**DOCUMENTAÇÃO**

Linguagem: Python

Bibliotecas: Docx2txt

Métodos:

* Percorrer pasta raiz e gravar em uma lista, o nome dos arquivos da pasta
* Criar arquivo vazio ‘paginasContadas.txt’
* Percorrer lista de nomes e caso for ‘.docx’, converter usando docx2txt, contar quantidade de caracteres – numero de quebras de linhas ‘\n’, adicionar contagem ao total, escrever no arquivo ‘paginasContadas.txt’

Procedimentos;

V1:

* Primeiros testes em arquivos simples, contagem correta com o word
* Após teste com arquivos mais complexos, houve divergência entre as contagens.
* Após isolar casos, foi constatado que há contagem dupla de caracteres em caixas de textos, uma solução será estudada.

V2:

* Após uma analise na biblioteca docx2txt, foi desenvolvido um método para eliminar linhas duplicadas no texto, que funcionou. Porém, ao analisar mais fundo o método aplicado, viu-se que ele eliminava linhas importantes do texto e não apenas duplicadas das caixas de texto. Outro método de abordagem será estudado.

V3:

* O método de abordagem foi alterando, passando a ser usada a extração de texto de arquivos pdf. Para isso, primeiro é feita a conversão do docx para pdf.
* O algoritmo conta os textos do cabeçalho e o word não, então é feita uma contagem e é subtraída no final do algoritmo.
* É gerado um arquivo txt com todo o texto do docx para possível conferencia de conteúdo e quantidade de caracteres.
* Os arquivos pdf e txt são excluídos após o uso dos mesmos pelo algoritmo.
* O algoritmo remove espaços duplos da contagem, otimizando o aproveitamento correto da contagem.
* Após testar com vários arquivos reais, o algoritmo teve uma contagem média de 1,34% menor em relação a contagem do Word, utilizando 19 arquivos.
* Um dos arquivos, teve uma diferença de 18%. Após testes unitários, mudanças no código e no arquivo, não foram encontrados os motivos de tal diferença.
* Foi verificado que o algoritmo conta caracteres de formulas indexadas do Matlab, o Word não considera esses caracteres.
* Há uma inconsistência não controlável em quebras de linhas e espaços em linhas não hifenizadas, onde um espaço é adicionado ao final da linha antes da quebra, adicionando assim um caractere a mais na contagem em situações assim.