Trabalho 2 - δ-Presença

Javam Machado

Outubro 2019

1 Objetivo

- Implementar um algoritmo que anonimize um conjunto de dados de tal forma que seja atendido o modelo de privacidade δ-Presença.

2 Especificação

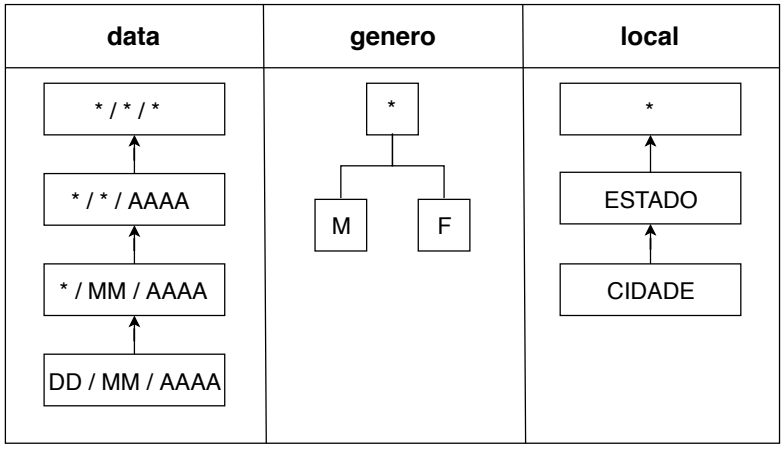
- Carregue o conjunto de dados "doencas.csv" (*private*), contendo doenças relacionadas a usuários de um hospital. Os atributos são:
  - *Identificadores Explícitos:* id;
  - *Semi-identificadores:* genero, data, cidade, estado;
  - *Sensíveis:* doenca.
- Carregue também o conjunto de dados "background.csv" (*public*), contendo informações externas a respeito de determinados usuários.
- O valor de δ deve variar da seguinte forma:

  $\delta = 40\%$ ($\delta_{min} = 10\%$ e $\delta_{max} = 50\%$);
  $\delta = 30\%$ ($\delta_{min} = 10\%$ e $\delta_{max} = 40\%$);
  $\delta = 20\%$ ($\delta_{min} = 10\%$ e $\delta_{max} = 30\%$);
  $\delta = 10\%$ ($\delta_{min} = 10\%$ e $\delta_{max} = 20\%$).
- Para cada conjunto de δ, o conjunto de dados deve ser anonimizado de tal forma que seja atendido o modelo δ-Presença.
- A métrica de utilidade a ser adotada deve ser a **Precisão**.

3 Requisitos

- Linguagem: Python

- Meio de entrega: criar um repositório chamado "disciplina_privacidade_2019" no Github e compartilhar com os seguintes e-mails: {andre.luis, iago.chaves, israel.vidal, javam.machado}@lsbd.ufc.br. **Todos os trabalhos da disciplina serão entregues através desse repositório**.
- Criar uma pasta "*delta*_presenca" no repositório "disciplina_privacidade_2019".
- Equipes de até 2 pessoas.
- Somente um repositório deve ser criado por equipe.
- O arquivo "README.md" no repositório deve conter os componentes da equipe.
- Utilize as seguintes hierarquias de generalização:



4 Avaliação

- O algoritmo implementado irá anonimizar um novo conjunto de dados "doencas_**x**.csv" (onde **x** = 40, 30, 20 e 10), que terá os mesmos atributos do arquivo "doencas.csv", mas **não necessariamente o mesmo número de linhas**.
- Na avaliação será considerada a:
  1. Corretude do algoritmo;
  2. Corretude da precisão;
  3. Apresentação da equipe;
  4. Qualidade da precisão.

5 Precisão

- Penaliza cada instância de um valor de atributo que é generalizado ou suprimido
- Quanto maior a precisão, maior a utilidade dos dados
- $Prec(D) = 1 - \frac{\sum_{i=1}^{N_a} \sum_{j=1}^{|D|} \frac{h}{|HGV_{A_i}|}}{|D| \times |N_a|}$
  - D: conjunto de dados
  - N_a: número de atributos semi-identificadores
  - h: altura da hierarquia de generalização de valor do atributo Ai após anonimização
  - $|HGV_{A_i}|$: altura máxima da hierarquia

6 Material de Apoio

`https://drive.google.com/open?id=1dcqpfVYoYQ1XGOFonhIRx5QAGWx1EfXN`

7 Entrega

- 15 de Outubro de 2019. *Commit* do código até 13:59.