S0 - HOPE extract data

November 15, 2020

0.1 Import data from DB.

[1]:

```
# pip install pymysql
from sqlalchemy import create_engine
import pymysql
import pandas as pd
import numpy as np
```

[2]:

```
dbConnectionURL = 'mysql+pymysql://root:hope@mysql-master/hope'
dbConnection = create_engine(dbConnectionURL)

df = pd.read_sql('SELECT id, pedido, respuesta FROM fed_hope_sugerencia',␣
 ↪con=dbConnection)
```

[3]:

```
df.head(10)
```

[3]:

```
   id                                            pedido  \
0  29  {"data":{"type":"emr--em","attributes":{"name"…
1  30  {"data":{"type":"emr--em","attributes":{"name"…
2  31  {"data":{"type":"emr--em","attributes":{"name"…
3  32  {"data":{"type":"emr--em","attributes":{"name"…
4  33  {"data":{"type":"emr--em","attributes":{"name"…
5  34  {"data":{"type":"emr--em","attributes":{"name"…
6  35  {"data":{"type":"emr--em","attributes":{"name"…
7  36  {"data":{"type":"emr--em","attributes":{"name"…
8  37  {"data":{"type":"emr--em","attributes":{"name"…
9  38  {"data":{"type":"emr--em","attributes":{"name"…

                                          respuesta
0  {"data":{"type":"emr--emr","id":"ef6d63fb-afe8…
1  {"data":{"type":"emr--emr","id":"0b8a1cc8-ce17…
2  {"data":{"type":"emr--emr","id":"25733e18-3245…
3  {"data":{"type":"emr--emr","id":"40320232-7510…
4  {"data":{"type":"emr--emr","id":"f686d89e-fc8e…
5  {"data":{"type":"emr--emr","id":"d94d7c78-9941…
6  {"data":{"type":"emr--emr","id":"0a14cc6b-af7b…
7  {"data":{"type":"emr--emr","id":"245bb87d-b52c…
8  {"data":{"type":"emr--emr","id":"fad05206-04f6…
```

```
9  {"data":{"type":"emr--emr","id":"a0f8dabe-a795…
```

0.2 Remplace ilegal charts in json

```
[4]: mapping = {'\r\n': '', '\n': '', '\t': '', r'\\.': '',}
     df = df.replace({'pedido': mapping, 'respuesta': mapping}, regex=True)
     df.head(10)
```

```
[4]:   id                                             pedido  \
     0  29  {"data":{"type":"emr--em","attributes":{"name"…
     1  30  {"data":{"type":"emr--em","attributes":{"name"…
     2  31  {"data":{"type":"emr--em","attributes":{"name"…
     3  32  {"data":{"type":"emr--em","attributes":{"name"…
     4  33  {"data":{"type":"emr--em","attributes":{"name"…
     5  34  {"data":{"type":"emr--em","attributes":{"name"…
     6  35  {"data":{"type":"emr--em","attributes":{"name"…
     7  36  {"data":{"type":"emr--em","attributes":{"name"…
     8  37  {"data":{"type":"emr--em","attributes":{"name"…
     9  38  {"data":{"type":"emr--em","attributes":{"name"…

                                           respuesta
     0  {"data":{"type":"emr--emr","id":"ef6d63fb-afe8…
     1  {"data":{"type":"emr--emr","id":"0b8a1cc8-ce17…
     2  {"data":{"type":"emr--emr","id":"25733e18-3245…
     3  {"data":{"type":"emr--emr","id":"40320232-7510…
     4  {"data":{"type":"emr--emr","id":"f686d89e-fc8e…
     5  {"data":{"type":"emr--emr","id":"d94d7c78-9941…
     6  {"data":{"type":"emr--emr","id":"0a14cc6b-af7b…
     7  {"data":{"type":"emr--emr","id":"245bb87d-b52c…
     8  {"data":{"type":"emr--emr","id":"fad05206-04f6…
     9  {"data":{"type":"emr--emr","id":"a0f8dabe-a795…
```

0.3 Flattening JSON

```
[5]: import ast
     import json
     from pandas import read_json, json_normalize #package for flattening json in␣
      ↪pandas df
     #https://stackoverflow.com/questions/39899005/
      ↪how-to-flatten-a-pandas-dataframe-with-some-columns-as-json

     pd.options.display.max_columns = None
```

```
[6]: # Flatterin column "Pedido"
```

```
pedidosData = json_normalize(df['pedido'].apply(json.loads).tolist()).
 ↪add_prefix('pedido.')

pedidosData
```

[6]:

```
    pedido.data.type pedido.data.attributes.name  \
0            emr--em                        None
1            emr--em                        None
2            emr--em                        None
3            emr--em                        None
4            emr--em                        None
..               …                          …
119          emr--em                        None
120          emr--em                        None
121          emr--em                        None
122          emr--em                        None
123          emr--em                        None

    pedido.data.attributes.affected_organ pedido.data.attributes.age  \
0                                                                 75
1                                                                 31
2                                                                 76
3                                                                 75
4                                                                 31
..                                    …                        …
119                                                               74
120                                                               48
121                                                               40
122                                                               43
123                                                               37

    pedido.data.attributes.diagnostic_main pedido.data.attributes.gender  \
0                       FISTULA PERITONEAL                          male
1               REHABILITACION NEUROLOGICA                          male
2                   INSUFICIENCIA CARDIACA                          male
3                       FISTULA PERITONEAL                          male
4               REHABILITACION NEUROLOGICA                          male
..                                     …                           …
119                DIFICULTAD RESPIRATORIA                          male
120             REHABILITACION NEUROLOGICA                          male
121             REHABILITACION NEUROLOGICA                          male
122                                    TEP                          male
123                        DOLOR ABDOMINAL                          male

                pedido.data.attributes.medical_history
0    Paciente de 75 au00f1os con antecedentes de ga…
1    Paciente estable clu00ednicamente, afebril. no…
```

```
2    Paciente de 76 au00f1os con antecednetes de ar…
3    Paciente de 75 au00f1os que cursa postoperator…
4    Paciente con lesion medular a nivel D12. Arrtr…
..                                                …
119  Paciente de 74 au00f1os con antecedentes hiper…
120  PACIENTE SE ENCUENTRA ESTABLE HEMODINAMICAMENT…
121  Paciente se encuentra cursando postquiru00farg…
122  Paciente de 43 au00f1os con antecedentes de co…
123  aciente de 36 au00f1os con antecedentes de ami…

[124 rows x 7 columns]
```

[7]:

```python
# Flatterin column "respuesta"

def get_articles_from_respuesta(ld):
    jsonData = json.loads(ld)
    pubmedKeys = jsonData['data']['attributes']['pubmed_mt_opt']
    if pubmedKeys is None : pubmedKeys = []

    articles = list(jsonData['data']['attributes']['pubmed']['articles'])
    articlesIDs = []
    articlesRevisedYear = []
    articlesRevisedMonth = []
    for article in articles:
        articlesIDs.append(article['id'])
        articlesRevisedYear.append(article['revisedDate']['Year'])
        articlesRevisedMonth.append(article['revisedDate']['Month'])


    return dict({
        'articles': articlesIDs,
        'articlesRevisedYear': articlesRevisedYear,
        'articlesRevisedMonth': articlesRevisedMonth,
        'pubmed_keys': ','.join(pubmedKeys)
    })


respuestaData = json_normalize(df['respuesta'].
 ↪apply(get_articles_from_respuesta).tolist()).add_prefix('respuesta.')

respuestaData
```

[7]:

```
                                    respuesta.articles  \
0    [27395425, 28560554, 28641726, 26245344, 28942…
1    [30210096, 27617939, 27210858, 26412482, 25487…
2    [21067951, 27616270, 27532500, 28426556, 27495…
3    [30179656, 28641726, 28694230, 27796647, 28867…
```

```
4    [29787536, 24840763, 28273653, 26836795, 26409…
..                                                 …
119  [28641726, 30179656, 28694230, 27796647, 28867…
120  [27128826, 30336861, 30226191, 29371130, 29587…
121  [30595510, 21554494, 26465238, 26875969, 30056…
122  [30081165, 30629460, 26220984, 25749853, 28545…
123  [30662053, 29879068, 26849395, 31061178, 31223…

                         respuesta.articlesRevisedYear  \
0    [2018, 2018, 2017, 2016, 2018, 2014, 2018, 201…
1    [2019, 2017, 2017, 2017, 2016, 2016, 2019, 201…
2    [2011, 2017, 2017, 2019, 2017, 2009, 2017, 201…
3    [2019, 2017, 2018, 2017, 2017, 2017, 2015, 201…
4    [2019, 2014, 2017, 2016, 2016, 2019, 2013, 201…
..                                                 …
119  [2017, 2019, 2018, 2017, 2017, 2017, 2017, 201…
120  [2017, 2019, 2019, 2018, 2018, 2019, 2019, 201…
121  [2019, 2012, 2016, 2016, 2019, 2019, 2016, 201…
122  [2018, 2019, 2016, 2016, 2018, 2020, 2015, 201…
123  [2019, 2018, 2016, 2019, 2019, 2019, 2018, 201…

               respuesta.articlesRevisedMonth  \
0    [01, 04, 12, 12, 06, 06, 09, 04, 01, 04]
1    [03, 04, 08, 06, 06, 09, 02, 03, 01, 11]
2    [03, 07, 05, 01, 07, 03, 06, 05, 03, 03]
3    [03, 12, 05, 02, 11, 05, 08, 06, 01, 08]
4    [05, 10, 11, 11, 05, 04, 03, 09, 02, 12]
..                                        …
119  [12, 03, 05, 02, 11, 05, 09, 09, 01, 03]
120  [04, 01, 09, 10, 08, 01, 06, 08, 11, 04]
121  [03, 04, 09, 08, 07, 10, 10, 05, 10, 06]
122  [12, 03, 06, 04, 12, 02, 02, 09, 05, 04]
123  [02, 06, 12, 12, 07, 03, 12, 05, 04, 04]

                                 respuesta.pubmed_keys
0    Intestines,Therapeutics,Catharsis,Wounds and I…
1    Back,Wounds and Injuries,Catheterization,Rest,…
2    Heart Murmurs,Intestines,Lactic Acid,Therapeut…
3    Intestines,Therapeutics,Catharsis,Lower Extrem…
4    Abdomen,Catheterization,Headache,Diuresis,Extr…
..                                                 …
119  Extremities,Catharsis,Tazobactam,Abdomen,Oxyge…
120                             Catharsis,Abdomen,Lung
121  Abdomen,Wounds and Injuries,Lung,Stroke,Aphasi…
122  Extremities,Catharsis,Thromboembolism,Foramen …
123  Fever,Catharsis,Infections,Abdominal Pain,Abdo…
```

```
[124 rows x 4 columns]
```

0.4 Reorder keys alphabetically

[8]:

```
def get_keys_ordered(keys):
    keysList = [x.strip() for x in keys.split(',')]
    keysList.sort()
    return ",".join(keysList)


respuestaData['respuesta.pubmed_keys'] = respuestaData['respuesta.pubmed_keys'].
 ↪apply(lambda x : get_keys_ordered(x))

respuestaData
```

[8]:

```
                                     respuesta.articles  \
0    [27395425, 28560554, 28641726, 26245344, 28942…
1    [30210096, 27617939, 27210858, 26412482, 25487…
2    [21067951, 27616270, 27532500, 28426556, 27495…
3    [30179656, 28641726, 28694230, 27796647, 28867…
4    [29787536, 24840763, 28273653, 26836795, 26409…
..                                                …
119  [28641726, 30179656, 28694230, 27796647, 28867…
120  [27128826, 30336861, 30226191, 29371130, 29587…
121  [30595510, 21554494, 26465238, 26875969, 30056…
122  [30081165, 30629460, 26220984, 25749853, 28545…
123  [30662053, 29879068, 26849395, 31061178, 31223…

                          respuesta.articlesRevisedYear  \
0    [2018, 2018, 2017, 2016, 2018, 2014, 2018, 201…
1    [2019, 2017, 2017, 2017, 2016, 2016, 2019, 201…
2    [2011, 2017, 2017, 2019, 2017, 2009, 2017, 201…
3    [2019, 2017, 2018, 2017, 2017, 2017, 2015, 201…
4    [2019, 2014, 2017, 2016, 2016, 2019, 2013, 201…
..                                                …
119  [2017, 2019, 2018, 2017, 2017, 2017, 2017, 201…
120  [2017, 2019, 2019, 2018, 2018, 2019, 2019, 201…
121  [2019, 2012, 2016, 2016, 2019, 2019, 2016, 201…
122  [2018, 2019, 2016, 2016, 2018, 2020, 2015, 201…
123  [2019, 2018, 2016, 2019, 2019, 2019, 2018, 201…

               respuesta.articlesRevisedMonth  \
0    [01, 04, 12, 12, 06, 06, 09, 04, 01, 04]
1    [03, 04, 08, 06, 06, 09, 02, 03, 01, 11]
2    [03, 07, 05, 01, 07, 03, 06, 05, 03, 03]
3    [03, 12, 05, 02, 11, 05, 08, 06, 01, 08]
4    [05, 10, 11, 11, 05, 04, 03, 09, 02, 12]
```

```
..                                       …
119  [12, 03, 05, 02, 11, 05, 09, 09, 01, 03]
120  [04, 01, 09, 10, 08, 01, 06, 08, 11, 04]
121  [03, 04, 09, 08, 07, 10, 10, 05, 10, 06]
122  [12, 03, 06, 04, 12, 02, 02, 09, 05, 04]
123  [02, 06, 12, 12, 07, 03, 12, 05, 04, 04]

                                  respuesta.pubmed_keys
0    Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…
1    Back,Catheterization,Dyshidrotic,Eczema,Micona…
2    Abdomen,Acidosis,Anemia,Arthritis,Arthritis,Be…
3    Abdomen,Blood Culture,Catharsis,Diuresis,Drug …
4    Abdomen,Catheterization,Diuresis,Extremities,H…
..                                                …
119  Abdomen,Catharsis,Cough,Diarrhea,Diuresis,Extr…
120                              Abdomen,Catharsis,Lung
121  Abdomen,Aphasia,Aphasia,Broca,Lung,Paresis,Sig…
122  Abdomen,Acenocoumarol,Arteries,Catharsis,Cathe…
123  Abdomen,Abdominal Pain,Analgesics,Catharsis,Cy…

[124 rows x 4 columns]
```

0.5 Join data

[9]:

```
dfCleaned = df[['id']].join([pedidosData,respuestaData])

dfCleaned.head(10)
```

[9]:

```
   id pedido.data.type pedido.data.attributes.name  \
0  29          emr--em                        None
1  30          emr--em                        None
2  31          emr--em                        None
3  32          emr--em                        None
4  33          emr--em                        None
5  34          emr--em                        None
6  35          emr--em                        None
7  36          emr--em                        None
8  37          emr--em                        None
9  38          emr--em                        None

  pedido.data.attributes.affected_organ pedido.data.attributes.age  \
0                                                                75
1                                                                31
2                                                                76
3                                                                75
4                                                                31
5                                                                52
```

```
6                                                               41
7                                                               52
8                                                               67
9                                                               71

  pedido.data.attributes.diagnostic_main pedido.data.attributes.gender  \
0                     FISTULA PERITONEAL                          male
1             REHABILITACION NEUROLOGICA                          male
2                 INSUFICIENCIA CARDIACA                          male
3                     FISTULA PERITONEAL                          male
4             REHABILITACION NEUROLOGICA                          male
5                        CEFALEA INTENSA                          male
6                                LEGRADO                          male
7                        CEFALEA INTENSA                          male
8             REHABILITACION NEUROLOGICA                          male
9             REHABILITACION NEUROLOGICA                          male

              pedido.data.attributes.medical_history  \
0  Paciente de 75 au00f1os con antecedentes de ga…
1  Paciente estable clu00ednicamente, afebril. no…
2  Paciente de 76 au00f1os con antecednetes de ar…
3  Paciente de 75 au00f1os que cursa postoperator…
4  Paciente con lesion medular a nivel D12. Arrtr…
5  Paciente de 52 au00f1os con antecedentes de me…
6  ANTECEDENTES:HIPERTENSION ARTERIAL HACE 15  EV…
7  Paciente con antecedentes de melanoma hace 4 a…
8  Paciente reingresa traido de CMIC tras recambi…
9  Paciente controlado en consultorio se observa …

                                  respuesta.articles  \
0  [27395425, 28560554, 28641726, 26245344, 28942…
1  [30210096, 27617939, 27210858, 26412482, 25487…
2  [21067951, 27616270, 27532500, 28426556, 27495…
3  [30179656, 28641726, 28694230, 27796647, 28867…
4  [29787536, 24840763, 28273653, 26836795, 26409…
5  [29618703, 29866680, 27209571, 25304079, 28474…
6  [26609031, 24002584, 20176351, 19521237, 30243…
7  [29618703, 24310475, 25304079, 28592387, 28474…
8  [28760799, 26099617, 22584817, 25559313, 27555…
9  [21665344, 29563376, 27881312, 26486203, 29621…

                       respuesta.articlesRevisedYear  \
0  [2018, 2018, 2017, 2016, 2018, 2014, 2018, 201…
1  [2019, 2017, 2017, 2017, 2016, 2016, 2019, 201…
2  [2011, 2017, 2017, 2019, 2017, 2009, 2017, 201…
3  [2019, 2017, 2018, 2017, 2017, 2017, 2015, 201…
4  [2019, 2014, 2017, 2016, 2016, 2019, 2013, 201…
```

```
5  [2019, 2018, 2017, 2015, 2018, 2018, 2014, 201…
6  [2016, 2014, 2010, 2009, 2019, 2019, 2019, 201…
7  [2019, 2014, 2015, 2017, 2018, 2016, 2010, 201…
8  [2018, 2016, 2013, 2015, 2017, 2011, 2014, 201…
9  [2011, 2018, 2018, 2017, 2019, 2018, 2013, 201…

             respuesta.articlesRevisedMonth  \
0  [01, 04, 12, 12, 06, 06, 09, 04, 01, 04]
1  [03, 04, 08, 06, 06, 09, 02, 03, 01, 11]
2  [03, 07, 05, 01, 07, 03, 06, 05, 03, 03]
3  [03, 12, 05, 02, 11, 05, 08, 06, 01, 08]
4  [05, 10, 11, 11, 05, 04, 03, 09, 02, 12]
5  [02, 11, 03, 09, 04, 12, 09, 03, 02, 09]
6  [11, 09, 08, 12, 02, 03, 02, 10, 02, 03]
7  [02, 07, 09, 09, 04, 03, 03, 09, 02, 12]
8  [04, 10, 06, 06, 08, 10, 09, 09, 01, 10]
9  [12, 04, 05, 05, 01, 10, 09, 08, 11, 04]

                               respuesta.pubmed_keys
0  Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…
1  Back,Catheterization,Dyshidrotic,Eczema,Micona…
2  Abdomen,Acidosis,Anemia,Arthritis,Arthritis,Be…
3  Abdomen,Blood Culture,Catharsis,Diuresis,Drug …
4  Abdomen,Catheterization,Diuresis,Extremities,H…
5  Abdomen,Anticonvulsants,Axilla,Biopsy,Extremit…
6  Cellulite,Cellulitis,Conjunctiva,Cysts,Edema,E…
7  Abdomen,Axilla,Catharsis,Conscience,Diuresis,E…
8  Abdomen,Central Nervous System,Heart Murmurs,H…
9               Cheek,Hematoma,Maxilla,Pain,Palpation
```

0.6 Expand Articles

[10]:

```python
dfCleaned = dfCleaned.explode('respuesta.articles').reset_index(drop=True)

# Expand Articles Year
dfCleanedArticlesRevisedYear = dfCleaned.explode('respuesta.
 ↪articlesRevisedYear').reset_index(drop=True)
dfCleaned['respuesta.articlesRevisedYear'] =␣
 ↪dfCleanedArticlesRevisedYear['respuesta.articlesRevisedYear']

# Expand Articles Month
dfCleanedArticlesRevisedMonth = dfCleaned.explode('respuesta.
 ↪articlesRevisedMonth').reset_index(drop=True)
dfCleaned['respuesta.articlesRevisedMonth'] =␣
 ↪dfCleanedArticlesRevisedMonth['respuesta.articlesRevisedMonth']
```

```
dfCleaned
```

[10]:

```
       id pedido.data.type pedido.data.attributes.name  \
0      29          emr--em                        None
1      29          emr--em                        None
2      29          emr--em                        None
3      29          emr--em                        None
4      29          emr--em                        None
…   …              …                         …
1235  152          emr--em                        None
1236  152          emr--em                        None
1237  152          emr--em                        None
1238  152          emr--em                        None
1239  152          emr--em                        None

     pedido.data.attributes.affected_organ pedido.data.attributes.age  \
0                                                                  75
1                                                                  75
2                                                                  75
3                                                                  75
4                                                                  75
…                                    …                        …
1235                                                               37
1236                                                               37
1237                                                               37
1238                                                               37
1239                                                               37

     pedido.data.attributes.diagnostic_main pedido.data.attributes.gender  \
0                        FISTULA PERITONEAL                          male
1                        FISTULA PERITONEAL                          male
2                        FISTULA PERITONEAL                          male
3                        FISTULA PERITONEAL                          male
4                        FISTULA PERITONEAL                          male
…                                     …                            …
1235                        DOLOR ABDOMINAL                          male
1236                        DOLOR ABDOMINAL                          male
1237                        DOLOR ABDOMINAL                          male
1238                        DOLOR ABDOMINAL                          male
1239                        DOLOR ABDOMINAL                          male

                 pedido.data.attributes.medical_history respuesta.articles  \
0     Paciente de 75 au00f1os con antecedentes de ga…          27395425
1     Paciente de 75 au00f1os con antecedentes de ga…          28560554
2     Paciente de 75 au00f1os con antecedentes de ga…          28641726
3     Paciente de 75 au00f1os con antecedentes de ga…          26245344
```

```
4     Paciente de 75 au00f1os con antecedentes de ga…          28942543
…                                                    …                …
1235  aciente de 36 au00f1os con antecedentes de ami…          29975804
1236  aciente de 36 au00f1os con antecedentes de ami…          26362243
1237  aciente de 36 au00f1os con antecedentes de ami…          30711130
1238  aciente de 36 au00f1os con antecedentes de ami…          27932159
1239  aciente de 36 au00f1os con antecedentes de ami…          19104113

     respuesta.articlesRevisedYear respuesta.articlesRevisedMonth  \
0                             2018                             01
1                             2018                             04
2                             2017                             12
3                             2016                             12
4                             2018                             06
…                              …                            …
1235                          2016                             01
1236                          2019                             04
1237                          2015                             10
1238                          2019                             05
1239                          2015                             02

                                    respuesta.pubmed_keys
0     Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…
1     Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…
2     Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…
3     Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…
4     Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…
…                                                    …
1235  Abdomen,Abdominal Pain,Analgesics,Catharsis,Cy…
1236  Abdomen,Abdominal Pain,Analgesics,Catharsis,Cy…
1237  Abdomen,Abdominal Pain,Analgesics,Catharsis,Cy…
1238  Abdomen,Abdominal Pain,Analgesics,Catharsis,Cy…
1239  Abdomen,Abdominal Pain,Analgesics,Catharsis,Cy…

[1240 rows x 12 columns]
```

0.7 Create column with join of id and article id to join with feedback

[11]:
```
dfCleaned['sugerencia_id_+_articles_id'] = dfCleaned['id'].astype(str) + "_" +␣
↪dfCleaned['respuesta.articles'].astype(str)

dfCleaned
```

[11]:
```
        id pedido.data.type pedido.data.attributes.name  \
0       29         emr--em                        None
1       29         emr--em                        None
2       29         emr--em                        None
```

```
3      29          emr--em                        None
4      29          emr--em                        None
…   …              …                         …
1235  152          emr--em                        None
1236  152          emr--em                        None
1237  152          emr--em                        None
1238  152          emr--em                        None
1239  152          emr--em                        None

     pedido.data.attributes.affected_organ pedido.data.attributes.age  \
0                                                                   75
1                                                                   75
2                                                                   75
3                                                                   75
4                                                                   75
…                                     …                        …
1235                                                                37
1236                                                                37
1237                                                                37
1238                                                                37
1239                                                                37

     pedido.data.attributes.diagnostic_main pedido.data.attributes.gender  \
0                        FISTULA PERITONEAL                          male
1                        FISTULA PERITONEAL                          male
2                        FISTULA PERITONEAL                          male
3                        FISTULA PERITONEAL                          male
4                        FISTULA PERITONEAL                          male
…                                      …                            …
1235                        DOLOR ABDOMINAL                          male
1236                        DOLOR ABDOMINAL                          male
1237                        DOLOR ABDOMINAL                          male
1238                        DOLOR ABDOMINAL                          male
1239                        DOLOR ABDOMINAL                          male

                 pedido.data.attributes.medical_history respuesta.articles  \
0     Paciente de 75 au00f1os con antecedentes de ga…          27395425
1     Paciente de 75 au00f1os con antecedentes de ga…          28560554
2     Paciente de 75 au00f1os con antecedentes de ga…          28641726
3     Paciente de 75 au00f1os con antecedentes de ga…          26245344
4     Paciente de 75 au00f1os con antecedentes de ga…          28942543
…                                                 …                …
1235  aciente de 36 au00f1os con antecedentes de ami…          29975804
1236  aciente de 36 au00f1os con antecedentes de ami…          26362243
1237  aciente de 36 au00f1os con antecedentes de ami…          30711130
1238  aciente de 36 au00f1os con antecedentes de ami…          27932159
1239  aciente de 36 au00f1os con antecedentes de ami…          19104113
```

```
     respuesta.articlesRevisedYear respuesta.articlesRevisedMonth  \
0                             2018                             01
1                             2018                             04
2                             2017                             12
3                             2016                             12
4                             2018                             06
…                            …                            …
1235                          2016                             01
1236                          2019                             04
1237                          2015                             10
1238                          2019                             05
1239                          2015                             02

                                  respuesta.pubmed_keys  \
0     Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…
1     Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…
2     Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…
3     Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…
4     Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…
…                                                …
1235  Abdomen,Abdominal Pain,Analgesics,Catharsis,Cy…
1236  Abdomen,Abdominal Pain,Analgesics,Catharsis,Cy…
1237  Abdomen,Abdominal Pain,Analgesics,Catharsis,Cy…
1238  Abdomen,Abdominal Pain,Analgesics,Catharsis,Cy…
1239  Abdomen,Abdominal Pain,Analgesics,Catharsis,Cy…

     sugerencia_id_+_articles_id
0                    29_27395425
1                    29_28560554
2                    29_28641726
3                    29_26245344
4                    29_28942543
…                          …
1235                152_29975804
1236                152_26362243
1237                152_30711130
1238                152_27932159
1239                152_19104113

[1240 rows x 13 columns]
```

0.8 Remove unnecessary columns

[12]:

```
dfCleaned = dfCleaned.drop([
    'pedido.data.type',
    'pedido.data.attributes.name',
    'pedido.data.attributes.affected_organ',
    'pedido.data.attributes.medical_history'
], axis=1)

dfCleaned
```

[12]:

```
       id pedido.data.attributes.age pedido.data.attributes.diagnostic_main  \
0      29                         75                     FISTULA PERITONEAL
1      29                         75                     FISTULA PERITONEAL
2      29                         75                     FISTULA PERITONEAL
3      29                         75                     FISTULA PERITONEAL
4      29                         75                     FISTULA PERITONEAL
…   …                        …                                    …
1235  152                         37                        DOLOR ABDOMINAL
1236  152                         37                        DOLOR ABDOMINAL
1237  152                         37                        DOLOR ABDOMINAL
1238  152                         37                        DOLOR ABDOMINAL
1239  152                         37                        DOLOR ABDOMINAL

     pedido.data.attributes.gender respuesta.articles  \
0                             male           27395425
1                             male           28560554
2                             male           28641726
3                             male           26245344
4                             male           28942543
…                            …                …
1235                          male           29975804
1236                          male           26362243
1237                          male           30711130
1238                          male           27932159
1239                          male           19104113

     respuesta.articlesRevisedYear respuesta.articlesRevisedMonth  \
0                             2018                             01
1                             2018                             04
2                             2017                             12
3                             2016                             12
4                             2018                             06
…                            …                            …
1235                          2016                             01
1236                          2019                             04
1237                          2015                             10
```

```
1238                              2019                                 05
1239                              2015                                 02

                                            respuesta.pubmed_keys  \
0     Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…
1     Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…
2     Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…
3     Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…
4     Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…
…                                                  …
1235  Abdomen,Abdominal Pain,Analgesics,Catharsis,Cy…
1236  Abdomen,Abdominal Pain,Analgesics,Catharsis,Cy…
1237  Abdomen,Abdominal Pain,Analgesics,Catharsis,Cy…
1238  Abdomen,Abdominal Pain,Analgesics,Catharsis,Cy…
1239  Abdomen,Abdominal Pain,Analgesics,Catharsis,Cy…

     sugerencia_id_+_articles_id
0                    29_27395425
1                    29_28560554
2                    29_28641726
3                    29_26245344
4                    29_28942543
…                            …
1235                152_29975804
1236                152_26362243
1237                152_30711130
1238                152_27932159
1239                152_19104113

[1240 rows x 9 columns]
```

0.9 Get feedback

[13]:
```
dfFeedback = pd.read_sql('SELECT fed_hope_sugerencia_id as id, articulo,␣
↪utilidad, concat(fed_hope_sugerencia_id, "_", articulo) as␣
↪"sugerencia_id_+_articles_id" FROM fed_hope_sugerencia_feedback',␣
↪con=dbConnection)

dfFeedback
```

[13]:
```
     id  articulo  utilidad sugerencia_id_+_articles_id
0    29  27395425         1                 29_27395425
1    32  28694230         1                 32_28694230
2    52  28805236         0                 52_28805236
3    58  27537587         0                 58_27537587
4    58  28148670         1                 58_28148670
5    59  25055513         1                 59_25055513
```

6	59	29279563	0	59_29279563
7	59	29279563	0	59_29279563
8	59	28065368	1	59_28065368
9	60	30762794	1	60_30762794
10	60	30762794	1	60_30762794
11	60	23424242	1	60_23424242
12	60	29587951	0	60_29587951
13	67	28720605	0	67_28720605
14	67	28956549	1	67_28956549
15	74	26362243	0	74_26362243
16	74	26362243	0	74_26362243
17	75	28552471	1	75_28552471
18	76	30711130	1	76_30711130
19	76	30662053	1	76_30662053
20	76	28294508	0	76_28294508
21	77	28429658	0	77_28429658
22	77	29600476	1	77_29600476
23	77	30910320	1	77_30910320
24	78	25989443	1	78_25989443
25	78	26429116	1	78_26429116
26	79	25904281	1	79_25904281
27	79	24526429	1	79_24526429
28	79	28188621	0	79_28188621
29	83	21055785	0	83_21055785
30	84	30182627	1	84_30182627
31	84	29196305	1	84_29196305
32	85	23850836	1	85_23850836
33	87	28956549	1	87_28956549
34	87	25070613	1	87_25070613
35	89	27480349	1	89_27480349
36	89	29937071	1	89_29937071
37	89	28270645	0	89_28270645
38	90	30413186	0	90_30413186
39	90	21598204	1	90_21598204
40	94	29975804	0	94_29975804
41	94	23761322	0	94_23761322
42	94	30413186	0	94_30413186
43	95	30567250	0	95_30567250
44	99	28803258	0	99_28803258
45	102	23587563	1	102_23587563
46	102	28516316	0	102_28516316
47	103	24079694	0	103_24079694
48	124	31515299	1	124_31515299
49	125	31515299	1	125_31515299
50	129	28251285	1	129_28251285

[14]:
```
dfJoined = pd.merge(dfCleaned, dfFeedback, how='left',␣
 ↪on=['sugerencia_id_+_articles_id'])

dfJoined
```

[14]:
```
      id_x pedido.data.attributes.age pedido.data.attributes.diagnostic_main  \
0       29                         75                     FISTULA PERITONEAL
1       29                         75                     FISTULA PERITONEAL
2       29                         75                     FISTULA PERITONEAL
3       29                         75                     FISTULA PERITONEAL
4       29                         75                     FISTULA PERITONEAL
…    …                        …                                     …
1238   152                         37                        DOLOR ABDOMINAL
1239   152                         37                        DOLOR ABDOMINAL
1240   152                         37                        DOLOR ABDOMINAL
1241   152                         37                        DOLOR ABDOMINAL
1242   152                         37                        DOLOR ABDOMINAL

     pedido.data.attributes.gender respuesta.articles  \
0                             male           27395425
1                             male           28560554
2                             male           28641726
3                             male           26245344
4                             male           28942543
…                            …                …
1238                          male           29975804
1239                          male           26362243
1240                          male           30711130
1241                          male           27932159
1242                          male           19104113

     respuesta.articlesRevisedYear respuesta.articlesRevisedMonth  \
0                             2018                             01
1                             2018                             04
2                             2017                             12
3                             2016                             12
4                             2018                             06
…                            …                            …
1238                          2016                             01
1239                          2019                             04
1240                          2015                             10
1241                          2019                             05
1242                          2015                             02

                                  respuesta.pubmed_keys  \
0     Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…
1     Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…
```

```
2     Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…
3     Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…
4     Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…
…                                                  …
1238  Abdomen,Abdominal Pain,Analgesics,Catharsis,Cy…
1239  Abdomen,Abdominal Pain,Analgesics,Catharsis,Cy…
1240  Abdomen,Abdominal Pain,Analgesics,Catharsis,Cy…
1241  Abdomen,Abdominal Pain,Analgesics,Catharsis,Cy…
1242  Abdomen,Abdominal Pain,Analgesics,Catharsis,Cy…

     sugerencia_id_+_articles_id  id_y  articulo  utilidad
0                    29_27395425  29.0  27395425       1.0
1                    29_28560554   NaN       NaN       NaN
2                    29_28641726   NaN       NaN       NaN
3                    29_26245344   NaN       NaN       NaN
4                    29_28942543   NaN       NaN       NaN
…                          …   …       …       …
1238                152_29975804   NaN       NaN       NaN
1239                152_26362243   NaN       NaN       NaN
1240                152_30711130   NaN       NaN       NaN
1241                152_27932159   NaN       NaN       NaN
1242                152_19104113   NaN       NaN       NaN

[1243 rows x 12 columns]
```

0.10 Remove nans and unnecessary columns

[15]:
```
dfJoined = dfJoined.drop([
    'id_x',
    'id_y',
    'articulo',
    'sugerencia_id_+_articles_id'
], axis=1)

# Rename column respuesta.articles to preserve back compatibility
dfJoined.rename(columns={'respuesta.articles':'articulo'}, inplace=True)

print(dfJoined.shape[0])
```

```
1243
```

[16]:
```
dfJoined.head(10)
```

[16]:
```
  pedido.data.attributes.age pedido.data.attributes.diagnostic_main  \
0                         75                     FISTULA PERITONEAL
1                         75                     FISTULA PERITONEAL
2                         75                     FISTULA PERITONEAL
```

```
3                         75                     FISTULA PERITONEAL
4                         75                     FISTULA PERITONEAL
5                         75                     FISTULA PERITONEAL
6                         75                     FISTULA PERITONEAL
7                         75                     FISTULA PERITONEAL
8                         75                     FISTULA PERITONEAL
9                         75                     FISTULA PERITONEAL

  pedido.data.attributes.gender  articulo respuesta.articlesRevisedYear  \
0                          male  27395425                          2018
1                          male  28560554                          2018
2                          male  28641726                          2017
3                          male  26245344                          2016
4                          male  28942543                          2018
5                          male  24782153                          2014
6                          male  28002229                          2018
7                          male  27505109                          2017
8                          male  24850546                          2015
9                          male  29371050                          2019

  respuesta.articlesRevisedMonth  \
0                             01
1                             04
2                             12
3                             12
4                             06
5                             06
6                             09
7                             04
8                             01
9                             04

                               respuesta.pubmed_keys  utilidad
0  Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…       1.0
1  Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…       NaN
2  Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…       NaN
3  Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…       NaN
4  Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…       NaN
5  Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…       NaN
6  Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…       NaN
7  Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…       NaN
8  Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…       NaN
9  Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…       NaN
```

0.11 Save data to csv

```
[17]: dfJoined.to_csv('hope_dataset_cleaned.csv',index=False)
```

```
[ ]:
```