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 [9]:
```
#
# 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 [10]:
```
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)
```

```
[['Me gusta bailar salsa.', 'espanhol'], ['Estou indo para o trabalho agora.', 'português'], ['Estoy aprendiendo a tocar el pia
no.', 'espanhol'], ['Hello, how are you?', 'inglês'], ['Vamos a dar un paseo.', 'espanhol'], ['Estoy emocionado por el conciert
o.', 'espanhol'], ['Estoy planeando unas vacaciones.', 'espanhol'], ["I'm studying for my exams.", 'inglês'], ['Disfruto ir a l
a playa.', 'espanhol'], ['O restaurante tem uma vista incrível.', 'português'], ['Estou cansado depois de um longo dia de traba
lho.', 'português'], ['Necesito comprar algunas frutas.', 'espanhol'], ['Can you help me with this?', 'inglês'], ['Vamos fazer
uma viagem de carro.', 'português'], ['Adoro passar tempo com minha família.', 'português'], ['Gosto de praticar esportes ao ar
livre.', 'português'], ['The cat is sleeping.', 'inglês'], ['Vamos fazer um churrasco no final de semana.', 'português'], ['Ten
ho uma reunião importante amanhã.', 'português'], ['Estou aprendendo a cozinhar pratos novos.', 'português'], ['Você gosta de d
ançar?', 'português'], ['I want to learn French.', 'inglês'], ["Let's go for a walk.", 'inglês'], ['Quiero aprender italiano.',
'espanhol'], ['Have a great day!', 'inglês'], ["I'm learning to play the guitar.", 'inglês'], ['Voy al parque todos los días.',
'espanhol'], ["I'm planning a vacation.", 'inglês'], ['Preciso fazer exercícios físicos regularmente.', 'português'], ['O parqu
e fica cheio aos finais de semana.', 'português'], ['Do you speak English?', 'inglês'], ['Me gusta cocinar la cena en casa.',
'espanhol'], ['She sings beautifully.', 'inglês'], ['I like to cook dinner at home.', 'inglês'], ['A comida estava deliciosa!',
'português'], ['O trânsito está terrível hoje.', 'português'], ['I enjoy going to the beach.', 'inglês'], ['O livro que estou l
endo é muito interessante.', 'português'], ["I'm excited for the concert.", 'inglês'], ['Este es mi libro favorito.', 'espanho
l'], ['Estou ansioso para as férias de verão.', 'português'], ["I'm going to the movies tonight.", 'inglês'], ['Me encanta leer
libros.', 'espanhol'], ['Lamento las molestias.', 'espanhol'], ['Gosto de fazer caminhadas na natureza.', 'português'], ['Vamos
ao cinema no sábado.', 'português'], ['El clima está agradable hoy.', 'espanhol'], ['¿Qué hora es?', 'espanhol'], ['Vamos sair
para jantar no sábado.', 'português'], ['Ella canta hermosamente.', 'espanhol'], ['Do you have any recommendations for restaura
nts?', 'inglês'], ['¿Dónde está el restaurante más cercano?', 'espanhol'], ['Você já visitou o Rio de Janeiro?', 'português'],
['Adoro explorar novos lugares.', 'português'], ['Gosto de ouvir música clássica.', 'português'], ['I need to buy some grocerie
s.', 'inglês'], ['A festa começa às 20h.', 'português'], ['¡Que tengas un buen día!', 'espanhol'], ['¿Me pasas la sal, por favo
r?', 'espanhol'], ['Vou viajar para o exterior nas férias.', 'português'], ['Estou estudando para o vestibular.', 'português'],
['Vou visitar meus avós no domingo.', 'português'], ["I'm sorry for the inconvenience.", 'inglês'], ['What is your favorite col
or?', 'inglês'], ['Hola, ¿cómo estás?', 'espanhol'], ['¿Tienes alguna recomendación de restaurantes?', 'espanhol'], ['I like li
stening to music.', 'inglês'], ['Preciso comprar leite e pão.', 'português'], ['I enjoy playing soccer.', 'inglês'], ['¿Dónde p
uedo encontrar un taxi?', 'espanhol'], ['Tengo una reunión a las 2 PM.', 'espanhol'], ['Meu time de futebol favorito ganhou o j
ogo.', 'português'], ['El perro está jugando.', 'espanhol'], ['Quero aprender a tocar violão.', 'português'], ['¿Cómo estuvo tu
fin de semana?', 'espanhol'], ['I have a meeting at 2 PM.', 'inglês'], ['The weather is nice today.', 'inglês'], ['How was your
weekend?', 'inglês'], ['What time is it?', 'inglês'], ['I love to read books.', 'inglês'], ['Me gustaría ir de vacaciones.', 'e
spanhol'], ['Hoje é meu aniversário!', 'português'], ['¿Hablas español?', 'espanhol'], ['Estoy estudiando para mis exámenes.',
'espanhol'], ['Where can I find a taxi?', 'inglês'], ['¿Puedes ayudarme con esto?', 'espanhol'], ['Preciso resolver esse proble
ma o mais rápido possível.', 'português'], ['¿Cuál es tu comida favorita?', 'espanhol'], ['Where is the nearest restaurant?',
'inglês'], ['Could you pass me the salt, please?', 'inglês'], ['This is a beautiful place.', 'inglês'], ['O filme que assisti o
ntem foi ótimo.', 'português']]
```

O DataFrame do pandas facilita a visualização.

In [11]:
```
import pandas as pd
dados = pd.DataFrame(pre_padroes)
dados
```

Out[11]:

	0	1
0	Me gusta bailar salsa.	espanhol
1	Estou indo para o trabalho agora.	português
2	Estoy aprendiendo a tocar el piano.	espanhol
3	Hello, how are you?	inglês
4	Vamos a dar un paseo.	espanhol
...	...	...
87	¿Cuál es tu comida favorita?	espanhol
88	Where is the nearest restaurant?	inglês
89	Could you pass me the salt, please?	inglês
90	This is a beautiful place.	inglês
91	O filme que assisti ontem foi ótimo.	português

92 rows × 2 columns

In [12]:
```
# 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)
```

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

Out[12]:
```
<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 [13]:
```
# 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
```

[[4.5, 0, 0, 0.0, 0, 1, 0, 0.25, 0.0, 0.0, 'espanhol'], [4.5, 0, 0, 0.0, 3, 0, 0, 0.0, 0.0, 0.0, 'português'], [4.8333333333333
33, 0, 0, 0.0, 1, 2, 1, 0.16666666666666666, 0.3333333333333333, 0.0, 'espanhol'], [3.5, 0, 1, 0.0, 0, 0, 1, 0.0, 0.0, 0.0, 'in
glês'], [3.2, 0, 0, 0.0, 1, 1, 1, 0.2, 0.0, 0.0, 'espanhol'], [5.8, 0, 0, 0.0, 0, 2, 0, 0.2, 0.2, 0.0, 'espanhol'], [7.0, 0, 0,
0.0, 0, 1, 0, 0.0, 0.75, 0.0, 'espanhol'], [3.3333333333333335, 0, 0, 0.0, 0, 0, 0, 0.0, 0.0, 0.16666666666666666, 'inglês'],
[3.6, 0, 0, 0.0, 1, 1, 1, 0.0, 0.0, 0.0, 'espanhol'], [5.166666666666667, 1, 0, 0.02702702702702703, 2, 0, 0, 0.333333333333333
3, 0.16666666666666666, 0.0, 'português'], [4.444444444444445, 0, 0, 0.0, 4, 2, 0, 0.1111111111111111, 0.0, 0.0, 'português'],
[7.0, 0, 0, 0.0, 0, 0, 0, 0.25, 0.0, 0.0, 'espanhol'], [3.3333333333333335, 0, 0, 0.0, 0, 0, 2, 0.0, 0.0, 0.0, 'inglês'], [4.33
3333333333333333, 0, 1, 0.0, 2, 1, 0, 0.0, 0.0, 0.0, 'português'], [5.166666666666667, 1, 1, 0.02702702702702703, 0, 0, 0, 0.16666
666666666666666, 0.0, 0.0, 'português'], [4.571428571428571, 0, 0, 0.0, 2, 1, 0, 0.2857142857142857, 0.0, 0.0, 'português'], [4.0,
0, 1, 0.0, 0, 0, 1, 0.0, 0.0, 0.25, 'inglês'], [4.5, 0, 1, 0.0, 3, 2, 0, 0.0, 0.0, 0.0, 'português'], [6.2, 1, 0, 0.05555555555
555555, 1, 0, 0, 0.0, 0.2, 0.0, 'português'], [5.833333333333333, 0, 0, 0.0, 2, 0, 1, 0.16666666666666666, 0.0, 0.0, 'portuguê
s'], [4.25, 1, 0, 0.047619047619047616, 2, 1, 0, 0.25, 0.0, 0.0, 'português'], [3.6, 0, 0, 0.0, 0, 0, 1, 0.0, 0.0, 0.0, 'inglê
s'], [2.3333333333333335, 0, 0, 0.0, 1, 0, 1, 0.0, 0.0, 0.0, 'inglês'], [7.333333333333333, 0, 0, 0.0, 0, 0, 0, 0.0, 0.33333333
33333333, 0.0, 'espanhol'], [3.25, 0, 0, 0.0, 1, 0, 2, 0.0, 0.0, 0.0, 'inglês'], [3.5714285714285716, 0, 0, 0.0, 0, 0, 2, 0.142
85714285714285, 0.0, 0.14285714285714285, 'inglês'], [3.8333333333333335, 1, 0, 0.034482758620689655, 0, 1, 0, 0.0, 0.166666666
66666666, 0.0, 'espanhol'], [3.8, 0, 1, 0.0, 1, 0, 1, 0.0, 0.0, 0.4, 'inglês'], [8.2, 1, 0, 0.043478260869565216, 0, 0, 0, 0.2,
0.4, 0.0, 'português'], [4.125, 0, 0, 0.0, 2, 1, 0, 0.0, 0.0, 0.0, 'português'], [4.25, 0, 0, 0.0, 0, 0, 1, 0.0, 0.0, 0.0, 'ing
lês'], [3.7142857142857144, 0, 0, 0.0, 0, 3, 0, 0.14285714285714285, 0.0, 0.0, 'espanhol'], [6.333333333333333, 0, 1, 0.0, 0,
0, 0, 0.0, 0.0, 0.3333333333333333, 'inglês'], [3.2857142857142856, 0, 1, 0.0, 0, 0, 1, 0.0, 0.0, 0.0, 'inglês'], [5.5, 0, 0,
0.0, 1, 0, 1, 0.0, 0.0, 0.0, 'português'], [5.0, 1, 1, 0.1, 1, 0, 0, 0.2, 0.0, 0.0, 'português'], [3.5, 0, 0, 0.0, 0, 0, 2, 0.
0, 0.16666666666666666, 0.16666666666666666, 'inglês'], [4.625, 1, 1, 0.022222222222222223, 3, 1, 0, 0.0, 0.125, 0.0, 'portuguê
s'], [3.6666666666666665, 0, 0, 0.0, 0, 0, 1, 0.0, 0.0, 0.16666666666666666, 'inglês'], [4.2, 0, 0, 0.0, 0, 1, 0, 0.0, 0.0, 0.
0, 'espanhol'], [4.428571428571429, 1, 0, 0.05263157894736842, 4, 1, 0, 0.14285714285714285, 0.0, 0.0, 'português'], [3.5714285
714285716, 0, 0, 0.0, 0, 0, 2, 0.0, 0.0, 0.14285714285714285, 'inglês'], [4.75, 0, 1, 0.0, 0, 0, 0, 0.0, 0.25, 0.0, 'espanho
l'], [6.333333333333333, 0, 0, 0.0, 0, 1, 0, 0.3333333333333333, 0.0, 0.0, 'espanhol'], [5.333333333333333, 0, 0, 0.0, 1, 1, 0,
0.0, 0.0, 0.0, 'português'], [4.2, 1, 0, 0.038461538461538464, 2, 1, 0, 0.2, 0.0, 0.0, 'português'], [4.6, 1, 0, 0.035714285714
28571, 0, 1, 0, 0.0, 0.2, 0.2, 'espanhol'], [3.0, 1, 0, 0.07692307692307693, 0, 1, 0, 0.0, 0.0, 0.0, 'espanhol'], [4.5, 1, 0,
0.030303030303030304, 2, 1, 0, 0.3333333333333333, 0.0, 0.0, 'português'], [7.0, 0, 1, 0.0, 0, 1, 0, 0.3333333333333333, 1.0,
0.0, 'espanhol'], [5.857142857142857, 0, 1, 0.0, 0, 0, 2, 0.0, 0.0, 0.0, 'inglês'], [5.333333333333333, 1, 0, 0.076923076923076
93, 0, 1, 0, 0.0, 0.16666666666666666, 0.0, 'espanhol'], [3.7142857142857144, 1, 0, 0.06060606060606061, 3, 1, 0, 0.14285714285
714285, 0.0, 0.0, 'português'], [6.25, 0, 0, 0.0, 0, 0, 0, 0.25, 0.0, 0.0, 'português'], [5.2, 1, 1, 0.06451612903225806, 1, 1,
0, 0.0, 0.0, 0.0, 'português'], [3.8333333333333335, 0, 1, 0.0, 0, 0, 1, 0.0, 0.0, 0.16666666666666666, 'inglês'], [3.4, 1, 0,
0.045454545454545456, 1, 0, 1, 0.0, 0.0, 0.0, 'português'], [3.6, 1, 0, 0.041666666666666664, 1, 2, 0, 0.0, 0.0, 0.0, 'espanho
l'], [3.3333333333333335, 0, 1, 0.0, 0, 1, 0, 0.0, 0.0, 0.0, 'espanhol'], [4.428571428571429, 1, 0, 0.02631578947368421, 2, 0,
0, 0.14285714285714285, 0.0, 0.0, 'português'], [5.8, 0, 0, 0.0, 3, 0, 0, 0.2, 0.2, 0.0, 'português'], [4.5, 1, 0, 0.0303030303
03030304, 1, 1, 0, 0.16666666666666666, 0.0, 0.0, 'português'], [4.333333333333333, 0, 1, 0.0, 0, 0, 1, 0.0, 0.0, 0.0, 'inglê
s'], [4.6, 0, 0, 0.0, 0, 0, 0, 0.0, 0.0, 0.0, 'inglês'], [4.333333333333333, 1, 1, 0.1111111111111111, 0, 1, 0, 0.0, 0.0, 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, 0, 1, 0.0, 0.0, 0.2, 'i
nglês'], [4.6, 1, 0, 0.03571428571428571, 1, 0, 0, 0.2, 0.0, 0.0, 'português'], [4.75, 0, 1, 0.0, 0, 0, 0, 0.0, 0.25, 0.25, 'in
glês'], [5.0, 1, 0, 0.03225806451612903, 0, 1, 0, 0.2, 0.0, 0.0, 'espanhol'], [3.142857142857143, 1, 0, 0.034482758620689655,
1, 1, 1, 0.0, 0.0, 0.0, 'espanhol'], [4.375, 0, 0, 0.0, 2, 1, 0, 0.0, 0.0, 0.0, 'português'], [4.5, 1, 1, 0.045454545454545456,

```
0, 1, 0, 0.0, 0.25, 0.0, 'espanhol'], [5.0, 1, 0, 0.03333333333333333, 1, 0, 1, 0.2, 0.2, 0.0, 'português'], [3.833333333333333
5, 1, 0, 0.03333333333333333, 1, 3, 0, 0.0, 0.0, 0.0, 'espanhol'], [2.5714285714285716, 0, 1, 0.0, 1, 0, 2, 0.0, 0.0, 0.1428571
4285714285, 'inglês'], [4.2, 0, 0, 0.0, 0, 0, 1, 0.0, 0.0, 0.0, 'inglês'], [4.25, 0, 1, 0.0, 0, 0, 0, 0.0, 0.0, 0.0, 'inglês'],
[3.0, 0, 0, 0.0, 0, 0, 0, 0.0, 0.0, 0.0, 'inglês'], [3.2, 0, 1, 0.0, 0, 0, 1, 0.0, 0.0, 0.0, 'inglês'], [4.8, 1, 0, 0.034482758
620689655, 1, 1, 0, 0.0, 0.2, 0.0, 'espanhol'], [4.75, 1, 0, 0.08695652173913043, 0, 0, 0, 0.0, 0.0, 0.0, 'português'], [6.5,
1, 0, 0.0, 0, 0, 0, 0.0, 0.0, 0.0, 'espanhol'], [6.0, 1, 0, 0.02857142857142857, 1, 1, 0, 0.0, 0.4, 0.0, 'espanhol'], [3.0, 0,
0, 0.0, 1, 0, 1, 0.0, 0.0, 0.0, 'inglês'], [5.25, 0, 0, 0.0, 0, 0, 0, 0.0, 0.0, 0.0, 'espanhol'], [5.75, 1, 1, 0.03703703703703
7035, 1, 0, 0, 0.125, 0.0, 0.0, 'português'], [4.4, 1, 0, 0.03571428571428571, 0, 3, 0, 0.0, 0.0, 0.0, 'espanhol'], [5.4, 0, 0,
0.0, 0, 0, 1, 0.0, 0.0, 0.0, 'inglês'], [3.857142857142857, 0, 1, 0.0, 0, 0, 2, 0.0, 0.0, 0.0, 'inglês'], [4.2, 0, 0, 0.0, 1,
0, 2, 0.0, 0.0, 0.0, 'inglês'], [4.142857142857143, 1, 1, 0.027777777777777776, 2, 1, 0, 0.0, 0.0, 0.0, 'português']]
```

Out[13]:

	0	1	2	3	4	5	6	7	8	9	10
0	4.500000	0	0	0.000000	0	1	0	0.250000	0.000000	0.0	espanhol
1	4.500000	0	0	0.000000	3	0	0	0.000000	0.000000	0.0	português
2	4.833333	0	0	0.000000	1	2	1	0.166667	0.333333	0.0	espanhol
3	3.500000	0	1	0.000000	0	0	1	0.000000	0.000000	0.0	inglês
4	3.200000	0	0	0.000000	1	1	1	0.200000	0.000000	0.0	espanhol
...	...	...	...	...	...	...	...	...	...	...	...
87	4.400000	1	0	0.035714	0	3	0	0.000000	0.000000	0.0	espanhol
88	5.400000	0	0	0.000000	0	0	1	0.000000	0.000000	0.0	inglês
89	3.857143	0	1	0.000000	0	0	2	0.000000	0.000000	0.0	inglês
90	4.200000	0	0	0.000000	1	0	2	0.000000	0.000000	0.0	inglês
91	4.142857	1	1	0.027778	2	1	0	0.000000	0.000000	0.0	português

92 rows × 11 columns

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

In [14]:
```
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 [15]:
```
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: 86.96%
[[19  3  0]
 [ 1 22  0]
 [ 4  1 19]]
              precision    recall  f1-score   support

    espanhol       0.79      0.86      0.83        22
      inglês       0.85      0.96      0.90        23
   português       1.00      0.79      0.88        24

    accuracy                           0.87        69
   macro avg       0.88      0.87      0.87        69
weighted avg       0.88      0.87      0.87        69

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

    espanhol       0.88      0.88      0.88         8
      inglês       0.88      1.00      0.93         7
   português       1.00      0.88      0.93         8

    accuracy                           0.91        23
   macro avg       0.92      0.92      0.91        23
weighted avg       0.92      0.91      0.91        23
```