Resumo 8 – PLN

A oitava aula de Processamento de Linguagem Natural foi iniciada com uma atividade, passada pelo professor Jesús, contemplando três questões a respeito do conteúdo a ser ministrado na aula. Nela, foi possível retirar alguns insights sobre similaridade entre palavras, significado de homógrafo, homófono e tesauro.

Assim, tendo em mente uma introdução do tema da aula, iniciamos a discussão sobre similaridade de palavras e algumas aplicações práticos disso, como na biologia em cadeias de ácidos nucleicos. Um ponto essencial para a definição desse conceito é a distância entre as palavras que, apesar de parecer abstrato no início, se faz um conceito extremamente claro com as explicações que se sucederam. A medida de distância de Hamming foi introduzida como forma de realizar esse cálculo, mostrando então, como exemplo, as diferenças entre “PLN” e “PNL”, “UFABC” e “UFRJ”, “UFABC e UFRRJ”. O resultado para os respectivos exemplos foram 1, -1 e 3, indicando assim que esse algoritmo para o cálculo da distância não leva em consideração palavras trocadas e outras particularidades.

Outro algoritmo que pode ser utilizado é a distância de Levenshtein em que, de uma forma mais robusta, consegue calcular uma espécie de diferença entre as palavras. O código responsável por definir o algoritmo de Levenshtein foi explicado pelo professor que, por sua vez, também pontuou sua principal desvantagem: custo computacional. Por definição, esse algoritmo trabalha com chamadas recursivas para realizar comparações de palavras e sucessivas alterações nas mesmas.

Para evitar esse aumento do custo computacional, pode ser utilizada uma tática de programação dinâmica (memoização), utilizando alguma estrutura auxiliar, como matrizes e vetores, para aumentar a velocidade de execução do algoritmo. Para ilustrar essa aplicação, foi pontuado o exemplo do cálculo de Fibonacci utilizando matrizes como estruturas auxiliares, evitando assim a recursão excessiva.

Em uma segunda etapa da aula, iniciou-se o debate sobre semântica e similaridade de palavras sob o macro tema de casamento aproximado entre strings. Ao invés de analisar palavras na ótica de troca de letras ou mudanças de posição, vamos agora aplicar os conhecimentos de PLN para aplicar questões de linguística de modo a identificar se realmente uma palavra é próxima a outra no universo semântico.

O primeiro conceito para o entendimento dessa similaridade entre palavras é o “Lemma”, ou seja, significado das palavras sem inflexão. Em seguida, o conceito de “Homônimos” foi apresentado como uma forma de dificultar nossa vida em PLN. Exemplo: a frase “banco quebrado” indica a instituição ou o assento? Um outro conceito apresentado foi o de Polissemia, seguindo com os exemplos onde mais de um significado pode ser encontrado para a mesma palavra: letra, vela.

Um teste a ser realizado a fim de identificar se uma palavra possui mais de um significado é o teste de Zeugma. Nele, dois contextos contendo a mesma utilização da palavra são comparados de modo a identificar se existe algum sentido nessa comparação. O exemplo utilizado em aula comparava a utilização da palavra “universidade” nas seguintes frases: “...construirá uma universidade de mármore” e “...pedirá à universidade de João”. Os tempos “universidade de mármore” e “universidade de João” podem não fazer sentido, indicando assim que o termo “universidade” foi utilizado e possui mais de um contexto.

Em seguida, os conceitos de sinônimos e antônimos foram passados como parte do conteúdo teórico. Por fim, temos a Hiponímia e Hiperonímia que indicam relação hierárquica de significados entre a palavras.

A definição de tesauro foi dada como sendo um dicionário de ideias comuns. Uma definição mais formal indica que o tesauro é um acervo ordenado de termos e conceitos relacionados entre si. Um dos tesauros mais comuns é o Wordnet.

Por que é importante avaliar a similaridade? A resposta para essa pergunta pode ser dada em três tópicos diferentes. O primeiro deles, para recuperação de informação. O segundo, para detecção de plágio. O terceiro, para agrupamento de texto.

A partir da criação de um agrupamento hierárquico, é possível assimilar conclusões sobre similaridades de palavras e a criação de um tesauros.