Recolector y Clasificador de Noticias Mediante Inteligencia Artificial

Trabajo Terminal No. ____-___

Alumnos: Hernández Gómez Carlos Andrés, Meza Martínez Luis Daniel
Directores: Juárez Gambino Joel Omar, García Mendoza Consuelo Varinia
Turno para la presentación del TT: Matutino
e-mail: ldanielmezam@gmail.com

Resumen – En el presente trabajo terminal se propone desarrollar un recolector (crawler) de noticias, que permita recuperar artículos publicados en diferentes sitios de información. El recolector permitirá establecer dos filtros: el periodo de fechas en el cual se publicaron las noticias y la sección o secciones a las cuales pertenecen estas noticias. Algunas fuentes de información no dividen sus noticias en secciones o el nombre de las secciones es distinto entre ellas a pesar de que el tipo de contenido sea el mismo (por ejemplo, en el periódico Excelsior la sección de deportes se llama adrenalina), esto complica la recuperación de la información. Por lo tanto,
se propone que las noticias recuperadas sean clasificadas en secciones de forma automática, de acuerdo con su contenido.

Palabras clave – Aprendizaje automático, Clasificación de texto, Web Crawler.

1. Introducción

La noticia es la información de un hecho de interés ocurrido recientemente. Constituye el elemento primordial de la información periodística y el género básico del periodismo [1]. Conocer los acontecimientos del mundo independientemente del tema, día, lugar en que se suscitan, tiene una gran importancia en la sociedad, se comparten por el conjunto de medios de comunicación, como televisión, redes sociales, diarios, blogs, radios, las cuales llamamos noticas, estas nos permiten conocer la situación económica del país, logros de la ciencia, desastres naturales, la situación en cuestión de inseguridad. En el ámbito de las inversiones, las noticias crean expectativas y eso a su vez puedes modificar los planes de inversión en cualquier sector, siendo así de suma importancia compartirlas de una forma eficaz [2].

Un Web Crawler es el proceso mediante el cual se reúnen páginas de la web, para indexarlas y respaldarlas con un motor de búsqueda. El objetivo del Crawler es reunir de manera rápida y eficiente tantas páginas web como sea posible [3]. En la actualidad las páginas web van incrementando día con día, por lo cual podemos consultar información relevante de distintos sitios, uno de ellos son los periódicos electrónicos, los cuales dividen sus artículos en secciones para facilitar la búsqueda del usuario, sin embargo, existen un sinfín de sitios en la red independientes, que nos proveen de una gran variedad de artículos, dichos sitios no cuentan con una clasificación particular, por lo que resulta difícil para el usuario realizar una búsqueda específica dentro de dicho sitio.

Dado que existe una gran cantidad de sitios web que publican noticias, se han creado algunas aplicaciones similares a la propuesta en este trabajo que permiten la recolección de noticias de interés para el usuario como Flipboard [4], Huffpost [5] y Google News [6].
En la Tabla 1 se muestran dichas aplicaciones con sus características más relevantes.

SOFTWARE	CARACTERÍSTICAS	
Flipboard	Consulta diferentes fuentes	Si
	Permite indicar secciones de interés	Si
	Permite establecer periodo de interés	No
	Clasifica los artículos automáticamente	No
Huffpost	Consulta diferentes fuentes	Si
	Permite indicar secciones de interés	Si
	Permite establecer periodo de interés	No
	Clasifica los artículos automáticamente	No

Google News	Consulta diferentes Fuentes	Si
	Permite indicar secciones de interés	Si
	Permite establecer periodo de interés	No
	Clasifica los artículos automáticamente	No

Tabla 1. Resumen de productos similares.

Una diferencia importante del trabajo propuesto con las aplicaciones es el hecho de clasificar la información de forma automática, cabe mencionar que existe un trabajo terminal realizado previamente el cual lleva como título "Clasificación de noticias de diarios de circulación nacional mediante aprendizaje automático" con número 2017-A02 como se puede ver en [7], en este trabajo se llevó acabo la clasificación de noticias en forma automática , el trabajo propuesto se basa en dicha clasificación para realizar el recolector (crawler).

2. Objetivo

Crear un recolector, el cual permita visualizar un conjunto de noticias, recopilanado información de diferentes fuentes como diarios, sitios de noticias, blogs, foros y mediante el análisis automático de su contenido muestre aquellas noticias que satisfagan los filtros de período y secciones establecidas por el usuario.

Objetivos particulares

- Desarrollar un buscador, el cual permita recopilar información de diferentes fuentes como diarios, sitios de noticias, blogs, foros.
- Analizar de forma automática el contenido de las noticias para satisfacer los filtros establecidos por el usuario.
- Mostrar el enlace (URL) de las noticias que cumplieron con los filtros establecidos
- Afinar el clasificador de noticias en secciones realizado en el trabajo terminal 2017-A02 para utilizarlo en el contexto de esta propuesta.

3. Justificación

En el mundo existen distintas problemáticas, de las cuales no siempre obtenemos la información adecuada o simplemente no existe dicha información, lo cual genera problemas en distintos ámbitos, social, cultural, político, económico, por ejemplo, cuando existe una desinformación sobre un acontecimiento reciente el cual afecta el ámbito económico a nivel nacional, se pueden suscitar inconformidades en la sociedad.

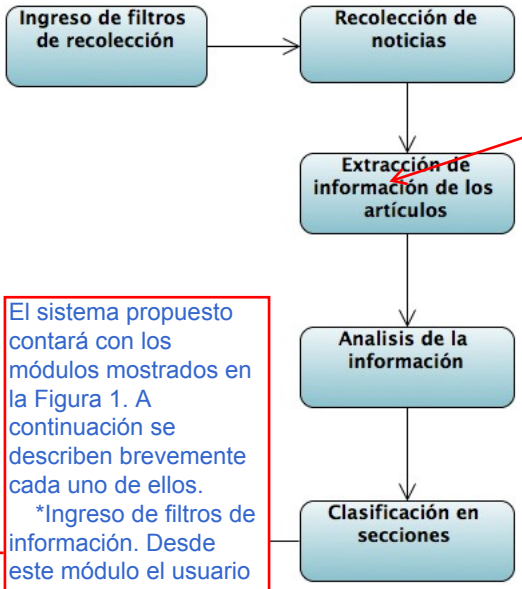
Hoy en día existen distintas maneras de informarnos acerca de s recientes, por ejemplo, en la televisión, blogs, redes sociales, foros, diarios, de los cuales podemos ob a día, sin embargo, hay ocasiones en las cuales nos informan de manera equivocada o simplemente nos y emos de acceder a muchos recursos para verificar la veracidad de la información, lo cual llega a ser po y esfuerzo.

Internet es una forma eficiente para conocer sobre los acontecim encionar que una gran parte de los sitios web no cuentan con la clasificación de sus noticas en seccion n con nombres especiales para llamar la atención del usuario, según El Economista [8] animalpolitico.c o cuatro en el ranking de medios nativos digitales, el cual clasifica sus noticias de una manera poco habi no la sección "El sabueso", "El plumaje", "Hablemos de … ", entre otras, lo que hace complicado obte evantes del momento, de igual forma las noticias mostradas tienen que ser clasificadas de manera manua cursos y tiempo para tener preparada una nueva edición del sitio.

Para ayudar en este problema existen herramientas que hacen la búsqueda de noticias de interés para el usuario en forma automática. Sin embargo dichas herramientas dependen de que los sitios a consultar cuenten con una etiquetación correcta y homogenea de la información.

Una solución es tener un lugar el cual nos proporcione distintas fuentes de información para su consulta de una forma ordenada y clasificada, esto nos permitiría tener el conocimiento de los hechos con una mayor velocidad. Se propone crear un buscador el cual permita recolectar noticias de distintas fuentes de información, posteriormente clasificarlas según la sección de interés y el periodo de fecha establecido de una forma automática y enseguida mostrarlas al usuario.

4. Productos o resultados esperados



Cambiar por: Extracción de información

El sistema propuesto contará con los módulos mostrados en la Figura 1. A continuación se describen brevemente cada uno de ellos.
*Ingreso de filtros de información. Desde este módulo el usuario podrá establecer el periodo y las secciones de las noticias de su interés.
*Recolección de noticias...
* ...

Al finalizar este trabajo se contará con:
*Documentación de...

...ctura del sistema.

- Documentación de análisis y dis...
- Recolector Web.
- Algoritmos de clasificación auto...

5. Metodología

La metodología que utilizaremos para la r... rabajo terminal será la incremental como lo explica [9], debido a que uno de los objetivos es el crecimiento progresivo, es decir se realizan entregas parciales en un periodo de tiempo corto y así reducir el peligro en el proyecto.

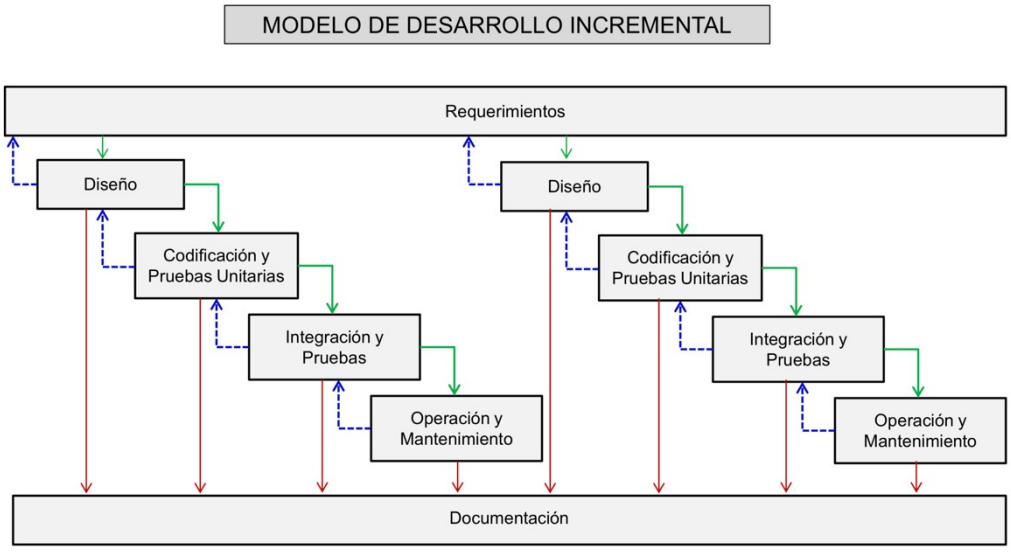


Figura 2. Modelo de desarrollo incremental [7].

6. Cronograma

Nombre del alumno(a): Hernández Gómez Carlos Andrés
Título del TT:

Actividad	FEB	MAR	ABR	MAYO	JUNI	AGO	SEP	OCT	NOV	DIC
Selección de fuentes de información										
Definición de requerimientos										
Análisis y diseño del módulo recolector de noticias										
Implementación del módulo recolector de noticias										
Pruebas del módulo recolector de noticias										
Evaluación de TT1										
Análisis y diseño del extractor de información										
Implementación del extractor de información										
Integración de los módulos										
Pruebas integrales										
Redacción de los manuales de usuario y técnico										
Evaluación de TT2										

Nombre del alumno(a): Meza Martínez Luis Daniel
Título del TT:

Actividad	FEB	MAR	ABR	MAYO	JUNI	AGO	SEP	OCT	NOV	DIC
Selección de fuentes de información										
Definición de requerimientos										
Análisis y diseño del módulo recolector de noticias										
Implementación del módulo recolector de noticias										
Pruebas del módulo recolector de noticias										
Evaluación de TT1										
Análisis y diseño del extractor de información										
Implementación del extractor de información										

Integración de los módulos										
Pruebas integrales										
Redacción de los manuales de usuario y técnico										
Evaluación de TT2										

7. Referencias

[1] S/A. (S/A). Importancia de las Noticias. 07/09/2018, de INNOVACION INTERNAUTICA Sitio web: https://innovainternetmx.com/2014/12/importancia-de-las-noticias/

[2] Manning, C., Raghavan, P. and Schütze, H. (2009). Introduction to information retrieval. New York: Cambridge University Press, pp.443-459.

[3] Bernabeu Morón, N. (2013). La noticia y el reportaje. España: Ministerio de Educación de España, p.9.

[4] MIKE MCCUE. (S/A). FLIPBOARD ES TU MOMENTO. 06/09/2018, de FLIPBOARD Sitio web: https://es-es.about.flipboard.com/?noredirect=es_ES

[5] HuffPost. (S/A). HuffPost México. 06/09/2018, de HuffPost Sitio web: https://www.huffingtonpost.com.mx/p/huffpost-mexico-about-us

Falta espacio

[6] Google. (S/A). Google News. 12/09/2018, de Google Sitio web: https://news.google.com/

[7] J. García ,L. Ramírez, M. Sanchéz, "Clasificación de noticias de diarios de circulación nacional mediante aprendizaje automático", trabajo terminal, ESCOM IPN, 2018.

[8] El Economista Y ComScore. (03/09/2018). Ranking de Medios Nativos Digitales. 12/09/2018, de El Economista Sitio web: https://www.eleconomista.com.mx/tecnologia/Ranking-de-Medios-Nativos-Digitales-20170830-0177.html

[9] D. Tapias, "Proyectos de desarrollo software", 2014, [En linea]. Disponible en : http://arantxa.ii.uam.es/~proyectos/teoria/C5_Proyectos%20de%20desarrollo%20software.pdf

8. Alumnos y Directores

Hernandez Gomez Carlos Andres.- Alumno de la carrera de ingeniería en sistemas computacionales en la Escuela Superior de Cómputo del Instituto Politécnico Nacional,

Boleta: 2015620193, Tel. 5546218045, email: carlosandreshg.ipn@gmail.com

Firma:_____

Meza Martínez Luis Daniel.- Alumno de la carrera de ingeniería en Sistemas Computacionales en la Escuela Superior de Cómputo del Instituto politécnico nacional , Boleta:2015630305 , Tel.5573994181 , email: *ldanielmezam@gmail.com*

Firma:_____

Joel Omar Juárez Gambino.- Licenciado en Informática por la Facultad de Informática, UAS. Maestro en Ciencias de la computación por el CIC, IPN. Sus áreas de estudios son: Inteligencia Artificial, Lenguaje Natural y Representación de conocimiento. Departamento de Ciencias e Ingeniería de la computación, ESCOM, Tel. 57296000 Ext. 52022, email: omarjg@gmail.com

Firma:_____

Consuelo Varinia García Mendoza.- Ingeniera en Sistemas Computacionales por la ESCOM, IPN , UAS. Maestra en ciencias en Tecnología Avanzada por el CICATA-Legaria, IPN, Doctora en Técnologias Avanzadas por la CICATA-Legaria, IPN. Sus áreas de estudios son: Análisis de algoritmos y Optimización. Departamento de Ciencias e ingeniería de la computación, ESCOM, Tel. 57296000 Ext. 52022, email: consuelo.varinia@gmail.com

Firma:_____