The first part will analyze and search the links on the main search page of the repository, where each link that leads to each educational object to be analyzed will be obtained, advancing by each page until reaching the last one, for this part a function called Parse is generated that will help with the automated search of all links.

Second, after obtaining the link of each educational object, the spider enters each link to obtain the information of the fields that were previously selected. Another function called parse_mongo is generated which receives the link sent in the previous function called parse, through request and response objects, and the necessary data is obtained.

After applying the scrapping techniques to obtain the data, but none of the learning objects have a label assigned and in order to process them, a certain amount of them need to have labels assigned to be able to train the MLP classification algorithm, and for this, so different people experts in the field will perform this assignment manually.

For the assignment were considered generic verbs that fit several taxonomies including SOLO taxonomy, in addition to these verbs containing cognitive processes that have similarities between them, so that the classification will take the cognitive processes that will help to understand to the teacher at the time of preparing his OER(Villalba, Cristina, Bender & Cuba, 2019). While the verbs will be used on the same platform at the time of the searches.

There are generic verbs in the right side that have numbers assigned under them, those numbers represent Cognitive Processes, in total there are twelve verbs and three or four Cognitive Processes assigned to each one. The relation for verb/Cognitive Processes is describe in Table 3.

Table 3
Generic Verbs and Cognitive Processes (Villalba, Cristina, Bender & Castro, 2019)

Verb	**#**	**Cognitive Processes**
Identify o Recognize	1	Reception of information
	2	Characterization
	3	Recognition
Discriminate	1	Reception of information

	4	manifestation of the differences
	5	identification and verification of characteristics
Compare	1	Reception of information
	6	Identification of individual characteristics
	7	Verification of characteristics of two or more objects of study
Select	8	Determination of criteria or specifications
	9	Information search
	10	Identification and verification of criteria or specifications with prototypes
	11	Choice
Organize	1	Reception of information
	12	Identification of the elements that will be organized
	13	Determination of criteria to organize
	14	Disposition of the elements considering established criteria and order
Analyze	1	Reception of information
	15	Selective observation
	16	Division of the whole into its parts
	17	Interrelation of the parties to explain or justify
Infer	1	Reception of information
	18	Premise identification
	19	Contrasting the premises in context
	20	Formulation of deductions
Judge or Prosecute	1	Reception of information
	21	Criteria Formulation
	22	Contrast criteria with the referent
	23	Issuance of opinion or judgment
Represent or Diagram or Schematize or Design or Graph or Draw	24	Observation of the object or situation to be represented
	25	Description of the form / situation and location of its elements
	26	Order generation and representation sequencing
	27	Representation of the external and internal form or situation
Argue	1	Reception of information
	28	Selective observation of the information that will support
	29	Presentation of the arguments
Apply or Use	1	Reception of information
	30	Identification and sequencing of the principle, process or concept to be applied
	31	Execution of processes and strategies
Formulate or Pose	1	Reception of information
	32	Item Identification
	33	Interrelation of the elements
	34	Presentation of the interrelations

3.2 *Pre processing*

With pre processing, we seek to improve the quality of the data set, through techniques or a series of steps that help to obtain more and better information, achieving quality data that will help us to improve the analysis of this data.

As part of the pre processing, the texts must be represented in a way that is maintainable for our MLP, A large amount of data is stored in different languages, containing strange characters, text that is not useful, etc. So you have to apply different techniques that help us improve data and make sense. For this stage you have the following techniques that will be applied for pre processing: Data cleaning and TF-IDF.

3.2.1 *Data Cleaning*

Data Cleaning will help us to ensure the quality of the data. In this step, all the data will be placed in lowercase, so that there is no duplication of data at the time of applying TF-IDF or in the classification algorithms, then the removal of different tags, special characters and digits will also be done.

Also the elimination of words that in the text are very repetitive or words that do not give a special meaning to the text (eg, from, after, before, the, them, etc) that are called stopwords, they will be removed from the data in order to have real data that has value and also reduce the amount of data for better processing.

3.2.2 *TF-IDF*

Through the python scikit-learn library, the TF-IDF algorithm will be applied through the TfidfVectorizer class, to build a frequency matrix, in addition to replacing the data that are “null" or “empty" obtained by the scraping technique for 0, which at the end of this processing will give us a matrix of *nxm*, where n will be the number of words for each description and m the number of descriptions of the OER.

Table 4
TF-IDF Words Weights

	Term	Weight
8029	science	0.039497
7428	reading	0.038644
5241	life	0.038016
8748	students	0.026403
110	activity	0.023320

After pre processing all the data a final matrix is composed by the set of documents as an index, the vocabulary created from the documents that are stored in the variable *cv* and is obtained in the form *cv.get_feature_names()* that will be the columns of the matrix and as content is will have the weights obtained with the tf-idf algorithm.

3.3 *Multi Label Classification*

For do out the multi-label classification, it receives the output matrix of the TF-IDF as input and, for the generation of the labels, we will use the recurrent actions defined in the table 3 which will allow us to carry out the multi-label classification, for which due to the large amount of data is that we have proceeded to divide the dataset, 40% for training our neural network, and 60% to perform the new classifications of the proposed model, then it is necessary to normalize the data so that they have zero mean and standard deviation one, After Normalizing, for the architecture of our neural network we use the following structure, an input layer with the quantity of neurons as words in the vocabulary, two hidden layer with 100 neurons, and many output layers as labels quantity. Once the multi-label

classification has been carried out as a result, we will generate a table of the associated labels for each REA, and these labels will be stored in the database in order to perform the corresponding search.

3.4 *Search Based on Purpose Learning*

Once we have implemented our proposal, we proceed to create an interface, through a web page to search for educational content, based on learning purpose, for which we receive the terms to search, and the categories based on the generic verbs in Table 3, so that it works is that through a REST API service, it is that a request is sent with the term to search and the categories, and as a result it returns a JSON with all the resources that contain those categories, based on our field “Label" that are the Cognitive processes in our database.

4. Result

As part of our results, we show the implemented platform, using the scraping techniques of three OER repositories, and storing in a database, in addition to obtaining the accuracy measure, with respect to training data and new data. To be classified by our MLP network, it is worth mentioning that without performing the StopWords, our frequency matrix increases its size excessively, for which dimensionality reduction techniques can be used, but one disadvantage is that we would lose the nature of the words already We are looking for verbs to make the classification of the OER. As a case study we have used OER, based on computational thinking and mathematics, because if we do not delimit the research, it would be very general what we would have to validate for all the cases that can be found in these repositories, so we will focus only on computational thinking as an alternative to search for specialized content in this area.

In the table 5, we show the Precision-Recall, that is a useful measure of success of prediction when the classes are very imbalanced, in information retrieval, precision is a measure of result relevancy, while recall is a measure of how many truly relevant results are returned, the precision-recall, we show only 6 labels, and we use 1759 elements as prediction data,

Table 5
Precision-recall of classification multilabel

	Precision	Recall	f1-score	support
0	0.56	0.51	0.53	768
1	0.41	0.24	0.30	87
2	0.57	0.48	0.52	414
3	0.17	0.10	0.12	21
4	0.52	0.40	0.45	400
5	0.35	0.16	0.22	69
Micro avg	0.54	0.44	0.49	1759
Macro avg	0.43	0.31	0.36	1759
Weighted avg	0.53	0.44	0.48	1759
Samples avg	0.04	0.04	0.04	1759

And as a final result, there is a search platform based on learning purpose with different labels to select and to enter a text to search, then shows results with the algorithms applied.

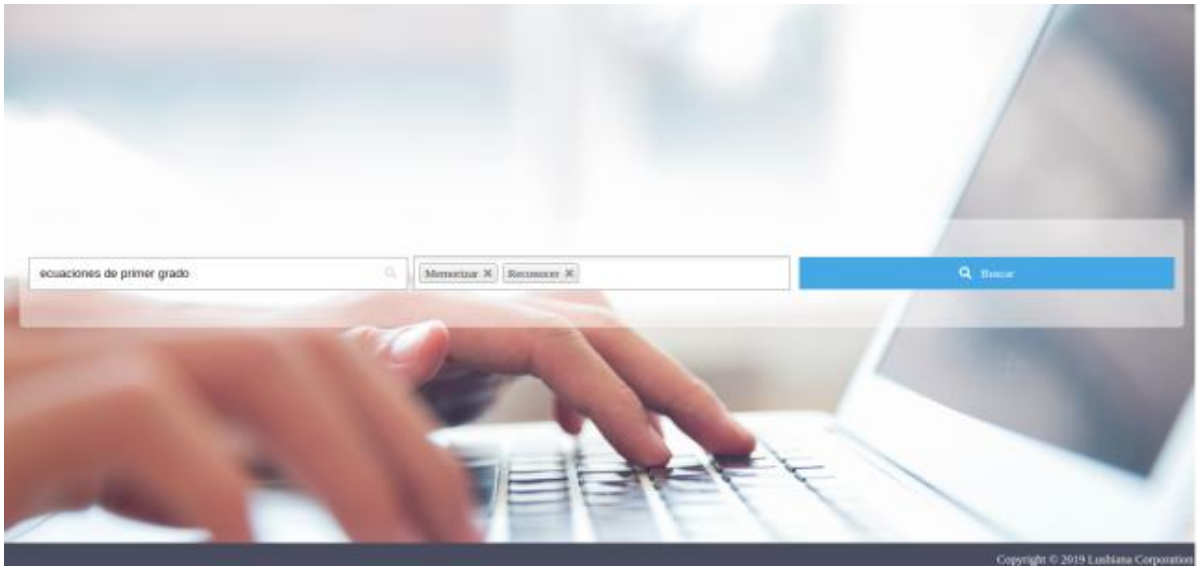

Figure 15: Screen of the proposal of the search platform based on learning purpose[4]

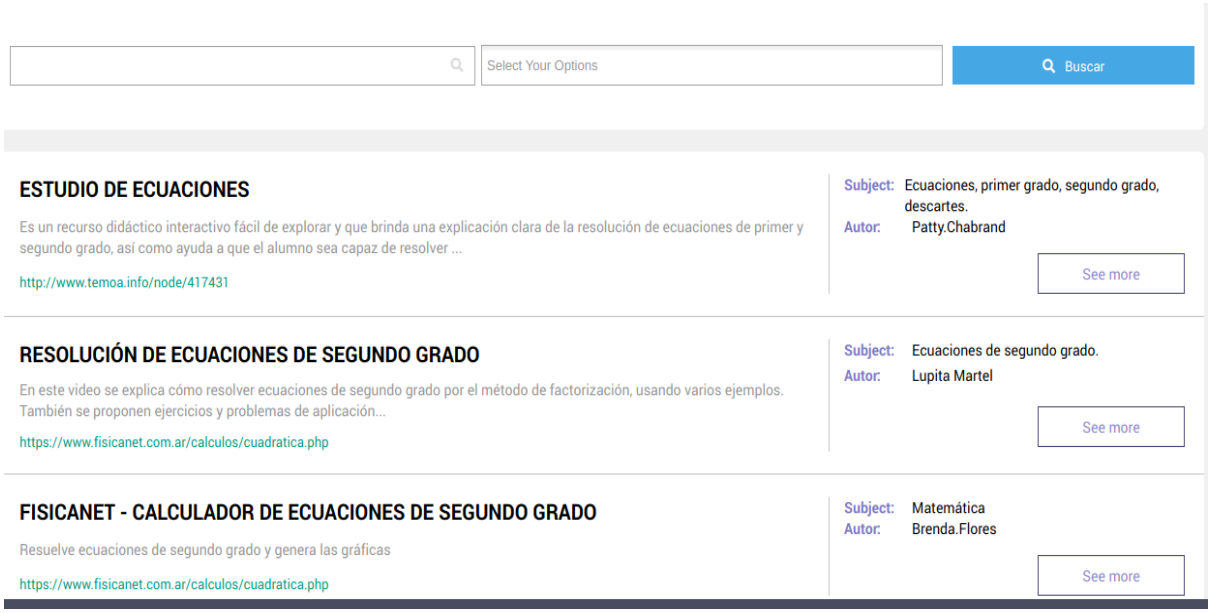

Figure 16: Screen of the proposal of the search platform based on learning purpose, results [4]

5. Conclusion

In recent years the development of specialized search systems has grown significantly, being very useful in different fields and particularly in education. In this work, an architecture of a content search system has been proposed under a SOLO taxonomy approach, which is focused on using the verb proposed as a learning purpose. This verb displays internally cognitive processes or learning processes that will serve as classification labels and allows more specialized searches of OER. In addition, a complete model is shown from the extraction of different repositories, until the conformation of a repository fed by external sources. In this way, the search system can be articulated by learning purpose, taking into account the teacher's intention of teaching and the need of the student focused on cognitive processes.

References

Aguila, J. V. B., y Burgos, V. (2010). Distribución de conocimiento y acceso libre a la información con recursos educativos abiertos (rea) . La educación.

Arias, D. P. H., Zermeño, M. G. G., y Chávez, M. M. P. (2015). Uso de recursos educativos abiertos en ambientes virtuales de aprendizaje para una educación inclusiva y de calidad. Revista de Investigación Educativa del Tecnológico de Monterrey, 6(11), 29–35.

Bishop, C. M. (2006). Pattern recognition and machine learning (1st ed. 2006. Corr. 2 nd printing ed.). Springer. Descargado de http://gen.lib.rus.ec/book/index.php?md5=6B552B24CAE380BB656F7AAEF7F81B46

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., . . . Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project. En Ecml pkdd workshop: Languages for data mining and machine learning (pp. 108–122).

Dadgar, S. M. H., Araghi, M. S., y Farahani, M. M. (2016, March). A novel text mining approach based on tf-idf and support vector machine for news classification. En 2016 ieee international conference on engineering and technology (icetech) (p. 112-116). doi: 10.1109/ICETECH.2016.7569223

Espinoza-Suarez, S., Falen, J. M. M., Chipana, S. S., Villalba-Condori, K. O., & Castro-Cuba-Sayco, S (2019) . Project-oriented learning for the achievement of educational outcomes . Volume 46, Pages 639-648

Kouzis-Loukas, D. (2016). Learning scrapy.

Packt Publishing Ltd. Levicoy, D. D., Obreque, A. S., Vásquez, C., y Salvatierra, M. O. (2017). Organización de las respuestas sobre tablas estadísticas por futuras maestras de educación infantil desde la taxonomía solo. Didasc@ lia: Didáctica y Educación , 8 (2), 193–212.

Liñán, L. C., y Pérez, Á. A. J. (2015). Educational data mining and learning analytics: differences, similarities, and time evolution. International Journal of Educational Technology in Higher Education, 12 (3), 98–112.

Myers, D., y McGuffee, J. W. (2015). Choosing scrapy. Journal of Computing Sciences in Colleges, 31 (1), 83-89. Ristoski, P., Paulheim, H.

Smedt, T. D., y Daelemans, W. (2012). Pattern for python. Journal of Machine Learning Research, 13 (Jun), 2063–2067.

Vargiu, E., y Urru, M. (2013). Exploiting web scraping in a collaborative filtering-based approach to web advertising. Artif. Intell. Research, 2 (1), 44–54.

Villalba, K., Cuba, S. C., Deco, C., Bender, C., y García-Peñalvo, F. J. (2017, Oct). A recommender system of open educational resources based on the purpose of learning. En 2017 twelfth latin american conference on learning technologies (laclo) (p. 1-4). doi: 10.1109/LACLO.2017.8120899

Wang, J., y Guo, Y. (2012). Scrapy-based crawling and user-behavior characteristics analysis on taobao. En 2012 international conference on cyber-enabled distributed computing and knowledge discovery (pp. 44–52).

Zhang, L., Towsey, M., Xie, J., Zhang, J., y Roe, P. (2016). Using multi-label classification for acoustic pattern detection and assisting bird species surveys. Applied Acoustics, 110, 91–98.

Villalba, K., Cristina, C., Bender, C., Castro, E. (2019). A Methodology to assign Educational Resources with Metadata based on the Purpose of Learning.

Villalba-Condori K. O. et al., "Learning Objects to Strengthen Learning. Experience in Regular Basic Education in Perú," 2018 XIII Latin American Conference on Learning Technologies (LACLO), São Paulo, Brazil, 2018, pp. 499-504. doi: 10.1109/LACLO.2018.00088

TUTORIALS