

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

1. Introducción	1
1.1. Problemática	2
1.2. Justificación	2
1.3. Solución Propuesta	3
1.4. Objetivo	3
1.5. Objetivos Específicos	3
2. Estado del arte	5
2.1. Introducción	5
2.2. Trabajos nacionales	5
2.3. Trabajos internacionales	6
2.4. Herramientas disponibles	6
3. Análisis y diseño	9
3.1. Requisitos funcionales	9
3.2. Requisitos No funcionales	10
3.3. Reglas de negocio	10
3.4. Casos de uso	12
3.4.1. Diagrama de casos de uso	12
3.4.2. CU1 Mostrar noticias	12
3.4.3. CU2 Recolectar noticias	15
3.4.4. CU3 Calasificar noticias	17
3.4.5. CU4 Tokenizar noticias	19
3.4.6. CU5 Lematizar noticias	21
3.4.7. CU6 Filtrar noticias por fecha	23
3.5. Pantallas	25
3.5.1. UI1- Inicio	25
3.5.2. UI2-Sección deportes	26
3.5.3. UI3-Busqueda por fecha	27
3.5.4. UI4-Página ejemplo	28

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 contenido 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 contenido 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 a su 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)

Capítulo 2

Estado del arte

2.1. Introducción

A continuación, se mostrarán distintos trabajos nacionales e internacionales, así como herramientas las cuales desempeñan una labor similar al propuesto en este trabajo.

2.2. Trabajos nacionales

El trabajo *Clasificación Automática de Textos de Desastres Naturales en México* propone clasificar noticias del ámbito en 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 preprocesamiento, 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.

2.3. Trabajos internacionales

La obra *Clasificación Automática de Textos Usando Redes de Palabras* 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 cocurrentes construida a partir de un conjunto de textos clasificados; para su realización el primer proceso 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.

El trabajo *Document Classification for Newspaper Articles* se ha enfocado en clasificar artículos del MIT (Massachusetts Institute of Technology) de las siguientes categorías:

- Arts
- Features
- News
- Opinion
- Sports
- World

Para los cuales utilizaron los algoritmos 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 utilizó un corpus de 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.

2.4. Herramientas disponibles

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

CAPÍTULO 2. ESTADO DEL ARTE2.4. HERRAMIENTAS DISPONIBLES

- *Google Cloud Natural Language*; Ha revelado 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 se puede utilizar para comprender las opiniones sobre sus 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[]
- Algoritmo Naive Bayes
- Procesamiento de lenguaje natural
- Árbol de decisión
- Clasificación máxima de entropía

Capítulo 3

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.

3.1. 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

3.2. 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

3.3. 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:**
- **Referenciado por:** nombre caso de uso

RN2 Lenguaje



- **Tipo:**
- **Descripción:** Las noticias deben estar redactadas en lenguaje español
- **Ejemplo:**
- **Referenciado por:** nombre caso de uso

3.4. Casos de uso

3.4.1. Diagrama de casos de uso

La figura 3.1 muestra el diagrama de casos de uso implementado en el sistema.

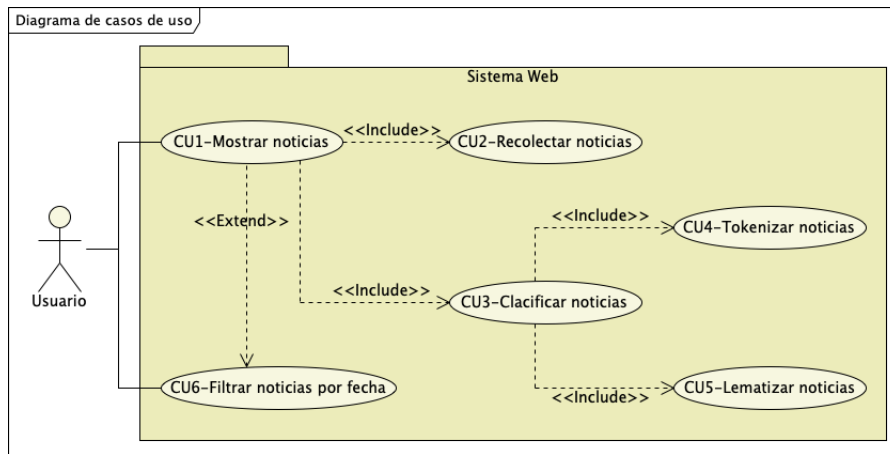


Figura 3.1: Diagrama de casos de uso

3.4.2. CU1 Mostrar noticias





Resumen

Brinda al usuario un punto de acceso para visualizar el portal web correspondiente a cada tipo clasificación, ya sea **Política**, **Deportes**, **Ciencia** o **Economía**; El sitio muestra al menos 15 noticias las cuales se encuentran ordenadas de forma descendente por la fecha de difusión, cada artículo contiene el **Título**, la **Fecha de publicación**, si es el caso un **Resumen de la información** y un **Link** el cual direcciona a la página fuente. Cabe mencionar que los datos mostrados datan desde un mes anterior a la fecha en el cual el actor ha ingresado a la página.

Descripción



Caso de uso:	CU1 Mostrar noticias
Actor:	Usuario
Propósito:	Proporciona una herramienta al usuario para acceder a los diferentes tipos de clasificaciones disponibles.
Entradas:	Ninguna
Salidas:	Noticias clasificadas y ordenadas
Precondición:	Lorem Ipsum
Postcondiciones:	Lorem Ipsum
Reglas de negocio:	Lorem Ipsum
Errores:	Lorem Ipsum
Autor:	Lorem Ipsum

Trayectoria principal

1.  lorem ipsum
2.  lorem ipsum
3.  lorem ipsum
4.  lorem ipsum
5. - - - *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





3.4.3. CU2 Recolectar noticias**Resumen**

Texto.

Descripción



Caso de uso:	CU-1 Recolectar noticias
Actor:	Lorem Ipsum
Propósito:	Lorem Ipsum
Entradas:	Lorem Ipsum
Salidas:	Lorem Ipsum
Precondición:	Lorem Ipsum
Postcondiciones:	Lorem Ipsum
Reglas de negocio:	Lorem Ipsum
Errores:	Lorem Ipsum
Autor:	Lorem Ipsum

Trayectoria principal

1.  lorem ipsum
2.  lorem ipsum
3.  lorem ipsum
4.  lorem ipsum
5. - - - - *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





3.4.4. CU3 Calasificar noticias**Resumen**

Texto.

Descripción



Caso de uso:	CU-1 Recolectar noticias
Actor:	Lorem Ipsum
Propósito:	Lorem Ipsum
Entradas:	Lorem Ipsum
Salidas:	Lorem Ipsum
Precondición:	Lorem Ipsum
Postcondiciones:	Lorem Ipsum
Reglas de negocio:	Lorem Ipsum
Errores:	Lorem Ipsum
Autor:	Lorem Ipsum

Trayectoria principal

1.  lorem ipsum
2.  lorem ipsum
3.  lorem ipsum
4.  lorem ipsum
5. - - - - *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





3.4.5. CU4 Tokenizar noticias**Resumen**

Texto.

Descripción



Caso de uso:	CU-1 Recolectar noticias
Actor:	Lorem Ipsum
Propósito:	Lorem Ipsum
Entradas:	Lorem Ipsum
Salidas:	Lorem Ipsum
Precondición:	Lorem Ipsum
Postcondiciones:	Lorem Ipsum
Reglas de negocio:	Lorem Ipsum
Errores:	Lorem Ipsum
Autor:	Lorem Ipsum

Trayectoria principal

1.  lorem ipsum
2.  lorem ipsum
3.  lorem ipsum
4.  lorem ipsum
5. - - - - *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





3.4.6. CU5 Lematizar noticias**Resumen**

Texto.

Descripción



Caso de uso:	CU-1 Recolectar noticias
Actor:	Lorem Ipsum
Propósito:	Lorem Ipsum
Entradas:	Lorem Ipsum
Salidas:	Lorem Ipsum
Precondición:	Lorem Ipsum
Postcondiciones:	Lorem Ipsum
Reglas de negocio:	Lorem Ipsum
Errores:	Lorem Ipsum
Autor:	Lorem Ipsum

Trayectoria principal

1.  lorem ipsum
2.  lorem ipsum
3.  lorem ipsum
4.  lorem ipsum
5. - - - - *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





3.4.7. CU6 Filtrar noticias por fecha**Resumen**

Texto.

Descripción



Caso de uso:	CU-1 Recolectar noticias
Actor:	Lorem Ipsum
Propósito:	Lorem Ipsum
Entradas:	Lorem Ipsum
Salidas:	Lorem Ipsum
Precondición:	Lorem Ipsum
Postcondiciones:	Lorem Ipsum
Reglas de negocio:	Lorem Ipsum
Errores:	Lorem Ipsum
Autor:	Lorem Ipsum

Trayectoria principal

1.  lorem ipsum
2.  lorem ipsum
3.  lorem ipsum
4.  lorem ipsum
5. - - - - *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

3.5. Pantallas

3.5.1. UI1- Inicio

Objetivo

Texto

Descripción

Texto

Comandos

- Lorem ipsum
- Lorem ipsum
- Lorem ipsum

Referencia

Nombre Caso de uso

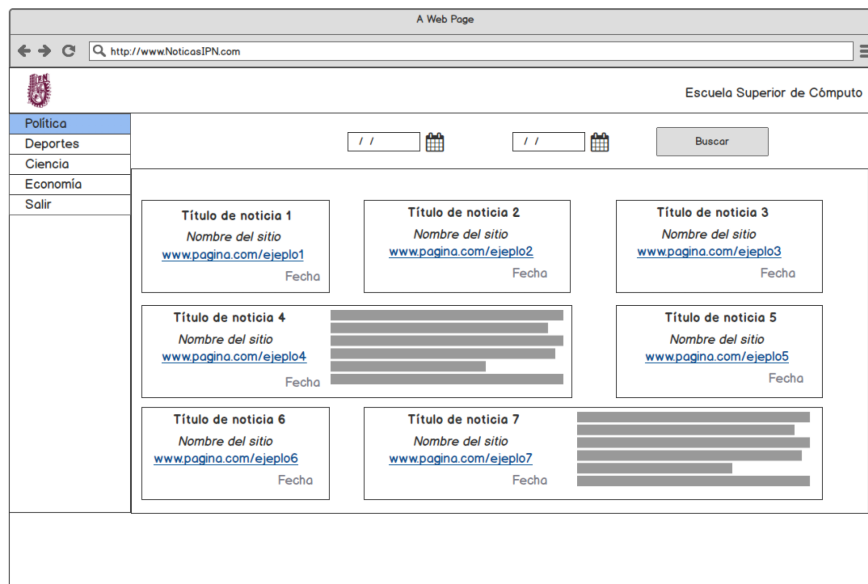


Figura 3.2: Pantalla IU1-Inico

3.5.2. UI2-Sección deportes

Objetivo

Texto

Descripción

Texto

Comandos

- Lorem ipsum
- Lorem ipsum
- Lorem ipsum

Referencia

Nombre Caso de uso

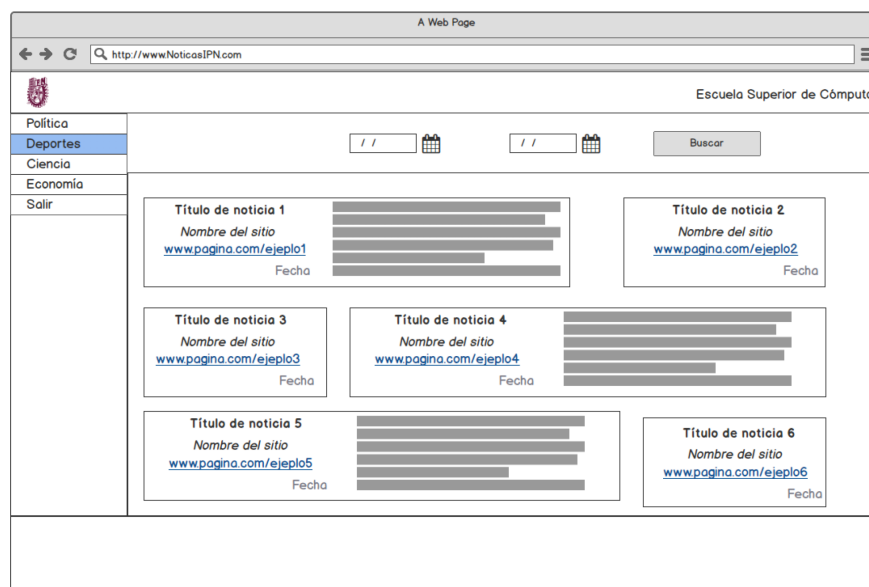


Figura 3.3: Pantalla IU2-Sección deportes

3.5.3. UI3-Búsqueda por fecha

Objetivo

Texto

Descripción

Texto

Comandos

- Lorem ipsum
- Lorem ipsum
- Lorem ipsum

Referencia

Nombre Caso de uso

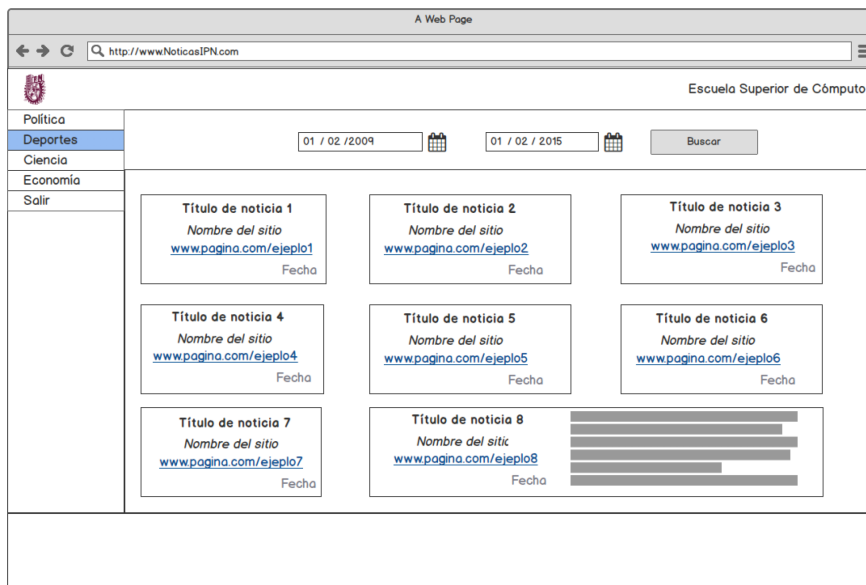


Figura 3.4: Pantalla IU3-Búsqueda por fecha

3.5.4. UI4-Página ejemplo

Objetivo

Texto

Descripción

Texto

Comandos

Texto

- Lorem ipsum
- Lorem ipsum
- Lorem ipsum

Referencia

Nombre Caso de uso

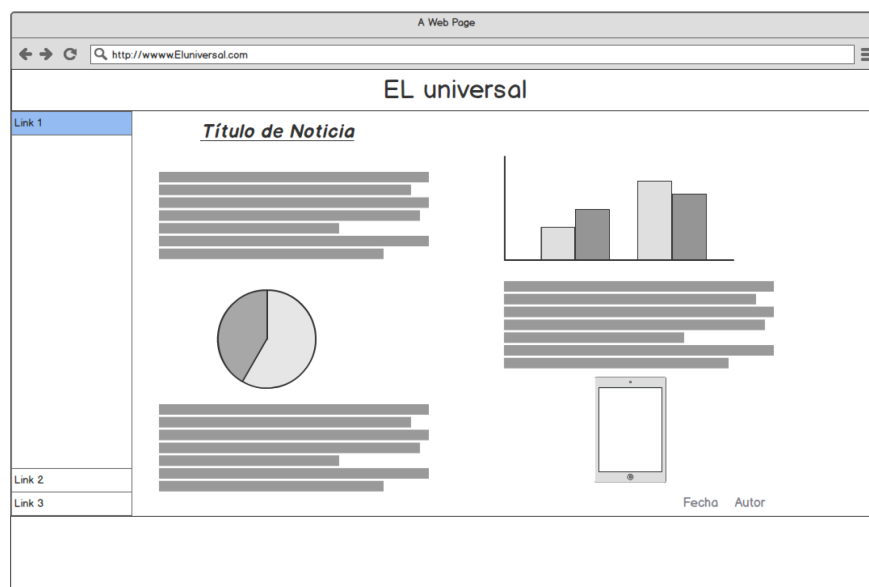


Figura 3.5: Pantalla IU4-Página ejemplo