

RECUPERACIÓN DE INFORMACIÓN EN INTERNET

P.O.P. Lengua y Usos Profesionales

Facultad de Filología, Universidad de A Coruña

Curso 2009/2010

TRABAJO PRÁCTICO

El siguiente trabajo práctico consta de 3 ejercicios que debes resolver utilizando las técnicas de recuperación de información que hemos visto a lo largo del curso.

El trabajo es individual. Debes elaborar una breve memoria explicativa de los pasos que has seguido para resolver cada ejercicio (aprox. 2 o 3 páginas para cada ejercicio). Posteriormente, debes enviar dicha memoria a la dirección de correo electrónico del profesor Carlos Gómez Rodríguez (cgomezr@udc.es).¹ La fecha límite de entrega es el 12 de enero de 2010, a las 23:59.

Los profesores de este curso revisarán los trabajos presentados, y comunicarán las calificaciones a los alumnos, reservándose el derecho de convocarlos previamente a una defensa personal, en caso de que algún aspecto del trabajo no esté expuesto de manera clara.

Por supuesto, desde hoy hasta la fecha de entrega, puedes consultar a los profesores del curso las dudas relacionadas con la realización del trabajo, o bien por correo electrónico, o bien en persona durante las horas de tutorías que tenemos reservadas para ello.

Los ejercicios son los siguientes:

1. Preparación de un viaje.

Queremos pasar una semana en Torremolinos, del 7 al 13 de diciembre de 2009². Queremos alojarnos en un camping de primera categoría o, en su defecto, en uno de segunda categoría. Queremos viajar en tren desde A Coruña hasta Málaga³ y, una vez en Málaga, queremos llegar hasta Torremolinos utilizando algún otro tipo de transporte público (autobús o tren de cercanías). Durante el tiempo que estemos alojados, para poder realizar alguna excursión, queremos alquilar un coche durante 2 o 3 días.

Averigua si todo esto es posible. Indica aquellas búsquedas que han sido fructíferas y aquellas que no. Haz mención explícita de aquellas páginas que permiten directamente efectuar reservas y de aquellas que no. En caso de que necesites más información, y que la única manera de obtenerla sea ponerse en contacto con alguna persona de alguna de las entidades consultadas, puedes hacerlo. Simplemente, hazlo constar y, por supuesto, no lo hagas por teléfono, sino por correo electrónico (la idea es no gastar absolutamente ningún dinero en los preparativos) y procura no molestar en exceso, ni ser demasiado insistente (ya que al final se trata de un ejercicio, no de efectuar reservas reales).

¹El servidor web de la UDC puede rechazar correos provenientes de Hotmail. Se recomienda no utilizar estas direcciones para asegurar la recepción.

²Si estas fechas ya han pasado cuando hagas la práctica, puedes cambiarlas por otra semana, por ejemplo 23 a 29 de marzo de 2010. Se aconseja que la semana sea del año en curso, ya que de lo contrario puede haber dificultades para buscar billetes (es decir, algunas páginas web de trenes o autobuses pueden no mostrar billetes/horarios para 2010 hasta que empiece el año 2010).

³En el caso de que no haya trenes en el día de ida o vuelta, podemos ir el día siguiente o volver el anterior.

Si lo deseas, puedes cambiar el destino, las fechas, el tipo de alojamiento y el medio de transporte, por otros que te parezcan más atractivos o que te interesen más. Eso sí, intenta que la *complejidad* del viaje y de las actividades que se van a realizar durante los días de estancia sea similar a la del supuesto que te proponemos.

Al final, queremos también que nos transmitas tu opinión sobre si consideras o no que Internet y las herramientas de recuperación de información son verdaderamente útiles para preparar este tipo de actividad, y queremos también que nos indiques cuál es el tiempo medio que consideras necesario para realizar toda la operación, desde que comienzas las primeras búsquedas, hasta que confirmas las reservas. Aunque puede ser muy variable, dependiendo de las fechas y del destino, este último dato te servirá para otra vez, para saber con cuánto tiempo de antelación se deben comenzar los preparativos.

2. Búsqueda de información personal.

En este ejercicio vamos a comprobar hasta qué punto las técnicas de recuperación de información en Internet se pueden emplear para encontrar información sobre personas individuales.

Para ello, tomaremos como referencia la siguiente lista de admitidos y excluidos:

http://www.iuii.ua.es/master_ti/Listadofinaladmitidos_junio.pdf

Se pueden encontrar numerosas listas similares en la Red, y hemos escogido ésta de forma totalmente aleatoria.

Escoge tres de las siete primeras personas de la lista, y averigua toda la información personal sobre ellas que puedas. Por supuesto, debes limitarte a información que esté disponible públicamente en Internet (es decir, no es válido contactar con nadie a través de mail, msn, facebook o cualquier otro medio, ya que no se trata de molestar a nadie). Si sobre alguna persona en particular te resulta muy difícil encontrar información, puedes cambiarla por otra dentro de las siete primeras.

Ejemplos de información personal que puedes intentar recopilar son: direcciones de correo electrónico, teléfonos, direcciones postales, aspecto físico/fotos de las personas, comunidades a las que pertenecen, trabajos que han llevado a cabo, asignaturas en las que se han matriculado, becas y otra información académica, participaciones en foros o blogs, o cualquier otra cosa que consideres que sería relevante si quisieras elaborar un perfil de esa persona.

Para cada dato que obtengas, incluye en la memoria de la práctica un enlace a la página web donde aparece, y explica la forma en que lo has encontrado, incluyendo las consultas en buscadores en caso de utilizarlos.

Comenta si buscar información sobre unas personas te ha resultado más difícil que sobre otras, y en tal caso cuáles crees que son los motivos de estas diferencias.

A continuación, puedes experimentar buscando en Internet información personal sobre ti mismo/a; aunque no la incluyas en la memoria.

Finalmente, queremos que nos transmitas tu opinión sobre la privacidad de los datos personales en Internet. ¿Cuáles son las ventajas y los inconvenientes de que esta información esté públicamente disponible en la Red? ¿Dónde crees que debería situarse el límite entre lo que se hace público y lo que no? ¿Piensas que en general deberíamos tomar precauciones para proteger nuestros datos personales en Internet, o simplemente aceptar que datos que solían considerarse privados ya no lo sean tanto?

3. Estudio comparativo del rendimiento de varios sistemas de recuperación de información.

Elige 3 buscadores. Elige una consulta relacionada con algún tema que conozcas bien. Lanza esa consulta en cada uno de los 3 buscadores. Fíjate en los 10 primeros resultados de cada buscador. Calcula la unión de esos documentos. Serán como mucho 30 documentos, aunque probablemente sean menos, ya que es de suponer que algún documento será recuperado por más de un buscador. Llamemos entonces D al número de documentos diferentes, siendo, como ya hemos dicho, $D \leq 30$. Imagina ahora que esos D documentos son toda la base de datos documental de la que disponemos. Accede a cada uno de esos D documentos, y decide si es relevante o no para la consulta que has hecho. En función de esas relevancias, calcula la precisión, el *recall* y la precisión no interpolada de cada buscador. Indícanos qué buscadores has utilizado, cuál ha sido la consulta, cómo son las tres listas sí/no de relevancias, cuál es el valor de cada una de las tres medidas de rendimiento, y qué conclusiones sacas al respecto.

Repite de nuevo el mismo experimento, pero esta vez con una consulta relacionada también con algún tema que conozcas, igual que antes, para que seas capaz de juzgar rápidamente la relevancia de cada documento recuperado, pero que sea de un tema completamente diferente. ¿Qué ha ocurrido esta vez? El buscador que más te satisface ahora, ¿es el mismo de antes o es otro? ¿Qué conclusiones sacas al respecto?

Por supuesto, en cada uno de los dos experimentos, tomando inicialmente un único buscador, puedes tantear con la consulta e intentar refinarla hasta que el buscador ofrezca resultados interesantes. Pero eso sí, una vez que encuentres la consulta que te gusta, lánzala tal cual en cada uno de los tres buscadores, para que todos ellos compitan al mismo nivel y ninguno se vea perjudicado respecto a los demás.

Miguel A. Alonso Pardo (alonso@udc.es)
Carlos Gómez Rodríguez (cgomezr@udc.es)
Jesús Vilares Ferro (jvilares@udc.es)