verato

Desafío de Programación: Generador de PII (Información de identificación personal)

Objetivo:

El objetivo de este reto será explorar soluciones al problema de generar Información de identificación personal de manera automatizada y realista, manteniendo ciertos patrones que pueden ser identificados en bases de datos de este tipo. Además, los estudiantes tendrán la oportunidad de usar sus habilidades de lógica, resolución de problemas, programación e integración con librerías externas, para generar código útil y confiable.

Introducción:

Desarrollar un código que genere información de identificación personal (en inglés PII - *Personally Identifiable Information*) de manera realista con datos ficticios. El término PII se utiliza para datos que contengan un conjunto de los siguientes atributos: prefijo, primer nombre, segundo nombre, apellido, sufijo, fecha de nacimiento (DOB – *Date of Birth*), número de seguridad social (SSN – *Social Security Number* en Estados Unidos), direcciones (actual y pasadas) y teléfonos (actual y pasados). Cada atributo podrá tener asociado un valor. El conjunto de atributos asociados a una persona se conoce como *récord*.

En el caso del nombre, una persona tiene un nombre legal y podrá tener cualquier alias asociado. Por ejemplo, en EE. UU. es común que las mujeres al casarse cambien su apellido por el de su pareja. En este caso, su nombre de soltera pasa a ser un alias. También consideramos como aliases cualquier apodo o variaciones de un nombre.

Dependiendo de la fuente de la que provengan los datos (PII) podrá tener un error de entre el 20 y 30%. Errores en los datos pueden ocurrir debido a una variedad de razones como errores de dedo al registrar los datos, información obsoleta, falsificación de datos o por imprecisiones en alguna de las diversas etapas de procesamiento que sufren los datos. Los errores más comunes son: transposición de primer y segundo nombre, errores de dedo en 1 o 2 caracteres del nombre, errores en DOB o SSN, nombres mezclados (*commingled,* uno de los aliases es completamente diferente al nombre), errores de transcripción (por ejemplo, cuando un 1 se transcribe erróneamente por un 7 o 9), y transposición de caracteres adyacentes. Estos errores deben verse representados en el generador de PII.

Para el Desarrollo del proyecto al alumno se le proporcionará:

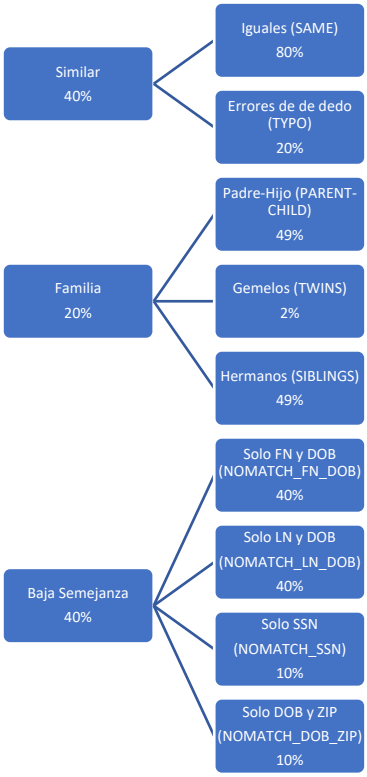
- *Records* semilla – estos récords contarán con todos los atributos base (primer nombre, apellido, DOB, SSN y dirección. El generador de PII será el responsable de leer los récords semilla y generar su correspondiente arco de identidades según los ajustes de configuración. En este caso, un arco será definido como el conjunto de récords generados a partir del récord semilla.
- Una versión abreviada de la lista de variantes de un nombre (primer nombre).

- Una lista de los 50 estados de conforman a Estados Unidos y la abreviación con las que éstos se conocen. También se proporcionará una lista de tipos de calle (*Street type*) válidos para las direcciones.
- Un diagrama en el que se muestre el formato de las direcciones en Estados Unidos.
- Una lista de prefijos y sufijos válidos para los nombres.

Consideraciones:
Cada arco asociado a un récord semilla generado debe contener diferentes casos según la proporción definida en la configuración. Los porcentajes para cada caso mencionado a continuación deben poder modificarse en las configuraciones del programa generado en el reto:

- **Similares:** Alto grado de semejanza con la semilla pero algunos valores en ciertos atributos con cierta "suciedad" o con nuevas direcciones. Si una persona fuera a comparar el récord con su semilla se sentiría cómodo con asumir que son la misma persona. Distribución de este caso en el arco: 40%
  - Casos que podemos encontrar entre récords similares:
    - SAME: Los récords generados son exactamente iguales a la semilla, ocurrencia del 80%
    - TYPO: Algunos errores de dedos en los datos comparados a la semilla, ocurrencia del 20%
- **Familia:** En estos casos el sexo del récord generado puede ser elegido de manera aleatoria (usando nombres asociados normalmente a este sexo). Esto significa que el sexo del récord generado puede ser opuesto al de la semilla de manera aleatoria. Distribución de este caso en el arco: 20%
  - Casos que podemos encontrar entre récords asociados a familiares:
    - TWINS / Gemelos: El apellido es el mismo, los SSNs difieren generalmente en un carácter en los últimos 4 dígitos, mismo DOB y al menos una de las direcciones es la misma. Ocurrencia del 2%
    - PARENT-CHILD / Padre-Hijo: El apellido es el mismo, el DOB debe estar separado por lo menos 20 años y al menos una de las direcciones es la misma. El nombre puede ser el mismo siempre y cuando el récord semilla tenga como sufijo Jr. o Sr es decir, si la semilla contiene el sufijo Jr. el nuevo récord generado para el arco puede tener el mismo nombre siempre y cuando tenga como sufijo Sr (o viceversa). Ocurrencia del 49%
    - SIBLINGS / Hermanos: El apellido es el mismo y al menos una de las direcciones es la misma. Ocurrencia del 49%
- **Baja semejanza**: Solo ciertos atributos compartidos con el récord semilla. Si una persona fuera a comparar el récord con su semilla no se sentiría cómodo con asumir que son la misma persona. Distribución de este caso en el arco: 40%
  - Casos que podemos encontrar entre récords con baja semejanza a su semilla:
    - NOMATCH_FN_DOB - Semejanza solo en su primer nombre (FN) y DOB. Ocurrencia del 40%

- NOMATCH_LN_DOB - Semejanza solo en su apellido (LN) y DOB. Ocurrencia del 40%
- NOMATCH_SSN - Semejanza solo en su SSN. Ocurrencia del 10%
- NOMATCH_DOB_ZIP - Semejanza solo en su DOB y código postal (*zip code*). Ocurrencia del 10%



Requerimientos Indispensables:

- El generador de PII debe crear un "arco" de récords que varíe en su parecido con el récord semilla entre alto y baja semejanza según los casos definidos anteriormente.
- El número de récords que generará como output final para cada arco deberá ser configurable.
- El porcentaje y tipo de casos que puede contener cada arco (ver consideraciones) debe ser configurable.
- El output esperado deberá contener al récord semilla junto con los récords generados por el programa.
- El valor de DOB deberá ser para personas mayores de 18 años.
- El output esperado deberá estar delimitado por *pipes* ("|" – Ver sección de output esperado para un ejemplo).
- Los nombres deberán incluir un nombre legal y hasta 3 aliases para el caso de similar.
- El generador de PII deberá medir la semejanza cada uno de los récords en el arco con su semilla asociada. Para esto podrán usar *embeddings*, análisis de clústers, semejanza de los textos y otros métodos que usen Machine Learning. Esta semejanza será parte del

output (Ver sección de output esperado para un ejemplo). Por ejemplo, el récord semilla tendrá una semejanza de 1 respecto a él mismo.

- Todos los récords generados deberán contener los siguientes atributos: Primer nombre (FN), Apellido (LN), Dirección completa, teléfono, SSN y DOB.
- El código de área asociado al teléfono deberá estar relacionado geográficamente al estado de alguna de las direcciones. (Ver *Links interesantes para el proyecto*)
- En el output, el atributo CASE Type deberá estar relacionado al caso que le corresponde según la descripción en la sección de consideraciones (como TWIN, PARENT-CHILD, NOMATCH_FN_DOB, etc.).

Estructura del archivo proporcionado (récord semilla) y output esperado

ID|Prefix|FirstName|MiddleName|LastName|Suffix|Name Alias-1|Name Alias-2|Name Alias-3|DOB|SSN|Address-1 Line 1|Address-1 Line 2| Address-1 City| Address-1 State| Address-1 Zip| Address-1 Zip4|Address-2 Line 1|Address-2 Line 2| Address-2 City| Address-2 State| Address-2 Zip| Address-2 Zip4|Phone-1 Area Code|Phone-1 Base Number|Phone-2 Area Code|Phone-2 Base Number|Gender|SimilarityScore|CASE Type

Ejemplo del output esperado:

123ABC||STANFORD||SMITH|MD|SMITH,STANFORD|S,F,SMOTH||1965-01-09|343679845|123 MAIN ST||MOSCOW|ID|83844||456 ELM RD||MOSCOW|ID|83844||208|3450998|208|4569845|M|1.0|SEED
123ABC||STANFORD||SMITH||SMITH,STANFORD|||1965-01-09|343679845|9456 STUDENT AVE||MOSCOW|ID|83844||208|3450998|208|4569845|M|0.95|SAME
123ABC ||LENNY||SMITH||LENORD,,SMITH|L,,SMITH||1965-01-09|343679846|123 MAIN ST||MOSCOW|ID|83844||208|3468863|||M|0.9|TWIN
123ABC ||MARY|M|BLACK||M,,SMITH|MARY,,SMITH||1995-04-01|346985557|123 MAIN ST||MOSCOW|ID|83844|||||M|0.89|PARENT-CHILD
123ABC ||STANFORD|P|JONES|||||1965-01-09|284348745|2302 SEMINOLE DR|APT 3|TALLAHASSEE|FL|32312||850|5468943|||M|0.3|NOMATCH

**Las líneas en negrita ejemplifican a un récord semilla en el output, los demás récords serán parte de su arco.*

¿Cómo esperamos ejecutar el código del desafío?

Nombre del ejecutable o endpoint, archivo con récord semillas, configuraciones (número de récords por arcos, porcentajes de cada grupo, etc.).

Notas:

- El ID será asignado por el código y será único para cada persona (todos los récords dentro del arco tendrán el mismo). Podrá ser alfanumérico y podrá tener hasta 10 caracteres.
- El código postal (*zip code*) será de 5 dígitos siempre.
- El DOB tendrá la siguiente estructura: AAAA-MM-DD
- El código de área para un teléfono siempre será de 3 dígitos.

- El número base para un teléfono siempre será de 7 dígitos.
- El SSN solo será válido si cumple con los siguientes criterios:
    - 9 dígito
    - Los primeros 3 dígitos no son 000, 666 o cualquier número entre 900 y 999.
    - Los caracteres 4 y 5 no pueden ser ceros (no pueden ser 00).
    - Los caracteres número 4 y 5 deben estar entre 01 y 99.
    - The digits 4-5 range from 01 to 99.
    - El número no contiene ceros en los últimos 4 dígitos.
    - El número no contiene 9 dígitos iguales (como 222222222) o 9 dígitos continuos consecutivos (como 123456789).

Links interesantes para el Proyecto:

1. https://www.census.gov/topics/population/genealogy/data/2010_surnames.html
2. https://namecensus.com/
3. https://www.allareacodes.com/area_code_listings_by_state.htm
4. https://pe.usps.com/text/pub28/28c1_001.htm