Resumo 10 – PLN

A décima aula de Processamento de Linguagem Natural foi iniciada com uma breve revisão da aula anterior, contemplando temas como a matriz de co-ocorrência onde era possível olhar linhas (como forma de verificar similaridade entre documentos) e colunas (como forma de verificar similaridade entre palavras). Essa matriz é chamada de matriz esparsa, uma vez que possui muitos elementos nulos.

Até o momento, vimos a matriz termo-documento. O mais natural seria considerar as definições de matriz termo-termo, matriz palavra-palavra ou matriz termo-contexto. Cada célula da matriz registra o número de vezes que a palavra (da linha) co-ocorre com outra palavra (da coluna) em um contexto, dado em um corpus de treinamento. E assim, temos a matriz-contexto. Analisando apenas a vizinhança do contexto, ou seja, apenas um número fio de palavras antes e depois em um trecho para identificar palavras semelhantes não em termos de grafia, mas sim em termos de semelhança. Como exemplo prático, o professor Jesús mostrou a frase relacionadas. “O carro é rápido? Sim, ele é muito veloz”. As palavras “rápido” e “veloz” possuem relação. Dessa forma, podemos definir que palavras similares estão circundadas (ao redor).

Como exemplo de corpus, o professor indicou o site do Kaggle como forma de obtenção de alguns corpus selecionados para que se possa trabalhar em um projeto relacionado. Existem dados de nomes, comentários, notícias, músicas, tweets, reviews em produtos e outros. Grande parte destes datasets estão fielmente documentados e estão disponíveis de forma gratuita.

No decorrer da aula, foram apresentados programas de teste contendo com informações sobre o número de documentos, o tamanho do vocabulário e o número de stopwords. Um código utilizando o matplotlib também foi utilizado para mostrar a plotagem de uma matriz esparsa (termo-documento) com diferentes números de “k”, parâmetro que, se maior que o vocabulário, menos esparsa a matriz e mais pontos preenchidos temos (estamos olhando menos palavras).

Em um segundo momento da aula, foram abordadas duas medidas para identificar o quão informativa é a palavra do contexto, sendo estas: Pointwise Mutual Information PMI e PPMI. Por conta de níveis próximos de proximidade, como no caso de preposições, algumas palavras precisam ser ponderadas (para, de, nosso) por não serem informativas pelo fato de estarem presentes em quase todos os contextos.

A medida Pointwise Mutual Information, ou PMI, é uma medida baseada na medida de Informação Mútua. A fórmula da Informação Mútua é construída sob o somatório de duas palavras (x e y) que identifica a quantidade de informação que uma variável aleatória contém sobre outra variável aleatória. O PMI terá valores entre menos infinito e mais infinito. Como a fórmula contém uma função logaritma, se o termo for menor que 1, o resultado é uma probabilidade negativa. Como tomar decisões a partir de probabilidades negativas?

Para sanar este pequeno percalço, foi criado o Positive Pointwise Mutual Information, ou PPMI, o que restringe a função logaritima apenas a parcela positiva, substituindo valores negativos por zero.

A aula teve sequência a partir da execução de uma atividade rápida para o cálculo dos conceitos dados em sala. Na atividade, calculamos as métricas PPMI para um conjunto de vocabulário, montando combinações entre linhas e colunas, as quais eram compostas por palavras similares, de certa forma.