REGRAS PARA EXCLUSÃO NO PYTHON

Coleta Violação Gramatical – Python

**1** **FILTROS PARA A MENSURAÇÃO DAS VIOLAÇÕES GRAMATICAIS DAS NOTAS EXPLICATIVAS**

**1.1 Filtros Externos**

**1.1.1 Informações excluídas na mensuração das Violações Gramaticais das Notas Explicativas**

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

**1.2 Filtros Internos**

Aproximadamente 1500 Notas Explicativas de 250 empresas (média de 6 NE por empresa) = atualizado em 20/03/2023.

**1.2.1 Regras** **excluídas na mensuração das Violações Gramaticais das Notas Explicativas**

VG1 - FILTROS ADICIONADOS DIA 09/01/2023

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.

VG2 - FILTROS ADICIONADOS DIA 18/01/2023

22 Rule ID: PORTUGUESE\_WORD\_REPEAT\_RULE

VG: repetição de palavras (empresa; empresa).

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).

VG3 - 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.

VG4 - FILTROS ADICIONADOS DIA 06/02/2023

76 Rule ID: PT\_WORDINESS\_REPLACE

VG: uso de linguagem informal que não configura violação gramatical (por meio de/através de).

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

VG: repetição de palavras que não configura violação gramatical (empresa; empresa).

78 Rule ID: DIFERENTES

VG: repetição de palavras que não configura violação gramatical (diferente; diferir).

79 Rule ID: DASH\_ENUMERATION\_SPACE\_RULE

VG: ausência de espaço entre palavras.

80 Rule ID: POR\_PÔ

VG: substituição de palavra que não se aplica ao contexto da nota explicativa (que tem por objetivo/que tem pôr objetivo).

81 Rule ID: SI\_UNITS\_EXPONENT

VG: despadronização de medidas.

82 Rule ID: CURRENCY

VG: despadronização de moedas.

83 Rule ID: DATE\_FUTURE\_VERB\_PAST

VG: despadronização de datas.

84 Rule ID: AO90\_CARDINAL\_POINTS\_CASING

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

85 Rule ID: PT\_WORD\_COHERENCY

VG: substituição de palavra que não se aplica ao contexto da nota explicativa (cotas/quotas).

86 Rule ID: SEMICOLON\_COLON\_SPACING

VG: ausência de espaço entre palavras.

87 Rule ID: UNITS\_OF\_MEASURE\_SPACING

VG: ausência de espaço entre palavras.

88 Rule ID: CACOPHONY

VG: frases com sonoridade incomuns.

89 Rule ID: PARONYM\_VARIA\_340

VG: acentuação inadequada que não configura violação gramatical (varia de 0 a 10/vária de 0 a 10).

90 Rule ID: IRREGULAR\_PAST\_PARTICIPLES

VG: erro de ortografia que não configura violação gramatical (são aceitos/são aceites).

91 Rule ID: LOOSE\_ACCENTS

VG: falta de sinal de aspas que não configura violação gramatical.

92 Rule ID: ANY20016\_2016

VG: despadronização de data que não configura violação gramatical.

93 Rule ID: À\_VERBO\_DE

VG: estilo de escrita que não configura violação gramatical (à taxa de/a taxar).

94 Rule ID: ESTAR\_DE\_ACORDO

VG: estilo de escrita que não configura violação gramatical (está de acordo/concorda).

95 Rule ID: EASILY\_CONFUSED\_RARE\_WORDS

VG: redação divergente não configura violação gramatical (1/48 avos/1/48 avós/1/48 avôs).

96 Rule ID: TIME\_FORMAT

VG: despadronização de horas.

97 Rule ID: MAIORES\_INFORMACOES\_BR

VG: redação divergente que não configura violação gramatical (para maiores informações/para mais informações).

98 Rule ID: DENTRO\_DE\_DA\_DAS\_DO\_DOS\_EM\_NA\_NAS\_NO\_NOS

VG: redação divergente que não configura violação gramatical (dentro da parcela/na parcela).

99 Rule ID: HONRAR\_COMPROMISSOS

VG: redação divergente que não configura violação gramatical (cumprir/honrar).

100 Rule ID: UNLIKELY\_OPENING\_PUNCTUATION

VG: despadronização de sinal de pontuação em tabelas.

VG5 - FILTROS ADICIONADOS DIA 23/02/2023

101 Rule ID: AS\_PONTUAL\_BR

VG: ausência de acentuação que não configura violação gramatical (as duas empresas/às duas horas).

102 Rule ID: DEVE\_DE

VG: redação divergente que não configura violação gramatical (de acordo/devido).

103 Rule ID: MONTH\_REDUNDANCY

VG: despadronização de datas.