Lead University
Bachillerato Ingeniería en Ciencia de Datos
Administración de datos
Jose Miguel Ugalde Rodriguez
Lina Acevedo Guiral

Segunda Prueba parcial
Porcentaje: 15%
Valor de la Tarea: 100 puntos
Fecha de Entrega: miércoles 11 de agosto de 2021 antes de las 9:00 pm

Instrucciones:

- La prueba parcial es en parejas. Cuando se presente el caso de dos o más pruebas parciales iguales se les anulará a todos los involucrados.
- Lea cuidadosamente la aprueba para completar todos los puntos que se solicitan.
- Se calificará únicamente lo que aparezca en el documento que debe entregar, se puede entregar en formato pdf o docx.

Sección 1. Preguntas sobre almacenamiento, preservación de datos y propiedad intelectual (40 puntos)

1. ¿Qué medidas recomienda para asegurar el correcto almacenamiento a corto plazo de sus datos de manera segura? Cite al menos 3 recomendaciones **(15 puntos)**

   Lo más recomendable es siempre tener respaldos de nuestros datos, por los menos que estén replicados en tres lugares.
   Recomendaciones
   - No compartir contraseñas o acceso a cuentas por terceros.
   - En el caso de información guardada en línea, la creación de contraseñas seguras, que incluyan, letras carácter y números.
   - De ser posible utilizar 2 métodos de autentificación ejemplo: el código bac

2. ¿Qué recomendaciones sugiere para la preservación a largo plazo de los datos, una vez finalizado el proyecto de investigación? Cite al menos 2 recomendaciones **(10 puntos)**
   Siempre tener varios respaldos por lo menos 3 y que los respaldos estén almacenados en diferentes respaldos, como en la nube, un disco duro y la computadora en uso.
   - Tener una clara gestión de los datos en caso de que sea necesario regresar para usar los datos recolectados.
   - Utilizar nomenclaturas que faciliten la comprensión de lo que guardar los archivos.
   - Implementar el uso de un read me, qué facilite entender cómo fueron organizados los datos en primera instancia.
   - Mantener cosistencia en la nomenclatura.

3. ¿Cuáles tipos de propiedad intelectual sugiere para la protección de los datos de un proyecto de investigación? ¿en qué consisten? Cite al menos 2 tipos de propiedad intelectual que aplican a los datos **(10 puntos)**
   La de derecho de autor: "en ella se establece un listado a manera de ejemplo de las obras objeto de tutela por derecho de autor, dentro de las cuales se puede citar: Libros, folletos, programas de cómputo" por esto creo que es la más relevante.
   Derechos patrimoniales: son susceptibles a de ser transmitidos y son temporales y el autor puede autorizar o prohibir cualquier forma de explotación económica de su obra.

4. ¿Qué aspectos legales debo considerar al realizar una investigación donde se almacene datos sencibles de personas en Costa Rica? **(5 puntos)**
   Tomado del sistema costarricense de información Jurídica

Cuando se soliciten datos de carácter personal será necesario informar de previo a las personas titulares o a sus representantes, de modo expreso, preciso e inequívoco,

a) De la existencia de una base de datos de carácter personal.
b) De los fines que se persiguen con la recolección de estos datos.
c) De los destinatarios de la información, así como de quiénes podrán consultarla.
d) Del carácter obligatorio o facultativo de sus respuestas a las preguntas que se le formulen durante la recolección de los datos.
e) Del tratamiento que se dará a los datos solicitados.
f) De las consecuencias de la negativa a suministrar los datos.
g) De la posibilidad de ejercer los derechos que le asisten.
h) De la identidad y dirección del responsable de la base de datos.

Cuando se utilicen cuestionarios u otros medios para la recolección de datos personales figurarán estas advertencias en forma claramente legible.

Se debe recibir un Otorgamiento del consentimiento

a) Exista orden fundamentada, dictada por autoridad judicial competente o acuerdo adoptado por una comisión especial de investigación de la Asamblea Legislativa en el ejercicio de su cargo.
b) Se trate de datos personales de acceso irrestricto, obtenidos de fuentes de acceso público general.
c) Los datos deban ser entregados por disposición constitucional o legal.

Se prohíbe el acopio de datos sin el consentimiento informado de la persona, o bien, adquiridos por medios fraudulentos, desleales o ilícitos.

5. (10) ¿Por qué es importante la consistencia de los datos dentro del ciclo de vida de los datos? **(10 puntos)** Ayuda a que otras personas dentro de la organización de forma rápida y fácil identifique de qué se está hablando, dónde encontrar qué parte de la información, en general permite el trabajo cooperativo. Y permiten el orden, fácil mantenimiento y la reutilización.

Sección 2. Consistencia de los datos (60 puntos)

6. Procesar el archivo dataset.xlxs en OpenRefine para mejorar la consistencia de los datos.
7. Los nombres de los autores: **(8 puntos)**
    a. Los nombres siguen la siguiente estructura: Apellido, nombre, Ejemplo Aaron, Jason *(Si tiene duda de los nombres adjunto la lista de nombres al final del examen)*
    b. Eliminar los caracteres especiales de los nombres
    c. Aplicar los métodos de cluster:
        i. Key collision – fingerprint
        ii. Key collision – metaphone3
        iii. Key Collision – Daitch-Mokotoff
        iv. Nearest neighbor – ppm

8. **Lugares de publicación: (5 puntos)**
    a. Quitar los caracteres especiales de los lugares de publicación
    b. Aplicar el método de cluster:
        i. Key collision – fingerprint
9. **Editorial: (5 puntos)**
    a. Quitar los caracteres especiales de la columna de editorial
    b. Aplicar los métodos de cluster:
        i. Key collision – fingerprint
        ii. Key collision – Metaphone3
        iii. Key collision – Daitch-mokotoff
        iv. Nearest neighbor – ppm
10. **Años de publicación: (5 puntos)**
    a. Los años deben de venir en formato de número, para esto quitar los caracteres especiales.
    b. Aplicar los métodos de cluster:
        i. Key collision – fingerprint
        ii. Key collision – ngram-fingerprint
    c. Transformar esta columna a número
11. Descarga el archivo procesado de OpenRefine **(2 puntos)**
12. Crear la estructura de carpetas para datos sin procesar y los datos procesados. **(5 puntos)**
13. Crear la documentación que considere necesaria para la correcta interpretación de la nomenclatura de las carpetas y los archivos. **(10 puntos)**
14. Subir los datos procesados y sin procesar utilizando el sistema de control de versiones de su preferencia y adjuntar el link. **(10 puntos)**

Datos sin inconsistencias

Nombres:

Aaron, Jason
Abnett, Dan
Adams, Art
Adams, Jeff
Adams, Neal
Adams, Scott
Adlard, Charlie
Anderson, Brent Eric
Anderson, Kevin J.
Andru, Ross
Appleby, Steven
Ashley, Bernard
Austin, David
Bachalo, Chris
Bagley, Mark
Baikie, Jim
Bair, Michael.
Barker, Martin
Barlow, Jeremy
Barreto, Luis Eduardo
Baxter, Glen
Beatty, Scott
Bedard, Tony
Belardinelli, Massimo
Bell, Anthea
Bell, Steve
Bellamy, Frank
Bellamy, Sasha
Bellus, Jean.
Bendis, Brian Michael
Benes, Ed
Bisley, Simon
Bolland, Brian
Bolton, John
Bond, Simon
Boothby, Ian
Boultwood, Dan
Brosseau, Pat
Bryant, Clive
Bryant, Mark

Bunnage, Mick
Campbell, Eddie
Carey, Mike
Cassaday, John
Chaykin, Howard V.
Christie, Agatha
Claremont, Chris
Clowes, Daniel

Place of Publication

Aldershot, Hants, England
Boston; London
Brighton
Edgeware
Edinburgh
Godalming
Harmondsworth
Holmes Chapel
La Jolla, CA
Leicester
London
Manchester
Milwaukie, OR
Mount Kisco; Watford
New York
New York ; Watford
New York, N.Y.
Nottingham
Oxford
Santa Monica, CA
Seattle
Southampton
Southend-on-Sea
Swindon
Sydney ; London
Todmorden
Towcester
Tresaith
Tunbridge Wells
Wadenhoe, Peterborough
Walton-on-Thames

Publisher

A. & C. Black
America's Best Comics
Arc
Arrow
Ashford
Ashgate
Assorted images
Bantam
Barker
Bellew
Black Library
Bloomsbury
Book Palace Books
Boxtree
Brockhampton Press
Cape
Chancellor
Classical
Classical Comics Ltd.
Cooper
Dark Horse Comics
DC Comics
Dorling Kindersley
Earthscan
Ebury
Eclipse Graphic Novels
Egmont Publishing Limited
Egmont-Methuen
Equinox
Escape
Exley
Fantagraphics
Flying Pig Enterprises
Foulsham
Franklin Watts
GeminiScan Publishing Co.
Grub Street
Hamilton & Co
Hamlyn
HarperCollins
Heinmann Kingswood
Hodder and Stoughton

Jonathan Cape
Knight Books
Knockabout
Knockabout Comics
Lion Publishing
Little, Brown
M. & J. Hobbs
Macdonald
Macmillan Children's
Mainstream
Manchester University Press
Mandarin
Manga Books
Marvel Comics
Methuen
Modern Toss
Nicholas Brealey
Old Street
Panini
Penguin
Peter Lang
Picador
Portico
Portrait
Prion
Private eye
Ravette
Rebellion
Reynolds & Hearn
Robson
Robson Books
Sam Books
Sanctuary
Scolar Press
Silver Link Publishing
Sphere
Thames and Hudson
Titan
Ty Mawr Publications
Usharp Comics
Vista
W.H. Smith
WildStorm Productions