Resumo 4 – PLN

A quarta aula de Processamento de Linguagem Natural foi iniciada com um breve resumo referente a terceira aula e introduziu o tema de normalização de palavras. Em inglês, português, ou qualquer idioma, as palavras são categorizadas de acordo com as regras linguísticas caraterísticas de cada dialeto. Em todos os casos, a normalização pode ser entendida como o processo de redução, simplificação ou radicalização das palavras. Existem técnicas específicas para a realização da normalização, sendo as principais conhecidas como *stemming* e *lemmatization.*

Após o link com o final da aula de número três, a aula quatro teve sequência detalhando destas duas principais abordagens responsáveis pela normalização de palavras. Stemming geralmente refere-se a um processo de heurística que corta as extremidades das palavras inclui frequentemente a remoção de afixos derivacionais. Já o processo de Lemmatization, por sua vez, pode ser considerado mais complexo que o Stemming (e também mais eficiente), visto que busca reduzir as palavras em “Lemmas”, ou seja, a forma básica da palavra e não em apenas uma parte da palavra, como no Stemming.

Entrando mais a fundo do algoritmo de Stemming, é possível encontrar algumas variações idealizadas e implementadas por estudiosos ao longo dos anos. Um deles, desenvolvido em 1968 por Julie Beth Lovins é conhecido como Lovins Stemmer e dado como o pioneiro e influenciador de muitos stemmer.

O algoritmo de Lovins Stemmer é composto por 294 sufixos, 29 condições e 34 regras de transformação e possui apenas duas etapas de processamento. Primeiro, deve-se procurar e remover o sufixo de maior tamanho contido na palavra. Para isso, existe uma lista de 294 sufixos ordenados pela quantidade e letras. A segunda etapa aplica-se na transformação do final do palavra e é utilizada se um sufixo é removido ou não na primeira etapa.

Com essa explicação, foi possível realizar a atividade proposta pelo professor para identificar o radical de cada palavra utilizando o algoritmo de Lovins Stemmer como base. Nela, foram propostas algumas palavras que deveriam ter seus sufixos e regras, se aplicáveis, identificados e transformados de acordo com as regras propostas. O objetivo foi presenciar, na prática, diferentes situações de normalização com palavras da língua inglesa.

Logo após a atividade, o algoritmo de Stemmer proposto por Martin F. Porter foi detalhado. Porter Stemmer, como é conhecido, foi inicialmente publicado em um relatório de projeto final de Recuperação de Informação. É considerado mais completo e simples que o Lovins Stemmer.

Para seu entendimento, o professor Jesus mostrou aos alunos o código em Python utilizado para a implementação do Porter Stemmer. Curioso pontuar que o algoritmo é formado por uma série de condicionais.

Até o momento, os alunos tinham em mente algoritmos responsáveis pela normalização de palavras que foram desenvolvidos levando em consideração o idioma inglês. Quando o português é o espaço amostral, existem uma série de regras diferentes que precisam ser aplicadas de acordo com as particularidades proporcionadas pelo idioma. Considerando essas dificuldades, Viviane Moreira Orengo propôs, em 2001, uma versão de um algoritmo de radicalização para a língua portuguesa. Seus detalhes foram evidenciados em aula.

Logo em seguida, alguns temas a respeito do significado das palavras de modo a contemplar o conceito das mesmas na normalização. Outras abordagens para stemming também foram pontuadas em aula, como por exemplo, a utilização de redes (grafos) de palavras.