Processamento de Linguagem Natural

Modelos de Linguagem

Marlo Souza[1]
[1]Universidade Federal da Bahia - Brasil

29 de julho de 2024

Sumário

Modelos de Linguagem Probabilísticos

O que é PLM

Modelos de Linguagem, ou modelos probabilísticos de linguagens, são modelos estatísticos que associam uma probabilidade à ocorrência de uma palavra, dado um contexto, ou de forma mais geral, a probabilidade de ocorrência de uma sequência de palavras na língua.

O que é PLM

Modelos de Linguagem, ou modelos probabilísticos de linguagens, são modelos estatísticos que associam uma probabilidade à ocorrência de uma palavra, dado um contexto, ou de forma mais geral, a probabilidade de ocorrência de uma sequência de palavras na língua.

- Qual a palavra mais provável de suceder a seguinte cadeia: “Por favor, após sair desliguem a ”

O que é PLM

Modelos de Linguagem, ou modelos probabilísticos de linguagens, são modelos estatísticos que associam uma probabilidade à ocorrência de uma palavra, dado um contexto, ou de forma mais geral, a probabilidade de ocorrência de uma sequência de palavras na língua.

- Qual a palavra mais provável de suceder a seguinte cadeia: “Por favor, após sair desliguem a ”
- Qual a palavra melhor se adequa ao contexto: “Xoxa, capenga, [MASK], anêmica, frágil e inconsistente”

Aplicações

Modelos de Linguagem possuem uma ampla aplicação a problemas práticos de PLN pois aspectos distribucionais da linguagem codificam muita informação sobre a sua estrutura.

Aplicações

Modelos de Linguagem possuem uma ampla aplicação a problemas práticos de PLN pois aspectos distribucionais da linguagem codificam muita informação sobre a sua estrutura.

- Tradução Automática: $P(\textbf{high}\ \text{winds tonite}) > P(\textbf{large}\ \text{winds tonite})$

Aplicações

Modelos de Linguagem possuem uma ampla aplicação a problemas práticos de PLN pois aspectos distribucionais da linguagem codificam muita informação sobre a sua estrutura.

- Tradução Automática: $P(\textbf{high}\text{ winds tonite}) > P(\textbf{large}\text{ winds tonite})$
- Correção ortográfica: "Estamos a cinco minuots da estação" $\rightarrow$ "Estamos a cinco **minutos** da estação"

Aplicações

Modelos de Linguagem possuem uma ampla aplicação a problemas práticos de PLN pois aspectos distribucionais da linguagem codificam muita informação sobre a sua estrutura.

- Tradução Automática: $P(\textbf{high}\ \text{winds tonite}) > P(\textbf{large}\ \text{winds tonite})$
- Correção ortográfica: "Estamos a cinco minuots da estação" $\rightarrow$ "Estamos a cinco **minutos** da estação"
- Inúmeras outras aplicações

Como fazemos um PLM

6 O objetivo de um PLN é calcular a probabilidade de uma sentença ou sequência de palavras.

$$Pr(W) = Pr(w_1, w_2, \ldots, w_n)$$

Como fazemos um PLM

6

O objetivo de um PLN é calcular a probabilidade de uma sentença ou sequência de palavras.

$$Pr(W) = Pr(w_1, w_2, \ldots, w_n)$$

Pela definição de probabilidade condicional, temos que

$$P(A, B) = P(A)P(B|A)$$

Como fazemos um PLM

O objetivo de um PLN é calcular a probabilidade de uma sentença ou sequência de palavras.

$$Pr(W) = Pr(w_1, w_2, \ldots, w_n)$$

Pela definição de probabilidade condicional, temos que

$$P(A, B) = P(A)P(B|A)$$

De forma geral:

$$P(x_1, x_2, x_3, \ldots, x_n) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2)\ldots P(x_n|x_1, \ldots, x_{n-1})$$

Então podemos usar a regra da cadeia para realizar a computação.

Cadeias de Markov

Cadeias de Markov são modelos estocásticos em que a probabilidade de um evento futuro depende somente do estado atual, ou seja, assumindo a condição de Markov, poderíamos reduzir o problema de calcular $P(x_1, x_2, x_3, \ldots, x_n)$ a:

$$P(x_1, x_2, x_3, \ldots, x_n) = P(x_1)P(x_2|x_1)P(x_3|x_2)\ldots P(x_n|x_{n-1})$$

Esse modelo é também chamado de modelo de bigramas.

Cadeias de Markov

Cadeias de Markov são modelos estocásticos em que a probabilidade de um evento futuro depende somente do estado atual, ou seja, assumindo a condição de Markov, poderíamos reduzir o problema de calcular $P(x_1, x_2, x_3, \ldots, x_n)$ a:

$$P(x_1, x_2, x_3, \ldots, x_n) = P(x_1)P(x_2|x_1)P(x_3|x_2)\ldots P(x_n|x_{n-1})$$

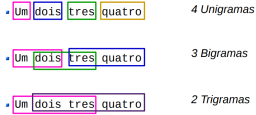
Esse modelo é também chamado de modelo de bigramas.
Modelos de bigramas são muito pobres para codificar relações sintáticas longas existentes no texto. Podemos generalizar a condição de Markov para que o resultado do processo estocástico depende de até n estados anteriores, ao qual damos o nomes de modelo de n-grama

Modelos de N-gramas

Um *n*-grama é uma sequência (contígua ou não) de *n* tokens observados no texto.

Modelos de N-gramas

Um *n*-grama é uma sequência (contígua ou não) de *n* tokens observados no texto.



Modelos de N-gramas

Um modelo de n-gramas pode ser calculado estimando as probabilidades $P(x_1, \cdots, x_n)$ e $P(x_n|x_1, \cdots, x_{n-1})$ ao maximizar a verossimilhança estimada do modelo a partir dos dados.

Modelos de N-gramas

Um modelo de n-gramas pode ser calculado estimando as probabilidades $P(x_1, \cdots, x_n)$ e $P(x_n|x_1, \cdots, x_{n-1})$ ao maximizar a verossimilhança estimada do modelo a partir dos dados.

$$P(w_i|w_{i-1}) = \frac{freq(w_{i-1}, w_i)}{freq(w_{i-1})}$$

Problemas:

- Contexto pode ser muito grande

Modelos de N-gramas

Um modelo de n-gramas pode ser calculado estimando as probabilidades $P(x_1, \cdots, x_n)$ e $P(x_n|x_1, \cdots, x_{n-1})$ ao maximizar a verossimilhança estimada do modelo a partir dos dados.

$$P(w_i|w_{i-1}) = \frac{freq(w_{i-1}, w_i)}{freq(w_{i-1})}$$

Problemas:

- Contexto pode ser muito grande
- Complexidade Computacional
- Escassez de dados

Google N-gram Viewer

https://books.google.com/ngrams/

Vamos para o Colab!

Avaliação de PLM

As métricas de classificação não se adéquam à tarefa de modelagem de linguagem, portanto, usualmente, empregamos a métrica de **perplexidade**.

Avaliação de PLM

As métricas de classificação não se adéquam à tarefa de modelagem de linguagem, portanto, usualmente, empregamos a métrica de **perplexidade**. A perplexidade de um modelo sobre um corpus de teste é dado por:

$$\left(\prod_{i=1}^{n} q(s_i)\right)^{-1/N}$$

com s_i as sentenças do corpus de teste e N o número de tokens no corpus.

Avaliação de PLM

As métricas de classificação não se adéquam à tarefa de modelagem de linguagem, portanto, usualmente, empregamos a métrica de **perplexidade**. A perplexidade de um modelo sobre um corpus de teste é dado por:

$$\left(\prod_{i=1}^{n} q(s_i)\right)^{-1/N}$$

com s_i as sentenças do corpus de teste e N o número de tokens no corpus. Minimizar a perplexidade corresponde, então, a maximizar a probabilidade de ocorrência das sentenças reais do corpus.

Aplicação: Geração de texto

Vamos para o Colab!