cocaClean

開発の意図

- COCA samples のファイルには、以下の「ゴミ」が含まれていて、単語の処理をする際に「ゴミ」となる
  - テキスト ID
  - 匿名記号
  - 段落記号
  - html タグ
  - @long, @longurl

開発方法

- Copilot に指示を出してスクリプトの原案を作ってもらい修正する

■指示のプロンプト

Python 言語で、フォルダー内の複数のテキストファイル（拡張子 ".txt"）を対象に、以下の文字列を削除するスクリプトを書いてください。

削除した結果のファイル名は、それぞれ元のファイル名の後ろに ".cln" という拡張子を付けて保存して下さい。

そして、そのスクリプトを windows 上で実行できる exe ファイルにコンパイルする方法を教えてください。

- その後の試行錯誤
  - 削除文字列を正規表現で表記できるように修正

■出力されたスクリプト

■修正箇所

- 指定するフォルダーをカレントディレクトリーに

```
folder_path = "."
```

- 削除する文字列を正規表現で

```
remove_patterns = [r"(@ )+", r"\<\/*[Pph]\>", r"@@\d{7}", r"&\w+;", r"@long(url)*"]
```

  - まだゴミがあった。↓

```
remove_patterns = [r"(@ )+", r"\<\/*\s*[Pph]\s*\>", r"@@\d{7}", r"&\w+;", r"@long(url)*"]
```

・保存するファイル名の拡張子を修正

```
new_filename = filename.replace(".txt", ".cln.txt")
```

・終了を日本語でなく英語 "Done" に

■修正後のスクリプト

実行ファイルにコンパイル

```
pip install pyinstaller
pyinstaller --onefile cocaClean.py
```

使い方

・処理したいフォルダーにテキストファイルを入れておく
・cocaClean.exe をダウンロードして、そのフォルダーに入れる
・ダブルクリックして実行
・拡張子「.cln.txt」がついた変換済みファイルが保存される