# TAREA 1 PRÁCTICAS UNIR:

**Código implementado en 4 puntos de filtrado y tras lematización:**

#Gold: Glosario de referencia revisado

gold\_data\_eng = json.loads(open("gold\_occupational\_therapy\_eng.json").read())

gold\_data\_fre = json.loads(open("gold\_occupational\_therapy\_fre.json").read())

gold\_data\_spa = json.loads(open("gold\_occupational\_therapy\_spa.json").read())

gold\_data\_ger = json.loads(open("gold\_occupational\_therapy\_ger.json").read())

gold\_terms\_eng = [term['forms']['eng'][0]['text'] for term in gold\_data\_eng]

gold\_terms\_fre = [term['forms']['fre'][0]['text'] for term in gold\_data\_fre]

gold\_terms\_spa = [term['forms']['spa'][0]['text'] for term in gold\_data\_spa]

gold\_terms\_ger = [term['forms']['ger'][0]['text'] for term in gold\_data\_ger]

def MatrizConfusion(terminos\_raw, idioma, paso\_actual):

import numpy as np

array\_raw = np.array(terminos\_raw)

if idioma == "eng":

array\_gold = np.array(gold\_terms\_eng)

elif idioma == "fre":

array\_gold = np.array(gold\_terms\_fre)

elif idioma == "spa":

array\_gold = np.array(gold\_terms\_spa)

elif idioma == "ger":

array\_gold = np.array(gold\_terms\_ger)

else:

print("ERROR: Idioma no reconocido")

# Inicializa los contadores de TP, FP, TN y FN en 0

TP = 0 #Están en ambos arrays

TN = 0 #No se puede calcular?

FP = 0 #Están en raw pero no en gold

FN = 0 #No están en raw pero si en gold

# Comparar los valores en los dos arrays

for extraido in array\_raw:

#print("extraido: ")

#print(extraido)

for valido in array\_gold:

#print(valido)

if extraido == valido:

TP += 1

FP = len(array\_raw) - TP

FN = len(array\_gold) - TP

# Crear la matriz de confusión

matriz\_confusion = [[TP, FP], [TN, FN]]

print("Estamos en el paso: " + paso\_actual)

print("Idioma: " + idioma)

print("Matriz de Confusión:")

print("|TP|FP|")

print("|TN|FN|\n")

for fila in matriz\_confusion:

print(fila)

#Precisión: Porcentaje de predicciones positivas correctas

P=TP/(FP+TP)\*100

#Sensibilidad: Porcentaje de casos positivos detectados

S=TP/(TP+FN)\*100

print("\nPrecisión (%):")

print(P)

print("\nSensibilidad (%):")

print(S)

print("\n----------------------------------------------------------------------------\n")

**Esta función se ha ejecutado en cada punto de filtrado de la siguiente manera:**

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

**Se calcula la matriz de confusión para los datos extraídos en 4 idiomas:**

a) English:

Estamos en el paso: Paso de filtrado 1

Idioma: eng

Matriz de Confusión:

|TP|FP|

|TN|FN|

[401, 1295]

[0, 348]

Precisión (%):

23.6438679245283

Sensibilidad (%):

53.538050734312414

----------------------------------------------------------------------------

Estamos en el paso: Paso de filtrado 2

Idioma: eng

Matriz de Confusión:

|TP|FP|

|TN|FN|

[393, 1241]

[0, 356]

Precisión (%):

24.05140758873929

Sensibilidad (%):

52.46995994659546

----------------------------------------------------------------------------

Estamos en el paso: Paso de filtrado 3

Idioma: eng

Matriz de Confusión:

|TP|FP|

|TN|FN|

[383, 1122]

[0, 366]

Precisión (%):

25.448504983388702

Sensibilidad (%):

51.134846461949266

----------------------------------------------------------------------------

Estamos en el paso: Paso de filtrado 4

Idioma: eng

Matriz de Confusión:

|TP|FP|

|TN|FN|

[344, 919]

[0, 405]

Precisión (%):

27.23673792557403

Sensibilidad (%):

45.92790387182911

----------------------------------------------------------------------------

Estamos en el paso: Paso de filtrado 5 (POST-LEMATIZACIÓN)

Idioma: eng

Matriz de Confusión:

|TP|FP|

|TN|FN|

[537, 657]

[0, 212]

Precisión (%):

44.9748743718593

Sensibilidad (%):

71.69559412550068

----------------------------------------------------------------------------

b) German:

Estamos en el paso: Paso de filtrado 1

Idioma: ger

Matriz de Confusión:

|TP|FP|

|TN|FN|

[134, 1898]

[0, 9]

Precisión (%):

6.594488188976378

Sensibilidad (%):

93.7062937062937

----------------------------------------------------------------------------

Estamos en el paso: Paso de filtrado 2

Idioma: ger

Matriz de Confusión:

|TP|FP|

|TN|FN|

[134, 1893]

[0, 9]

Precisión (%):

6.610754810064135

Sensibilidad (%):

93.7062937062937

----------------------------------------------------------------------------

Estamos en el paso: Paso de filtrado 3

Idioma: ger

Matriz de Confusión:

|TP|FP|

|TN|FN|

[134, 1208]

[0, 9]

Precisión (%):

9.985096870342772

Sensibilidad (%):

93.7062937062937

----------------------------------------------------------------------------

Estamos en el paso: Paso de filtrado 4

Idioma: ger

Matriz de Confusión:

|TP|FP|

|TN|FN|

[134, 1076]

[0, 9]

Precisión (%):

11.074380165289256

Sensibilidad (%):

93.7062937062937

----------------------------------------------------------------------------

Estamos en el paso: Paso de filtrado 5 (POST-LEMATIZACIÓN)

Idioma: ger

Matriz de Confusión:

|TP|FP|

|TN|FN|

[67, 1090]

[0, 76]

Precisión (%):

5.790838375108038

Sensibilidad (%):

46.85314685314685

----------------------------------------------------------------------------

c) French:

Estamos en el paso: Paso de filtrado 1

Idioma: fre

Matriz de Confusión:

|TP|FP|

|TN|FN|

[92, 1577]

[0, 51]

Precisión (%):

5.512282804074296

Sensibilidad (%):

64.33566433566433

----------------------------------------------------------------------------

Estamos en el paso: Paso de filtrado 2

Idioma: fre

Matriz de Confusión:

|TP|FP|

|TN|FN|

[91, 1529]

[0, 52]

Precisión (%):

5.617283950617284

Sensibilidad (%):

63.63636363636363

----------------------------------------------------------------------------

Estamos en el paso: Paso de filtrado 3

Idioma: fre

Matriz de Confusión:

|TP|FP|

|TN|FN|

[91, 1164]

[0, 52]

Precisión (%):

7.250996015936255

Sensibilidad (%):

63.63636363636363

----------------------------------------------------------------------------

Estamos en el paso: Paso de filtrado 4

Idioma: fre

Matriz de Confusión:

|TP|FP|

|TN|FN|

[90, 925]

[0, 53]

Precisión (%):

8.866995073891626

Sensibilidad (%):

62.93706293706294

----------------------------------------------------------------------------

Estamos en el paso: Paso de filtrado 5 (POST-LEMATIZACIÓN)

Idioma: fre

Matriz de Confusión:

|TP|FP|

|TN|FN|

[76, 905]

[0, 67]

Precisión (%):

7.747196738022426

Sensibilidad (%):

53.14685314685315

----------------------------------------------------------------------------

d) Spanish:

Estamos en el paso: Paso de filtrado 1

Idioma: spa

Matriz de Confusión:

|TP|FP|

|TN|FN|

[88, 2765]

[0, 29]

Precisión (%):

3.0844724851034

Sensibilidad (%):

75.21367521367522

----------------------------------------------------------------------------

Estamos en el paso: Paso de filtrado 2

Idioma: spa

Matriz de Confusión:

|TP|FP|

|TN|FN|

[82, 2674]

[0, 35]

Precisión (%):

2.9753265602322205

Sensibilidad (%):

70.08547008547008

----------------------------------------------------------------------------

Estamos en el paso: Paso de filtrado 3

Idioma: spa

Matriz de Confusión:

|TP|FP|

|TN|FN|

[82, 1469]

[0, 35]

Precisión (%):

5.286911669890394

Sensibilidad (%):

70.08547008547008

----------------------------------------------------------------------------

Estamos en el paso: Paso de filtrado 4

Idioma: spa

Matriz de Confusión:

|TP|FP|

|TN|FN|

[81, 1144]

[0, 36]

Precisión (%):

6.612244897959184

Sensibilidad (%):

69.23076923076923

----------------------------------------------------------------------------

Estamos en el paso: Paso de filtrado 5 (POST-LEMATIZACIÓN)

Idioma: spa

Matriz de Confusión:

|TP|FP|

|TN|FN|

[84, 1113]

[0, 33]

Precisión (%):

7.017543859649122

Sensibilidad (%):

71.7948717948718

----------------------------------------------------------------------------

e) Italian:

**No dispongo de fichero gold.**

**Nota**: No se ha podido calcular el idioma italiano, no tengo fichero gold para comparación tras cada paso de filtrado.