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 fuentes 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, adrenalina y deportes), esto complica la recuperación de la información. Por lo tanto, en este trabajo terminal 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**

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 [1].

En la actualidad los sitios web van incrementando día con día, por lo cual existen muchos sitios en los cuales podemos consultar información, 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, los cuales nos proveen de distintos tipos 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 [2], Huffpost [3] y Google News [4] . 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 [6], 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).

### Objetivo

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

Objetivos particulares

* Desarrollar un sitio web el cual permita recopilar información de diferentes fuentes como diarios, sitios de noticias, blogs, etc. y mediante el análisis automático de su contenido muestre aquellas noticias que satisfagan los filtros de período y secciones establecidos por el usuario.
* 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.

1. **Justificación**

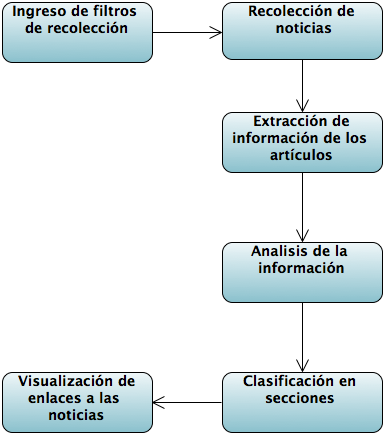
En el mundo existen distintas problemáticas, de las cuales no siempre obtenemos la información adecuada o simplemente no existe dicha información, lo cual genera problemas en distintos ámbitos, social, cultural, político, económico, por ejemplo, cuando existe una desinformación sobre un acontecimiento 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 los acontecimientos más recientes, por ejemplo, en la televisión, blogs, redes sociales, foros, diarios, de los cuales podemos obtener las noticas del día a día, sin embargo, hay ocasiones en las cuales nos informan de manera equivocada o simplemente nos dan una noticia a medias y debemos de acceder a muchos recursos para verificar la veracidad de la información, por lo cual llega a ser gasto de recursos, tiempo y esfuerzo.

Internet es una forma eficiente para conocer sobre los acontecimientos del mundo, cabe mencionar que una gran parte de los sitios web no cuentan con la clasificación de sus noticas en secciones o lo hacen con nombres especiales para llamar la atención del usuario, según [5] animalpolitico.com ocupa el lugar número cuatro en el ranking de medios nativos digitales, el cual clasifica sus noticias un tanto extravagante como la sección “El sabueso”, “El plumaje”, “Hablemos de … “, entre otras, lo que hace complicado obtener los artículos mas relevantes del momento, de igual forma las paginas mostradas en [5] tiene que ser clasificadas de una forma manual es decir, se invierten recursos y tiempo para tener preparada una nueva edición del sitio.

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 sitio web 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.

1. **Productos o resultados esperados**

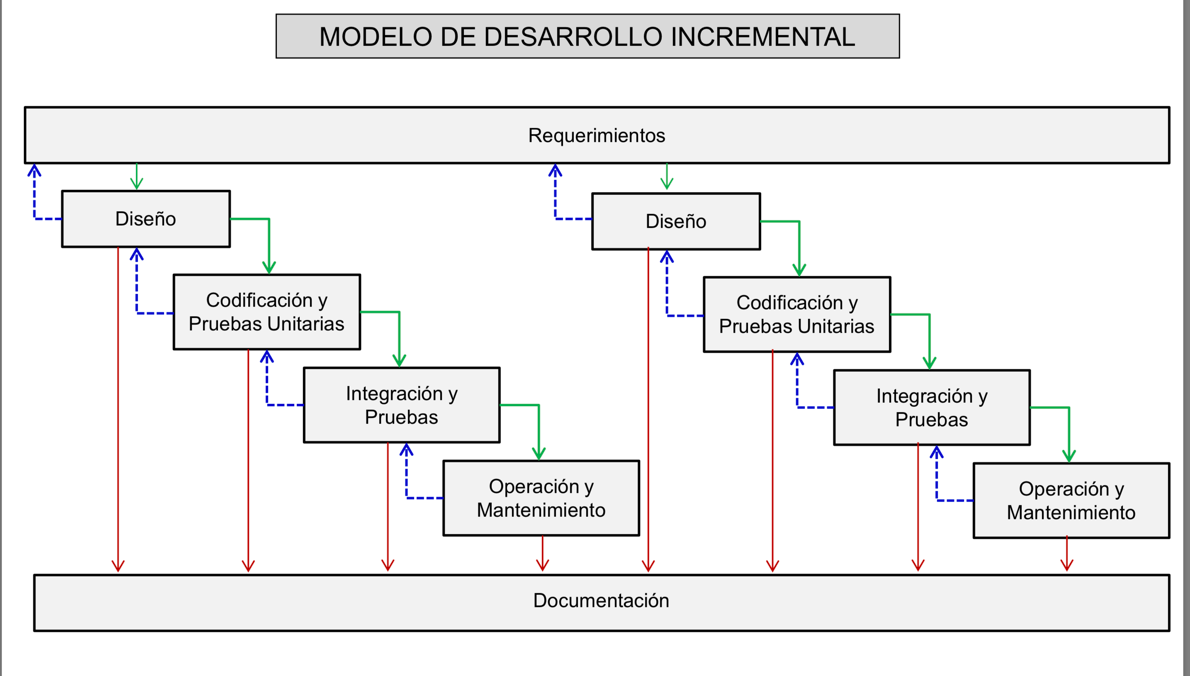


**Figura 1**. Arquitectura del sistema.

* Documentación de análisis y diseño.
* Recolector Web.
* Algoritmos de clasificación automática.

1. **Metodología**

La metodología que utilizaremos para la realización del presente trabajo terminal será la incremental como lo explica [7], 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].

1. **Cronograma**

Nombre del alumno(a): Hernadez Gómez Carlos Andres

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 |
| Definición de secciones a utilizar en el filtrado |  |  |  |  |  |  |  |  |  |  |
| Definición de requerimientos |  |  |  |  |  |  |  |  |  |  |
| Adaptación del módulo clasificador de noticias realizado en el trabajo previo (TT 2017-A02) |  |  |  |  |  |  |  |  |  |  |
| Pruebas del módulo clasificador de noticias |  |  |  |  |  |  |  |  |  |  |
| Evaluación de TT1 |  |  |  |  |  |  |  |  |  |  |
| Análisis y diseño del módulo visualizador de noticias filtradas |  |  |  |  |  |  |  |  |  |  |
| Implementación del módulo visualizador de noticias filtradas |  |  |  |  |  |  |  |  |  |  |
| Intregración de módulos |  |  |  |  |  |  |  |  |  |  |
| Pruebas integrales |  |  |  |  |  |  |  |  |  |  |
| Redacción de los manuales de usuario y técnico |  |  |  |  |  |  |  |  |  |  |
| Evaluación de TT2 |  |  |  |  |  |  |  |  |  |  |

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

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

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

[5] 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

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

[7] D. Tapias, “Proyectos de desarrollo software”, 2014, [En linea ]. Disponible en : http://arantxa.ii.uam.es/~proyectos/teoria/C5\_Proyectos%20de%20desarrollo%20software.pdf

1. **Alumnos y Directores**

CARÁCTER: Confidencial

FUNDAMENTO LEGAL: Art. 3, fracc. II, Art. 18, fracc. II y Art. 21, lineamiento 32, fracc. XVII de la L.F.T.A.I.P.G.

PARTES CONFIDENCIALES: No. de boleta y Teléfono.

|  |
| --- |
| *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:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_