BLOQUE 1. RECUPERACIÓN EN INTERNET

- **MC-F-001**. Tema 1. **Fundamentos de Recuperación en Internet** (PDF).
- **MC-F-002**. Tema 2. **Posicionamiento de recursos en Internet** (PDF).
- **MC-F-003**. Tema 3. **Sistemas de Recuperación. Crawlers** (PDF).
- **MC-F-004**. Tema 4. **Acceso y Recuperación de datos en la Web** (PDF).
- **MC-F-005**. Tema 5. **Adquisición de datos en la Web Semántica** (PDF).

1 Indique la respuesta correcta en relación a los tipos de Web.

1 Existen distintos tipos de web, como la web de datos, la web 1.0, la web social, la web profunda y la web semántica

2 Actualmente nos encontramos en la web 2.6

3 La web semántica hace inservible los demás tipos de Web

4 La web semántica se centra en dar semántica a los usuarios de la web 2.0

2 Indique la respuesta correcta.

1 En el ciclo de BigData la fase de enriquecimiento es la que muestra la rentabilidad del proyecto

2 La limpieza consiste en técnicas centradas en eliminar los registros que contienen difamaciones e insultos

3 La limpieza consiste en eliminar errores ortográficos, tipográficos y datos inexactos

4 Enriquecimiento y limpieza son tareas que se implementan de forma rápida y sin apenas coste, por lo que merece la pena incluirlas en los proyectos

3 Los sistemas pregunta respuesta

1 Son una alternativa a la web automatizada que remunera a algunos usuarios por responder consultas

2 Pueden precisar de herramientas PLN, ontologías, reglas heurísticas (obtenidas manualmente o por inteligencia artificial) y bancos de consultas realizadas en el pasado

3 Son el futuro de todo Internet, y harán inservibles todas las páginas web actuales

4 Es un proyecto que pretende ser implantado en el futuro, pero en el que buscadores como Google o el propio de iOS no han implementado hasta el momento

4 En posicionamiento

1 Los factores indirectos son los que tienen un impacto sobre los directos, por ejemplo aumentando el número de visitas potenciales de la página

2 El PageRank es una medida de la credibilidad de una página otorgada por el W3C (WorldWideWeb Consortium)

3 Una página con muchos enlaces entrantes siempre es más creíble

4 Los principales criterios que considera Google para mostrar una página en primera posición en los resultados es el número de visitas que esta tiene y el importe que pagan sus propietarios a Google para aparecer en esta posición

5 Un crawler

1 Solo existen en sistemas que implementan el modelo booleano

2 Son sistemas que mediante inteligencia artificial determinan el mínimo recorrido para escoger las páginas más relevantes

3 Es aconsejable que sea distribuido, escalable, eficiente y extensible, pero sin dejar de ser robusto y respetuoso con la política del sitio

4 El fichero robots.txt es una alternativa a los crawlers

6 En el BigData

1 Es irrelevante la calidad de los datos ya que en caso de ser escasa se puede compensar con el tratamiento de un mayor volumen de datos

2 El mayor coste de un proyecto BigData es el análisis de datos

3 La integración de distintas fuentes de datos no es precisa en los proyectos BigData, ya que se trata de datos homogéneos, muy estructurados y procedentes de una única fuente de datos.

4 La mayoría de los proyectos BigData fallan por una planificación deficiente

7 Las bases de datos BigData

1 Garantizan las cuatro propiedades ACID en las transacciones con la base de datos (atomicidad, consistencia, secuencialidad y durabilidad de las transacciones)

2 Como sistema de cómputo distribuido, se pueden garantizar los tres criterios del teorema CAP simultaneamente (Consistencia, Disponibilidad y Tolerancia a particiones)

3 Una base de datos NoSQL, como su nombre indica, no permite interrogar en el lenguaje de consulta SQL

4 Cassandra, MongoDB y CouchDB son bases de datos NoSQL

8 Para limpiar los registros es preciso:

1 Analizar manualmente los registros uno por uno, validándolos según un recurso de referencia externo

2 Distancia de edición, similitud fonética y frecuencia de subcadenas (p.e. con fingerprint) son técnicas frecuentemente utilizadas para limpiar registros

3 La limpieza de registros no es necesaria en BigData, ya que el gran volumen de datos procesado hace innecesario eliminar unos pocos registros con errores

4 Utilizar trabajadores reclutados por Internet (crowdsourcing) para limpiar registros es una técnica que ha demostrado ser ineficaz

9 Indique que característica de RDF es correcta:

1 Es un lenguaje de programación similar a Java que permite acceder más velozmente a los datos que con una base de datos relacional

2 Son tripletas para expresar hechos con la forma <sujeto><predicado><valor>

3 El lenguaje específico e idóneo para interrogar RDF es SQL

4 Si bien es una alternativa prometedora, no existen apenas recursos en Internet expresados en RDF

10 Indique la opción correcta en relación al lenguaje SPARQL:

1 Las tripletas son siempre simétricas en su significado, así "?pais dbo:leaderTitle ?gobernante" tiene una redacción equivalente a "?gobernante dbo:leaderTitle ?pais"

2 El elemento "FILTER regex(?x, “^p”, “i”)" significa que solo debe mostrar los registros que empiecen por x, contengan una p y una i es su cadena.

3 El elemento OPTIONAL indica que su contenido puede ser ignorado por no ser relevante para nuestra necesidad de información

4 Existen cuatro tipos básicos de consultas: SELECT, CONSTRUCT, ASK y DESCRIBE