Resumo 3 – PLN

A terceira aula de Processamento de Linguagem Natural, foram apresentados conceitos referentes a normalização de textos. Iniciando a aula com uma rápida atividade de 4 questões, o professor Jesus Mena introduziu o tema com definições sobre *stopwords, stoplists* e normalização de palavras (stemmer).

Em seguida, continuando a matéria a partir da segunda parte do capítulo 2 do livro Speech and Language Processing de Dan Jurafsky e James H. Martin, a aula teve sequência com um breve resumo da aula anterior sobre expressões regulares. O conceito de “grupos” foi passado como forma de demonstrar diferentes formas de se aplicar expressões regulares para a identificação de trechos textuais, inclusive aqueles com características linguísticas diferentes do inglês, como a presença de acentuação, cedilha, entre outros.

A tokenização de palavras, assunto de uma das questões do teste inicial aplicado na aula, foi o tema abordado em sequência. Considerando a quantidade de palavras (tokens) e a quantidade de palavras diferentes presentes em determinado trecho textual (types), a denominação de corpus foi apresentada como uma prática de aprendizado (análise) e validação (verificação).

Em um exemplo prático, o livro A Semana, de Machado de Assis, foi alvo de uma análise em Python utilizando ferramentas de expressões regulares e tokenização. Como resultado, foi possível realizar um levantamento das palavras mais comuns presentes no texto e, em conclusão, foi visto que tais palavras, muitas vezes, não representavam significado algum para um algoritmo de aprendizado, ou seja, muitas das palavras mais comuns resultantes foram preposições, artigos e demais grupos que, sozinhos ou em contextos, não causam alterações significativas no entendimento da frase e/ou são irrelevantes. Tais palavras são conhecidas como stopwords. Um conjunto de stopwords é conhecido como stoplist. Tratas as stopwords é um procedimento necessário durante o processamento de linguagem natural.

Na sequência, o processo de normalização de palavras foi exemplificado de forma a transformar termos e manter seus radicais. Existem técnicas específicas para normalização: Stemming e Lemmatization. Dentro dos algoritmos de Stemming, existem Lovins Stemmer, Porter Stemmer, Paice Stemmer, cada qual com sua respectiva particularidade. Durante o curso, iremos trabalhar com Porter Stemmer.