# Processamento de Linguagem Natural

**Aluno:** Leonam Rezende Soares de Miranda
**matrícula**: 2020681492

1. Cite três vantagens da representação distribuída para palavras, ao invés da representação one-hot .

- Os modelos que a representação distribuída para palavras generalizam melhor para palavras raras do que aqueles que usam vetores esparsos (one-hot);
- A representação distribuída para palavras codifica semelhanças entre as palavras, enquanto a representação one-hot não;
- A representação distribuída para palavras é mais fácil de ser incluída em sistemas de aprendizado de máquina do que representação one-hot, pois a dimensão da representação é menor.

2. Sobre o Skip-Gram , marque as alternativas corretas.
a. O algoritmo prediz a palavra central a partir das palavras que formam o contexto.
b. O vetor final é dado pela média dos vetores de entrada.
c. Seu desempenho é pior do que o algoritmo CBOW, quando o corpus é relativamente pequeno.

3. Suponha que você queira classificar comentários sobre filmes em positivos e negativos. Proponha um algoritmo para realizar essa tarefa. Explique suas escolhas em termos de evitar overfitting e justifique que essas escolhas irão levar a bons resultados

Primeiramente eu utilizaria um modelo de linguagem denso pré treinado, ou treinaria um do zero (utilizando CBOW ou Skip-Gram) a partir de um corpus representativo.
Assim, a partir do modelo de linguagem, iria converter as palavras de cada comentário em suas respectivas representações densas, que serão as entradas do modelo. Para esta tarefa eu utilizaria o modelo LSTM fim-a-fim (muitos para um), pois este modelo é capaz de integrar informações sobre o ordenamento das palavras, e evitaria que ocorresse o desaparecimento do gradiente. Dividiria os dados entre treinamento e teste, treinando o modelo com os dados de treinamento.

Para evitar o overfitting, inicialmente utilizaria somente uma camada de LSTM, caso ocorra underfitting, aumentaria para 2 ou 3 camadas. Realizaria também a operação de dropout e regularização dos pesos das nas células LSTMs para evitar que ocorra overfitting.

4. Suponha que você produziu, com o algoritmo Skip-Gram, vetores semânticos de palavras utilizando textos de artigos do Wikipedia. Agora você tem uma tarefa

específica, para a qual você tem um pequeno corpus, e você se depara com a seguinte questão:
a. Utilizar os vetores da forma como eles estão.
b. Re-treinar os vetores no corpus específico, mas ao invés de iniciar os vetores aleatoriamente, usa-se os vetores pré-treinados.
Qual a escolha correta? Justifique.

Opção **b.** Apesar de ser possível obter vetores com boas representações semânticas, a partir do corpus da Wikipedia, no corpus específico (por exemplo um texto técnico) podem existir termos ou combinações de termos que não existem na Wikipedia, mas que são representativos para a tarefa desejada.