COLLABORATIVE WEB-BASED TAGGER FOR NAMED ENTITIES IN THE TASK OF INFORMATION EXTRACTION

ETIQUETADOR COLABORATIVO BASADO EN WEB PARA ENTIDADES NOMBRADAS EN LA TAREA DE EXTRACCIÓN DE LA INFORMACIÓN

David Efraín Muñoz Morales
Institute of Technology Tallaght
David.Efrain@postgrad.ittdublin.ie

Fernando Pérez Téllez
Institute of Technology Tallaght
Fernando.PerezTellez@it-tallaght.ie

David Eduardo Pinto Avendaño
Benemérita Universidad Autónoma de Puebla
dpinto@cs.buap.mx

Abstract

Nowadays, there exists a huge amount of information on the World Wide Web and since every day is mainly generated a lot of text data, the problem of information overload arise. In this way, the task of extracting meaningful information from text has gained the significant attention of researchers. In this paper, we propose a collaborative tagging system to help users in the task of highlighting important information in plain text files. Additionally, it allows converting tagged texts into a structured format. The web-based system is proposed in order to exploit the relevant content information provided by tagger users, since actual collaborative tagging systems suffer from issues such as tag scarcity or ambiguous labeling. Approaches such as the proposed here can facilitate to obtain better quality in tags and in any domain, allowing to achieve significant improvements in information extraction through named entities extraction, avoiding the noise of information overload.

Keywords: Collaborative tagging, information extraction, information overload, named entities, web-based tagger.

Resumen

Hoy en día existe una gran cantidad de información en Internet y ya que cada día se genera mucha información principalmente en forma de texto, el problema de sobrecarga de información se hace presente. En este sentido, la tarea de extraer información significativa de los textos ha ganado la atención de investigadores. En este artículo, proponemos un sistema de etiquetamiento colaborativo para ayudar a los usuarios en la tarea de resaltar información importante en archivos de texto plano. Adicionalmente, el sistema permite convertir textos etiquetados a un formato estructurado. El sistema basado en web es propuesto con el fin de explotar el contenido relevante de la información proporcionada por los usuarios etiquetadores, ya que los sistemas de etiquetamiento colaborativos actuales sufren de algunos problemas tales como la escasez de etiquetas o el etiquetado ambiguo. Enfoques como el propuesto aquí pueden facilitar la obtención de etiquetas con mejor calidad y en cualquier dominio, permitiendo lograr mejoras significativas en la extracción de información a través de la extracción de entidades nombradas, evitando el ruido en la sobrecarga de información.

Palabras Claves: *Entidades nombradas, etiquetamiento colaborativo, etiquetador basado en web, extracción de la información, sobrecarga de información.*

1. Introduction

Most of the existing information in the world is presented in a textual form which is unstructured information. In this way, information is significantly harder for machines to understand because of the complexity of natural language. So, with a huge amount of textual data, the information overload problem comes resulting important the extraction of meaningful data over the noise. Here is when the task of extracting information takes importance since Information Extraction (IE) is responsible for allowing to build a very general representation of meaning from unrestricted text [Bird, 2009].

IE is considered as a limited form of full natural language understanding, where the information we are looking for is known beforehand. So, IE includes two fundamental tasks, namely, named entity recognition and relation extraction [Allahyari, 2017]. The effective identification of Named Entities (NE) represents an important aspect in Information Retrieval (IR) related tasks.

In the last years, the task of extracting meaningful data of text has gained the attention in researcher and industry fields. Collaborative tagging systems have emerged as a solution for avoiding noise on information content since every day a huge amount of information is generated, and the majority is textual data. Collaborative tagging consists in assigning labels to a set of information resources. After that, tags can be used for many purposes such as retrieval, browsing, and categorization [Bischoff, 2008].

There exists some works presenting alternatives as a solution in the problem of highlighting relevant information over the information overload, one of them are some actual works in collaborative tagging systems, however, they have some issues in quality of tags such as scarcity or ambiguous labeling. As a result, the quality of recommendations is far to be excellent.

Previously, researchers proposed several wrapper inductions approaches for the rapid generation of extractors for Web pages, where the wrapper induction programs provide users with a GUI to click and highlight strings on a rendered Web page to produce a training example. This action of clicking and highlighting is referred to as “labeling” [Chang, 2003].

On the other hand, Halpin et al claimed that there are three main entities in any tagging system: users, items, and tags. They produced a generative model of collaborative tagging in order to understand the basic dynamics behind tagging. So, they showed how tag co-occurrence networks for a sample domain of tags can be used to analyze the meaning of particular tags given their relationship to others tags [Halpin, 2012]. In this way, [Nanopoulos, 2011] proposed to model data from collaborative tagging systems with three-mode tensors, in order to capture the three-way correlations between users, tags, and items. He said that by applying multiway analysis, latent correlations are revealed, which help to improve the quality recommendations. He also developed a hybrid scheme that additionally considers content-based information that is extracted from items.

In the task of universal semantic tagging, [Abzianidze, 2017] contributes to better semantic analysis for wide-coverage multilingual text. The authors said that, besides their application in semantic parsing demonstrated in the PMB project, sem-tags can contribute to other NLP tasks, e.g. POS tagging, or research lines rooted in compositional semantics. In their work, the authors have shown that the tags provide semantically fine-grained information, and they are suitable for cross-lingual semantic parsing.

In this way, getting better tags can improve the recommender systems performance, since the philosophy behind the success of recommendation technology is the fact that it is human tendency to rely on experiences of their neighbors and friends prior to making decision of any kind, especially regarding purchase of any items, taking admissions in institutes for higher education, opting an apartment for rent or buying it, spending weekend at some holiday places, etc [Saquib, 2017]. Font et al, proposed a general scheme for building a folksonomy-based tag recommendation system to help users tagging online content resources. They achieved this by using 3 independent steps: 1) Getting candidate tags, selecting a number of candidate tags for every input tag based on a tag-tag similarity matrix derived from a folksonomy, then 2) Aggregating candidate tags, assigning scores to the candidates of step 1 and merging them all in a single list of candidates tags, and finally 3) Selecting which tags to recommend, automatically selecting the candidates that will be part of the final recommendation by determining a threshold and filtering out those candidates whose score is below the threshold [Font, 2013]. In the task of recommending items, the classification and prediction have an important role by analyzing data. It is important to know what classification algorithm to use depending on the application to be developed. As noted by Sheshasaayee & Thailambal, Classification is in supervised Learning of Machine Learning where a set of correctly predicted observation is available [Sheshasaayee, 2017].

Chavaltada et al, proposed a framework for automatic product categorization and explained that each classification method is affected in the efficiency of the model since each method have different the parameters [Chavaltada, 2017].

The rest of this paper is organized as follows. First, we show the methods used and methodology followed. After we present a use of the Web-based Tagger developed, followed by the discussions. Finally, we present the conclusions of this paper.

2. Methods

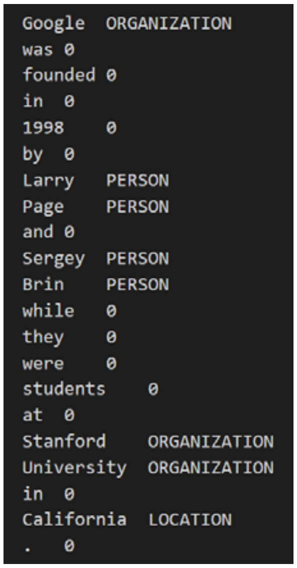
Since most of the textual data exist in an unstructured form it is important to produce structured data ready for post-processing, which is crucial to many applications of text mining such as text categorization, entity extraction, learning relations between named entities, etc. That is why the proposed web-based tool aims to help users in the task of highlighting the relevant information in plain text files and then by producing a structured version of the data ready to be used as corpora in the training of some models such as NERclassifiers.

There are so many reasons to have the tool tagger based on the web because it allows access to information at any device with internet connection. Also, it facilitates a huge number of users who can tag different text files simultaneously, increasing the number of tagged texts to be used as training data.

Tokenization scheme

In order to prepare the training data to be used in text mining applications, a fundamental step is tokenization of the text. Tokenization is the task of breaking up a string into identifiable linguistic units that constitute a piece of language data (words and punctuation) [Bird, 2009].

The tagged data is formatted by the web-based tool into a predefined format by the Sanford NLP Group, where data needs to be in tab-separated columns, with word tokens in one column and the class labels in another column [NLP, 2018]. An original plain text file is converted into a tab-separated columns format by tokenizing the continuous text into one column and then in a second column by assigning the class labeled for each word, if a token has not been labeled in one of the classes as a named entity, then the value in the second column for that token should be 0. A short illustration of the format is shown in figure 1, where the text has been labeled in 3 classes (Organization, Person and Location).



```
Google  ORGANIZATION
was 0
founded 0
in  0
1998    0
by  0
Larry   PERSON
Page    PERSON
and 0
Sergey  PERSON
Brin    PERSON
while   0
they    0
were    0
students    0
at  0
Stanford    ORGANIZATION
University  ORGANIZATION
in  0
California  LOCATION
.   0
```

Figure 1 Stanford highlighted format (tab-separated columns).

Learning algorithm

The need to segment and label sequences arises in many different problems in several scientific fields. In the field of computer sciences, some generative models such as Hidden Markov Models (HMMs) and stochastic models have been applied in a wide variety of problems in text and speech processing. However, these generative models are not the most optimal in the task of labeling data.

Conditional random fields

Conditional random fields (CRFs) are a probabilistic framework for labeling and segmenting structured data, such as sequences (that is why the tokenization scheme presented in this work is important since the tokenized plain text file by the web-based tool is a continuous text). The CRF model presents some advantages over the HMMs, and the main advantage is their conditional nature, resulting in the relaxation of the independence assumptions required by HMMs in order to ensure tractable inference. CRFs avoid the label bias problem, a weakness exhibited by maximum entropy Markov models (MEMMs) [Lafferty, 2001].

The mathematics behind this model are defined as follows, considering to X as a random variable over data sequences to be labeled, and Y is a random variable over corresponding label sequences. All components Y_i of Y are assumed to range over a finite alphabet λ. For example, X might range over natural language sentences and Y range over part-of-speech tagging of those sentences, with λ the set of possible part-of-speech tags [Lafferty, 2001].

Definition: Let $G = (V, E)$ be a graph such that $Y = (Yv)v \in v$, so that Y is indexed by the vertices of G. Then (X, Y) is a conditional random field in case, when conditioned on X, the random variables Y_v obey the Markov property with respect to the graph*:* $p(Yv|X, Yw, w \neq v) = p(Yv \mid X, Yw, w \sim v)$, where $w \sim v$ means that *w* and *v* are neighbors in *G*.

So, a CRF is a random field globally conditioned on the observation *X*.

Proposed system

As there are three main entities in any tagging system: users, items, and tags [Halpin, 2012], the proposed work takes into consideration the affirmation. On the other hand, since the main task of tagging is to get identified concepts for users, and in any kind of domains, so, the web-based tool needs to allow access to many users and from any place they are. Also, users should download the tagged files when they need it. On the other hand, it needs to facilitate the creation of new labels, since there exist a lot of domains in which users can tag concepts. Following these statements, the system architecture is illustrated in figure 2.

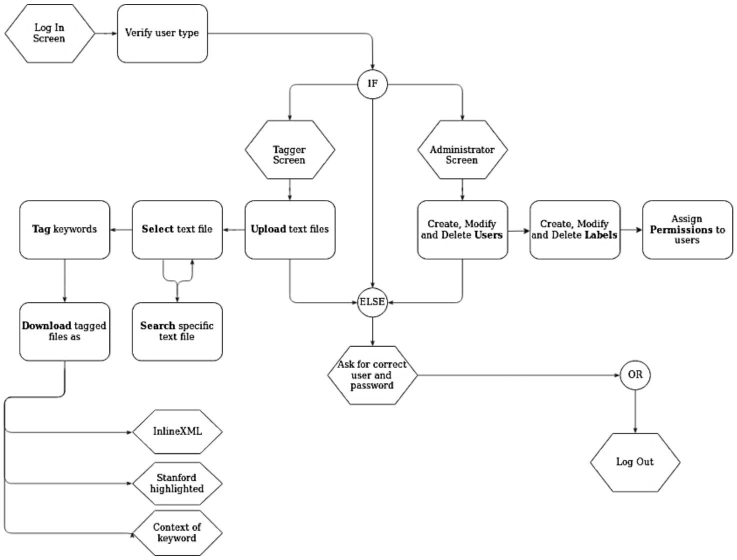


Figure 2 System architecture.

Figure 2 shows the flow of the system, where users first need to authenticate and depending on the type of user they are, is the flow they will follow.

There are two types of users: Administrator and Tagger. If the authenticated user is an administrator, then the user is able to create, modify and delete as many users and labels as necessary; and also he can assign permissions of which labels can use every user depending on the domain they will be working.

If the user is authenticated as a tagger, then the user can upload plain text files in his account, as well as search and open a specific plain text file to start tagging the keywords in that document. After that, the user is able to download the tagged plain text files in 3 different formats: InlineXML, Stanford highlighted (tab-separated columns) and a special format where the file includes the words preceding and proceeding the keywords tagged, so it can help to understand the context in which the keywords are taking place. In the next sections, we describe a more detailed explanation related to each part of the system.

User authentication

As there is a huge number of people tagging text files, it is necessary to authenticate the user session to know what plain text files and classes are allowed to use for a specific user. On the other hand, it facilitates access from any device, it means, once a text file was uploaded, the user can use it at any time, in any place and any device.

Tagger user

Taggers are main generators of high quality tagged data, since they have a good understanding in finding key concepts in textual data, differentiating from one and another context.

Once they are logged in the system, they are allowed to: upload, search and select specific plain text files, tag keywords or concepts in the text and then download tagged files. Alternatively, tagged files are available to download in three formats:

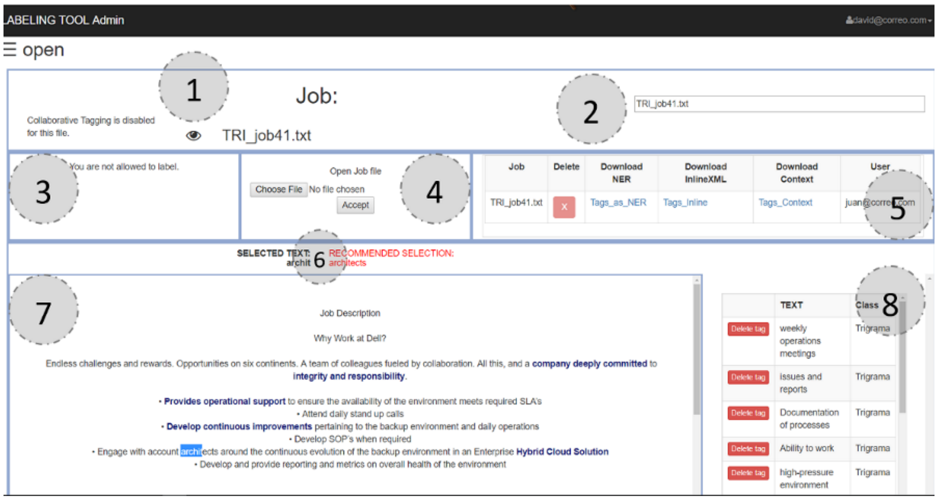
- InlineXML format. It is a conversion of the original text into an XML representation where the output file contains the complete original text, but where each previously tagged concept appears between tags (similar to the HTML style) where the tag is the same that was assigned by the tagger user.

- Stanford format (tab-separated columns). The original text is converted into a structured format established by the NLP Group, where the first column contains the tokens of the plain text file and the second column contains the belonging to each token. The first thing that the web-based tool makes is tokenizing the whole text into words and punctuation. After each token is stored in the first column of the output file, in this way if the token is part of a tagged concept then the second column will store the class assigned to that concept, in another way the token will have the value of 0 as the class in the second column.
- The context of Keyword. This format is similar to the tab-separated columns format, however, only the tagged concepts are exported. In this format, the first column contains the class and the second column contains the tagged concept as well as the words that proceed and precede to that concept. It is because the meaning of the concept can be different depending on the context in which it is identified.

With this approach is possible to obtain more consistent tagged texts, avoiding some problems as tag scarcity, the use of multiple labels to refer to a single concept, and the ambiguity in the meaning of certain labels. In figure 3 is displayed the interface built for the tagger user, and it is composed of eight sections. The first section (denoted by the circle with the number 1 inside) indicates if the selected plain text file is available to be tagged by other users since every owner decides if he allows each one of his files to be tagged collaboratively. If the owner does not allow his files to be tagged collaboratively, then each of his files will only be visible in reading mode to other users, however, any user can download the generated formats of any file, because it can be used as training data. Section two contains a search field in order to help the users in finding a specific plain text file.

Section three contains the classes (as buttons) allowed to each user in the task of tagging plain text files, however, if the user selects a plain text file that belongs to other user and it is not enabled to be tagged collaboratively, then the user will not have active buttons to tag this plain text file.

Section four allows the users to upload plain text files in order to be tagged. Every tagger user can upload files as needed. Section five is one of the most important sections since it offers the possibility to download the formatted plain text files in each one of the three available formats. This section also displays the information of the current plain text file selected, such as the name, the owner and the possibility to delete the file in case that the owner is the current user. Section six have been integrated in order to provide help to each user at the time of highlighting the most important concepts in the text. When a user selects one keyword or concept in the text, he can omit part of the complete real concept so in that case, the web-tool offers an alternative to the highlighted text, such as in figure 3, where the user selected *arch'* and the web-tool offers the correct alternative for this case: *architects*.



Job:

Collaborative Tagging is disabled for this file.

TRI_job41.txt

TRI_job41.txt

You are not allowed to label.

Open Job file

No file chosen

Job	Delete	Download NER	Download InlineXML	Download Context	User
TRI_job41.txt	X	Tags_as_NER	Tags_Inline	Tags_Context	juan@correo.com

SELECTED TEXT: archit

RECOMMENDED SELECTION: architects

Job Description

Why Work at Dell?

Endless challenges and rewards. Opportunities on six continents. A team of colleagues fueled by collaboration. All this, and a **company deeply committed** to **integrity and responsibility**.

- **Provides operational support** to ensure the availability of the environment meets required SLA's
- Attend daily stand up calls
- **Develop continuous improvements** pertaining to the backup environment and daily operations
- Develop SOP's when required
- Engage with account architects around the continuous evolution of the backup environment in an Enterprise **Hybrid Cloud Solution**
- Develop and provide reporting and metrics on overall health of the environment

	TEXT	Class
Delete tag	weekly operations meetings	Trigrama
Delete tag	issues and reports	Trigrama
Delete tag	Documentation of processes	Trigrama
Delete tag	Ability to work	Trigrama
Delete tag	high-pressure environment	Trigrama

Figure 3 Tagger Interface.

Section seven is the most important of all since it displays the content of the selected plain text file to be tagged. Additionally, when the user has tagged a key concept then it is highlighted in the text with font bold and with the color assigned to the class with which the concept was tagged. This special feature helps to improve the reading of the text.

Finally, in section eight, a list with the tagged key concepts as well as the class assigned to each one are shown. Additionally, the user can delete one or more of the tagged concepts with the possibility of reassigning it to other class.

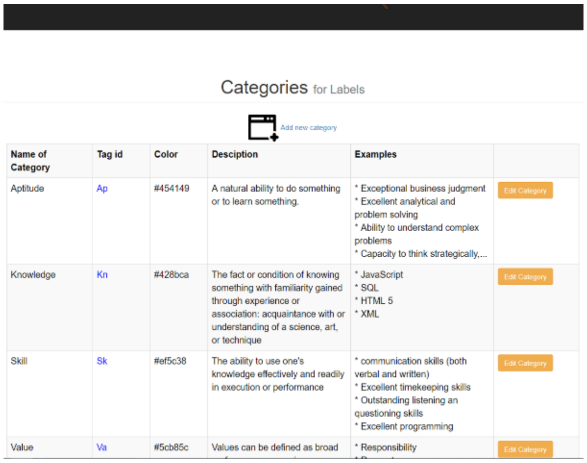
Administrator user

An administrator is allowed to create, modify and delete users, as well as administrate and assign classes (kind of labels) to each tagger. The administrator has an important role because he allows users to have access only to certain classes, and he can create as many as classes are required for users, allowing to get tagged texts with high quality in tags through human feedback. Additionally, the administrator is allowed to view the plain text files and is allowed to change the user tagger password if the tagger asks for that.

In figure 4 the administration panel for classes is shown. When a new class is created it needs five features to be considered:

- Name. It denotes the name of the class and will be used to format the output files in each one of the three available.
- Short identifier. It can be a little contraction to represent the class when the original name is too long.
- Color. The color is used to highlight the tagged concepts by the tagger user in the text (displayed in section seven). It facilitates the reading to the users.
- Description. A short description needs to be provided in order to guide users to understand the meaning of the class, since sometimes some classes can be similar, e.g. skill and aptitude.
- Examples. In order to facilitate help for taggers, some examples should be included.

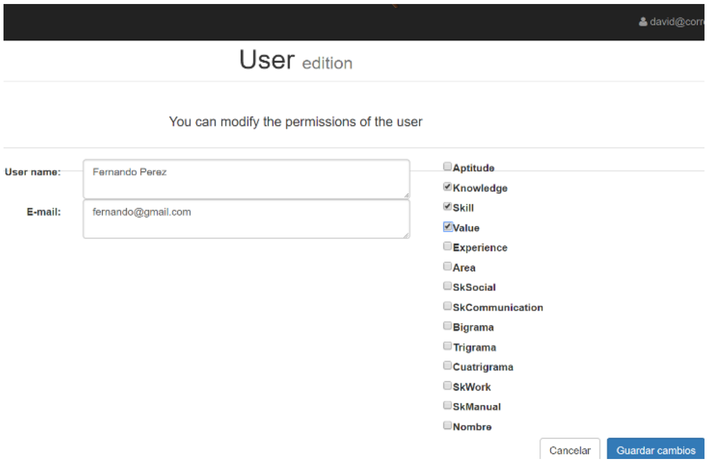
In figure 5 the interface corresponding to the administrator user is shown. In this screen, an administrator can assign permissions of classes to every tagger. This is because not all users will be working in the same domains, so they can request the creation and assignment of new classes to expand and improve the accuracy of the labeling for each text.



Categories for Labels

Name of Category	Tag id	Color	Desciption	Examples	
Aptitude	Ap	#454149	A natural ability to do something or to learn something.	* Exceptional business judgment * Excellent analytical and problem solving * Ability to understand complex problems * Capacity to think strategically,...	Edit Category
Knowledge	Kn	#428bca	The fact or condition of knowing something with familiarity gained through experience or association: acquaintance with or understanding of a science, art, or technique	* JavaScript * SQL * HTML 5 * XML	Edit Category
Skill	Sk	#ef5c38	The ability to use one's knowledge effectively and readily in execution or performance	* communication skills (both verbal and written) * Excellent timekeeping skills * Outstanding listening an questioning skills * Excellent programming	Edit Category
Value	Va	#5cb85c	Values can be defined as broad	* Responsibility	Edit Category

Figure 4 Screen to administrate the classes.



User edition

You can modify the permissions of the user

User name: Fernando Perez

E-mail: fernando@gmail.com

□ Aptitude
☑ Knowledge
☑ Skill
☑ Value
□ Experience
□ Area
□ SkSocial
□ SkCommunication
□ Bigrama
□ Trigrama
□ Cuatrigrama
□ SkWork
□ SkManual
□ Nombre

Figure 5 Panel to assign labeling permissions for users.

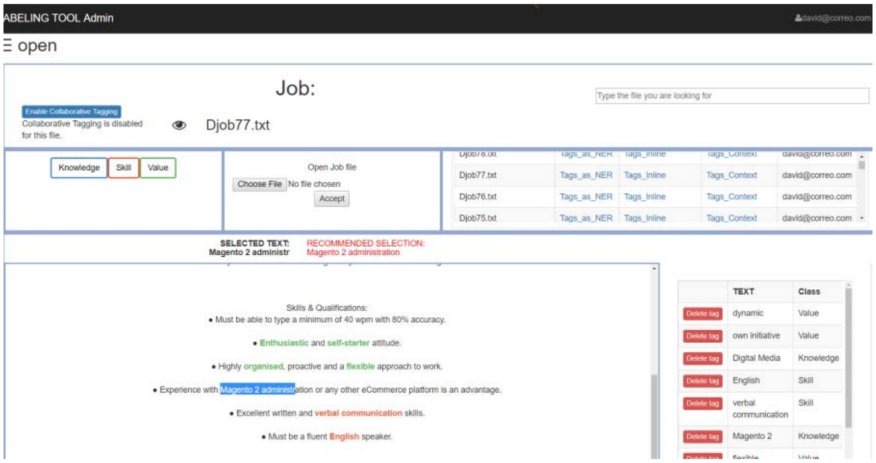
3. Results

This section shows the use of the system applied in an example scenario where the main task is to highlight the skills, values, and knowledge in some plain text files.

In order to tag the text, the procedure followed was (as illustrated in figure 6):

- First, the user david@correo.com logged into the system.
- Then, in section four he uploaded a plain text file called *Djob77.txt*.
- After, in section five he selected the file to be displayed.
- Once the content file was displayed in section seven, the user started to identify and highlight the relevant concepts in the text by assigning one of the

classes allowed. The user had recommendations for the selected text every time that the web-based tool identified the selection as incomplete. On the other hand, the user always had the possibility to delete some key concepts already tagged.



Djob77.txt	Tags_as_NER	Tags_Inline	Tags_Context	david@correo.com
Djob76.txt	Tags_as_NER	Tags_Inline	Tags_Context	david@correo.com
Djob75.txt	Tags_as_NER	Tags_Inline	Tags_Context	david@correo.com

SELECTED TEXT: Magento 2 administr
RECOMMENDED SELECTION: Magento 2 administration

Skills & Qualifications:
- Must be able to type a minimum of 40 wpm with 80% accuracy.
- Enthusiastic and self-starter attitude.
- Highly organised, proactive and a flexible approach to work.
- Experience with Magento 2 administration or any other eCommerce platform is an advantage.
- Excellent written and verbal communication skills.
- Must be a fluent English speaker.

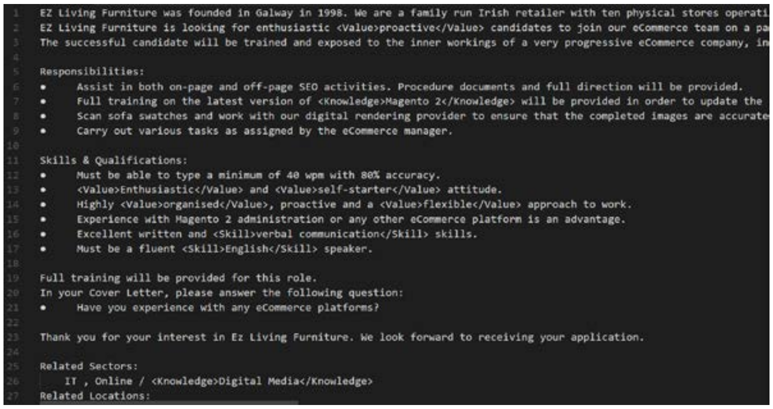
	TEXT	Class
Delete tag	dynamic	Value
Delete tag	own initiative	Value
Delete tag	Digital Media	Knowledge
Delete tag	English	Skill
Delete tag	verbal communication	Skill
Delete tag	Magento 2	Knowledge

Figure 6 The web-based tagger tool in use.

Since the goal of the web-based tagger tool is to convert the unstructured information in the texts into structured data, users are interested in obtaining the keywords and their labels to which each one belongs in a structured form. That is why the next step is to get the plain text files formatted in one of the three available formats by the web-based tool.

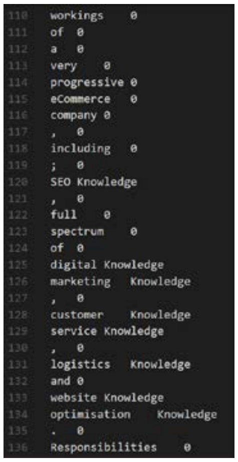
So, after tagger users identified the most important concepts of the text related to skill, knowledge, and value they can use the generated file (structured data) for identifying named entities in new text files. Named entity recognition [NLP, 2018] have the main task of label sequences of words in a text, which are the keywords of interest. So, the web-based tagger allows us to download the named entities in the InlineXML format as well as the Stanford format as illustrated in figures 7 and 8, respectively. And the special format which includes the words preceding and proceeding the tagged keywords is illustrated, in figure 9.

As it is possible to observe obtained results demonstrate the advantage of using the web-based tagger tool in order to convert unstructured information from texts into structured information. After, this structured data can be used in some tasks such as the training of some classifiers in the recognition of named entities, since the web-based tagger allow us to obtain the data in two formats from those offered by the Stanford named entity tagger [http://nlp.stanford.edu:8080/ner/, 2018].



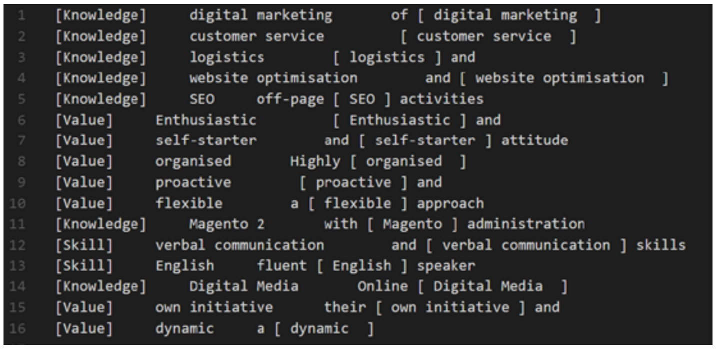
```
EZ Living Furniture was founded in Galway in 1998. We are a family run Irish retailer with ten physical stores operati
EZ Living Furniture is looking for enthusiastic <Value>proactive</Value> candidates to join our eCommerce team on a pa
The successful candidate will be trained and exposed to the inner workings of a very progressive eCommerce company, in

Responsibilities:
•     Assist in both on-page and off-page SEO activities. Procedure documents and full direction will be provided.
•     Full training on the latest version of <Knowledge>Magento 2</Knowledge> will be provided in order to update the
•     Scan sofa swatches and work with our digital rendering provider to ensure that the completed images are accurate
•     Carry out various tasks as assigned by the eCommerce manager.

Skills & Qualifications:
•     Must be able to type a minimum of 40 wpm with 80% accuracy.
•     <Value>Enthusiastic</Value> and <Value>self-starter</Value> attitude.
•     Highly <Value>organised</Value>, proactive and a <Value>flexible</Value> approach to work.
•     Experience with Magento 2 administration or any other eCommerce platform is an advantage.
•     Excellent written and <Skill>verbal communication</Skill> skills.
•     Must be a fluent <Skill>English</Skill> speaker.

Full training will be provided for this role.
In your Cover Letter, please answer the following question:
•     Have you experience with any eCommerce platforms?

Thank you for your interest in Ez Living Furniture. We look forward to receiving your application.

Related Sectors:
    IT , Online / <Knowledge>Digital Media</Knowledge>
Related Locations:
```

Figure 7 InlineXML format of named entities in web-based tagger tool.



```
workings    0
of  0
a   0
very    0
progressive 0
eCommerce   0
company 0
,   0
including   0
;   0
SEO Knowledge
,   0
full    0
spectrum    0
of  0
digital Knowledge
marketing   Knowledge
,   0
customer    Knowledge
service Knowledge
,   0
logistics   Knowledge
and 0
website Knowledge
optimisation    Knowledge
.   0
Responsibilities    0
```

Figure 8 Stanford format (tab-separated columns).



```
[Knowledge]     digital marketing       of [ digital marketing  ]
[Knowledge]     customer service        [ customer service  ]
[Knowledge]     logistics       [ logistics ] and
[Knowledge]     website optimisation        and [ website optimisation  ]
[Knowledge]     SEO     off-page [ SEO ] activities
[Value]     Enthusiastic        [ Enthusiastic ] and
[Value]     self-starter        and [ self-starter ] attitude
[Value]     organised       Highly [ organised  ]
[Value]     proactive       [ proactive ] and
[Value]     flexible        a [ flexible ] approach
[Knowledge]     Magento 2       with [ Magento ] administration
[Skill]     verbal communication        and [ verbal communication ] skills
[Skill]     English     fluent [ English ] speaker
[Knowledge]     Digital Media       Online [ Digital Media  ]
[Value]     own initiative      their [ own initiative ] and
[Value]     dynamic     a [ dynamic  ]
```

Figure 9 Context of keyword format.

4. Discussion

The problem of information overload and poor quality in tags affecting recommender systems has been examined. The Stanford named entity tagger has been studied since it represents a widely used tool by training models of labeled data. However, the low quality in tags affects to classification models in improving the recommendations. With the proposed web-based tagger is possible to improve the quality of tags in key concepts since these are labeled by people, it means, is achieved a human feedback, in this way better tags can help to improve recommender systems. On the other hand, as was mentioned in the section of tokenization scheme, the provided format (tab-separated columns) by the web-based tool tagger allows us to use the generated structured data in the training some classifier models in the task of named entities recognition.

5. Conclusion

The proposed system offers some advantages in the improving quality of tags and in a fast way, as follows:

- The web-based tagger makes the information accessible since any device, allowing us to create as many labels as necessary to work in any domain. Additionally, the administrator can create as many users as necessary.
- Since tags are provided by humans, the quality of tags is high, in this way recommender systems can improve their recommendations.
- In the task of information extraction, the web-based tagger allows downloading the representation of the original text into structured data, in 3 possible formats: InlineXML, Stanford format (tab-separated columns), and a special format to knowing the context of a keyword.

Future work considers new features, such as the recognition of named entities immediately upon uploading a new plain text file for labeling. On the other hand, future work involves the use of these named entities in job descriptions and resumes, where through the use of classification algorithms and a recommender system, the companies can find the best applicant for each job offer.

6. Bibliography and References

[1] Abzianidze, L. & Bos, J. Towards Universal Semantic Tagging. International Conference on Computational Semantics, 2017.

[2] Allahyari, M., Safaei, S., Pouriyeh, S., Trippe, E., Kochut, K., Assefi, M. & Gutierrez, J. A brief survey of text mining: classification, clustering and extraction techniques. KDD Bigdas, 2017.

[3] Bird, S., Klein E. & Loper, E. Natural Language Processing with Python. O'Reilly, 109-112, 261-285, 2009.

[4] Bischoff, K., Firan, C., Nejdl, W., & Paiu, R. Can all tags be used for search? in Proceedings of the 17th acm conference on information and knowledge management, 193-202, 2008.

[5] Chang, C., Kayed, M., Girgis, M. R. & Shaalan, K. F., A Survey of Web Information Extraction Systems, vol. 18, 1411-1428, 2006.

[6] Chang, C.-H., Hsu C.-N. & Lui S.-C., Automatic Information Extraction from Semi-Structured Web Pages by Pattern Discovery. Decision Support Systems J., vol. 35, NO. 1, pp. 129-147, 2003.

[7] Chavaltada, C., Pasupa, K., & Hardoon, D. A comparative study of machine learning techniques for automatic product categorization. Springer international publishing, 10-17, 2017.

[8] Font, F., Serrà, J. & Serra, X. Folksonomy-based tag recommendation for collaborative tagging systems. International Journal on Semantic Web and Information Systems, 1-27, 2013.

[9] Halpin, H., Robu, V. & Shepherd, H. The complex dynamics of collaborative tagging. WWW 2007, 211-220, 2012.

[10] Lafferty, J., McCallum, A. & C.N. Pereira, F., Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, in Proceedings of the 18th International Conference on Machine Learning, 282-289, 2001.

[11] NLP. Stanford Named Entity Recognizer. The Stanford natural language processing group. May, 2018: https://nlp.stanford.edu/software/CRF-NER.html.

[12] Nanopoulos, A. Item recommendation in collaborative tagging systems. IEEE transactions on systems, man, and cybernetics, NO. 4, 760-771, 2011.

[13] Saquib, S., Siddiqui, J. & Ali, R. Classifications of Recommender Systems: A review. Journal of engineering and technology review, 132-153, 2017.

[14] Sheshasaayee, A. & Thailambal, G. Comparison of Classification Algorithms in Text Mining. International journal of pure and applied mathematics, 425-433, 2017.