Anonimizando variáveis

Importando biblioteca hashlib

In [0]:
```
import hashlib
```

Simplificando o comando

In [0]:
```
m = hashlib.sha256()
```

Teste de anonimização

Variável singular

In [0]:
```
m.update(b"Nobody inspects")
```

In [0]:
```
m.update(b" the spammish repetition")
```

Resultado da anonimização

In [0]:
```
m.digest()
```

Out[5]:
```
b'\x03\x1e\xdd}Ae\x15\x93\xc5\xfe\\\x00o\xa5u+7\xfd\xdf\xf7\xbcN\x84:\xa6\xaf\x0c\x95\x0fK\x94\x06'
```

Simplificação do algorítmo

In [0]:
```
h = hashlib.new('ripemd160')
h.update(b"fabiano")
h.hexdigest()
```

Out[6]:
```
'926fd0b43701a3e720e635cfded14ee6c8e8d2ce'
```

Teste em DataFrame

In [0]:
```
import pandas as pd
```

In [6]:
```
df_sinasc_es_1 = pd.read_csv('https://docs.google.com/spreadsheets/d/e/2PACX-1vQm3T1J0pWoiG-JofcT-K4gvs
df_sinasc_es_2 = pd.read_csv('https://docs.google.com/spreadsheets/d/e/2PACX-1vQIyzKl_u0t_YN_stW1Nr4VHW
df_sinasc_es_3 = pd.read_csv('https://docs.google.com/spreadsheets/d/e/2PACX-1vTIPBsiEtG3heF7zGO6HRkVae
df_sinasc_es_4 = pd.read_csv('https://docs.google.com/spreadsheets/d/e/2PACX-1vROjHq3_oX3W_j6GEVjNBoKdn
df_sinasc_es_5 = pd.read_csv('https://docs.google.com/spreadsheets/d/e/2PACX-1vRsmVmV6LY0nRKPr6MeqbNmH7
df_sinasc_es = pd.concat([df_sinasc_es_1, df_sinasc_es_2, df_sinasc_es_3, df_sinasc_es_4, df_sinasc_es_
df_sinasc_es.columns
```

Out[6]:
```
Index(['APGAR1', 'APGAR5', 'CODANOMAL', 'CODCART', 'CODESTAB', 'CODINST',
       'CODMUNCART', 'CODMUNNASC', 'CODMUNNATU', 'CODMUNRES', 'CODOCUPMAE',
       'CODPAISRES', 'CODUFNATU', 'CONSPRENAT', 'CONSULTAS', 'DIFDATA',
       'DTCADASTRO', 'DTDECLARAC', 'DTNASC', 'DTNASCMAE', 'DTRECEBIM',
       'DTRECORIG', 'DTRECORIGA', 'DTREGCART', 'DTULTMENST', 'ESCMAE',
       'ESCMAE2010', 'ESCMAEAGR1', 'ESTCIVMAE', 'GESTACAO', 'GRAVIDEZ',
       'HORANASC', 'IDADEMAE', 'IDADEPAI', 'IDANOMAL', 'KOTELCHUCK', 'LOCNASC',
       'MESPRENAT', 'NATURALMAE', 'NUMERODN', 'NUMERODV', 'NUMEROLOTE',
       'NUMREGCART', 'ORIGEM', 'PARIDADE', 'PARTO', 'PESO', 'PREFIXODN',
       'QTDFILMORT', 'QTDFILVIVO', 'QTDGESTANT', 'QTDPARTCES', 'QTDPARTNOR',
       'RACACOR', 'RACACORMAE', 'RACACORN', 'SEMAGESTAC', 'SERIESCMAE', 'SEXO',
       'STCESPARTO', 'STDNEPIDEM', 'STDNNOVA', 'STTRABPART', 'TPAPRESENT',
       'TPDOCRESP', 'TPFUNCRESP', 'TPMETESTIM', 'TPNASCASSI', 'TPROBSON',
       'Unnamed: 0', 'VERSAOSIST'],
      dtype='object')
```

In [0]:
```
df_sinasc_es.dtypes
```

Out[22]:
```
APGAR1         float64
APGAR5         float64
CODANOMAL       object
CODCART        float64
CODESTAB       float64
CODINST         object
CODMUNCART     float64
CODMUNNASC       int64
CODMUNNATU     float64
CODMUNRES        int64
CODOCUPMAE     float64
CODPAISRES     float64
CODUFNATU      float64
CONSPRENAT     float64
CONSULTAS      float64
DIFDATA          int64
DTCADASTRO       int64
DTDECLARAC     float64
DTNASC           int64
DTNASCMAE      float64
DTRECEBIM      float64
DTRECORIG      float64
DTRECORIGA     float64
DTREGCART      float64
DTULTMENST     float64
ESCMAE         float64
ESCMAE2010     float64
ESCMAEAGR1     float64
ESTCIVMAE      float64
GESTACAO       float64
                ...
NUMEROLOTE       int64
NUMREGCART     float64
ORIGEM           int64
PARIDADE       float64
PARTO          float64
PESO           float64
PREFIXODN      float64
QTDFILMORT     float64
QTDFILVIVO     float64
QTDGESTANT     float64
```

```
QTDPARTCES    float64
QTDPARTNOR    float64
RACACOR       float64
RACACORMAE    float64
RACACORN      float64
SEMAGESTAC    float64
SERIESCMAE    float64
SEXO           object
STCESPARTO    float64
STDNEPIDEM      int64
STDNNOVA        int64
STTRABPART    float64
TPAPRESENT    float64
TPDOCRESP     float64
TPFUNCRESP    float64
TPMETESTIM    float64
TPNASCASSI    float64
TPROBSON      float64
Unnamed: 0      int64
VERSAOSIST     object
Length: 71, dtype: object
```

In [10]:
```
df_sinasc_es['CODINST']
```

Out[10]:
```
0        MBA2910720001
1        MBA2913600002
2        MBA2922000001
3        MBA2922000001
4        MBA2922000001
5        MBA2922000001
6        MBA2922000001
7        MBA2931350001
8        MMG3101100001
9        MMG3101100001
10       MMG3101100001
11       MMG3101100001
12       MMG3101100001
13       MMG3101100001
14       MMG3101100001
15       MMG3101100001
16       MMG3101100001
17       MMG3101100001
18       MMG3101100001
19       MMG3101100001
20       MMG3101100001
21       MMG3101100001
22       MMG3101100001
23       MMG3101100001
24       MMG3101100001
25       MMG3101100001
26       MMG3101100001
27       MMG3101100001
28       MMG3101100001
29       MMG3101100001
             ...
52805    MES3204900001
52806    MES3204900001
52807    MES3204900001
52808    MES3204900001
52809    MES3204900001
52810    MES3204900001
52811    MES3204900001
52812    MES3204900001
52813    MES3204900001
52814    MES3204900001
```

```
52815    MES3204900001
52816    MES3204900001
52817    MES3205000001
52818    MES3205000001
52819    MES3205000001
52820    MES3205000001
52821    MES3205000001
52822    MES3204900001
52823    MES3205000001
52824    MES3205000001
52825    MES3201300001
52826    MES3201300001
52827    MES3201300001
52828    MES3201300001
52829    MES3201300001
52830    MES3201300001
52831    MES3201300001
52832    MES3201300001
52833    MES3205200001
52834    MES3205000001
Name: CODINST, Length: 273802, dtype: object
```

In [1]:
```
pip install unicode
```

```
Collecting unicode
  Downloading https://files.pythonhosted.org/packages/bd/e8/d30276462ce627d1fd2c547cdf64a7eccec20e86f4fcaa145e1ea8758641/unicode-2.7-py2.py3-none-any.whl (https://files.pythonhosted.org/packages/bd/e8/d30276462ce627d1fd2c547cdf64a7eccec20e86f4fcaa145e1ea8758641/unicode-2.7-py2.py3-none-any.whl)
Installing collected packages: unicode
Successfully installed unicode-2.7
```

In [3]:
```
df = df_sinasc_es
print(unidecode(df['XXXX']))
df
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-3-f7e40a4f75d0> in <module>()
----> 1 df = df_sinasc_es
      2 print(unidecode(df['DTNASCMAE']))
      3 df

NameError: name 'df_sinasc_es' is not defined
```

In [18]:
```
import pandas as pd
import hashlib
import random
import string


d1 = df_sinasc_es
display(d1.head())



# Get a unique list of the clear text, as a List
tmplist = list(set(d1['CODINST']))

# Add some random characters before and after the DRNASC
# Structured them in a Dictionary
# Example -- XXXX -> CODINST
mapping1 = {i : (''.join(random.choice(string.hexdigits) for i in range(12)))+i+(''.join(random.choice(s

# Create a DF to leave the original DF intact.
d2 = d1.copy()

# Create a new column containing clear_text_Nonce
d2['newname'] = [mapping1[i] for i in d2['CODINST']]
display(d2.head())




# Hash the clear_text+Nonce string
d2['hash'] = [hashlib.sha1(str.encode(str(i))).hexdigest() for i in d2['newname']]
display(d2.head())




# To proof that the same clear text has the same Hash string.
display(d2[d2['CODINST']=='MBA2910720001'].tail(30))


# Create a new DF with the clear-text column removed.
d3 = d2[['APGAR1', 'APGAR5', 'CODANOMAL', 'CODCART', 'CODESTAB', 'hash',
```

```
       'CODMUNCART', 'CODMUNNASC', 'CODMUNNATU', 'CODMUNRES', 'CODOCUPMAE',
       'CODPAISRES', 'CODUFNATU', 'CONSPRENAT', 'CONSULTAS', 'DIFDATA',
       'DTCADASTRO', 'DTDECLARAC', 'DTNASC', 'DTNASCMAE', 'DTRECEBIM',
       'DTRECORIG', 'DTRECORIGA', 'DTREGCART', 'DTULTMENST', 'ESCMAE',
       'ESCMAE2010', 'ESCMAEAGR1', 'ESTCIVMAE', 'GESTACAO', 'GRAVIDEZ',
       'HORANASC', 'IDADEMAE', 'IDADEPAI', 'IDANOMAL', 'KOTELCHUCK', 'LOCNASC',
       'MESPRENAT', 'NATURALMAE', 'NUMERODN', 'NUMERODV', 'NUMEROLOTE',
       'NUMREGCART', 'ORIGEM', 'PARIDADE', 'PARTO', 'PESO', 'PREFIXODN',
       'QTDFILMORT', 'QTDFILVIVO', 'QTDGESTANT', 'QTDPARTCES', 'QTDPARTNOR',
       'RACACOR', 'RACACORMAE', 'RACACORN', 'SEMAGESTAC', 'SERIESCMAE', 'SEXO',
       'STCESPARTO', 'STDNEPIDEM', 'STDNNOVA', 'STTRABPART', 'TPAPRESENT',
       'TPDOCRESP', 'TPFUNCRESP', 'TPMETESTIM', 'TPNASCASSI', 'TPROBSON',
       'Unnamed: 0', 'VERSAOSIST']].rename(columns={'hash':'CODINSTHash'})

display(d3.head())
```

	APGAR1	APGAR5	CODANOMAL	CODCART	CODESTAB	CODINST	CODMUNCART	CODMUNNASC	CODMUNNATU	CODMUNF
0	8.0	9.0	NaN	NaN	2507447.0	MBA2910720001	NaN	291072	291630.0	320
1	9.0	10.0	NaN	NaN	2802112.0	MBA2913600002	NaN	291360	291360.0	320
2	9.0	10.0	NaN	NaN	2506122.0	MBA2922000001	NaN	292200	320405.0	320
3	10.0	10.0	NaN	NaN	2506122.0	MBA2922000001	NaN	292200	320501.0	320
4	8.0	9.0	NaN	NaN	2498804.0	MBA2922000001	NaN	292200	314430.0	320

5 rows × 71 columns

	APGAR1	APGAR5	CODANOMAL	CODCART	CODESTAB	CODINST	CODMUNCART	CODMUNNASC	CODMUNNATU	CODMUNF
0	8.0	9.0	NaN	NaN	2507447.0	MBA2910720001	NaN	291072	291630.0	320
1	9.0	10.0	NaN	NaN	2802112.0	MBA2913600002	NaN	291360	291360.0	320
2	9.0	10.0	NaN	NaN	2506122.0	MBA2922000001	NaN	292200	320405.0	320
3	10.0	10.0	NaN	NaN	2506122.0	MBA2922000001	NaN	292200	320501.0	320
4	8.0	9.0	NaN	NaN	2498804.0	MBA2922000001	NaN	292200	314430.0	320

5 rows × 72 columns

	APGAR1	APGAR5	CODANOMAL	CODCART	CODESTAB	CODINST	CODMUNCART	CODMUNNASC	CODMUNNATU	CODMUNF
0	8.0	9.0	NaN	NaN	2507447.0	MBA2910720001	NaN	291072	291630.0	320
1	9.0	10.0	NaN	NaN	2802112.0	MBA2913600002	NaN	291360	291360.0	320
2	9.0	10.0	NaN	NaN	2506122.0	MBA2922000001	NaN	292200	320405.0	320
3	10.0	10.0	NaN	NaN	2506122.0	MBA2922000001	NaN	292200	320501.0	320
4	8.0	9.0	NaN	NaN	2498804.0	MBA2922000001	NaN	292200	314430.0	320

5 rows × 73 columns

	APGAR1	APGAR5	CODANOMAL	CODCART	CODESTAB	CODINST	CODMUNCART	CODMUNNASC	CODMUNNATU	CODM
0	8.0	9.0	NaN	NaN	2507447.0	MBA2910720001	NaN	291072	291630.0	
19288	2.0	2.0	Q742	NaN	2507447.0	MBA2910720001	NaN	291072	291180.0	
38096	9.0	10.0	NaN	NaN	2507447.0	MBA2910720001	NaN	291072	320000.0	
38097	8.0	9.0	NaN	NaN	2507447.0	MBA2910720001	NaN	291072	320305.0	
38098	8.0	9.0	NaN	NaN	2507447.0	MBA2910720001	NaN	291072	291465.0	
2067	9.0	10.0	NaN	NaN	2507447.0	MBA2910720001	NaN	291072	NaN	
2076	6.0	8.0	NaN	NaN	2507447.0	MBA2910720001	NaN	291072	292540.0	
2077	8.0	9.0	NaN	NaN	2507447.0	MBA2910720001	NaN	291072	291072.0	
2078	4.0	2.0	NaN	NaN	2507447.0	MBA2910720001	NaN	291072	292090.0	
2079	8.0	7.0	NaN	NaN	2507447.0	MBA2910720001	NaN	291072	291072.0	
2090	8.0	10.0	NaN	NaN	2507447.0	MBA2910720001	NaN	291072	292270.0	
2092	4.0	8.0	NaN	NaN	2507447.0	MBA2910720001	NaN	291072	291072.0	
2095	8.0	10.0	NaN	NaN	2507447.0	MBA2910720001	NaN	291072	291072.0	
29150	8.0	9.0	NaN	NaN	2507447.0	MBA2910720001	NaN	291072	291072.0	
7716	9.0	10.0	NaN	NaN	2507447.0	MBA2910720001	291072.0	291072	NaN	
7717	8.0	10.0	NaN	NaN	2507447.0	MBA2910720001	291072.0	291072	291072.0	
7718	9.0	10.0	NaN	NaN	2507447.0	MBA2910720001	291072.0	291072	291560.0	

	APGAR1	APGAR5	CODANOMAL	CODCART	CODESTAB	CODINST	CODMUNCART	CODMUNNASC	CODMUNNATU	CODM
1465	NaN	NaN	NaN	NaN	2507447.0	MBA2910720001	NaN	291072	291630.0	
1466	NaN	NaN	NaN	NaN	2507447.0	MBA2910720001	NaN	291072	292805.0	
1467	NaN	NaN	NaN	NaN	2507447.0	MBA2910720001	NaN	291072	291630.0	

20 rows × 73 columns

	APGAR1	APGAR5	CODANOMAL	CODCART	CODESTAB	CODINSTHash	CODMUNCART	CODMUNNASC
0	8.0	9.0	NaN	NaN	2507447.0	bdb169acdb1be9e450326225b2257af0dba966d4	NaN	291072
1	9.0	10.0	NaN	NaN	2802112.0	4cca40948e1d43d333e5b55c96f2683134ac0a32	NaN	291360
2	9.0	10.0	NaN	NaN	2506122.0	4724522994c8ed9ec9e26c4f4e71e47f9541d5b5	NaN	292200
3	10.0	10.0	NaN	NaN	2506122.0	4724522994c8ed9ec9e26c4f4e71e47f9541d5b5	NaN	292200
4	8.0	9.0	NaN	NaN	2498804.0	4724522994c8ed9ec9e26c4f4e71e47f9541d5b5	NaN	292200

5 rows × 71 columns

Drop NA?

In [0]:
```
d1 = d0[['APGAR1', 'APGAR5', 'CODANOMAL', 'CODCART', 'CODESTAB', 'CODINST',
       'CODMUNCART', 'CODMUNNASC', 'CODMUNNATU', 'CODMUNRES', 'CODOCUPMAE',
       'CODPAISRES', 'CODUFNATU', 'CONSPRENAT', 'CONSULTAS', 'DIFDATA',
       'DTCADASTRO', 'DTDECLARAC', 'DTNASC', 'DTNASCMAE', 'DTRECEBIM',
       'DTRECORIG', 'DTRECORIGA', 'DTREGCART', 'DTULTMENST', 'ESCMAE',
       'ESCMAE2010', 'ESCMAEAGR1', 'ESTCIVMAE', 'GESTACAO', 'GRAVIDEZ',
       'HORANASC', 'IDADEMAE', 'IDADEPAI', 'IDANOMAL', 'KOTELCHUCK', 'LOCNASC',
       'MESPRENAT', 'NATURALMAE', 'NUMERODN', 'NUMERODV', 'NUMEROLOTE',
       'NUMREGCART', 'ORIGEM', 'PARIDADE', 'PARTO', 'PESO', 'PREFIXODN',
       'QTDFILMORT', 'QTDFILVIVO', 'QTDGESTANT', 'QTDPARTCES', 'QTDPARTNOR',
       'RACACOR', 'RACACORMAE', 'RACACORN', 'SEMAGESTAC', 'SERIESCMAE', 'SEXO',
       'STCESPARTO', 'STDNEPIDEM', 'STDNNOVA', 'STTRABPART', 'TPAPRESENT',
       'TPDOCRESP', 'TPFUNCRESP', 'TPMETESTIM', 'TPNASCASSI', 'TPROBSON',
       'Unnamed: 0', 'VERSAOSIST']].dropna()
```

In [0]:
```
d1
```