

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 **11 de diciembre de 2014 a las 9 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_Haskell_"+"número de expediente del alumn@".
Las prácticas que no sigan este formato no se evaluarán. Ejemplo: PP_Haskell_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

Se quiere disponer de un programa en Haskell que sea capaz de procesar una colección de noticias de actualidad con el fin de conocer diferente información relacionada con cada noticia y también para agrupar de forma automática las noticias que estén relacionadas.

Cada noticia viene dada en un fichero de texto como el que se presenta en la Figura 1.


```
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 B
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 1. Formato del fichero de texto con una noticia.

Tal y como se puede ver en la imagen de la Figura 1, el formato de cada noticia es el siguiente: la primera línea del fichero de texto contiene el título de la noticia; la segunda línea contiene la fuente (el nombre del periódico que ha publicado la noticia) y la fecha de publicación, y el resto de líneas contienen el cuerpo de la noticia, donde cada línea representa un párrafo en la noticia original.

Se deberán procesar las diferentes noticias de la colección para consultar cierta información, pero también se debe 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 2 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 2. 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 un módulo en Haskell (*OrthographicMeasures*) que contiene una función (*similars*) que puede comparar la ortografía de dos cadenas. Esta función 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. Algunos ejemplos de aplicación de esta función son:

```
*OrthographicMeasures> similars "Barack Obama" "Obama" 70
False
*OrthographicMeasures> similars "Barack Obama" "Barack H. Obama" 70
True
*OrthographicMeasures> similars "Barack Obama" "Obama" 40
True
*OrthographicMeasures>
```

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 de los ficheros proporcionados. La información leída se estructurará adecuadamente construyendo nuevos tipos de datos (datos para representar una noticia, una colección de noticias, etc.).

Se valorará positivamente la claridad y extensibilidad del código, así como la definición de los tipos de datos más adecuados para la resolución del problema y la estructuración del código mediante diferentes funciones y módulos. Igualmente se valorarán de forma positiva todas aquellas propuestas novedosas y que permitan posibles extensiones sobre el enunciado propuesto.

Se pide lo siguiente:

1. Listado de fuentes que han publicado las noticias de la colección. Por ejemplo:

```
*Main> showNewspapers newsItems
El Mundo
El País
ABC
La Voz de Galicia
La Vanguardia
*Main>
```

Donde `showNewspapers` es el nombre de la función encargada de mostrar las diferentes fuentes de noticias y `newsItems` representa la colección de noticias.

2. Dada una fuente, mostrar los titulares de las noticias publicadas por dicha fuente. Por ejemplo:

```
*Main> showTitlesBySource newsItems "La Vanguardia"
Varios detenidos en la operación Madeja contra la corrupcion en Andalucía
El capitán del Sewol es condenado a 36 años de cárcel, pero es absuelto de
los cargos por homicidio
Suben a once las mujeres muertas tras operaciones de ligadura de trompas
en India
Roger Federer, a un paso de las semifinales
*Main>
```

3. Dado un número de párrafos, mostrar los titulares de todas las noticias que tengan un número de párrafos igual o mayor al dado. Por ejemplo:

```
*Main> showTitlesByParagraphs newsItems 8
La Guardia Civil detiene a 32 personas en una nueva redada contra la
corrupción
Once mujeres mueren en una campaña de esterilización en India
Alaya desmantela una "red criminal" para amañar contratos públicos
Mueren once mujeres en una campaña de esterilización en India
Condenado a 36 años de prisión el capitán del ferri surcoreano 'Sewol'
```

Descalifican a Marruecos por no acoger la Copa de África por el ébola
La Guardia Civil deja libre a cuatro de los 32 detenidos por la red de funcionarios corruptos
Varios detenidos en la operación Madeja contra la corrupción en Andalucía
Suben a once las mujeres muertas tras operaciones de ligadura de trompas en India
Roger Federer, a un paso de las semifinales
*Main>

4. 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

Por ejemplo:

```
*Main> showNewsBySearch newsItems "Roger"
```

```
- - - - -
```

Roger Federer, a un paso de las semifinales

La Vanguardia - "11"/"11"/"2014"

Londres. (EFE).- El suizo Roger Federer derrotó al japonés Kei Nishikori por 6-3 y 6-2, logró la victoria 70 en lo que va de año y se situó a un paso de las semifinales del Masters de Londres, clasificación que puede lograr esta misma noche, dependiendo del partido entre el británico Andy Murray y el canadiense Milos Raonic.

Si Raonic vence a Murray o el británico se impone al canadiense en tres sets, Federer lograría la clasificación por duodécima ocasión para las semifinales de este torneo, que ha ganado ya en seis ocasiones (récord), la última hace tres años.

Aunque se escucharon gritos de apoyo al japonés, uno de los tres novatos este año, el público del O2 se decantó de nuevo hacia Federer, excepto una aficionada que mostraba una pequeña pancarta con la frase: "Rafa ponte bien pronto", dirigida al español Rafael Nadal, ausente del torneo debido a su operación de apendicitis.

A Federer le bastó con apretar en los momentos claves del partido para imponerse físicamente al japonés, que pidió fisioterapeuta dos veces por un problema en la muñeca derecha.

La cara del estadounidense de origen chino Michael Chang y la del argentino Dante Bottini, ambos entrenadores de Nishikori, reflejaban la impotencia de Kei sobre la pista azul del O2, donde intentaba sin fortuna la tercera victoria sobre el helvético en cinco encuentros.

El japonés, que venció en la jornada inaugural a Murray por un doble 6-4, se vio mermado por su problema físico y lo notó especialmente en su saque. Lo cedió en tres ocasiones y cometió cinco dobles faltas, una de ellas crucial cuando concedió su servicio en el séptimo juego del segundo parcial.

Federer conservó el suyo durante todo el partido y se hizo con la victoria en 69 minutos, con siete saques directos.

- - - - -

*Main>

5. 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.
6. 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.

7. 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:

*Main> showGroups newsItems

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"

...

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