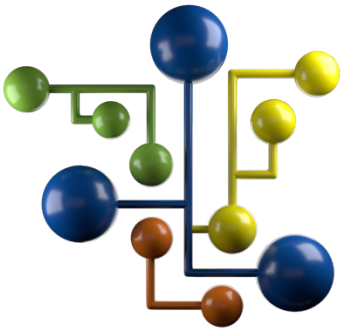


Data Science Academy

www.datascienceacademy.com.br

Introdução à Inteligência Artificial

Modelos de Linguagem

Em termos de processamento de linguagem natural, os modelos de linguagem geram cadeias de saída que ajudam a avaliar a probabilidade de um conjunto de strings ser uma sentença em uma linguagem específica. Se descartamos a sequência de palavras em todas as frases de um corpus de texto e basicamente o tratamos como um saco de palavras, então a eficiência de diferentes modelos de linguagem pode ser estimada pela precisão com que um modelo restaurou a ordem das strings nas frases. Qual frase é mais provável: *Eu estou aprendendo mineração de texto* ou *Eu texto mineração aprendizagem estou*? Qual palavra é mais provável de seguir Eu...?

Modelos de linguagem são amplamente utilizados em tradução automática, correção ortográfica, reconhecimento de fala, resumo de texto, questionários e assim por diante. Basicamente, um modelo de linguagem atribui a probabilidade de uma sentença estar em uma ordem correta. A probabilidade é atribuída ao longo da sequência de termos usando a probabilidade condicional. Vamos definir um problema de modelagem de linguagem simples. Suponha que um saco de palavras contém palavras W1, W2, ..., Wn.. Um modelo de linguagem pode ser definido para calcular qualquer um dos seguintes:

Estimar a probabilidade de uma frase S1:
P (S1) = P (W1, W2, W3, W4, W5)

Estimar a probabilidade da próxima palavra em uma frase ou conjunto de strings:
P(W3|W2, W1)

Como calcular a probabilidade? Usaremos a regra da cadeia, ao decompor a probabilidade da sentença como um produto de probabilidades de strings menores:

P(W1W2W3W4) = P(W1) P(W2 | W1) P(W3 | W1W2) P(W4 | W1W2W3)

Modelos N-gramas

N-gramas são usados em uma ampla gama de aplicações. Eles podem ser usados para construir modelos de linguagem simples. Consideremos um texto T com tokens W. Seja SW uma janela deslizante. Se a janela deslizante consiste em uma célula, então a coleção de strings é chamada de unigrama. Se a janela deslizante consiste de duas células, a saída é,

$$(w_1, w_2)(w_3, w_4)(w_5 w_5, w_5)(w_1, w_2)(w_3, w_4)(w_5, w_5)$$

isso é chamado de bigrama. Usando probabilidade condicional, podemos definir a probabilidade de uma palavra ter visto a palavra anterior. Isso é conhecido como probabilidade bigrama. Assim, a probabilidade condicional de um elemento, dado o elemento anterior, $(w_i\text{-}1)$ é $P(w_i|w_{i-1})$.

Estendendo a janela deslizante, podemos generalizar essa probabilidade n-grama como a probabilidade condicional de um elemento dado elemento n-1 anterior:

$$P(w_i|w_{i-n-1} \ldots w_{i-1})$$

A fim de obter bigramas mais significativo, podemos executar o corpus através de um part-of-speech (POS) tagger. Isso filtraria os bigramas em pares mais relacionados ao conteúdo, como desenvolvimento de infraestrutura, subsídios agrícolas, taxas bancárias; Esta pode ser uma maneira de filtrar bigramas menos significativa.

Uma maneira melhor de abordar esse problema é levar em conta as colocações (collocations). Uma colocação é a sequência criada quando duas ou mais palavras co-ocorrem em uma linguagem com mais frequência. Uma maneira de fazer isso em um corpus é a informação mútua pontual (PMI). O conceito por trás do PMI é para duas palavras, A e B, gostaríamos de saber o quanto uma palavra nos diz sobre a outra. Por exemplo, dada uma ocorrência de “A, a”, e uma ocorrência de “B, b”, quanto a sua probabilidade conjunta difere do valor esperado de assumir que eles são independentes. Isto pode ser expresso como segue:

$$PMI(a,b) = \ln \frac{P(a,b)}{P(a)P(b)} PMI(a,b) = \ln \frac{P(a,b)}{P(a)P(b)}$$

Modelo Unigrama:

Punigram(W1W2W3W4) = P(W1) P(W2) P(W3) P(W4)

Modelo Bigrama:

Pbu(W1W2W3W4) = P(W1) P(W2 | W1) P(W3 | W2) P(W4 | W3) P(w1w2… wn) = P(wi | w1w2… wi" 1)

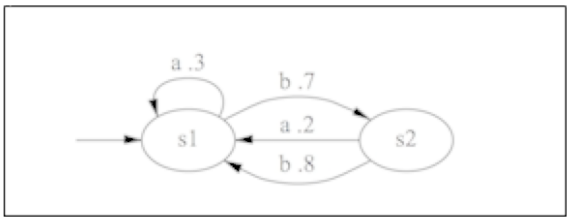
A aplicação da regra da cadeia em n contextos pode ser difícil de estimar; A suposição de Markov é aplicada para lidar com tais situações.

Suposição (Hipótese) de Markov

Se a previsão de que uma sequência atual é independente de alguma sequência de palavras no passado, podemos desconsiderar essa sequência para simplificar a probabilidade. Digamos que a história consiste em três palavras, Wi, Wi-1, Wi-2, em vez de estimar a probabilidade P (Wi + 1) usando P (Wi, i-1, i-2), podemos aplicar diretamente P (Wi + 1 | Wi, Wi-1).

Modelos Ocultos de Markov

Cadeias de Markov são usadas para estudar sistemas que estão sujeitos a influências aleatórias. As cadeias de Markov modelam sistemas que se movem de um estado para outro em etapas governadas por probabilidades. O mesmo conjunto de resultados em uma sequência de ensaios é chamado estados. Conhecer as probabilidades dos estados é chamado de distribuição de estado. A distribuição de estado em que o sistema inicia é a distribuição de estado inicial. A probabilidade de ir de um estado para outro é chamada probabilidade de transição. Uma cadeia de Markov consiste em uma coleção de estados junto com probabilidades de transição. O estudo das cadeias de Markov é útil para entender o comportamento a longo prazo de um sistema. Cada arco se associa a determinado valor de probabilidade e todos os arcos que saem de cada nó devem ter uma distribuição de probabilidade. Em termos simples, há uma probabilidade associada a cada transição nos estados:



Os modelos ocultos de Markov são cadeias de Markov não determinísticas. Eles são uma extensão dos modelos de Markov em que o símbolo de saída não é o mesmo que o estado. Os modelos ocultos de Markov são frequentemente usados em atividades de tradução automática de idiomas, através de processamento de linguagem natural.

Referências:

Livro: Inteligência Artificial
Autor: Peter Norvig