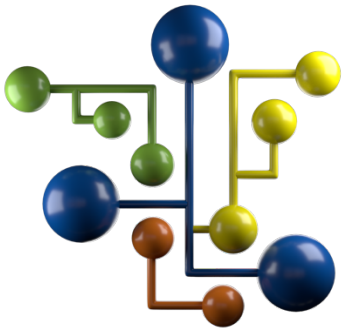

Data Science Academy

www.datascienceacademy.com.br

Processamento de Linguagem Natural

Técnicas de Smoothing

Como algumas estruturas de palavras menos frequentes podem não aparecer no corpus de treinamento, são necessários métodos que estimem a sua probabilidade. Estes métodos associam valores não nulos às probabilidades dos eventos não encontrados, que seria considerada zero. Para isso, diminui-se a probabilidade dos eventos encontrados, para que reste uma fatia de probabilidade para aqueles eventos não encontrados. A técnica que prevê esse desconto de probabilidades é referida como smoothing. Vejamos uma breve descrição das principais técnicas de smoothing.

Lei de Laplace

A técnica de smoothing baseada na lei de Laplace tem como efeito reservar uma pequena porção do espaço de probabilidade para os eventos não conhecidos. Para isso, adiciona 1 à frequência de cada n-grama encontrado no corpus de treinamento.

$$P_{LAP}(w_1, ..., w_n) = \frac{C(w_1, ..., w_2) + 1}{N + B}$$

onde C(w1, ..., wn) é a frequência do n-grama w1, ..., wn e B é o número de classes em que as instâncias de treinamento estão divididas.

Pode-se notar que a lei de Laplace considera o tamanho do vocabulário N. Em conjuntos de dados muito esparsos, como os das aplicações de PLN, esta técnica destina muito do espaço de probabilidade para eventos não encontrados no corpus de treinamento, subestimando a probabilidade dos eventos observados.

Estimação Held Out

Esta técnica divide os dados de treinamento em duas partes: uma para obtenção das frequências de ocorrência dos n-gramas, e outra, held out data, também denominados dados de validação, para estimação da probabilidade de n-gramas desconhecidos.

Para cada n-grama w1, ..., wn, deve ser computada a sua frequência nos dados de treinamento, C1(w1, ..., wn), e sua frequência nos dados held out, C2(w1, ..., wn). Calcula-se Nr, que corresponde à quantidade de n-gramas cuja frequência nos dados de treinamento é r. Então, calcula-se Tr , que corresponde a soma das frequências dos n-gramas, cuja frequência nos dados de treinamento é r, nos dados held out. A estimativa de probabilidade de um n-grama é:

$$P_{ho}(w_1, ..., w_n) = \frac{T_r}{N_r N}$$