
**ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA INFORMÁTICA**

**Proyecto fin de Carrera Máster**

**Máster en Ingeniería Informática**

# Lector de noticias inteligente

**Realizado por**
**(ponente): Andrés M. Jiménez Ríos**

**Dirigido por**
**Alejandro Fernández-Montes**

**Departamento**
**Lenguaje de Sistemas Informáticos**

**Sevilla, 13 de octubre de 2018**

# Resumen

¿Por qué LT-News? Surge de la necesidad de todos de informarse, pero la enormidad de medios y el poquísimo tiempo nos hace excusarnos de la gran tarea de leer y formarse una opinión propia de lo que ocurre a nuestro alrededor.

Este trabajo se propone ser un lector RSS. Pero no uno típico, no, sino uno que te acabe conociendo tanto, que te averigüe tus gustos, te muestre las noticias en las que estés realmente interesado. Además, poseerá funcionalidades tan útiles como la de buscar noticias en los periódicos que sigas, o hacer un análisis diario de las noticias más importantes de tus secciones.

Además, podrás ir viendo gráficas en bases a tu uso de la aplicación o de la evolución de la aparición de una determinada noticia a lo largo del tiempo.

# Agradecimientos

Sero te amavi, pulchritudo tam antiqua et tam nova, sero te amavi! et ecce intus eras et ego foris, et ibi te quaerebam, et in ista formosa, quae fecisti, deformis inruebam. mecum eras, et tecum non eram. ea me tenebant longe a te, quae si in te non essent, non essent. vocasti et clamasti et rupisti surditatem meam: coruscasti, splenduisti et fugasti caecitatem meam: fragrasti, et duxi spiritum, et anhelo tibi, gustavi et esurio et sitio, tetigisti me, et exarsi in pacem tuam.

# Índice general

# Índice de cuadros

# Índice de figuras

# Índice de código

# CAPÍTULO 1

# Introducción

## 1.1– Justificación del proyecto

Todos hemos observado cómo ha evolucionado Internet. Ahora, en unos minutos, cualquiera puede hacer un blog sobre cualquier tema. Ese mismo, además, suele accesible desde cualquier parte del Globo las 24 horas del día, los 365 días del año. Esto ha llevado a la enorme especialización en muchísimos temas, a la transmisión de información de manera inmediata, justo en el momento, así como a una globalización cada vez más plástica, si cabe.

Si a esto le sumamos la cantidad de periódicos que se encuentran accesibles en internet, podemos llegar a la conclusión de que estamos ante un mar insondable de información. Así, un usuario que quiere estar al día de una cantidad notable de temas, se encontrará que ha de estar pegado a una pantalla recargando quizá una decena de webs para conseguirlo. Cosa sencilla, pero que puede agotar a cualquiera.

¿Cómo solucionar esto? Sería ideal una herramienta que me permitiese unir todas las fuentes de información que sigo o me interesan para no recurrir a todas las webs, sino a una sola. Esto es un problema que ya se solucionó en 1995 con la sindicación de contenido1. Gracias a estándares como RSS (aparecido en 1999) o Atom (2003), un usuario podía y puede reunir todo el contenido de sus periódicos o blogs favoritos en una página web, una aplicación o u programa.

Una vez que conseguimos esto, nos encontramos con que no siempre nuestros feeds, que es el nombre que recibe el archivo que genera la web con las últimas noticias, nos sorprenden con noticias interesantes. Puede llegar incluso el momento en que entre toda la información que veamos en nuestro lector RSS nos interese solo una pequeña minoría. ¿Cómo saber lo que nos gusta entre tanta cantidad de información?

Esto es el problema que viene a resolver nuestro proyecto. Este será un lector RSS enriquecido, que sepa descubrir los gustos del usuario, que sepa relacionar las noticias. En definitiva, analizar lo que se le muestra al usuario, e incluso a este mismo. No encuentro mejor forma de explicarlo que con una cita del profesor Lev Manovich.

> Tras la novela, y posteriormente la narrativa cinematográfica como forma clave de expresión cultural de la era moderna, la era digital introduce su correlato: las bases de datos. Es natural, entonces, que queramos desarrollar una poética, una estética y una ética de los datos.
>
> Manovich (2001)

Esto es lo que quiere hacer LT-News: *una poética de los datos*.

## 1.2– Objetivos

Dentro de los objetivos, los englobo en dos categorías. En la primera categoría agrupo todo aquello que me propongo que haga la aplicación resultante. En la segunda, lo que quiero aprender llevando a cabo dicho trabajo.

### 1.2.1. Objetivos funcionales

El objetivo principal del proyecto es hacer un lector RSS enriquecido; esto es, que de un valor añadido a los que ya existen. En concreto, me propongo tres características a alto nivel que, creo, puedan dar este nuevo valor:

- Ver las noticias de los periódicos que elija el usuario.
- Buscar las noticias en base a distintos filtros que el usuario quiera.
- Recomendar noticias a los usuarios.

El primer objetivo, como vemos, es, y debe ser, común a todos los lectores RSS. Es por ello, que nuestro sistema debe ser capaz de extraer las noticias regularmente de los medios que estén en la base de datos y de mostrar dichas noticias de una manera clara y atractiva. El segundo objetivo es claro: nuestra aplicación debe analizar las noticias para permitir su búsqueda.

Por último, se propone un valor añadido sobre los demás lectores: la capacidad de recomendar noticias a los usuarios. Para ello, es necesario un doble aspecto. El primero es que, gracias al punto anterior, tenemos analizadas las noticias que poseemos. Gracias a esto, podemos relacionar noticias entre sí según los temas que traten. El segundo aspecto es el análisis de los usuarios de la aplicación: se extraerá un perfil del mismo según sus gustos. Gracias a esta relación entre usuarios-temas, y noticias-temas, podemos relacionar usuarios con noticias, es decir, recomendar.

### 1.2.2. Objetivos docentes

Bajo este punto de vista englobo tres grandes objetivos que me planteo a conseguir gracias al trabajo, desde un punto de vista docente.

El primero es el aprendizaje de la técnica web scraping. Gracias a esta se podrá obtener una noticia de cualquier periódico, incluyendo, además de los datos que obtengamos del feed, su cuerpo principal o la imagen de la noticia.

La segunda meta es la capacidad de trabajar con texto libre. A través de este trabajo quiero entrar, aunque sea un poco, dentro del Procesamiento de Lenguaje Natural. Gracias a esto, extraeré las keywords en base al texto de la noticia y podré relacionar noticias.

El tercer objetivo es la implementación de un sistema de recomendación. Gracias a esto, podré recomendar a los usuarios noticias en base a sus gustos, tal y como se ha explicado anteriormente.

## 1.3– Alcance

Con este proyecto quiero realizar una aplicación web capaz de leer feeds RSS. Tendrá una vista que permita la búsqueda de noticias de los diarios, en base a múltiples filtros. Además, deberá extraer los gustos de cada usuario en base al uso que haga de la aplicación, para, más tarde, recomendarle noticias. Se quiere hacer también uso de la relación que haya entre noticias, por ejemplo, agrupándolas según sus temas.

La aplicación que se desee hacer deberá ser accesible desde la web y será compatible con los principales navegadores y con los diferentes dispositivos (ordenadores, tabletas o smartphones). Además, tendrá soporte de feeds multilingüe, en el sentido que se podrán añadir medios en diferentes idiomas, funcionando igual de bien la aplicación. Por último, la página deberá estar, como mínimo, en dos idiomas: español e inglés.

## 1.4– Estado del arte

Como se ha dicho a lo largo de la introducción, nuestro sistema de información viene a resolver un problema actual: navegar por la enorme cantidad de información disponible en la red. Por eso se ha dicho es un lector RSS enriquecido. Para ver qué significa ese término de enriquecido, veamos cuáles son y qué hacen los principales lectores RSS.

# CAPÍTULO 2

# Partes y plantillas

## 2.1– Partes

Todas las memorias de Trabajo fin de Grado deberán constar de las siguientes partes

- Portada (según formato oficial). No debe incluir número de página. Debe incluir:
  - Sello de la universidad de Sevilla a dos tintas
  - ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA INFORMÁTICA
  - Trabajo fin de Grado
  - *Denominación del Grado*
  - Realizado por: *Nombre y apellidos del estudiante*
  - Dirigido por: *Nombre y apellidos del tutor o tutores*
  - Departamento *Nombre del departamento en el que se lee el TfG*
  - Sevilla, *Mes y año de la convocatoria de entrega*
- Preliminares: las páginas no se numeran o se numeran con números romanos.
  - Resumen en castellano (máximo una hoja)
  - Abstract (resumen en inglés, obligatorio para el caso de las memorias escritas en inglés, opcional para las escritas en castellano)
  - Agradecimientos (opcional)
  - Índice general (contenido de la memoria, con mención de las partes en que está dividida)
  - Índice de figuras (opcional)
  - Índice de cuadros (mal llamados en general, tablas) (opcional)
  - Índice de código o algoritmos (opcional).
- Cuerpo de la memoria, dividida en capítulos. El contenido de la memoria ha de incluir los elementos característicos de un proyecto de ingeniería o de un estudio o trabajo en el ámbito de una investigación, los cuales, en el sentido más amplio, son:
  - Definición de objetivos.
  - Análisis de antecedentes y aportación realizada.
  - Análisis temporal y de costes de desarrollo.

  - Análisis de requisitos, diseño e implementación.
  - Manual de usuario, en su caso.
  - Pruebas.
  - Comparación con otras alternativas.
  - Conclusiones y desarrollos futuros

  Estos puntos podrán ser ajustados y modificados en función de la naturaleza del proyecto realizado.

- Bibliografía: se deben documentar las fuentes bibliográficas utilizadas en el formato APA 2009.
- Índice alfabético o glosario (Lista ordenada de los conceptos, los nombres propios, etc.; que aparecen en la memoria, con las indicaciones necesarias para su localización)(opcional)
- Apéndices: Si la memoria contiene algún artículo de investigación o similar, este se incluirá en los Apéndices. Es necesario, en ese caso, incluir en la página anterior una hoja con la citación bibliográfica. En esta citación, el título del artículo se debe enlazar con la página web de la revista en la que aparecen el resumen o abstract y el acceso al texto completo.

La numeración de las páginas de la bibliografía y del glosario debe continuar la del cuerpo de la memoria. Los apéndices pueden llevar su propia numeración independiente o usar la general del cuerpo de la memoria.

## 2.2– Impresión

La entrega del memoria y en su caso, el depósito en biblioteca, se hacen en formato electrónico. Debido a ello, la memoria se presentará *a una cara*. Si se requiriera por algún motivo la impresión de la misma, se recomienda vivamente preparar la memoria adecuadamente para su impresión. Algunas sugerencias:

- Dejar una página en blanco cuando sea necesario para que los capítulos comiencen siempre en página impar (derecha)
- Ajustar los márgentes para que el exterior sea ligeramente más grande que el interior.
- Ajustar las cabeceras y pies de página (en su caso). Por ejemplo, si el número de página ocurriese en un lateral de la cabecera o pie, este debe ser siempre el exterior (derecho para las páginas impares, izquierdo para las pares)

## 2.3– Plantillas

Se encuentra disponible una plantilla LaTex en la sección de documentos de la plataforma de la aplicación de TfG de la ETSII `https://tfc.eii.us.es/TfG/`. Esa plantilla ha sido utilizada para preparar este documento. Se espera en breve disponer de plantillas de ejemplo para las aplicaciones OpenOffice Writer y Office Word.

Nota: Las plantillas se proporcionan como ejemplo, las condiciones obligatorias son las que constan en este procedimiento.

# CAPÍTULO 3

# Ejemplos

## 3.1– Manejo de la bibliografía

En esta sección mostramos brevemente ejemplos de referencia a la bibliografía citando un libro (de Sousa, 2004), un artículo (Bezos, 2007) y una página web (Autores, 2014).

Se recuerda que son campos obligatorios en todos los ítems de la bibliografía: autor(es), título del libro o artículo y año de publicación. En el caso de páginas web, es obligatoria la fecha de la última consulta. En general, la bibliografía debe ayudar al lector a encontrar fácilmente los ítems citados.

## 3.2– Código

En general se debe evitar incluir código o seudocódigo en la memoria. Si fuese preciso, se destacará de forma que sea fácilmente identificable y se indexarán los trozos de código incluidos. Un ejemplo puede verse a continuación.

```
%COMANDO PARA INSERTAR UN CUADRO UTILIZANDO EL FORMATO:
%1---> especificar numero de columnas y su alineacion ejm:
   % |r||c|c| r=right, c=center,l=left
%2---> especificar el caption o titulo de la figura
%3---> label para hacer referencia a la tabla insertada
%4---> contenido de tabla separando columnas con & y filas con \\
\newcommand{\cuadro}[4]{
      \begin{table}[htb]
         \centering
            \begin{tabular}{#1}
               \hline
                  #4
               \hline
            \end{tabular}
         \caption{#2}
         \label{#3}
      \end{table}
}
```

Código 3.1: Código de ejemplo en LaTeX

## 3.3– Imágenes

Este es un ejemplo de inclusión de figura en el texto (véase la figura 3.1).


Figura 3.1: Ada Lovelace

Figuras y cuadros se colocarán preferentemente tras el párrafo en el que son llamados por primera vez. Si no cupieran, se colocarán (en orden de preferencia):

- Al final de la página en que se llaman
- Al principio de la siguiente página
- Al final del capítulo

siempre respetando el orden de aparición en el texto.

## 3.4– Cuadros (mal llamados Tablas)

Este es un ejemplo de inclusión de cuadro en el texto. Véase el cuadro 3.1

| | | |
|---|---|---|
| elemento<br>elemento | elemento<br>elemento | elemento<br>elemento |

Cuadro 3.1: Cuadro de prueba

## Referencias

Autores, V. (2014). *Escuela Técnica Superior de Ingeniería Informática.* ETSII. Descargado de `http://www.informatica.us.es` (fecha de consulta: 24 de Noviembre de 2014)

Bezos, J. (2007). The titlesec and titletoc packages. *TexEmplares*, *8*, 283–298.

de Sousa, J. M. (2004). *Ortografía y ortotipografía del español actual*. Trea.

Manovich, L. (2001). *The language of new media.* MIT Press.