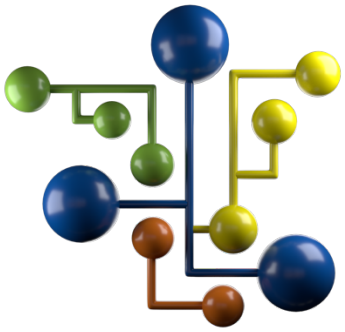

Data Science Academy

www.datascienceacademy.com.br

Processamento de Linguagem Natural

Estimando as Probabilidades em Modelos N-gramas

Modelos n-gramas

Um n-grama é simplesmente uma sequência de tokens. No contexto da linguística computacional, esses tokens são geralmente palavras, embora possam ser caracteres ou subconjuntos de caracteres. O n simplesmente se refere ao número de tokens.

Se estivermos contando palavras, a string "Amanhã vai chover forte" é um 4-gramas. Este 4-gramas contém os 3 gramas “Amanhã vai chover” e “vai chover forte”. Um unigrama é um único token, por exemplo, “Amanhã”.

Esses tokens não precisam ser palavras ou caracteres. Podem ser códigos de DNA, números binários ou potencialmente qualquer tipo de dados sequenciais imagináveis. Eles geralmente são usados para capturar informações estatísticas de alguns conjuntos de dados e são muito eficazes ao fazê-lo, muitas vezes superando as abordagens mais complexas.

Mais precisamente, podemos usar modelos n-gramas para derivar uma probabilidade da sentença, W, como a probabilidade conjunta de cada palavra individual na sentença, wi.

$$P(W) = P(w1, w2, ..., wn)$$

Isso pode ser reduzido a uma sequência de n-gramas usando a Regra de Cadeia (Chain Rule) de probabilidade condicional.

$$P(x1, x2, ..., xn) = P(x1)P(x2|x1)...P(xn|x1,...xn-1)$$

Como um exemplo concreto, vamos prever a probabilidade da sentença. Havia forte chuva ontem.

P(‘Havia forte chuva ontem’) = P('Havia', 'forte', 'chuva', 'ontem')

P(‘Havia forte chuva ontem’) = P('Havia')P('forte'|'Havia')P('chuva'|'Havia forte')P('ontem'|'Havia forte chuva')

Cada um dos termos do lado direito desta equação são probabilidades de n-grama que podemos estimar usando as contagens de n-gramas em nosso corpus. Para calcular a probabilidade de toda a frase, precisamos procurar as probabilidades de cada componente na probabilidade condicional.

Infelizmente, esta fórmula não escala, pois não podemos calcular n-gramas de cada comprimento. Por exemplo, considere o caso em que temos apenas bigramas em nosso modelo; nós não temos como saber a probabilidade de 'P' ('chuva' | 'Havia forte ') de bigramas.

Ao usar a Suposição de Markov, podemos simplificar nossa equação assumindo que os estados futuros em nosso modelo dependem apenas do estado atual do nosso modelo. Essa

suposição significa que podemos reduzir nossas probabilidades condicionais de ser aproximadamente iguais para que:

P('ontem'|'Havia forte chuva') ~ P('ontem'|'chuva')

Assim, podemos estimar a probabilidade de uma sentença pelas probabilidades de cada componente.

Para que podemos usar modelos n-grama? Dadas as probabilidades de uma frase, podemos determinar a probabilidade de uma tradução automática automatizada estar correta, podemos prever que a próxima palavra mais provável ocorrerá em uma frase, podemos gerar automaticamente texto da fala, automatizar correção ortográfica ou determinar o sentimento relativo de um texto.

Um exemplo de problema que necessita de inferência estatística é a previsão da palavra seguinte em uma frase, dadas as palavras anteriores. Uma sequência de palavras pode começar de uma maneira conhecida, mas terminar por uma palavra desconhecida.

Um modo de agrupar todas as sequências de tamanho n que começam pelas mesmas n - 1 palavras em uma classe de equivalência é supor que o contexto local prévio afeta a palavra seguinte, e construir o modelo de Markov de ordem (n - 1) ou modelo de n-gramas (sendo a última palavra do n-grama a que está sendo prevista). Os casos de n-gramas mais utilizados são com n = 2, 3 e 4, particularmente denominados bigramas, trigramas e tetragramas.

Quanto maior o valor de n, isto é, maior o número de classes que dividem os dados, maior a confiabilidade da inferência. No entanto, o número de parâmetros a serem estimados cresce exponencialmente em relação a n. Por isso, geralmente são utilizados bigramas ou trigramas em sistemas dessa natureza.

Estimadores estatísticos

Tendo-se os dados de treinamento já divididos em classes de equivalência, o passo seguinte é descobrir, para os dados de cada classe, como derivar uma boa estimativa de probabilidade para uma característica, com base nestes dados. No exemplo do modelo de n-gramas, deseja-se conhecer a probabilidade de ocorrência do n-grama w1, ..., wn, notada como P (w1, ..., wn), e prever a probabilidade de ocorrência da palavra wn dada a ocorrência da sequência de palavras w1, ..., wn1 , notada como P (wn | w1, ..., wn1). Alguns dos métodos de estimação existentes são: Maximum Likelihood Estimation, leis de Laplace, Lidstone e Jeffreys-Perks, Held Out Estimation, Cross-Validation e Good-Turing Estimation.