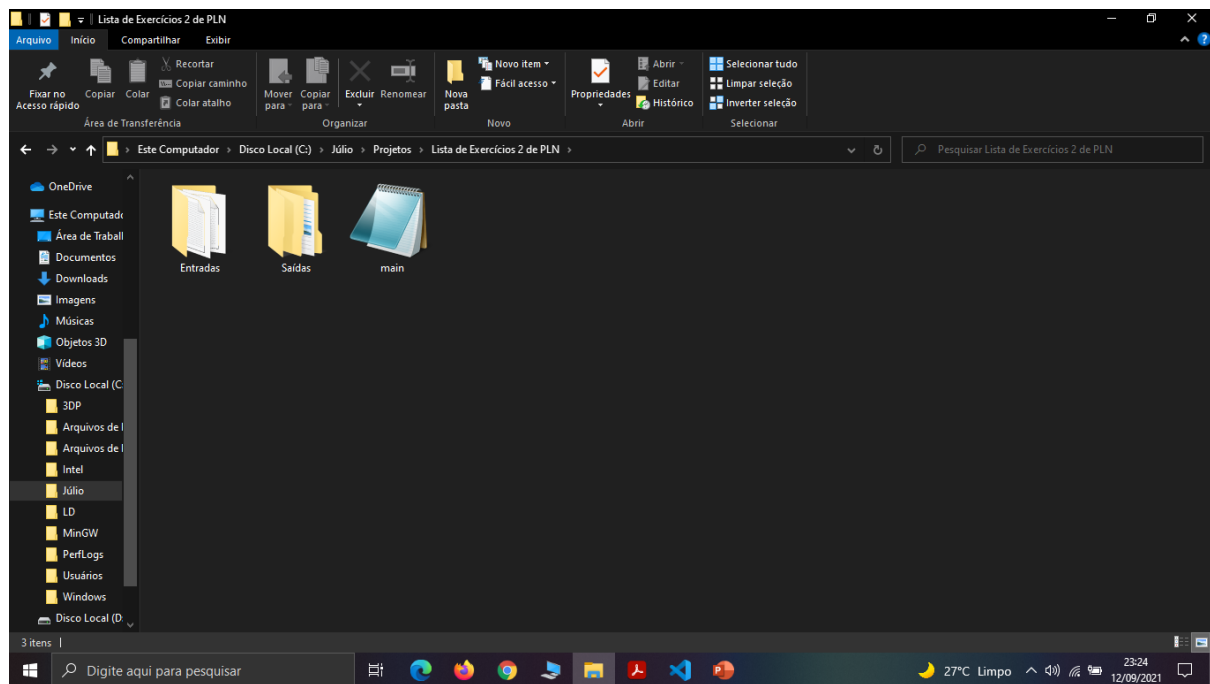


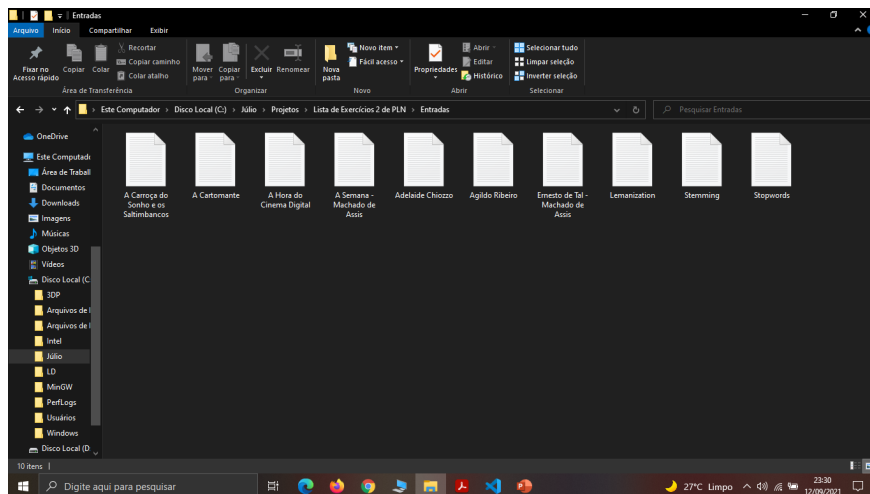
Júlio Cabral
Rodrigo Araújo
Pedro

Lista de Exercício 2 de PLN.

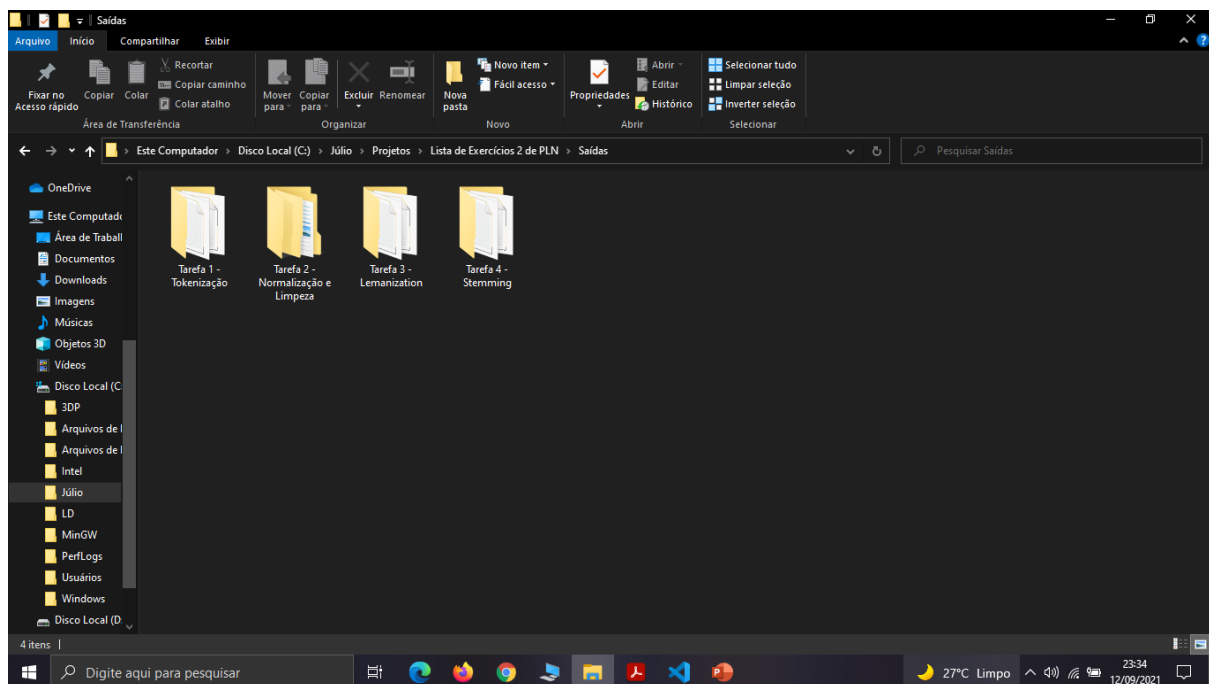
O trabalho conta com 2 pastas, uma pasta é a de Entrada e a outra é a de saída, para cada função usamos os mesmos arquivos de entrada e geramos saídas diferentes, cada função tem saídas diferentes para cada um dos livros utilizados como Entrada, abaixo segue a imagem das 2 pastas e da main:



Dentro da Pasta Entrada temos os 7 livros utilizados como base de dados para as funções, além de mais 3 arquivos que foram utilizados como auxiliares (Stemming, Lemanization e Stopwords). O Stemming é um arquivo contendo diversas palavras utilizadas na função Stemming para buscarmos nos livros, já no arquivo Lemanization temos uma lista de palavras que usamos para fazer a função do Lemanization e por último temos o arquivo Stopwords que contém uma lista de Stopwords utilizadas para fazer a função de retirada de Stopwords, é válido lembrar que assim como os livros utilizados são em português tanto o Stemming, Lemanization e Stopwords utilizadas também foram da língua portuguesa, segue a foto dos arquivos de Entrada:

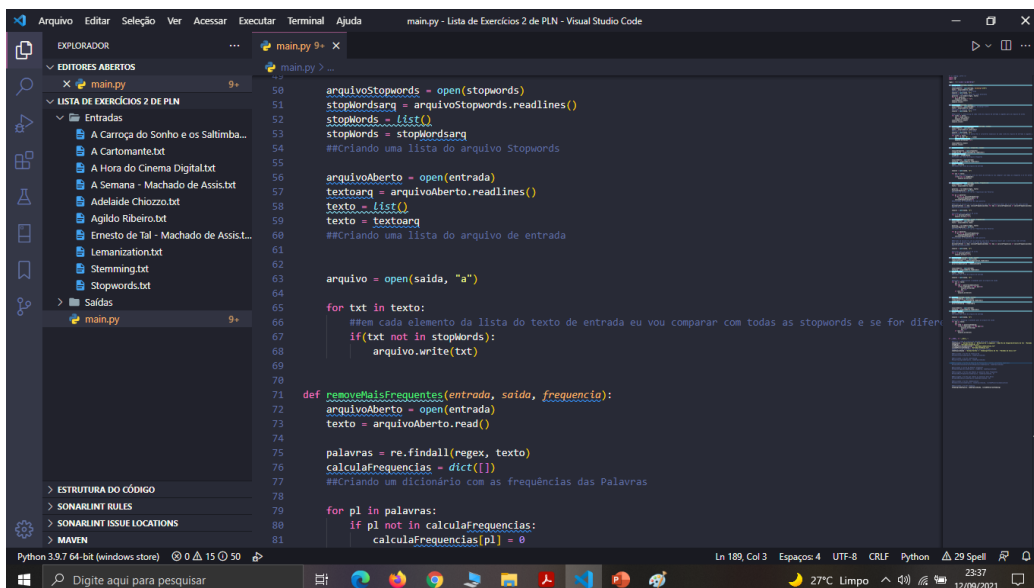
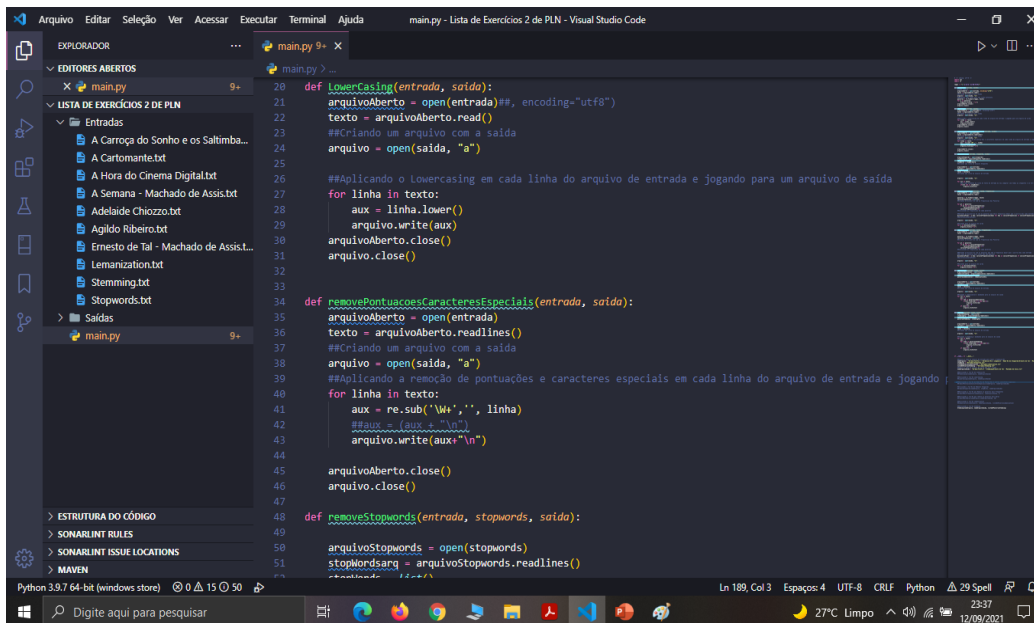


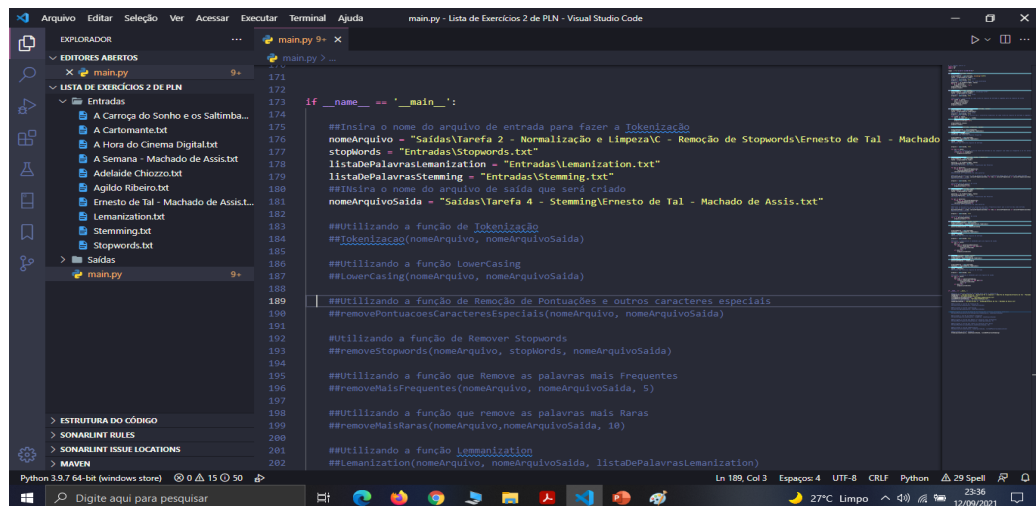
No arquivo de Saída temos todas as funções da Atividade, a cada função nova executada foi gerado 7 saídas diferentes dos livros utilizados na Entrada, para melhor verificação dos processos, com o decorrer das funções até o final vemos que as últimas funções contém todas as funcionalidades completas, como tokenização, normalização de texto, remoção de stopWords, ou seja, os arquivos de saídas vão aglutinando funções anteriores, sendo assim o último arquivo de saída da última Tarefa já possui o pré-processamento completo, segue a foto dos arquivos de Saída divididos por tarefas:



O código da main está todo comentado para facilitar o entendimento, as funções estão todas no começo do arquivo, na main está alguns comentários sobre possíveis chamadas da função, para melhor entendimento deve-se ler os comentários da main, segue algumas imagens da main:

```
main.py 9+ X
main.py > ...
1  # -*- coding: utf-8 -*-
2  import sys
3  import re
4
5  regex = r"[-'a-zA-ZÀ-ÖØ-öø-ÿ0-9]+"
6
7  def Tokenizacao(entrada, saida):
8      #Lendo Arquivo
9      arquivoAberto = open(entrada, encoding="utf8")
10     texto = arquivoAberto.read()
11     ##Criando um arquivo com a saida
12     arquivo = open(saida, "a")
13     ##Mandando cada Token para uma linha diferente
14     palavras = re.findall(regex, texto)
15     for p in palavras:
16         arquivo.write(p + "\n")
17     arquivoAberto.close()
18     arquivo.close()
```





```
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202

if __name__ == '__main__':
    #Insira o nome do arquivo de entrada para fazer a Tokenização
    nomeArquivo = "Saídas\Tarefa 2 - Normalização e Limpeza\Ernesto de Tal - Machado
    stopwords = "Entradas\stopwords.txt"
    listaDePalavrasLemmatization = "Entradas\lemmatization.txt"
    listaDePalavrasStemming = "Entradas\stemming.txt"
    #Insira o nome do arquivo de saída que será criado
    nomeArquivoSaída = "Saídas\Tarefa 4 - Stemming\Ernesto de Tal - Machado de Assis.txt"

    #Utilizando a função de Tokenização
    #tokenizacao(nomeArquivo, nomeArquivoSaída)

    #Utilizando a função LowerCasing
    #lowerCasing(nomeArquivo, nomeArquivoSaída)

    #Utilizando a função de Remoção de Pontuações e outros caracteres especiais
    #removePontuacoesCaracteresEspeciais(nomeArquivo, nomeArquivoSaída)

    #Utilizando a função de Remover Stopwords
    #removeStopwords(nomeArquivo, stopwords, nomeArquivoSaída)

    #Utilizando a função que Remove as palavras mais frequentes
    #removeMaisFrequentes(nomeArquivo, nomeArquivoSaída, 5)

    #Utilizando a função que remove as palavras mais raras
    #removeMaisRaras(nomeArquivo, nomeArquivoSaída, 10)

    #Utilizando a função lemmatization
    #lemmatization(nomeArquivo, nomeArquivoSaída, listaDePalavrasLemmatization)
```

Segue em anexo um link no google drive de todo o trabalho, caso não queira fazer o download do arquivo Winrar enviado:

<https://drive.google.com/drive/folders/1pO65GBiexyBH7587LvEEr42WDMf8qf6C?usp=sharing>