

# Análise de Textos Científicos utilizando Python

**Disciplina:** Introdução a Inteligência Artificial  
**Professor:** Professor Doutor Wagner Igarashi  
**Curso:** Informática

**Equipe:**  
Álvaro de Araújo Ferreira Lima Neto  
Karoline Harummy Romero Moriya  
Rafael Prado Torres

# Agenda

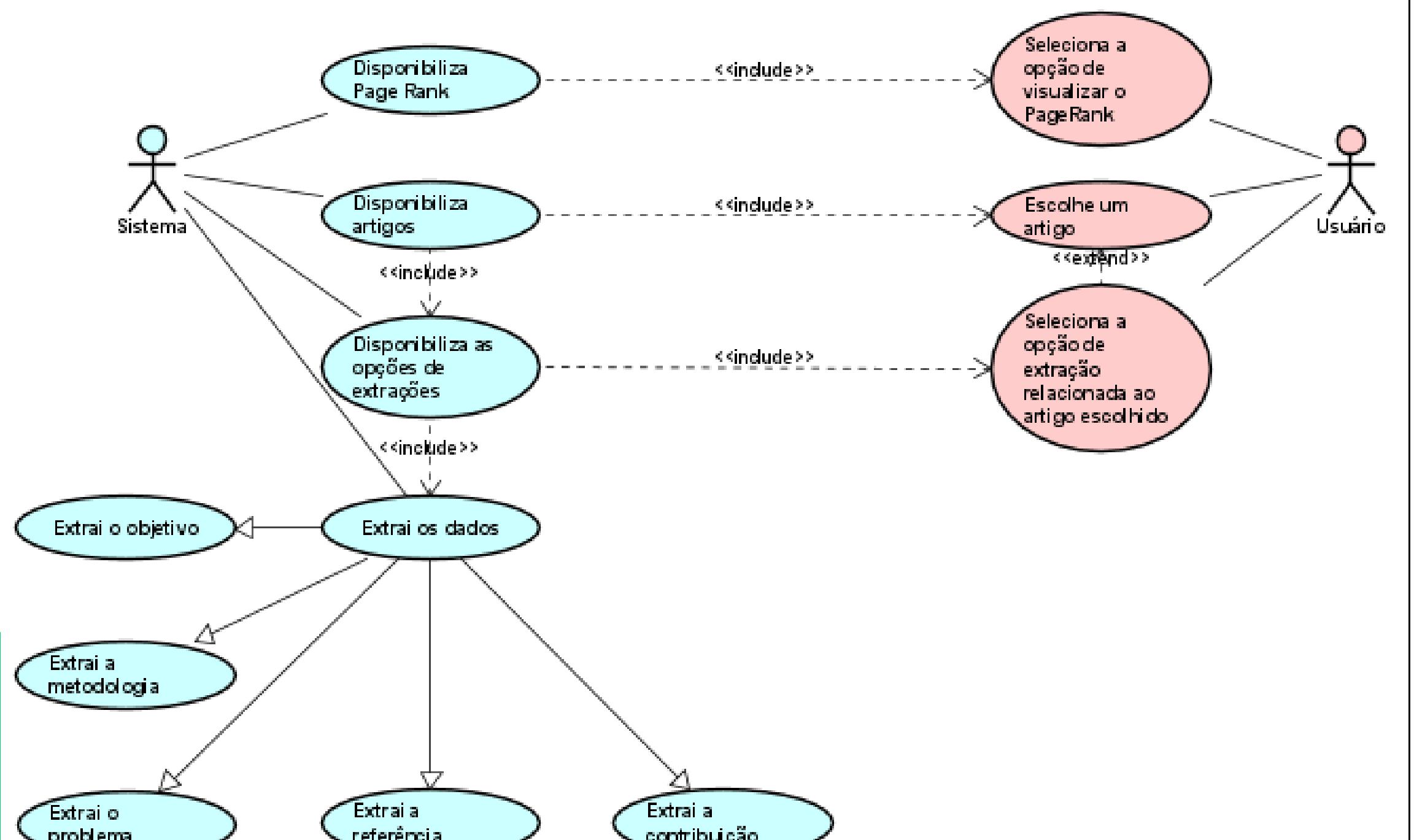
- |   |                       |    |                       |
|---|-----------------------|----|-----------------------|
| 3 | Descrição do Problema | 7  | Instruções para Teste |
| 4 | Casos de Uso          | 8  | Código                |
| 5 | Funcionalidades       | 12 | Simulação             |
| 6 | Plataforma            | 16 | Bibliografia          |

# **Descrição do Problema**

O presente trabalho apresenta uma ontologia sobre a estrutura de um artigo científico e a implementação de um software que utiliza técnicas de Processamento de Linguagem Natural para extrair dados destes artigos.



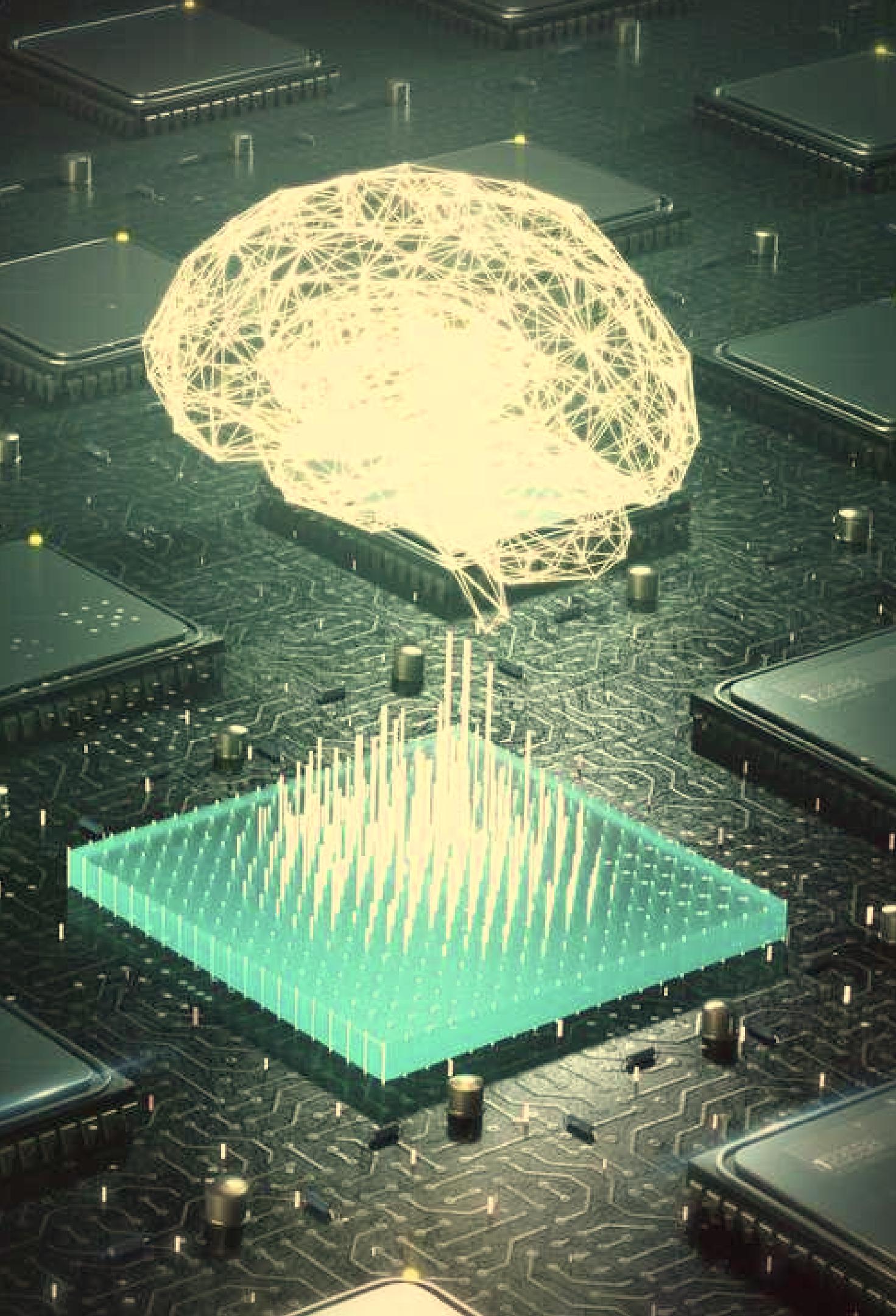
# Casos de Uso



# Funcionalidades

O sistema conta com uma interface, que permite que o usuário selecione um arquivo para extrair os atributos ou selecionar o pageRank para todos os artigos. Caso o usuário escolha algum artigo, abre-se uma opção para que o usuário selecione qual extração de dado deseja, ou seja, qual informação ele quer saber sobre o artigo, estando dentro das possíveis informações:

- Objetivo do artigo
- Problema do artigo
- Metodologia do artigo
- Contribuição do artigo
- Referências do artigo



# Plataforma

Esclarecer os principais objetivos e metas gerais do projeto.



## Dados do computador

Macbook Air 2017, i5, 8gb

RAM

Acer Aspire 5, 8gb RAM

Dell, i5, 8gb RAM



## S.O.

MacOS

Windows 10

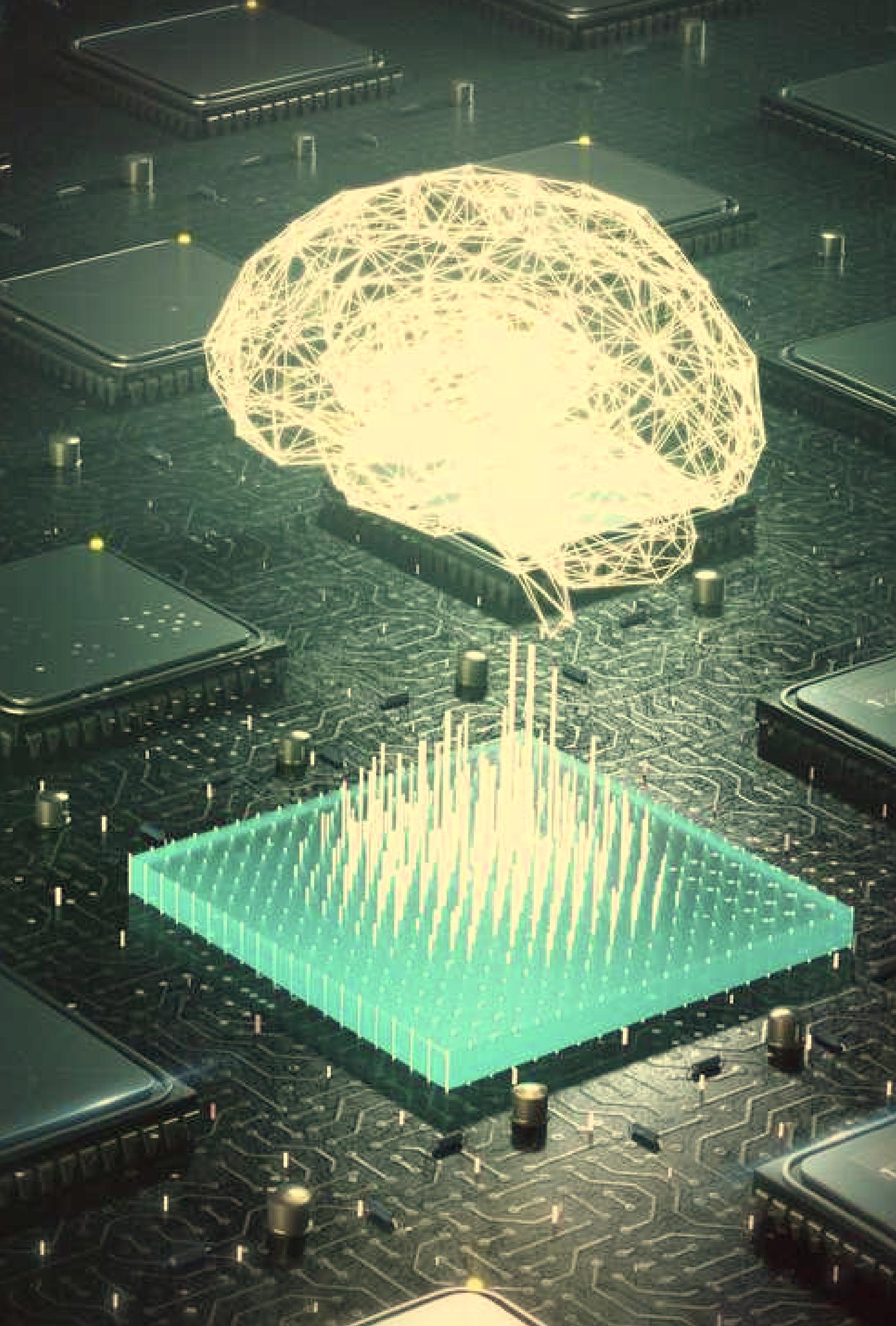


## Linguagem

Python 3.10.9

# Instruções para Teste

O programa funciona a partir da interface.  
Abra o diretório da interface, rode o arquivo interface.py (feito com tkinter) e aparecerá o botão PgeRank e um comboBox para escolher o artigo para extrair os dados desejados.



# Trechos de Código

# Manipular PDF

```
# if busca0:  
#     termina = busca0.end()  
#     pagina = pagina[termina:]  
if busca1:  
    termina = busca1.end()  
    pagina = pagina[termina:]  
elif busca2:  
    termina = busca2.end()  
    pagina = pagina[termina:]  
elif busca3:  
    termina = busca3.end()  
    pagina = pagina[termina:]  
  
return pagina  
  
def removerPontuacao(texto :str) -> str:  
    return "".join(caractere for caractere in texto if caractere not in punctuation)  
  
def limparTexto(texto :str) -> str:  
    texto = texto.lower()  
    texto = removerNumeroPagina(texto)  
    texto = removerPontuacao(texto)  
    return texto  
  
# é usado na classe problema. Será que é necessário?  
def removerBarraN(texto :str) -> str:  
    return texto.replace('\n', ' ')  
  
def lerPDF(arquivo :str) -> str:  
    return PyPDF2.PdfReader(arquivo)
```

# Sumário

```
import re
from manipularPDF import limparTexto

class Sumario:
    def __init__(self, pdfLido :object) -> None:
        self.__sumario = self.__extrairSumario(pdfLido)

    def getSumario(self) -> dict:
        return self.__sumario

    def getPaginasTopico(self, topicoRegex: re) -> list:
        posicaoPaginas = []

        for keys, values in self.__sumario.items():
            if re.match(topicoRegex, keys) and len(posicaoPaginas) == 0:
                posicaoPaginas.append(values)
            elif len(posicaoPaginas) == 1:
                posicaoPaginas.append(values)
        return posicaoPaginas

    def __extrairTextoSumario(self, pdfLido :object) -> list:
        textoFinal = ''
        reInicioTopico = r'sum\s*á\s*rio'
        reFimTopico = r'referências\s*\.*\b'
        achouTopico = False
        posicaoSumario = None
```

# Referências

```
from sumario import Sumario
from extrairTopico import ExtrairTopico

class Referencia():

    def __init__(self, pdfLido: object, sumario :Sumario) -> None:
        self.__topico = ExtrairTopico(sumario, self.__getPadroes())
        self.__referencia = self.__extrairReferencia(pdfLido)

    def getReferencia(self) -> list:
        return self.__referencia

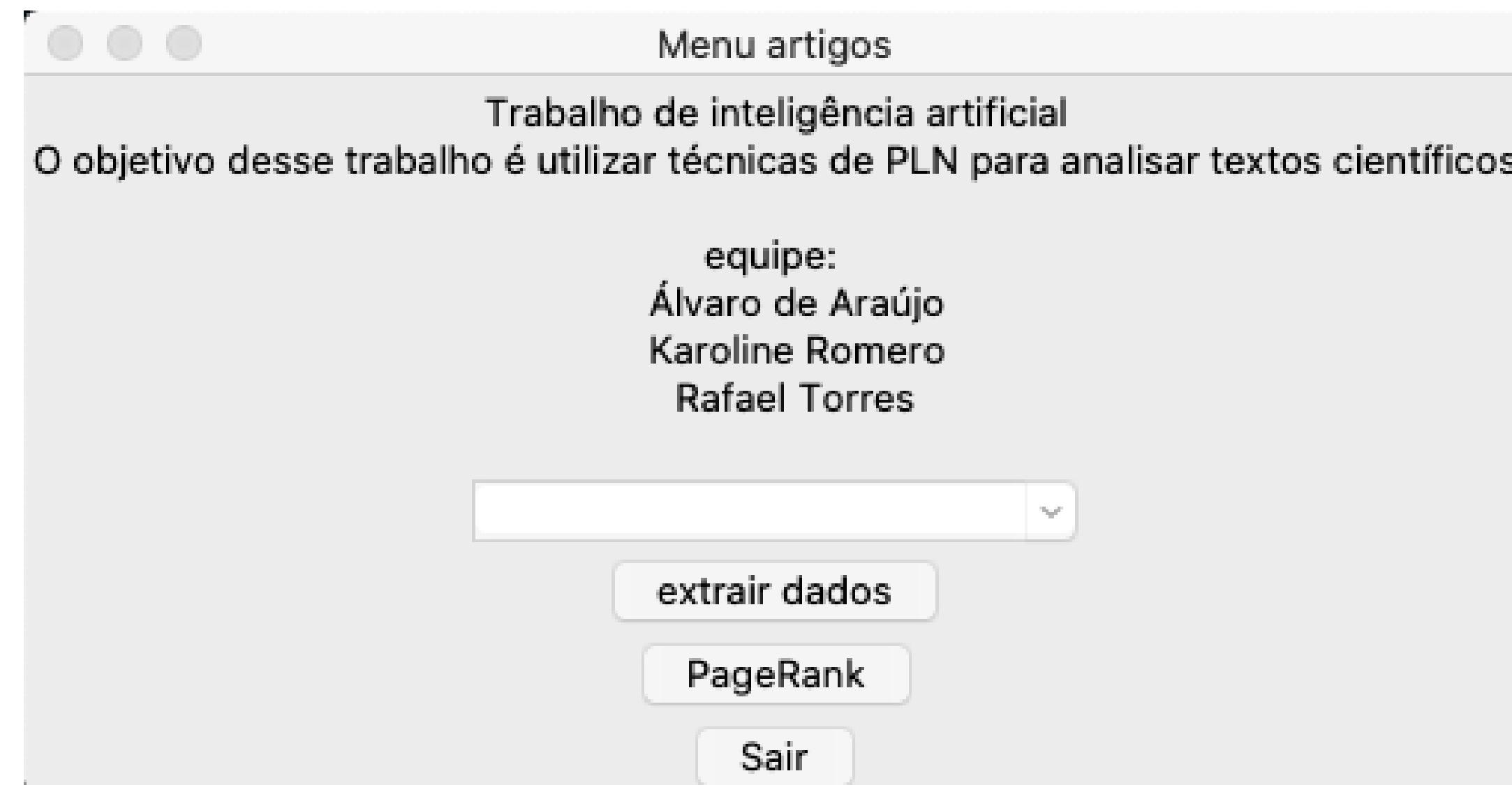
    def __getPadroes(self):
        dictPadroes = {'topico': r'referências\b',
                      'reComecoTopico': r'referências\b',
                      'reFimTopico': r'\s*(apêndice(|s)|anexo)\b'}

        return dictPadroes

    def __extrairReferencia(self, pdfLido: object) -> list:
        textoTopico = self.__topico._getTopico(pdfLido)
        textoTopico = textoTopico.split('.\n')

        return textoTopico
```

# Simulação

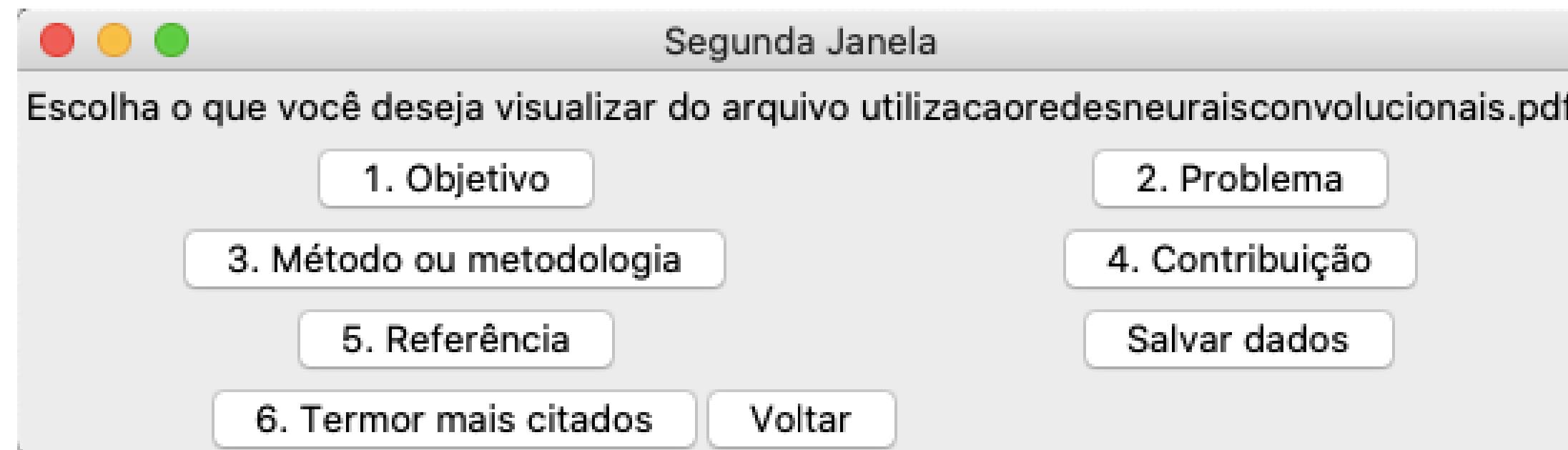


# Simulação

The screenshot shows a web interface titled "PageRank". At the top left are three colored circles (red, yellow, green). The title "PageRank" is centered above a list of document entries. On the right side of the list is a "Voltar" button. The list contains the following entries:

- {{Redes Neurais Convolucionais para detecção de câncer de mama utilizando regiões de interesse de imagens infravermelhas.pdf} 0.004545454545454546} {{Detecção de câncer de mama por meio de imagens infravermelhas utilizando Redes Neurais Convolucionais.pdf} 0.003260869565217392} {{REDE NEURAL CONVOLUCIONAL COM DOIS CANAIS PARA CLASSIFICAÇÃO AUTOMÁTICA DE ARRITMIAS EM SINAIS DE ECG.pdf} 0.002238805970149254} {{Classificação de Arritmias Cardíacas em Sinais de ECG Utilizando Redes Neurais Profundas.pdf} 0.0018518518518521} {{Classificação de Distúrbios e em Folhas de Macieiras Utilizando Redes Neurais Convolucionais.pdf} 0.0017857142857142859} {{CLASSIFICAÇÃO DE VEÍCULOS BASEADA EM DEEP LEARNING PARA APLICAÇÃO EM SEMIFORNOS INTELIGENTES.pdf} 0.001666666666666667} {{Estudo do Desempenho de Redes Neurais Convolucionais Aplicada ao Reconhecimento de Símbolos Musicais, Glaucoma e Texto.pdf} 0.0016483516483516486} {{Sistema para detecção de estradas e obstáculos baseado em imagens RGB e nuvem de pontos para equipamentos de mineração.pdf} 0.0015306122448979593} {{UTILIZAÇÃO DE REDES NEURAIS COMPLETAMENTE CONVOLUCIONAIS PARA IDENTIFICAÇÃO E MEDIÇÃO DE CRÂNIOS FETAIS.pdf} 0.0014851485148514854}

# Simulação



# Simulação

The screenshot shows a software window with a light gray header bar. On the left side of the header are three small colored circles: red, yellow, and green. To the right of these circles, the word "Objetivo" is centered in a bold black font. In the bottom right corner of the header, there is a small rectangular button with the word "Voltar" in black text. The main content area of the window is white and contains the following text in black font:

o objetivo geral deste trabalho é a segmentação de crânio s fetais em imagens de ultrassonografia bidimensionais por meio da rede completamente convolucional v-net.

# Bibliografia

- <https://www.insightlab.ufc.br/pln-processamento-de-linguagem-natural-para-iniciantes/>