7506R_TP2_GRUPO16_ENTREGA_PLN

December 8, 2022

[23]:
```
!pip install dtreeviz
!pip install pyreadstat
!pip install visualkeras
!pip install keras_tuner
!pip install gdown
!pip install matplotlib==3.1.1
!pip install $(spacy info es_core_news_sm --url)
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-
wheels/public/simple/
Requirement already satisfied: dtreeviz in /usr/local/lib/python3.8/dist-
packages (1.4.1)
Requirement already satisfied: graphviz>=0.9 in /usr/local/lib/python3.8/dist-
packages (from dtreeviz) (0.10.1)
Requirement already satisfied: matplotlib in /usr/local/lib/python3.8/dist-
packages (from dtreeviz) (3.1.1)
Requirement already satisfied: pytest in /usr/local/lib/python3.8/dist-packages
(from dtreeviz) (3.6.4)
Requirement already satisfied: pandas in /usr/local/lib/python3.8/dist-packages
(from dtreeviz) (1.3.5)
Requirement already satisfied: colour in /usr/local/lib/python3.8/dist-packages
(from dtreeviz) (0.1.5)
Requirement already satisfied: numpy in /usr/local/lib/python3.8/dist-packages
(from dtreeviz) (1.21.6)
Requirement already satisfied: scikit-learn in /usr/local/lib/python3.8/dist-
packages (from dtreeviz) (1.0.2)
Requirement already satisfied: python-dateutil>=2.1 in
/usr/local/lib/python3.8/dist-packages (from matplotlib->dtreeviz) (2.8.2)
Requirement already satisfied: kiwisolver>=1.0.1 in
/usr/local/lib/python3.8/dist-packages (from matplotlib->dtreeviz) (1.4.4)
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.8/dist-
packages (from matplotlib->dtreeviz) (0.11.0)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.1 in
/usr/local/lib/python3.8/dist-packages (from matplotlib->dtreeviz) (3.0.9)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.8/dist-
packages (from python-dateutil>=2.1->matplotlib->dtreeviz) (1.15.0)
Requirement already satisfied: pytz>=2017.3 in /usr/local/lib/python3.8/dist-
packages (from pandas->dtreeviz) (2022.6)
```

```
Requirement already satisfied: setuptools in /usr/local/lib/python3.8/dist-
packages (from pytest->dtreeviz) (57.4.0)
Requirement already satisfied: attrs>=17.4.0 in /usr/local/lib/python3.8/dist-
packages (from pytest->dtreeviz) (22.1.0)
Requirement already satisfied: more-itertools>=4.0.0 in
/usr/local/lib/python3.8/dist-packages (from pytest->dtreeviz) (9.0.0)
Requirement already satisfied: py>=1.5.0 in /usr/local/lib/python3.8/dist-
packages (from pytest->dtreeviz) (1.11.0)
Requirement already satisfied: pluggy<0.8,>=0.5 in
/usr/local/lib/python3.8/dist-packages (from pytest->dtreeviz) (0.7.1)
Requirement already satisfied: atomicwrites>=1.0 in
/usr/local/lib/python3.8/dist-packages (from pytest->dtreeviz) (1.4.1)
Requirement already satisfied: joblib>=0.11 in /usr/local/lib/python3.8/dist-
packages (from scikit-learn->dtreeviz) (1.2.0)
Requirement already satisfied: scipy>=1.1.0 in /usr/local/lib/python3.8/dist-
packages (from scikit-learn->dtreeviz) (1.7.3)
Requirement already satisfied: threadpoolctl>=2.0.0 in
/usr/local/lib/python3.8/dist-packages (from scikit-learn->dtreeviz) (3.1.0)
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-
wheels/public/simple/
Requirement already satisfied: pyreadstat in /usr/local/lib/python3.8/dist-
packages (1.2.0)
Requirement already satisfied: pandas>=1.2.0 in /usr/local/lib/python3.8/dist-
packages (from pyreadstat) (1.3.5)
Requirement already satisfied: pytz>=2017.3 in /usr/local/lib/python3.8/dist-
packages (from pandas>=1.2.0->pyreadstat) (2022.6)
Requirement already satisfied: numpy>=1.17.3 in /usr/local/lib/python3.8/dist-
packages (from pandas>=1.2.0->pyreadstat) (1.21.6)
Requirement already satisfied: python-dateutil>=2.7.3 in
/usr/local/lib/python3.8/dist-packages (from pandas>=1.2.0->pyreadstat) (2.8.2)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.8/dist-
packages (from python-dateutil>=2.7.3->pandas>=1.2.0->pyreadstat) (1.15.0)
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-
wheels/public/simple/
Requirement already satisfied: visualkeras in /usr/local/lib/python3.8/dist-
packages (0.0.2)
Requirement already satisfied: numpy>=1.18.1 in /usr/local/lib/python3.8/dist-
packages (from visualkeras) (1.21.6)
Requirement already satisfied: aggdraw>=1.3.11 in /usr/local/lib/python3.8/dist-
packages (from visualkeras) (1.3.15)
Requirement already satisfied: pillow>=6.2.0 in /usr/local/lib/python3.8/dist-
packages (from visualkeras) (7.1.2)
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-
wheels/public/simple/
Requirement already satisfied: keras_tuner in /usr/local/lib/python3.8/dist-
packages (1.1.3)
Requirement already satisfied: packaging in /usr/local/lib/python3.8/dist-
packages (from keras_tuner) (21.3)
```

```
Requirement already satisfied: kt-legacy in /usr/local/lib/python3.8/dist-
packages (from keras_tuner) (1.0.4)
Requirement already satisfied: requests in /usr/local/lib/python3.8/dist-
packages (from keras_tuner) (2.23.0)
Requirement already satisfied: numpy in /usr/local/lib/python3.8/dist-packages
(from keras_tuner) (1.21.6)
Requirement already satisfied: ipython in /usr/local/lib/python3.8/dist-packages
(from keras_tuner) (7.9.0)
Requirement already satisfied: tensorboard in /usr/local/lib/python3.8/dist-
packages (from keras_tuner) (2.9.1)
Requirement already satisfied: backcall in /usr/local/lib/python3.8/dist-
packages (from ipython->keras_tuner) (0.2.0)
Requirement already satisfied: pickleshare in /usr/local/lib/python3.8/dist-
packages (from ipython->keras_tuner) (0.7.5)
Requirement already satisfied: traitlets>=4.2 in /usr/local/lib/python3.8/dist-
packages (from ipython->keras_tuner) (5.6.0)
Requirement already satisfied: prompt-toolkit<2.1.0,>=2.0.0 in
/usr/local/lib/python3.8/dist-packages (from ipython->keras_tuner) (2.0.10)
Requirement already satisfied: decorator in /usr/local/lib/python3.8/dist-
packages (from ipython->keras_tuner) (4.4.2)
Requirement already satisfied: pygments in /usr/local/lib/python3.8/dist-
packages (from ipython->keras_tuner) (2.6.1)
Requirement already satisfied: setuptools>=18.5 in
/usr/local/lib/python3.8/dist-packages (from ipython->keras_tuner) (57.4.0)
Requirement already satisfied: jedi>=0.10 in /usr/local/lib/python3.8/dist-
packages (from ipython->keras_tuner) (0.18.2)
Requirement already satisfied: pexpect in /usr/local/lib/python3.8/dist-packages
(from ipython->keras_tuner) (4.8.0)
Requirement already satisfied: parso<0.9.0,>=0.8.0 in
/usr/local/lib/python3.8/dist-packages (from jedi>=0.10->ipython->keras_tuner)
(0.8.3)
Requirement already satisfied: wcwidth in /usr/local/lib/python3.8/dist-packages
(from prompt-toolkit<2.1.0,>=2.0.0->ipython->keras_tuner) (0.2.5)
Requirement already satisfied: six>=1.9.0 in /usr/local/lib/python3.8/dist-
packages (from prompt-toolkit<2.1.0,>=2.0.0->ipython->keras_tuner) (1.15.0)
Requirement already satisfied: pyparsing!=3.0.5,>=2.0.2 in
/usr/local/lib/python3.8/dist-packages (from packaging->keras_tuner) (3.0.9)
Requirement already satisfied: ptyprocess>=0.5 in /usr/local/lib/python3.8/dist-
packages (from pexpect->ipython->keras_tuner) (0.7.0)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in
/usr/local/lib/python3.8/dist-packages (from requests->keras_tuner) (1.24.3)
Requirement already satisfied: idna<3,>=2.5 in /usr/local/lib/python3.8/dist-
packages (from requests->keras_tuner) (2.10)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.8/dist-packages (from requests->keras_tuner) (2022.9.24)
Requirement already satisfied: chardet<4,>=3.0.2 in
/usr/local/lib/python3.8/dist-packages (from requests->keras_tuner) (3.0.4)
Requirement already satisfied: tensorboard-plugin-wit>=1.6.0 in
```

```
/usr/local/lib/python3.8/dist-packages (from tensorboard->keras_tuner) (1.8.1)
Requirement already satisfied: wheel>=0.26 in /usr/local/lib/python3.8/dist-
packages (from tensorboard->keras_tuner) (0.38.4)
Requirement already satisfied: werkzeug>=1.0.1 in /usr/local/lib/python3.8/dist-
packages (from tensorboard->keras_tuner) (1.0.1)
Requirement already satisfied: grpcio>=1.24.3 in /usr/local/lib/python3.8/dist-
packages (from tensorboard->keras_tuner) (1.51.1)
Requirement already satisfied: google-auth-oauthlib<0.5,>=0.4.1 in
/usr/local/lib/python3.8/dist-packages (from tensorboard->keras_tuner) (0.4.6)
Requirement already satisfied: protobuf<3.20,>=3.9.2 in
/usr/local/lib/python3.8/dist-packages (from tensorboard->keras_tuner) (3.19.6)
Requirement already satisfied: absl-py>=0.4 in /usr/local/lib/python3.8/dist-
packages (from tensorboard->keras_tuner) (1.3.0)
Requirement already satisfied: tensorboard-data-server<0.7.0,>=0.6.0 in
/usr/local/lib/python3.8/dist-packages (from tensorboard->keras_tuner) (0.6.1)
Requirement already satisfied: google-auth<3,>=1.6.3 in
/usr/local/lib/python3.8/dist-packages (from tensorboard->keras_tuner) (2.15.0)
Requirement already satisfied: markdown>=2.6.8 in /usr/local/lib/python3.8/dist-
packages (from tensorboard->keras_tuner) (3.4.1)
Requirement already satisfied: rsa<5,>=3.1.4 in /usr/local/lib/python3.8/dist-
packages (from google-auth<3,>=1.6.3->tensorboard->keras_tuner) (4.9)
Requirement already satisfied: pyasn1-modules>=0.2.1 in
/usr/local/lib/python3.8/dist-packages (from google-
auth<3,>=1.6.3->tensorboard->keras_tuner) (0.2.8)
Requirement already satisfied: cachetools<6.0,>=2.0.0 in
/usr/local/lib/python3.8/dist-packages (from google-
auth<3,>=1.6.3->tensorboard->keras_tuner) (5.2.0)
Requirement already satisfied: requests-oauthlib>=0.7.0 in
/usr/local/lib/python3.8/dist-packages (from google-auth-
oauthlib<0.5,>=0.4.1->tensorboard->keras_tuner) (1.3.1)
Requirement already satisfied: importlib-metadata>=4.4 in
/usr/local/lib/python3.8/dist-packages (from
markdown>=2.6.8->tensorboard->keras_tuner) (4.13.0)
Requirement already satisfied: zipp>=0.5 in /usr/local/lib/python3.8/dist-
packages (from importlib-
metadata>=4.4->markdown>=2.6.8->tensorboard->keras_tuner) (3.11.0)
Requirement already satisfied: pyasn1<0.5.0,>=0.4.6 in
/usr/local/lib/python3.8/dist-packages (from pyasn1-modules>=0.2.1->google-
auth<3,>=1.6.3->tensorboard->keras_tuner) (0.4.8)
Requirement already satisfied: oauthlib>=3.0.0 in /usr/local/lib/python3.8/dist-
packages (from requests-oauthlib>=0.7.0->google-auth-
oauthlib<0.5,>=0.4.1->tensorboard->keras_tuner) (3.2.2)
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-
wheels/public/simple/
Requirement already satisfied: gdown in /usr/local/lib/python3.8/dist-packages
(4.4.0)
Requirement already satisfied: filelock in /usr/local/lib/python3.8/dist-
packages (from gdown) (3.8.0)
```

```
Requirement already satisfied: requests[socks] in /usr/local/lib/python3.8/dist-
packages (from gdown) (2.23.0)
Requirement already satisfied: six in /usr/local/lib/python3.8/dist-packages
(from gdown) (1.15.0)
Requirement already satisfied: beautifulsoup4 in /usr/local/lib/python3.8/dist-
packages (from gdown) (4.6.3)
Requirement already satisfied: tqdm in /usr/local/lib/python3.8/dist-packages
(from gdown) (4.64.1)
Requirement already satisfied: idna<3,>=2.5 in /usr/local/lib/python3.8/dist-
packages (from requests[socks]->gdown) (2.10)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in
/usr/local/lib/python3.8/dist-packages (from requests[socks]->gdown) (1.24.3)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.8/dist-packages (from requests[socks]->gdown) (2022.9.24)
Requirement already satisfied: chardet<4,>=3.0.2 in
/usr/local/lib/python3.8/dist-packages (from requests[socks]->gdown) (3.0.4)
Requirement already satisfied: PySocks!=1.5.7,>=1.5.6 in
/usr/local/lib/python3.8/dist-packages (from requests[socks]->gdown) (1.7.1)
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-
wheels/public/simple/
Requirement already satisfied: matplotlib==3.1.1 in
/usr/local/lib/python3.8/dist-packages (3.1.1)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.1 in
/usr/local/lib/python3.8/dist-packages (from matplotlib==3.1.1) (3.0.9)
Requirement already satisfied: numpy>=1.11 in /usr/local/lib/python3.8/dist-
packages (from matplotlib==3.1.1) (1.21.6)
Requirement already satisfied: kiwisolver>=1.0.1 in
/usr/local/lib/python3.8/dist-packages (from matplotlib==3.1.1) (1.4.4)
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.8/dist-
packages (from matplotlib==3.1.1) (0.11.0)
Requirement already satisfied: python-dateutil>=2.1 in
/usr/local/lib/python3.8/dist-packages (from matplotlib==3.1.1) (2.8.2)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.8/dist-
packages (from python-dateutil>=2.1->matplotlib==3.1.1) (1.15.0)
/usr/local/lib/python3.8/dist-packages/torch/cuda/__init__.py:497: UserWarning:
Can't initialize NVML
  warnings.warn("Can't initialize NVML")
2022-12-08 21:18:45.742466: E
tensorflow/stream_executor/cuda/cuda_driver.cc:271] failed call to cuInit:
CUDA_ERROR_NO_DEVICE: no CUDA-capable device is detected
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-
wheels/public/simple/
Collecting es-core-news-sm==3.4.0
  Downloading https://github.com/explosion/spacy-
models/releases/download/es_core_news_sm-3.4.0/es_core_news_sm-3.4.0-py3-none-
any.whl (12.9 MB)
     |                | 12.9 MB 1.9 MB/s
Requirement already satisfied: spacy<3.5.0,>=3.4.0 in
```

```
/usr/local/lib/python3.8/dist-packages (from es-core-news-sm==3.4.0) (3.4.3)
Requirement already satisfied: setuptools in /usr/local/lib/python3.8/dist-
packages (from spacy<3.5.0,>=3.4.0->es-core-news-sm==3.4.0) (57.4.0)
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in
/usr/local/lib/python3.8/dist-packages (from spacy<3.5.0,>=3.4.0->es-core-news-
sm==3.4.0) (1.0.9)
Requirement already satisfied: spacy-loggers<2.0.0,>=1.0.0 in
/usr/local/lib/python3.8/dist-packages (from spacy<3.5.0,>=3.4.0->es-core-news-
sm==3.4.0) (1.0.3)
Requirement already satisfied: pathy>=0.3.5 in /usr/local/lib/python3.8/dist-
packages (from spacy<3.5.0,>=3.4.0->es-core-news-sm==3.4.0) (0.10.0)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in
/usr/local/lib/python3.8/dist-packages (from spacy<3.5.0,>=3.4.0->es-core-news-
sm==3.4.0) (3.0.8)
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.10 in
/usr/local/lib/python3.8/dist-packages (from spacy<3.5.0,>=3.4.0->es-core-news-
sm==3.4.0) (3.0.10)
Requirement already satisfied: typer<0.8.0,>=0.3.0 in
/usr/local/lib/python3.8/dist-packages (from spacy<3.5.0,>=3.4.0->es-core-news-
sm==3.4.0) (0.7.0)
Requirement already satisfied: thinc<8.2.0,>=8.1.0 in
/usr/local/lib/python3.8/dist-packages (from spacy<3.5.0,>=3.4.0->es-core-news-
sm==3.4.0) (8.1.5)
Requirement already satisfied: langcodes<4.0.0,>=3.2.0 in
/usr/local/lib/python3.8/dist-packages (from spacy<3.5.0,>=3.4.0->es-core-news-
sm==3.4.0) (3.3.0)
Requirement already satisfied: srsly<3.0.0,>=2.4.3 in
/usr/local/lib/python3.8/dist-packages (from spacy<3.5.0,>=3.4.0->es-core-news-
sm==3.4.0) (2.4.5)
Requirement already satisfied: jinja2 in /usr/local/lib/python3.8/dist-packages
(from spacy<3.5.0,>=3.4.0->es-core-news-sm==3.4.0) (2.11.3)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in
/usr/local/lib/python3.8/dist-packages (from spacy<3.5.0,>=3.4.0->es-core-news-
sm==3.4.0) (2.23.0)
Requirement already satisfied: pydantic!=1.8,!=1.8.1,<1.11.0,>=1.7.4 in
/usr/local/lib/python3.8/dist-packages (from spacy<3.5.0,>=3.4.0->es-core-news-
sm==3.4.0) (1.10.2)
Requirement already satisfied: wasabi<1.1.0,>=0.9.1 in
/usr/local/lib/python3.8/dist-packages (from spacy<3.5.0,>=3.4.0->es-core-news-
sm==3.4.0) (0.10.1)
Requirement already satisfied: catalogue<2.1.0,>=2.0.6 in
/usr/local/lib/python3.8/dist-packages (from spacy<3.5.0,>=3.4.0->es-core-news-
sm==3.4.0) (2.0.8)
Requirement already satisfied: numpy>=1.15.0 in /usr/local/lib/python3.8/dist-
packages (from spacy<3.5.0,>=3.4.0->es-core-news-sm==3.4.0) (1.21.6)
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in
/usr/local/lib/python3.8/dist-packages (from spacy<3.5.0,>=3.4.0->es-core-news-
sm==3.4.0) (4.64.1)
```

```
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.8/dist-
packages (from spacy<3.5.0,>=3.4.0->es-core-news-sm==3.4.0) (21.3)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in
/usr/local/lib/python3.8/dist-packages (from spacy<3.5.0,>=3.4.0->es-core-news-
sm==3.4.0) (2.0.7)
Requirement already satisfied: pyparsing!=3.0.5,>=2.0.2 in
/usr/local/lib/python3.8/dist-packages (from
packaging>=20.0->spacy<3.5.0,>=3.4.0->es-core-news-sm==3.4.0) (3.0.9)
Requirement already satisfied: smart-open<6.0.0,>=5.2.1 in
/usr/local/lib/python3.8/dist-packages (from
pathy>=0.3.5->spacy<3.5.0,>=3.4.0->es-core-news-sm==3.4.0) (5.2.1)
Requirement already satisfied: typing-extensions>=4.1.0 in
/usr/local/lib/python3.8/dist-packages (from
pydantic!=1.8,!=1.8.1,<1.11.0,>=1.7.4->spacy<3.5.0,>=3.4.0->es-core-news-
sm==3.4.0) (4.4.0)
Requirement already satisfied: idna<3,>=2.5 in /usr/local/lib/python3.8/dist-
packages (from requests<3.0.0,>=2.13.0->spacy<3.5.0,>=3.4.0->es-core-news-
sm==3.4.0) (2.10)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in
/usr/local/lib/python3.8/dist-packages (from
requests<3.0.0,>=2.13.0->spacy<3.5.0,>=3.4.0->es-core-news-sm==3.4.0) (1.24.3)
Requirement already satisfied: chardet<4,>=3.0.2 in
/usr/local/lib/python3.8/dist-packages (from
requests<3.0.0,>=2.13.0->spacy<3.5.0,>=3.4.0->es-core-news-sm==3.4.0) (3.0.4)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.8/dist-packages (from
requests<3.0.0,>=2.13.0->spacy<3.5.0,>=3.4.0->es-core-news-sm==3.4.0)
(2022.9.24)
Requirement already satisfied: blis<0.8.0,>=0.7.8 in
/usr/local/lib/python3.8/dist-packages (from
thinc<8.2.0,>=8.1.0->spacy<3.5.0,>=3.4.0->es-core-news-sm==3.4.0) (0.7.9)
Requirement already satisfied: confection<1.0.0,>=0.0.1 in
/usr/local/lib/python3.8/dist-packages (from
thinc<8.2.0,>=8.1.0->spacy<3.5.0,>=3.4.0->es-core-news-sm==3.4.0) (0.0.3)
Requirement already satisfied: click<9.0.0,>=7.1.1 in
/usr/local/lib/python3.8/dist-packages (from
typer<0.8.0,>=0.3.0->spacy<3.5.0,>=3.4.0->es-core-news-sm==3.4.0) (7.1.2)
Requirement already satisfied: MarkupSafe>=0.23 in
/usr/local/lib/python3.8/dist-packages (from jinja2->spacy<3.5.0,>=3.4.0->es-
core-news-sm==3.4.0) (2.0.1)
```

[24]:
```
import string
import pandas as pd
import numpy as np
import csv
import statistics
```

```
#Visualización
from seaborn import color_palette
import matplotlib.pyplot as plt
import seaborn as sns

# Pickle
import pickle

import pyreadstat

#configuración warnings
import warnings
warnings.simplefilter(action='ignore', category=FutureWarning)
warnings.simplefilter(action='ignore', category=UserWarning)


#
import spacy
from collections import Counter

#USAR PARA VER LAS DESCRIPCIONES COMPLETAS
pd.set_option('display.max_colwidth', None)
```

Importamos dataset varios

[25]:
```
url = 'https://drive.google.com/file/d/1pjDOWMUmUzX871xIXfkqqAEEi17C87rd/view?
 ↪usp=sharing'
path = 'https://drive.google.com/uc?export=download&id='+url.split('/')[-2]

ds_propiedades = pd.read_csv(path)
```

[ ]:
```
url = 'https://drive.google.com/file/d/1kHwjnroz-B9e56X8h-FCErXwyJocjnyf/view?
 ↪usp=share_link'
path = 'https://drive.google.com/uc?export=download&id='+url.split('/')[-2]

ds_train = pd.read_csv(path)
ds_train.loc[:, ~ds_train.columns.str.contains('^Unnamed')]
```

[ ]:
```
url = 'https://drive.google.com/file/d/1BtGrZB4SDa-Ta083wigArlxC81TAbYta/view?
 ↪usp=sharing'
path = 'https://drive.google.com/uc?export=download&id='+url.split('/')[-2]

ds_test = pd.read_csv(path)
ds_test.loc[:, ~ds_test.columns.str.contains('^Unnamed')]
```

[28]:
```
url = 'https://drive.google.com/file/d/1xjZFkcBtrJkfAdtzyBGl_wtAIiLzv-RL/view?
 ↪usp=sharing&confirm=t'
```

```
path = 'https://drive.google.com/uc?export=download&id='+url.split('/
  ↪')[-2]+"&confirm=t"

ds_descripciones = pd.read_csv(path)
ds_descripciones = ds_descripciones[ds_descripciones['id'].
  ↪isin(ds_propiedades['id'])]
```

0.1 Analisis del lenguaje natural

Unimos los dos datasets: el de descripciones y el preprocesado en el tp1. Para esto usamos el atributo 'id' de ambos.

[29]:
```
ds_pln_train = ds_train.merge(ds_descripciones, left_on='id', right_on='id')
ds_pln_test = ds_test.merge(ds_descripciones, left_on='id', right_on='id')
```

Normalizamos las expresiones regulares de la columna de descripciones de nuestro dataset

[30]:
```
def normalizar(descripcion):
  descripcion = descripcion.lower()
  descripcion = descripcion.replace('á', 'a')
  descripcion = descripcion.replace('é', 'e')
  descripcion = descripcion.replace('í', 'i')
  descripcion = descripcion.replace('ó', 'o')
  descripcion = descripcion.replace('ú', 'u')
  descripcion = descripcion.replace('ü', 'u')
  descripcion = descripcion.replace('Á', 'a')
  descripcion = descripcion.replace('É', 'e')
  descripcion = descripcion.replace('Í', 'i')
  descripcion = descripcion.replace('Ó', 'o')
  descripcion = descripcion.replace('Ú', 'u')
  descripcion = descripcion.replace('ñ', 'ni')
  descripcion = descripcion.replace('Ñ', 'Ni')
  descripcion = descripcion.replace('m²','m2')
  descripcion = descripcion.replace('M²', 'M2')
  descripcion = descripcion.replace('&', ' ')
  descripcion = descripcion.replace(' y ', ' ')
  descripcion = descripcion.replace(' el ', ' ')
  descripcion = descripcion.replace(' los ', ' ')
  descripcion = descripcion.replace(' la ', ' ')
  descripcion = descripcion.replace(' las ', ' ')
  descripcion = descripcion.replace(' a ', ' ')
  descripcion = descripcion.replace(' o ', ' ')
  descripcion = descripcion.replace('\n', ' ')
  descripcion = descripcion.replace('para', ' ')
  descripcion = descripcion.replace('tiene', ' ')
  descripcion = descripcion.replace('como', ' ')
  descripcion = descripcion.replace('esta', ' ')
  descripcion = descripcion.replace('este', ' ')
```

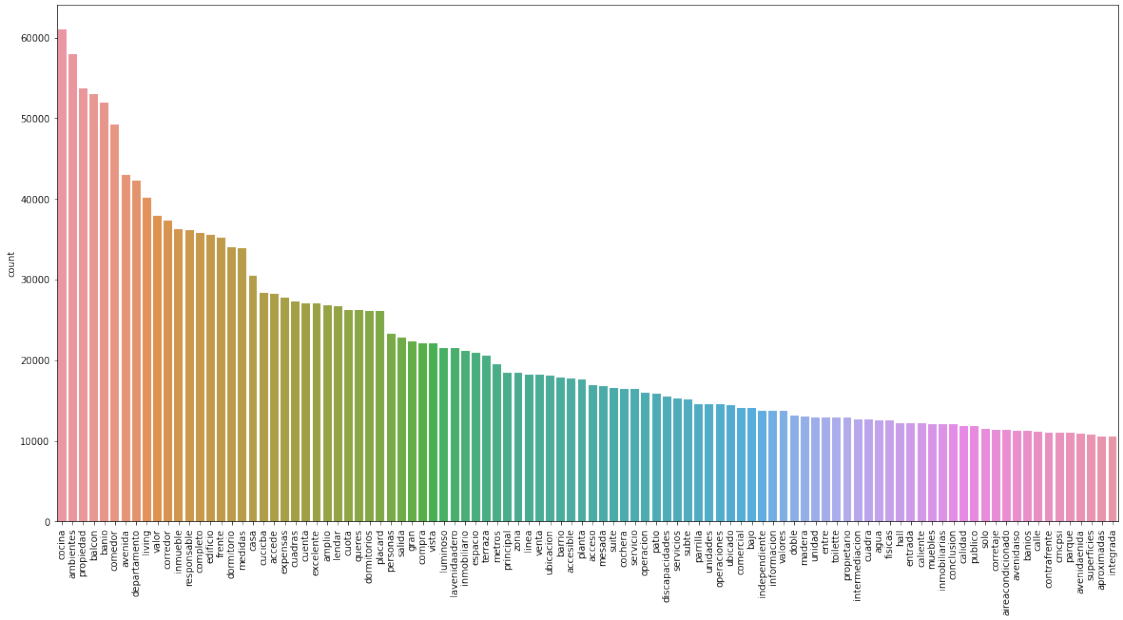
```
  descripcion = descripcion.replace('hasta', ' ')
  descripcion = descripcion.replace('aire acondicionado', 'aire-acondicionado')
  descripcion = descripcion.replace('podes', '')
  descripcion = descripcion.replace('simula', '')
  descripcion = descripcion.replace('gran vista', 'gran-vista')
  descripcion = descripcion.replace('parte', '')
  descripcion = descripcion.replace('encuentra', '')
  descripcion = descripcion.replace('presente', '')
  descripcion = descripcion.replace('todas', '')
  descripcion = descripcion.replace('todos', '')
  descripcion = descripcion.replace('sobre', '')
  descripcion = descripcion.replace('titulo', '')
  descripcion = descripcion.replace('piso', '')
  descripcion = descripcion.replace('tipo', '')
  descripcion = descripcion.replace('todo', '')
  descripcion = descripcion.replace('total', '')
  descripcion = descripcion.replace('aproximado', '')
  descripcion = descripcion.replace(' cion ', '')
  descripcion = descripcion.replace('puede', '')
  descripcion = descripcion.replace('Gimnasio', 'gimnasio')
  descripcion = descripcion.replace('a estrenar', 'a_estrenar')
  descripcion = descripcion.replace('av', 'avenida')
  descripcion = descripcion.replace('av.', 'avenida')

  return descripcion



def borrar_signos_puntuacion_caracteres_especiales(row):
  descripcion = row['property_description']
  descripcion = normalizar(descripcion)

  return descripcion.translate(str.maketrans('', '', string.punctuation))
```

Aplicamos las funciones de limpieza y usamos una regexp para quedarnos nada mas con numeros, letras y simbolos de puntuacion.

[31]:
```
ds_pln_train.property_description = ds_pln_train.apply(lambda row:␣
 ↪borrar_signos_puntuacion_caracteres_especiales(row), axis=1)
ds_pln_test.property_description = ds_pln_test.apply(lambda row:␣
 ↪borrar_signos_puntuacion_caracteres_especiales(row), axis=1)

ds_pln_train['property_description'] = ds_pln_train['property_description'].
 ↪replace(r'[^A-Za-z0-9,.!]', ' ', regex=True)
ds_pln_test['property_description'] = ds_pln_test['property_description'].
 ↪replace(r'[^a-zA-Z0-9,.!]', ' ', regex=True)
```

Vemos la frecuencia de las palabras más usadas en las descripciones de las propiedades y las graficamos.

[32]:
```
word_count = {}
ds_pln_train.reset_index(drop=True, inplace=True)

for i in range(len(ds_pln_train['property_description'])):
    if type(ds_pln_train['property_description'][i]) == str:
        for word in ds_pln_train['property_description'][i].split():
            if len(word) < 4:
                continue
            if word in word_count:
                word_count[word] += 1
            else:
                word_count[word] = 1

# save dictionary in a dataframe

word_count_aux = word_count.copy()

for key, value in word_count_aux.items():
    if value <= 300:
        del word_count[key]

df_word_count = pd.DataFrame.from_dict(word_count, orient='index',␣
 ↪columns=['count'])
df_word_count.sort_values(by=['count'], ascending=False, inplace=True)
df_word_count



# barplot of the 20 most common words
plt.figure(figsize=(20,10))
sns.barplot(x=df_word_count.index[:100], y=df_word_count['count'][:100])
plt.xticks(rotation=90)
plt.show()
```



Generamos las nuevas columnas usando palabras claves en el dominio del problema.

[33]:
```
def crear_columnas_pnl(row, col):
  if col in row['property_description']:
    return 1

  return 0

columnas_pnl = ["cochera", "aire-acondicionado", "gran-vista", "parque",␣
 ↪"balcon", "amplio", "luminoso", "terraza", "a_estrenar", "gimnasio"]
for col in columnas_pnl:
  ds_pln_train[col] = ds_pln_train.apply(lambda row:␣
 ↪crear_columnas_pnl(row,col), axis=1)
  ds_pln_test[col] = ds_pln_test.apply(lambda row: crear_columnas_pnl(row,col),␣
 ↪axis=1)


ds_pln_train.drop('property_description', axis =1, inplace = True)
ds_pln_test.drop('property_description', axis =1, inplace = True)
ds_pln_train.drop('Unnamed: 0', axis =1, inplace = True)
ds_pln_test.drop('Unnamed: 0', axis =1, inplace = True)
```

Exportamos los Datasets a sus correspondientes CSVs

[34]:
```
ds_pln_train.to_csv('ds_pln_train')
ds_pln_test.to_csv('ds_pln_test')
```

Usamos una tecnica basada en frecuencia y palabras que indican “carga de valor” y creamos sus columnas correspondientes al estilo “one-hot encoding”, usamos una expresion regular para realizar una limpieza de las descripciones y normlizamos el texto para hacer mas facil su analisis.