Processamento de Linguagem Natural

Prova 1

Aluno(a): _____

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

1. Mais dados para o embedding, facilitando na comparação de similares dados entre as palavras
2. Representação Semantica mais completa
3. Possibilidade de analises ampliada para similaridade e correlações ao usar convoluções e auxiliar a rede neural com parametros e dados para classificações

- Densidade de informações
- Capacidade de generalização
- Redução de dimensionalidade
- Relações semanticas capturadas

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.

Primeiro utilizar o word2vec para gerar os embeddings;
Segundo concatenar os vetores dos embeddings em uma matriz de embedding

Terceiro Usar uma convnet com uma convolução 1D, uma camada de max pooling e uma camada Softmax para realizar a classificação entre positivo e negativo para o embedding de filmes
2 classes.
regularização e dropout

Essa escolha leva a bons resultados pois com a junção do word2vec e as convoluções é possível extrair informações dos embeddings onde as convoluções e o max pooling irão executar a analise e classificação pela similaridade entre os filmes para cada classe.

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.

Ao utilizar o Skip-gram, creio que a melhor abordagem é a letra b, uma vez que o corpus é pequeno o Skip-Gram pode ter dificuldades para lidar com um corpus pequeno, então vetores pré-treinados podem ajudar na tarefa

Correção GPT

3º Utilizar Word2Vec para gerar embeddings é uma ótima escolha, pois captura semântica e relações contextuais entre as palavras. Concatenar esses vetores em uma matriz cria uma representação rica do texto. O uso de uma ConvNet com convolução 1D é eficaz para extrair padrões locais nos dados, enquanto o MaxPooling ajuda a reduzir a dimensionalidade e destacar os recursos mais importantes. Por fim, a camada Softmax é adequada para classificação binária. Essa combinação deve ser capaz de evitar overfitting, especialmente com a aplicação adequada de regularização e dropout na rede neural.Quando se trata de um corpus pequeno, treinar vetores semânticos do zero pode levar a resultados fracos devido à falta de dados.

4º Usar vetores pré-treinados como ponto de partida para o treinamento com o Skip-gram no corpus específico pode aproveitar os benefícios dos dados maiores usados inicialmente, garantindo que as representações semânticas sejam mais precisas e adaptadas ao novo corpus. Isso geralmente resulta em melhores vetores de palavras para tarefas específicas.