TRABALHO DE IAA006 – Arquitetura de Dados

Atividade 1 Construção de Características: Identificador automático de idioma

Equipe 03

- Gustavo Costa de Souza
- Marcos Vinicius de Melo
- Marcus Eneas Silveira Galvao do Rio Apa II
- Patrícia Verdugo Pascoal
- Rodrigo de Araujo
- William de Souza Alencar

Identificador automático de idioma

Problema: Dados um texto de entrada, é possível identificar em qual língua o texto está escrito?

Entrada: "texto qualquer"
Saída: português ou inglês ou francês ou italiano ou...

O processo de Reconhecimento de Padrões

O objetivo desse trabalho é demonstrar o processo de "construção de atributos" e como ele é fundamental para o **Reconhecimento de Padrões (RP)**.

Primeiro um conjunto de "amostras" previamente conhecido (classificado)

In [30]:
```
#
# amostras de texto em diferentes línguas
#
ingles = [
"Hello, how are you?",
"I love to read books.",
"The weather is nice today.",
"Where is the nearest restaurant?",
"What time is it?",
"I enjoy playing soccer.",
"Can you help me with this?",
"I'm going to the movies tonight.",
"This is a beautiful place.",
"I like listening to music.",
"Do you speak English?",
"What is your favorite color?",
"I'm learning to play the guitar.",
"Have a great day!",
"I need to buy some groceries.",
"Let's go for a walk.",
"How was your weekend?",
"I'm excited for the concert.",
"Could you pass me the salt, please?",
"I have a meeting at 2 PM.",
"I'm planning a vacation.",
"She sings beautifully.",
"The cat is sleeping.",
"I want to learn French.",
"I enjoy going to the beach.",
"Where can I find a taxi?",
"I'm sorry for the inconvenience.",
"I'm studying for my exams.",
"I like to cook dinner at home.",
"Do you have any recommendations for restaurants?",
]

espanhol = [
"Hola, ¿cómo estás?",
"Me encanta leer libros.",
"El clima está agradable hoy.",
```

```
"¿Dónde está el restaurante más cercano?",
"¿Qué hora es?",
"Voy al parque todos los días.",
"¿Puedes ayudarme con esto?",
"Me gustaría ir de vacaciones.",
"Este es mi libro favorito.",
"Me gusta bailar salsa.",
"¿Hablas español?",
"¿Cuál es tu comida favorita?",
"Estoy aprendiendo a tocar el piano.",
"¡Que tengas un buen día!",
"Necesito comprar algunas frutas.",
"Vamos a dar un paseo.",
"¿Cómo estuvo tu fin de semana?",
"Estoy emocionado por el concierto.",
"¿Me pasas la sal, por favor?",
"Tengo una reunión a las 2 PM.",
"Estoy planeando unas vacaciones.",
"Ella canta hermosamente.",
"El perro está jugando.",
"Quiero aprender italiano.",
"Disfruto ir a la playa.",
"¿Dónde puedo encontrar un taxi?",
"Lamento las molestias.",
"Estoy estudiando para mis exámenes.",
"Me gusta cocinar la cena en casa.",
"¿Tienes alguna recomendación de restaurantes?",
]

portugues = [
"Estou indo para o trabalho agora.",
"Adoro passar tempo com minha família.",
"Preciso comprar leite e pão.",
"Vamos ao cinema no sábado.",
"Gosto de praticar esportes ao ar livre.",
"O trânsito está terrível hoje.",
"A comida estava deliciosa!",
"Você já visitou o Rio de Janeiro?",
"Tenho uma reunião importante amanhã.",
"A festa começa às 20h.",
"Estou cansado depois de um longo dia de trabalho.",
```

```
"Vamos fazer um churrasco no final de semana.",
"O livro que estou lendo é muito interessante.",
"Estou aprendendo a cozinhar pratos novos.",
"Preciso fazer exercícios físicos regularmente.",
"Vou viajar para o exterior nas férias.",
"Você gosta de dançar?",
"Hoje é meu aniversário!",
"Gosto de ouvir música clássica.",
"Estou estudando para o vestibular.",
"Meu time de futebol favorito ganhou o jogo.",
"Quero aprender a tocar violão.",
"Vamos fazer uma viagem de carro.",
"O parque fica cheio aos finais de semana.",
"O filme que assisti ontem foi ótimo.",
"Preciso resolver esse problema o mais rápido possível.",
"Adoro explorar novos lugares.",
"Vou visitar meus avós no domingo.",
"Estou ansioso para as férias de verão.",
"Gosto de fazer caminhadas na natureza.",
"O restaurante tem uma vista incrível.",
"Vamos sair para jantar no sábado.",
]
```

A "amostras" de texto precisa ser "transformada" em **padrões**

Um padrão é um conjunto de características, geralmente representado por um vetor e um conjunto de padrões no formato de tabela. Onde cada linha é um padrão e as colunas as características e, geralmente, na última coluna a **classe**

In [31]:
```
import random

pre_padroes = []
for frase in ingles:
  pre_padroes.append( [frase, 'inglês'])

for frase in espanhol:
  pre_padroes.append( [frase, 'espanhol'])

for frase in portugues:
  pre_padroes.append( [frase, 'português'])
```

```
random.shuffle(pre_padroes)
print(pre_padroes)
```

```
[['Hola, ¿cómo estás?', 'espanhol'], ['O parque fica cheio aos finais de semana.', 'português'], ['O trânsito está terrível hoj
e.', 'português'], ['Estoy estudiando para mis exámenes.', 'espanhol'], ['Me gusta cocinar la cena en casa.', 'espanhol'], ["Le
t's go for a walk.", 'inglês'], ['Vamos ao cinema no sábado.', 'português'], ['Do you speak English?', 'inglês'], ['I have a me
eting at 2 PM.', 'inglês'], ['Estou cansado depois de um longo dia de trabalho.', 'português'], ['I like to cook dinner at hom
e.', 'inglês'], ['¿Me pasas la sal, por favor?', 'espanhol'], ['O livro que estou lendo é muito interessante.', 'português'],
['Gosto de ouvir música clássica.', 'português'], ["I'm studying for my exams.", 'inglês'], ['This is a beautiful place.', 'ing
lês'], ['O restaurante tem uma vista incrível.', 'português'], ['Gosto de fazer caminhadas na natureza.', 'português'], ['Lamen
to las molestias.', 'espanhol'], ['¿Puedes ayudarme con esto?', 'espanhol'], ['Preciso comprar leite e pão.', 'português'], ['H
ello, how are you?', 'inglês'], ['The cat is sleeping.', 'inglês'], ['Voy al parque todos los días.', 'espanhol'], ['Preciso re
solver esse problema o mais rápido possível.', 'português'], ['Quero aprender a tocar violão.', 'português'], ['Adoro passar te
mpo com minha família.', 'português'], ["I'm going to the movies tonight.", 'inglês'], ["I'm excited for the concert.", 'inglê
s'], ["I'm planning a vacation.", 'inglês'], ['Me gustaría ir de vacaciones.', 'espanhol'], ['Preciso fazer exercícios físicos
regularmente.', 'português'], ['Me encanta leer libros.', 'espanhol'], ['How was your weekend?', 'inglês'], ['¿Qué hora es?',
'espanhol'], ['Vamos a dar un paseo.', 'espanhol'], ['Have a great day!', 'inglês'], ['El perro está jugando.', 'espanhol'],
['¿Tienes alguna recomendación de restaurantes?', 'espanhol'], ['Este es mi libro favorito.', 'espanhol'], ['¿Cómo estuvo tu fi
n de semana?', 'espanhol'], ['Estoy aprendiendo a tocar el piano.', 'espanhol'], ['Do you have any recommendations for restaura
nts?', 'inglês'], ['Estou aprendendo a cozinhar pratos novos.', 'português'], ['Vamos fazer um churrasco no final de semana.',
'português'], ['Adoro explorar novos lugares.', 'português'], ['Vamos sair para jantar no sábado.', 'português'], ['¿Hablas esp
añol?', 'espanhol'], ['I like listening to music.', 'inglês'], ['Quiero aprender italiano.', 'espanhol'], ['Estou indo para o t
rabalho agora.', 'português'], ['Estou ansioso para as férias de verão.', 'português'], ['Tenho uma reunião importante amanh
ã.', 'português'], ['Could you pass me the salt, please?', 'inglês'], ['She sings beautifully.', 'inglês'], ['I need to buy som
e groceries.', 'inglês'], ['What time is it?', 'inglês'], ['Vamos fazer uma viagem de carro.', 'português'], ['I love to read b
ooks.', 'inglês'], ['¿Dónde está el restaurante más cercano?', 'espanhol'], ['Estou estudando para o vestibular.', 'português
s'], ['The weather is nice today.', 'inglês'], ['Meu time de futebol favorito ganhou o jogo.', 'português'], ['¿Dónde puedo enc
ontrar un taxi?', 'espanhol'], ['Você gosta de dançar?', 'português'], ['Gosto de praticar esportes ao ar livre.', 'português
s'], ['El clima está agradable hoy.', 'espanhol'], ['What is your favorite color?', 'inglês'], ['Você já visitou o Rio de Janei
ro?', 'português'], ["I'm learning to play the guitar.", 'inglês'], ['Vou visitar meus avós no domingo.', 'português'], ['A com
ida estava deliciosa!', 'português'], ['Estoy planeando unas vacaciones.', 'espanhol'], ['Disfruto ir a la playa.', 'espanho
l'], ['Me gusta bailar salsa.', 'espanhol'], ['Tengo una reunión a las 2 PM.', 'espanhol'], ['Where can I find a taxi?', 'inglê
s'], ['Necesito comprar algunas frutas.', 'espanhol'], ['¿Cuál es tu comida favorita?', 'espanhol'], ['¡Que tengas un buen dí
a!', 'espanhol'], ['I want to learn French.', 'inglês'], ['O filme que assisti ontem foi ótimo.', 'português'], ['I enjoy playi
ng soccer.', 'inglês'], ["I'm sorry for the inconvenience.", 'inglês'], ['Ella canta hermosamente.', 'espanhol'], ['Vou viajar
para o exterior nas férias.', 'português'], ['I enjoy going to the beach.', 'inglês'], ['A festa começa às 20h.', 'português'],
['Estoy emocionado por el concierto.', 'espanhol'], ['Where is the nearest restaurant?', 'inglês'], ['Can you help me with thi
s?', 'inglês'], ['Hoje é meu aniversário!', 'português']]
```

O DataFrame do pandas facilita a visualização.

In [32]:
```
import pandas as pd
dados = pd.DataFrame(pre_padroes)
dados
```

Out[32]:

	0	1
0	Hola, ¿cómo estás?	espanhol
1	O parque fica cheio aos finais de semana.	português
2	O trânsito está terrível hoje.	português
3	Estoy estudiando para mis exámenes.	espanhol
4	Me gusta cocinar la cena en casa.	espanhol
...	...	...
87	A festa começa às 20h.	português
88	Estoy emocionado por el concierto.	espanhol
89	Where is the nearest restaurant?	inglês
90	Can you help me with this?	inglês
91	Hoje é meu aniversário!	português

92 rows × 2 columns

In [33]:
```
# A criação de ngrams com o modelo TfidfVectorizer foi uma tentativa, porém não melhorou a acurácio do treinamento, por isso n
import numpy as np
import re
from sklearn.feature_extraction.text import TfidfVectorizer

pre_padroes_np = np.array(pre_padroes)
pattern_regex = re.compile('[^\w+]', re.UNICODE)
corpus = [re.sub(pattern_regex, ' ', str(item)) for item in pre_padroes_np[:, 0]]
print(corpus)
```

```
ngram_model = TfidfVectorizer(analyzer='char', ngram_range=(3, 5))
ngram_model.fit_transform(corpus)
```

```
['Hola   cómo estás ', 'O parque fica cheio aos finais de semana ', 'O trânsito está terrível hoje ', 'Estoy estudiando para mi
s exámenes ', 'Me gusta cocinar la cena en casa ', 'Let s go for a walk ', 'Vamos ao cinema no sábado ', 'Do you speak English
', 'I have a meeting at 2 PM ', 'Estou cansado depois de um longo dia de trabalho ', 'I like to cook dinner at home ', ' Me pas
as la sal  por favor ', 'O livro que estou lendo é muito interessante ', 'Gosto de ouvir música clássica ', 'I m studying for m
y exams ', 'This is a beautiful place ', 'O restaurante tem uma vista incrível ', 'Gosto de fazer caminhadas na natureza ', 'La
mento las molestias ', ' Puedes ayudarme con esto ', 'Preciso comprar leite e pão ', 'Hello  how are you ', 'The cat is sleepin
g ', 'Voy al parque todos los días ', 'Preciso resolver esse problema o mais rápido possível ', 'Quero aprender a tocar violão
', 'Adoro passar tempo com minha família ', 'I m going to the movies tonight ', 'I m excited for the concert ', 'I m planning a
vacation ', 'Me gustaría ir de vacaciones ', 'Preciso fazer exercícios físicos regularmente ', 'Me encanta leer libros ', 'How
was your weekend ', ' Qué hora es ', 'Vamos a dar un paseo ', 'Have a great day ', 'El perro está jugando ', ' Tienes alguna re
comendación de restaurantes ', 'Este es mi libro favorito ', ' Cómo estuvo tu fin de semana ', 'Estoy aprendiendo a tocar el pi
ano ', 'Do you have any recommendations for restaurants ', 'Estou aprendendo a cozinhar pratos novos ', 'Vamos fazer um churras
co no final de semana ', 'Adoro explorar novos lugares ', 'Vamos sair para jantar no sábado ', ' Hablas español ', 'I like list
ening to music ', 'Quiero aprender italiano ', 'Estou indo para o trabalho agora ', 'Estou ansioso para as férias de verão ',
'Tenho uma reunião importante amanhã ', 'Could you pass me the salt  please ', 'She sings beautifully ', 'I need to buy some gr
oceries ', 'What time is it ', 'Vamos fazer uma viagem de carro ', 'I love to read books ', ' Dónde está el restaurante más cer
cano ', 'Estou estudando para o vestibular ', 'The weather is nice today ', 'Meu time de futebol favorito ganhou o jogo ', ' Dó
nde puedo encontrar un taxi ', 'Você gosta de dançar ', 'Gosto de praticar esportes ao ar livre ', 'El clima está agradable hoy
', 'What is your favorite color ', 'Você já visitou o Rio de Janeiro ', 'I m learning to play the guitar ', 'Vou visitar meus a
vós no domingo ', 'A comida estava deliciosa ', 'Estoy planeando unas vacaciones ', 'Disfruto ir a la playa ', 'Me gusta bailar
salsa ', 'Tengo una reunión a las 2 PM ', 'Where can I find a taxi ', 'Necesito comprar algunas frutas ', ' Cuál es tu comida f
avorita ', ' Que tengas un buen día ', 'I want to learn French ', 'O filme que assisti ontem foi ótimo ', 'I enjoy playing socc
er ', 'I m sorry for the inconvenience ', 'Ella canta hermosamente ', 'Vou viajar para o exterior nas férias ', 'I enjoy going
to the beach ', 'A festa começa às 20h ', 'Estoy emocionado por el concierto ', 'Where is the nearest restaurant ', 'Can you he
lp me with this ', 'Hoje é meu aniversário ']
```

Out[33]:
```
<Compressed Sparse Row sparse matrix of dtype 'float64'
        with 7296 stored elements and shape (92, 4748)>
```

Construção dos atributos

Esse é o coração desse trabalho e que deverá ser desenvolvido por vocês. Pensem em como podemos "medir" cadas frase/sentença e extrair características que melhorem o resultado do processo de identificação.

Após a criação de cada novo atributo, execute as etapas seguintes e registre as métricas da matriz de confusão. Principalmente acurácia e a precisão.

In [34]:
```
# a entrada é o vetor pre_padroes e a saída desse passo deverá ser "padrões"
import re

from scipy.sparse import hstack

def tamanhoMedioFrases(texto):
  palavras = re.split("\s",texto)
  #print(palavras)
  tamanhos = [len(s) for s in palavras if len(s)>0]
  #print(tamanhos)
  soma = 0
  for t in tamanhos:
    soma=soma+t
  return soma / len(tamanhos)


def caracteresEspeciais(frase):
    return 1 if any(ord(char) > 127 for char in frase) else 0

def possuiDoisSimbolosSucessivos(frase):
    frase = frase.lower();
    for i in range(len(frase) - 1):
        if frase[i] == frase[i+1]:
            return 1;
    return 0;

def proporcaoDeSufixos(frase):
    sufixos_por_idioma = {
        0: ["mente", "dade", "eiro", "ista", "oso", "ável", "ível", "ência", "idade", "ização", "amento", "imento", "ar", "adc
        1: ["ción", "ciones", "dad", "tad", "ero", "encia", "iendo", "amiento", "imiento", "mente", "oy", "ando", "nte", "nta"
        2: ["ing", "ed", "ly", "ness", "ment", "able", "ible", "tion", "sion", "ous"]   # Inglês
    }
    palavras = frase.lower().split()
    contador = [0.0,0.0,0.0]
    for palavra in palavras:
        for posicao, sufixos in sufixos_por_idioma.items():
```

```
            for sufixo in sufixos:
                if palavra.endswith(sufixo):
                    contador[posicao] += 1
    total_palavras = len(palavras)
    for i in range(len(contador)):
        if (contador[i] != 0):
            contador[i] /= total_palavras
    return contador

# retorna um array com a contagem de cada lingua
# posição 0 - pt, posição 1 - es, posição 2 - en
# logo a função será quebrada em 3 caracteristicas
def contagemPalavrasFrequentes(frase):
    palavras_pt = ["que", "não", "porém", "então", "porque", "lá", "saudade", "para", "de", "no", "ao", "estou", "você", "um",
    palavras_es = ["no", "pero", "entonces", "el", "la", "es", "un", "estoy", "cómo", "tu", "gusta", "cuál", "ella", "de", "en
    palavras_en = ["the", "and", "but", "because", "there", "add", "sadness", "you", "i m", "to", "this", "have", "of", "in",

    frase = frase.lower();
    palavras = frase.lower().split()
    contador = [0,0,0]
    for p in palavras:
        if p in palavras_pt:
            contador[0] += 1
        if p in palavras_es:
            contador[1] += 1
        if p in palavras_en:
            contador[2] += 1
    return contador

def proporcaoAcentosPtEs(frase):
    acentos_pt_es = ['â', 'ê', 'ô', 'ã', 'õ', 'á', 'é', 'í', 'ó', 'ú', 'à']
    frase = frase.lower();
    total_acentos = sum(frase.count(c) for c in acentos_pt_es)
    return total_acentos / len(frase)

def criaNGrans(frase):
    x = ngram_model.transform([frase])
    return x.toarray()

def extraiCaracteristicas(frase):
  # frase é um vetor [ 'texto', 'lingua' ]
```

```
  texto = frase[0]
  #print(texto)
  pattern_regex = re.compile('[^\w+]', re.UNICODE)
  texto = re.sub(pattern_regex,' ',texto)
  #print(texto)
  caracteristica1=tamanhoMedioFrases(texto)
  caracteristica2=caracteresEspeciais(texto)
  caracteristica3=possuiDoisSimbolosSucessivos(texto)
  caracteristica4=proporcaoAcentosPtEs(texto)
  caracteristica6, caracteristica7, caracteristica8 = contagemPalavrasFrequentes(texto)
  caracteristica9, caracteristica10, caracteristica11 = proporcaoDeSufixos(texto)
  # acrescente as suas funcoes no vetor padrao
  padrao = [caracteristica1, caracteristica2, caracteristica3, caracteristica4, caracteristica6, caracteristica7, caracteristi
  padrao_array = np.array(padrao)
  return padrao


def geraPadroes(frases):
  padroes = []
  for frase in frases:
    padrao = extraiCaracteristicas(frase)
    padroes.append(padrao)
  return padroes

# converte o formato [frase classe] em
# [caracteristica_1, caracteristica_2,... caracteristica n, classe]
padroes = geraPadroes(pre_padroes)

#
# apenas para visualizacao
print(padroes)

dados = pd.DataFrame(padroes)
dados
```

[[4.333333333333333, 1, 1, 0.1111111111111111, 0, 1, 0, 0.0, 0.0, 0.0, 'espanhol'], [4.125, 0, 0, 0.0, 2, 1, 0, 0.0, 0.0, 0.0, 'português'], [5.0, 1, 1, 0.1, 1, 0, 0, 0.2, 0.0, 0.0, 'português'], [6.0, 1, 0, 0.02857142857142857, 1, 1, 0, 0.0, 0.4, 0.0, 'espanhol'], [3.7142857142857144, 0, 0, 0.0, 0, 3, 0, 0.14285714285714285, 0.0, 0.0, 'espanhol'], [2.3333333333333335, 0, 0, 0.0, 1, 0, 1, 0.0, 0.0, 0.0, 'inglês'], [4.2, 1, 0, 0.038461538461538464, 2, 1, 0, 0.2, 0.0, 0.0, 'português'], [4.25, 0, 0, 0.0, 0, 0, 1, 0.0, 0.0, 0.0, 'inglês'], [2.5714285714285716, 0, 1, 0.0, 1, 0, 2, 0.0, 0.0, 0.14285714285714285, 'inglês'], [4.4444444444444445, 0, 0, 0.0, 4, 2, 0, 0.1111111111111111, 0.0, 0.0, 'português'], [3.2857142857142856, 0, 1, 0.0, 0, 0, 1, 0.0, 0.0, 0.0, 'inglês'], [3.3333333333333335, 0, 1, 0.0, 0, 1, 0, 0.0, 0.0, 0.0, 'espanhol'], [4.625, 1, 1, 0.022222222222222223, 3, 1, 0, 0.0, 0.125, 0.0, 'português'], [5.2, 1, 1, 0.06451612903225806, 1, 1, 0, 0.0, 0.0, 0.0, 'português'], [3.3333333333333335, 0, 0, 0.0, 0, 0, 0, 0.0, 0.0, 0.16666666666666666, 'inglês'], [4.2, 0, 0, 0.0, 1, 0, 2, 0.0, 0.0, 0.0, 'inglês'], [5.166666666666667, 1, 0, 0.02702702702702703, 2, 0, 0, 0.3333333333333333, 0.16666666666666666, 0.0, 'português'], [5.333333333333333, 0, 0, 0.0, 1, 1, 0, 0.0, 0.0, 0.0, 'português'], [6.333333333333333, 0, 0, 0.0, 0, 1, 0, 0.3333333333333333, 0.0, 0.0, 'espanhol'], [5.25, 0, 0, 0.0, 0, 0, 0, 0.0, 0.0, 0.0, 'espanhol'], [4.6, 1, 0, 0.03571428571428571, 1, 0, 0, 0.2, 0.0, 0.0, 'português'], [3.5, 0, 1, 0.0, 0, 0, 1, 0.0, 0.0, 0.0, 'inglês'], [4.0, 0, 1, 0.0, 0, 0, 1, 0.0, 0.0, 0.25, 'inglês'], [3.8333333333333335, 1, 0, 0.034482758620689655, 0, 1, 0, 0.0, 0.16666666666666666, 0.0, 'espanhol'], [5.75, 1, 1, 0.037037037037037035, 1, 0, 0, 0.125, 0.0, 0.0, 'português'], [5.0, 1, 0, 0.03333333333333333, 1, 0, 1, 0.2, 0.2, 0.0, 'português'], [5.166666666666667, 1, 1, 0.02702702702702703, 0, 0, 0, 0.16666666666666666, 0.0, 0.0, 'português'], [3.5714285714285716, 0, 0, 0.0, 0, 0, 2, 0.0, 0.0, 0.14285714285714285, 'inglês'], [3.6666666666666665, 0, 0, 0.0, 0, 0, 1, 0.0, 0.0, 0.16666666666666666, 'inglês'], [3.8, 0, 1, 0.0, 1, 0, 1, 0.0, 0.0, 0.4, 'inglês'], [4.8, 1, 0, 0.034482758620689655, 1, 1, 0, 0.0, 0.2, 0.0, 'espanhol'], [8.2, 1, 0, 0.043478260869565216, 0, 0, 0, 0.2, 0.4, 0.0, 'português'], [4.75, 0, 1, 0.0, 0, 0, 0, 0.0, 0.25, 0.0, 'espanhol'], [4.25, 0, 1, 0.0, 0, 0, 0, 0.0, 0.0, 0.0, 'inglês'], [3.0, 1, 0, 0.07692307692307693, 0, 1, 0, 0.0, 0.0, 0.0, 'espanhol'], [3.2, 0, 0, 0.0, 1, 1, 1, 0.2, 0.0, 0.0, 'espanhol'], [3.25, 0, 0, 0.0, 1, 0, 2, 0.0, 0.0, 0.0, 'inglês'], [4.5, 1, 1, 0.045454545454545456, 0, 1, 0, 0.0, 0.25, 0.0, 'espanhol'], [7.8, 1, 0, 0.022222222222222223, 1, 1, 0, 0.0, 0.2, 0.0, 'espanhol'], [4.2, 0, 0, 0.0, 0, 1, 0, 0.0, 0.0, 0.0, 'espanhol'], [3.8333333333333335, 1, 0, 0.03333333333333333, 1, 3, 0, 0.0, 0.0, 0.0, 'espanhol'], [4.833333333333333, 0, 0, 0.0, 1, 2, 1, 0.16666666666666666, 0.3333333333333333, 0.0, 'espanhol'], [5.857142857142857, 0, 1, 0.0, 0, 0, 2, 0.0, 0.0, 0.0, 'inglês'], [5.833333333333333, 0, 0, 0.0, 2, 0, 1, 0.16666666666666666, 0.0, 0.0, 'português'], [4.5, 0, 1, 0.0, 3, 2, 0, 0.0, 0.0, 0.0, 'português'], [6.25, 0, 0, 0.0, 0, 0, 0, 0.25, 0.0, 0.0, 'português'], [4.5, 1, 0, 0.030303030303030304, 2, 1, 0, 0.3333333333333333, 0.0, 0.0, 'português'], [6.5, 1, 0, 0.0, 0, 0, 0, 0.0, 0.0, 0.0, 'espanhol'], [4.2, 0, 0, 0.0, 0, 0, 1, 0.0, 0.0, 0.2, 'inglês'], [7.333333333333333, 0, 0, 0.0, 0, 0, 0, 0.0, 0.3333333333333333, 0.0, 'espanhol'], [4.5, 0, 0, 0.0, 3, 0, 0, 0.0, 0.0, 0.0, 'português'], [4.428571428571429, 1, 0, 0.05263157894736842, 4, 1, 0, 0.14285714285714285, 0.0, 0.0, 'português'], [6.2, 1, 0, 0.05555555555555555, 1, 0, 0, 0.0, 0.2, 0.0, 'português'], [3.857142857142857, 0, 1, 0.0, 0, 0, 2, 0.0, 0.0, 0.0, 'inglês'], [6.333333333333333, 0, 1, 0.0, 0, 0, 0, 0.0, 0.0, 0.3333333333333333, 'inglês'], [3.8333333333333335, 0, 1, 0.0, 0, 0, 1, 0.0, 0.0, 0.16666666666666666, 'inglês'], [3.0, 0, 0, 0.0, 0, 0, 0, 0.0, 0.0, 0.0, 'inglês'], [4.333333333333333, 0, 1, 0.0, 2, 1, 0, 0.0, 0.0, 0.0, 'português'], [3.2, 0, 1, 0.0, 0, 0, 1, 0.0, 0.0, 0.0, 'inglês'], [5.333333333333333, 1, 0, 0.07692307692307693, 0, 1, 0, 0.0, 0.16666666666666666, 0.0, 'espanhol'], [5.8, 0, 0, 0.0, 3, 0, 0, 0.2, 0.2, 0.0, 'português'], [4.2, 0, 0, 0.0, 0, 0, 1, 0.0, 0.0, 0.0, 'inglês'], [4.375, 0, 0, 0.0, 2, 1, 0, 0.0, 0.0, 0.0, 'português'], [5.0, 1, 0, 0.03225806451612903, 0, 1, 0, 0.2, 0.0, 0.0, 'espanhol'], [4.25, 1, 0, 0.047619047619047616, 2, 1, 0, 0.25, 0.0, 0.0, 'português'], [4.571428571428571, 0, 0, 0.0, 2, 1, 0, 0.2857142857142857, 0.0, 0.0, 'português'], [4.6, 1, 0, 0.03571428571428571, 0, 1, 0, 0.0, 0.2, 0.2, 'espanhol'], [4.6, 0, 0, 0.0, 0, 0, 0, 0.0, 0.0, 0.0, 'inglês'], [3.7142857142857144, 1, 0, 0.06060606060606061, 3, 1, 0, 0.14285714285714285, 0.0, 0.0, 'português'], [3.5714285714285716, 0, 0, 0.0, 0, 0, 2, 0.14285714285714285, 0.0, 0.14285714285714285, 'inglês'], [4.5, 1, 0, 0.030303030303030304, 1, 1, 0, 0.16666666666666666, 0.0, 0.0, 'português'], [5.5, 0, 0, 0.0, 1, 0, 1, 0.0, 0.0, 0.0, 'português'], [7.0, 0, 0, 0.0, 0, 1, 0, 0.0, 0.75, 0.0, 'espanhol'], [3.

```
6, 0, 0, 0.0, 1, 1, 1, 0.0, 0.0, 0.0, 'espanhol'], [4.5, 0, 0, 0.0, 0, 1, 0, 0.25, 0.0, 0.0, 'espanhol'], [3.142857142857143,
1, 0, 0.034482758620689655, 1, 1, 1, 0.0, 0.0, 0.0, 'espanhol'], [3.0, 0, 0, 0.0, 1, 0, 1, 0.0, 0.0, 0.0, 'inglês'], [7.0, 0,
0, 0.0, 0, 0, 0, 0.25, 0.0, 0.0, 'espanhol'], [4.4, 1, 0, 0.03571428571428571, 0, 3, 0, 0.0, 0.0, 0.0, 'espanhol'], [3.6, 1, 0,
0.041666666666666664, 1, 2, 0, 0.0, 0.0, 0.0, 'espanhol'], [3.6, 0, 0, 0.0, 0, 0, 1, 0.0, 0.0, 0.0, 'inglês'], [4.1428571428571
43, 1, 1, 0.027777777777777776, 2, 1, 0, 0.0, 0.0, 0.0, 'português'], [4.75, 0, 1, 0.0, 0, 0, 0, 0.0, 0.25, 0.25, 'inglês'],
[4.333333333333333, 0, 1, 0.0, 0, 0, 1, 0.0, 0.0, 0.0, 'inglês'], [7.0, 0, 1, 0.0, 0, 1, 0, 0.3333333333333333, 1.0, 0.0, 'espa
nhol'], [4.428571428571429, 1, 0, 0.02631578947368421, 2, 0, 0, 0.14285714285714285, 0.0, 0.0, 'português'], [3.5, 0, 0, 0.0,
0, 0, 2, 0.0, 0.16666666666666666, 0.16666666666666666, 'inglês'], [3.4, 1, 0, 0.045454545454545456, 1, 0, 1, 0.0, 0.0, 0.0, 'p
ortuguês'], [5.8, 0, 0, 0.0, 0, 2, 0, 0.2, 0.2, 0.0, 'espanhol'], [5.4, 0, 0, 0.0, 0, 0, 1, 0.0, 0.0, 0.0, 'inglês'], [3.333333
3333333335, 0, 0, 0.0, 0, 0, 2, 0.0, 0.0, 0.0, 'inglês'], [4.75, 1, 0, 0.08695652173913043, 0, 0, 0, 0.0, 0.0, 0.0, 'portuguê
s']]
```

Out[34]:

	0	1	2	3	4	5	6	7	8	9	10
0	4.333333	1	1	0.111111	0	1	0	0.000000	0.0	0.0	espanhol
1	4.125000	0	0	0.000000	2	1	0	0.000000	0.0	0.0	português
2	5.000000	1	1	0.100000	1	0	0	0.200000	0.0	0.0	português
3	6.000000	1	0	0.028571	1	1	0	0.000000	0.4	0.0	espanhol
4	3.714286	0	0	0.000000	0	3	0	0.142857	0.0	0.0	espanhol
...	...	...	...	...	...	...	...	...	...	...	...
87	3.400000	1	0	0.045455	1	0	1	0.000000	0.0	0.0	português
88	5.800000	0	0	0.000000	0	2	0	0.200000	0.2	0.0	espanhol
89	5.400000	0	0	0.000000	0	0	1	0.000000	0.0	0.0	inglês
90	3.333333	0	0	0.000000	0	0	2	0.000000	0.0	0.0	inglês
91	4.750000	1	0	0.086957	0	0	0	0.000000	0.0	0.0	português

92 rows × 11 columns

Treinando o modelo com SVM
Separando o conjunto de treinamento do conjunto de testes

In [35]:
```
from sklearn.model_selection import train_test_split
import numpy as np


#from sklearn.metrics import confusion_matrix

vet = np.array(padroes)
classes = vet[:,-1]         # classes = [p[-1] for p in padroes]
#print(classes)
padroes_sem_classe = vet[:,0:-1]
#print(padroes_sem_classe)
X_train, X_test, y_train, y_test = train_test_split(padroes_sem_classe, classes, test_size=0.25, stratify=classes, random_stat
```

Com os conjuntos separados, podemos "treinar" o modelo usando a SVM.

In [36]:
```
from sklearn import svm
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report

treinador = svm.SVC(random_state=42)  #algoritmo escolhido
modelo = treinador.fit(X_train, y_train)

#
# score com os dados de treinamento
acuracia = modelo.score(X_train, y_train)
print("Acurácia nos dados de treinamento: {:.2f}%".format(acuracia * 100))

#
# melhor avaliar com a matriz de confusão
y_pred = modelo.predict(X_train)
cm = confusion_matrix(y_train, y_pred)
print(cm)
print(classification_report(y_train, y_pred))
```

```
#
# com dados de teste que não foram usados no treinamento
print('métricas mais confiáveis')
y_pred2 = modelo.predict(X_test)
cm = confusion_matrix(y_test, y_pred2)
print(cm)
print(classification_report(y_test, y_pred2))
```

```
Acurácia nos dados de treinamento: 88.41%
[[19  2  1]
 [ 1 22  0]
 [ 3  1 20]]
              precision    recall  f1-score   support

    espanhol       0.83      0.86      0.84        22
      inglês       0.88      0.96      0.92        23
   português       0.95      0.83      0.89        24

    accuracy                           0.88        69
   macro avg       0.89      0.88      0.88        69
weighted avg       0.89      0.88      0.88        69

métricas mais confiáveis
[[6 1 1]
 [0 7 0]
 [2 0 6]]
              precision    recall  f1-score   support

    espanhol       0.75      0.75      0.75         8
      inglês       0.88      1.00      0.93         7
   português       0.86      0.75      0.80         8

    accuracy                           0.83        23
   macro avg       0.83      0.83      0.83        23
weighted avg       0.83      0.83      0.82        23
```

Conclusões

Criamos as características representadas nas funções abaixo:

1. tamanhoMedioFrases(texto) Descrição da característica: Essa função calcula o tamanho médio das palavras no texto, com base na contagem de caracteres por palavra. Idiomas diferentes tendem a ter padrões distintos de comprimento médio das palavras. Por exemplo, palavras do inglês costumam ser mais curtas, enquanto o português pode apresentar palavras mais longas, especialmente por causa de sufixos como “mente”, “dade”, etc. Essa métrica pode ajudar o modelo a distinguir entre línguas com base na estrutura lexical.

2. caracteresEspeciais(frase) Descrição da característica: Essa função verifica se a frase contém caracteres especiais (Unicode > 127). Isso captura a presença de letras acentuadas ou símbolos não-ASCII, que são comuns em português (ex: “ção”, “é”) e espanhol (ex: “niño”, “estás”), mas geralmente ausentes no inglês, que usa apenas caracteres básicos do alfabeto latino. Assim, essa característica ajuda a separar idiomas com e sem acentos.

3. possuiDoisSimbolosSucessivos(frase) Descrição da característica: Essa função verifica se existem letras repetidas consecutivamente, como “ll” ou “ss”. Isso pode ajudar a identificar padrões ortográficos típicos de certos idiomas. Por exemplo, o espanhol frequentemente usa “ll” (“llamar”), e o português pode ter "ss" ("passar"). Embora essa característica seja fraca isoladamente, ela pode contribuir quando combinada com outras.

4. proporcaoDeSufixos(frase) Descrição da característica: Essa função calcula a proporção de palavras que terminam com sufixos comuns de cada idioma (português, espanhol e inglês). Os sufixos são fortes indicadores morfológicos. Por exemplo, "mente" é comum em português e espanhol para advérbios, "ing" em inglês para verbos contínuos.

5. contagemPalavrasFrequentes(frase) Descrição da característica: Essa função conta quantas palavras da frase estão em listas pré-definidas de palavras comuns de português, espanhol e inglês. Como palavras funcionais e muito frequentes são altamente características de cada idioma, essa é uma das features mais importantes.

6. proporcaoAcentosPtEs(frase) Descrição da característica: Essa função calcula a proporção de letras acentuadas típicas do português e espanhol em relação ao total de caracteres da frase. O português tende a usar mais acentos variados (“á”, “â”, “õ”), enquanto o espanhol usa com menos variedade. Portanto, essa caracteristica auxilia na separação do idioma português e espanhol.

7. criaNGrans(frase) Descrição da característica: Essa função transforma a frase em um vetor baseado em n-grams de palavras ou caracteres. Os n-grams capturam padrões locais de coocorrência ou estrutura do idioma, permitindo que o modelo aprenda combinações de letras ou palavras comuns em um idioma. Por exemplo, "que", "est", "the", "you" são sequências comuns em português, espanhol e inglês. Apesar de muito útil, essa característica não trouxe ganhos para a acurácia do modelo por isso não foi considerada, porém deixamos aqui para mostrar que foi explorada.

- A combinação das características selecionadas com o dataset permitiu ao modelo atingir uma acurácia de 88%.