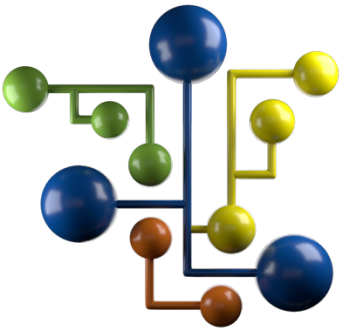

Data Science Academy

www.datascienceacademy.com.br

Processamento de Linguagem Natural

Processo de Derivação de Tokens

As frases são formadas por fluxos de palavras e, a partir de uma frase, precisamos derivar pedaços significativos individuais que são chamados tokens e o processo de derivação de token é chamado de Tokenização.

O processo de derivação de tokens de um fluxo de texto tem duas etapas. Se você tiver muitos parágrafos, então primeiro você precisa fazer tokenização de sentença, para na sequência fazer tokenização de palavras e gerar o significado dos tokens.

Tokenização e lemmatização são processos que são úteis para análise léxica. Tokenização pode ser definida como a identificação do limite de frases ou palavras. A lemmatização pode ser definida como um processo que identifica o POS (Part-of-Speech) pretendido correto e o significado das palavras que estão presentes em frases. A lemmatização também inclui marcação POS (POS-Tagging) para desambiguar o significado dos tokens. Neste processo, a janela de contexto é em nível de frase ou nível de sentença.

Diferença entre Stemming e Lemmatization

Stemming e Lemmatization são conceitos usados para normalizar a palavra dada, removendo sufixos e considerando seu significado. Veja a tabela abaixo para uma comparação entre essas duas técnicas:

Stemming	Lemmatization
Stemming geralmente opera em uma única palavra sem conhecimento do contexto.	Lemmatization geralmente considera palavras e o contexto da palavra na frase.
Em Stemming nós não consideramos POS tags.	Em Lemmatization nós consideramos POS tags.
Stemming é usado para agrupar palavras que tenham um significado similar.	Lemmatization é usado para criar dicionários ou WordNets.

É demorado construir um dicionário lexical. Se você quer construir uma ferramenta de Lemmatização que pode considerar um contexto maior, levando em consideração o contexto das frases precedentes, saiba que esta ainda é uma área ativa de pesquisa.