Processamento de linguagem natural - aula prática

O objetivo desse exercício é implementar um gerador de texto baseado em cadeias de Markov.

Tarefas

- Copiar e gravar textos de alguma fonte (sites de notícias, blogs, livros, etc.)
- Computar as probabilidades dos textos (deve ser possível escolher usar unigrama, bigrama e trigrama)
- Gerar texto

Resposta

A resposta deve ser o texto gerado. Para esse exercício, o programa deve escrever um número N de frases ou palavras.

FAQ

1. **Em qual linguagem eu posso implementar o algoritmo?** Java ou Python.
2. **É para fazer individualmente ou em grupo?** Pode ser feito sozinho ou em dupla.
3. **O texto precisa fazer sentido?** Não necessariamente. Em muitos casos, principalmente usando unigrama e bigrama, os textos não farão muito sentido.
4. **Devo considerar as pontuações como palavras?** Sim.
5. **Como eu sei que uma frase terminou?** Ao escrever uma pontuação final (ponto, exclamação e interrogação) você pode contar que uma frase foi terminada.
6. **Estou fazendo o gerador usando bigramas, mas para a primeira palavra da frase não existe nenhuma anterior. O que eu faço?** Mesmo usando bigrama ou trigrama, é importante computador as probabilidades para contextos menores. Por exemplo, usando bigrama, é importante computar as probabilidades usando unigrama também. Dessa forma, no início da frase, quando não existir nenhuma palavra anterior, você deve usar as probabilidades do unigrama.
7. **Vale ponto?** Sim. O gerador de texto é um dos exercícios de implementação cobrados que poderão adicionar pontos à prova. Entretanto, ele não precisa ser finalizado em sala de aula, pois o prazo de entrega dos exercícios de implementação será numa data próxima a 2a avaliação.