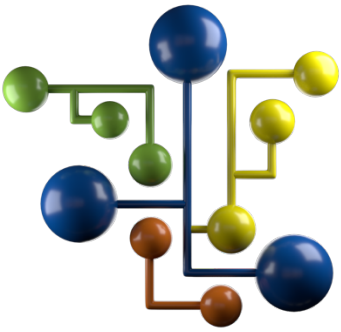

Data Science Academy

www.datascienceacademy.com.br

Processamento de Linguagem Natural

Modelagem Probabilística da Linguagem

Discutiremos um dos modelos probabilísticos mais famosos da PLN, que tem sido usado para uma variedade de aplicações - o modelo de linguagem. Teremos uma ideia intuitiva sobre como o modelo de linguagem funciona e onde podemos usá-lo.

Existem dois objetivos básicos do modelo de linguagem (LM – Language Model):

1. O objetivo do LM é atribuir probabilidade a uma sentença ou sequência de palavras.
2. LM também nos fala sobre a probabilidade da próxima palavra, o que significa que indica qual é a próxima palavra mais provável, observando a sequência de palavras anterior.

Se qualquer modelo pode calcular qualquer uma das tarefas anteriores, ele é chamado de modelo de linguagem. O LM usa a regra da cadeia de probabilidade condicional (chain rule). A regra da cadeia de probabilidade condicional é apenas uma extensão da probabilidade condicional. Nós já vimos a equação:

$$P(A|B) = P(A \text{ and } B) / P(B)$$

$$P(A \text{ and } B) = P(A,B) = P(A|B)\, P(B)$$

Aqui, P (A, B) é chamado de probabilidade conjunta. Suponha que você tenha vários eventos que são dependentes e, em seguida, a equação para calcular a probabilidade de junção se torna mais geral:

$$P(A, B, C, D) = P(A)\, P(B|A)\, P(C|A, B)\, P(D|A, B, C)$$

$$P(x1, x2, x3,..., xn) = P(x1)\, P(x2 \mid x1)\, P(x3 \mid x1, x2)... P(xn \mid x1, x2, x3,... xn-1)$$

A equação anterior é chamada de regra de cadeia para probabilidade condicional. LM usa essa regra para prever a probabilidade das próximas palavras. Costumamos calcular a probabilidade contando o número de vezes que um evento particular ocorre e dividindo-o pelo número total de combinações

possíveis, mas não podemos aplicá-lo à linguagem porque, com certas palavras, você pode gerar milhões de sentenças. Então, não vamos usar a equação de probabilidade. Estamos usando uma hipótese chamada ***Markov Assumption*** para calcular a probabilidade. Vamos entender o conceito intuitivamente antes de olhar para uma definição técnica dele. Se você tem uma frase muito longa e está tentando prever qual será a próxima palavra na sequência da sentença, você realmente precisa considerar todas as palavras que já estão presentes na sentença para calcular a probabilidade da próxima palavra. Este cálculo é muito entediante, então consideramos apenas a última, as duas últimas ou três últimas palavras para calcular a probabilidade da próxima palavra; isso é chamado de suposição de Markov. A suposição é que você pode calcular a probabilidade da próxima palavra que vem em uma sequência da sentença, olhando para as duas últimas palavras por exemplo.

Um LM simples usa um unigrama, o que significa que estamos apenas considerando a própria palavra e calculando a probabilidade de uma palavra individual; você simplesmente pega a probabilidade de palavras individuais e gera uma sequência aleatória de palavras. Se você pegar um modelo bigrama, então você considera que uma palavra anterior decidirá a próxima palavra na sequência. Você pode ver o resultado do modelo bigrama abaixo:

$$P(w_i \mid w_1 w_2 \ldots w_{i-1}) \approx P(w_i \mid w_{i-1})$$

Como podemos contar a probabilidade n-gram que é uma parte essencial do LM? Vamos dar uma olhada no modelo do bigrama. Vamos ver a equação e depois passar pelo exemplo. Veja a equação abaixo:

$$P(w_i \mid w_{i-1}) = \frac{count(w_{i-1}, w_i)}{count(w_{i-1})}$$

A equação é fácil de compreender. Precisamos calcular quantas vezes as palavras wi-1 e wi ocorrem juntas e quantas vezes wi-1 ocorre. Veja o exemplo abaixo:

$$P(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i)}{c(w_{i-1})}$$

<s> I am Sam </s>
<s> Sam I am </s>
<s> I do not like green eggs and ham </s>

$P(\texttt{I} \mid \texttt{<s>}) = \frac{2}{3} = .67$ $P(\texttt{Sam} \mid \texttt{<s>}) = \frac{1}{3} = .33$ $P(\texttt{am} \mid \texttt{I}) = \frac{2}{3} = .67$

$P(\texttt{</s>} \mid \texttt{Sam}) = \frac{1}{2} = 0.5$ $P(\texttt{Sam} \mid \texttt{am}) = \frac{1}{2} = .5$ $P(\texttt{do} \mid \texttt{I}) = \frac{1}{3} = .33$

Usando LM, nós sabemos como os pares de palavras são descritos no Corpus, assim como qual é o par mais popular. E se usarmos um four-grama ou five-grama, podemos ter resultados ainda melhores, pois algumas sentenças têm longas dependências na sintaxe entre verbos e predicados.

O LM possui várias aplicações em PLN: Sistemas de tradução automática usam LM para descobrir a probabilidade de cada uma das sentenças traduzidas para decidir qual frase traduzida é a melhor tradução possível para a sentença de entrada. Para soletrar a aplicação correta, podemos usar um bigrama LM para fornecer a sugestão de palavra mais provável. Podemos usar LM para sumarização de texto. Podemos usar LM em um sistema de perguntas para classificar as respostas conforme sua probabilidade.