

INSTITUTO POLITÉCNICO NACIONAL

---

ESCUELA SUPERIOR DE CÓMPUTO

*TRABAJO TERMINAL*

**Recolector y clasificador de  
noticias**

**2018-B013**

**PRESENTAN:**

CARLOS ANDRES HERNANDEZ GOMEZ  
LUÍS DANIEL MEZA MARTÍNEZ

**DIRECTORES:**

M. en C. JOEL OMAR JUÁREZ GAMBINO  
Dra. CONSUELO VARINIA GARCÍA MENDOZA

**CIUDAD DE MÉXICO**

2 de abril de 2019

---

---

# Índice general

<b>1. Introducción</b>	<b>1</b>
1.1. Problemática . . . . .	2
1.2. Justificación . . . . .	2
1.3. Solución Propuesta . . . . .	2
1.4. Objetivo . . . . .	3
1.5. Objetivos Específicos . . . . .	3
<b>2. Estado del arte</b>	<b>5</b>
2.1. Introducción . . . . .	5
2.2. Trabajos nacionales . . . . .	5
2.3. Trabajos internacionales . . . . .	6
2.4. Herramientas disponibles . . . . .	7
<b>3. Marco teórico</b>	<b>9</b>
3.1. Crawler . . . . .	9
3.2. Sitios web . . . . .	9
3.3. Página web . . . . .	9
3.4. Blog . . . . .	10
3.5. Foro . . . . .	10
3.6. Lenguaje . . . . .	10
3.7. Procesamiento de lenguaje natural . . . . .	10
3.8. Tokenización . . . . .	10
3.9. Lematización . . . . .	10
3.10. Representación del texto . . . . .	11
3.11. Corpus . . . . .	11
3.12. Algoritmos de clasificación . . . . .	11
3.12.1. Naive bayes . . . . .	11
3.12.2. Máquina de soporte vectorial . . . . .	11
<b>4. Análisis y diseño</b>	<b>13</b>
4.1. Requisitos funcionales . . . . .	13
4.2. Requisitos no funcionales . . . . .	14
4.3. Reglas de negocio . . . . .	14
4.4. Casos de uso . . . . .	21

4.4.1.	Diagrama de casos de uso . . . . .	21
4.4.2.	CU1 Seleccionar sección . . . . .	22
4.4.3.	CU2 Buscar noticias . . . . .	24
4.4.4.	CU3 Recolectar noticias . . . . .	27
4.4.5.	CU7 Ingrear a sitio web . . . . .	29
4.5.	Mensajes . . . . .	29
4.6.	Pantallas . . . . .	32
4.6.1.	UI1 Sitio . . . . .	32

# Capítulo 1

## Introducción

El objetivo de estudio de este trabajo es clasificar noticias utilizando un recolector (Crawler), en el cual se implementará procesamiento de lenguaje natural, así como algoritmos de aprendizaje automático. La noticia es la información de un hecho de interés ocurrido durante un periodo de tiempo determinado. 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, estos se comparten por distintos medios de comunicación, tales como la televisión, redes sociales, diarios, blogs y la radio; Nos permiten conocer la situación económica del país, logros de la ciencia, desastres naturales, la situación en cuestión de inseguridad y otros acontecimientos. En el ámbito de las inversiones, crean expectativas y eso a su vez puede modificar los planes de inversión en cualquier sector, siendo así de suma importancia compartirlas de una forma eficaz [2].

El uso de páginas web como medio de comunicación 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 portal, incluso teniendo 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 consultados en México, con una buena segmentación en su contenido y se han homogeneizado las secciones en común, para poder extraer la información necesaria de cada noticia y así poder llevar a cabo el entrenamiento del algoritmo de clasificación.

## 1.1. Problemática

La clasificación de textos es una tarea que se lleva a cabo de forma manual, lo cual representa un costo en términos del tiempo ocupado y el dinero invertido en la contratación del personal.

Los métodos tradicionales para la recopilación de información de los recolectores web (Crawler), están basados en las etiquetas o marcadores que los sitios añaden a su código fuente, por ejemplo, algunos artículos periodísticos son etiquetados a la sección que pertenecen (Política, ciencia, tecnología, etc). Sin embargo, existen muchas fuentes de información que no etiquetan sus publicaciones, incluso si la tarea es realizada, dicha segmentación no indica claramente el tipo de contenido; Al consultar los portales mas visitados en México (En el giro del periodismo) [3] se encuentra definida la sección deportes con varios sinónimos como *Universal deportes* (Diario (El universal)), *La afición* (Portal de *Milenio*), *Adrenalina* (**Excelsior**), etc. Como este ejemplo se encuentran mas; Las noticias son segmentadas de forma tan diversa que ha complicado su búsqueda en la internet.

## 1.2. Justificación

Hoy en día existen distintas maneras de informarse acerca de los acontecimientos más recientes, por ejemplo, la televisión, blogs, redes sociales, foros, diarios, etc. Esto ha provocado que la información se encuentre dispersa y se deba acceder a múltiples recursos para ser recopilada, implicando una inversión de tiempo y esfuerzo. Para ayudar en está problemática existen herramientas que hacen la búsqueda de noticias de interés para el usuario en forma automática. Sin embargo, dichas herramientas dependen de los sitios a consultar los cuales deben contener etiquetas definidas y homogéneas.

Según El Economista [3] el sitio web “Animal Político” ([www.animalpolitico.com](http://www.animalpolitico.com)) ocupa el lugar número cuatro en el ranking de medios nativos digitales, 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 con los métodos tradicionales de recopilación que, se basan sólo en las etiquetas que identifican cada sección y no el contenido de las noticias. Se propone crear un sitio web, el cual permite recolecta noticias de la internet de forma automática; las cuales serán clasificadas de acuerdo con su contenido y posteriormente serán mostradas al usuario. Cabe mencionar que el sitio permite filtrar las noticias de acuerdo a su contenido y a su fecha publicación.

## 1.3. Solución Propuesta

Se propone crear un portal web el cual recolecte y clasifique noticias de acuerdo a su contenido y periodo de difusión, con el uso de recopiladores web ( Crawler

) y la implementación de algoritmos de procesamiento de lenguaje natural. Finalmente, las noticias que satisfagan ambos filtros (Tipo de contenido y fecha de publicación) serán mostradas al usuario.

## 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)





## Capítulo 2

# Estado del arte

### 2.1. Introducción

A continuación, se mostrarán distintos trabajos nacionales e internacionales, así como herramientas las cuales desempeñan un trabajo similar al propuesto por nosotros.

### 2.2. Trabajos nacionales

#### **Clasificación Automática de Textos de Desastres Naturales en México**

En este trabajo se propone clasificar noticias del ámbito Desastres Naturales utilizando estrategias de reducción de dimensionalidad conocidas como umbral en la frecuencia y ganancia en la información, los métodos de clasificación utilizados fueron el clasificador simple de Bayes y vecinos más cercanos.

Se utilizaron 375 noticias del periódico Reforma como conjunto de entrenamiento, para posteriormente clasificarlas (relevantes e irrelevantes), de los cuales el 11.5 de noticias eran relevantes y el 88.5 % restante eran irrelevantes. Una vez obtenido el conjunto de noticias se procedió con un pre-procesamiento, el cual reducía el tamaño de los documentos, eliminando las partes de los textos que no se consideraban relevantes; posteriormente se realizó el indexado, el cual los documentos son representados por vectores de palabras en un espacio de dimensionalidad  $n$  en el cual se logró una reducción de dimensionalidad en donde finalmente se utilizaron técnicas de clasificación como el algoritmo simple de Bayes en el cual se obtuvo un resultado del 97 % de efectividad al clasificar noticias de desastres naturales [4].

#### **Clasificación de Texto Mediante Atributos Probabilísticos de Concurrencia de Palabras**

En este trabajo se propone clasificar artículos en dos categorías (relevantes y no relevantes), los cuales se encuentran almacenados en una base de datos llamada Regulon DB, la cual contiene más de 2,000 artículos clasificados manualmente relacionados a los genes de regulación de la bacteria *Escherichia coli*; con el objetivo de facilitar la búsqueda de información acerca de dicha bacteria. Proponen medir los atributos es decir las palabras de un documento, utilizando representaciones probabilísticas utilizadas en el entorno de análisis de información. Se utilizaron 1,823 documentos; en la etapa de clasificación se entrenaron a los clasificadores Bayes simple y Bayes multinominal; obteniendo mejores resultados con el método Bayes Naive [5].

### 2.3. Trabajos internacionales

#### Clasificación Automática de Textos Usando Redes de Palabras

En este trabajo se propone un algoritmo para la clasificación automática de textos basado en una representación y clasificación distinta utilizada en los algoritmos de clasificación supervisada, utilizando redes de palabras. Se utilizaron 1000 mensajes de texto de la plataforma Twitter, en el idioma español y correspondiente a distintos contextos, para posteriormente clasificar el tipo de contenido de los mensajes (positivos, negativos y neutrales), se definió un grafo como aquella red de palabras co-currentes construida a partir de un conjunto de textos clasificados; para su realización el primero de estos procesos es llevar distintas variantes de una misma palabra a su raíz, esto para reducir la variabilidad del lenguaje posteriormente se considera las palabras plurales (terminadas con 's' o 'es'). A estas se les elimina el sufijo para compararlas con su equivalente singular, realizando el cambio de manera automática. Los resultados mostraron que el clasificador presenta un 80 % cercanía respecto a la clasificación realizada por una persona; su nivel de desempeño fue mayor al obtenido con el algoritmo Naive Bayes [6].

#### Document Classification for Newspaper Articles

En este trabajo se enfocaron en clasificar artículos del MIT (Massachusetts Institute of Technology) de las siguientes categorías:

- Arts
- Features
- News
- Opinion
- Sports

- World

Para los cuales utilizaron el algoritmo de clasificación como el Naive Bayes ya que era uno de los clasificadores más simples y eficaces que otras técnicas de clasificación, de igual manera utilizaron la clasificación máxima de entropía el cual provee segmentación de texto, modelado de lenguaje. Se utilizaron un corpus 3000 artículos en total, siendo 500 artículos de cada sección mencionada, para el entrenamiento se utilizaron 120 artículos siendo 20 de cada sección y teniendo como resultado un 77 % de exactitud [7].

## **2.4. Herramientas disponibles**

Entre las herramientas utilizadas para el procesamiento de lenguaje natural y aprendizaje automático se encuentran:

### **Cloud Natural Language**

Google Cloud Natural Language [8] revela la estructura y el significado del texto con modelos potentes de aprendizaje automático previamente entrenados en una API de REST fácil de usar y con modelos personalizados se puede utilizar para extraer información sobre personas, lugares, eventos y muchos otros datos, que se mencionan en documentos de texto, artículos periodísticos o entradas de blog. También puedes utilizarla para comprender las opiniones sobre tus productos expresadas en los medios sociales o analizar la intención en las conversaciones de los clientes que se den en un centro de atención telefónica o una aplicación de mensajería.

### **GoogleBot**

El crawler más famoso del mundo es Googlebot, el software diseñado por Google para indexar el contenido nuevo o actualizado de Internet. Googlebot [9] no sólo tiene la capacidad de rastrear e indexar los sitios web de internet, sino que además puede extraer información de ficheros como pueden ser PDF, XLS, DOC, etc. Una vez el contenido está indexado, el servidor lo clasifica y establece un orden de relevancia para las distintas búsquedas que pueda efectuar un usuario, es decir, lo posiciona.

### **Watson Natural Language Classifier**

Watson NLC [10] aplica técnicas de computación cognitiva para analizar un texto y proporcionar la clase que mejor encaja entre un conjunto de clases predefinidas a partir de un texto corto. Al ser un clasificador, esta compuesto de ciertos pasos, en primera instancia se necesitan de clases las cuales son etiquetas que identificarán el texto analizado y será la salida proporcionada por el clasificador; posteriormente se debe tomar en cuenta que se necesita de una colección

#### 2.4. HERRAMIENTAS DISPONIBLES CAPÍTULO 2. ESTADO DEL ARTE

de textos, los cuales proporcionarán apoyo para que el clasificador logre identificar las clases ingresadas posteriormente teniendo todos estos datos se logra entrenar al clasificador, el cual proporcionará una salida dependiendo a los datos que fueron utilizados.

## Capítulo 3

# Marco teórico

En esta sección se expondrán de manera detallada conceptos los cuales son esenciales para la elaboración de este trabajo

### 3.1. Crawler

Un crawler [11] es una herramienta la cual analiza sitios web, permitiendo recolectar las páginas web para así posteriormente extraer la información que contengan. Un crawler también conocido como robot o spider, es un sistema para la descarga masiva de páginas web. Son uno de los componentes principales de los motores de búsqueda web, los sistemas que reúnen un conjunto de páginas web, las indexan y permiten a los usuarios realizar consultas contra el índice y encontrar las páginas web que coincidan con las consultas.

### 3.2. Sitios web

Un sitio web es un conjunto de páginas web

### 3.3. Página web

Una página web es un documento electrónico el cual forma parte de la WWW (*World Wide Web*) generalmente construido en el lenguaje HTML (Hyper Text Markup Language). Este documento puede contener enlaces que nos direcciona a otra página web. Para visualizar una página web es necesario de un browser o un navegador[12]. Dentro de las páginas web podemos encontrar un sinfin de sitios los cuales pueden ser de nuestro interés.

### 3.4. Blog

Un blog es una página web en la cual el usuario no necesita conocimientos específicos del medio electrónico ni del formato digital para poder aportar contenidos de forma inmediata, ágil y constante desde cualquier punto de conexión a Internet [13]. En un blog el usuario puede compartir cualquier tipo de información que sea de su agrado, teniendo una mayor libertad de expresión lo cual permite que otras personas compartan y comenten su manera de expresarse.

### 3.5. Foro

Un foro es una herramienta de comunicación asíncrona. Los foros permiten la comunicación de los participantes desde cualquier lugar en el que esté disponible una conexión a Internet sin que éstos tengan que estar dentro del sistema al mismo tiempo, de ahí su naturaleza asíncrona [14]. Brindando una mayor interacción entre distintos participantes y permitiendo conocer la opinión sobre un tema de distintas personas.

### 3.6. Lenguaje

El lenguaje es un medio de comunicación a través de un sistema de símbolos[15]. La Real Academia Española define al lenguaje como la facultad del ser humano de expresarse y comunicarse con los demás a través del sonido articulado o de otros sistemas de signos.

### 3.7. Procesamiento de lenguaje natural

El procesamiento de lenguaje natural es una disciplina de la Inteligencia Artificial que se ocupa de la formulación e investigación de mecanismos computacionales para la comunicación entre personas y máquinas mediante el uso de Lenguajes Naturales[16].

### 3.8. Tokenización

Es la acción de separar el texto en sus unidades mínimas (Palabras), se les asigna un código como el ASCII o hexadecimal para ser reconocidas de forma única, son almacenadas para su posterior análisis y reconocimiento. Cabe mencionar que los signos de puntuación son eliminados.

### 3.9. Lematización

Es el proceso lingüístico que, dada una palabra flexionada se encuentra su lema. Una palabra flexionada es cuando está en el plural, en femenino conjugada,

diminutivo o en superlativo. El lema es la palabra que esta en singular para sustantivo, singular masculino para adjetivo e infinitivo para un verbo. Ejemplo:

- amigos, amiga, amiguitos->Amigo
- soy, son, es->Ser

Cabe mencionar que existen diversos grados de lematización

- Morfológica: Es la anterior mente explicada
- Sintáctica: Toma en cuenta el contexto donde se encuentra la palabra

Una opción para lematizar es Freeling, este es un lematizador hecho por la universidad de cataluña.

### **3.10. Representación del texto**

Los métodos de aprendizaje automático, requieren que la información este representada en un formato que facilite su procesamiento. Un método utilizado es representar los datos en un vector de valores numéricos.

### **3.11. Corpus**

Se le llama corpus a la recopilación de un conjunto de textos, de materiales escritos y/o hablados, agrupados bajo un conjunto de criterios mínimos, para realizar ciertos análisis lingüísticos.

### **3.12. Algoritmos de clasificación**

#### **3.12.1. Naive bayes**

Es una aproximación probabilística, las cuales hacen especulaciones sobre como deben ser generados los datos. Generalmente utiliza aprendizaje supervisado sobre el conjunto de entrenamiento para estimar sus parámetros. Con el conjunto de entrada se aplica el teorema de bayes.

#### **3.12.2. Máquina de soporte vectorial**

Las máquinas de soporte vectorial son sistemas de aprendizaje los cuales se basan en el uso de un espacio de funciones lineales, el cual se encuentra con mas dimensiones inducido por un kernel, en el que las hipotesis son las entradas para el algoritmo. El algoritmo induce separadores lineales ya sea en el espacio original de los ejemplos de entrada, si los datos no son separables se busca un hiperplano en el que si lo sean, se hace de forma implícita con las funciones kernel.





## Capítulo 4

# 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.

### 4.1. Requisitos funcionales

#### RF1 Recolectar noticias

---



- **Descripción:** El sistema debe recolectar noticias de forma automática en la internet

#### RF2 Clasificar noticias

---



- **Descripción:** El sistema debe clasificar las noticias recolectadas de acuerdo a su contenido

#### RF3 Filtrar por fecha

---



- **Descripción:** El sistema debe permitir filtrar las noticias de acuerdo a su fecha de publicación

#### RF4 Entorno web

---



- **Descripción:** El sistema debe mostrar las noticias recolectadas y clasificadas al usuario en un entorno web

#### RF5 Link a noticia



- **Descripción:** Cada noticia mostrada debe contener un hipervínculo que redirija al usuario a su sitio de origen

## 4.2. Requisitos no funcionales

#### RNF1 Tiempo de clasificación



- **Descripción:** La clasificación de una noticia no debe tardar mas de un segundo

#### RNF2 Número de palabras



- **Descripción:** Las noticias recolectads deberán tener un mínimo de 180 palabras en ellas

#### RNF3 Número de noticias mostradas



- **Descripción:** En el sitio web, se den visualizar almenos 15 noticias

#### RNF4 Tiempo de actualización



- **Descripción:** El tiempo para mostrar las 15 noticias clasificadas no debe exceder los 3 segundos

## 4.3. Reglas de negocio

En esta sección se describen las reglas de negocio implementadas en el trabajo propuesto.

### RN1 Número de palabras

---



- **Tipo:**
- **Descripción:** La noticia debe tener al menos 180 palabras
- **Ejemplo:**
- **Refer:**

### RN2 Lenguaje de direcciones web

---



- **Tipo:**
- **Descripción:** Las direcciones de los sitios a consultar deben estar redactadas en lenguaje español.
- **Referenciado por:** CU1 [Seleccionar sección](#)

### RN3 Lenguaje de noticias

---



- **Tipo:**
- **Descripción:** Las noticias deben estar redactadas en lenguaje español mexicano.
- **Ejemplo:**
- **Referenciado por:**

### RN4 Restricción en la recolección

---



- **Tipo:**
- **Descripción:** Solo se puede recolectar información de los sitios que lo permitan.
- **Ejemplo:**
- **Referenciado por:**

### RN5 Porcentaje de aceptación



- **Tipo:**
- **Descripción:** Solo se puede mostrar una noticia si cumple con un porcentaje de aceptación mayor a 60 %.
- **Ejemplo:**
- **Referenciado por:**

### RN6 Formato de fecha



- **Tipo:** Derivación.
- **Descripción:** El formato de fecha se define como:

$$F = D/M/A$$

donde

$$D = \{x/x \in N, 1 \leq x \leq 31\}$$

$$M = \{y/y \in N, 1 \leq y \leq 12\}$$

$$A = \{z/z \in N, 1990 \leq z \leq \Lambda_i\}$$

$F : Fecha$

$\Lambda_i : Anio\_actual$

- **Ejemplo:**
- **Referenciado por:** [CU2 Buscar noticias](#)

### RN7 Perido preestablecido



- **Tipo:** Cálculo
- **Descripción:** El formato de fecha es el descrito en la [RN6 Formato de fecha](#) ; El periodo preestablecido se define de la siguiente forma.

**Fecha fin:** Toma el valor de la fecha actual.

**Fecha inicio:** Es colocada 5 día antes de la fecha actual; Se muestra la forma completa para el cálculo del día, mes y año:

*Sea*

$D_a$  : *Día\_actual*

$M_a$  : *Mes\_actual*

$A_a$  : *Año\_actual*

$F_i$  : *Fecha\_inicio*

*mod* : *Operacion modulo*

$\Psi(M_a)$  : *Funcion dias de mes*

$$\xi = \frac{D_a - 5}{|D_a - 5|}$$

$$\delta = \frac{\xi(|\xi - 1|)}{2}$$

La fecha toma el siguiente valor:

$$F_i = D_c / M_c / A_c$$

Existen 4 casos para calcular el día, mes y años; Dependen del mes y el día actual:

Caso:	1	2
<b>Restricción:</b>	$2 \leq M_a \leq 12; D_a \neq 5$	$2 \leq M_a \leq 12; D_a = 5$
$D_c$ :	$(D_a - 5) \pmod{\Psi(M_a - 1) + 1}$	$\Psi(M_a - 1)$
$M_c$ :	$M_a + \delta$	$M_a - 1$
$A_c$ :	$A_a$	$A_a$

Caso:	3	4
<b>Restricción:</b>	$M_a = 1; D_a \neq 5$	$M_a = 1; D_a = 5$
$D_c$ :	$(D_a - 5) \pmod{32}$	31
$M_c$ :	$\xi \pmod{13}$	12
$A_c$ :	$A_a + \delta$	$A_a - 1$

- **Ejemplo:**
- **Referenciado por:** [CU1 Seleccionar sección](#), [CU2 Buscar noticias](#)

## RN8 Periodo válido



- **Tipo:** Derivación.
- **Descripción:**  
*Sea*

$F_i : Fecha\_inicio$

$F_f : Fecha\_fin$

$T = \{Valido, Invalido\}$

$\psi \in T$

Un periodo de fecha se defino como:

$(\psi = Valido) \leftrightarrow (F_i \leq F_f)$

$(\psi = Invalido) \leftrightarrow (F_i > F_f)$

- **Ejemplo:**
- **Referenciado por:** CU2 Buscar noticias

## RN9 Límite de periodo



- **Tipo:** Derivación.
- **Descripción:**

$Sea$

$F_i : Fecha\_inicio$

$F_f : Fecha\_fin$

$F_a : Fecha\_actual$

$F_c : 01/01/1990$

$T = \{Valido, Invalido\}$

$\psi \in T$

$\Phi = [F_c, F_a]$

$I = [F_i, F_f]$

Un intervalo de tiempo dentro de los limites del sistema se define como:

$(\psi = Valido) \leftrightarrow (I \subseteq \Phi)$

$(\psi = Invalido) \leftrightarrow (I \not\subseteq \Phi)$

- **Referenciado por:** CU2 Buscar noticias

### RN10 Campos obligatorios



- **Tipo:** Restricción.
- **Descripción:** Los campos marcados con no se deben omitir.
- **Ejemplo:**
- **Referenciado por:**

### RN11 Sitios restringidos



- **Tipo:**m
- **Descripción:** No se debe acceder a las siguientes páginas:
  - Facebook
  - Youtube
  - Twitter
  - Instagram
- **Ejemplo:**
- **Referenciado por:** CU3 Recolectar noticias

### RN12 Orden de publicación



- **Tipo:** Descripción
- **Descripción:** Las noticias se mostrarán de forma descendente de acuerdo a su fecha de difusión; Es decir la primera publicación en mostrarse es aquella que tiene la fecha y hora mas cercana a la del sistema.
- **Ejemplo:**
- **Referenciado por:** CU2 Buscar noticias

### RN13 Profundidad de búsqueda



- **Tipo:** Cálculo.
- **Descripción:** Se muestra la forma correcta de realizar el cálculo para obtener la profundidad de búsqueda asociada al número de noticias por recolectar, y la cantidad de hiperbínculos que cada página contiene.

*Sea*

$$\Lambda(\mu, \psi) = \log_{\psi} (\mu(\psi - 1) + \psi) - 1$$

*donde*

$\psi$  : Numero de URL por pagina

$\mu$  : Numero de noticias recolectadas

$\Lambda$  : Profundidad de busqueda

- **Ejemplo:**
- **Referenciado por:** [CU3 Recolectar noticias](#)

## RN14 Número de peticiones



- **Tipo:** Restricción
- **Descripción:** El número de peticiones realizadas a una página no debe exceder el permitido por el dominio.
- **Ejemplo:**
- **Referenciado por:** [CU3 Recolectar noticias](#)



## 4.4. Casos de uso

### 4.4.1. Diagrama de casos de uso

La figura 4.1 muestra el diagrama de casos de uso implementado en el sistema.

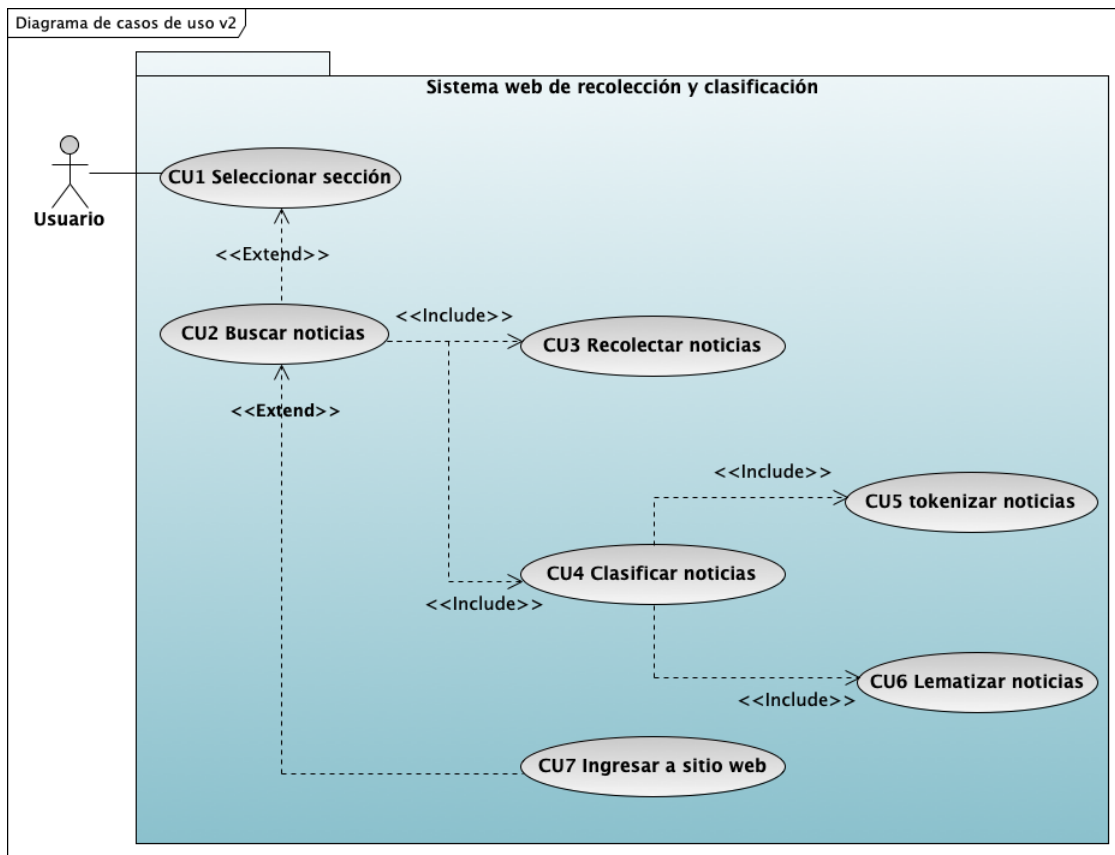


Figura 4.1: Diagrama de casos de uso

#### 4.4.2. CU1 Seleccionar sección


##### Resumen







Brinda al usuario un punto de acceso para elegir una sección; Las clasificaciones son, **Política**, **Deportes**, **Ciencia** y **Economía** en cada una se podrá consultar noticias, artículos y publicaciones dentro de un intervalo de tiempo específico; La fuente de información será un diccionario de los sitios web mas consultados y confiables en México. Cabe mencionar que el intervalo de tiempo por defecto es anterior en cinco dias de la fecha que se ha ingresa al portal.

##### Descripción

Caso de uso:	CU1 Seleccionar sección
<b>Actor:</b>	Usuario
<b>Propósito:</b>	Proporcionar una herramienta para acceder a los diferentes tipos de clasificaciones disponibles.
<b>Entradas:</b>	Ninguna.
<b>Salidas:</b>	<ul style="list-style-type: none"> <li>• <i>Fecha inicio</i></li> <li>• <i>Fecha fin</i></li> <li>• <a href="#">MSG1 Catálogo vacío</a></li> <li>• <a href="#">MSG2 Lenguaje de sitio</a></li> </ul>
<b>Precondición:</b>	El catálogo <b>Direcciones web</b> debe estar poblado.
<b>Postcondiciones:</b>	<ul style="list-style-type: none"> <li>• El usuario tendrá la facultad de buscar noticias de la sección elegida</li> <li>• El usuario tendrá la facultad de establecer un intervalo de tiempo para buscar los artículos</li> </ul>
<b>Reglas de negocio:</b>	<ul style="list-style-type: none"> <li>• <a href="#">RN2 Lenguaje de direcciones web</a></li> <li>• <a href="#">RN7 Período preestablecido</a></li> </ul>
<b>Errores:</b>	<ul style="list-style-type: none"> <li>• <b>Uno:</b> Cuando el catálogo <b>Direcciones web</b> no contiene información se muestra el mensaje <a href="#">MSG1 Catálogo vacío</a>, fin del caso de uso</li> <li>• <b>Dos:</b> Cuando los sitios proporcionados no se encuentran redactados en lenguaje español se muestra el mensaje <a href="#">MSG2 Lenguaje de sitio</a>, fin del caso de uso</li> </ul>
<b>Autor:</b>	Carlos Andres Hernandez Gomez

##### Trayectoria principal

1.  Selecciona una opción de la pantalla [UI1 Inicio](#); **Política**, **Economía**, **Deportes** o **Ciencia**.

2.  Obtiene el catálogo **Direcciones web**.
3.  Verifica que el catálogo **Direcciones web** contenga información. [\[Error Uno\]](#)
4.  Verifica que al menos un sitio cumpla con la regla de negocio [RN2 Lenguaje de direcciones web](#) . [\[Error Dos\]](#)
5.  Obtiene la fecha actual.
6.  Calcula el campo **Fecha inicio** y **Fecha fin** de acuerdo a la regla de negocio [RN7 Perido preestablecido](#) .
7.  Muestra deshabilitado y con fechas los campos **Fecha inicio** y **Fecha fin** con lo antes calculado, como se visualiza en la pantalla [UI2 Sección política](#).
8. - - - *Fin del caso de uso.*

**Causa de la extensión:** El actor desea consultar las noticias de una sección.

**Región de la trayectoria:** Paso 3 de la trayectoria principal.

**Extiende a :** [CU2 Buscar noticia](#).

#### 4.4.3. CU2 Buscar noticias

##### Resumen












Permite al actor realizar una búsqueda de noticias correspondiente a la sección elegida, ya sea **Política**, **Deportes**, **Ciencia** o **Economía**; La consulta se realiza en un periodo preestablecido el cual data 5 días antes a la fecha actual o se ingresar un intervalo de tiempo diferente. El sitio muestra 15 noticias ordenadas de forma descendente de acuerdo a la fecha de difusión; Cada artículo contiene el **Título**, la **Fecha de publicación**, un **Link** el cual direcciona a la página fuente y de contar con ello un **Resumen de la información**.

##### Descripción

Caso de uso:	CU2 Buscar noticias
<b>Actor:</b>	Usuario
<b>Propósito:</b>	Brindar una herramienta que permita recolectar y clasificar noticias de una sección específica en un periodo de tiempo.
<b>Entradas:</b>	<ul style="list-style-type: none"> <li>• <i>Fecha inicio</i></li> <li>• <i>Fecha fin</i></li> </ul>
<b>Salidas:</b>	Noticias clasificadas; De cada una se muestra: <ul style="list-style-type: none"> <li>• <b>Título</b></li> <li>• <b>Nombre de la página fuente</b></li> <li>• <b>Link al artículo</b></li> <li>• <b>Fecha de difusión</b></li> <li>• <b>Resumen</b></li> <li>• MSG3 Faltan campos obligatorios</li> <li>• MSG4 Formato de fecha inválido</li> <li>• MSG5 Periodo no válido</li> <li>• MSG6 Límites fuera de rango</li> </ul>
<b>Precondición:</b>	Una sección debe estar seleccionada.
<b>Postcondiciones:</b>	<ul style="list-style-type: none"> <li>• El usuario tendrá la facultad de consultar las noticias</li> <li>• El usuario tendrá la facultad de acceder a los sitios web de las noticias recolectadas</li> </ul>
<b>Reglas de negocio:</b>	<ul style="list-style-type: none"> <li>• RN6 Formato de fecha</li> <li>• RN7 Periodo preestablecido</li> <li>• RN8 Periodo válido</li> <li>• RN9 Límite de periodo</li> <li>• RN10 Campos obligatorios</li> <li>• RN12 Orden de publicación</li> </ul>
<b>Errores:</b>	<ul style="list-style-type: none"> <li>• <b>Uno:</b> Cuando hay campos vacíos marcados como obligatorios se muestra el mensaje MSG3 Faltan campos obligatorios y continua en el paso 2 de la trayectoria alternativa A</li> </ul>






Caso de uso:	CU2 Buscar noticias
<b>Errores:</b>	<ul style="list-style-type: none"> <li>• <b>Dos:</b> Cuando el usuario ha ingresado una fecha no válido se muestra el mensaje <a href="#">MSG4 Formato de fecha inválido</a> y continua en el paso 2 de la trayectoria alternativa A</li> <li>• <b>Tres:</b> Cuando existe incongruencias en los periodos de la fecha se muestra el mensaje <a href="#">MSG5 Periodo no válido</a> y continua en el paso 2 de la trayectoria alternativa A</li> <li>• <b>Cuatro:</b> Cuando el usuario ha introducido una intervalo de tiempo fuera de los límites se muestra el mensaje <a href="#">MSG6 Límites fuera de rango</a> y continua en el paso 2 de la trayectoria alternativa A</li> </ul>
<b>Autor:</b>	Carlos Andres Hernandez Gomez

### Trayectoria principal

1.  Da click en el botón **Buscar** de la pantalla [UI2 Sección política](#). [[Trayectoria A](#)] [[Trayectoria B](#)]
2.  Verifica la regla de negocio [RN10 Campos obligatorios](#) . [[Error Uno](#)]
3.  Verifica la regla de negocio [RN6 Formato de fecha](#) . [[Error Dos](#)]
4.  Verifica la regla de negocio [RN8 Periodo válido](#) . [[Error Tres](#)]
5.  Verifica la regla de negocio [RN9 Límite de periodo](#) . [[Error Cuatro](#)]
6.  Incluye el caso de uso [CU3 Recolectar noticias](#).
7.  Incluye el caso de uso [CU4 Calasificar noticias](#).
8.  Obtiene de cada noticia clasificada en el paso 7 de la trayectoria principal el **Título**, **Nombre de la página fuente**, **Link al artículo**, **Fecha de difusión** y de contar con ello el **Resumen**.
9.  Ordena de forma descendente las noticias clasificadas del paso 7 de la trayectoria principal de acuerdo a su fecha de difusión.
10.  Muestra 15 noticias de las ordenadas, de acuerdo a la regla de negocio [RN12 Orden de publicación](#) con la información obtenida en el paso 8 de la trayectoria principal, como se visualiza en la pantalla [UI3 Resultados de búsqueda](#)
11.  Consulta la información.
12. - - - *Fin del caso de uso.*





**Trayectoria alternativa A:**

**Condición:** *El botón **Cambiar periodo** es presionado cuando está en estado Off*

- A-1.  Habilita y limpia el campo **Fecha inicio** y **Fecha fin**, como se muestra en la pantalla [UI2.1 Cambio de periodo](#).
- A-2.  Ingresa el campo **Fecha inicio**.
- A-3.  Ingresa el campo **Fecha fin**.
- A-4.  Da click en el botón **Buscar** de la pantalla [UI2.1 Cambio de periodo](#).
- A-5.  Continúa en el paso 2 de la trayectoria principal.
- A-6. - - - *Fin de la trayectoria.*

**Trayectoria alternativa B:**

**Condición:** *El botón **Cambiar periodo** es presionado cuando está en estado On*

- B-1.  Obtiene la fecha actual.
- B-2.  Calcula el campo **Fecha inicio** y **Fecha fin** de acuerdo a la regla de negocio [RN7 Perido preestablecido](#).
- B-3.  Muestra deshabilitado y con fechas los campos **Fecha inicio** y **Fecha fin** con lo antes calculado, como se visualiza en la pantalla [UI2 Sección política](#).
- B-4.  Continúa en el paso 1 de la trayectoria principal.
- B-5. - - - *Fin de la trayectoria.*

**Puntos de extensión**

**Causa de la extensión:** El usuario desea leer la noticia completa

**Región de la trayectoria:** 11

**Extiende a :** [CU7 Ingresar a sitio web](#)

#### 4.4.4. CU3 Recolectar noticias




##### Resumen

Brinda un servicio para recolectar noticias en la web (Crawler); Toma como punto de partida los portales registrados en el diccionario **URL's**. Se crea un proceso de trabajo independiente por cada liga en el diccionario, para simular un ambiente de extracción en paralelo; De cada sitio se recolecta las noticias publicadas; De cada artículo se obtiene **Fecha de publicación**, **Título**, **Contenido**, **URL de la noticia**, **Nombre de la página fuente** y de contar con ello el **Resumen**. Cabe destacar que las ligas contenidas en los sitios visitados son extraídas para su posterior análisis.

##### Descripción



Caso de uso:	CU3 Recolectar noticias
Actor:	Usuario
Propósito:	Brindar una herramienta de recolección del ciber espacio(Crawler).
Entradas:	URL de las paginas por consultar
Salidas:	<ul style="list-style-type: none"> <li>• Noticias; De cada una se obtiene               <ul style="list-style-type: none"> <li>◦ <b>Fecha de publicación</b></li> <li>◦ <b>Título</b></li> <li>◦ <b>Contenido</b></li> <li>◦ <b>Resumen</b></li> <li>◦ <b>URL de la noticia</b></li> <li>◦ <b>Nombre de la página fuente</b></li> </ul> </li> <li>• <a href="#">MSG7 Petición vacía</a></li> <li>• <a href="#">MSG9 Fallo en la recolección</a></li> </ul>
Precondición:	EL <b>Diccionario de URL'S</b> debe contener los vínculos de los sitios a consultar
Postcondiciones:	<ul style="list-style-type: none"> <li>• Las noticias recolectadas serán filtradas por su fecha de difusión</li> <li>• Las noticias recolectadas serán clasificadas</li> </ul>
Reglas de negocio:	<ul style="list-style-type: none"> <li>• <a href="#">RN11 Sitios restringidos</a></li> <li>• <a href="#">RN13 Profundidad de búsqueda</a></li> <li>• <a href="#">RN14 Número de peticiones</a></li> </ul>
Errores:	<ul style="list-style-type: none"> <li>• <b>Uno:</b> Cuando no se ha encontrado noticias en el periodo establecido se muestra el mensaje <a href="#">MSG7 Petición vacía</a>, fin del caso de uso</li> <li>• <b>Dos:</b> Cuando no se puede extraer información de los sitios brindados, se muestra el mensaje <a href="#">MSG9 Fallo en la recolección</a>, fin del caso de uso</li> </ul>
Autor:	Carlos Andres Hernandez, Luis Daniel Meza

### Trayectoria principal

1.  Solicita realizar una búsqueda de noticias con el botón **Buscar** de la pantalla [UI2 Sección política](#).
2.  Obtiene los vínculos de los sitios web registrados en el diccionario **URL's**.
3. 
4. - - - - *Fin del caso de uso.*



### Trayectoria alternativa A:

**Condición:** *Se escribe la condición*

- A-1.  lorem ipsum
- A-2.  lorem ipsum
- A-3. - - - - *Fin de la trayectoria.*

### Trayectoria alternativa B:

**Condición:** *Se escribe la condición*

- B-1.  lorem ipsum
- B-2.  lorem ipsum
- B-3. - - - - *Fin de la trayectoria.*

### Puntos de extensión

**Causa de la extensión:** Lorem ipsum

**Región de la trayectoria:** Lorem ipsum

**Extiende a :** Lorem ipsum

**Causa de la extensión:** Lorem ipsum

**Región de la trayectoria:** Lorem ipsum

**Extiende a :** Lorem ipsum



#### 4.4.5. CU7 Ingrear a sitio web




##### Resumen

Permite al usuario acceder al portal web de las noticias mostradas.

##### Descripción

<b>Caso de uso:</b>	CU7 Ingresar a sitio web
<b>Actor:</b>	Usuario.
<b>Propósito:</b>	Brindar un punto de acceso a los sitios web que han proporcionado la información.
<b>Entradas:</b>	URL:Es seleccionado con el mouse.
<b>Salidas:</b>	<ul style="list-style-type: none"> <li>• Sitio web de la URL seleccionada</li> <li>• <a href="#">MSG8 Fallo al ingresar a sitio web</a></li> </ul>
<b>Precondición:</b>	Debe existir almenos una noticia mostrada en el portal.
<b>Postcondiciones:</b>	El usuario tendra la facultad de visualizar la noticia completa.
<b>Reglas de negocio:</b>	Ninguna.
<b>Errores:</b>	<b>Uno:</b> Cuando no es posible ingresar al sitio web seleccionado, se muestra el mensaje <a href="#">MSG8 Fallo al ingresar a sitio web</a> en la pantalla <a href="#">UI3 Resultados de búsqueda</a> , fin del caso de uso.
<b>Autor:</b>	Carlos Andres Hernandez Gomez

##### Trayectoria principal

1.  Solicita ver la noticia completa dando click en la **URL** de la publicación deseada, en la pantalla [UI3 Resultados de búsqueda](#).
2.  Abre una ventana en el navegador y en ella se direcciona a la URL seleccionada. [\[Error Uno\]](#)
3.  Muestra la pantalla [UI4 Página de sitio web](#)
4. - - - *Fin del caso de uso.*

## 4.5. Mensajes

### MSG1 Catálogo vacio



- **Tipo:** Error.
- **Objetivo:** Dar a conocer que no se tiene las l gas a los sitios web.
- **Redacci n:** El cat logo **Direcciones web** se encuentra vac o.
- **Referenciado por:** CU1 Seleccionar secci n

### MSG2 Lenguaje de sitio



- **Tipo:** Error.
- **Objetivo:** Dar a conocer que los sitios a los cuales desea ingresar, no est n redactados en lenguaje espa ol.
- **Redacci n:** Los sitios no se encuentran en lenguaje espa ol, por lo cual no ser n consultados.
- **Referenciado por:** CU1 Seleccionar secci n

### MSG3 Faltan campos obligatorios



- **Tipo:** Error.
- **Objetivo:** Dar a conocer que hay campos obligatorios vac os.
- **Redacci n:** Los campos marcados con \* no pueden omitirse.
- **Referenciado por:** CU2 Buscar noticias

### MSG4 Formato de fecha inv lido



- **Tipo:** Error.
- **Objetivo:** Informar que se ha ingrsado una fecha no v lida.
- **Redacci n:** Se ha ingresado una fecha inv lida; El formato correcto es DD/MM/AAAA.
- **Referenciado por:** CU2 Buscar noticias

### MSG5 Periodo no v lido



- **Tipo:** Error.
- **Objetivo:** Informar que se ha ingresado una un periodo incongruente.
- **Redacción:** El periodo ingresado es incorrecto.
- **Referenciado por:** CU2 [Buscar noticias](#)

### MSG6 Límites fuera de rango



- **Tipo:** Error.
- **Objetivo:** Informar que el intervalo de tiempo está fuera de los límites del sistema.
- **Redacción:** El periodo ingresado está fuera de los límites permitidos, la fecha debe estar entr 01/01/1990 y el día en curso.
- **Referenciado por:** CU2 [Buscar noticias](#)

### MSG7 Petición vacía



- **Tipo:** Error.
- **Objetivo:** Informar al usuario que no se ha encontrado resultados en el tiempo de busca permitido.
- **Redacción:** El tiempo de búsqueda máximo se ha alcanzado y no se ha encontrado infomración en el periodo ingresado.
- **Referenciado por:** CU3 [Recolectar noticias](#)

### MSG8 Fallo al ingresar a sitio web



- **Tipo:** Error.
- **Objetivo:** Informar al usuario sobre el fallo ocurrido en la conexión del portal web seleccionado.
- **Redacción:** Ha ocurrido un fallo en la conexión con el sitio seleccionado, intente mas tarde.
- **Referenciado por:** CU7 [Ingresar a sitio web](#)

## MSG9 Fallo en la recolección



- **Tipo:** Error.
- **Objetivo:** Informar al usuario que las ligas registradas en el diccionario de **URL's** no permiten extraer información.
- **Redacción:** No se puede extraer noticias de los sitios registrados en este portal web.
- **Referenciado por:** [CU3 Recolectar noticias](#)

## 4.6. Pantallas

### 4.6.1. UI1 Sitio

#### Objetivo

Texto

#### Descripción

Texto

#### Comandos

- Lorem ipsum
- Lorem ipsum
- Lorem ipsum

#### Referencia

[CU1 Mostrar noticias](#), [CU7 Ingresar a sitio web](#)



Figura 4.2: Pantalla UI1 Inicio

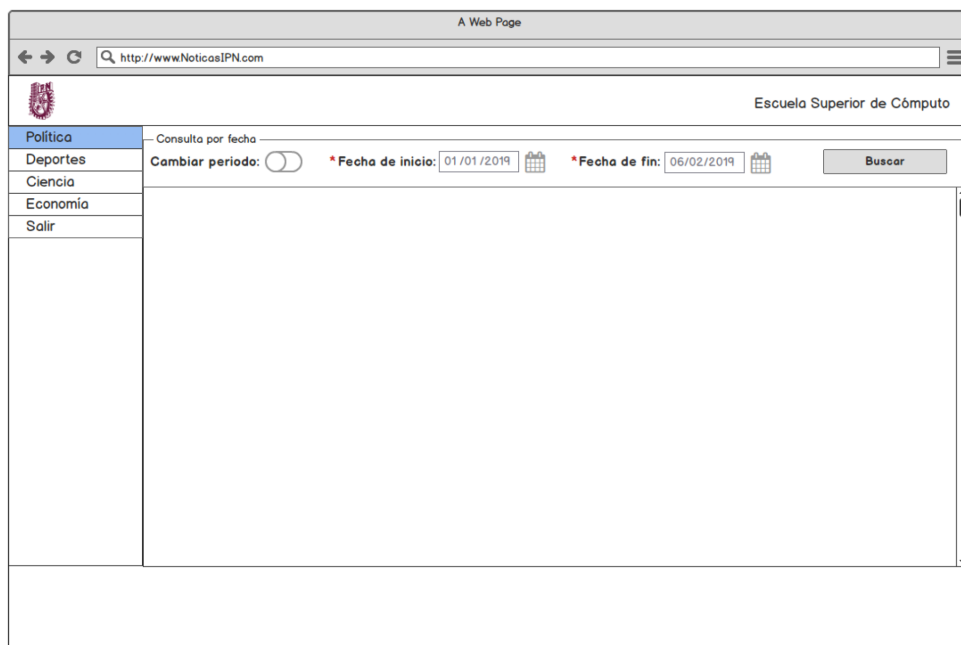


Figura 4.3: Pantalla UI2 Sección política

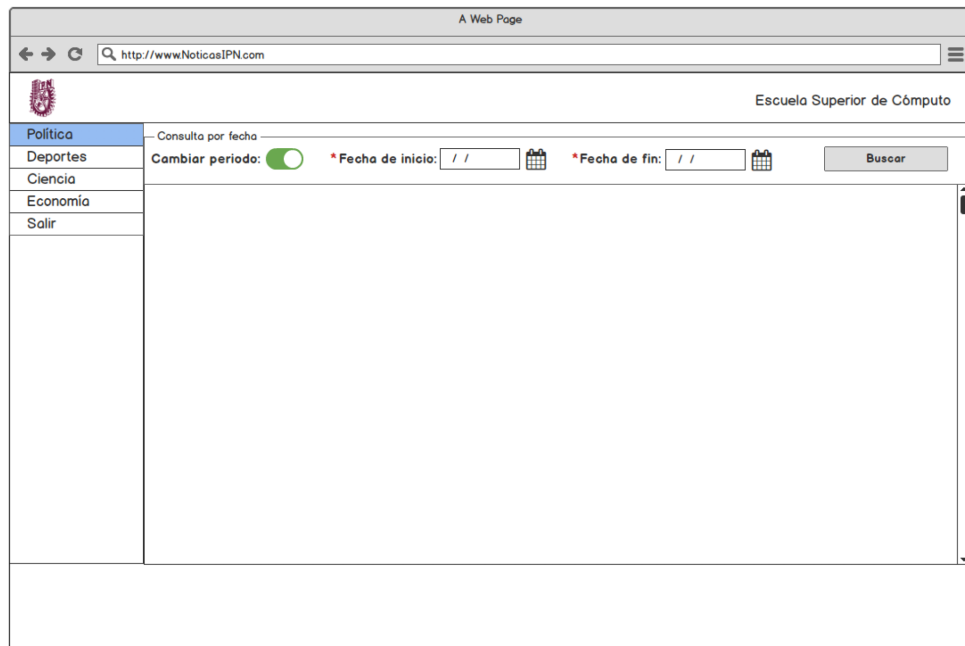


Figura 4.4: Pantalla IU2.1 Cambio de periodo

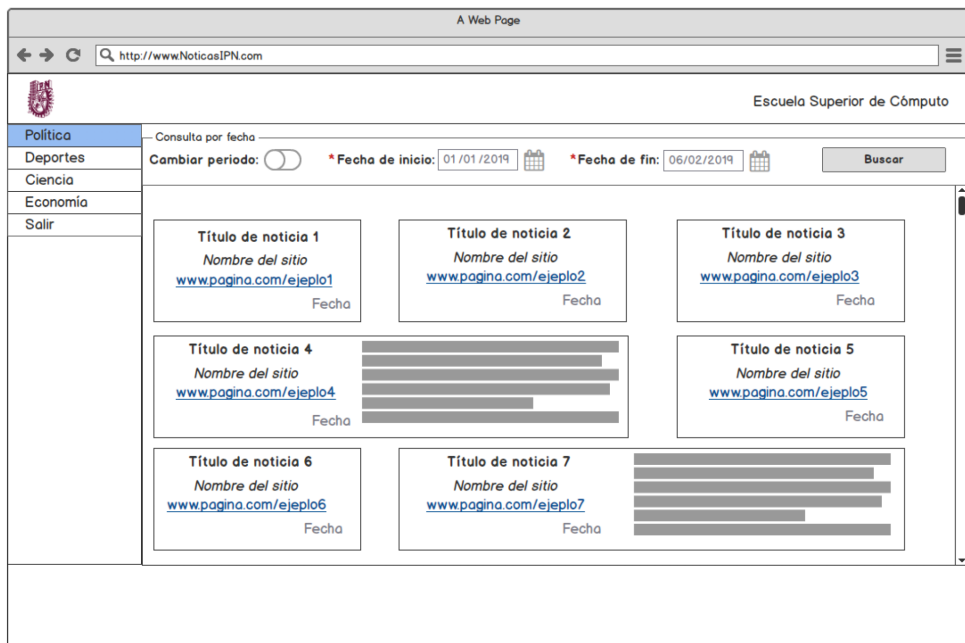


Figura 4.5: Pantalla UI3 Resultados de búsqueda

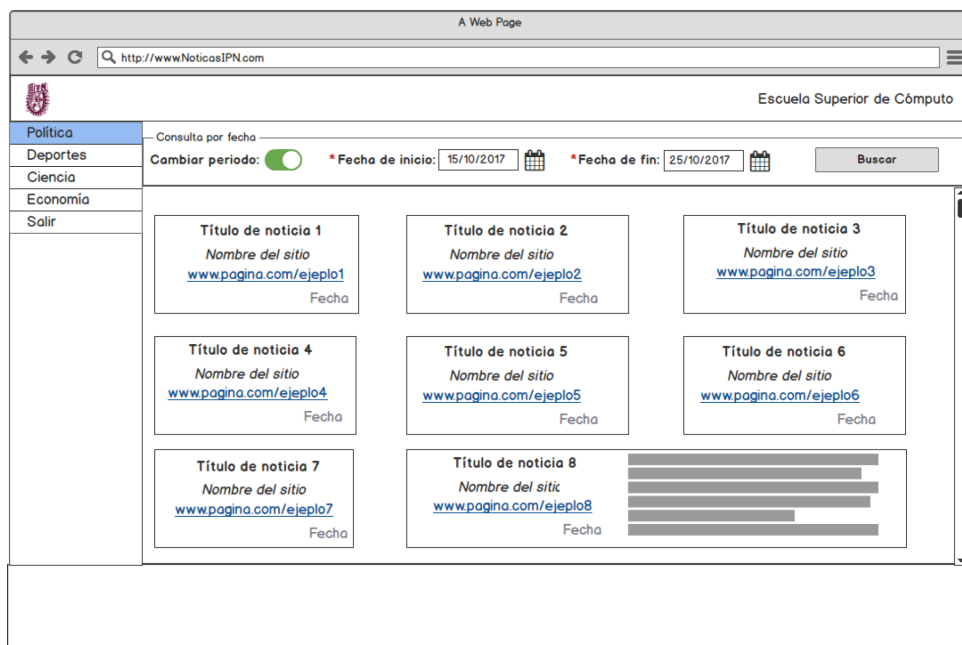


Figura 4.6: Pantalla UI3.1 Resultados de sección con periodo diferente

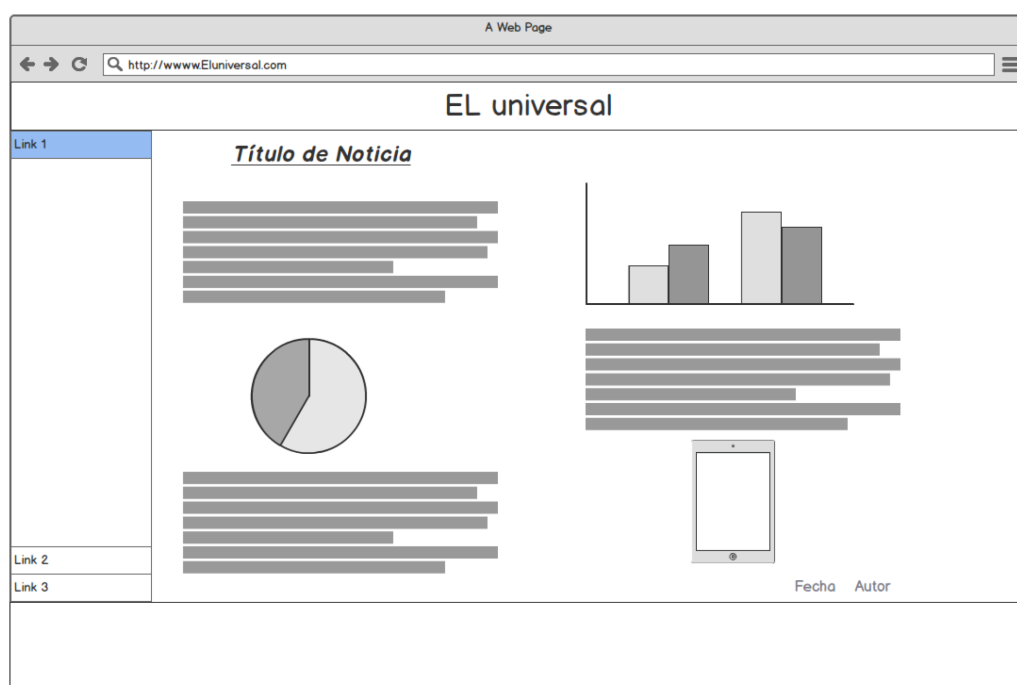


Figura 4.7: Pantalla UI4 Página de sitio web