

Práctica obligatoria evaluable: Procesamiento de noticias

Normativa de entrega

La entrega de esta práctica es *obligatoria* y debe realizarse *individualmente*. A continuación se detallan otros datos de interés relacionados con la normativa de entrega:

- La fecha límite de entrega de la práctica será el día **16 de Enero de 2015 a las 09 am**.
- La entrega de la práctica se hará a través del Campus Virtual, empleando el **formulario** habilitado para ello (se habilitará más adelante, se pondrá una noticia cuando esté disponible).
- Se deberá subir el proyecto Eclipse generado para la resolución de la práctica en un único fichero zip. Este fichero deberá presentar el siguiente nombre:
“PP_Ruby_”+“número de expediente del alumn@”.
Las prácticas que no sigan este formato no se evaluarán. Ejemplo: PP_Ruby_123.zip
- Sobre la **defensa de la práctica** se informará más adelante de los detalles a través de una noticia en el Campus Virtual.

Descripción

Una empresa internacional de marketing, como parte de su estrategia empresarial analiza o revisa la prensa diaria, procesando noticias de actualidad de diferentes fuentes para ver oportunidades de negocio. Actualmente este análisis lo hace “manualmente”, de forma que varios de sus empleados se encargan del análisis y procesamiento de las noticias, pero para ser más eficientes quieren automatizar todo el proceso. Saben que las técnicas de Procesamiento del Lenguaje Natural (PLN) se utilizan cada vez más para el análisis automático de textos, sean del tipo que sean, por lo que quieren utilizarlas para automatizar su proceso.

En esta práctica se pretende desarrollar el software que esta empresa necesita, por lo que se va a implementar un programa en Ruby que sea capaz de procesar una colección de noticias de actualidad con el fin de conocer diferente información relacionada con cada noticia, con las fuentes que publican las noticias, si hay noticias relacionadas o no, etc. En definitiva, la empresa es la que decidirá qué necesita conocer de las noticias y cómo lo quiere.

La información de la que se partirá es una colección de noticias, donde cada noticia se almacena en un fichero de texto diferente (con extensión `.txt`). Algunos ficheros contienen una versión reducida de la noticia original, incluyendo título, fuente, fecha de publicación y un resumen. En la Figura 1 se muestra un ejemplo.

```
Condenan a 36 años de cárcel al capitán del ferri surcoreano «Sewol»  
<La Voz de Galicia, 11/11/2014>  
El capitán del buque Sewol, cuyo naufragio dejó 304 muertos en una de las mayores tragedias
```

Figura 1. Formato del fichero de texto con una versión reducida de noticia.

Otros ficheros contienen la noticia al completo, por lo que además de la información contenida en la versión reducida, contienen además todos los párrafos que completan el contenido de la noticia original, tal y como se puede ver en la Figura 2.

```
La Guardia Civil detiene a 32 personas en una nueva redada contra la corrupción  
<El Mundo, 11/11/2014>  
La Guardia Civil la ha bautizado como 'operación Enredadera'. Va más allá de la denominada 'or  
  
De las 30 detenciones, 22 se han efectuado en Andalucía y casi la mitad de ellas en Sevilla. I  
Diez de las detenciones han sido en Sevilla, cuatro en Córdoba, tres en Jaén, otras tantas en  
Además, los agentes han arrestado a dos personas en Zaragoza y otras tantas en Madrid y la mis  
La Guardia Civil está desarrollando registros desde primeras horas de la mañana en las provinc  
Entre los detenidos, por el momento, se encuentran el funcionario de la Diputación de Sevilla,  
También ha sido detenido el portavoz de Coalición Canaria en el Cabildo de Lanzarote, Sergio M  
Se han registrado, entre otros sitios, las diputaciones provinciales de Sevilla, Córdoba y Jaé  
En Jaén han sido detenidas tres personas: el teniente de alcalde de Deportes del Ayuntamiento  
En el caso de Huelva, la Guardia Civil ha confirmado la detención de tres personas. Se trata d
```

Figura 2. Formato del fichero de texto con una noticia completa.

Toda la colección de noticias se proporciona en una carpeta llamada “newsCorpus”, que contiene las noticias en versión reducida y completas indistintamente. Además, los nombres de los archivos que contienen las noticias siguen unas reglas determinadas en función de la fuente a la que pertenezcan. A continuación se detalla el listado de fuentes y el formato de los nombres de archivos que contienen las noticias (en todos los casos XX representa un número entero):

- El Mundo – MUNXX.txt.
- El País – PAXX.txt.
- ABC – ABCXX.txt.
- La Voz de Galicia – VGAXX.txt
- La Vanguardia – VANXX.txt
- El Adelantado – ADELXX.txt

La empresa de marketing suele realizar análisis particulares dependiendo de la fuente de las noticias y como se persigue la eficiencia, puede resultar muy interesante conocer rápidamente a qué fuente pertenece cada noticia para almacenarlo debidamente acorde con esta información.

Por otra parte, todas las redacciones de los diferentes periódicos utilizan el mismo software de edición de textos, el cual tiene un bug y de forma aleatoria hay veces que no guarda bien las palabras de los títulos, cometiendo errores tales como: pérdida de letras, intercambio de consonantes homófonas como b/v o c/q/k, acortamientos o contracciones, etc. Las técnicas de PLN a aplicar sobre los textos no funcionarán bien si de vez en cuando se producen alteraciones en los títulos de algunas noticias. Por lo tanto, mientras se soluciona el bug del software de edición, al procesar las noticias hay que comprobar primero si el título es correcto o no. Es necesario entonces transformar el texto con errores a lenguaje estándar, es decir, hay que normalizar el texto. Para poder hacerlo se proporciona un fichero de texto (`normalization.txt`) con una tabla con posibles equivalencias entre errores que se pueden encontrar y lo que sería su correspondencia correcta. La Tabla 1 presenta el contenido del fichero para normalizar.

Palabra incorrecta	Palabra correcta
d	de
q	que
dj	deja
lbre	libre
x	por
l	la
u	un
n	en
cntr	contra
sbre	sobre
dl	del
xra	para
+	suma

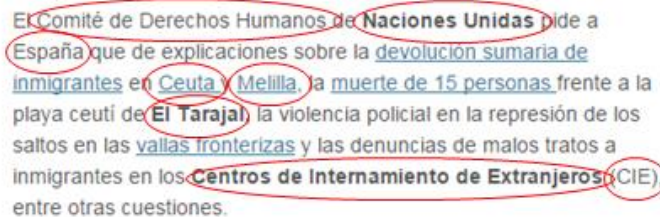
Tabla 1. Equivalencias entre errores léxicos y las palabras correctas.

El proceso de normalización debe ser lo más eficiente posible, ya que constantemente hay que consultar la tabla de equivalencias anterior varias veces para cada noticia que se procesa.

Algo de vital importancia para la empresa es poder consultar qué noticias están relacionadas de diferentes fuentes, e incluso, de diferentes días. Es decir, les interesa saber qué noticias están hablando de lo mismo, por lo que será necesario diseñar e implementar un algoritmo de agrupación de noticias, de manera que las noticias que informen del mismo suceso se puedan agrupar en el mismo grupo. Para ello será necesario establecer los criterios o condiciones que

deberán cumplir dos noticias para agruparse juntas. Dado que una agrupación real es subjetiva y puede variar dependiendo de quién la haga (diferentes personas leyendo las mismas noticias pueden agruparlas en grupos distintos), *cada alumno establecerá los criterios que le parezcan más convenientes para agrupar las noticias.*

Una forma de comparar si dos noticias están relacionadas es comparando las Entidades Nombradas (nombres propios de personas, lugares, organizaciones, etc.) que contienen, de tal manera que si comparten un buen número de Entidades Nombradas se podría considerar que las noticias narran el mismo suceso. En la Figura 3 se presenta una noticia destacando las Entidades Nombradas que contiene.



El Comité de Derechos Humanos de Naciones Unidas pide a España que de explicaciones sobre la devolución sumaria de inmigrantes en Ceuta y Melilla, la muerte de 15 personas frente a la playa ceutí de El Tarajal, la violencia policial en la represión de los saltos en las vallas fronterizas y las denuncias de malos tratos a inmigrantes en los Centros de Internamiento de Extranjeros (CIE), entre otras cuestiones.

Figura 3. Noticia con las Entidades Nombradas resaltadas en el texto.

Dado que se puede hacer referencia a una misma Entidad Nombrada de formas diferentes (*Barack Obama*, *Obama*, *Presidente Obama*), una forma de comparar si dos Entidades Nombradas se refieren a lo mismo es comparando su similitud ortográfica. Para ello se proporciona una clase en Ruby (`LCS`) con un método (`similar`) que permite comparar la ortografía de dos cadenas. Este método recibe dos cadenas y un porcentaje mínimo de similitud, de tal forma que devolverá `True` si la ortografía de las dos cadenas se parece al menos en el porcentaje indicado y devolverá `False` en caso contrario.

Para poder llevar a cabo el listado de cosas que se piden a continuación, lo primero que hay que hacer es leer la información de cada noticia. Se proporciona una clase en Ruby (`FileUtils`) con dos métodos que se pueden utilizar, respectivamente, para listar el contenido de un directorio (`list_files`) y para leer el contenido de un fichero de texto (`read_file`).

Se valorará positivamente un diseño Orientado a Objetos adecuado, donde la extensibilidad del código sea sencilla. También se valorará positivamente un acceso eficiente a la información. Igualmente se valorarán de forma positiva todas aquellas propuestas novedosas y que permitan posibles extensiones sobre el enunciado propuesto.

Todos los recursos necesarios para la realización de la práctica se proporcionan en un archivo adicional (`recursos.zip`).

Se pide lo siguiente:

1. Listado de fuentes que han publicado las noticias de la colección.
2. Normalizar los títulos de las noticias. De esta forma aquellos que tuvieran errores léxicos serán corregidos. Mostrar, para cada fuente, un listado de noticias mostrando el título erróneo y el título normalizado.
3. Dada una fuente, mostrar los titulares de las noticias publicadas por dicha fuente. En el caso de las noticias que son una versión reducida, al lado del título debe aparecer "(R)". Por ejemplo:

Alaya desmantela una "red criminal" para amañar contratos públicos (R)

4. Dada una fuente y una fecha, mostrar los titulares de las noticias publicadas en dicha fecha por la fuente. En el caso de las noticias que son una versión reducida, al lado del título debe aparecer "(R)".
5. Dado un número de párrafos, mostrar los titulares de todas las noticias que tengan un número de párrafos igual o mayor al dado.
6. Dado una cadena mostrar todas las noticias que contengan dicha cadena en el título. Se deben mostrar las noticias completas con el siguiente formato:

```
- - - - -  
Título  
Fuente - dd/mm/aaaa  
Texto de la noticia  
- - - - -
```

Y las noticias de version reducida con el siguiente format:

```
- - - - -  
Título (R)  
Fuente - dd/mm/aaaa  
Texto de la noticia  
- - - - -
```

7. Mostrar un listado de noticias ordenadas por su fecha de publicación. A igual fecha de publicación el listado deberá presentarse por orden alfabético de los títulos de las noticias. De cada noticia habrá que mostrar el título, la fuente y la fecha de publicación.
8. Dada una fuente, se deberá mostrar para cada noticia las Entidades Nombradas que contiene. Para ello será necesario identificar dichas entidades en el texto (título y cuerpo de la noticia). Se puede hacer buscando aquellas palabras que comiencen por mayúscula y no sean *stopwords* (palabras vacías de contenido como preposiciones, artículos). Se proporciona un listado de palabras vacías en un fichero adicional (*stopwords.txt*). Por ejemplo, en el siguiente texto se han marcado en rojo dos palabras que comienzan por mayúscula pero no son Entidades Nombradas, sino que son palabras vacías:

El capitán del ferry surcoreano Sewol, el buque cuyo hundimiento el pasado abril causó 304 muertos, fue condenado hoy a 36 años de cárcel por "negligencia grave". **El** tribunal

En este caso la única Entidad Nombrada que aparece es "Sewol". Dada la variabilidad de las Entidades Nombradas en los textos se puede asumir que habrá errores en la identificación de las entidades, por lo tanto es posible que no se puedan identificar correctamente todas las entidades en cada noticia.

9. Diseñar e implementar un algoritmo para agrupar noticias que informen del mismo evento o suceso de forma automática. Mostrar un listado de los grupos creados, donde en cada grupo aparezcan los titulares de las noticias. Por ejemplo:

Grupo 1

Alaya desmantela una "red criminal" para amañar contratos públicos
El País - "11"/"11"/"2014"
La Guardia Civil detiene a 32 personas en una nueva redada contra la corrupción
El Mundo - " 11"/"11"/"2014"
La Guardia Civil deja libre a cuatro de los 32 detenidos por la red de funcionarios corruptos
ABC - "11"/"11"/"2014"
Varios detenidos en la operación Madeja contra la corrupcion en Andalucía
La Vanguardia - "11"/"11"/"2014"

Grupo 2

Roger Federer, a un paso de las semifinales
La Vanguardia - "11"/"11"/"2014"

Grupo 3

China y EE UU acuerdan levantar aranceles de productos tecnológicos
El País - "11"/"11"/"2014"

...

10. Calcular ciertas estadísticas relacionadas con los grupos de noticias similares: número de grupos; número medio de noticias por grupo; número medio de noticias resumidas y noticias completas por grupo; número de grupos con todas las noticias de la misma fecha y número de grupos con noticias de fechas variadas y número de "grupos" que contienen una única noticia.
11. Por cada grupo mostrar un listado de palabras clave que identifiquen de alguna forma el tema de sus noticias.