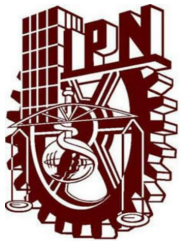

INSTITUTO POLITÉCNICO NACIONAL

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 clasificar noticias implementando procesamiento de lenguaje natural, algoritmos de aprendizaje automático y el manejo tecnologías web (Crawler). El artículo periodístico es la información de un hecho ocurrido durante un lapso determinado, permite conocer el estado económico de un país, avances en la ciencia, desastres naturales, nivel de inseguridad, dichos acontecimientos independientemente del tema, día y lugar en el cual ocurrieron, tienen un impacto en la sociedad. Las características principales de este tipo de información es depender de un medio de comunicación como la televisión, redes sociales, diarios, blogs, los cuales crean expectativas en los sectores económicos (Educación, industria, turismo, etc.), modifica los planes de inversión de las empresas o naciones, siendo así de suma importancia su distribución de una forma eficaz. Cabe mencionar que el medio más utilizado es la internet.

El uso de las páginas web está en incremento, permitiendo consultar noticias de distintos sitios como los periódicos electrónicos; su información al igual que un diario tradicional se encuentra dividida en secciones para facilitar la consulta, sin embargo, la clasificación suele variar en cada compañía de prensa, incluso con el mismo contenido, un problema mayor se encuentra en los sitios independientes, los cuales no cuentan con una segmentación particular, haciendo difícil realizar una búsqueda eficaz.

Se han seleccionado los diarios más utilizados en México [1], con una buena segmentación en su contenido y se ha homogenizado las secciones en común, para obtener los datos necesarios (Noticias clasificadas) y realizar el entrenamiento del algoritmo de clasificación.

1.1. Problemática

Hoy en día existen distintas maneras de informarse acerca de los acontecimientos más recientes, por ejemplo la televisión, 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. Justificación

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" (www.animalpolitico.com) 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.

1.3. Solución Propuesta

Se propone crear un sitio web, el cual permite recolecta noticias de la internet de forma automática; las cuales serán clasificadas de acuerdo a su contenido y posteriormente son mostradas al usuario. El sitio permite filtrar las noticias de acuerdo a su contenido y asu fecha publicación.

1.4. Objetivo

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

1.5. 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)

1.6. Estructura del Documento

Capítulo 2

Análisis y diseño

En este capítulo se describe el análisis y el diseño del trabajo propuesto para la recolección, clasificación de noticias y el entorno web.

2.1. Requisitos funcionales

- RF1: El sistema debe recolectar noticias de forma automática en la internet
- RF2: El sistema debe clasificar las noticias recolectadas deacuerdo a su contenido
- RF3: El sistema debe permitir filtrar las noticias deacuerdo a su fecha de publicación
- RF4: El sistema debe mostrar las noticias recolectadas y clasificadas al usuario en un entorno web
- RF5: Cada noticia mostrada debe contener un hipervínculo que rediriga al usuario a su sitio de origen
- RNF6:

2.2. Requisitos No funcionales

2.3. Estructura del Documento

- RNF1: La clasificación de una noticia no debe tardar mas de un segundo
- RNF2: Las noticias recolectads deberán tener un mínimo de 180 palabras en ellas
- RNF3: El sistema

2.4. Reglas de negocio

2.5. Estructura del Documento

- RDN1: La notica debe tener almenos 180 palabras
- RDN2: Las noticias deben estar redactadas en lenguaje español