

Processamento de Linguagem Natural

Yuri Malheiros

Exercício - Modelo de Linguagem

Descrição

O objetivo deste exercício é criar um modelo de linguagem e executar o método de visualização de Shannon para geração de texto.

O modelo deve ser criado utilizando trigramas, ou seja, a probabilidade da próxima palavra deve ser escolhida levando em consideração duas palavras anteriores: $P(w_n | w_{n-2}, w_{n-1})$.

O texto deve ser gerado através de amostragem. Isto é, a próxima palavra deve ser escolhida de forma aleatória, mas levando em consideração as probabilidades. Assim, palavras mais prováveis são escolhidas com maior frequência, mas palavras pouco prováveis também podem ser escolhidas.

Usem a criatividade de vocês para criar gerações interessantes!

Dados

O modelo escreverá de acordo com os dados que for passado para ele, então preste muita atenção nos dados usados. Lembre que é necessário um volume considerável de dados para que o modelo calcule probabilidades de forma precisa.

Entrega

Este exercício pode ser feito em dupla ou individualmente. Todo o código e resultados devem ser reportados usando um Jupyter Notebook. Crie uma célula no início com nomes e matrículas.

Coloque o seguinte nome no seu notebook: `pln-01-nome1-sobrenome1-nome2-sobrenome2` e envie para o meu e-mail ou pelo Discord.