Processamento de Linguagem Natural

Prova 1

Aluno: Leonardo Santos Miranda

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

- Diferente da representação one-hot que possui vetores esparsos, na representação distribuída os vetores são densos, ou seja, os vetores não dependem do tamanho do vocabulário.
- Baixa dimensionalidade (menor que no one-hot)
- Na representação distribuída os vetores são contínuos no espaço, ou seja, as palavras do vocabulário são representadas em um espaço de baixa dimensão D e operações como o cálculo da distância de similaridades entre palavras se tornam possíveis.

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.
R: Somente a alternativa c) está correta:

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
R: Classificar se um comentário é positivo ou negativo é um problema de análise de sentimento. Uma Rede Neural Convolucional para Textos é uma opção mas eu particularmente optaria por uma Rede Neural Recorrente LSTM por geralmente gerar resultados melhores se comparados a outros classificadores nesse tipo de problema (por possuir um alto poder de memorização), além de evitar o problema do gradiente desaparecer durante o treino no backpropagation caso os comentários sejam longos demais. Para evitar o overfitting, eu usaria as técnicas de normalização dos dados, adicionaria a técnica de dropout nas camadas e tentaria usar uma base de dados diferente da que o modelo foi anteriormente treinado (mas com o mesmo tema) como

treino para que o modelo possa generalizar o máximo possível os diferentes comentários (visto que os usuários podem ser informais, usar abreviações, etc)..

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.

R: A alternativa b) está correta: Apesar do corpus (contexto) do wikipédia ser muito grande e gerando por consequência ótimos vetores semânticos, quando falamos de um corpus específico não há a garantia de que o modelo conheça esse corpus. Por isso, o modelo deve ser treinado novamente para que tenha a certeza de que seja considerado possíveis peculiaridades que não tem no Wikipédia e possam possuir e serem úteis neste corpus em específico