PROYECTO DE SISTEMAS INFORMÁTICOS

CURSO 2010/2011

Agrupación y resúmenes multi-documento de noticias de periódicos web

Autores:

Fernando García-Mauriño González-Conde
Eduardo Gordillo Berlanga
Carlos Puebla Sainz

Directores:

Alberto Díaz Esteban
Laura Plaza Morales

Autorización

Los abajo firmantes: Fernando García-Mauriño González-Conde, Eduardo Gordillo Berlanga y Carlos Puebla Sainz, autorizan a la Universidad Complutense de Madrid a difundir y utilizar con fines académicos, no comerciales, y, mencionando expresamente a los autores, tanto la presente memoria, como el código, la documentación, y/o el prototipo desarrollado.

Fernando García-Mauriño González-Conde

Eduardo Gordillo Berlanga

Carlos Puebla Sainz

Índice

Resumen

Este proyecto consiste en el diseño y desarrollo de una aplicación capaz de generar de forma automática resúmenes multi-documento a partir de noticias extraídas de las páginas web de distintos periódicos.

Para ello hemos desarrollado varios módulos independientes, capaces de capturar las noticias de las páginas web de los periódicos, agrupar esas noticias en grupos que traten un mismo tema y finalmente generar los resúmenes multi-documento de cada una de las agrupaciones. Además, realizamos una evaluación de las agrupaciones y de la calidad de los resúmenes.

Se trata de una aplicación completamente configurable en la que, entre otras cosas, se pueden añadir periódicos donde buscar noticias, seleccionar los pesos de las heurísticas de generación de resúmenes, seleccionar el umbralde similitud para la agrupación de noticias. Todo esto con el fin de que el usuario pueda probar distintas configuraciones y quedarse con la que mejor funcione para su caso concreto.

Palabras clave: Noticias, Cluster, Agrupación, Resumen, Multi-documento.

Abstract

This project is an application that automatically generates multi-document summaries from news that come from newspapers websites.

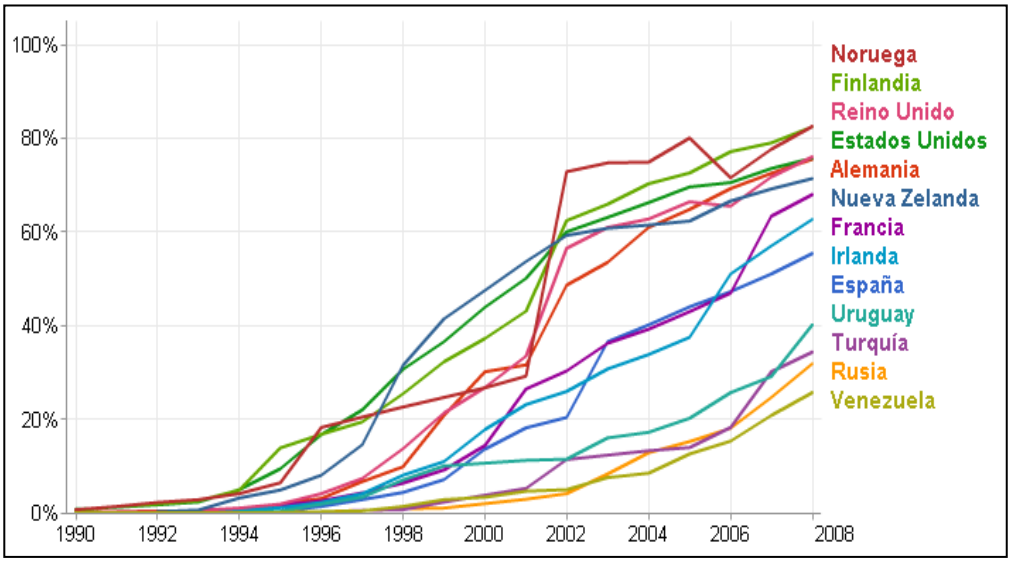
To this end, we have developed several independent modules. These modules capturethe news from the websites and make single issue groups. Finally they generate a multi-document summary for each group. We have also developed modules that evaluate the clustering process and the quality of the summaries.

This is a configurable application that allows the user toadd or delete newspapers, modify heuristic weight-setting and choose the threshold that generates the best cluster. In this way, the user can test different configurations and choose the one that best fits their needs.

Capítulo 1: Introducción

1.1 -Actualidad

Con el gran avance tecnológico que se está produciendo en los últimos años cada vez somos más gente los que utilizamos Internet y con ello la exploración de páginas web. Si observamos la gráfica siguiente podemos comprobar que más de un 50% de la población desarrollada o semi-desarrollada accede a Internet.

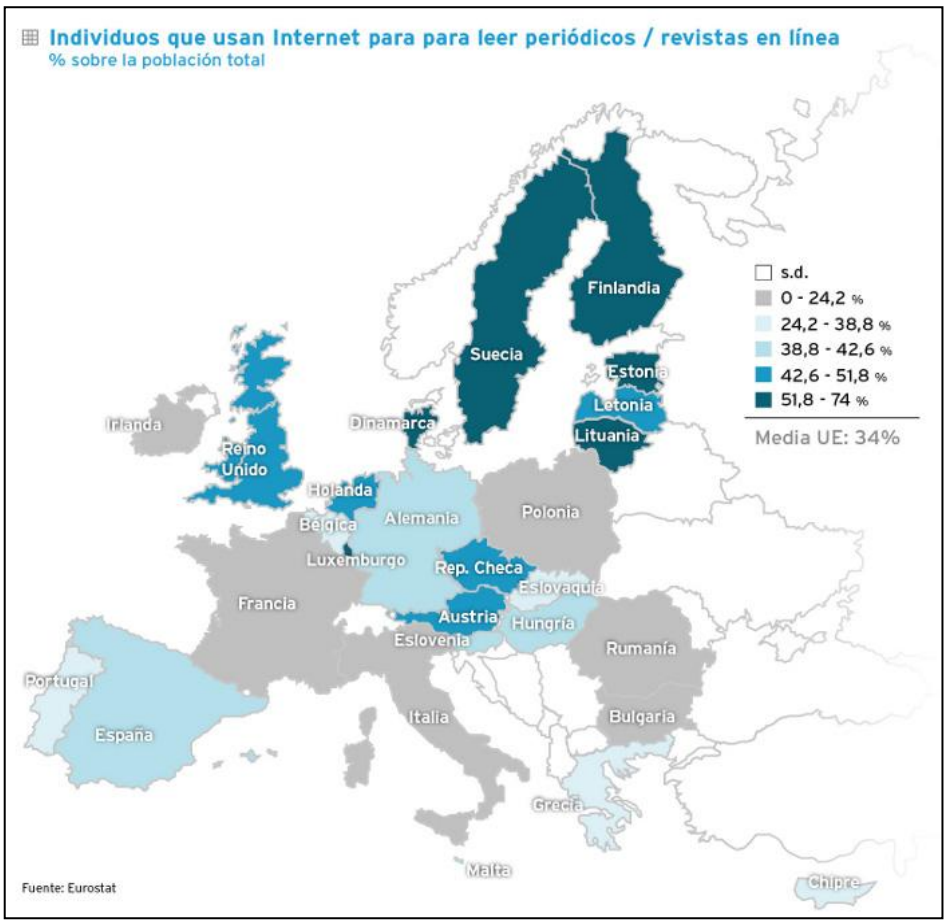


Gráfica 1.1: Porcentaje de la población que usa Internet

Esto es una cifra enorme de usuarios que utilizan Internet y navegan por páginas web, y por ello es muy necesaria la reorganización de las páginas web y su clasificación para poder encontrar rápidamente la información que cualquiera de estos usuarios necesite. La mayor parte de los buscadores se ven obligados a mostrar la información de una manera más intuitiva y accesible para los usuarios. Surge por lo tanto la idea del clustering y procesamiento de lenguaje natural. Es necesario **recopilar información, clasificarla y resumirla**.

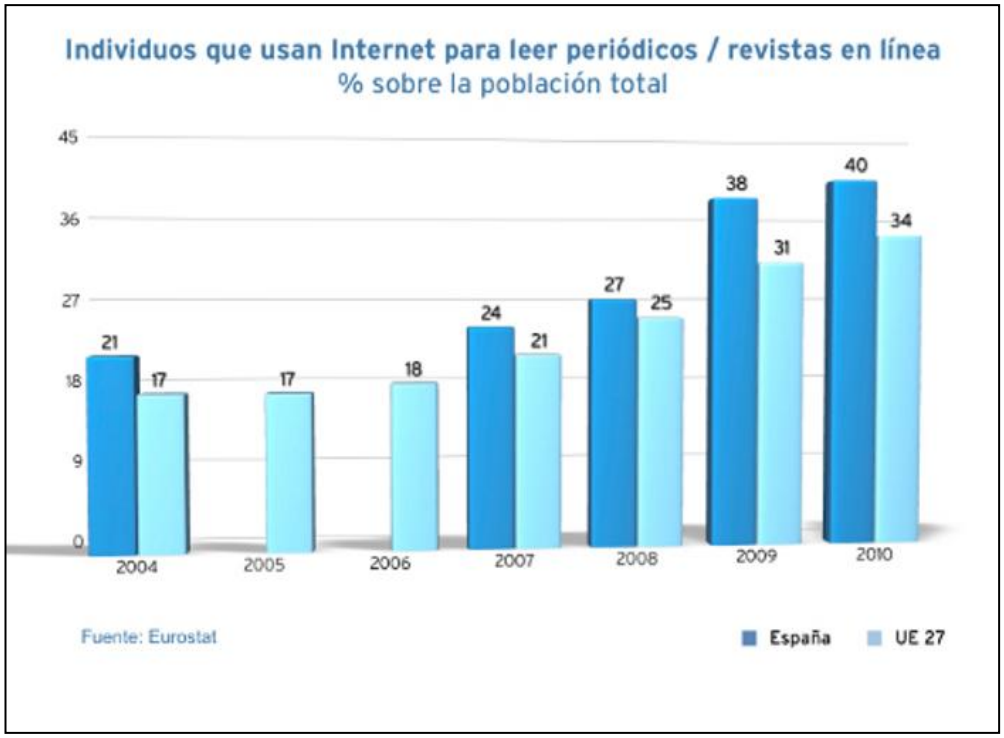
En este proyecto vamos a enfocar esta recopilación hacia las noticias de periódicos publicadas en la web.

En los últimos años, la utilización de Internet para leer periódicos es muy extensa. En la gráfica 1.2 se observa el porcentaje de usuarios que utilizan la red para leer periódicos en Europa.



Gráfica 1.2: Usuarios que leen periódicos en Internet

Si nos centramos en España, el incremento en el número de usuarios ha sido mayor respecto a Europa según se puede ver en la gráfica 1.3.



Gráfica 1.3: Porcentaje de españoles que usan Internet para leer periódicos

1.2 - Problema y objetivo

Existe una gran cantidad de artículos de periódicos relacionados unos con otros, es decir, que tratan el mismo tema. El usuario para conocer en detalle un acontecimiento concreto tiene que leer varios periódicos para estar bien informado, lo que supone un gran esfuerzo. Es por ello que el principal objetivo de este trabajo es facilitar la labor al usuario recopilando las noticias que traten el mismo contenido y realizando un único resumen de todas ellas.

De esta forma, el usuario no necesita consultar los distintos periódicos para poder conocer los diferentes puntos de vista y enfoques de cada uno de ellos, bastará con que lea un solo documento.

Todo este proceso se abordausando técnicas relacionadas con el procesamiento de lenguaje natural (Natural Language Processing). Es uno de los campos relacionados con la inteligencia artificial y con los procesadores del lenguaje en los que más se ha

trabajado e investigado a lo largo de los últimos años.El lenguaje natural, entendido como la herramienta que utilizan las personas para expresarse, posee propiedades que dificultan a los sistemas de recuperación de información textual. Los principales problemas son la gran variedad que tiene nuestro vocabulario y la aparición de la ambigüedad. Cuando hablamos de variedad nos referimos a que podemos usar diferentes palabras o expresiones para intentar decir una misma idea. En cambio, la ambigüedad lingüística aparece cuando una palabra o frase permite más de una interpretación.

Por estos motivos, es más conveniente orientar el reconocimiento del lenguaje natural al análisis de frases, oraciones y textos en su conjunto, que al reconocimiento de palabras aisladas. Es decir, dar más prioridad al reconocimiento del sistema en su conjunto que al reconocimiento de cada una de las partes que lo conforman.

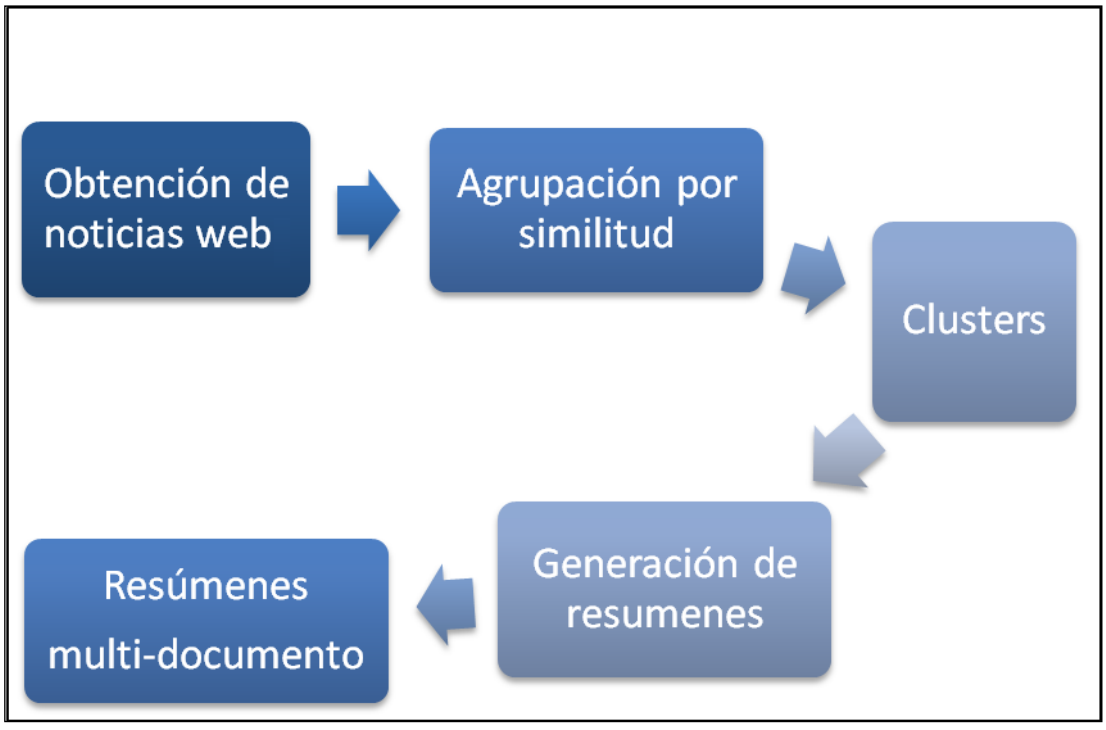
De las múltiples aplicaciones del procesamiento del lenguaje natural, en este trabajo se abordan las siguientes:

- Recuperación de la información
- Extracción de información
- Comprensión de texto
- Resumen automático

Capítulo 2: Sistema

El sistema desarrollado se divide en tres grandes etapas. En la primera, se capturan las noticias desde las webde los periódicos seleccionados; en la segunda, se agrupan esas noticias capturadas en grupos de noticias que tratan el mismo tema; y en la tercera, a partir de esas agrupaciones, se generan resúmenes multi-documento.

Fijándonos en el esquema 2.1, observamos todos los módulos en los que se ha dividido el proyecto.



Esquema 2.1- Módulos del sistema

2.1- Etapa 1: Obtención de noticias

Tiene como principal objetivo extraer todas las noticias de los distintos periódicos y almacenarlas correctamente en distintos archivos XML para su utilización en la siguiente etapa, la de agrupación.

La primera parte de nuestro trabajo consiste en observar y estudiar las distintas páginas web de varios periódicos para saber cómo están estructurados. Nos concentramos especialmente en uno (El País) para luego ampliar nuestro estudio a todos los demás.

Basándonos en el estudio de los HTML de las noticias de los periódicos acordamos crear un diseño para guardar su información, que consiste en dividir las distintas partes de las noticias en: **título, cabecera, entradilla y contenido**.

Además, la entradilla se divide en oraciones, mientras que el contenido se divide en párrafos y estos, a su vez, en oraciones. Para no detectar las abreviaturas como final de oración tenemos un archivo de configuración que almacena una lista de abreviaturas.

Para finalizar, una vez que guardamos todo el contenido de las noticias creamos sus archivos de tipo XML que explicaremos más adelante.

En este apartado explicaremos detalladamente los pasos seguidos y su implementación. La figura 2.1 muestra la estructura de la etapa.

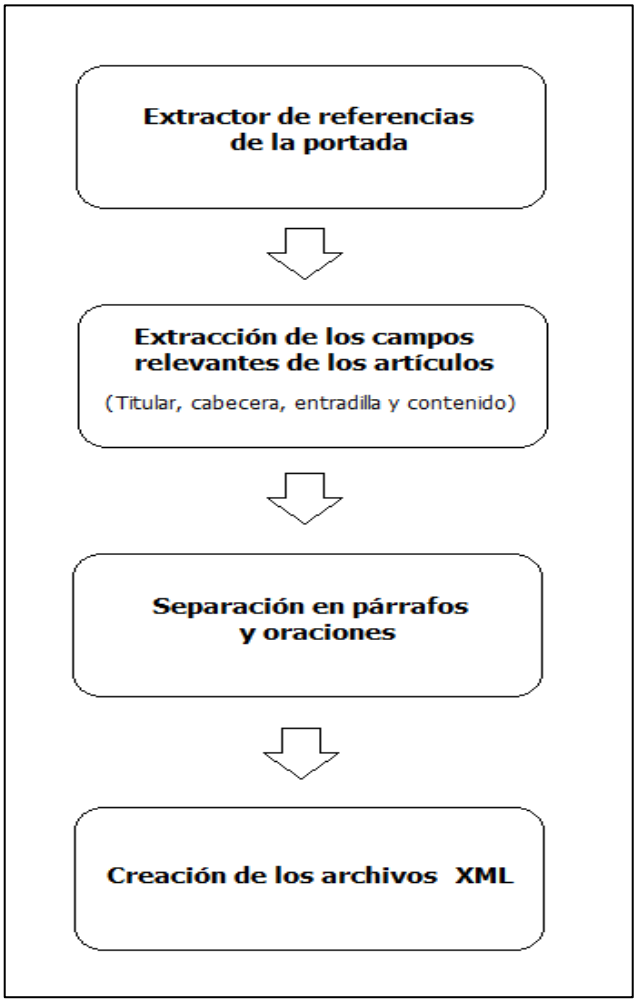


Figura2.1: Obtención

2.1.1– Configuración de la captura de noticias

En esta parte leemos un fichero de configuración en el que se encuentran la información que el capturador de noticias necesita para su ejecución. Para ello se apoya en tres ficheros, el de **configuración de periódicos**, el de **abreviaturas** y el de **decodificaciones**.

El primero le da información acerca de en qué periódicos se quiere buscar información, y qué palabras clave se van a utilizar para extraer el artículo y poder ignorar el resto de información irrelevante para su trabajo. Un ejemplo de cómo está organizado el fichero de configuraciónse muestra en la figura 2.2.

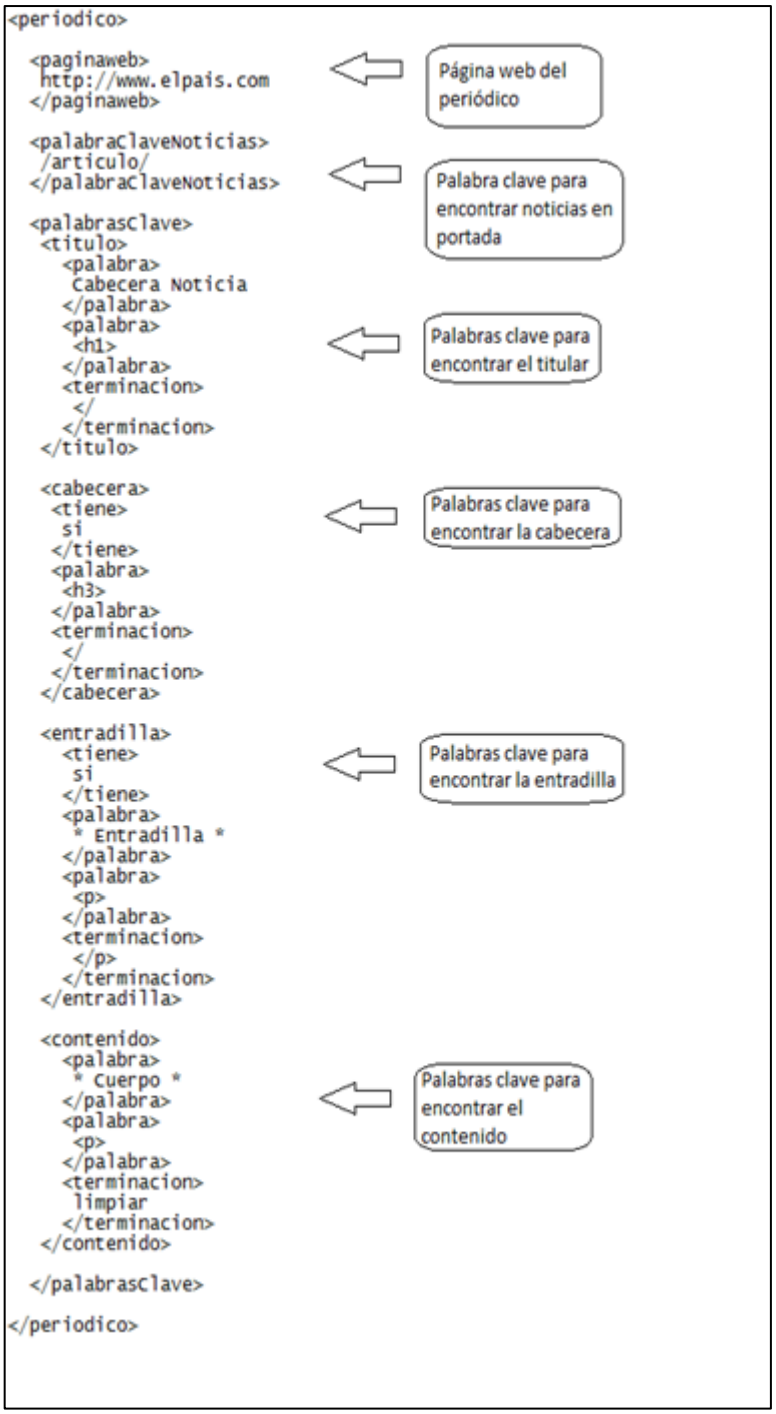


Figura 2.2: Ejemplo de archivo de configuración

Gracias a la incorporación de este archivo, es muy fácil añadir periódicos para la búsqueda de noticias. En primer lugar se debe introducir entre las etiquetas periódico, toda la información de configuración que se desea introducir. Lo primero que se debe rellenar con la etiqueta <*paginaweb*>, es la URL de la portada del periódico, por ejemplo http://www.elpais.com. Seguido de esto se añadirá la palabra clave que indicará cuales de todas las referencias de portada son enlaces a noticias, en el caso de El País, todas las referencias a noticias empiezan por "/articulo/". Empieza la configuración de cada noticia en particular, para ello es necesario configurar los cuatro campos en los que dividimos las noticias (titular, cabecera, entradilla y contenido). Todas se configuran de una forma similar, primero, mediante una etiqueta <*palabra*>, se introducen las palabras que tiene que buscar en el HTML de la noticia hasta llegar a donde empieza exactamente la parte que queremos. En el caso de los titulares de El País, lo primero que tiene que buscar es la palabra "Cabecera Noticia", una vez encontrado esto, tiene que avanzar hasta que lea "<h1>" y sabemos que a partir de ahí se encuentra el titular de la noticia. Para la búsqueda en este caso hemos necesitado dos palabras, pero puede ocurrir que solo necesitemos una, o más de dos. En cualquier caso, el titular siempre debe empezar inmediatamente después de la última palabra que introduzcamos. La siguiente etiqueta (<*terminacion*>) nos dirá exactamente dónde acaba exactamente el titular, que en este caso será al encontrarse con "</". Esto mismo se aplica a la cabecera, la entradilla y el contenido.

Además, la configuración de la cabecera y la entradilla tienen una etiqueta especial para indicar si el periódico que estamos configurando tiene esa sección en sus artículos. Para ello, sólo tendremos que indicárselo mediante un “sí” o un “no”. En el caso de El País, al contener cabecera y entradilla pondremos “sí” en ambos casos.

Adicionalmente, el capturador separa las noticias en párrafos y estos, a su vez en oraciones como podemos ver en la figura 2.3.

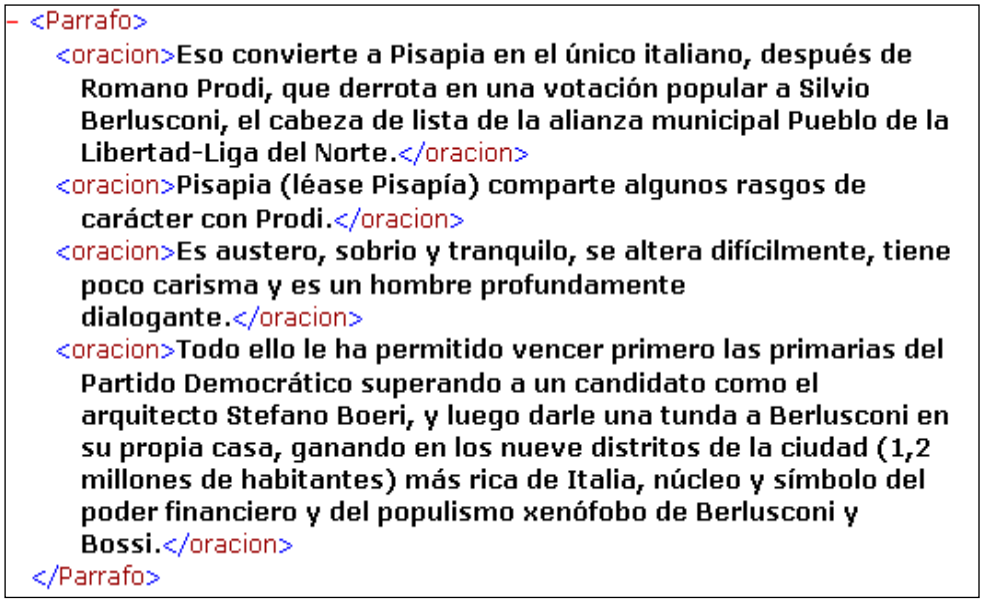


Figura 2.3: Párrafo y oraciones

También tenemos un fichero de configuración (abreviaturas.xml) que contiene una lista de abreviaturas, el propósito de este fichero es facilitarle al capturador información sobre cuando un punto es final de oración, ya que en ocasiones se puede encontrar con una abreviatura acabada con un punto y dar lugar a un mal funcionamiento al separar las oraciones. En este archivo se pueden introducir todas las abreviaturas respetando el formato XML, cada abreviatura irá entre las etiquetas <*abreviatura*> y </*abreviatura*>.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
- <abreviaturas>
    <abreviatura>E.E.U.U.</abreviatura>
    <abreviatura>www.</abreviatura>
    <abreviatura>etc</abreviatura>
    <abreviatura>...</abreviatura>
    <abreviatura>U.S.A.</abreviatura>
    <abreviatura>www.</abreviatura>
    <abreviatura>CC.</abreviatura>
    <abreviatura>AA.</abreviatura>
  </abreviaturas>
```

Figura 2.4: Archivo abreviaturas

El último fichero de configuraciónque tenemos para la etapa de captura es el de decodificación, ya que no todos los periódicos usan el mismo formato, y necesitamos tener la equivalencia de algunos caracteres que usan (figura 2.5).

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<Configuracion>
<Codigo>
<Palabra>&aacute;</Palabra>
<Correspondencia>á</Correspondencia>
</Codigo>
<Codigo>
<Palabra>&eacute;</Palabra>
<Correspondencia>é</Correspondencia>
</Codigo>
<Codigo>
<Palabra>&iacute;</Palabra>
<Correspondencia>í</Correspondencia>
</Codigo>
<Codigo>
<Palabra>&oacute;</Palabra>
<Correspondencia>ó</Correspondencia>
</Codigo>
```

Figura 2.5: Archivo de decodificación

2.1.3– Proceso: extracción de información relevante, segmentación en oraciones y creación de archivos XML

Como ya hemos comentado, lo primero es estudiar una página web y nos decidimos por el periódico El País. Podemos comprobar que en el HTML de la portada se encuentran todas las direcciones a noticias actuales que contiene ese periódico.

Nos centramos en recuperar solo las **referencias** a noticias, pero hay que tener cuidado ya que existen una gran cantidad de enlaces que no redireccionan a ninguna de ellas como pueden ser fotos, secciones del periódico, etc. Los enlaces en HTML están marcados con la etiqueta "*<a href=”* seguido de la dirección web.

Para conseguir las referencias correctas tenemos que averiguar cómo se diferencian en cada periódico las que son noticias de las que no.Normalmente, es una palabra determinada distinta en cada diario que está contenida en la referencia. Por ejemplo, en la página web de elpais.com la palabra clave es "/noticia/" y por eso el siguiente enlace *<a href="/pagina/home"* no es lo que estamos buscando porque no se refiere a ningún tipo de noticia sino que nos redirecciona a la portada. Por el contrario, el enlace "*<a href="/noticia/9110-opacidad-y-abusos-en-la-ayuda-regional-europea"* sí se refiere a una noticia.

El siguiente paso es la investigación del HTML de las noticias y su estructura. En estas páginas, existe bastante información que no nos interesa de momento para nuestra aplicación como pueden ser fotos, artículos relacionados, comentarios realizados por los usuarios, etc. Concentrándonos en la información que pretendemos obtener, observamos que los artículos se dividen normalmente en un título, con una cabecera, una entradilla y todo el contenido, tal y como se observa en la figura 2.6 para una noticia del periódico El País.

Figura 2.6- Noticia extraída de www.elpais.com

Comprobamos que cada parte está declarada de distinta forma dentro del código fuente de la página.

Como cada una se obtiene de una forma, debemos averiguar cuáles son las palabras que nos indican que comienza el título, cabecera y demás partes. Pero también se debe saber cuál es el final de cada una de las partes, y así no obtener datos que no pertenecen a su fragmento. Para esta recuperación nos facilita las cosas que los datos están organizados en orden, esto quiere decir que una vez que extraemos el título, más tarde se encuentra la cabecera, luego la entradilla y por último el contenido y nunca se altera este orden.

Además, optamos por separar el texto en **párrafos** y estos, a su vez, por **oraciones**. Para la separación por párrafos nos basamos en el diseño de todos los HTML de las noticias, en las cuales siempre aparece la etiqueta " " para denotar el comienzo del párrafo y para el final utilizan su cierre, es decir, " ". Las oraciones no están explícitamente delimitadas, sino que vienen separadas por un punto, teniendo en cuenta algunas excepciones, como puntos suspensivos o abreviaturas.

Una parte importante de este apartado es el control de la recolección de contenido erróneo, ya que en un artículo pueden aparecer enlaces, fotos en medio de un texto, código para indicar un estilo, etc. Como ya sabemos, el lenguaje HTML se basa en etiquetas, así que para saber qué información guardar comprobamos que dentro de un párrafo siempre que aparece una apertura de etiqueta ("<") no nos interesa su interior hasta que aparece su cierre de etiqueta (">").

Por ejemplo, considérese el siguiente párrafo:

* El jefe del Ejecutivo español, José Luis Rodríguez Zapatero, expresó su…… *

En este caso se guarda todo lo que aparece después de <p> (comienzo de un nuevo párrafo) hasta llegar al final del párrafo (etiqueta </*p*>), salvo <*strong*> y </*strong*> que indican el tamaño que quiere que aparezca José Luis Rodríguez Zapatero en la web.

Después de recuperar la información relevante tenemos que guardarla en **archivos** de tipo **XML** para cada utilización posterior en las siguientes etapas.

La estructura elegida es sencilla y fácil de modificar en un futuro si fuera necesario.

Cada sección de las noticias siempre va precedida de una etiqueta que indica el comienzo de texto. En la figura 2.7 se corresponde con <*Titular*>, <*Cabecera*>, <*Entradilla*> y <*Contenido*> y todo englobado por la etiqueta <*Noticia*>.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
- <Noticia>
    <Titular> Zapatero respalda al Constitucional y pide
      respeto a sus decisiones </Titular>
    <Cabecera> El presidente no ha especificado si comparte
      la sentencia del Alto Tribunal pero ha defendido
      expresamente su política antiterrorista </Cabecera>
  - <Entradilla>
      <oracion> El presidente del Gobierno, José Luis
        Rodríguez Zapatero, ha pedido en el Congreso
        "respeto" al Tribunal Constitucional y a sus
        decisiones. </oracion>
      <oracion> En respuesta al diputado de UPN Carlos
        Salvador Zapatero ha respaldado expresamente al
        Tribunal y ha hecho un llamamiento "a la
        responsabilidad" para no deslegitimarlo.
        </oracion>
    </Entradilla>
  - <Contenido>
    - <Parrafo>
        <oracion> No ha especificado, sin embargo, si
          comparte la sentencia que permitirá que Bildu
          esté en las elecciones municipales y
          autonómicas, pero sí ha defendido
          expresamente su política antiterrorista.
          </oracion>
      </Parrafo>
    - <Parrafo>
        <oracion> Según el presidente, "es posible
          discrepar y criticar las sentencias de los
          tribunales, lo que no es apropiado y es grave
          es deslegitimar una sentencia del máximo
          intérprete de la Constitución""Con Bildu o sin
          Bildu, el Gobierno seguirá con la misma
          firmeza y los mismos éxitos en su eficaz tarea
          de acorralar a ETA; con Bildu o sin Bildu el
          Gobierno no consentirá que nadie se
          aproveche de las instituciones; con Bildu o sin
          Bildu el Gobierno seguirá defendiendo la
          libertad democrática", ha asegurado.
          </oracion>
        <oracion> El presidente del Gobierno ha advertido
          de que nunca consentirá un cambio del estatus
          de Navarra, sin que lo quieran sus ciudadanos.
          </oracion>
      </Parrafo>
    - <Parrafo>
        <oracion> ETA ha vuelto luego al pleno cuando el
          diputado del PP Ignacio Gil lázaro ha insistido
          en preguntar por enésima vez a Alfredo Pérez
          Rubalcaba sobre El Faisán. </oracion>
        <oracion> El vicepresidente ha acusado al principal
          partido de la oposición de intentar desgastar al
          Gobierno usando este asunto. </oracion>
      </Parrafo>
    </Contenido>
</Noticia>
```

Figura 2.7- Archivo noticia XML

Las etiquetas </*Titular*>, </*Cabecera*>, </*Entradilla*> y </*Contenido*> indican el final de su sección, con esto sabemos hasta donde llega cada fragmento.

Cada oración aparece delimitada por las etiquetas <*oracion*> y </oración>, tal y como se puede ver en la entradilla. Dentro de la parte contenido, como ya hemos comentado anteriormente, las oraciones forman párrafos y para indicarlo en el fichero se escriben entre las marcas <*Parrafo*> y </*Parrafo*>.

Otro aspecto interesante es el nombre que vamos a dar a estos ficheros. Pensamos que lo mejor es que tengan la dirección web del periódico al que pertenecen y un número distinto en cada uno que indique el orden en el que han sido creados. Por ejemplo, si tenemos una noticia de El País y es el primero en crearse, su nombre de archivo es "elpais.com0". Nos decidimos por esta opción para poder saber en cualquier momento el periódico del que procede cada noticia.

Finalmente, el capturador genera un archivo llamado "archivoConfClasificador.xml" (figura 2.8) que será necesario para poder configurar correctamente el agrupador.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<Configuracion>
    <NumPeriodicos> 3</NumPeriodicos>
    <Periodico>elpais.com</Periodico>
    <Noticias>25</Noticias>
    <Periodico>elmundo.es</Periodico>
    <Noticias>23</Noticias>
    <Periodico>larazon.es</Periodico>
    <Noticias>19</Noticias>
</Configuracion>
```

Figura 2.8- Salida del capturador

2.2- Etapa 2: Agrupación de noticias

Esta etapa consiste en leer las noticias proporcionadas en los archivos XML que resultan de la ejecución del módulo de obtención de noticias, y agruparlas, de manera que relacione las noticias que traten de los mismos temas. Las agrupaciones creadas serán la entrada de la siguiente etapa, en la que se creará un resumen multi-documento de las noticias de cada agrupación.

En un caso ideal agrupará las noticias que hablan del mismo tema dados por los distintos periódicos y así poder hacer un resumen del tema en sí y quitar la parte subjetiva que añade cada uno de los diferentes periódicos.

El agrupador de noticias está claramente estructurado en dos pasos: representación de documentos y función de similitud (ver figura 2.9).

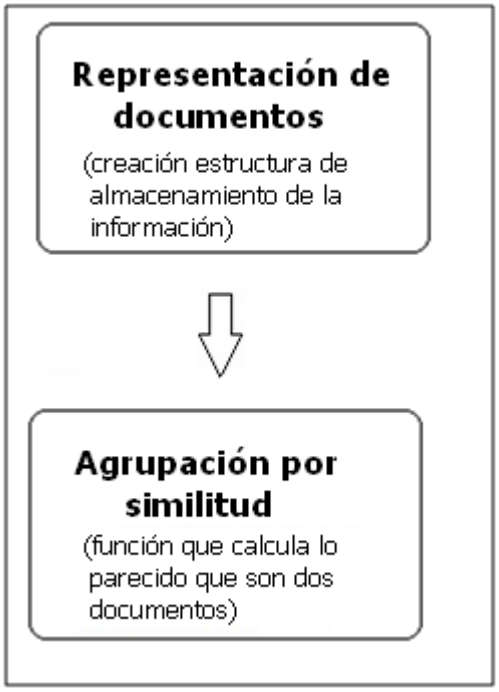


Figura 2.9- Estructura de la agrupación

2.2.1 - Configuración del agrupador

Lo primero que necesitamos para hacer la agrupación de documentos, es averiguar cuáles son los nombres que el módulo de obtención de noticias ha dado a los archivos XML de cada noticia. Para ello en el módulo anterior, generamos un documento XML en el que le indicamos el número de periódicos de los que hemos obtenido noticias y el número de noticias que obtenemos de cada uno, con lo que ya sabemos qué nombres le ha dado a los archivos. Por ejemplo, si tenemos cinco noticias de elpais.com, le habrá dado los nombres elpais.com0.xml, elpais.com1.xml, hasta elpais.com4.xml.

La apariencia del contenido del documento será la mostrada en la figura 2.8. Este documento es generado por la etapa anterior.

Una vez tenemos los nombres de los documentos, ya sólo nos falta leer un archivo de palabras comunes configurable, para que no las tenga en cuenta en la agrupación. Estas palabras son por ejemplo, preposiciones o determinantes, que no aportan ninguna información, es más, dificultarían el proceso de agrupación dando lugar a confusiones y pudiendo provocar falsos positivos o hacer que algún caso en el que los ficheros hablan de lo mismo, nos diga que no es así.

Este fichero también está en formato XML y es fácil añadir nuevas palabras, tan solo hay que introducirlas con la etiquetas <*Palabra*>. La figura 2.10 muestra un fragmento del fichero de palabras comunes.

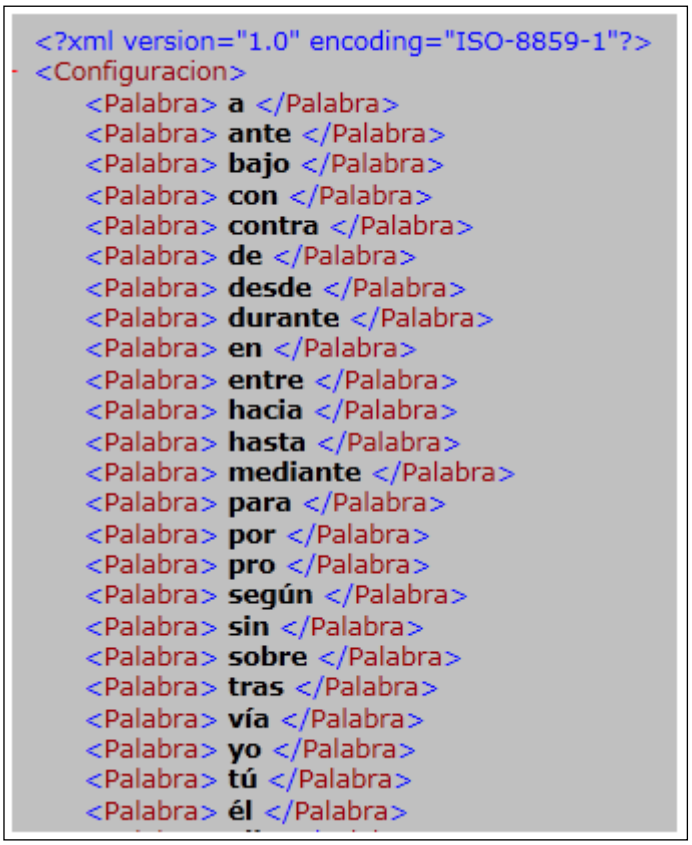


Figura 2.10- Palabras sin significado

2.2.2 –Proceso: Lectura de los documentos, rellenado de la estructura de datos y agrupación por similitud

Una vez configurado el agrupador, pasamos a la lectura de los documentos XML, leeremos palabra a palabra y las iremos almacenando de la siguiente manera:

En primer lugar, nos definimos una estructura "contpalabra", en la que tenemos un String con la palabra que representa junto con un vector en el que indicaremos en cada posición el número de veces que aparece en cada documento. Para tener constancia de todas las palabras, usaremos un vector que almacene en cada posición una estructura de las anteriormente definidas, teniendo así el número de veces que aparece cada palabra en cada documento.

La inserción de información se hace de la siguiente forma: si la palabra está en la lista, entonces aumentamos en uno el número

de veces que aparece en el documento actual, y en caso de que no aparezca introducimos un nuevo elemento en el vector con la palabra nueva y el vector de apariciones inicializado. Para dar más importancia a las palabras según aparezcan en las distintas partes de la noticia, cada vez que se lee una palabra, dependiendo de dónde se haya leído, se indicará que aparece más o menos veces. De esta forma, se tendrá más en cuenta si una palabra aparece en el titular de la noticia, que en el contenido, ya que en el titular y la entradilla hay mucha información sobre el tema a tratar.

Para mejorar la eficiencia, la inserción en el vector se hará de forma alfabética, de manera que si queremos saber si una palabra ya ha aparecido, la búsqueda será mucho más eficiente.

Como hemos dicho en el punto 2.2.1, no nos interesan todas las palabras, así que las palabras que aparezcan en la lista de palabras comunes no las introduciremos en la lista.

En resumen, realizamos una representación de las noticias como vectores de frecuencia.

Tenemos en cuenta el poder de resolución de los términos, que está basado en la frecuencia total de aparición de los términos en el corpus de documentos. De esta forma, los términos que tienen una alta o baja frecuencia de aparición en la colección no se utilizan en la representación (ver figura 2.11).

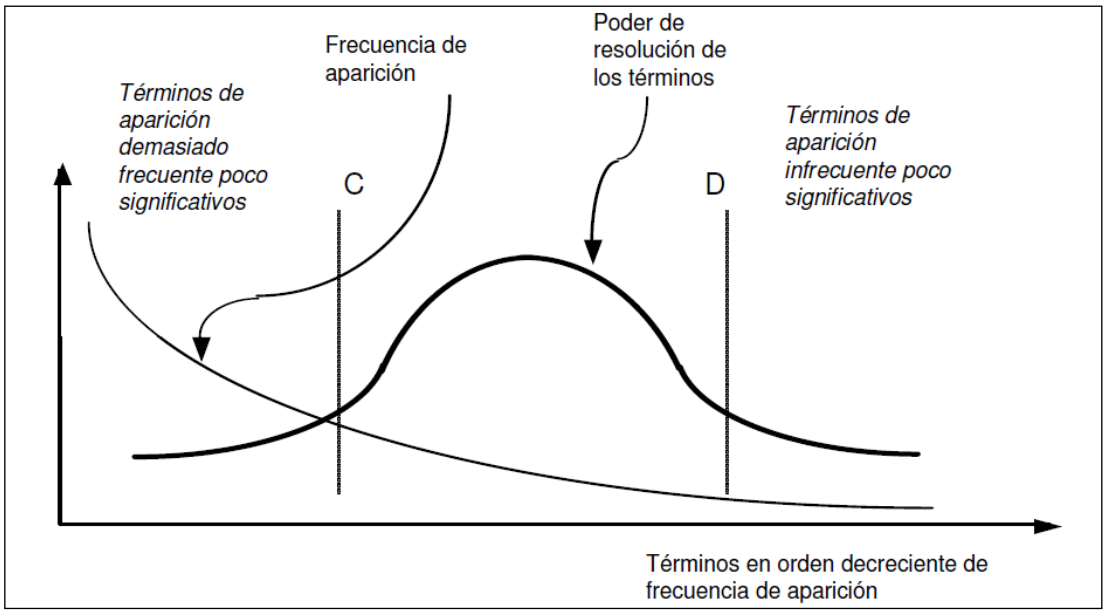


Figura 2.11-Frecuencia de aparición y poder de resolución de los términos.

Para calcular las agrupaciones usamos los pesos TF-IDF (*Term Frequency – Inverse Document Frequency*), lo primero es el cálculo de la estructura IDF, que será un vector en el que cada posición contendrá elvalor IDF de esa palabra en concreto. IDF es la frecuencia inversa de una palabra en los documentos, y se calcula dividiendo el número total de documentos entre el número de documentos que contienen esa palabra. El número de apariciones en distintos documentos lo calculamos previamente.

El TF es la frecuencia del término en el documento y se calcula de la siguiente manera:

$$\mathbf{TF(pal,doc)=}\frac{\text{nº de veces que aparece la palabra pal en el documento doc}}{\text{nº de palabras que tiene el documento doc}}$$

El IDF valora la importancia de un término según sus apariciones en todo el corpus:

$$IDF(p) = \text{Log}_2 \frac{\text{nº total de palabras del corpus}}{\text{nº de documentos en los que aparece la palabra}}$$

Una vez que tenemos calculado el IDF, podemos calcular la matriz TF, definiéndola con tantas filas como documentos y tantas columnas como palabras. En cada posición introducimos la frecuencia de esa palabra en el documento, es decir, el número de apariciones de la palabra entre el número total de palabras del documento.

Una vez calculados los valores TF e IDF, tendremos un vector para cada documento en el que cada posición será el valor de TF-IDF de cada palabra que aparezca en cada uno de los documentos. Estos vectores son los que compararemos para saber si dos documentos son lo suficientemente parecidos mediante la función de similitud.

A la vez que vamos calculando los valores TF-IDF, nos vamos quedando con los tres valores más altos para cada noticia, que se corresponderán con las palabras más relevantes de la noticia. De este modo, al crear la agrupación podremos saber qué palabras la identifican mejor para darle al usuario una idea del tema que trata la agrupación.

Agrupación por función de similitud

Ahora que sabemos el peso de cada palabra en cada documento, podemos saber qué documentos hablan del mismo tema. Cuanto más parecidos sean los vectores de pesos de dos documentos, más se parecerán entre sí. Entonces lo que tenemos que averiguar es

qué vectores de aparición son parecidos. Para ello, calculamos el producto escalar que forman esos dos vectores, tal y como se define en la figura 2.12.

Para resolver esa fórmula, lo primero que hacemos es calcular el módulo de cada vector para no tener que calcularlo en cada comparación, y así hacer la ejecución más rápida.

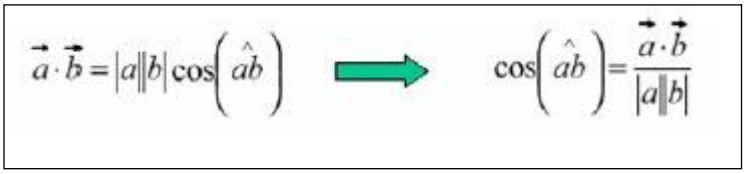


Figura 2.12- Producto escalar

Una vez calculados los módulos, hacemos un bucle que compara los documentos uno a uno y va calculando el coseno con la fórmula de la figura 2.12, y si ese coseno no supera un umbral, significa que los documentos hablan de cosas diferentes; en caso contrario, entenderemos que tratan el mismo tema. A medida que se aumenta ese umbral, se hace al agrupador más exigente, aunque no es conveniente reducirlo mucho, ya que puede dar lugar a que no agrupe ningún documento.

Para saber el valor de cada coseno y de qué documentos lo forman, nos apoyamos en una clase que almacena estos tres valores, de esta manera podemos crear un vector ordenado de esta clase, para facilitar el trabajo a la hora de crear los clusters.

Ahora ya sabemos lo que se parecen los documentos dos a dos, pero lo que nos interesa es crear clusters en los que tengamos un grupo de documentos que traten el mismo tema, así que lo primero que hacemos es seleccionar cuáles de ellos superan un cierto umbral configurable y ordenarlos de mayor a menor, para saber cuáles son los documentos que más se parecen, después recorremos cada pareja de documentos ordenados, uniendo en el

cluster los dos documentos, de la siguiente forma: si los dos documentos tienen ya un cluster asociado, fundimos los dos clusters añadiendo los documentos que se encuentran en uno de ellos al del otro. Si solo uno de los dos documentos tiene cluster, se une el objeto que no tiene al cluster que ya estaba creado, y si ninguno de los dos tiene, se crea uno nuevo con los dos documentos. Por ejemplo:

Si tenemos la siguiente tabla de correspondencia cluster-documento (tabla 2.1):

Documento	0	1	2	3	4	5
Cluster	-1	1	-1	2	2	1

Tabla 2.1- Correspondencia cluster-documento en un momento de la ejecución

Deducimos que tenemos dos clusters formados por el momento, uno con los documentos uno y cinco, y otro con los documentos tres y cuatro. Los otros dos documentos no tendrían cluster asignado en este momento.

Si comparamos ahora el documento uno y cuatro, y resulta que nos da un resultado por encima del umbral de similitud, fusionaríamos los dos cluster quedando la correspondencia mostrada en la tabla 2.2.

Documento	0	1	2	3	4	5
Cluster	-1	1	-1	1	1	1

Tabla 2.2- Correspondencia cluster-documento tras la fusión del cluster uno y dos

Sin embargo, si partiendo de la tabla 2.1 comparamos el fichero dos y cinco y resulta estar el resultado por encima del umbral de similitud, al no tener el documento dos cluster asignado, lo añadiríamos al cluster que contenga el grupo cinco, quedando el resultado mostrado en la tabla 2.3.

Documento	0	1	2	3	4	5
Cluster	-1	1	1	2	2	1

Tabla 2.3- Correspondencia cluster-documento tras la unión del documento dos al cluster uno

Para la implementación de los clusters, en primer lugar tenemos un array de control, que almacena en cada posición el cluster al que pertenece cada documento.

Una vez que ya sabemos a qué cluster tiene que ir cada documento pasamos a crear en cluster en sí. Esta clase tendrá el número de documentos que contiene, un array con los nombres de estos documentos, y otro array con el contenido de cada uno de los documentos, estructurado en forma de noticia (titular, cabecera, entradilla y contenido). También tendrá un array con las tres palabras con mayor TF-IDF de las noticias que contiene, que se calcularán a partir de los valores almacenados durante la generación de la tabla TF-IDF (serán las palabras clave de la agrupación).

Para acabar la parte del clustering ya sólo queda guardar el resultado de la agrupación en un documento, para que el generador de resúmenes sepa de qué documentos tiene que hacer el resumen. Este documento tendrá formato XML y en él aparecerá el número total de agrupaciones. De cada agrupación, indicará cuántas noticias tiene, en qué ficheros están almacenadas, y

cuáles son sus palabras clave. Un fragmento de este documento se muestra en la figura 2.13.

Figura 2.13 - Fragmento del fichero de configuración del generador de resúmenes multi-documento

2.3- Etapa 3: Generación del resumen multi-documento

Esta etapa se encarga de la generación de resúmenes multi-documento a partir de los clusters generados en la etapa 2. Se genera un resumen a partir de documentos que hablan del mismo tema.

2.3.1 - Configuración del generador de resumenes

En esta parte, se lee el fichero que genera el agrupador en el que indica cuantos clusters se han generado y qué documentos los componen (figura 2.8).

Ahora que sabemos qué documentos van en cada cluster, leemos los ficheros de cada noticia y los guardamos en una estructura "Noticia" de manera que el contenido sea accesible de una forma sencilla.

2.3.2 –Proceso: Generación del resumen

Para este apartado, nos apoyamos en las estructuras creadas en el apartado anterior, en las que tenemos almacenado el contenido de cada noticia, y como ya sabemos también cuales son las noticias que formarán cada resumen, podemos empezar a generarlo.

Lo primero es encontrar un titular para el grupo de noticias, para ello hemos decidido que la mejor opción será la de elegir el titular que más se parezca a los demás titulares del grupo, así que lo que hacemos es comparar todos los titulares dos a dos y sumar todos los valores de similitud, con lo que el que mayor puntuación saque será el titular escogido.

Después de seleccionar el titular, pasamos a seleccionar el cuerpo del resumen, para lo que tendremos que puntuar cada una

de las frases de todas las noticias del cluster. La puntuación será la suma de una serie de heurísticas. El usuario podrá elegir cuales de ellas se aplican o si alguna debe tener mayor o menor peso que el resto, seleccionándolo desde la interfaz. Las heurísticas a elegir son las siguientes:

- **Posición:** Da un valor a cada frase dependiendo de si aparece antes o después en el texto, con lo que las primeras frases de la noticia conseguirán una puntuación máxima, y las ultimas una valoración mínima.

- **Similitud con su titular:** Compara cada frase con el titular de la noticia a la que pertenece mediante una función de similitud similar a la aplicada para la comparación de dos textos en la agrupación de noticas. Así, la frase que más se parezca a su titular obtendrá una valoración más alta.

- **Comparación con el titular del resumen:** Similar a la heurística anterior, pero en vez de comparar cada línea con el titular al que pertenece, la comparamos con el titular del resumen que hemos seleccionado anteriormente.

- **Comparación con todos los titulares del cluster:** Compara cada frase con todos los titulares de todas las noticias del cluster, y calcula la media.

Con estas dos últimas heurísticas se consigue disminuir un posible error en el agrupador, ya que si alguna noticia se ha introducido en un cluster, y no debería estar en él, en estas

heurísticas sus frases tendrán muy poco valor, por lo que será más difícil que sean seleccionadas para el resumen.

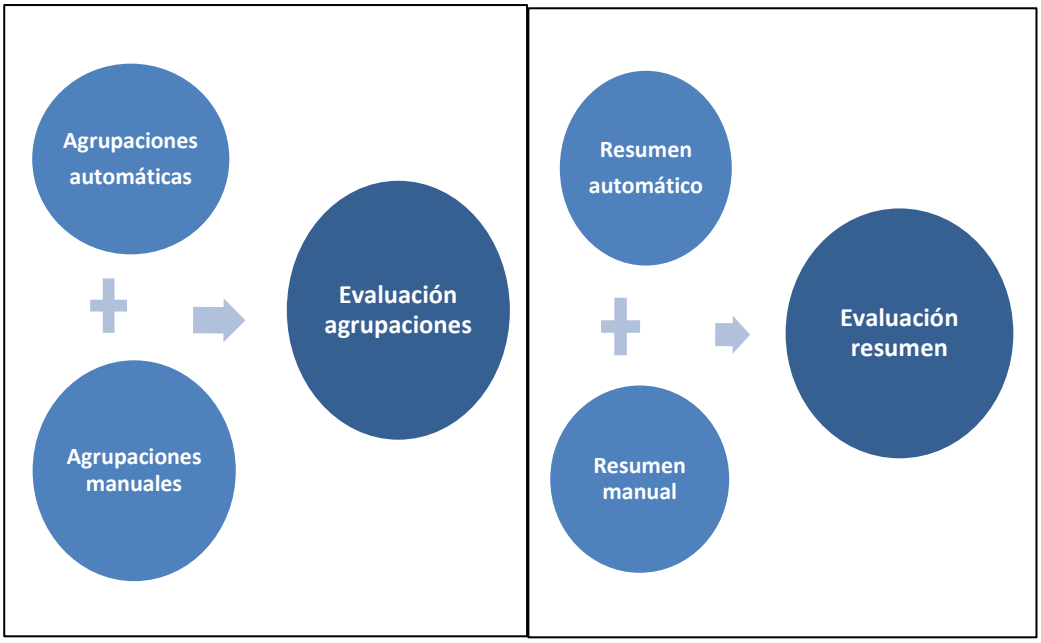
A medida que terminamos de puntuar una noticia, introducimos estas puntuaciones en una estructura que almacena las frases ordenadas de mayor a menor puntuación. En el momento que terminamos con la última noticia del cluster, solo tenemos que hacer un pequeño cálculo con el número de oraciones que tenemos y el porcentaje de resumen que quiere el usuario para seleccionar ese número de oraciones con mayor puntuación y formar el resumen.

Además, para que no haya frases muy parecidas, cuando se coge una oración para introducirla en el resumen, se compara con las ya introducidas para intentar eliminar las redundancias y si se parece mucho entonces no se añade.

Finalmente, para intentar evitar que se dé el caso de que dos oraciones, en la que la segunda habla sobre lo explicado en la primerahayan sido seleccionadas, y estén en orden inverso, se ordenan todas las frases seleccionadas mediante un algoritmo de quicksort por la posición en su periódico, de modo que, por ejemplo si se seleccionan las oraciones ocho y once de una noticia, la ocho aunque tenga menos puntuación en el resumen aparecerá antes que la once para intentar evitar incoherencias.

Capítulo 3: Evaluación

Nuestro sistema cuenta con una etapa adicional (Esquema 3.1) en la que se evalúa la calidad de las agrupaciones. Además se han realizado evaluaciones de los resúmenes generados automáticamente.



Esquema 3.1- Módulos de evaluación, para las agrupaciones y los resúmenes

3.1-Evaluación del algoritmo de agrupación

En esta sección vamos a tratar la forma de evaluación de las agrupaciones sacadas de la etapa 2.2 apoyándonos en el software de Lingpipe, que sirve para el procesamiento de textos de lenguaje natural. Algunas de sus principales funciones son: categorizar por temas, clustering, corrección de ortografía, base de datos de minería de texto, comparación de cadenas de texto, detección de frases interesantes, etc.

3.1.1 – Configuración de la evaluación de agrupaciones

La configuración de esta parte se necesitan dos ficheros: el archivo generado automáticamente por el agrupador (figura 2.13) y el archivo construido manualmente (figura 3.1).

Como se ha comentado en el capítulo 2, el agrupador genera un archivo en el cual se encuentran las diferentes agrupaciones resultantes con sus correspondientes noticias.

Es muy importante para que la evaluación se pueda llevar a cabo que el documento generado manualmente tenga el formato mostrado en la figura 3.1 y contenga los mismos ficheros de noticias que el generado por el programa. Es decir, al principio con la etiqueta <*NumCluster*> se especifica el número de agrupaciones totales. Seguido con una etiqueta <*Agrupacion*> se engloba cada una de las agrupaciones que constan de dos subetiquetas, la primera <*NumNoticias*> en la que indicas el número de noticias que tendrá la agrupación y la otra <*Noticia*> en la que indicas el nombre del archivo que contiene la noticia (de ésta habrá tantas como noticias haya en el grupo).

Figura 3.1- Fragmento del archivo Evaluacion_Manual_Agrupacion.xml

3.1.2 - Proceso

Para entender mejor esta parte desarrollaremos un pequeño ejemplo:

A partir de este momento, por simplicidad, las noticias del ejemplo las representaremos con los siguientes números: Elpais0.com→ 1, Elpais2.com → 2, Larazon0.es→ 3, ElPais11.com→ 4, Larazon29.es→ 5 y Larazon20.es→ 6.

Primero construimos manualmente las agrupaciones que deben resultar para poder compararlas con el resultado de la aplicación. Las agrupaciones que deben resultar son tres: {1,2,3}, {4} y {5,6}. Al conjunto {{1, 2, 3}, {4}, {5, 6}} lo denominaremos conjunto de referencia.

Por otro lado el resultado de nuestra aplicación cuando aplicamos la parte del clustering son dos agrupaciones: {1,2} y {3, 4, 5, 6}; estos dos conjuntos se unen formando el conjunto de respuesta.

Conjunto **Referencia**: {{1, 2, 3}, {4}, {5,6}}

Conjunto **Respuesta**: {{1,2}, {3, 4, 5, 6}}

Se evalúan cada par de elementos según las medidas proporcionadas por el software de Lingpipe. Las distintas categorías que existen para realizar la puntuación entre dos elementos son:

- **Verdaderos Positivos (VP).** Están en el mismo grupo de referencia y también en el mismo grupo de respuesta.

- **Falsos Negativos (FN).** Se encuentran en el mismo grupo de referencia pero en distintos grupos en el conjunto de respuesta.

- **Falsos Positivos (FP).** Aparecen en distintos grupos del conjunto de referencia y en el mismo grupo de respuesta.

- **Verdaderos Negativos (VN).** Tanto en el conjunto de referencia como en el de respuesta aparecen en distintos grupos.

		Conjunto Respuesta		Referencias Totales
		Verdadero	*Falso*	
Conjunto Referencia	*Verdadero*	VP	FN	Positivas = VP+FN
	Falso	FP	VN	Negativas = FP+VN
Respuestas Totales		Positivas = VP + FP	Negativas = FN + VN	TOTAL = VP + FN + FP+ VN

Tabla 3.1.1- Categorías

Con la tabla 3.1.1 se pueden realizar algunos cálculos y estadísticas:

- **Accuracy→** Es el número de respuestas correctas dividido por el número total de casos.

- **Recall→** Son el número de Verdaderos Positivos (VP) dividido por las referencias positivas (VP + FN).

- **Precision→** Es el resultado de dividir los verdaderos positivos (VP) por las respuestas positivas.

- **Rejection recall→** Es el número de Verdaderos Negativos (VN) entre las referencias negativas (FP+VN).

- **Rejection precision→** El resultado de dividir los Verdaderos Negativos (VN) por las repuestas negativas (FN+VN).

Siguiendo con el ejemplo podemos observar que el par (1,2) pertenece a la categoría de verdadero positivo porque los elementos aparecen en el mismo grupo del conjunto de referencia, {1, 2, 3}, y en el mismo grupo del conjunto de respuesta {1,2}. Tanto el par (2, 1) como el par (1,1) se cuentan como verdaderos positivos. Por ello habrá tantos verdaderos positivos como elementos poniendo como pares los elementos idénticos.

Los pares (5,6) y (6,5) pertenecen también a verdaderos positivos; en total existen diez, seis para elementos idénticos y cuatro para los pares anteriores.

Un ejemplo de la categoría falso positivo es el par (3, 4), al igual que su contrario (4, 3); aparecen en distintos grupos de referencia pero en el mismo grupo de respuesta ({3,4,5,6}).

Los pares (1,3), (3,1) y (2,3) y (3,2) son los falsos negativos, que hacen un total de seis. El resto de pares de elementos como (1,6) pertenecen a verdaderos negativos, están en distintos grupos tanto en el conjunto de referencia como en el de respuesta.

Mostramos los cálculos obtenidos aplicados al ejemplo:

Total = 36

VerdaderosPositivos = 10

Falsos Negativo = 4

Falsos Positivos = 10

Verdaderos Negativos = 12

Referencias positivas = 14

Respuestas positivas = 20

Referencias Negativas = 22

Respuestas Negativas = 16

Accuracy = 0.6111111111111112

Recall = 0.7142857142857143

Precision = 0,5

Rejection recall = 0.5454545454545454

Rejection precision= 0,75

3.1.3 – Resultados de la evaluación del agrupador

A continuación se muestran los resultados con algunos casos de prueba. Se estructura en las siguientes secciones: la sección 1.1 agrupaciones de las que partimos, sección 1.2 agrupación manual, sección 1.3 se muestran los resultados de la evaluación. Por último, la sección 1.4 recopila las conclusiones.

Sección 1.1- Agrupaciones automáticas

Para realizar la evaluación de las agrupaciones se parte de las siguientes agrupaciones generadas por nuestro sistema con distintos umbrales de similitud:

Las tablas se encuentran en el anexo 1, con las siguientes correspondencias (tabla 3.1.2):

Tablas	Tabla 1.1	Tabla 1.2	Tabla 1.3	Tabla 1.4
Umbral de similitud	0.1	0.2	0.3	0.5
Número de clusters	6	20	32	50

Tabla 3.1.2- Tabla correspondencia figura-umbral similitud

Sección 1.2 – Agrupación manual

Se construye el archivo de agrupación manual partiendo de las mismas noticias que el sistema. Este archivo se puede ver en el anexo 1, tabla 1.5.

Número de clusters agrupación manual→ 12.

Sección 1.3- Resultados de evaluación de agrupación

La tabla 3.1.3 muestra los resultados obtenidos para los distintos umbrales de similitud utilizados.

Umbral de similitud	**0.1**	**0.2**	**0.3**	**0.5**
Total	3721	3721	3721	3721
Verdaderos Positivos	335	241	151	89
Falsos Negativos	42	136	226	288
Falsos Positivos	2592	108	62	12
Verdaderos Negativos	752	3236	3282	3332
Referencias Positivas	377	377	377	377
Respuestas Positivas	2927	349	213	101
Referencias Negativas	3344	3344	3344	3344

RespuestasNegativas	794	3372	3508	3620
Accuracy	0.292	0.934	0.923	0.919
Recall	0.889	0.639	0.401	0.236
Precision	0.114	0.691	0.709	0.881
Rejection Recall	0.225	0.968	0.981	0.996
Rejection Precision	0.947	0.960	0.936	0.920

Tabla 3.1.3 – Resultados de la evaluación de la agrupación para diferentes umbrales de similitud

Sección 1.4 – Conclusiones

Observando la tabla 3.1.3 se pueden obtener varias conclusiones. Nos vamos a centrar principalmente en los valores obtenidos para la medida "accuracy", que nos indica la precisión de cada uno de los casos de prueba. Con el valor del umbral de similitud igual a 0.2 se obtiene la mayor precisión (aproximadamente un 93%), mientras que para el umbral igual a 0.1 se puede ver que es el de menor valor (29%). Observando con más detalle, para valores de umbrales más grandes que 0.2 comienza a disminuir la precisión.

Llegamos a la conclusión que para el valor 0.1 se obtienen pocos clusters, ya que no se necesitan que las noticias de una misma agrupación sean muy parecidas.

Mientras que para valores del umbral mayores que 0.3 hacen que las formaciones de las agrupaciones sean más restrictivas, es decir, se necesitan que las noticias se parezcan mucho más para que pertenezcan al mismo cluster, y de esta forma, a medida que se aumenta el umbral de similitud se crean una gran cantidad de agrupaciones a la que pertenecen pocas noticias.

Por lo tanto, la elección para obtener las mejores agrupaciones con una cantidad de noticias elevadas (superiores a 50) es escoger un valor del umbral próximo a 0.2.

3.2 -Evaluación del algoritmo de generación de resúmenes

Para la evaluación de este parte del sistema nosotros empleamosuno de los paquetes de evaluación de resúmenes automatizados más extendido como es el sistema ROUGE (Lin 2004). Este módulo de evaluación no está incluido en nuestra aplicación.

3.2.1 – Configuración de la evaluación del generador de resúmenes

Se necesitan tanto el resumen generado automáticamente por el sistema como el resumen escrito manualmente, para poder realizar correctamente la comparación. Se recomienda tener el mismo número de oraciones en ambos resúmenes.

3.2.2 – Proceso: Métricas ROUGE

ROUGE representa las siglas *Recall-Oriented Understudy for Gisting Evaluation* y fue desarrollado por el *Information Science Institute* en la *University of Southern California.* Es una herramienta que compara un resumen generado por un sistema con otro u otros denominados modelos generados por humanos obteniendo diferentes métricas. Las más importantes son: ROUGE-N, ROUGE-L, ROUGE-W, ROUGE-SN.

- **ROUGE-N:** contabiliza el número de secuencias de palabras (ngramas)que coinciden entre un resumen candidato y uno o más modelos, por lo que se pueden calcular las medidas ROUGE-1, -2, -3,etc.

- **ROUGE-L:** emplea la longitud de las secuencias más largas quecoinciden en el candidato y en el modelo.

- **ROUGE-W:** una versión ponderada de ROUGE-L que, además de la longitud de la secuencia, valora la ausencia de "huecos" en la misma.

- **ROUGE-SN:** que tiene en cuenta bi-gramas que no necesariamente han de aparecer consecutivos en el texto, sino que pueden presentar hasta un máximo de N términos entre ellos.

3.2.3 – Resultados de la evaluación del generador de resúmenes automáticos

A continuación se muestran los resultados con algunos casos de prueba. Se organiza en las siguientes secciones: la sección 1.1 noticias de las que partimos, Sección 1.2 resumen manual, Sección 1.3 se muestran los resultados de la evaluación. Por último, la sección 1.4 recopila las conclusiones.

Sección 1.1- Noticias de prueba

Para realizar la evaluación de los resúmenes se parte de una agrupación que trata sobre un mismo tema. La agrupación que se ha elegido para realizar el proceso de prueba trata sobre el tema de Japón (explosión nuclear). Las noticias se encuentran en el anexo 2 y sus titulares son los siguientes:

Noticia 1: "En directo: Los niveles de radiación suben en Japón, tras una nueva explosión y un nuevo incendio."(Tabla 2.1)

Noticia 2: "Japón admite fugas radiactivas "que pueden afectar a la salud" tras un incendio y una nueva explosión en Fukushima."(Tabla 2.2)

Noticia 3: "Alarma nuclear."(Tabla 2.3)

Noticia 4: "El último balance estima en 4.000 los muertos en Japón."(Tabla 2.4)

Noticia 5: "La alarma radiactiva se dispara tras una nueva explosión en el reactor dos de Fukushima."(Tabla 2.5)

Noticia 6: "Japón, en vilo ante una emergencia nuclear que no hace sino empeorar."(Tabla 2.6)

Noticia 7: "Los vientos están dispersando hacia el océano la amenaza nuclear."(Tabla 2.7)

Noticia 8: "El único destino de los reactores de Fukushima es su desmantelamiento."(Tabla 2.8)

Sección 1.2 – Resumen manual

A continuación, se realiza el resumen manual partiendo de las noticias de la sección 1.1. Ver tabla 2.9 del anexo 2.

Titular resumen manual: "La alarma radiactiva se dispara tras una nueva explosión en el reactor dos de Fukushima."

Sección 1.3- Resultados de evaluación de resúmenes

La tabla 3.2.1 recoge las configuraciones utilizadas para generar los resúmenes, con los valores asignados a cada heurística.

Configuración	Conf.1	Conf.2	Conf.3	Conf.4	Conf.5
Posición Titular	1	0	0	1	1
Comparación su titular	1	0	0	0	1
Comparación todos titulares	3	1	1	2	1
Comparación titular resumen	3	1	0	1	1
Porcentaje número de oraciones	5	5	5	5	5

Tabla 3.2.1: Parámetros de las configuraciones de resúmenes

Aplicando las configuraciones de la tabla 3.2.1 con el resultado obtenido del módulo del agrupador de nuestro sistema obtenemos los resúmenes automáticos. Se encuentran en el anexo 1, tablas 1.10-1.14.

En la tabla 3.2.2 se muestran los resultados obtenidos aplicando las distintas métricas de ROUGE, en cada una de ellas solo nos centraremos en la medida del recall.

	R-1	R-2	R-3	R-4	R-L	R-W	R-S4
Conf.1	0.584	0.301	0.213	0.191	0.555	0.161	0.299
Conf.2	0.575	0.268	0.184	0.165	0.542	0.158	0.278
Conf.3	0.447	0.121	0.046	0.023	0.418	0.105	0.139
Conf.4	0.500	0.242	0.180	0.165	0.464	0.140	0.242
Conf.5	0.307	0.088	0.042	0.033	0.294	0.078	0.106

Tabla 3.2.2: Evaluación de resúmenes generados por el sistema con distintas heurísticas.

Sección 1.4 -Conclusiones

Observando la tabla 3.2.2 se pueden llegar a algunas conclusiones interesantes. En las configuraciones 1, 2 y 4 se obtienen los valores más elevados para las diferentes métricas ROUGE, estas configuraciones tienen en común que el valor del parámetro de la heurística de posición respecto al titular es menor que el valor de la comparación con todos los titulares o en el caso de la configuración 4 no se tiene en cuenta la posición. Además ocurre lo mismo con la heurística "comparación con su titular".

Esto ocurre debido a que no es tan importante la posición que ocupa la frase en la noticia, sino que lo quemás interesa es que la oración sea lo más parecida al tema que se habla en el titular elegido para el resumen (se corresponde con la heurística "comparación titular resumen"); o sea, lo más semejante con todos los titulares de las noticias del conjunto.

Por ejemplo, si por algún motivo en la agrupación aparece una noticia en el que su contenido no se relaciona con el tema de las demás noticias del cluster, dando más valor a los parámetros de las heurísticas mencionadas anteriormente nos aseguramos que en el resumen automático sólo aparezcan oraciones relacionadas entre sí.

Capítulo 4: Funcionalidad

4.1 - Diagrama de casos de uso

En este apartado se explican las distintas funcionalidades de nuestro software a través de la figura 4.1 que representa un diagrama de casos de uso.

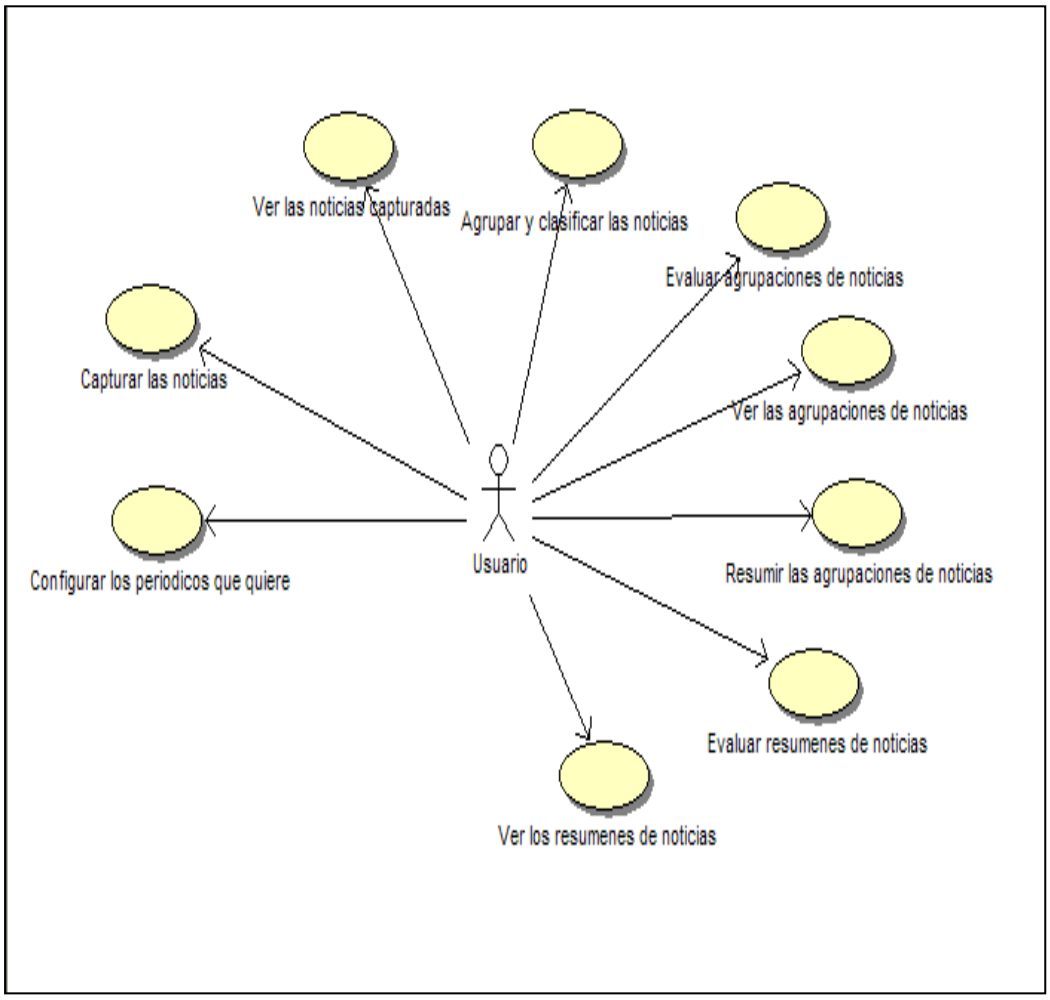


Figura 4.1: Diagrama caso de uso

En nuestro sistema, el único actor que existe es el usuario de la aplicación, y las distintas funcionalidades que puede realizar son:

1. Crear o modificar el archivo de configuración con los periódicos que se quieran utilizar.
2. Capturar las noticias, es decir, obtener las noticias de Internet que son almacenadas en nuestro sistema en archivos XML.
3. Visualización del contenido de todas las noticias capturadas anteriormente.
4. Agrupar las noticias por temas similares.
5. Visualización del resultado de las agrupaciones.
6. Evaluación de las diferentes agrupaciones realizadas anteriormente.
7. Resumir los grupos de noticias.
8. Visualización de los resúmenes.
9. Evaluar los resúmenes obtenidos.

4.2 - Diagrama de clases

En la figura 4.2 se muestra el diagrama de clases de nuestra aplicación sin tener en cuenta la parte de las interfaces gráficas del proyecto. En esta figura hemos incluido los principales atributos de cada clase y los métodos que consideramos más importantes, también se puede observar las distintas asociaciones que existen entre las clases.

La parte superior izquierda de la figura 4.2 se corresponde con el primer módulo que es el del capturador de noticias, en los siguientes puntos se explica detalladamente cada clase:

- Clase **Configuracion_periodico**: tiene como atributos la dirección de la página web del periódico ("paginaWeb") y los atributos necesarios que nos indican las distintas palabras clave para poder obtener correctamente las noticias ("palabrasClave", "Terminaciones").

- Clase **Capturador**: posee el nombre del archivo de salida de configuración ("ArchivoConfiguracionSalida") en el que se almacena la siguiente información: número de periódicos capturados, y para cada uno de ellos el número de noticias. Los métodos principales son capturaNoticias (se encarga de realizar todo el proceso de la captura), downloadURL (descarga el contenido de la pagina web que se desee), obtenerNoticia, crearArchivosXML (crea los archivos de tipo XML de todas las noticias). Esta clase se conecta a Internet a través del protocolo http. Además, el Capturador puede tener una o más clases Configuración_Periódico, dependiendo del número de periódicos.

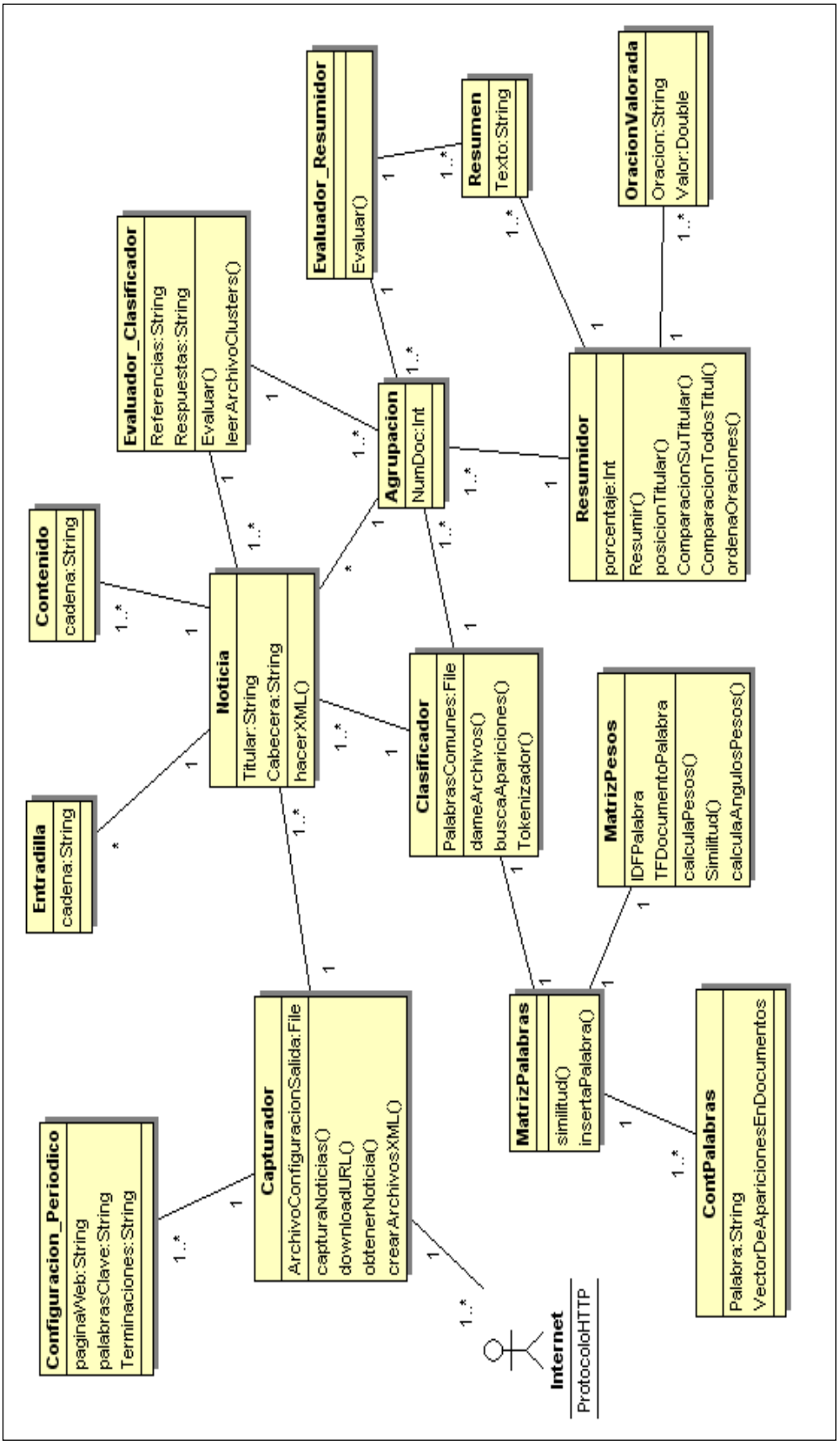


Figura 4.2: Diagrama de clases

- Clase **Noticia**: tiene los atributos que representan la estructura de la noticia: titular, cabecera, puede tener o no entradilla y al contenido le sucede lo mismo.

Las siguientes clases corresponden al módulo de la parte de la agrupación:

- Clase **Clasificador**: posee un atributo de tipo archivo en el que se encuentran las palabras comunes y una sola clase MatrizPalabras. El resultado que se obtiene como mínimo es una clase Agrupación.
  Los métodos más importantes son: dameArchivos (obtiene todos los archivos XML de las noticias), Tokenizador (se encarga de separar las palabras de cada documento) y buscaApariciones (cuenta el número de repeticiones de las palabras).

- Clase **MatrizPalabras**: contiene una sola matrizPesos y una o varias clases ContPalabras. Los métodos principales son: similitud (calcula la similitud entre los documentos ayudándose de la matriz de Pesos) e insertaPalabra (introduce la palabra en los correspondientes vectores).

- Clase **ContPalabras:** contiene la palabra y el vector de apariciones en todos los documentos.

- Clase **MatrizPesos**: posee los atributos IDF por palabra y el TF por documento/palabra. Los métodos principales son calculaPesos, similitud y calculaAngulosPesos.

- Clase **Agrupacion**: contiene todas las noticias pertenecientes a esa agrupación y el atributo "NumDoc" que representa el número de documentos del conjunto.

El módulo de la evaluación de las agrupaciones relaciona las clases Agrupacion,Noticia y Evaluador_Clasificador:

- Clase **Evaluador_Clasificador:** contiene todas las agrupaciones creadas por el sistema ("respuestas") y las realizadas manualmente ("referencias"), que son leídas de un archivo de texto a través del método "leerArchivoClusters".

El módulo de la parte del resumidor se corresponde con las clases Generación de resúmenes, Agrupación, Resumen, OracionValorada:

- Clase **Resumidor**: es la encargada de realizar toda la funcionalidad de este módulo. Para cada agrupación se crea una clase Resumen. Contiene un atributo porcentaje que indica el número de oraciones totales que resulta en el resumen. Los métodos posicionTitular, ComparacionSuTitular, ComparacionTodosTitulares calculan las distintas heurísticas. El método ordenaOraciones organiza las oraciones de mayor a menor según el valor calculado por las heurísticas.

- Clase **OracionValorada**: contiene un atributo "oracion" que representa una sola frase y el atributo "valor" que almacena el peso de la frase.

- Clase **Resumen**: contiene el resumen generado por la aplicación.

Capítulo 5: Manual de usuario

El presente manual está organizado de acuerdo a la secuencia de ingreso a las pantallas del sistema de la siguiente manera:

- Inicio del Sistema
- Captura de noticias
- Agrupación de noticias
- Resumen automático de noticias
- Requisitos mínimos del sistema

5.1- Inicio del sistema

En esta pantalla el usuario puede elegir entre tres opciones: capturar noticias, agrupar y resumir. Antes de seleccionar la opción de agrupación es necesario asegurarse de que hayan sido capturadas noticias previamente y que el archivo "archivoConfClasificador.xml" esté configurado correctamente de manera que el usuario puede añadir las noticias que desee a la carpeta noticias rellenando adecuadamente este archivo (más adelante detallaremos su construcción). Análogamente, antes de elegir la opción de resumir, previamente se tiene que agrupar para obtener los clusters que utiliza el generador de resúmenes. A continuación se muestra la interfaz principal (figura 5.1)

Figura 5.1 - Interfaz principal

Al pasar el cursor por encima de los textos, se muestra una pequeña ayuda dinámica. A continuación se muestra un ejemplo (Figura 5.2).

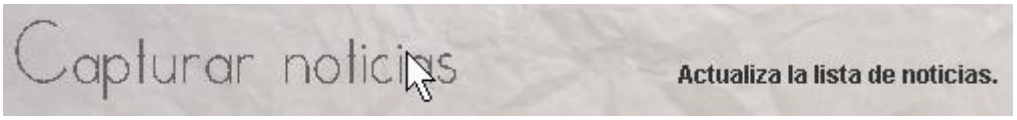


Figura 5.2 Fragmento de la interfaz principal

5.2 - Captura de noticias

Una vez seleccionada la opción de "Capturar noticias" en la ventana principal, se accede a la configuración de este apartado.

Figura 5.3- Interfaz de configuración de la captura de noticias

En la parte de la derecha de la figura 5.3 se pueden observar que periódicos están configurados actualmente.

Una vez hecha la configuración de los periódicos, en la pantalla de configuración de la captura (figura 5.3) podemos elegir el número máximo de noticias que deseas capturar por cada periódico configurado. Además cuentas con una opción para que se capturen todas las que el agrupador encuentre.

Al seleccionar la opción de configurar periódicos se muestra la siguiente pantalla (figura 5.4). En esta pantalla es posible configurar, para cada periódico del que se desee extraer noticias, la siguiente información:

Figura 5.4- Interfaz de configuración de los periódicos

- Página web.- Dirección de la página web del periódico, incluyendo el protocolo "http://".

- Palabra clave referencias.- Todos los enlaces que contengan esta cadena de texto son considerados referencias a noticias.

- Palabras clave de título.- Pueden aparecer una o más cadenas de texto separadas por comas. Cuando el sistema encuentra en el HTML de las noticias estas palabras claves, entonces comienza el contenido del título.

- Terminación título.- Cadena de texto que indica el final del contenido del título.

- ¿Tiene entradilla?.- Se marca cuando la noticia tiene entradilla.

- Palabras clave de entradilla.- Análogo a "Palabras clave de título" solo que se continúa procesando desde la terminación del título.

- Terminación entradilla.- Análogo a la terminación de título.

- ¿Tiene cabecera?.- Se marca cuando la noticia tiene cabecera.

- Palabras clave de cabecera.- Análogo a "Palabras clave de entradilla" solo que se continúa procesando desde la última terminación.

- Terminación cabecera.- Análogo a la terminaciones explicadas para el título y la entradilla.

- Palabras clave de contenido.- Análogo a "Palabras clave de cabecera" solo que se continúa procesando desde la última terminación.

- Terminación contenido.- Análogo a las terminaciones ya explicadas.

- Lista de periódicos.- Lista de periódicos ya configurados.

A continuación se explica las funcionalidades de esta interfaz.

- Menú – Archivo:

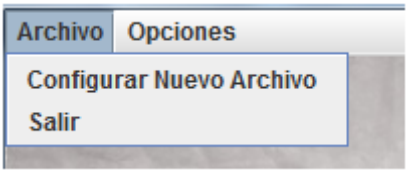


Figura 5.5- Menú de la configuración de periódicos

- Configurar Nuevo Archivo.- Al pulsar se borra el contenido del archivo de configuración de XML.

- Salir.- Cierra la ventana actual.

- Menú – Opciones:

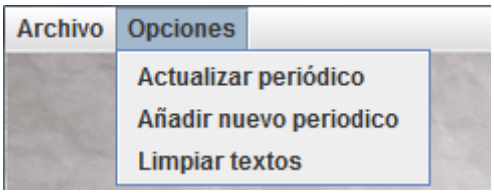


Figura 5.6- Menú de la configuración de periódicos

- Actualizar periódico.- Actualiza los contenidos del periódico seleccionado en la "Lista de periódicos" con los valores que haya en los campos de texto. Esta opción actualiza el fichero de configuración de periódicos.

- Añadir nuevo periódico.- Inserta un nuevo periódico con los valores que haya en los campos de texto.

- Limpiar textos.- Borra los campos de texto.

Una vez hecha la configuración deseada, regresamos a la pantalla de la figura 5.3, donde podemos pulsar el botón capturar y realizar una extracción de noticias. El tiempo que consume este proceso depende de la cantidad de periódicos configurados y el número de noticias seleccionadas por periódico. Al finalizar el proceso de extracción, aparece la siguiente ventana:

Figura 5.7 - Interfaz mostrando los resultados de la captura de noticias

Donde se muestra la información de la captura, es decir, se indica cuantos periódicos se han consultado y cuantas noticias han capturado de cada uno. Además, este proceso genera un fichero llamado "archivoConfClasificador.xml" (figura 5.8) que contiene esta misma información, que se puede modificar de una forma sencilla como se explicará más adelante por si se quiere añadir alguna noticia para que la agrupe con las ya capturadas (se deberá generar el fichero XML de forma manual e introducirlo en la carpeta noticias). Todas las noticias se guardan en la carpeta "Noticias" en formato XML.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
- <Configuracion>
    <NumPeriodicos> 3</NumPeriodicos>
    <Periodico>elpais.com</Periodico>
    <Noticias>102</Noticias>
    <Periodico>elmundo.es</Periodico>
    <Noticias>28</Noticias>
    <Periodico>larazon.es</Periodico>
    <Noticias>270</Noticias>
  </Configuracion>
```

Figura 5.8- Archivo de configuración para la agrupación de noticias

Para volver a la pantalla principal existen dos opciones, cerrar la ventana o presionar el botón Aceptar.

Explicación del "archivoConfClasificador.xml" (figura 5.8) para el clasificador: En la primera línea debe aparecer la cabecera XML que será la mostrada en la figura 5.8, a continuación aparece la etiqueta general <*Configuracion*>, seguida de la etiqueta <*NumPeriodicos*> que indica el número de periódicos que se explorarán. Las siguientes líneas, con las etiquetas <*Periodico*> indican el nombre del periódico y las que tienen la etiqueta <*Noticias*> representan la cantidad de noticias que va a leer el clasificador de cada periódico.

El clasificador toma cada línea para leer los archivos "nombre del periódico" + i +.xml siendo i desde 0 hasta el número de noticias que tiene el periódico.

En este ejemplo, en la segunda línea el clasificador interpreta que tiene que leer los archivos "elpais.com0.xml" hasta "elpais.com101.xml".

Es muy importante tener este archivo coordinado con las noticias en nuestro formato XML en el caso de que el usuario quisiera configurarlo manualmente.

5.3 - Agrupación de noticias

Al pulsar la opción agrupar, aparece la ventana mostrada en la figura 5.9. En la parte de la derecha observamos los resultados de la captura que tiene configurada en ese momento (número de periódicos y número de noticias por periódico), además también aparecen campos donde se puede configurar los parámetros del agrupador:

Figura 5.9 - Interfaz de la configuración de la agrupación de noticias

- Umbral similitud.- Valor entre 0 y 1 para comprobar la similitud de los vectores.

Al igual que en la pantalla principal, al situar el puntero del ratón sobre los textos aparecen ayudas dinámicas.

Pulsamos el botón Agrupar con los parámetros elegidos y se muestra el resultado en la siguiente pantalla (figuras 5.10 o 5.11):

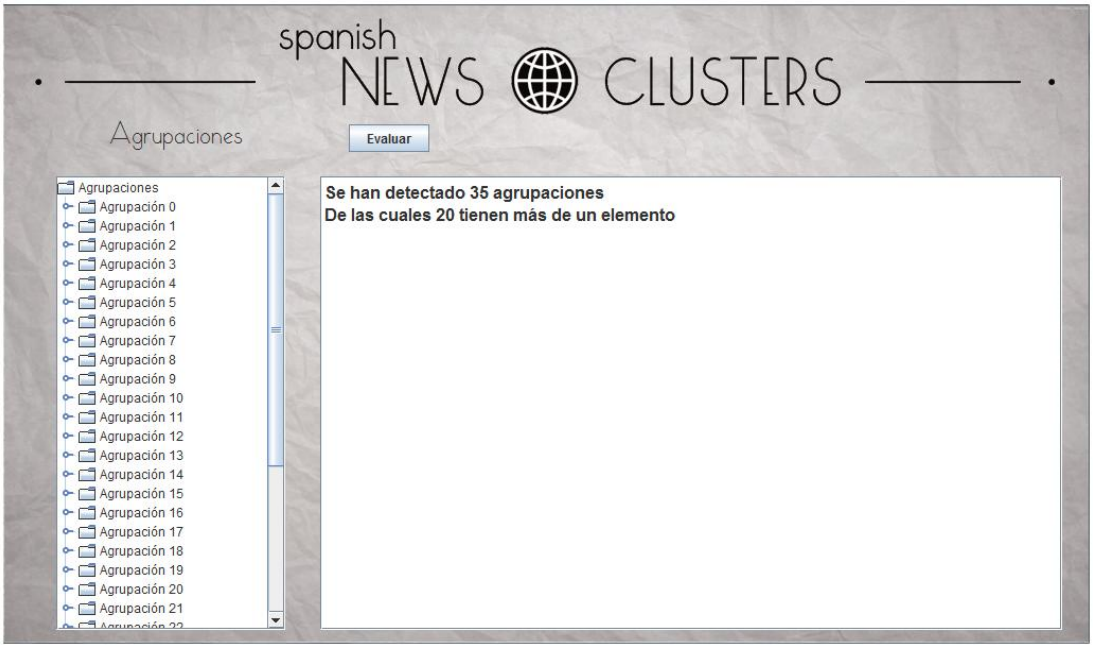


Figura 5.10- Resultados de la agrupación

En primer lugar aparece el resultado de la agrupación, es decir, cuántos grupos ha generado y cuáles de ellos tienen más de un elemento (figura 5.10). Para ver las noticias que componen un conjunto se pulsa la pestaña de la agrupación deseada y aparece un subárbol. Seleccionamos una noticia del subárbol haciendo doble click sobre ella se muestra su contenido (figura 5.11).

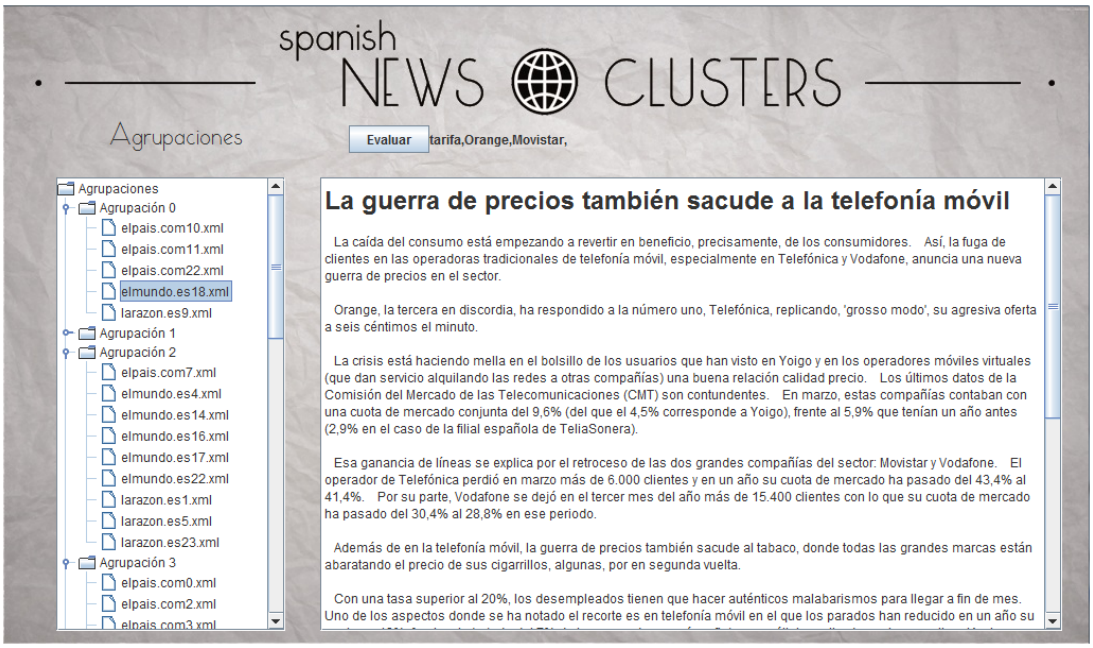


Figura 5.11- Resultados de la agrupación

En la parte superior podemos ver cuáles son las palabras claves de la agrupación que contiene a la noticia seleccionada y podemos seleccionar la opción evaluar, para evaluar la calidad de la agrupación. Para ello debemos asegurarnos de haber rellenado correctamente el fichero que contiene la agrupación de referencia, tal como se explica en el apartado 3.1-Evaluación del algoritmo de agrupación. El resultado obtenido será el siguiente (figura 5.12).

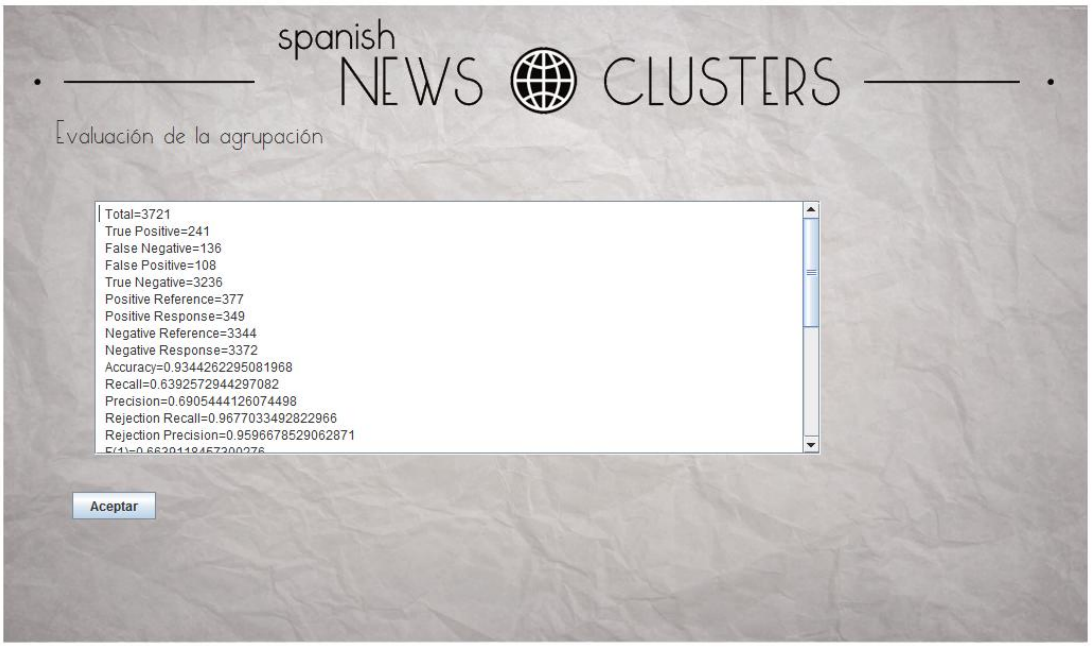


Figura 5.12 - Interfaz de la evaluación de la clasificación

5.4 - Resumen automático de noticias

Desde la interfaz principal presionamos la opción Resumir y surge la siguiente ventana (figura 5.13) para realizar la configuración:

Figura 5.13 - Interfaz configuración resúmenes multi-documento

El módulo del generador de resúmenes utiliza las heurísticas seleccionadas en la pantalla. Se introduce el valor del peso que se quiere dar a cada heurística. Todos los valores tienen que ser enteros mayores o iguales que cero.

El porcentaje de número de oraciones es un valor comprendido entre 1 y 100 que determina el número de oraciones que tendrá el resumen, y su valor es el tanto por ciento del total de las oraciones de todos los documentos recogidos en el grupo.

En la parte derecha de la ventana nos encontramos con las opciones para guardar, cargar y borrar heurísticas, con esta opción podemos guardar los pesos que deseamos para cada heurística y cargarlos cuando queramos.

Al completar las heurísticas se presiona el botón resumir y a continuación emerge la siguiente ventana (figura 5.14 o 5.15):

Figura 5.14 - Interfaz resultado de resúmenes multi-documento

Análogamente a la interfaz del clasificador, podemos expandir cada resumen en las noticias que lo han generado. Como se ha comentado anteriormente, se puede visualizar el resumen o las noticias haciendo doble click o seleccionándola y pulsando el botón mostrar. Al seleccionar una noticia o resumen en la parte superior aparecerán las palabras clave que lo identifican (Figura 5.15).

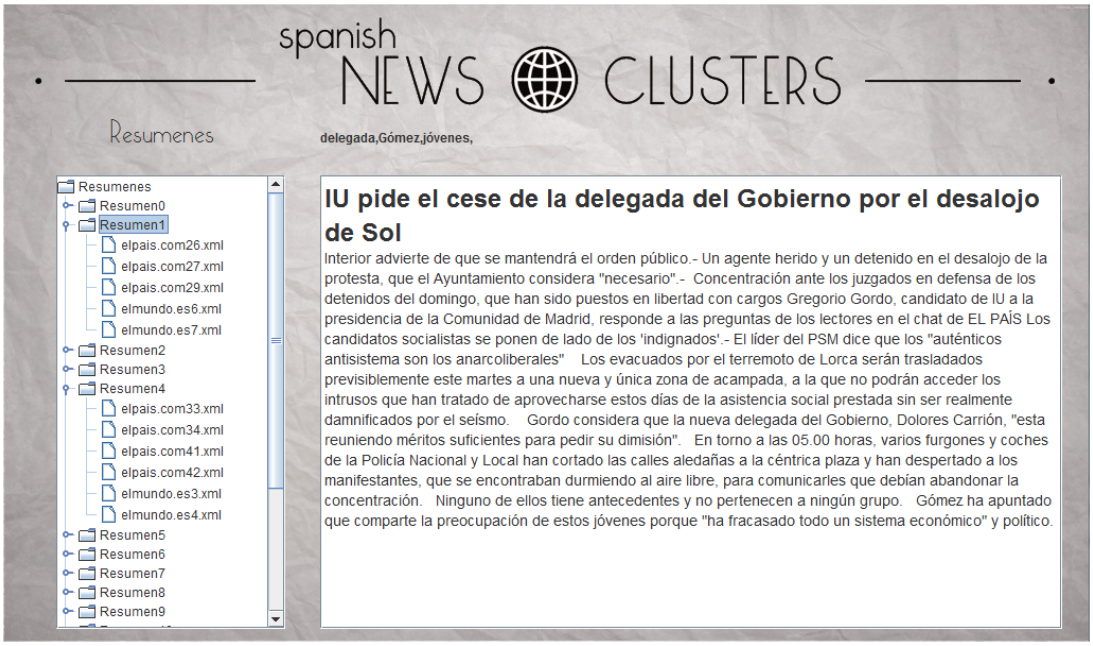


Figura 5.15 - Interfaz resultado de resúmenes multi-documento

5.5 - Requerimientos.

- Sistema Operativo Windows XP o Superior
- Resolución gráfica mínima 1024*600
- Conexión a Internet (para el capturador)

Capítulo 6: Conclusiones

Nuestros dos objetivos principales en este proyecto han sido el desarrollo de un generadorautomático de resúmenes de las noticias de los principales periódicos españoles, y poder evaluar de una manera automática los resultados generados por la aplicación.

Para conseguir el generador automático de resúmenes, hemos necesitado desarrollar un capturador para obtener las noticias de las páginas web de los periódicos y un agrupador de documentos, al que le introducimos las noticias capturadas y nos devuelve grupos de noticias que tratan un mismo tema.

Además una de las principales características del sistema es que es muy configurable, de manera que el usuario pueda añadir nuevos periódicos o cambiar distintos valores para tratar de obtener un resultado óptimo.

En el caso de la evaluación de los resultados, no solo hemos conseguido evaluarlos, sino que en el caso de la evaluación de las agrupaciones la hemos integrado en nuestro sistema, siendo accesible desde la interfaz.

6.1- Resultados obtenidos

La aplicación desarrollada tiene diferentes propiedades que pueden hacer variar el resultado de la ejecución. En este apartado explicaremos cuales pueden ser los mejores valores para conseguir un buen funcionamiento.

6.1.1 Resultados de la captura de noticias

Para poder capturar noticias se pueden configurar dos aspectos, la primera es el número de noticias que se capturan como máximo de

cada periódico, y la segunda los periódicos en los que buscar las noticias.

Es el módulo que más tarda en ejecutarse, ya que es muy costoso en tiempo descargar el código HTML de cada noticia y parsearlo para obtener el contenido. Por eso, está pensado para ser ejecutado una vez al día, por ejemplo, por la mañana realizar una descarga de la web y a partir de esas noticias capturadas, realizar distintos experimentos tanto con las agrupaciones como con la generación de resúmenes. Una manera de reducir el tiempo de ejecución es disminuir el número de noticias máximas a buscar.

6.1.2 Resultados de la agrupación de noticias

El agrupador se encarga de generar las matrices de vectores para la generación de las agrupaciones. Esta es posible gracias a la comparación vectorial de documentos, permitiendo modificar el umbral de similitud para obtener agrupaciones más precisas. Normalmente, cuantos más documentos, más vectores se crearán y la matriz de palabras será mucho mayor, con lo que podremos aumentar el valor del umbral para obtener mejores resultados.

Si queremos obtener una mayor precisión en las agrupaciones basta con aumentar el valor de similitud, aunque también aumentamos el riesgo de no detectar alguna agrupación correcta. Si por el contrario, queremos hacerlo menos restrictivo, solo tendremos que disminuir el valor, con lo que aumenta también la posibilidad de que se forme alguna agrupación con documentos que no traten el mismo contenido.

En conclusión, a medida que se aumenta el número de noticias en el sistema, éste funcionará mejor, ya que al haber más noticias hay más variaciones de contenido, con lo que los vectores que contienen los valores TF e IDF se parecerán mucho más entre

sí, si contienen palabras parecidas, con lo que la función de similitud funcionará mejor.

6.1.3 Resultados de la generación de resúmenes automáticos

Un aspecto importante en la generación de resúmenes es partir de una buena agrupación, ya que si tenemos documentos que tratan contenidos distintos, el resumen nunca será coherente.

El generador de resúmenes usa distintas heurísticas para funcionar, en la aplicación se permite elegir entre cuatro distintas. Es muy importante combinar correctamente la elección de las heurísticas, ya que el resumen obtenido será de mejor calidad.

Lo mejor para obtener un buen resumen es ir probando con distintos pesos en las heurísticas hasta obtener un resultado satisfactorio. Normalmente se generan mejores resúmenes dando mayor peso a las heurísticas "comparación con todos los titulares" y "comparación titular resumen", luego a "comparación con su titular", y por último a "posición".

6.2- Problemas encontrados

El primer problema al que nos enfrentamos fue a la hora de capturar noticias, debido a que cada periódico cuenta con un código fuente estructurado de forma totalmente distinta para mostrar sus noticias, por lo que nos vimos obligados a añadir un fichero de configuración en el que indicar cómo poder encontrar el contenido de cada noticia.

Otro inconveniente de la captura de noticias es el tiempo de espera a la hora de descargar el código fuente de cada noticia, que a pesar de ser alto es un problema menor, ya que este módulo está pensado para ser ejecutado una vez al día.

En cuanto al umbral de similitud en la agrupación, en ocasiones, es un problema que no haya valores predeterminados. Es una configuración que depende de las noticias capturadas, del número de documentos, número de palabras y no se puede obtener de ninguna forma de antemano.

La generación de resúmenes es un módulo complejo, ya que un resumen multi-documento es una tarea muy subjetiva, y el sistema a diferencia de un ser humano solo puede basarse en la posición que ocupa cierta frase o lo similar de una frase a otra, pero no es capaz de conocer el significado en sí de cada frase. Para solucionar este problema pensamos que la mejor manera era disponer de varias heurísticas, y poder seleccionar distintos pesos para cada una, de manera que se puedan ver distintos resultados y decidir cuál es el más acertado. También nos encontramos con que las heurísticas no son perfectas, es decir, son simplemente maneras para valorar con un peso cada oración de cada documento y así saber qué posición toman de cara al resumen, por lo que, la elección de ellas también depende de la cantidad de noticias que haya en la agrupación, de la cantidad de oraciones que haya en cada documento y, por lo tanto, no existen valores predeterminados.

Por último, la inclusión de las evaluaciones en el sistema nos ha generado grandes dificultades, sobre todo a la hora de evaluar la generación de resúmenes, que requiere de un software (*ActivePerl*) que nos ha causado bastantes problemas su instalación, por lo que dicha evaluación la hemos realizado de forma independiente del sistema.

6.3- Trabajo futuro

Durante la etapa de la agrupación ya hemos comentado anteriormente que el tamaño de las matrices puede ser muy elevado (ya que es fácil encontrarnos con noticias con más de 500 palabras). Una forma de disminuir el tamaño puede ser utilizar algoritmos ya existentes (por ejemplo el algoritmo de *Porter*), para poder extraer las raíces de las palabras. De esta forma, palabras como "persona" y "personas" son la misma en la matriz.

Otra mejora, que reduce el tamaño de las matrices, es tener en cuenta los sinónimos, de modo que eliminaríamos palabras que tengan el mismo significado en la matriz. Existe una librería Java, denominada *Lucene*, que obtiene una representación matricial de frecuencias de los documentos más eficiente que nuestro sistema. Este software realiza la extracción de raíces, una eliminación muy completa de las palabras sin significado, etc. Esta librería está incluida en el proyecto, pero finalmente ha sido eliminada de las funcionalidades de la interfaz debido a que, a pesar de ser más eficiente, no estaba suficientemente comprobado que el funcionamiento fuese el correcto en nuestro sistema.

La principal mejora que se puede llevar a cabo en la generación de resúmenes es desarrollar nuevas heurísticas, por ejemplo, tener en cuenta los valores TF e IDF de cada palabra de cada frase. Además se integraría la evaluación de los resúmenes en el sistema, de manera que se pudiese ejecutar desde la interfaz de la aplicación.

Finalmente, una vez obtenido el resumen, el sistema podría tener una opción que obtuviera un documento en HTML con los resultados del sistema, las capturas de las noticias, las agrupaciones y los resúmenes de cada agrupación, para poder subirlo a la web.

Anexo 1: Casos de prueba para la evaluación del agrupador

Agrupación con umbral de similitud 0.1

```
<?xmlversion="1.0" encoding="ISO-8859-1"?>
<Configuracion>
        <NumCluster>6</NumCluster>
        <Agrupacion>
                <NumNoticias>54</NumNoticias>
                <Noticia>elpais.com0.xml</Noticia>
                <Noticia>elpais.com1.xml</Noticia>
                <Noticia>elpais.com2.xml</Noticia>
                <Noticia>elpais.com3.xml</Noticia>
                <Noticia>elpais.com4.xml</Noticia>
                <Noticia>elpais.com5.xml</Noticia>
                <Noticia>elpais.com6.xml</Noticia>
                <Noticia>elpais.com7.xml</Noticia>
                <Noticia>elpais.com8.xml</Noticia>
                <Noticia>elpais.com9.xml</Noticia>
                <Noticia>elpais.com10.xml</Noticia>
                <Noticia>elpais.com11.xml</Noticia>
                <Noticia>elpais.com12.xml</Noticia>
                <Noticia>elpais.com13.xml</Noticia>
                <Noticia>elpais.com14.xml</Noticia>
                <Noticia>elpais.com15.xml</Noticia>
                <Noticia>elpais.com16.xml</Noticia>
                <Noticia>elpais.com17.xml</Noticia>
                <Noticia>elpais.com19.xml</Noticia>
                <Noticia>elpais.com20.xml</Noticia>
                <Noticia>elpais.com22.xml</Noticia>
                <Noticia>elpais.com24.xml</Noticia>
                <Noticia>elpais.com25.xml</Noticia>
                <Noticia>elpais.com26.xml</Noticia>
                <Noticia>elpais.com27.xml</Noticia>
                <Noticia>elpais.com29.xml</Noticia>
                <Noticia>elpais.com30.xml</Noticia>
                <Noticia>elpais.com31.xml</Noticia>
                <Noticia>elpais.com32.xml</Noticia>
                <Noticia>elpais.com33.xml</Noticia>
                <Noticia>elpais.com34.xml</Noticia>
                <Noticia>elpais.com36.xml</Noticia>
                <Noticia>elpais.com37.xml</Noticia>
```

```
                <Noticia>elpais.com38.xml</Noticia>
                <Noticia>elpais.com40.xml</Noticia>
                <Noticia>elpais.com41.xml</Noticia>
                <Noticia>elpais.com42.xml</Noticia>
                <Noticia>elpais.com43.xml</Noticia>
                <Noticia>elmundo.es0.xml</Noticia>
                <Noticia>elmundo.es1.xml</Noticia>
                <Noticia>elmundo.es3.xml</Noticia>
                <Noticia>elmundo.es4.xml</Noticia>
                <Noticia>elmundo.es5.xml</Noticia>
                <Noticia>elmundo.es6.xml</Noticia>
                <Noticia>elmundo.es7.xml</Noticia>
                <Noticia>larazon.es0.xml</Noticia>
                <Noticia>larazon.es1.xml</Noticia>
                <Noticia>larazon.es2.xml</Noticia>
                <Noticia>larazon.es3.xml</Noticia>
                <Noticia>larazon.es4.xml</Noticia>
                <Noticia>larazon.es5.xml</Noticia>
                <Noticia>larazon.es6.xml</Noticia>
                <Noticia>larazon.es7.xml</Noticia>
                <Noticia>larazon.es8.xml</Noticia>
                <Claves>
                        <Palabra>decisiones</Palabra>
                        <Palabra>Qaeda</Palabra>
                        <Palabra>enero</Palabra>
                </Claves>
</Agrupacion>
<Agrupacion>
                <NumNoticias>2</NumNoticias>
                <Noticia>elpais.com18.xml</Noticia>
                <Noticia>elpais.com21.xml</Noticia>
                <Claves>
                        <Palabra>Skype</Palabra>
                        <Palabra>Microsoft</Palabra>
                        <Palabra>operadoras</Palabra>
                </Claves>
</Agrupacion>
<Agrupacion>
                <NumNoticias>2</NumNoticias>
                <Noticia>elpais.com35.xml</Noticia>
                <Noticia>elmundo.es2.xml</Noticia>
                <Claves>
                        <Palabra>Endeavour</Palabra>
                        <Palabra>Stormy</Palabra>
                        <Palabra>Mondays</Palabra>
                </Claves>
</Agrupacion>
```

```
<Agrupacion>
		<NumNoticias>1</NumNoticias>
		<Noticia>elpais.com23.xml</Noticia>
		<Claves>
			<Palabra>IP</Palabra>
			<Palabra>canal</Palabra>
			<Palabra>acciones</Palabra>
		</Claves>
</Agrupacion>

<Agrupacion>
		<NumNoticias>1</NumNoticias>
		<Noticia>elpais.com28.xml</Noticia>
		<Claves>
			<Palabra>calle</Palabra>
			<Palabra>manifestaciones</Palabra>
			<Palabra>principales</Palabra>
		</Claves>
</Agrupacion>
<Agrupacion>
		<NumNoticias>1</NumNoticias>
		<Noticia>elpais.com39.xml</Noticia>
		<Claves>
			<Palabra>denunció</Palabra>
			<Palabra>Barça</Palabra>
			<Palabra>resolución</Palabra>
		</Claves>
</Agrupacion>
</Configuracion>
```

Tabla 1.1 - Agrupación con umbral de similitud 0.1

Agrupación con umbral de similitud 0.2

```
<?xmlversion="1.0" encoding="ISO-8859-1"?>
<Configuracion>
	<NumCluster>20</NumCluster>
	<Agrupacion>
		<NumNoticias>11</NumNoticias>
		<Noticia>elpais.com10.xml</Noticia>
		<Noticia>elpais.com11.xml</Noticia>
		<Noticia>elpais.com12.xml</Noticia>
		<Noticia>elpais.com13.xml</Noticia>
		<Noticia>elpais.com17.xml</Noticia>
		<Noticia>elpais.com20.xml</Noticia>
		<Noticia>elpais.com25.xml</Noticia>
		<Noticia>elpais.com40.xml</Noticia>
		<Noticia>elpais.com43.xml</Noticia>
		<Noticia>larazon.es4.xml</Noticia>
		<Noticia>larazon.es5.xml</Noticia>
		<Claves>
			<Palabra>Qaeda</Palabra>
			<Palabra>peor</Palabra>
			<Palabra>pederastas</Palabra>
		</Claves>
</Agrupacion>
	<Agrupacion>
		<NumNoticias>5</NumNoticias>
		<Noticia>elpais.com26.xml</Noticia>
		<Noticia>elpais.com27.xml</Noticia>
		<Noticia>elpais.com29.xml</Noticia>
		<Noticia>elmundo.es6.xml</Noticia>
		<Noticia>elmundo.es7.xml</Noticia>
		<Claves>
			<Palabra>jóvenes</Palabra>
			<Palabra>delegada</Palabra>
			<Palabra>Gómez</Palabra>
		</Claves>
</Agrupacion>
	<Agrupacion>
		<NumNoticias>8</NumNoticias>
		<Noticia>elpais.com0.xml</Noticia>
		<Noticia>elpais.com1.xml</Noticia>
		<Noticia>elpais.com3.xml</Noticia>
		<Noticia>elpais.com15.xml</Noticia>
		<Noticia>elpais.com16.xml</Noticia>
		<Noticia>elmundo.es0.xml</Noticia>
		<Noticia>larazon.es0.xml</Noticia>
		<Noticia>larazon.es1.xml</Noticia>
		<Claves>
			<Palabra>Manuel</Palabra>
```

```
                        <Palabra>sísmica</Palabra>
                        <Palabra>escombros</Palabra>
                </Claves>
</Agrupacion>
        <Agrupacion>
                <NumNoticias>7</NumNoticias>
                <Noticia>elpais.com4.xml</Noticia>
                <Noticia>elpais.com6.xml</Noticia>
                <Noticia>elpais.com7.xml</Noticia>
                <Noticia>elpais.com8.xml</Noticia>
                <Noticia>elpais.com19.xml</Noticia>
                <Noticia>elmundo.es1.xml</Noticia>
                <Noticia>larazon.es2.xml</Noticia>
                <Claves>
                        <Palabra>decisiones</Palabra>
                        <Palabra>Zapatero</Palabra>
                        <Palabra>Constitucional</Palabra>
                </Claves>
</Agrupacion>
        <Agrupacion>
                <NumNoticias>6</NumNoticias>
                <Noticia>elpais.com33.xml</Noticia>
                <Noticia>elpais.com34.xml</Noticia>
                <Noticia>elpais.com41.xml</Noticia>
                <Noticia>elpais.com42.xml</Noticia>
                <Noticia>elmundo.es3.xml</Noticia>
                <Noticia>elmundo.es4.xml</Noticia>
                <Claves>
                        <Palabra>Strauss</Palabra>
                        <Palabra>FMI</Palabra>
                        <Palabra>Kahn</Palabra>
                </Claves>
</Agrupacion>
        <Agrupacion>
                <NumNoticias>2</NumNoticias>
                <Noticia>elpais.com9.xml</Noticia>
                <Noticia>larazon.es3.xml</Noticia>
                <Claves>
                        <Palabra>empresas</Palabra>
                        <Palabra>bancos</Palabra>
                        <Palabra>exigir</Palabra>
                </Claves>
</Agrupacion>
        <Agrupacion>
                <NumNoticias>4</NumNoticias>
                <Noticia>elpais.com36.xml</Noticia>
                <Noticia>elpais.com37.xml</Noticia>
                <Noticia>elpais.com38.xml</Noticia>
                <Noticia>elmundo.es5.xml</Noticia>
```

```
                <Claves>
                        <Palabra>Pitt</Palabra>
                        <Palabra>árbol</Palabra>
                        <Palabra>Malick</Palabra>
                </Claves>
</Agrupacion>
        <Agrupacion>
                <NumNoticias>4</NumNoticias>
                <Noticia>elpais.com30.xml</Noticia>
                <Noticia>elpais.com31.xml</Noticia>
                <Noticia>elpais.com32.xml</Noticia>
                <Noticia>larazon.es7.xml</Noticia>
                <Claves>
                        <Palabra>Isabel</Palabra>
                        <Palabra>visita</Palabra>
                        <Palabra>Irlanda</Palabra>
                </Claves>
</Agrupacion>
        <Agrupacion>
                <NumNoticias>2</NumNoticias>
                <Noticia>elpais.com18.xml</Noticia>
                <Noticia>elpais.com21.xml</Noticia>
                <Claves>
                        <Palabra>Skype</Palabra>
                        <Palabra>Microsoft</Palabra>
                        <Palabra>operadoras</Palabra>
                </Claves>
</Agrupacion>
        <Agrupacion>
                <NumNoticias>2</NumNoticias>
                <Noticia>elpais.com35.xml</Noticia>
                <Noticia>elmundo.es2.xml</Noticia>
                <Claves>
                        <Palabra>Endeavour</Palabra>
                        <Palabra>Stormy</Palabra>
                        <Palabra>Mondays</Palabra>
                </Claves>
</Agrupacion>
        <Agrupacion>
                <NumNoticias>1</NumNoticias>
                <Noticia>elpais.com2.xml</Noticia>
                <Claves>
                        <Palabra>reconstrucción</Palabra>
                        <Palabra>Zapatero</Palabra>
                        <Palabra>apoyo</Palabra>
                </Claves>
       </Agrupacion>
```

```
<Agrupacion>
		<NumNoticias>1</NumNoticias>
		<Noticia>elpais.com5.xml</Noticia>
		<Claves>
			<Palabra>PNV</Palabra>
			<Palabra>EA</Palabra>
			<Palabra>Navarra</Palabra>
		</Claves>
	</Agrupacion>
	 <Agrupacion>
		<NumNoticias>1</NumNoticias>
		<Noticia>elpais.com14.xml</Noticia>
		<Claves>
			<Palabra>Ambos</Palabra>
			<Palabra>Google</Palabra>
			<Palabra>Samsung</Palabra>
		</Claves>
	</Agrupacion>
	 <Agrupacion>
		<NumNoticias>1</NumNoticias>
		<Noticia>elpais.com22.xml</Noticia>
		<Claves>
			<Palabra>partir</Palabra>
			<Palabra>Google</Palabra>
			<Palabra>versión</Palabra>
		</Claves>
	</Agrupacion>
	 <Agrupacion>
		<NumNoticias>1</NumNoticias>
		<Noticia>elpais.com23.xml</Noticia>
		<Claves>
			<Palabra>IP</Palabra>
			<Palabra>canal</Palabra>
			<Palabra>acciones</Palabra>
		</Claves>
	</Agrupacion>
	 <Agrupacion>
		<NumNoticias>1</NumNoticias>
		<Noticia>elpais.com24.xml</Noticia>
		<Claves>
			<Palabra>Ley</Palabra>
			<Palabra>enero</Palabra>
			<Palabra>resoluciones</Palabra>
		</Claves>
	</Agrupacion>
	 <Agrupacion>
		<NumNoticias>1</NumNoticias>
		<Noticia>elpais.com28.xml</Noticia>
		<Claves>
```

```
                        <Palabra>calle</Palabra>
                        <Palabra>manifestaciones</Palabra>
                        <Palabra>principales</Palabra>
                </Claves>
        </Agrupacion>
         <Agrupacion>
                <NumNoticias>1</NumNoticias>
                <Noticia>elpais.com39.xml</Noticia>
                <Claves>
                        <Palabra>denunció</Palabra>
                        <Palabra>Barça</Palabra>
                        <Palabra>resolución</Palabra>
                </Claves>
        </Agrupacion>
         <Agrupacion>
                <NumNoticias>1</NumNoticias>
                <Noticia>larazon.es6.xml</Noticia>
                <Claves>
                        <Palabra>hoja</Palabra>
                        <Palabra>presenta</Palabra>
                        <Palabra>Navarra</Palabra>
                </Claves>
        </Agrupacion>
         <Agrupacion>
                <NumNoticias>1</NumNoticias>
                <Noticia>larazon.es8.xml</Noticia>
                <Claves>
                        <Palabra>economía</Palabra>
                        <Palabra>empleo</Palabra>
                        <Palabra>saliendo</Palabra>
                </Claves>
</Agrupacion>
</Configuracion>
```

Tabla 1.2: Agrupación con umbral similitud 0.2

Agrupación con umbral de similitud 0.3

```
<?xmlversion="1.0" encoding="ISO-8859-1"?>
<Configuracion>
        <NumCluster>32</NumCluster>
        <Agrupacion>
                <NumNoticias>3</NumNoticias>
                <Noticia>elpais.com4.xml</Noticia>
                <Noticia>elpais.com6.xml</Noticia>
                <Noticia>elpais.com8.xml</Noticia>
                <Claves>
                        <Palabra>decisiones</Palabra>
                        <Palabra>Zapatero</Palabra>
                        <Palabra>Constitucional</Palabra>
                </Claves>
        </Agrupacion>
        <Agrupacion>
                <NumNoticias>7</NumNoticias>
                <Noticia>elpais.com10.xml</Noticia>
                <Noticia>elpais.com11.xml</Noticia>
                <Noticia>elpais.com12.xml</Noticia>
                <Noticia>elpais.com17.xml</Noticia>
                <Noticia>elpais.com20.xml</Noticia>
                <Noticia>elpais.com25.xml</Noticia>
                <Noticia>larazon.es5.xml</Noticia>
                <Claves>
                        <Palabra>preparado</Palabra>
                        <Palabra>Mallorca</Palabra>
                        <Palabra>pederastas</Palabra>
                </Claves>
        </Agrupacion>
        <Agrupacion>
                <NumNoticias>7</NumNoticias>
                <Noticia>elpais.com0.xml</Noticia>
                <Noticia>elpais.com1.xml</Noticia>
                <Noticia>elpais.com15.xml</Noticia>
                <Noticia>elpais.com16.xml</Noticia>
                <Noticia>elmundo.es0.xml</Noticia>
                <Noticia>larazon.es0.xml</Noticia>
                <Noticia>larazon.es1.xml</Noticia>
                <Claves>
                        <Palabra>Carreño</Palabra>
                        <Palabra>técnicos</Palabra>
                        <Palabra>sísmica</Palabra>
                </Claves>
        </Agrupacion>
        <Agrupacion>
                <NumNoticias>6</NumNoticias>
                <Noticia>elpais.com33.xml</Noticia>
```

```
            <Noticia>elpais.com34.xml</Noticia>
            <Noticia>elpais.com41.xml</Noticia>
            <Noticia>elpais.com42.xml</Noticia>
            <Noticia>elmundo.es3.xml</Noticia>
            <Noticia>elmundo.es4.xml</Noticia>
            <Claves>
                  <Palabra>Strauss</Palabra>
                  <Palabra>FMI</Palabra>
                  <Palabra>Kahn</Palabra>
            </Claves>
      </Agrupacion>
       <Agrupacion>
            <NumNoticias>2</NumNoticias>
            <Noticia>elpais.com37.xml</Noticia>
            <Noticia>elmundo.es5.xml</Noticia>
            <Claves>
                  <Palabra>Malick</Palabra>
                  <Palabra>media</Palabra>
                  <Palabra>Terrence</Palabra>
            </Claves>
      </Agrupacion>
       <Agrupacion>
            <NumNoticias>4</NumNoticias>
            <Noticia>elpais.com26.xml</Noticia>
            <Noticia>elpais.com27.xml</Noticia>
            <Noticia>elpais.com29.xml</Noticia>
            <Noticia>elmundo.es6.xml</Noticia>
            <Claves>
                  <Palabra>jóvenes</Palabra>
                  <Palabra>delegada</Palabra>
                  <Palabra>Gómez</Palabra>
            </Claves>
      </Agrupacion>
       <Agrupacion>
            <NumNoticias>2</NumNoticias>
            <Noticia>elpais.com40.xml</Noticia>
            <Noticia>larazon.es4.xml</Noticia>
            <Claves>
                  <Palabra>Qaeda</Palabra>
                  <Palabra>llegar</Palabra>
                  <Palabra>peor</Palabra>
            </Claves>
      </Agrupacion>
       <Agrupacion>
            <NumNoticias>4</NumNoticias>
            <Noticia>elpais.com30.xml</Noticia>
            <Noticia>elpais.com31.xml</Noticia>
            <Noticia>elpais.com32.xml</Noticia>
            <Noticia>larazon.es7.xml</Noticia>
```

```
            <Claves>
                  <Palabra>Isabel</Palabra>
                  <Palabra>visita</Palabra>
                  <Palabra>Irlanda</Palabra>
            </Claves>
      </Agrupacion>
       <Agrupacion>
            <NumNoticias>2</NumNoticias>
            <Noticia>elpais.com18.xml</Noticia>
            <Noticia>elpais.com21.xml</Noticia>
            <Claves>
                  <Palabra>Skype</Palabra>
                  <Palabra>Microsoft</Palabra>
                  <Palabra>operadoras</Palabra>
            </Claves>
      </Agrupacion>
       <Agrupacion>
            <NumNoticias>2</NumNoticias>
            <Noticia>elpais.com35.xml</Noticia>
            <Noticia>elmundo.es2.xml</Noticia>
            <Claves>
                  <Palabra>Endeavour</Palabra>
                  <Palabra>Stormy</Palabra>
                  <Palabra>Mondays</Palabra>
            </Claves>
      </Agrupacion>
       <Agrupacion>
            <NumNoticias>1</NumNoticias>
            <Noticia>elpais.com2.xml</Noticia>
            <Claves>
                  <Palabra>reconstrucción</Palabra>
                  <Palabra>Zapatero</Palabra>
                  <Palabra>apoyo</Palabra>
            </Claves>
      </Agrupacion>
       <Agrupacion>
            <NumNoticias>1</NumNoticias>
            <Noticia>elpais.com3.xml</Noticia>
            <Claves>
                  <Palabra>Manuel</Palabra>
                  <Palabra>escombros</Palabra>
                  <Palabra>Juan</Palabra>
            </Claves>
      </Agrupacion>
       <Agrupacion>
            <NumNoticias>1</NumNoticias>
            <Noticia>elpais.com5.xml</Noticia>
            <Claves>
                  <Palabra>PNV</Palabra>
```

```
                    <Palabra>EA</Palabra>
                    <Palabra>Navarra</Palabra>
            </Claves>
    </Agrupacion>
     <Agrupacion>
            <NumNoticias>1</NumNoticias>
            <Noticia>elpais.com7.xml</Noticia>
            <Claves>
                    <Palabra>Errandonea</Palabra>
                    <Palabra>cárcel</Palabra>
                    <Palabra>María</Palabra>
            </Claves>
    </Agrupacion>
     <Agrupacion>
            <NumNoticias>1</NumNoticias>
            <Noticia>elpais.com9.xml</Noticia>
            <Claves>
                    <Palabra>empresas</Palabra>
                    <Palabra>suelo</Palabra>
                    <Palabra>Aguirre</Palabra>
            </Claves>
    </Agrupacion>
     <Agrupacion>
            <NumNoticias>1</NumNoticias>
            <Noticia>elpais.com13.xml</Noticia>
            <Claves>
                    <Palabra>Laden</Palabra>
                    <Palabra>Unidos</Palabra>
                    <Palabra>recursos</Palabra>
            </Claves>
    </Agrupacion>
     <Agrupacion>
            <NumNoticias>1</NumNoticias>
            <Noticia>elpais.com14.xml</Noticia>
            <Claves>
                    <Palabra>Ambos</Palabra>
                    <Palabra>Google</Palabra>
                    <Palabra>Samsung</Palabra>
            </Claves>
    </Agrupacion>
     <Agrupacion>
            <NumNoticias>1</NumNoticias>
            <Noticia>elpais.com19.xml</Noticia>
            <Claves>
                    <Palabra>Constitucional</Palabra>
                    <Palabra>participación</Palabra>
                    <Palabra>Tribunal</Palabra>
            </Claves>
    </Agrupacion>
```

```
 <Agrupacion>
        <NumNoticias>1</NumNoticias>
        <Noticia>elpais.com22.xml</Noticia>
        <Claves>
            <Palabra>partir</Palabra>
            <Palabra>Google</Palabra>
            <Palabra>versión</Palabra>
        </Claves>
</Agrupacion>
 <Agrupacion>
        <NumNoticias>1</NumNoticias>
        <Noticia>elpais.com23.xml</Noticia>
        <Claves>
            <Palabra>IP</Palabra>
            <Palabra>canal</Palabra>
            <Palabra>acciones</Palabra>
        </Claves>
</Agrupacion>
 <Agrupacion>
        <NumNoticias>1</NumNoticias>
        <Noticia>elpais.com24.xml</Noticia>
        <Claves>
            <Palabra>Ley</Palabra>
            <Palabra>enero</Palabra>
            <Palabra>resoluciones</Palabra>
        </Claves>
</Agrupacion>
 <Agrupacion>
        <NumNoticias>1</NumNoticias>
        <Noticia>elpais.com28.xml</Noticia>
        <Claves>
            <Palabra>calle</Palabra>
            <Palabra>manifestaciones</Palabra>
            <Palabra>principales</Palabra>
        </Claves>
</Agrupacion>
 <Agrupacion>
        <NumNoticias>1</NumNoticias>
        <Noticia>elpais.com36.xml</Noticia>
        <Claves>
            <Palabra>cine</Palabra>
            <Palabra>mirar</Palabra>
            <Palabra>P</Palabra>
        </Claves>
</Agrupacion>
 <Agrupacion>
        <NumNoticias>1</NumNoticias>
        <Noticia>elpais.com38.xml</Noticia>
        <Claves>
```

```
                    <Palabra>Malick</Palabra>
                    <Palabra>Pitt</Palabra>
                    <Palabra>árbol</Palabra>
            </Claves>
    </Agrupacion>
     <Agrupacion>
            <NumNoticias>1</NumNoticias>
            <Noticia>elpais.com39.xml</Noticia>
            <Claves>
                    <Palabra>denunció</Palabra>
                    <Palabra>Barça</Palabra>
                    <Palabra>resolución</Palabra>
            </Claves>
    </Agrupacion>
     <Agrupacion>
            <NumNoticias>1</NumNoticias>
            <Noticia>elpais.com43.xml</Noticia>
            <Claves>
                    <Palabra>feria</Palabra>
                    <Palabra>Barça</Palabra>
                    <Palabra>equipo</Palabra>
            </Claves>
    </Agrupacion>
     <Agrupacion>
            <NumNoticias>1</NumNoticias>
            <Noticia>elmundo.es1.xml</Noticia>
            <Claves>
                    <Palabra>expreso</Palabra>
                    <Palabra>Agirre</Palabra>
                    <Palabra>Jon</Palabra>
            </Claves>
    </Agrupacion>
     <Agrupacion>
            <NumNoticias>1</NumNoticias>
            <Noticia>elmundo.es7.xml</Noticia>
            <Claves>
                    <Palabra>trasladados</Palabra>
                    <Palabra>campamento</Palabra>
                    <Palabra>evacuados</Palabra>
            </Claves>
    </Agrupacion>
     <Agrupacion>
            <NumNoticias>1</NumNoticias>
            <Noticia>larazon.es2.xml</Noticia>
            <Claves>
                    <Palabra>impugnar</Palabra>
                    <Palabra>M</Palabra>
                    <Palabra>apenas</Palabra>
            </Claves>
```

```
        </Agrupacion>
         <Agrupacion>
                <NumNoticias>1</NumNoticias>
                <Noticia>larazon.es3.xml</Noticia>
                <Claves>
                        <Palabra>bancos</Palabra>
                        <Palabra>Aguirre</Palabra>
                        <Palabra>exigir</Palabra>
                </Claves>
        </Agrupacion>
         <Agrupacion>
                <NumNoticias>1</NumNoticias>
                <Noticia>larazon.es6.xml</Noticia>
                <Claves>
                        <Palabra>hoja</Palabra>
                        <Palabra>presenta</Palabra>
                        <Palabra>Navarra</Palabra>
                </Claves>
        </Agrupacion>
         <Agrupacion>
                <NumNoticias>1</NumNoticias>
                <Noticia>larazon.es8.xml</Noticia>
                <Claves>
                        <Palabra>economía</Palabra>
                        <Palabra>empleo</Palabra>
                        <Palabra>saliendo</Palabra>
                </Claves>
        </Agrupacion>
</Configuracion>
```

Tabla 1.3: Agrupación con umbral de similitud 0.3

Agrupación con umbral de similitud 0.5

```
<?xmlversion="1.0" encoding="ISO-8859-1"?>
<Configuracion>
        <NumCluster>50</NumCluster>
        <Agrupacion>
                <NumNoticias>5</NumNoticias>
                <Noticia>elpais.com33.xml</Noticia>
                <Noticia>elpais.com34.xml</Noticia>
                <Noticia>elpais.com41.xml</Noticia>
                <Noticia>elpais.com42.xml</Noticia>
                <Noticia>elmundo.es3.xml</Noticia>
                <Claves>
                        <Palabra>Strauss</Palabra>
                        <Palabra>FMI</Palabra>
                        <Palabra>Kahn</Palabra>
                </Claves>
        </Agrupacion>
        <Agrupacion>
                <NumNoticias>4</NumNoticias>
                <Noticia>elpais.com30.xml</Noticia>
                <Noticia>elpais.com31.xml</Noticia>
                <Noticia>elpais.com32.xml</Noticia>
                <Noticia>larazon.es7.xml</Noticia>
                <Claves>
                        <Palabra>Isabel</Palabra>
                        <Palabra>visita</Palabra>
                        <Palabra>Irlanda</Palabra>
                </Claves>
        </Agrupacion>
        <Agrupacion>
                <NumNoticias>2</NumNoticias>
                <Noticia>elpais.com26.xml</Noticia>
                <Noticia>elmundo.es6.xml</Noticia>
                <Claves>
                        <Palabra>Sol</Palabra>
                        <Palabra>protestas</Palabra>
                        <Palabra>manifestantes</Palabra>
                </Claves>
        </Agrupacion>
        <Agrupacion>
                <NumNoticias>2</NumNoticias>
                <Noticia>elpais.com18.xml</Noticia>
                <Noticia>elpais.com21.xml</Noticia>
                <Claves>
                        <Palabra>Skype</Palabra>
                        <Palabra>Microsoft</Palabra>
                        <Palabra>operadoras</Palabra>
                </Claves>
```

```
</Agrupacion>
 <Agrupacion>
        <NumNoticias>2</NumNoticias>
        <Noticia>elpais.com25.xml</Noticia>
        <Noticia>larazon.es5.xml</Noticia>
        <Claves>
            <Palabra>abusar</Palabra>
            <Palabra>Mallorca</Palabra>
            <Palabra>pederastas</Palabra>
        </Claves>
</Agrupacion>
 <Agrupacion>
        <NumNoticias>2</NumNoticias>
        <Noticia>elpais.com35.xml</Noticia>
        <Noticia>elmundo.es2.xml</Noticia>
        <Claves>
            <Palabra>Endeavour</Palabra>
            <Palabra>Stormy</Palabra>
            <Palabra>Mondays</Palabra>
        </Claves>
</Agrupacion>
 <Agrupacion>
        <NumNoticias>1</NumNoticias>
        <Noticia>elpais.com0.xml</Noticia>
        <Claves>
            <Palabra>fallecidos</Palabra>
            <Palabra>confirmado</Palabra>
            <Palabra>Dos</Palabra>
        </Claves>
</Agrupacion>
 <Agrupacion>
        <NumNoticias>1</NumNoticias>
        <Noticia>elpais.com1.xml</Noticia>
        <Claves>
            <Palabra>80%</Palabra>
            <Palabra>Zapatero</Palabra>
            <Palabra>viviendas</Palabra>
        </Claves>
</Agrupacion>
 <Agrupacion>
        <NumNoticias>1</NumNoticias>
        <Noticia>elpais.com2.xml</Noticia>
        <Claves>
            <Palabra>reconstrucción</Palabra>
            <Palabra>Zapatero</Palabra>
            <Palabra>apoyo</Palabra>
        </Claves>
</Agrupacion>
```

```
<Agrupacion>
        <NumNoticias>1</NumNoticias>
        <Noticia>elpais.com3.xml</Noticia>
        <Claves>
                <Palabra>Manuel</Palabra>
                <Palabra>escombros</Palabra>
                <Palabra>Juan</Palabra>
        </Claves>
</Agrupacion>
 <Agrupacion>
        <NumNoticias>1</NumNoticias>
        <Noticia>elpais.com4.xml</Noticia>
        <Claves>
                <Palabra>decisiones</Palabra>
                <Palabra>Zapatero</Palabra>
                <Palabra>Constitucional</Palabra>
        </Claves>
</Agrupacion>
 <Agrupacion>
        <NumNoticias>1</NumNoticias>
        <Noticia>elpais.com5.xml</Noticia>
        <Claves>
                <Palabra>PNV</Palabra>
                <Palabra>EA</Palabra>
                <Palabra>Navarra</Palabra>
        </Claves>
</Agrupacion>
 <Agrupacion>
        <NumNoticias>1</NumNoticias>
        <Noticia>elpais.com6.xml</Noticia>
        <Claves>
                <Palabra>Errandonea</Palabra>
                <Palabra>etarra</Palabra>
                <Palabra>coalición</Palabra>
        </Claves>
</Agrupacion>
 <Agrupacion>
        <NumNoticias>1</NumNoticias>
        <Noticia>elpais.com7.xml</Noticia>
        <Claves>
                <Palabra>Errandonea</Palabra>
                <Palabra>cárcel</Palabra>
                <Palabra>María</Palabra>
        </Claves>
</Agrupacion>
 <Agrupacion>
        <NumNoticias>1</NumNoticias>
        <Noticia>elpais.com8.xml</Noticia>
        <Claves>
```

```
                    <Palabra>Constitucional</Palabra>
                    <Palabra>Batasuna</Palabra>
                    <Palabra>Supremo</Palabra>
            </Claves>
    </Agrupacion>
     <Agrupacion>
            <NumNoticias>1</NumNoticias>
            <Noticia>elpais.com9.xml</Noticia>
            <Claves>
                    <Palabra>empresas</Palabra>
                    <Palabra>suelo</Palabra>
                    <Palabra>Aguirre</Palabra>
            </Claves>
    </Agrupacion>
     <Agrupacion>
            <NumNoticias>1</NumNoticias>
            <Noticia>elpais.com10.xml</Noticia>
            <Claves>
                    <Palabra>Laden</Palabra>
                    <Palabra>hijos</Palabra>
                    <Palabra>padre</Palabra>
            </Claves>
    </Agrupacion>
     <Agrupacion>
            <NumNoticias>1</NumNoticias>
            <Noticia>elpais.com11.xml</Noticia>
            <Claves>
                    <Palabra>Laden</Palabra>
                    <Palabra>hijos</Palabra>
                    <Palabra>padre</Palabra>
            </Claves>
    </Agrupacion>
     <Agrupacion>
            <NumNoticias>1</NumNoticias>
            <Noticia>elpais.com12.xml</Noticia>
            <Claves>
                    <Palabra>Laden</Palabra>
                    <Palabra>Unidos</Palabra>
                    <Palabra>preparado</Palabra>
            </Claves>
    </Agrupacion>
     <Agrupacion>
            <NumNoticias>1</NumNoticias>
            <Noticia>elpais.com13.xml</Noticia>
            <Claves>
                    <Palabra>Laden</Palabra>
                    <Palabra>Unidos</Palabra>
                    <Palabra>recursos</Palabra>
            </Claves>
```

```
</Agrupacion>
 <Agrupacion>
        <NumNoticias>1</NumNoticias>
        <Noticia>elpais.com14.xml</Noticia>
        <Claves>
            <Palabra>Ambos</Palabra>
            <Palabra>Google</Palabra>
            <Palabra>Samsung</Palabra>
        </Claves>
</Agrupacion>
 <Agrupacion>
        <NumNoticias>1</NumNoticias>
        <Noticia>elpais.com15.xml</Noticia>
        <Claves>
            <Palabra>Carreño</Palabra>
            <Palabra>sísmica</Palabra>
            <Palabra>fallas</Palabra>
        </Claves>
</Agrupacion>
 <Agrupacion>
        <NumNoticias>1</NumNoticias>
        <Noticia>elpais.com16.xml</Noticia>
        <Claves>
            <Palabra>sur</Palabra>
            <Palabra>terremotos</Palabra>
            <Palabra>500</Palabra>
        </Claves>
</Agrupacion>
 <Agrupacion>
        <NumNoticias>1</NumNoticias>
        <Noticia>elpais.com17.xml</Noticia>
        <Claves>
            <Palabra>Laden</Palabra>
            <Palabra>Qaeda</Palabra>
            <Palabra>septiembre</Palabra>
        </Claves>
</Agrupacion>
 <Agrupacion>
        <NumNoticias>1</NumNoticias>
        <Noticia>elpais.com19.xml</Noticia>
        <Claves>
            <Palabra>Constitucional</Palabra>
            <Palabra>participación</Palabra>
            <Palabra>Tribunal</Palabra>
        </Claves>
</Agrupacion>
 <Agrupacion>
        <NumNoticias>1</NumNoticias>
        <Noticia>elpais.com20.xml</Noticia>
```

```
            <Claves>
                  <Palabra>versión</Palabra>
                  <Palabra>noticia</Palabra>
                  <Palabra>Laden</Palabra>
            </Claves>
      </Agrupacion>
       <Agrupacion>
            <NumNoticias>1</NumNoticias>
            <Noticia>elpais.com22.xml</Noticia>
            <Claves>
                  <Palabra>partir</Palabra>
                  <Palabra>Google</Palabra>
                  <Palabra>versión</Palabra>
            </Claves>
      </Agrupacion>
       <Agrupacion>
            <NumNoticias>1</NumNoticias>
            <Noticia>elpais.com23.xml</Noticia>
            <Claves>
                  <Palabra>IP</Palabra>
                  <Palabra>canal</Palabra>
                  <Palabra>acciones</Palabra>
            </Claves>
      </Agrupacion>
       <Agrupacion>
            <NumNoticias>1</NumNoticias>
            <Noticia>elpais.com24.xml</Noticia>
            <Claves>
                  <Palabra>Ley</Palabra>
                  <Palabra>enero</Palabra>
                  <Palabra>resoluciones</Palabra>
            </Claves>
      </Agrupacion>
       <Agrupacion>
            <NumNoticias>1</NumNoticias>
            <Noticia>elpais.com27.xml</Noticia>
            <Claves>
                  <Palabra>Sol</Palabra>
                  <Palabra>IU</Palabra>
                  <Palabra>delegada</Palabra>
            </Claves>
      </Agrupacion>
       <Agrupacion>
            <NumNoticias>1</NumNoticias>
            <Noticia>elpais.com28.xml</Noticia>
            <Claves>
                  <Palabra>calle</Palabra>
                  <Palabra>manifestaciones</Palabra>
                  <Palabra>principales</Palabra>
```

```
            </Claves>
        </Agrupacion>
         <Agrupacion>
            <NumNoticias>1</NumNoticias>
            <Noticia>elpais.com29.xml</Noticia>
            <Claves>
                <Palabra>pide</Palabra>
                <Palabra>jóvenes</Palabra>
                <Palabra>Gómez</Palabra>
            </Claves>
        </Agrupacion>
         <Agrupacion>
            <NumNoticias>1</NumNoticias>
            <Noticia>elpais.com36.xml</Noticia>
            <Claves>
                <Palabra>cine</Palabra>
                <Palabra>mirar</Palabra>
                <Palabra>P</Palabra>
            </Claves>
        </Agrupacion>
         <Agrupacion>
            <NumNoticias>1</NumNoticias>
            <Noticia>elpais.com37.xml</Noticia>
            <Claves>
                <Palabra>Malick</Palabra>
                <Palabra>media</Palabra>
                <Palabra>Terrence</Palabra>
            </Claves>
        </Agrupacion>
         <Agrupacion>
            <NumNoticias>1</NumNoticias>
            <Noticia>elpais.com38.xml</Noticia>
            <Claves>
                <Palabra>Malick</Palabra>
                <Palabra>Pitt</Palabra>
                <Palabra>árbol</Palabra>
            </Claves>
        </Agrupacion>
         <Agrupacion>
            <NumNoticias>1</NumNoticias>
            <Noticia>elpais.com39.xml</Noticia>
            <Claves>
                <Palabra>denunció</Palabra>
                <Palabra>Barça</Palabra>
                <Palabra>resolución</Palabra>
            </Claves>
        </Agrupacion>
```

```
<Agrupacion>
        <NumNoticias>1</NumNoticias>
        <Noticia>elpais.com40.xml</Noticia>
        <Claves>
                <Palabra>peor</Palabra>
                <Palabra>jugadores</Palabra>
                <Palabra>Mourinho</Palabra>
        </Claves>
</Agrupacion>
 <Agrupacion>
        <NumNoticias>1</NumNoticias>
        <Noticia>elpais.com43.xml</Noticia>
        <Claves>
                <Palabra>feria</Palabra>
                <Palabra>Barça</Palabra>
                <Palabra>equipo</Palabra>
        </Claves>
</Agrupacion>
 <Agrupacion>
        <NumNoticias>1</NumNoticias>
        <Noticia>elmundo.es0.xml</Noticia>
        <Claves>
                <Palabra>casco</Palabra>
                <Palabra>técnicos</Palabra>
                <Palabra>urbano</Palabra>
        </Claves>
</Agrupacion>
 <Agrupacion>
        <NumNoticias>1</NumNoticias>
        <Noticia>elmundo.es1.xml</Noticia>
        <Claves>
                <Palabra>expreso</Palabra>
                <Palabra>Agirre</Palabra>
                <Palabra>Jon</Palabra>
        </Claves>
</Agrupacion>
 <Agrupacion>
        <NumNoticias>1</NumNoticias>
        <Noticia>elmundo.es4.xml</Noticia>
        <Claves>
                <Palabra>murcianos</Palabra>
                <Palabra>Gibson</Palabra>
                <Palabra>Mel</Palabra>
        </Claves>
</Agrupacion>
 <Agrupacion>
        <NumNoticias>1</NumNoticias>
        <Noticia>elmundo.es5.xml</Noticia>
        <Claves>
```

```
                    <Palabra>habla</Palabra>
                    <Palabra>película</Palabra>
                    <Palabra>cine</Palabra>
            </Claves>
    </Agrupacion>
     <Agrupacion>
            <NumNoticias>1</NumNoticias>
            <Noticia>elmundo.es7.xml</Noticia>
            <Claves>
                    <Palabra>trasladados</Palabra>
                    <Palabra>campamento</Palabra>
                    <Palabra>evacuados</Palabra>
            </Claves>
    </Agrupacion>
     <Agrupacion>
            <NumNoticias>1</NumNoticias>
            <Noticia>larazon.es0.xml</Noticia>
            <Claves>
                    <Palabra>magnitud</Palabra>
                    <Palabra>evaluar</Palabra>
                    <Palabra>grados</Palabra>
            </Claves>
    </Agrupacion>
     <Agrupacion>
            <NumNoticias>1</NumNoticias>
            <Noticia>larazon.es1.xml</Noticia>
            <Claves>
                    <Palabra>grados</Palabra>
                    <Palabra>localidad</Palabra>
                    <Palabra>Granada</Palabra>
            </Claves>
    </Agrupacion>
     <Agrupacion>
            <NumNoticias>1</NumNoticias>
            <Noticia>larazon.es2.xml</Noticia>
            <Claves>
                    <Palabra>impugnar</Palabra>
                    <Palabra>M</Palabra>
                    <Palabra>apenas</Palabra>
            </Claves>
    </Agrupacion>
     <Agrupacion>
            <NumNoticias>1</NumNoticias>
            <Noticia>larazon.es3.xml</Noticia>
            <Claves>
                    <Palabra>bancos</Palabra>
                    <Palabra>Aguirre</Palabra>
                    <Palabra>exigir</Palabra>
            </Claves>
```

```
        </Agrupacion>
         <Agrupacion>
                <NumNoticias>1</NumNoticias>
                <Noticia>larazon.es4.xml</Noticia>
                <Claves>
                        <Palabra>llegar</Palabra>
                        <Palabra>Qaeda</Palabra>
                        <Palabra>peor</Palabra>
                </Claves>
        </Agrupacion>
         <Agrupacion>
                <NumNoticias>1</NumNoticias>
                <Noticia>larazon.es6.xml</Noticia>
                <Claves>
                        <Palabra>hoja</Palabra>
                        <Palabra>presenta</Palabra>
                        <Palabra>Navarra</Palabra>
                </Claves>
        </Agrupacion>
         <Agrupacion>
                <NumNoticias>1</NumNoticias>
                <Noticia>larazon.es8.xml</Noticia>
                <Claves>
                        <Palabra>economía</Palabra>
                        <Palabra>empleo</Palabra>
                        <Palabra>saliendo</Palabra>
                </Claves>
        </Agrupacion>
</Configuracion>
```

Tabla 1.4: Agrupación con umbral de similitud 0.5

Agrupación realizada manualmente

```
<?xmlversion="1.0" encoding="ISO-8859-1" ?>
<Configuracion>
      <NumCluster>12</NumCluster>
      <Agrupacion>
      <NumNoticias>8</NumNoticias>
            <Noticia>elpais.com0.xml</Noticia>
            <Noticia>elpais.com1.xml</Noticia>
            <Noticia>elpais.com2.xml</Noticia>
            <Noticia>elpais.com3.xml</Noticia>
            <Noticia>elpais.com15.xml</Noticia>
            <Noticia>elpais.com16.xml</Noticia>
            <Noticia>larazon.es0.xml</Noticia>
            <Noticia>larazon.es1.xml</Noticia>
      </Agrupacion>
      <Agrupacion>
            <NumNoticias>7</NumNoticias>
            <Noticia>elpais.com14.xml</Noticia>
            <Noticia>elpais.com18.xml</Noticia>
            <Noticia>elpais.com21.xml</Noticia>
            <Noticia>elpais.com22.xml</Noticia>
            <Noticia>elpais.com23.xml</Noticia>
            <Noticia>elpais.com25.xml</Noticia>
            <Noticia>larazon.es5.xml</Noticia>
      </Agrupacion>
      <Agrupacion>
            <NumNoticias>7</NumNoticias>
            <Noticia>elpais.com10.xml</Noticia>
            <Noticia>elpais.com11.xml</Noticia>
            <Noticia>elpais.com12.xml</Noticia>
            <Noticia>elpais.com13.xml</Noticia>
            <Noticia>elpais.com17.xml</Noticia>
            <Noticia>elpais.com20.xml</Noticia>
            <Noticia>larazon.es4.xml</Noticia>
      </Agrupacion>
      <Agrupacion>
            <NumNoticias>10</NumNoticias>
            <Noticia>elpais.com4.xml</Noticia>
            <Noticia>elpais.com5.xml</Noticia>
            <Noticia>elpais.com6.xml</Noticia>
            <Noticia>elpais.com7.xml</Noticia>
            <Noticia>elpais.com8.xml</Noticia>
            <Noticia>elpais.com19.xml</Noticia>
            <Noticia>elmundo.es0.xml</Noticia>
            <Noticia>elmundo.es1.xml</Noticia>
            <Noticia>larazon.es2.xml</Noticia>
            <Noticia>larazon.es6.xml</Noticia>
      </Agrupacion>
      <Agrupacion>
            <NumNoticias>3</NumNoticias>
            <Noticia>elpais.com9.xml</Noticia>
            <Noticia>elpais.com24.xml</Noticia>
            <Noticia>larazon.es3.xml</Noticia>
```

```
      </Agrupacion>
      <Agrupacion>
            <NumNoticias>4</NumNoticias>
            <Noticia>elpais.com30.xml</Noticia>
            <Noticia>elpais.com31.xml</Noticia>
            <Noticia>elpais.com32.xml</Noticia>
            <Noticia>larazon.es7.xml</Noticia>
      </Agrupacion>
      <Agrupacion>
            <NumNoticias>4</NumNoticias>
            <Noticia>elpais.com27.xml</Noticia>
            <Noticia>elpais.com39.xml</Noticia>
            <Noticia>elpais.com40.xml</Noticia>
            <Noticia>elpais.com43.xml</Noticia>
      </Agrupacion>
      <Agrupacion>
            <NumNoticias>4</NumNoticias>
            <Noticia>elpais.com33.xml</Noticia>
            <Noticia>elpais.com42.xml</Noticia>
            <Noticia>elmundo.es3.xml</Noticia>
            <Noticia>larazon.es8.xml</Noticia>
      </Agrupacion>
      <Agrupacion>
            <NumNoticias>5</NumNoticias>
            <Noticia>elpais.com26.xml</Noticia>
            <Noticia>elpais.com28.xml</Noticia>
            <Noticia>elpais.com29.xml</Noticia>
            <Noticia>elmundo.es6.xml</Noticia>
            <Noticia>elmundo.es7.xml</Noticia>
      </Agrupacion>
      <Agrupacion>
            <NumNoticias>5</NumNoticias>
            <Noticia>elpais.com36.xml</Noticia>
            <Noticia>elpais.com37.xml</Noticia>
            <Noticia>elpais.com38.xml</Noticia>
            <Noticia>elmundo.es4.xml</Noticia>
            <Noticia>elmundo.es5.xml</Noticia>
      </Agrupacion>
      <Agrupacion>
            <NumNoticias>2</NumNoticias>
            <Noticia>elpais.com35.xml</Noticia>
            <Noticia>elmundo.es2.xml</Noticia>
      </Agrupacion>
      <Agrupacion>
            <NumNoticias>2</NumNoticias>
            <Noticia>elpais.com34.xml</Noticia>
            <Noticia>elpais.com41.xml</Noticia>
      </Agrupacion>
</Configuracion>
```

Tabla 1.5: Agrupación manual

Anexo 2: Casos de prueba para la evaluación de resúmenes

Noticia del periódico elpais.com
En directo: Los niveles de radiación suben en Japón, tras una nueva explosión y un incendio. Se han registrado dos nuevas explosiones, en los reactores 2 y 3, y un incendio en el 4.- La situación es de máxima alerta.- En Tokio, a 240 kilómetros de la central nuclear de Fukushima, se han detectado partículas radiactivas y se teme que el fuerte viento lleve una nube tóxica al interior del país Japón va de mal en peor. Del caos y la desolación provocados por el terremoto de magnitud 9 que sacudió el país el viernes 11 y provocó un devastador tsunami, el país ha pasado a estar al borde de una catástrofe nuclear. La alarma crece en torno a la central de Fukushima, afectada en cuatro de sus seis reactores. Un incendio se ha desencadenado en el reactor 4, aunque ya está controlado, y se han producido explosiones en los reactores 2 y 3. El Gobierno japonés admite que "puede haberse producido una fuga de materiales radiactivos", especialmente por causa del incendio, "que pueden afectar a la salud humana", y cuyas partículas tóxicas ya han llegado hasta Tokio, a 240 kilómetros de la central. Hay 50 operarios tratando de controlar la situación, los únicos que no han sido evacuados.

Tabla 2.1: Noticia del país.com

Noticia del periódico elpais.com

Japón admite fugas radiactivas "que pueden afectar a la salud" tras un incendio y una nueva explosión en Fukushima

Evacuaciones 30 kilómetros alrededor de la central nuclear. -La alerta llega a Tokio porque el viento podría arrastrar las partículas. -El Gobierno reconoce que podría haber grietas en la vasija del reactor 2.- La agencia nuclear japonesa pide ayuda a la ONU y EE UU

La alarma de un desastre nuclear crece en torno a la central de Fukushima, afectada en cuatro de sus seis reactores por el terremoto que devastó el país el viernes.

Un incendio se ha desencadenado en el reactor 4, aunque ya está controlado, mientras que en el número 2 se produjo una explosión en torno a las seis de la mañana locales (22.00 en España).

Aunque el agua utilizada para rebajar la temperatura en el interior está sufriendo el efecto contrario y podría estar empezando a hervir, según informa la agencia de noticias Kyodo, Japón acaba de informar a la ONU de que los niveles de radiactividad en la puerta de la central está descendiendo, aunque siguen estando muy por encima de lo normal.

Además, la Agencia de Seguridad Nuclear japonesa (NISA, por sus siglas en inglés) ha confirmado que el incendio ha dejado dos agujeros de ocho metros en el muro del reactor 4.

Mientras tanto, la zona continúa viviendo réplicas del temblor.

Fukushima ha vivido hoy una réplica de magnitud 6,3.

Desde el viernes, se han producido más de 200 réplicas del grave terremoto.

El primer ministro japonés, Naoto Kan, ha reprendido a varios ejecutivos de la compañía eléctrica que gestiona la central -Tokyo Electric Power Co.

(Tepco)- inglés) por haber tardado demasiado tiempo en informarle de que se había producido la última explosión en un reactor de Fukushima 1.

"¿Qué demonios está pasando?", les preguntó Kan, según Kyodo.

"La televisión ha informado de una explosión, pero durante una hora no se ha dicho nada a la oficina del primer ministro", añadió.

Kan ha ordenado, además, que Tepco no saque a sus empleados de la central.

Hay 50 operarios trabajando en la central, los únicos que no han sido evacuados.

Todo está en sus manos: tienen que refrigerar las piscinas de los reactores.

En las centrales había 800 operarios trabajando, pero Tepco está desbordada y ha pedido que se retiren todos menos este medio centenar.

Por otro lado, el gobernador de Fukushima, Yuhei Sato, ha llamado por teléfono a Kan para decirle que "el miedo y el enfado de los habitantes de la prefectura está llegando a un límite", informa EP.

El Gobierno japonés admitió a primera hora que "puede haberse producido una fuga de materiales radiactivos", especialmente por causa del incendio, "que pueden afectar a la salud humana".

Sin embargo, la Organización Mundial de la Salud ha querido enviar un mensaje tranquilizador: "Japón está tomando las medidas de salud públicas adecuadas para proteger a la población de la radiación", ha dicho Gregory Hartl, portavoz citado por la agencia Reuters.

Además, añaden que dicho organismo no ha recibido ninguna petición de ayuda por parte de este país, aunque sus expertos en radiación están alerta.

A las tres de la mañana (hora española) el primer ministro japonés, Naoto Kan, ha comparecido para hacer un llamamiento a la población y anunciar nuevas evacuaciones, la

de los residentes en torno a 10 kilómetros de la central Fukushima I ya está completada; los que viven entre 10 y 20 serán rescatados en breve; y los que residen entre 20 y 30 kilómetros de la central no deben salir a la calle.

El portavoz del Ejecutivo que sucedió a Naoto Kan en la tribuna de oradores pasó del mensaje de calma de los días previos a inequívocas señales de alarma.

"Cuanto más lejos estén de la central, más seguros estarán", advirtió YukioEdano, que apareció ante las cámaras con muestras evidentes de sudor en la frente.

La radiación en los alrededores de la central ha llegado a sobrepasar diez mil veces los límites legales.

La situación ha generado una gran preocupación en el país; los locutores de televisión repiten mensajes que parecen salidos de una película de serie B sobre una catástrofe nuclear: "Cierren las ventanas, no utilicen sistemas de ventilación y tiendan la ropa en casa".

Unas 200.000 dosis de yodo (que ayudan a proteger la glándula tiroides de los efectos de la radiación) se han repartido ya entre la población.

Mientras, la Embajada francesa en Japón ha recomendado a sus nacionales que vivan en Tokio que no salgan al exterior, porque el viento que sopla hacia la capital podría arrastrar hasta allí las partículas radiactivas.

Hacia las cinco de la mañana (hora española) ya se habían detectado pequeñas cantidades de radiación en Tokio, alertó Kyodo.

Incendio y una tercera explosión.

Unos minutos después de la tercera explosión los niveles de radiación en los alrededores de Fukushima subieron hasta 8.217 microsieverts por hora, frente a los 1.941 que se registraban 40 minutos antes, según mediciones de la Agencia de Seguridad Nuclear japonesa.

Estos 8.000 microsieverts por hora serían el triple de la cantidad de radiación a la que está sometida una persona en un año.

En un clima de confusión, se pensó primero que era por culpa de la explosión del reactor 2, pero posteriormente se aclaró que era consecuencia de un incendio en el reactor 4 en el que estaban ardiendo sustancias radiactivas.

La buena noticia, si las hay, es que el reactor 4 estaba inoperativo en el momento del terremoto y no contenía barras de combustible.

YukioEdano ha reconocido que "hay una alta probabilidad" de que la vasija de contención del reactor 2 se haya agrietado, pero el Ejecutivo insiste en que el edificio de contención -el último muro ante una fuga, y de cuya resistencia depende que Fukushima no sea Chernóbil- no ha quedado dañado, descartando la posibilidad de una fuga de radiactividad de grandes dimensiones.

El Gobierno admite también que puede haber daños en la cámara de despresurización, el sistema circular de refrigeración dentro del edificio de la contención.

Un portavoz de Tepco explicó en una confusa rueda de prensa retransmitida y doblada al inglés por Al Yazira: "Hay una posibilidad de que haya daño", pero inmediatamente añadió que eso no tenía por qué significar una fuga o que podía tratarse simplemente de una válvula que estuviera midiendo mal la presión.

Aún con la confusión reinante, la situación parece cualitativamente distinta -más grave- que la de los días previos.

La eléctrica responsable de la planta, Tepco ha constatado es que el nivel del agua ha bajado sensiblemente dentro del reactor 2, lo que denotaría daños en la piscina de condensación destinada a enfriar el reactor y controlar las condiciones en el interior del recinto.

Al menos 2,7 metros de las varillas de combustible (de los cuatro que miden) no están cubiertas por el agua, y Tepco no puede confirmar si el nivel del agua está subiendo aunque haya vuelto a inyectar agua de mar.

Esto implica que la mitad del uranio está sin refrigeración, el paso previo a la fusión del núcleo de la central.

Japón pide ayuda.

La central de Fukushima parece un boxeador sonado: encajaba y encajaba golpes mientras la grada -en este caso el planeta entero- contemplaba con angustia deseando que no cayera a la lona ni tirara la toalla.

Si una explosión en una nuclear es una imagen insólita, Fukushima suma tres en tres días, por eso desde horas antes de la última deflagración la crisis ya había desbordado a Japón, y Tokio había pedido ayuda a la agencia nuclear de EE UU (NRC, en sus siglas en inglés) y a la Agencia Internacional de la Energía Atómica (OIEA), con la que debatía los detalles de cómo sería esa misión técnica.

Tras la explosión junto al reactor 1, ocurrida el pasado sábado, a las 11.01 de ayer (las 3.01 del lunes hora peninsular española) estalló el hidrógeno junto al reactor número 3 y se llevó de nuevo parte del edificio.

Las autoridades insistieron en que en los dos casos había aguantado la contención.

Tepco insistió en que la explosión se debió a la salida de hidrógeno, un gas que, en contacto con el oxígeno del aire, produce una deflagración.

Así que la explosión no fue nuclear pero sí reveló que las autoridades estaban dejando salir gases del interior de la planta -con la consiguiente radiactividad- para reducir la excesiva presión.

Los trabajos se centraban en conseguir refrigerar esos dos reactores hasta que el problema saltó en el reactor número 2.

Ese reactor puede acabar siendo el más problemático.

Su explosión, ocurrida cuando Japón amanecía al martes, ha sido diferente.

Cada día que pasa el riesgo de que cedan los edificios de la contención de los dos primeros reactores afectados es menor, según coinciden todos los expertos.

Aunque sea de forma precaria y a la desesperada con agua de mar, Japón estaba consiguiendo enfriar los núcleos de esos dos reactores.

María Teresa Domínguez, presidenta del Foro Nuclear, el lobby de las seis nucleares españolas, afirmó que el problema en el primer reactor estaba casi solucionado.

"Cuando paró la central, en el núcleo quedaba un 7% del calor residual del núcleo.

Ya solo queda el 0,05%", afirmó en una concurrida rueda de prensa.

El uso de agua de mar, que dejará inservibles los reactores, demuestra lo desesperado de la situación.

Domínguez defendió que Fukushima estaba resistiendo a la combinación terremoto-tsunami y defendió que esa era la prueba de la fiabilidad atómica.

Ese es el argumento que usa insistentemente el lobby nuclear.

Los detractores de esta energía, en cambo, ven en el accidente la prueba de que la seguridad total no existe y de que el excesivo riesgo no compensa su uso.

Tabla 2.2: Noticia del país.com

Noticia del periódico elpais.com
Alarma nuclear Las explosiones en Fukushima avivan la polémica sobre el desarrollo de este tipo de energía La terrible catástrofe sufrida por Japón en estos últimos días ha afectado a todos los sectores productivos y a la vida de cientos de miles de personas. Pero lo que quizá ha despertado mayor inquietud ha sido el daño sufrido por algunas plantas nucleares situadas en la zona más castigada por los terremotos, empezando por el mayor, de magnitud 8,9, y el tsunami posterior. La inquietud se debe, más que a los efectos nocivos sobre la población o el medio ambiente, menores hasta este momento, sobre todo a la potencialidad de graves emisiones de radiactividad al medio ambiente y al efecto que puede tener sobre el debate mundial acerca del papel de la energía nuclear en el futuro. Ante los problemas de seguridad de suministro, volatilidad de precios y emisiones de gases de efecto invernadero, se discute sobre la necesidad de impulsar un profundo cambio en nuestro paradigma energético para las próximas décadas, tanto desde el lado de la demanda, con medidas de ahorro y eficiencia energética, como desde el de la oferta, con fuentes de energía libres de carbono. La energía nuclear es uno de los candidatos a complementar el creciente papel que deben jugar las renovables en nuestro futuro esquema de suministro energético. Los sucesos de Japón ya han afectado al debate y han suscitado reservas sobre el uso de esta energía y, dependiendo de lo que ocurra con los reactores dañados del complejo de Fukushima, podrían suponer un nuevo parón de décadas, tal como ocurrió tras los accidentes de ThreeMile Island, en 1979, y Chernóbil, en 1986, o incluso un abandono definitivo de la alternativa nuclear. El primer impacto político se ha producido en Alemania: la canciller Merkel ha decidido suspender la prolongación del funcionamiento de sus 17 centrales nucleares en tanto se revisan los estándares de seguridad de las plantas. Lo que les ha ocurrido a los reactores de la central de Fukushima es probablemente lo peor que podía imaginarse, con un terremoto de inusitado poder destructivo y un tsunami que, además de agravar los daños, ha dificultado el acceso a las instalaciones y el transporte del equipamiento necesario para paliar los daños. En general, los reactores han respondido con seguridad excepto dos, o quizá tres, en los que está siendo difícil extraer el calor residual generado dentro del núcleo debido a las desintegraciones del material radiactivo en su interior. Si dicho material escapa de los sistemas de contención y se difunde por el exterior, es muy probable que se produzca una reacción contraria a cualquier desarrollo de nuevas plantas, por más seguras y perfeccionadas que sean. Si, por el contrario, el inventario de materiales radiactivos se mantiene confinado dentro de los recintos de las centrales, los daños a la salud de las personas serán reducidos, y el debate adoptará formas distintas aunque, en todo caso, supondrán una clara inflexión en la actual tendencia a considerar la energía nuclear como una tecnología valiosa para el futuro.

Tabla 2.3: Noticia del país.com

Noticia del periódico elpais.com

El último balance estima en 4.000 los muertos en Japón

La cifra, no oficial, incluye los 2.000 cadáveres hallados hoy en Miyagi.- Decenas de equipos de rescate internacionales tratan de socorrer a los supervivientes y buscar fallecidos

Japón continúa con el recuento de víctimas mortales del terremoto que el pasado viernes causó un devastador tsunami.

La cifra no para de crecer: el hallazgo hoy de 2.000 cadáveres en la prefectura de Miyagi, al noreste del país, duplica el balance anterior de 1.897 muertos -la cifra oficial que por ahora reconoce el Gobierno-. Las autoridades temen que lleguen a 10.000.

Además, 1.419 personas han resultado heridas, más de 10.000 están desaparecidas y más de 400.000 han sido evacuadas, la mayoría cerca de las centrales nucleares dañadas por el temblor, según la agencia local Kyodo.

Alrededor de un millar de cuerpos fueron hallados en una playa de la península de Ojika, mientras que otros tantos fueron encontrados en la ciudad de MinamiSanriku, donde al menos 9.500 personas -más de la mitad de la población- están en paradero desconocido.

Sin embargo, algunos medios creen que es posible que muchos de estos desaparecidos huyeran a tiempo a la vecina localidad de Tome, también en Miyagi.

Tampoco se conoce el paradero de otros 8.000 residentes del pueblo costero de Otsuchi, en la provincia de Iwate.

Entretanto, los equipos de emergencia se afanan por rescatar cerca de 300 cadáveres atrapados entre los escombros en la ciudad de Sendai, capital de dicha prefectura, que tampoco han sido incluidos en el recuento oficial de víctimas.

En muchos núcleos urbanos continúan apareciendo cuerpos sin vida en las playas y la labor de los equipos de rescate se ve dificultada por las constantes réplicas y la magnitud de la devastación causada por el terremoto, el mayor que ha sufrido Japón desde que comenzó a registrar datos hace 140 años.

El Gobierno de Miyagi ha solicitado ayuda a otras prefecturas para comenzar con la quema de los cuerpos con el fin de evitar la propagación de enfermedades entre los supervivientes, muchos de los cuales afrontan hoy su cuarta noche sin agua, comida o electricidad. Ayuda internacional.

Unos 100.000 militares al mando del operativo de salvamento siguen peinando la zona en busca de víctimas atrapadas bajo los escombros o arrastradas mar adentro por la ola gigante de 10 metros de altura.

Japón cuenta con la colaboración de EE UU, que distribuye alimentos y material de socorro a través de un portaaviones y helicópteros, y ha recibido ofertas de ayuda de cerca de 70 países, que han aportado bomberos, médicos o especialistas en el manejo de grúas para retirar los restos de edificios y llegar a los atrapados.

Además de países como Australia, India, Corea del Sur, España, México o Francia, están colaborando otros que tradicionalmente mantienen unas tensas relaciones con Japón, como China o Rusia, que mantiene un conflicto territorial con Tokio por la soberanía de las islas Kuriles.

China enviará 30 millones de yuanes (3,2 millones de euros) en ayuda humanitaria a la zona afectada.

Pekín ya envió ayer un grupo de 15 miembros del Equipo Chino de Búsqueda y Rescate Internacional para ayudar a localizar supervivientes.

"China es también un país propenso a sufrir terremotos, y empatizamos totalmente con los sentimientos del pueblo japonés en estos momentos", ha declarado hoy el primer ministro chino, WenJiabao, que ha transmitido sus "profundas condolencias por la pérdida

de vidas" y ha expresado la "sincera simpatía con el pueblo japonés".
Wen, en una conferencia de prensa anual celebrada en Pekín, ha recordado que tras el terremoto de Sichuan de 2008, que mató a más de 80.000 personas, "el Gobierno japonés envió un equipo de rescate a China y ofreció suministros".

Y ha asegurado que Pekín "continuará proveyendo de más ayuda a Japón de acuerdo con sus necesidades".

Por su parte, la Agencia de Turismo de Japón ha declinado informar sobre los 2.500 extranjeros que se encontraban visitando la zona afectada por el seísmo, según recoge la agencia de noticias Kyodo.

Varias embajadas han recomendado a sus ciudadanos no viajar al país, donde una nueva réplica de 6,3 grados en la escala Richter ha hecho temblar de nuevo la zona nororiental a las 15.13 hora local (ocho horas menos en la España peninsular).

Desde el viernes, se han registrado casi 300 réplicas del devastador seísmo, y la Agencia Meteorológica nipona indicó anoche que hay un 70% de posibilidades de que se produzcan réplicas de hasta 7 grados en la escala Richter hasta el miércoles, por lo que las autoridades siguen pidiendo precaución a las poblaciones de la costa ante la posibilidad de que se vuelvan a repetir los tsunamis.

Situación caótica.

Ante la gravedad de la situación, hoy se suspenderán los trabajos en instituciones como el Parlamento de Japón, algo inusual en una de las naciones más avanzadas del mundo.

Tampoco abrirán sus puertas las plantas de los gigantes de la industria automovilística nipona Honda, Nissan, Mitsubishi, Suzuki o Toyota, líder mundial del motor, ante la dificultad de continuar operando sin recibir las piezas que necesitan y la petición del Ejecutivo de que conserven energía para evitar más cortes de suministro en los próximos días.

Así, Toyota ha anunciado la suspensión de toda su producción en Japón al menos hasta el miércoles, lo que se traducirá en 40.000 vehículos menos.

Honda también detendrá la fabricación al menos hasta el día 20, y ya ha cerrado todas sus plantas excepto una de motocicletas en la isla de Kyushu (sur), que parará mañana.

También los ciudadanos, sobre todo los de grandes urbes como Tokio, se intentan adaptar ante el anuncio de cortes eléctricos programados para afrontar las crisis de las centrales nucleares.

El sector tecnológico también se ha visto afectado: compañías como Sony, Canon y Nikon ya han anunciado que sufrirán retrasos en las entregas de fábrica de sus productos.

La devastación y el caos también han afectado a competiciones deportivas, como los Mundiales de patinaje artístico que se iban a disputar en Tokio del 21 al 27 de marzo, y que han quedado suspendidos, según la Unión Internacional de Patinaje (ISU).

La cancelación definitiva del campeonato o el aplazamiento está pendiente de una posterior evaluación, según la federación internacional.

También se han pospuesto todos los partidos de la liga japonesa de fútbol, y la selección nacional podría retirarse de la aparición como equipo invitado en la Copa América, que se celebrará en julio en Argentina.

El archipiélago de Japón está asentado en el llamado Anillo de Fuego del Pacífico, una zona de gran actividad volcánica y telúrica, y Tokio se encuentra en uno de los lugares más peligrosos, donde tres placas continentales se están frotando unas con otras, lo que genera una enorme presión sísmica.

El Gobierno ha advertido desde hace tiempo de la posibilidad de que se produzca un terremoto de magnitud 8 antes de 30 años en la zona urbana de la capital.

Tabla 2.4: Noticia del país.com

Noticia del periódico larazon.es

La alarma radiactiva se dispara tras una nueva explosión en el reactor dos de Fukushima.

El primer ministro japonés, Naoto Kan, ha admitido que podría producirse una fuga radiactiva de la central nuclear de Fukushima-1 después de la explosión registrada esta madrugada en el reactor número dos, según informa la agencia de noticias Kiodo.

El Gobierno de Japón ha establecido una zona de exclusión aérea de un radio de 30 kilómetros sobre la central nuclear de Fukushima-1, ubicada en el noreste del país, según informa la agencia de noticias Kiodo.

En una breve comparecencia televisada, el mandatario ha subrayado la necesidad de evacuar a las personas que viven a menos de 20 kilómetros de la planta, al tiempo que les ha instado a permanecer en el interior de sus casas hasta que se complete el desalojo.

A pesar de lo complicado de la situación, el 'premier' ha solicitado a la población de esta prefectura que mantenga la calma.

Los niveles de radiación en torno a la central se sitúan en torno a los 8.217 microsievert por hora, ocho veces más que la cantidad a la que se encuentra expuesta una persona en un año.

En el caso del reactor número tres, la contaminación supera 400 veces el límite anual.

Confirmación de la OIEA

El Organismo Internacional de Energía Atómica (OIEA) ha confirmado que hubo una explosión en el reactor 2 de la planta nuclear Fukushima Diichi y que hay emanación de radiactividad a la atmósfera debido a un incendio en un depósito de combustible en el reactor 4.

En un comunicado, el OIEA precisa que ha obtenido la información de las autoridades japonesas, y que la explosión en el reactor 2 se produjo en torno a las 06:20 hora local de Japón.

Además, hay fuego en el depósito de almacenamiento de combustible usado del reactor 4, en la misma planta atómica, seriamente dañada por el terremoto y posterior tsunami del viernes, y está saliendo radiactividad directamente a la atmósfera.

En el lugar de los hechos se registraron hasta 400 millisievert por hora.

"Las autoridades japonesas están diciendo que hay una posibilidad de que el fuego haya sido causado por una explosión de hidrógeno", añade la nota.

Defecto en el contenedor del reactor 2

Los técnicos han hallado este martes un defecto en el contenedor del reactor número dos de la central nuclear de Fukushima-1, según ha confesado el jefe del Gabinete japonés, YukioEdano.

Concretamente, el defecto ha sido encontrado en la sala utilizada para pasar el vapor a líquido.

Mientras tanto, continúa sin ser extinguido un incendio en el cuarto reactor de la planta, y no se está logrando enfriar correctamente los demás.

En este sentido, el Gobierno de Japón ha negado que haya una fuga continua y elevada de radiación en torno a este reactor.

El ministro portavoz, YukioEdano, aseguró que los niveles de radiación en torno al reactor han disminuido, después de que tras el incendio superaran cien veces el límite legal permitido.

El fuego, en el cuarto piso del edificio, hizo que algunos objetos cayeran a la estructura

del reactor, que tiene barras de combustible ya utilizadas y no se encontraba en funcionamiento desde antes del seísmo del viernes.

Tampoco se encontraban activos los reactores 5 y 6 de esa misma central, en los que, según Edano, también se han detectado posibles problemas con su sistema de refrigeración.

El reactor número dos registró el martes a primera hora una nueva explosión y al parecer se podría haber producido una fusión parcial en el núcleo.

Además se teme una fuga de radiación elevada.

Se teme que la deflagración pueda haber provocado la fuga de una cantidad indeterminada de material radiactivo, informó la Agencia de Seguridad Nuclear.

La agencia Kyodo señaló que los niveles de radiación "superaron el límite legal" tras la explosión hasta llegar durante un instante a los 8.127 microsievert, ocho veces por encima del tope recomendado para la salud.

En la provincia de Ibaraki, al sur de Fukushima, también se detectó un aumento de la ionización del aire.

De acuerdo a la misma fuente, el receptáculo de seguridad que protege al núcleo ha resultado dañado por la caída de la presión en su interior a raíz de la combustión de hidrógeno.

El estallido ocurrió a primera hora de la mañana, las 6.10 hora local (21.10 GMT del lunes), poco después de que el Gobierno admitiera que el reactor continuaba inestable y, según la agencia Kyodo, ha comenzado la evacuación de los empleados de la central.

Los operarios de la planta estuvieron toda la noche trabajando para inyectar agua salada en el contenedor secundario del reactor en un intento de enfriar el núcleo y evitar una fusión radiactiva.

Los trabajadores de esta central de 40 años pretendían mantener intactos los recipientes primarios de contención de los reactores (las "capas" que los protegen) y evitar una peligrosa fuga de radiactividad en la zona, en la que se han evacuado a más de 200.000 personas.

Si el núcleo comenzara a fundirse, provocaría una situación de emergencia por emisión de radiaciones.

El reactor número 2 de Fukushima sufrió el lunes un fallo en una de sus diez válvulas que afectó al sistema de refrigeración, algo similar a lo ocurrido antes de que explotaran los reactores 1 y 3 de la misma central después del seísmo de 9 grados de magnitud en la escala Richter del viernes TEPCO no descarta una fusión parcial del núcleo del reactor número dos de Fukushima-1

La Compañía Eléctrica de Tokio (TEPCO) ha advertido este martes de que es probable que se haya producido una fusión parcial del núcleo del reactor número dos de la central nuclear de Fukushima-1, tras la explosión sufrida hace escasas horas, según informa la agencia de noticias Kiodo.

Asimismo, la entidad ha confirmado que la detonación en dicho reactor ha provocado daños en su contenedor, por lo que ahora se teme que haya una fuga de radiación elevada.

Los niveles se sitúan ahora en los 8.217 microsievert por hora, ocho veces más que la cantidad anual a la que se encuentra expuesta una persona.

Esta es la tercera explosión que se produce en la central desde el terremoto de 9 grados en la escala de Richter que el pasado viernes azotó el litoral noreste de Japón.

El sábado una afectó al reactor número uno y el lunes otra al número dos.

Tabla 2.5: Noticia de la razon.es

Noticia del periódico larazon.es

Japón, en vilo ante una emergencia nuclear que no hace sino empeorar.

El pueblo japonés, el más ordenado y disciplinado del mundo, parece incapaz de sobreponerse a una catástrofe que empieza a parecer un retorcido guión cinematográfico.

Un rosario de desgracias interconectadas que ha dejado ya miles de muertos, cuyos cuerpos la marea empezaba ayer a devolver a la costa.

«Es demasiado.

Ya basta», se derrumbaba ayer MichikoIkezawa, una voluntaria que repartía víveres entre los desplazados.

En las costas del este, los equipos de rescate y los militares desplegados siguen encontrando muertos entre los escombros y flotando entre los charcos.

Pero el horror que ha dejado a su paso el tsunami ha quedado en un segundo plano frente a la otra amenaza, que tiene un peligro potencial si cabe mayor.

Los ingenieros y técnicos desplegados no consiguen normalizar la situación en la central nuclear de Fukushima, donde ya han reventado las cúpulas de dos reactores, mientras que se dañó el muro de contención de un tercero.

La tercera explosión, que afectó al reactor número 2, fue la más grave.

En esta ocasión, por primera vez las autoridades reconocieron que podrían existir daños en la vasija de contención del reactor, compuesta de acero y hormigón y de forma de bombilla.

En las otras dos explosiones en Fukushima la vasija quedó intacta.

Según la Agencia de Seguridad Nuclear, la deflagración ha provocado una fuga de una cantidad indeterminada de material radiactivo.

La agencia Kyodo informó de que los niveles de radiación «superaron el límite legal» tras la explosión.

El portavoz del Gobierno, YukioEdano, confirmó que se han producido «posibles daños en la vasija», que se halla en la parte inferior de la caja de contención que sirve para refrigerar el reactor y controlar la presión en el interior.

Mientras tanto, un motivo para el optimismo: las otras dos centrales que presentaron fallos, Onagawa y Tokai, parecen haberse estabilizado.

En la ciudad de Kashima, a escasos kilómetros de Tokai, se recibió con alivio la noticia.

Los técnicos están intentado enfriar los reactores con agua marina, una medida que según los expertos debería ser la última de las soluciones, porque inutiliza para siempre el reactor.

Según los últimos informes, se estaría consiguiendo elevar el nivel de refrigerante en la piscina que contiene las barras de combustible nuclear.

Es decir: la medida parece estar funcionando, aunque ayer empezaron a surgir problemas logísticos para seguir bombeando agua marina hasta la central.

Las malas noticias se suceden cada hora y la población japonesa empieza a tener la sensación de que el Gobierno y los responsables de enfrentar la emergencia están improvisando soluciones a medida que aparecen nuevas fases de la crisis.

También en la oposición y en la Prensa siguen aumentando las voces críticas, que acusan al Ejecutivo de no tomar medidas de seguridad drásticas desde el primer momento.

Igualmente se les reprocha que no hayan ofrecido una información transparente y actualizada a la población.

Son más bien notas al pie de página: todos tiene claro, en todo caso, que la situación es lo suficientemente grave como para dejar a un lado temporalmente debates superfluos y centrarse en sacar adelante la «peor crisis que ha vivido Japón desde la II Guerra Mundial», en palabras del primer ministro, Naoto Kan.

Mientras tanto, Japón pidió ayer al Organismo Internacional de Energía Atómica (OIEA) el envío de un equipo de expertos para que aporten ideas que sirvan para contener el sobrecalentamiento de los reactores afectados.

La OIEA, dependiente de la ONU, rebajó la amenaza y sostuvo que el accidente no Chernóbil.

Pero quizá la opinión más repetida es la que apunta que las posibilidades de que se desate una tragedia es reducida, aunque no inexistente.

También es mayoritaria la idea de que la emergencia entra entre hoy y mañana en un punto de inflexión: o los reactores se fusionan, o tenderán a ir enfriándose, lo que acabaría con el peligro inminente.

Millones de personas cruzan los dedos para que el desenlace traiga la segunda opción.

Tabla 2.6: Noticia de la razon.es

Noticia del periódico larazon.es
Los vientos están dispersando hacia el océano la amenaza nuclear. La actual situación meteorológica, con vientos hacia el este, está alejando de Japón la amenaza de una nube radiactiva como consecuencia del accidente en la planta nuclear de Fukushima, indicó hoy la Organización Mundial de la Meteorología (OMM). "En el momento actual no hay implicaciones en tierra para Japón u otros países. Por ahora todas las implicaciones meteorológicas son hacia el mar abierto. El viento está dispersando las partículas radiactivas hacia el océano", señaló una portavoz del organismo en conferencia de prensa. No obstante, indicó que las condiciones de los vientos son muy cambiantes y que "la situación puede fluctuar en los próximos días". Por su parte, la Organización Mundial de la Salud (OMS) consideró hoy que no existe riesgo para la salud como consecuencia de la crisis nuclear en Japón ya que el gobierno nipón ha tomado las medidas correctas de evacuación. "Si estás expuesto a la radiactividad habría riesgo, pero con las medidas de evacuación que ha adoptado el gobierno japonés, la población no está expuesta", aseguró la doctora María Neira, responsable de Salud Pública y Medioambiente de la OMS. Neira señaló que las recomendaciones sanitarias para estos casos hablan de evacuar en un radio de cinco kilómetros, por lo que la decisión del gobierno japonés de ampliar la zona de evacuación hasta los veinte kilómetros en torno a la planta nuclear de Fukushima "es una medida de precaución adicional". En un documento elaborado acerca de la actual crisis nuclear en Japón tras el devastador terremoto del 11 de marzo, la OMS afirma que "en caso de accidente en una planta nuclear son los "equipos de rescate, los que primeros responden a la emergencia y los trabajadores de la planta los más expuestos a dosis de radiación capaces de causar efectos en la salud". Las OMS indica que la exposición a la radiación puede incrementar el riesgo de contraer cáncer, especialmente de tiroides. Por ello, recomienda que las personas expuestas a la radiación ingieran píldoras de yoduro de potasio en un corto lapso después de la exposición, una medida que ya han adoptado las autoridades japonesas entre la población más cercana a la central de Fukushima.

Tabla 2.7: Noticia de la razon.es

Noticia del periódico larazon.es
El único destino de los reactores de Fukushima es su desmantelamiento. Las autoridades japonesas continuaron ayer aplicando medidas de evacuación en el entorno de las centrales afectadas por el seísmo. Las más dañadas tendrán que ser desmanteladas. «Tras las explosiones en dos de los tres reactores de la nuclear de Fukushima es descabellado que vuelvan a funcionar. El proceso de desmantelamiento será largo y como es habitual antes de destruir habrá que «construir» y sobre todo ser paciente, ya que -como en cualquier proceso de parada-, los reactores mantienen energía latente (en torno al dos por ciento de la energía que tenían en operación) durante meses, explicaron fuentes vinculadas en su día a entidades nucleares y que coinciden con Gallego en que los reactores dañados son irrecuperables. «Antes de desmantelar los reactores, lo que hay que hacer ahora -prosigue Gallego- es restablecer el tendido eléctrico. Después, para poder acceder a las centrarles será necesario esperar más de un año para que dé tiempo a que el nivel de radiación vaya bajando. Será un proceso largo, ya que estas situaciones exigen una planificación meticulosa. Transcurrido ese año, habrá que reconstruir la planta de operación. Entonces tendrán que proceder a cubrir la zona de recarga de combustible de aquellos reactores (el 1 y el 3) que tienen esta zona al descubierto y limpiar la contaminación». Después deberá abrirse la zona de contención, que «en este tipo de reactores, de diseño similar al de Garoña, cuenta con una tapa sujeta con pernos. Tras esto, tendrán que enfriar la vasija hasta que no saquen todo el combustible. Para entonces habrá transcurrido un año y medio, y podrán sacar el material usado para generar energía siempre bajo agua, unos cuatro o cinco metros, y meterlo en contenedores también bajo este recurso para atenuar la radiación, ya que tendrán más contaminación que los de un desmantelamiento habitual». La ubicación posterior de los contenedores tendrán que resolverla las autoridades japonesas, concluyó Gallego, que añadió que «el uso de agua marina para el enfriamiento de reactores no es ideal, pero es la solución a una situación desesperada».

Tabla 2.8: Noticia de la razon.es

Resumen Manual
La alarma radiactiva se dispara tras una nueva explosión en el reactor dos de Fukushima. La alarma de un desastre nuclear crece en torno a la central de Fukushima, afectada en cuatro de sus seis reactores por el terremoto que devastó el país el viernes. La alarma crece en torno a la central de Fukushima, afectada en cuatro de sus seis reactores. Un incendio se ha desencadenado en el reactor 4, aunque ya está controlado, mientras que en el número 2 se produjo una explosión en torno a las seis de la mañana locales (22.00 en España). Los ingenieros y técnicos desplegados no consiguen normalizar la situación en la central nuclear de Fukushima, donde ya han reventado las cúpulas de dos reactores, mientras que se dañó el muro de contención de un tercero. Alrededor de un millar de cuerpos fueron hallados en una playa de la península de Ojika, mientras que otros tantos fueron encontrados en la ciudad de MinamiSanriku, donde al menos 9.500 personas -más de la mitad de la población- están en paradero desconocido. El Gobierno de Japón ha establecido una zona de exclusión aérea de un radio de 30 kilómetros sobre la central nuclear de Fukushima-1, ubicada en el noreste del país, según informa la agencia de noticias Kiodo. Las autoridades japonesas continuaron ayer aplicando medidas de evacuación en el entorno de las centrales afectadas por el seísmo. El primer impacto político se ha producido en Alemania: la canciller Merkel ha decidido suspender la prolongación del funcionamiento de sus 17 centrales nucleares en tanto se revisan los estándares de seguridad de las plantas. Fukushima ha vivido hoy una réplica de magnitud 6,3.

Tabla 2.9: Resumen manual

Configuración 1
Japón admite fugas radiactivas "que pueden afectar a la salud" tras un incendio y una nueva explosión en Fukushima. Se han registrado dos nuevas explosiones, en los reactores 2 y 3, y un incendio en el 4.- La situación es de máxima alerta.- En Tokio, a 240 kilómetros de la central nuclear de Fukushima, se han detectado partículas radiactivas y se teme que el fuerte viento lleve una nube tóxica al interior del país. Incendio y una tercera explosión. Japón va de mal en peor. Unos minutos después de la tercera explosión los niveles de radiación en los alrededores de Fukushima subieron hasta 8.217 microsieverts por hora, frente a los 1.941 que se registraban 40 minutos antes, según mediciones de la Agencia de Seguridad Nuclear japonesa. Los detractores de esta energía, en cambo, ven en el accidente la prueba de que la seguridad total no existe y de que el excesivo riesgo no compensa su uso. Alrededor de un millar de cuerpos fueron hallados en una playa de la península de Ojika, mientras que otros tantos fueron encontrados en la ciudad de MinamiSanriku, donde al menos 9.500 personas -más de la mitad de la población- están en paradero desconocido. Sin embargo, algunos medios creen que es posible que muchos de estos desaparecidos huyeran a tiempo a la vecina localidad de Tome, también en Miyagi. Japón continúa con el recuento de víctimas mortales del terremoto que el pasado viernes causó un devastador tsunami. Sin embargo, la Organización Mundial de la Salud ha querido enviar un mensaje tranquilizador: "Japón está tomando las medidas de salud públicas adecuadas para proteger a la población de la radiación", ha dicho Gregory Hartl, portavoz citado por la agencia Reuters. Fukushima ha vivido hoy una réplica de magnitud 6,3.

Tabla 2.10: Resumen automático generado por la configuración 1

Configuración 2
Japón admite fugas radiactivas "que pueden afectar a la salud" tras un incendio y una nueva explosión en Fukushima. Se han registrado dos nuevas explosiones, en los reactores 2 y 3, y un incendio en el 4.- La situación es de máxima alerta.- En Tokio, a 240 kilómetros de la central nuclear de Fukushima, se han detectado partículas radiactivas y se teme que el fuerte viento lleve una nube tóxica al interior del país. Incendio y una tercera explosión. Unos minutos después de la tercera explosión los niveles de radiación en los alrededores de Fukushima subieron hasta 8.217 microsieverts por hora, frente a los 1.941 que se registraban 40 minutos antes, según mediciones de la Agencia de Seguridad Nuclear japonesa. Japón va de mal en peor. Los detractores de esta energía, en cambo, ven en el accidente la prueba de que la seguridad total no existe y de que el excesivo riesgo no compensa su uso. Alrededor de un millar de cuerpos fueron hallados en una playa de la península de Ojika, mientras que otros tantos fueron encontrados en la ciudad de MinamiSanriku, donde al menos 9.500 personas -más de la mitad de la población- están en paradero desconocido. Sin embargo, algunos medios creen que es posible que muchos de estos desaparecidos huyeran a tiempo a la vecina localidad de Tome, también en Miyagi. Un portavoz de Tepco explicó en una confusa rueda de prensa retransmitida y doblada al inglés por Al Yazira: "Hay una posibilidad de que haya daño", pero inmediatamente añadió que eso no tenía por qué significar una fuga o que podía tratarse simplemente de una válvula que estuviera midiendo mal la presión. Sin embargo, la Organización Mundial de la Salud ha querido enviar un mensaje tranquilizador: "Japón está tomando las medidas de salud públicas adecuadas para proteger a la población de la radiación", ha dicho Gregory Hartl, portavoz citado por la agencia Reuters.

Tabla 2.11: Resumen automático generado por la configuración 2

Configuración 3
Japón admite fugas radiactivas "que pueden afectar a la salud" tras un incendio y una nueva explosión en Fukushima. «Tras las explosiones en dos de los tres reactores de la nuclear de Fukushima es descabellado que vuelvan a funcionar. Mientras tanto, continúa sin ser extinguido un incendio en el cuarto reactor de la planta, y no se está logrando enfriar correctamente los demás. Japón va de mal en peor. El viento está dispersando las partículas radiactivas hacia el océano", señaló una portavoz del organismo en conferencia de prensa. Honda también detendrá la fabricación al menos hasta el día 20, y ya ha cerrado todas sus plantas excepto una de motocicletas en la isla de Kyushu (sur), que parará mañana. La cifra no para de crecer: el hallazgo hoy de 2.000 cadáveres en la prefectura de Miyagi, al noreste del país, duplica el balance anterior de 1.897 muertos -la cifra oficial que por ahora reconoce el Gobierno-. Por su parte, la Agencia de Turismo de Japón ha declinado informar sobre los 2.500 extranjeros que se encontraban visitando la zona afectada por el seísmo, según recoge la agencia de noticias Kyodo. "La televisión ha informado de una explosión, pero durante una hora no se ha dicho nada a la oficina del primer ministro", añadió. Incendio y una tercera explosión. Japón continúa con el recuento de víctimas mortales del terremoto que el pasado viernes causó un devastador tsunami.

Tabla 2.12: Resumen automático generado por la configuración 3

Configuración 4
Japón admite fugas radiactivas "que pueden afectar a la salud" tras un incendio y una nueva explosión en Fukushima. Se han registrado dos nuevas explosiones, en los reactores 2 y 3, y un incendio en el 4.- La situación es de máxima alerta.- En Tokio, a 240 kilómetros de la central nuclear de Fukushima, se han detectado partículas radiactivas y se teme que el fuerte viento lleve una nube tóxica al interior del país. Incendio y una tercera explosión. Japón va de mal en peor. Alrededor de un millar de cuerpos fueron hallados en una playa de la península de Ojika, mientras que otros tantos fueron encontrados en la ciudad de MinamiSanriku, donde al menos 9.500 personas -más de la mitad de la población- están en paradero desconocido. Sin embargo, algunos medios creen que es posible que muchos de estos desaparecidos huyeran a tiempo a la vecina localidad de Tome, también en Miyagi. Japón continúa con el recuento de víctimas mortales del terremoto que el pasado viernes causó un devastador tsunami. Pero lo que quizá ha despertado mayor inquietud ha sido el daño sufrido por algunas plantas nucleares situadas en la zona más castigada por los terremotos, empezando por el mayor, de magnitud 8,9, y el tsunami posterior. Será un proceso largo, ya que estas situaciones exigen una planificación meticulosa. Un rosario de desgracias interconectadas que ha dejado ya miles de muertos, cuyos cuerpos la marea empezaba ayer a devolver a la costa. Las autoridades temen que lleguen a 10.000.

Tabla 2.13: Resumen automático generado por la configuración 4

Configuración 5
Japón admite fugas radiactivas "que pueden afectar a la salud" tras un incendio y una nueva explosión en Fukushima. «Tras las explosiones en dos de los tres reactores de la nuclear de Fukushima es descabellado que vuelvan a funcionar. El Gobierno japonés admitió a primera hora que "puede haberse producido una fuga de materiales radiactivos", especialmente por causa del incendio, "que pueden afectar a la salud humana". Japón va de mal en peor. Mientras tanto, continúa sin ser extinguido un incendio en el cuarto reactor de la planta, y no se está logrando enfriar correctamente los demás. Incendio y una tercera explosión. El viento está dispersando las partículas radiactivas hacia el océano", señaló una portavoz del organismo en conferencia de prensa. Japón continúa con el recuento de víctimas mortales del terremoto que el pasado viernes causó un devastador tsunami. Fukushima ha vivido hoy una réplica de magnitud 6,3. Ya basta», se derrumbaba ayer MichikoIkezawa, una voluntaria que repartía víveres entre los desplazados.

Tabla 2.14: Resumen automático generado por la configuración 5

Bibliografía

- Casillas Rubio, Arantza; Fresno Fernández, Víctor; Martínez Unanue, Raquel; Montalvo Herranz, Soto*. Evaluación del clustering de páginas web mediante funciones de peso y combinación heurística de criterios*. Sociedad Española para el Procesamiento del Lenguaje Natural2005.

- Plaza Morales, Laura. *Generación automática de resúmenes con apoyo en ontologías aplicada al dominio biomédico.* Departamento de Ingeniería del Software e Inteligencia Artificial de la Universidad Complutense de Madrid 2008.

- Documentación sobre Lucene: http://lucene.apache.org/

- Documentación sobre Lingpipe: http://alias-i.com/lingpipe/index.html