|  |  |
| --- | --- |
| **Lead University**  **Bachillerato Ingeniería en Ciencia de Datos**  **Administración de datos** | **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)**
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)**
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)**
4. ¿Qué aspectos legales debo considerar al realizar una investigación donde se almacene datos sencibles de personas en Costa Rica? **(5 puntos)**

**Sección 2.** **Consistencia de los datos (60 puntos)**

1. Procesar el archivo dataset.xlxs en OpenRefine para mejorar la consistencia de los datos.
2. Los nombres de los autores: **(8 puntos)**
   1. 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)*
   2. Eliminar los caracteres especiales de los nombres
   3. Aplicar los métodos de cluster:
      1. Key collision – fingerprint
      2. Key collision – metaphone3
      3. Key Collision – Daitch-Mokotoff
      4. Nearest neighbor – ppm
3. **Lugares de publicación: (5 puntos)**
   1. Quitar los caracteres especiales de los lugares de publicación
   2. Aplicar el método de cluster:
      1. Key collision – fingerprint
4. **Editorial: (5 puntos)**
   1. Quitar los caracteres especiales de la columna de editorial
   2. Aplicar los métodos de cluster:
      1. Key collision – fingerprint
      2. Key collision – Metaphone3
      3. Key collision – Daitch-mokotoff
      4. Nearest neighbor – ppm
5. **Años de publicación: (5 puntos)**
   1. Los años deben de venir en formato de número, para esto quitar los caracteres especiales.
   2. Aplicar los métodos de cluster:
      1. Key collision – fingerprint
      2. Key collision – ngram-fingerprint
   3. Transformar esta columna a número
6. Descarga el archivo procesado de OpenRefine **(2 puntos)**
7. Crear la estructura de carpetas para datos sin procesar y los datos procesados. **(5 puntos)**
8. Crear la documentación que considere necesaria para la correcta interpretación de la nomenclatura de las carpetas y los archivos. **(10 puntos)**
9. Subir los datos procesados y sin procesar utilizando el sistema de control de versiones de su preferencia y adjuntar el link. **(10 puntos)**
10. ¿Por qué es importante la consistencia de los datos dentro del ciclo de vida de los datos? **(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 |