In [64]:

```
import pandas as pd

df=pd.read_fwf(r'C:/Users/cast800390/Documents/IA_AP_202110/proyecto/secop/secop_train.csv',skiprows=0)
     #          ,names=['categorical_label','Z','ID','A','B','C','proyecto','D','valor'])
print(type(data))
print(df)



#category_labels = list(df['categorical_label'].value_counts().index)
category_labels = list(df[0])

#print(category_labels)
#df.columns.tolist()


#df=pd.read_csv(r'C:/Users/cast800390/Documents/IA_AP_202110/proyecto/secop/secop_train.csv')
```

```
<class 'pandas.core.frame.DataFrame'>
             categorical_label;;ID;;;;proyecto;;valor;        Unnamed: 1  \
0    0;225;TMSA-CD-1144-2021;Contratos y convenios ...               NaN
1    0;227;SED_CDM_026_2021;Régimen Especial;Convoc...               NaN
2    0;228;ITI - 030 -2021;Régimen Especial;Convoca...  contratista debe
3    0;230;Seleccionabreviada_35_2018;Selección Abr...  QUE CONFORMAN LA
4    0;231;CMZOA-23-2021;Contratación Mínima Cuantí...               NaN
..                                                 ...               ...
235  9;18;58-2021;Régimen Especial;Convocado;BOGOTÁ...               NaN
236  9;24;35;Régimen Especial;Convocado;BOGOTÁ D.C....               NaN
237  9;5;CE-08-21;Régimen Especial;Convocado;BOGOTÁ...               NaN
238  9;7;CE-015-21;Régimen Especial;Convocado;BOGOT...               NaN
239  9;9;CE-016-21;Régimen Especial;Convocado;BOGOT...               NaN

             Unnamed: 2                        Unnamed: 3  \
0                   NaN                               NaN
1                   NaN                               NaN
2    garantizar que los  espacios intervenidos queden sin
3    JURISDICCIÓN DE LA  CORPORACIÓN AUTÓNOMA REGIONAL DE
4                   NaN                               NaN
..                  ...                               ...
235                 NaN                               NaN
236                 NaN                               NaN
237                 NaN                               NaN
238                 NaN                               NaN
239                 NaN                               NaN

                          Unnamed: 4 Unnamed: 5 Unnamed: 6  \
0                                NaN        NaN        NaN
1                                NaN        NaN        NaN
2    filtraciones, ni goteras.;Bogotá     D.C. :     Bogotá
3    CUNDINAMARCA – CAR;Bogotá D.C. :     Bogotá     D.C.;$
4                                NaN        NaN        NaN
..                               ...        ...        ...
```

```
235                             NaN        NaN        NaN
236                             NaN        NaN        NaN
237                             NaN        NaN        NaN
238                             NaN        NaN        NaN
239                             NaN        NaN        NaN

                        Unnamed: 7
0                              NaN
1                              NaN
2    D.C.;$ 17.556.000,00;12-nov-21
3         389.997.387,00;12/11/2021
4                              NaN
..                             ...
235                            NaN
236                            NaN
237                            NaN
238                            NaN
239                            NaN

[240 rows x 8 columns]
---------------------------------------------------------------------------
KeyError                                  Traceback (most recent call last)
~\Anaconda3\lib\site-packages\pandas\core\indexes\base.py in get_loc(self, key, method, tolerance)
   3079             try:
-> 3080                 return self._engine.get_loc(casted_key)
   3081             except KeyError as err:

pandas\_libs\index.pyx in pandas._libs.index.IndexEngine.get_loc()

pandas\_libs\index.pyx in pandas._libs.index.IndexEngine.get_loc()

pandas\_libs\hashtable_class_helper.pxi in pandas._libs.hashtable.PyObjectHashTable.get_item()

pandas\_libs\hashtable_class_helper.pxi in pandas._libs.hashtable.PyObjectHashTable.get_item()

KeyError: 0

The above exception was the direct cause of the following exception:

KeyError                                  Traceback (most recent call last)
<ipython-input-64-00c6a91233e0> in <module>
      9
     10 #category_labels = list(df['categorical_label'].value_counts().index)
---> 11 category_labels = list(df[0])
     12
     13 #print(category_labels)

~\Anaconda3\lib\site-packages\pandas\core\frame.py in __getitem__(self, key)
   3022             if self.columns.nlevels > 1:
   3023                 return self._getitem_multilevel(key)
-> 3024             indexer = self.columns.get_loc(key)
   3025             if is_integer(indexer):
   3026                 indexer = [indexer]

~\Anaconda3\lib\site-packages\pandas\core\indexes\base.py in get_loc(self, key, method, tolerance)
```

```
   3080                return self._engine.get_loc(casted_key)
   3081            except KeyError as err:
-> 3082                raise KeyError(key) from err
   3083
   3084        if tolerance is not None:
```

KeyError: 0

In [67]:

```
df=pd.read_csv(r'C:/Users/cast800390/Documents/IA_AP_202110/proyecto/secop/secop_train.csv', sep=';')
print(df)
```

```
     categorical_label  Unnamed: 1                          ID  \
0                    0         225           TMSA-CD-1144-2021
1                    0         227            SED_CDM_026_2021
2                    0         228             ITI - 030 -2021
3                    0         230  Seleccionabreviada_35_2018
4                    0         231               CMZOA-23-2021
..                 ...         ...                         ...
235                  9          18                     58-2021
236                  9          24                          35
237                  9           5                   CE-08-21
238                  9           7                   CE-015-21
239                  9           9                   CE-016-21

                                        Unnamed: 3 Unnamed: 4  \
0          Contratos y convenios con más de dos partes  Convocado
1                                  Régimen Especial  Convocado
2                                  Régimen Especial  Convocado
3    Selección Abreviada de Menor Cuantía (Ley 1150...  Convocado
4                        Contratación Mínima Cuantía  Convocado
..                                               ...        ...
235                                Régimen Especial  Convocado
236                                Régimen Especial  Convocado
237                                Régimen Especial  Convocado
238                                Régimen Especial  Convocado
239                                Régimen Especial  Convocado

                                        Unnamed: 5  \
0                        BOGOTÁ D.C. - TRANSMILENIO
1                   BOGOTÁ D.C. - IED. DIVINO MAESTRO
2      BOGOTÁ D.C. - IED. ITI FRANCISCO JOSE DE CALDAS
3    CAR - CORPORACIÓN AUTÓNOMA REGIONAL DE CUNDINA...
4     BOGOTÁ D.C. - COLEGIO MANUEL ZAPATA OLIVELLA IED
..                                                 ...
235    BOGOTÁ D.C. - IED. INEM FCO. DE PAULA SANTANDER
236   BOGOTÁ D.C. - IED. EL SALITRE SAN CARLOS DE SUBA
237                          BOGOTÁ D.C. - IED. ESPAÑA
238                          BOGOTÁ D.C. - IED. ESPAÑA
239                          BOGOTÁ D.C. - IED. ESPAÑA

                                             proyecto  \
0    AUNAR ESFUERZOS PARA EL MEJORAMIENTO, ADECUACI...
1    MANTENIMIENTO GENERAL DE LAS INSTALACIONES DEL...
2    Mantenimiento integral de cubiertas, canales y...
```

```
3    PRESTACIÓN DEL SERVICIO DE MANTENIMIENTO PREVE...
4    EL COLEGIO Manuel Zapata Olivella IED REQUIERE...
..                                                 ...
235  suministro y cambio de tejas arreglode filtrac...
236  TRABAJOS DE RESANE PARTE INTERNA Y PINTURA PAR...
237  CONTRATACIN TRABAJOS DE MANTENIMIENTO A TODO C...
238  Contrato a todo costo incluye mano de obra y m...
239  TRABAJOS VARIOS A TODO COSTO INCLUYE MANO DE O...

                   Unnamed: 7             valor  Unnamed: 9
0    Bogotá D.C. : Bogotá D.C.           $ 0,00  12/11/2021
1    Bogotá D.C. : Bogotá D.C.   $ 17.757.000,00  12/11/2021
2    Bogotá D.C. : Bogotá D.C.   $ 17.556.000,00   12-nov-21
3    Bogotá D.C. : Bogotá D.C.  $ 389.997.387,00  12/11/2021
4    Bogotá D.C. : Bogotá D.C.   $ 28.900.000,00  12/11/2021
..                         ...               ...         ...
235  Bogotá D.C. : Bogotá D.C.   $ 17.872.000,00  25/11/2021
236  Bogotá D.C. : Bogotá D.C.   $ 17.443.200,00  25/11/2021
237  Bogotá D.C. : Bogotá D.C.    $ 5.065.000,00  27/11/2021
238  Bogotá D.C. : Bogotá D.C.    $ 5.780.800,00  27/11/2021
239  Bogotá D.C. : Bogotá D.C.    $ 6.200.000,00  27/11/2021

[240 rows x 10 columns]
```

In [75]:

```python
category_labels = list(df['categorical_label'])
print(category_labels)
```

```
[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3,
3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 6, 6, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8,
8, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9]
```

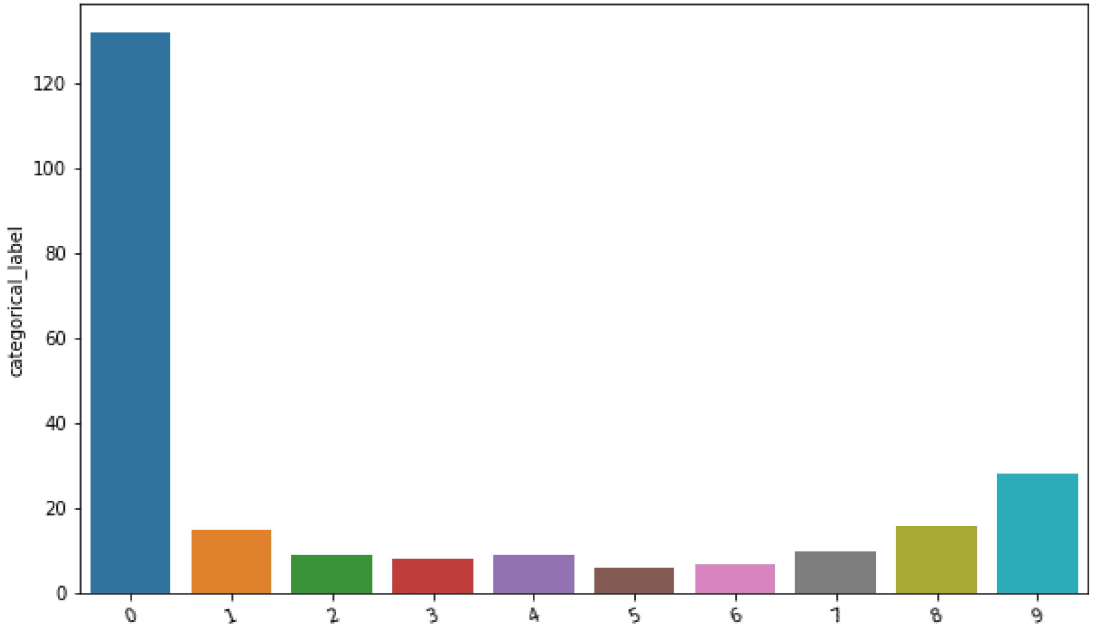
In [77]:

```python
import matplotlib.pyplot as plt
import seaborn as sns



fig = plt.figure(figsize=(10,6))
sns.barplot(df['categorical_label'].value_counts().index, df['categorical_label'].value_counts())
plt.xticks(rotation=20)
plt.show()
```

```
C:\Users\cast800390\Anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variables as keyword args: x, y. From version 0.1
2, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```



In [98]:

```
category_labels = list(df['categorical_label'].value_counts().index)
```

In [100…

```
#from sklearn.feature_extraction.stop_words import ENGLISH_STOP_WORDS
#from spacy.lang.en import English
import nltk  #**20211210
nltk.download('stopwords')  #**20211210
from nltk.corpus import stopwords
#from sklearn.feature_extraction.stop_words import SPANISH_STOP_WORDS #**
import spacy
#from spacy.lang.es #**
import string
from collections import Counter


#parser = Spanish()
nlp = spacy.load('es_core_news_md')
punctuations = string.punctuation

STOPLIST = set(stopwords.words('spanish'))# + list(SPANISH_STOP_WORDS))
SYMBOLS = " ".join(string.punctuation).split(" ") + ["'",":",";","-", "…","(",")",",","."]

# preprocesado de cada texto-proyecto
def cleanup_text(docs, logging=False):
  texts = []
  counter = 1
  for doc in docs:
```

```
    if counter % 1000 == 0 and logging:

      print("procesa %d desde %d proyectos." % (counter, len(docs)))
      counter += 1

     # doc = nlp(doc, disable=['parser', 'ner']) // # tokens = [tok.lemma_.lower().strip() for tok in doc if tok.lemma_ != '-PRON-']  // # tokens = [tok for t
      tokens = [tok for tok in tokens if tok not in SYMBOLS]
      tokens = ' '.join(tokens)
      texts.append(tokens)

  return pd.Series(texts)

#por cada categoría palabras comunes
def find_common_words_by_category(categories, N):
  for category in categories:

    category_text = [proyecto for proyecto in df[df['categorical_label'] == category]['proyecto']]
    cleanup_category_text = cleanup_text(category_text)
    cleanup_category_text = ' '.join(cleanup_category_text).split()
    category_counter = Counter(cleanup_category_text)

    common_words_by_category = [word[0] for word in category_counter.most_common(N)]
    word_count_by_category = [word[1] for word in category_counter.most_common(N)]

    word_statement = f"{category} has {word_count_by_category} words : {common_words_by_category}"
  #return word_statement --no aplica return
  print(word_statement)
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\cast800390\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping corpora\stopwords.zip.
```

In [ ]:

In [ ]: