https://repository.uaeh.edu.mx/revistas/index.php/investigium/issue/archive

Boletín Científico INVESTIGIUM de la Escuela Superior de Tizayuca

InvESTigium

Publicación semestral No. 9 (2019) 6-11

Empleo de expresiones regulares como herramienta para la identificación y corrección ortográfica / gramatical en el idioma español

Use of regular expressions as a tool for identifying and correcting spelling / grammar in Spanish

Alonso E. Solís-Galindo [a], Evangelina Lezama-León [b]

Abstract:

The learning of a language requires the presence of someone who has better command of the language in question, so that it is the latter who can provide feedback to the language learner about the errors discussed. However, it is possible to use technology to make the language learning process efficient, mainly when there are digital communication tools that cause users to use different words, emoticons, images and / or abbreviations that would make it difficult to understand the language of any apprentice. These vices are committed more frequently by young people who are the main consumers of the media. That is why it is proposed that it is through Natural Language Processing tools, such as the Regular Expressions, that allow defining an alternative to perform the syntactic and grammatical analysis of messages sent through digital communication media with the objective of providing an alternative that allows the user to be informed that he is making a spelling / grammar error in the Spanish language.

Keywords:

Regular expressions, natural language processing, disortography, spelling / grammar of Spanish

Resumen:

El aprendizaje de un idioma requiere la presencia de alguien que tenga mejor dominio sobre el idioma en cuestión, de tal manera que sea esta última quien pueda retroalimentar al aprendiz del idioma sobre los errores se comentan. Sin embargo, es posible hacer uso de la tecnología para hacer eficiente el proceso de aprendizaje un idioma, principalmente cuando existen herramientas de comunicación digital que provocan que los usuarios empleen palabras distintas, emoticonos, imágenes y/o abreviaciones que dificultarían la comprensión del idioma de cualquier aprendiz. Dichos vicios son cometidos con mayor frecuencia por jóvenes quienes son los principales consumidores de los medios de comunicación. Es por ello que se propone que sea a través de herramientas del Procesamiento del Lenguaje Natural, como los son las Expresiones Regulares, las que permitan definir una alternativa para realizar el análisis sintáctico y gramatical de los mensajes enviados a través de medios de comunicación digital con el objetivo de brindar una alternativa que permita indicarle al usuario que está cometiendo algún error ortográfico / gramatical en el idioma español.

Palabras Clave:

Expresiones regulares, procesamiento del lenguaje natural, disortografía, ortografía / gramática del español.

Introducción

Un problema que se presenta de manera cotidiana en estudiantes universitarios es la mala ortografía con la escriben, además de los errores gramaticales que con frecuencia incurren. Tal situación se ve acentuada por el uso de redes sociales, además del empleo de herramientas de mensajería instantánea como los es el WhatsApp, o bien, por el empleo de redes sociales. Sin embargo, la comunicación actual de los jóvenes se basa en el envío y recepción de mensajes de texto a través de dispositivos de comunicación digital, convirtiéndose éstos en los principales medios de comunicación escrita en la actualidad [1].

Esta diversidad en la forma de escribir está basada en la situación comunicativa, es decir, la situación formal o informal en la que se encuentre la persona que desea

[a] Autor de Correspondencia, Universidad Autónoma del Estado de Hidalgo, Escuela Superior de Tizayuca, https://orcid.org/0000-0002-3999-006X, Email: soliser@uaeh.edu.mx

[b] Universidad Autónoma del Estado de Hidalgo, Escuela Superior de Tizayuca, https://orcid.org/0000-0003-0818-0897, Email: evangeli@uaeh.edu.mx

comunicarse. La comunicación informal se le ha denominado disortográfica y es la que principalmente se emplea en los medios de comunicación digital [1].

Estudios realizados por [1], demuestran que el empleo de los medios de comunicación digital para el envío de mensajes basados en texto, además de otros elementos como emoticonos, imágenes fijas, audios o videos no repercuten cuando un estudiante universitario desea comunicarse a través de una escritura formal. Se encuentra que la comunicación formal se hace de manera aceptable, cometiendo errores ortográficos de manera esporádica, que se consideran aceptables debido al descuido o al desconocimiento de cómo se escriben algunas palabras, pese a que la misma persona se comunica de forma disortográfica en medios de comunicación digital.

Sin embargo, ¿qué pasa cuando la comunicación escrita la desea realizar algún aprendiz del idioma español? Para ello será necesario hacer uso de alguna herramienta que le permita corregir el error una vez que se ha cometido, principalmente cuando en ese momento no hay alguien que conozca el idioma y pueda corregir al revisar lo escrito, o bien, si es a través de alguna herramienta de software correctora ortográfica. Para ello en el presente trabajo se propone que se haga uso las Expresiones Regulares (ER) como herramienta que pueda ser implementada como aplicación que facilite identificar el error y así notificar al usuario para que corrija y aprenda del error.

Desde hace ya algunos años se han estado aplicando las ER como mecanismo a través del cual se aplica la tecnología para contar con aplicaciones que permitan corregir algunos de los errores que cometen los aprendices de algún idioma, principalmente para el aprendizaje del idioma inglés.

La tarea anteriormente descrita puede ser demasiado trivial para el cerebro humano, ya que es fácilmente, para el ser humano, distinguir una letra de un número, o bien, el conjunto de palabras que cumplen con un determinado patrón. Sin embargo, si dicha tarea se le deja a que lo haga una computadora, no será nada fácil implementarla. Es por ello que uno de los temas que ha sido tratado por las ciencias de la computación, es el procesamiento de texto. De allí que hayan tomado importancia para este campo el empleo de las ER, ya que con ellas es posible realizar operaciones de validación, de búsqueda, de extracción y de sustitución de texto.

Las ER son empleadas de forma cotidiana en el ámbito de la computación debido a que tienen una aplicación importante al trabajar con cadenas alfanuméricas sin tener que enumerar todos y cada uno de sus elementos. Lo anterior se puede realizar empleando una sintaxis o un lenguaje propio que también va de la mano del lenguaje de programación a emplear, ya que de ello depende el conjunto de instrucciones para invocar el empleo y definición de ER en alguna aplicación [2].

¿Qué son las ER?

Las ER son mecanismos flexibles y eficientes que permiten el procesamiento de textos. Por ejemplo, si se necesitara realizar la búsqueda de una palabra o palabras dentro de un texto extenso, considerando que dicha palabra o palabras pueda tener uno u otro caracter distinto. Independientemente de la variación en el caracter que se tenga, es posible realizar la búsqueda haciendo uso de metacaracteres, que generalmente es un paréntesis ya que suele ocuparse para delimitar la alternancia de caracteres que hacen que la o las palabras a buscar tengan algunas variantes.

Por otro lado [3] define una ER como una forma abreviada de representar cadenas de caracteres que se ajustan a un determinado patrón. Y es que el empleo de ER conlleva a definir los caracteres que conformarán un alfabeto específico (que pueden ser llamados literales o caracteres normales de texto), además de un conjunto de operadores a los cuales se les denominan metacaracteres.

Es así como a través del empleo de los caracteres y operadores de las ER, es posible definir secuencias de caracteres más complejas, las cuales permitirán realizar en análisis de los mensajes escritos por una persona en un medio de comunicación digital con el objetivo de verificar si cumple con las reglas ortográficas y/o gramaticales del idioma español, para llegar a ser una herramienta que permita a los aprendices del idioma verificar si cumple o no con las reglas del mismo, principalmente cuando la herramienta es implementada en forma de aplicación. De tal manera que daría origen a un aprendizaje asistido por un dispositivo digital.

La alternancia consiste en la posibilidad de buscar una secuencia de caracteres que pueden contener uno y otro caracter de forma indistinta. Para definir ello es necesario hacer uso de metacaracteres que permitan indicar cuáles son los caracteres que pudieran estar alternándose en la palabra a buscar [2].

Componentes de las ER

Una ER como tal es un lenguaje que también tiene cierta sintaxis para poder acceder a ellas. Por lo que de acuerdo con [4] los principales componentes son:

Literales
Clases de caracteres
Metacaracteres

Las literales son aquellos caracteres que conforman cualquier palabra, a menos de que se trate de un metacaracter. Las clases de caracteres son aquellas que

son definidas a través de una lista de caracteres entre corchetes. Los metacaracteres son aquellos especiales que permiten delimitar, iterar, alternar, etc. a las literales. En los metacaracteres radica la esencia de las ER.

Algunos de los metacaracteres más empleados descritos por [5] se muestran en la tabla 1.

Metacaracter	**Explicación**
\	Expresa que el siguiente caracter es especial.
^	Indica la posición del comienzo de la cadena de entrada.
$	Indica la posición del final de la cadena de entrada.
*	Indica que el caracter que le precede se presenta cero o más veces.
+	Indica que el caracter que le precede se presenta una o más veces.
()	Indica la creación de subcadenas.
{n}	Indica la repetición de un carácter n veces.
{n,}	Indica la repetición de un carácter como mínimo n veces.
{n,m}	Indica la repetición de un carácter como mínimo n veces y como máximo m veces.
x|y	Indica x o y
[xyz]	Indica la coincidencia con cualquiera de los caracteres entre los corchetes.
[^xyz]	Indica que cualquier carácter que no esté entre corchetes.
[a-z]	Indica cualquier caracter dentro del rango especificado.

Tabla 1. Ejemplos de metacaracteres. Creación propia

Los metacaracteres descritos en la tabla 1 son solo unos cuantos de los muchos que están definidos para establecer patrones de texto a buscar.

Un ejemplo de cómo se aplican las ER se muestra de la siguiente manera \sca(s|z|d)a\s. Con dicha ER estamos indicando que antes de la palabra lleva un espacio para luego comenzar "ca", seguida de 3 posibles letras (una s, o z, o d), seguida de la letra "a", para luego terminar con otro espacio. De acuerdo con la ER anterior, las palabras que pueden ser descritas por ella son *casa*, *caza*, *cada*. En la ER puede verse dónde está definida la alternancia a través del uso de paréntesis y de carácter "|".

Con base en lo anterior, es como el empleo de ER se encuentra muy difundida dentro del ámbito del Procesamiento del Lenguaje Natural, ya que el análisis de la lingüística del corpus se realiza a través de las ER para encontrar coincidencias, o bien, para establecer alguna alternancia entre las palabras a buscar.

Inclusive empleando ER es posible realizar equivalencias entre diferentes formatos de codificación de los caracteres, principalmente cuando se emplean acentos o el uso de la eñe en el idioma español. En este caso en particular, es posible omitir el uso de herramientas o codificadores/decodificadores para mantener en una secuencia alfanumérica los caracteres especiales empleados en el español.

Aplicación de las ER en la enseñanza del idioma español

El proceso para el empleo de las ER para la enseñanza del idioma español puede verse descrito en la figura 1.

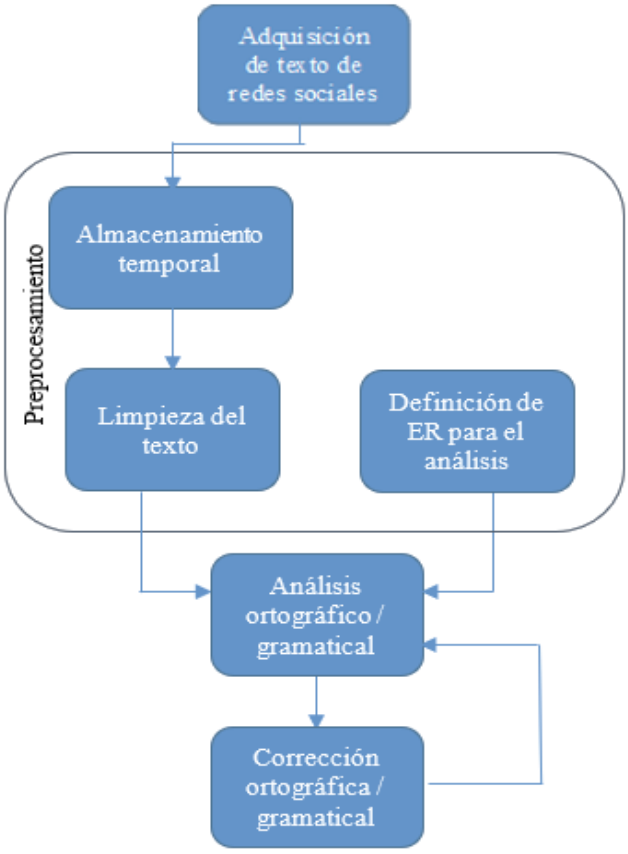

Figura 1. Fases del proceso de corrección ortográfica / gramatical. Creación propia

En la fase de preprocesamiento se realiza el almacenamiento de los mensajes de texto que son obtenidos desde la interacción con alguna red social. Para el presente trabajo cabe mencionar que la obtención de mensajes se hace sobre la red social Twitter a través de la API (Interfaz de Programación de Aplicaciones, *Application Programming Interface*) que dicha red ofrece para generar aplicaciones por parte de terceros.

El almacenamiento de los mensajes será útil para poder realizar una limpieza de los mismos, cuyo objetivo es el eliminar todo carácter y/o imagen adicional que pueda acompañarlo. De tal manera que los datos almacenados sean en texto plano. Un ejemplo de los primeros diez mensajes obtenidos de Twitter sin haber recibido algún tipo de limpieza se muestra en la figura 2.

Los mensajes de texto enviados por la red social, una vez que son limpiados para eliminar los caracteres que no son necesarios para la revisión ortográfica y/o gramatical, lucen como se muestra en la figura 4.

1	['2018-03-08 23:58:49', "b'Esta claro que les importa poco la comunidad estudiantil, la econom\\xc3\\xada del estado y la prosperidad del mismo\\xe2\\x80\\xa6 https://t.co/TdtQLWfhKO'"]
2	['2018-03-08 23:56:04', "b'Urge, resuelvan que va a pasar con las clases , neta no tienen madre, ustedes un d\\xc3\\xada fueron estudiantes tambien cab\\xe2\\x80\\xa6 https://t.co/JHeboskIH0'"]
3	['2018-03-08 23:53:14', "b'RT @Luz_Macm:¿que\\xc3\\xb1a con la presidencia de M\\xc3\\xa9xico en un futuro ?\\nNo logra ni siquier\\xe2\\x80\\xa6'"]
4	['2018-03-08 23:51:43', "b'De verdad creen que pueden enga\\xc3\\xb1ar a la gente con su simulaci\\xc3\\xb3n de huelga?https://t.co/JnqWstDsFU'"]
5	['2018-03-08 23:51:35', "b'RT @Luz_Macm: Disculpe y con esa apat\\xc3\\xada para con ellos sue\\xc3\\xb1a con la presidencia de M\\xc3\\xa9xico en un futuro ?\\nNo logra ni siquier\\xe2\\x80\\xa6'"]
6	['2018-03-08 23:53:14', "b'RT @Luz_Macm:¿que\\xc3\\xb1a con la presidencia de M\\xc3\\xa9xico en un futuro ?'"]
7	['2018-03-08 23:46:29', "b'RT @gabolazca: @CarlosLoret Oye Carlos, porque no cubres la nota, es la misma\\xe2\\x80\\xa6'"]
8	['2018-03-08 23:44:30', "b'@MirnaLopezU \\xf0\\x9f\\xa4\\x94\\xf0\\x9f\\xa4\\x94 eso lo dijo Sim\\xc3\\xb3n Vargas @SeGobHidalgo , y luego sali\\xc3\\xb3 corriendo \\xf0\\x9f\\x98\\x82'"]
9	['2018-03-08 23:33:26', "b\'RT @ikuronomi: Algo no me cuadra, intent\\xc3\\xa9 verificar \\xc3\\xa9sta informaci\\xc3\\xb3n en el portal del SAT y llegu\\xc3\\xa9 a un "callej\\xc3\\xb3n sin salida" \\xe2\\x80\\xa6\'"]
10	['2018-03-08 23:33:01', "b'¿Que todo esto del la huelga es una simulaci\\xc3\\xb3n?

Figura 2. Mensajes obtenidos de la red social de Twitter.

En la misma fase de preprocesamiento se definen las ER que serán empleadas para realizar el análisis ortográfico / gramatical. Pueden definirse tantas como se requieran, ya que estarán sujetas al grado exactitud que se quiera con las reglas del idioma español.

Una vez definidas las ER, es posible realizar el análisis ortográfico / gramatical de los mensajes de texto, de tal manera que pueda emitirse un resultado de la revisión. En caso de haber algún error, será notificado en la fase de corrección ortográfica / gramatical.

El análisis se realiza por oración, de tal manera que se realiza de manera iterativa el análisis ortográfico / gramatical tantas veces como se requiera.

Para el presente trabajo las expresiones regulares han sido definidas en Python, por lo que un ejemplo de ellas se muestra a continuación:

¿(Qué|Cómo|Quién|Cuándo)\s\w+(\s\w+)*\?

A través de dicha ER se está definiendo que ante el empleo de las palabras Qué, Cómo, Quién o Cuándo (el metacaracter "|" indica que la operación lógica "O", es decir, la presencia de una u otra de las palabras que inician el cuestionamiento), deberán iniciar con el signo de interrogación "¿" dado que, al estar acentuadas dichas palabras, se infiere que se realiza un cuestionamiento. Enseguida se emplea el metacaracter "\s", el cual indica que después de dichas palabras que inician un cuestionamiento debe estar presente un espacio, seguido de al menos una o más palabras representadas por "\w+". Entre paréntesis se encuentra "(\s\w+)*" debido a que se indica que la oración podría tener nuevamente cero o más espacios, seguidos de cero o más palabras, pero al final la oración debe finalizar con el signo de interrogación "?".

La expresión regular mostrada, permite leer cualquier frase independientemente de su longitud.

Otra expresión regular que define el uso correcto de los dos puntos ":" se muestra a continuación:

\w+:\s[:lower:]\w+

El uso de los dos puntos es utilizado para establecer enumeraciones, a las explicaciones y a las ampliaciones. Por lo que al emplearlos la siguiente palabra debe estar en minúsculas. Es por ello que la expresión regular define que ante una o más palabras ("\w+") puede presentarse el uso de los dos puntos (":"), para luego estar presente un espacio ("\s") seguido de una o más palabras en minúsculas "\w+".

La figura 3 muestra parte del código que describe el funcionamiento de la expresión regular que define cómo deben realizarse los cuestionamientos en el idioma español.

```
with open("/media/alsolmtl/UUI/Investigacion/Expresiones regulres/Programas/result_norm.csv") as read:
    reader=csv.reader(read)
    print("\nSe realizará la revisión del empleo de signos de interrogación en el mensaje enviado...")
    for texto in reader:
        regex=re.compile(r"(¿|\?)")
        if (regex.search(str(texto))):
                regex=re.compile(r"¿\w+\s+\w+(\w+\s+\w+)*\?")
                if(regex.search(str(texto))):
                        regex=re.compile(r"¿(Qué|Cómo|Quién|Cuándo)\s\w+(\s\w+)*\?")
                        if (regex.search(str(texto))):
                            print("El cuestionamiento está correctamente escrito.")
                        else:
                            if (SinMayusc()==0):
                                if(SinAcento()==0):
                                    SinAcentoMinus()
                else:
                        regex=re.compile(r"\?")
                        if(regex.search(str(texto))):
                            print("Te falta el signo ¿ en el mensaje.")
```

Figura 3. Código implementado en Python.

Como se puede observar, antes de llegar a la ER de un cuestionamiento se aplican diferentes filtros para validar si la lectura del mensaje enviado por la red social se aproxima a una pregunta. En caso de no serlo simplemente se descarta del análisis.

Antes de realizar el análisis del mensaje primero se revisa, a través de una ER, si el mensaje contiene signos de interrogación, éstos deben estar en el orden adecuado. Si se cumplen ambas cosas, se procede al análisis del cuestionamiento empleando nuevamente ER.

Dentro del análisis que se realiza, se pueden identificar algunos casos, por ejemplo:

Las palabras que inician el cuestionamiento (qué, cómo, quién, cuándo) no inician con mayúscula.

Las palabras que inician el cuestionamiento no están acentuadas.

Las palabras que inician el cuestionamiento no inician con mayúsculas, ni tampoco se encuentran acentuadas.

Un ejemplo de los mensajes que se obtienen de la red social de Twitter ya limpios se muestra en la figura 4. Las respuestas obtenidas del sistema se muestran en la figura 5.

1	['['2018-03-08 23:58:49', "b'Esta claro que les importa poco la comunidad estudiantil, la economia del estado y la prosperidad del mismo '"]']
2	['['2018-03-08 23:56:04', "b'Urge, resuelvan que va a pasar con las clases , neta no tienen madre, ustedes un dia fueron estudiantes tambien cab '"]']
3	['['2018-03-08 23:53:14', "b'RT @Luz_Macm: Â¿que suena con la presidencia de Mexico en un futuro? nNo logra ni siquier'"]']
4	['['2018-03-08 23:51:43', "b'De verdad creen que pueden enganar a la gente con su simulacion de huelga?
5	['['2018-03-08 23:51:35', "b'RT @Luz_Macm: Disculpe y con esa apatia para con ellos suena con la presidencia de Mexico en un futuro? No logra ni siquier'"]']
6	['['2018-03-08 23:53:14', "b'RT @Luz_Macm: Â¿que suena con la presidencia de Mexico en un futuro?
7	['['2018-03-08 23:46:29', "b'RT @gabolazca: @CarlosLoret Oye Carlos, porque no cubres la nota, es la misma Â¿QuÃ© con eso?'"]']
8	['['2018-03-08 23:44:30', "b'@MirnaLopezU eso lo dijo Simon Vargas @SeGobHidalgo , y luego salio corriendo '"]']
9	['['2018-03-08 23:33:26', 'b'RT @ikuronomi: Algo no me cuadra, intente verificar esta informacion en el portal del SAT y llegue a un "callejon sin salida"
10	['['2018-03-08 23:33:01', "b' v Â¿Que todo esto del la huelga es una simulacion?

Figura 4. Diez mensajes obtenidos de redes sociales ya limpios.

En dicha figura es posible ver la variedad de resultados que se generan al realizar el análisis de los mensajes que contienen un cuestionamiento. En los resultados que genera el sistema, también se indica a qué mensaje de la red social corresponde el resultado generado, para facilitar la revisión y depuración del sistema en caso de que dichos resultados no correspondan a los mensajes obtenidos de la red social.

```
NewPythonProject3 ✕   NewPythonProject3 #2 ✕

Se realizará la revisión del empleo de signos de interrogación en el mensaje enviado...
La palabra Qué lleva acento y empieza con mayúscula. Línea . 3
Te falta el signo ¿ en el mensaje. Línea . 4
Te falta el signo ¿ en el mensaje. Línea . 5
La palabra Qué lleva acento y empieza con mayúscula. Línea . 6
El cuestionamiento está correctamente escrito. Línea . 7
La palabra Qué lleva acento. Línea . 10
Te falta el signo ¿ en el mensaje. Línea . 22
Te falta el signo ¿ en el mensaje. Línea . 24
```

Figura 5. Resultados generados tras la revisión de mensajes de la red social.

Conclusiones

De manera cotidiana, el aprendizaje de un idioma se hace a través del apoyo de un profesor que se encarga de desarrollar las competencias necesarias para el manejo de un segundo idioma. Sin embargo, pese a que

hoy en día es posible encontrar apps para casi todo, son pocas las que hacen uso de herramientas de inteligencia artificial para retroalimentar el uso del idioma de personas que estén aprendiendo el español. Por lo tanto, contar con una app que cuando el estudiante del idioma esté hablando o se encuentre escribiendo en algún medio de comunicación digital, puede resultar provechoso, ya que no es necesario que en ese preciso instante se encuentre alguien que domine el idioma para poder corregir o retroalimentar al estudiante.

Como pudo observarse, el empleo de ER puede ser una herramienta que facilite el análisis de los mensajes que pueden ser enviados en distintas redes sociales y tener la oportunidad de contar con una herramienta que haga la revisión ortográfica y/o gramatical para recibir en ese instante la retroalimentación correspondiente.

Para ejemplificar la aplicación de ER para la revisión ortográfica y/o gramatical del presente trabajo, tan solo se empleó la definición de un cuestionamiento en el idioma español. Para realizarlo se requirió de establecer algunos filtros previos que permitieran identificar si el mensaje en cuestión correspondía a un cuestionamiento o no. Por lo tanto, para realizar la revisión de más reglas ortográficas y gramaticales requiere de la programación de diferentes filtros y casos que generaría un trabajo arduo al momento de programar, por lo rico que resulta ser el idioma español y por citar un ejemplo de ello, la prueba está en la variedad de formas que adoptan, por ejemplo, los verbos, en todas las personas gramaticales empleadas en el idioma.

Referencias

[1] A. Gómez Camacho y M. T. Gómez del Castillo, «Escritura ortográfica y mensajes de texto en estudiantes universitarios,» *Perfiles educativos*, vol. 37, nº 150, pp. 91-104, 2015.

[2] R. Chacón Beltrán, «El uso de expresiones regulares en la detección de errores escritos: implicaciones para el diseño de un corrector gramatical,» *Actas completas del VIII Congreso de Lingüística General 2008*, 2008.

[3] M. Alfonseca Moreno, M. de la Cruz Echeandía, A. Ortega de la Puente y E. Pulido Cañabate, Compiladores e intérpretes. Teoría y práctica, Madrid: Pearson Prentice Hall, 2006.

[4] J. Goyvaerts y S. Levithan, Regular expressions cookbook, O'Reilly, 2012.

[5] R. E. López Briega, «Matemáticas, análisis de datos y python,» 19 Julio 2015. [En línea]. Available: https://relopezbriega.github.io/blog/2015/07/19/expresiones-regulares-con-python/. [Último acceso: 26 Febrero 2019].

[6] J. E. F. Friedl, Mastering Regular Expressions, California: O'Reilly, 2006.

[7] J. Goyvaerts, «Regular Expressions. The Complete Tutorial,» 2007. [En línea]. Available: https://www.princeton.edu/~mlovett/reference/Regular-Expressions.pdf. [Último acceso: 8 Marzo 2019].