|  |  |
| --- | --- |
| **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 prueba 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)**

* En cuanto a la preservación de los datos, hay que tomar en cuenta el formato en el que se van a estar preservando, porque hay datos físicos, además de datos digitales, asegurándonos que se mantenga la seguridad e integridad de los datos utilizando diferentes métodos que pueden variar según el tipo de dato e incluso el presupuesto que se tenga.
* Tomar en cuenta el sistema de control de versiones a utilizar, por lo que es importante definir desde el inicio cómo se va a estar manejando ese versionamiento durante el proyecto, de tal forma que se puedan gestionar los datos de forma ágil, esto en funcional debido a que, por ejemplo, se puede regresar a un estado anterior de la información o de los datos y seguir toda su evolución.
* Tener un respaldo siempre es importante y esto aplica para cualquier tipo de almacenamiento de datos, por lo que anticipando cualquier inconveniente y estando preparados para el peor escenario, debemos crear un respaldo que va a depender del tipo de datos con el que estemos trabajando.

1. **¿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)**

* Como primer punto, es de vital importancia tener en cuenta las diferentes políticas y regulaciones que existen para la preservación de datos, de tal manera que podamos evitar cualquier tipo de problema legal que esto pudiera ocasionar, esto se podría evitar contando con un asesor legal, que pueda guiarnos y mantenernos alineados a las políticas, de igual forma, esto depende del tipo de información que estemos manejando y el uso que le vayamos a dar.
* Uno de los puntos más importantes es tener un Backup para evitar pérdida de datos o de información, por lo que se podría realizar una copia de seguridad por medio de un servidor web, donde esta copia se haga de manera diaria, esto podría hacerse de manera remota desde el servidor, manteniendo dos copias de cada respaldo, estas se pueden guardar en un disco casi en línea.

1. **¿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)**

* **Los derechos de autor:** Estas son facultades morales y patrimoniales que tiene un creador sobre una “obra literaria o artística" que elabora. Este derecho nace con la creación de la obra, es decir, su goce y protección es automática, sin embargo, la ley refuerza su reconocimiento y establece procedimientos para su registro, en caso de que sea una violación al derecho, el autor cuenta con material probatorio para proceder legalmente. Algunos de los objetos que protege son publicaciones científicas, programas de cómputo, bases de datos, obras dramáticas, bases de datos, programas de cómputo, etc.
* **Patentes**: Es un derecho exclusivo concedido a una invención, puede ser un producto o procedimiento que aporta una nueva manera de hacer algo o una nueva solución técnica a un problema. Entre sus ventajas está que dan al titular de los derechos la exclusividad para la explotación comercial e industrial de su invención, brinda protección al inventor o el titular del derecho de manera que puede tener ventaja competitiva, además, es un medio de divulgación del conocimiento para que terceros puedan realizar sus propias invenciones.

1. **¿Qué aspectos legales debo considerar al realizar una investigación donde se almacenen datos sensibles de personas en Costa Rica?** **(5 puntos)**

* Se debe tener en cuenta que quien recopila datos personas deberá obtener el consentimiento expreso de la persona titular de los datos o de su representante. Este consentimiento deberá constar por escrito, ya sea en un documento físico o electrónico.
* Cuando se soliciten datos personales será necesario informar de previo a las personas titulares o a sus representantes, de modo expreso, preciso e inequívoco de la existencia de una base de datos de carácter personal, de los fines que se persiguen con la recolección de estos datos, de los destinatarios de la información, así como de quiénes podrán consultarla, entre otras

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

1. Procesar el archivo dataset.xlsx 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)**

La Consistencia de datos en el ciclo de los datos es importante ya que se encarga de la eliminación y verificación de la redundancia de datos y reduce en gran medida el riesgo de inconsistencias. Si los datos se archivan solo una vez, cualquier actualización debe realizarse solo una vez y está disponible de inmediato para todos los usuarios. Si los datos están duplicados y el sistema conoce esta redundancia, el propio sistema puede garantizar que todas las copias sigan siendo coherentes.

# 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 |