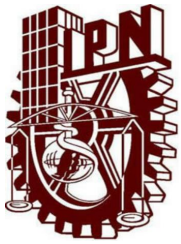

INSTITUTO POLITÉCNICO NACIONAL

ESCUELA SUPERIOR DE COMPUTO

TRABAJO TERMINAL

RECOLECTOR Y CLASIFICADOR DE NOTICIAS

2018-B013

PRESENTAN:

CARLOS ANDRES HERNANDEZ GOMEZ
LUIS DANIEL MEZA MARTINEZ

DIRECTORES:

M. en C. JOEL OMAR JUÁRES GAMBINO
Dra. CONSULO VARINIA GARCIA MENDOZA

CIUDAD DE MÉXICO

Índice general

Capítulo 1

Introducción

El objetivo de estudio de este trabajo es la clasificación de noticias implementando procesamiento de lenguaje natural y algoritmos de aprendizaje automático. El artículo periodistico es la información de un hecho ocurrido durante un lapso de tiempo determinado, permite conocer el estado económico de un país, logros de la ciencia, desastres naturales, grado dd inseguridad, dichos acontecimientos independientemente del tema, día y lugar en el cual ocurrieron tienen un impacto en la sociedad. Las caracteristicas principales de este tipo de artículo es depender de un medio de comunicación como la televisión, redes sociales, diarios, blogs, radio, entre otros, crean expectativas en los sectores económicos (Educación, industria, turismo, etc) esto a su vez modifica los planes de inversión de cual quier empresa o nación, siendo así de suma importancia su distribución de una forma eficaz [1].

En el siglo XXI la técnologia juega un papel importante en transmitir las noticias meidante páginas web, son accesibles gracias a una herramienta llamada crawler que las reúne para indexarlas y hacer más sencillo que los motores de búsqueda puedan recuperarlas [2]. En la actualidad las páginas web van incrementando día con día, por lo cual se pueden consultar noticias de distintos sitios, alguno de estos son los periódicos electrónicos, los cuales dividen sus artículos en secciones para facilitar la búsqueda del usuario, sin embargo, el nombre de las secciones no coincide en todos los periódicos a pesar de que el tipo de contenido sea el mismo. Existen un sinfín de sitios independientes en la red, que proveen 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 dichos sitios.

1.1. Problemática

Hoy en día existen distintas maneras de informarse acerca de los acontecimientos más recientes, por ejemplo, en la televisión, blogs, redes sociales, foros,

diarios, etc. Esto ha provocado que la información se encuentre más dispersa y se tenga que acceder a muchos recursos para recopilarla. Esta situación implica un gran esfuerzo y tiempo. 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 homogénea de la información.

Según El Economista [8] el sitio web "Animal Político" ocupa el lugar número cuatro en el ranking de medios nativos digitales y clasifica sus noticias de una manera poco habitual para los lectores, como la sección "El sabueso", "El plumaje", "Hablemos de ... ", entre otras, lo que hace complicado obtener los artículos para los métodos tradicionales de recopilación que se basan sólo en las etiquetas que identifican cada sección y no el contenido de las noticias. Debido a lo anterior se propone crear un recolector de noticias el cual permita recopilar noticias de distintas fuentes de información, y mediante el análisis automático de su contendido determine si este guarda relación con las secciones de interés del usuario y el periodo establecido. Finalmente, las noticias que satisfagan ambos filtros serán las que se le mostrarán al usuario.

1.2. Solución Propuesta

1.3. Objetivo

Crear un recolector de noticias, el cual permita recopilar 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 establecidos por el usuario.

1.4. Objetivos Específicos

- Desarrollar un recolector de noticias, el cual permita recopilar información de diferentes fuentes como diarios, sitios de noticias, blogs y 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 realizado en el trabajo terminal 2017-A02 para utilizarlo en el contexto de esta propuesta (filtro de sección)

1.5. Estructura del Documento

Capítulo 2

Conclusiones

Vestibulum faucibus nibh ac felis dapibus, ac efficitur ipsum fermentum. Nullam sapien ligula, varius sed neque nec, dictum scelerisque ipsum. Praesent pellentesque tristique lorem non cursus. In et urna lectus. Cras porttitor ipsum sed ullamcorper faucibus. Curabitur sapien turpis, vulputate sed enim a, feugiat aliquet eros. Donec ut dui a libero dapibus dictum ac eget justo. Duis tristique luctus diam, faucibus eleifend neque tincidunt at. Sed eget risus dolor. Quisque auctor tellus eget ipsum maximus, at sodales nisi maximus. Fusce nisi lectus, ornare sit amet mi et, fermentum vestibulum turpis. Nulla et rhoncus nulla. Suspendisse arcu ligula, mollis sed diam a, consequat sollicitudin ipsum. Donec eget aliquet nisi.

Capítulo 3

Bibliografía

3.1. Referencias

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

(2) Bernabeu Morón, N. (2013). La noticia y el reportaje. Espania, Ministerio de Educación de Espania, p.9.