Coleta Violação Gramatical – Python

**1** **FILTROS DAS NOTAS EXPLICATIVAS**

**1.1 Filtros Externos**

**1.1.1 Informações a serem descartadas nas Notas Explicativas**

1. Cabeçalho.
2. Rodapé.
3. Sumário.
4. Tabelas com dados numéricos.
5. Gráficos.
6. Figuras.

<blob:https://web.whatsapp.com/af5b03d2-6ef1-4463-a5aa-62ce6891f406>

O link acima serve como exemplo de uma tabela com dados numéricos, toda tabela com essas características não deve ser mensurada na análise da Violação Gramatical. No mesmo link, abaixo da tabela com dados numéricos, existe um texto explicativo sobre a planilha, esse texto deve ser mantido na análise da Violação Gramatical.

**1.2 Filtros Internos**

Aproximadamente 228 Notas Explicativas de 38 empresas (média de 6 NE por empresa) = atualizado em 23/01/2023.

**1.2.1 Informações a serem descartadas nas Notas Explicativas**

1 Rule ID: DASH\_RULE

VG: ausência de hífen entre palavras.

2 Rule ID: SPACE\_BEFORE\_PUNCTUATION

VG: espaço antes de sinal de pontuação.

3 Rule ID: WHITESPACE\_RULE

VG: espaço repetido entre palavras.

4 Rule ID: ORDINAL\_ABREVIATION

VG: despadronização em abreviatura de números ordinais.

5 Rule ID: SENT\_START\_NUM

VG: números no início de frase devem ser escritos por extenso (dois ao invés de 2).

6 Rule ID: ROMAN\_NUMBERS\_CHECKER

VG: despadronização ao escrever números romanos.

7 Rule ID: DECIMAL\_COMMA

VG: separar casas decimais com ponto no lugar de vírgula.

8 Rule ID: PT\_COMPOUNDS\_POST\_REFORM

VG: ausência de hífen entre palavras.

9 Rule ID: UNPAIRED\_BRACKETS

VG: ausência de aspas em palavras e símbolos.

10 Rule ID: PT\_BARBARISMS\_REPLACE

VG: uso de estrangeirismos.

11 Rule ID: PT\_BR\_SIMPLE\_REPLACE

VG: uso de estrangeirismos.

12 Rule ID: CHEMICAL\_FORMULAS\_TYPOGRAPHY

VG: despadronização de fórmulas.

13 Rule ID: GENERAL\_NUMBER\_FORMAT

VG: despadronização de datas.

14 Rule ID: GENERAL\_NUMBER\_AGREEMENT\_ERRORS

VG: despadronização de datas.

15 Rule ID: PHRASE\_REPETITION

VG: possível frase duplicada (Passivo Circulante e Passivo Não Circulante).

16 Rule ID: ABREVIATIONS\_PUNCTUATION

VG: despadronização em abreviação de palavras.

17 Rule ID: COPYRIGHT

VG: despadronização de identificação de elementos.

18 Rule ID: BARBARISMS

VG: uso de estrangeirismos.

19 Rule ID: UPPERCASE\_AFTER\_COMMA

VG: palavra iniciada em maiúscula após dois pontos (demonstrados abaixo: Controladora).

20 Rule ID: GENERAL\_GENDER\_AGREEMENT\_ERRORS

VG: possível erro de concordância de gênero (Nota Explicativa/Explicativo).

21 Rule ID: ZERO\_IN\_DAYS\_OF\_THE\_MONTH

VG: despadronização de datas.

FILTROS ADICIONADOS DIA 18/01/2023

22 Rule ID: PORTUGUESE\_WORD\_REPEAT\_RULE

VG: repetição de palavras (empresa; empresa).

|  |
| --- |
| Message: Possível erro de escrita: você repetiu uma palavra |
|  | Suggestion: 100,0% |
|  | ... da Canela. 31 de dezembro de 2016 100,0% 100,0% 100,0% 100,0% 100,0% 99.9% 100,0% ... |

23 Rule ID: GENTILICOS\_LINGUAS

VG: nacionalidade escrito com inicial maiúscula.

24 Rule ID: LOWERCASE\_RARE\_WORDS\_GEOGRAPHICAL

VG: nome de país escrito com inicial minúscula.

25 Rule ID: UPPERCASE\_SENTENCE\_START

VG: palavras iniciadas em maiúsculo no meio de frase.

26 Rule ID: REPEATED\_WORDS\_3X

VG: repetição de palavras 3 vezes (conta; conta; conta).

27 Rule ID: SI\_UNITS\_CASE

VG: unidade de medida escrita em letra maiúscula.

28 Rule ID: PERCENT\_WITHOUT\_SPACE

VG: excesso de espaço entre número e sinal de porcentagem.

29 Rule ID: ARTICLES\_PRECEDING\_LOCATIONS

VG: ausência de artigo antes de nome de rua ou avenida.

30 Rule ID: UNPAIRED\_BRACKET\_SUGGESTIONS

VG: ausência de aspas em palavras e símbolos.

31 Rule ID: COMMA\_PARENTHESIS\_WHITESPACE

VG: espaço repetido entre palavras.

32: Rule ID: DOUBLE\_PUNCTUATION

VG: sinal de pontuação duplicado.

33 Rule ID: A\_WORD

VG: erro de concordância de gênero (a/o empresa de auditoria).

34 Rule ID: HOMONYM\_ASSO\_10

VG: despadronização na escrita de nome de empresa.

35 Rule ID: AO90\_CARDINAL\_POINTS\_CASING

VG: letra inicial maiúscula de ponto cardeal (Norte).

36 Rule ID: AO90\_MONTHS\_CASING

VG: letra inicial maiúscula de mês (Agosto).

37 Rule ID: ABREVIATIONS\_PUNCTUATION

VG: despadronização em abreviação de números.

38 Rule ID: NO\_SPACE\_CLOSING\_QUOTE

VG: falta de espaço após sinal de aspas.

39 Rule ID: CERCA\_DE\_NR

VG: despadronização de números aproximados.

40 Rule ID: NUMBER\_ABREVIATION

VG: despadronização em abreviatura de números (nº/n.º).

FILTROS ADICIONADOS DIA 23/01/2023

41 Rule ID: INFORMALITIES

VG: uso de linguagem informal que não configura violação gramatical.

42 Rule ID: ENUMERATIONS\_AND\_AND

VG: uso de linguagem informal que não configura violação gramatical.

43 Rule ID: ARCHAISMS

VG: uso de português arcaico que não configura violação gramatical.

44 Rule ID: INVALID\_DATE

VG: despadronização de datas.

45 Rule ID: PT\_CLICHE\_REPLACE

VG: uso de linguagem informal que não configura violação gramatical (tipo/categoria).

46 Rule ID: QUE\_SER\_ESTAR\_PARTPASSADO

VG: uso de linguagem informal que não configura violação gramatical (tipo/categoria).

47 Rule ID: ALTERNATIVE\_CONJUNCTIONS\_COMMA

VG: uso de linguagem informal que não configura violação gramatical (tipo/categoria).

48 Rule ID: E\_É\_SÃO\_FOI\_FORAM\_SENDO\_SIDO

VG: uso de linguagem informal que não configura violação gramatical (tipo/categoria).

49 Rule ID: QUE\_FORAM\_FOI\_SÃO\_É\_SENDO

VG: uso de linguagem informal que não configura violação gramatical (tipo/categoria).

50 Rule ID: INTERNET\_ABBREVIATIONS

VG: despadronização em abreviação de palavras.

51 Rule ID: PORTUGUESE\_WORD\_REPEAT\_BEGINNING\_RULE

VG: repetição de palavras (empresa; empresa).

52: Rule ID: INTERJECTIONS\_PUNTUATION

VG: uso de linguagem informal que não configura violação gramatical (./!).

53 Rule ID: ERRO\_DE\_CONCORDNCIA\_DO\_GÉNERO\_MASCULINO\_O

VG: possível erro de concordância de gênero (a empresa beta/o empresa beto).

54: Rule ID: FRAGMENT\_TWO\_PREPOSITIONS

VG: uso de linguagem informal que não configura violação gramatical (tipo/categoria).

55: Rule ID: ADVERBIOS\_MODO\_EM\_SEQUENCIA

VG: uso de linguagem informal que não configura violação gramatical (tipo/categoria).

56: Rule ID: CUJO\_LIGACAO\_NOME\_ADJETIVO\_NUMERAL

VG: uso de linguagem informal que não configura violação gramatical (tipo/categoria).

57: Rule ID: PP\_OBJ\_IND

VG: uso de linguagem informal que não configura violação gramatical (pertencem a ele/pertencem-lhe).

58 Rule ID: PT\_SIMPLE\_REPLACE

VG: uso de estrangeirismos.

59: Rule ID: REDUNDANT\_CONJUNCTIONS

VG: uso de linguagem informal que não configura violação gramatical (e, portanto,/e,).

60: Rule ID: COLOCAÇÃO\_ADVÉRBIO

VG: uso de linguagem informal que não configura violação gramatical (o ágio gerado internamente é contabilizado/ o ágio gerado é internamente contabilizado).

61: Rule ID: INTERROGATIVES\_PUNTUATION

VG: uso de linguagem informal que não configura violação gramatical (./?).

62 Rule ID: YEAR\_NUMBER\_FORMAT

VG: despadronização de datas.

63 Rule ID: VERB\_COMMA\_CONJUNCTION

VG: despadronização entre número e sinal de porcentagem.

64: Rule ID: REDUNDANCY\_JUNTO\_COM

VG: uso de linguagem informal que não configura violação gramatical (tipo/categoria).

65: Rule ID: PORTUGUESE\_WRONG\_WORD\_IN\_CONTEXT

VG: uso de linguagem informal que não configura violação gramatical (diferido/deferido).

66 Rule ID: PT\_WEASELWORD\_REPLACE

VG: uso de linguagem informal que não configura violação gramatical (claramente/inegavelmente).

67 Rule ID: COLOCACAO\_ADVERBIOS\_LUGAR

VG: uso de linguagem informal que não configura violação gramatical (nos próximos cinco anos/cinco anos nos próximos).

68 Rule ID: SENTENCE\_WHITESPACE

VG: ausência de espaço entre palavras.

69 Rule ID: DATE\_NEW\_YEAR

VG: despadronização de datas.

70 Rule ID: kWh

VG: despadronização de medidas.

71 Rule ID: ALÉM\_AQUÉM\_RECÉM

VG: ausência de hífen entre palavras.

72 Rule ID: PROFANITY

VG: uso de expressão não reconhecida (“pro rata dia”).

73 Rule ID: PARENTESESE\_AND\_QUOTES\_SPACING

VG: excesso de espaço dentre do parêntesis.

74 Rule ID: EM\_LONGO\_PRAZO

VG: uso de linguagem informal que não configura violação gramatical (em longo prazo/a longo prazo).

75 Rule ID: SUBSTANTIVO\_PLURAL\_E\_CHAVE

VG: ausência de hífen entre palavras.