

Buscador web de noticias clasificadas 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 buscador web de noticias, mediante la clasificación de texto con técnicas de aprendizaje automático. Se analizarán las noticias de distintos sitios web y se clasificarán, para su posterior visualización.

Palabras clave – Aprendizaje automático, Clasificación de texto, Desarrollo web, Procesamiento de lenguaje natural.

1. Introducción

El conocimiento da significado al mundo en que vivimos, al ignorar los hechos de nuestra vida y de la sociedad caminamos por un sendero oscuro, las decisiones importantes son tomadas por un conjunto reducido la población, se habla de política y se critica sin conocer los hechos que se suscitaron, se juzga la economía de una nación y su poco aporte científico, se reprueban las protestas de los estudiantes, los escándalos de los gobiernos son callados por los medios de comunicación como la desaparición de 43 estudiantes, o el genocidio de Guatemala en 1980, ¿Quién puede regresar a la vida las víctimas del 2 de octubre del 1968 ?, no podemos negar que una nación que no conoce su historia está condenada a repetirla.

Conocer los acontecimientos del mundo independientemente del tema, día, lugar en que se suscitan, tiene una gran importancia en la sociedad, se comparte por el conjunto de medios de comunicación, como televisión, redes sociales, diarios, blogs, radios, las cuales llamamos noticias, 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.

En la actualidad los sitios web[1] 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 filtran sus artículos por categorías, para facilitar la búsqueda del usuario, sin embargo, existen un sinnúmero 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.

Sistemas similares que se han desarrollado son.

1. Flipboard
2. Huffpost
3. Clasificación de noticias de diarios de circulación nacional mediante aprendizaje automático 2017-A02.

SOFTWARE	CARACTERÍSTICAS	PRECIO EN EL MERCADO
Flipboard	<ul style="list-style-type: none"> Revista personalizable Tiene blogs creados por personas en los cuales pueden subir información y se pueden Actualizable Gran cantidad de contenido así como de una gran cantidad Referencias de sitios españoles. 	<div>Consulta diferentes fuentes Permite indicar secciones de interés Permite establecer período de interés </div>
Huffpost	<ul style="list-style-type: none"> Sitio de noticias y blog Ofrece cobertura sobre un conjunto grande de secciones. Es internacional 	Gratis
Clasificación de noticias de diarios de circulación nacional mediante aprendizaje automático 2017-A02	<ul style="list-style-type: none"> Permite clasificar texto en un conjunto de secciones de noticia. Ocupa algoritmos de aprendizaje automático. 	No definido

Tabla 1. Resumen de productos similares.

Una diferencia importante con las dos primeras aplicaciones del trabajo propuesto es el hecho de clasificar la información de forma automática, y con la última referencia es el hecho de tener una plataforma web que recopila la información.

2. Objetivo

Crear un buscador web el cual permita visualizar un conjunto de noticias, mediante la implementación de algoritmos de clasificación los cuales pueden provenir de distintos sitios como blogs, foros, diarios,

Objetivos particulares

- Construir un sitio web para búsqueda de los distintos textos para su análisis y clasificación.
- Utilizar técnicas para la extracción de las noticias de los distintos sitios web.
- Obtener características de las noticias o de los textos recuperados para poder realizar así su clasificación.
- Implementar técnicas de aprendizaje para la clasificación de textos.

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 los acontecimientos más recientes, por ejemplo, en la televisión, blogs, redes sociales, foros, diarios, de los cuales podemos obtener las noticias 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.

Una solución sería tener un lugar el cual nos proporcione distintas fuentes de información para su consulta, lo permitiría el conocimiento de los hechos con una mayor veracidad.

Se propone crear un sitio web el cual permita extraer noticias de distintas fuentes de información, posteriormente clasificarlas y mostrarlas al usuario.

4. Productos o Resultados esperados

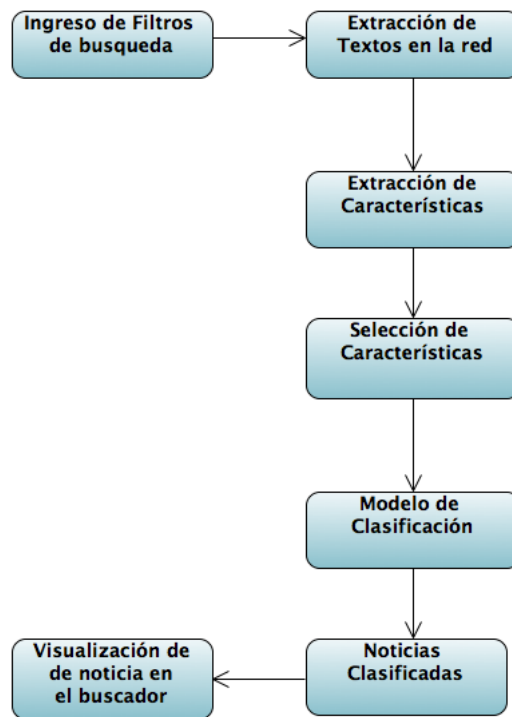
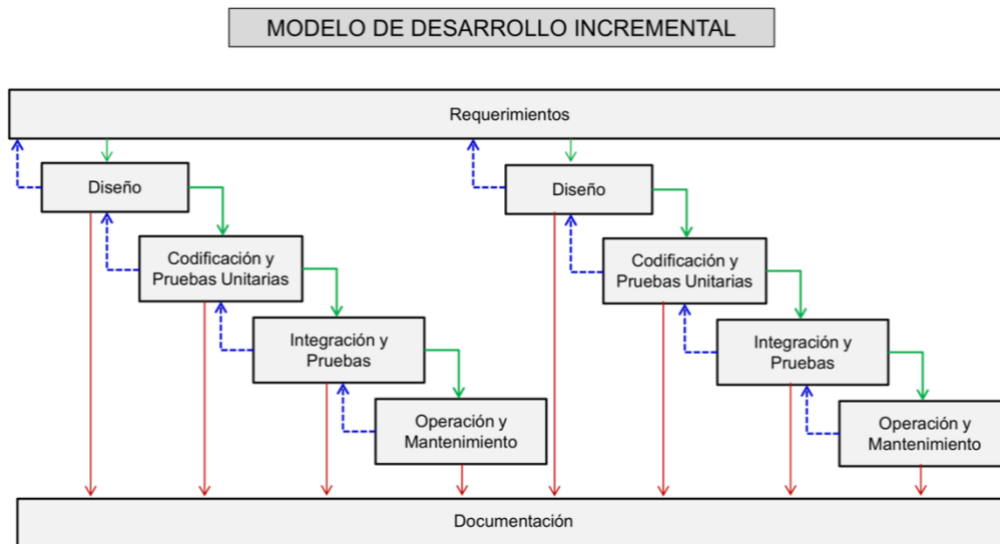


Figura 1. Arquitectura del sistema.

1. Documentación de análisis y diseño.
2. Buscador Web.
3. Algoritmos de clasificación automática.

5. Metodología

La metodología que utilizaremos para la realización del presente trabajo terminal será la incremental, 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.



6. Cronograma

Nombre del alumno(a): Hernandez Gómez Carlos Andres

Título del TT:

[illegible]

Nombre del alumno(a): Meza Martínez Luis Daniel

Título del TT:

Actividad	AGO	SEP	OCT	NOV	DIC	ENE	FEB	MAR	ABR	MAY
Análisis y diseño del sistema										
Evaluación de TT I.										
Generación del código.										
Pruebas.										
Reingeniería.										
Generación del Manual de Usuario y la Página web.										
Generación el Reporte Técnico.										
Presentar los resultados en congresos.										
Evaluación de TT II.										

7. Referencias

- [1] G. Baena Paz, S. Montero Olivares, Tesis en 30 días, Cd. Neza, EDOMEX, México, 1999.
[2] S. Mercado, ¿Cómo hacer una Tesis?, México, D.F., México, 1994.

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

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.

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 Tecnologías 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:_____

