Resumo 11 – PLN

A décima primeira aula de Processamento de Linguagem Natural foi iniciada a partir da explicação do conceito de matriz esparsa, onde cada palavra é representada por um longo vetor com muitos valores nulos. A partir da apresentação de uma tabela com os tamanhos da matriz (sem stopwords), é possível notar o valor obtido com alguns dos exemplos clássicos utilizados pelo professor ao longo do curso: obras de Machado de Assis, matriz curricular do Bacharelado em Ciência e Tecnologia, entre outros exemplos.

Logo em seguida, trabalhamos no conceito de Vetor esparso e vetor denso no sentido de facilitar o trabalho nos algoritmos de aprendizado de máquina. Um método extremamente conhecido para gerar vetores de dimensão pequena, porém densos, é o SVD que, por sua vez, trata da redução de dimensionalidade via Singular Value Decomposition (ou Decomposição de Valor Singular). Existem outros métodos e abordagens como Skip-gram ou CBOW (utilizando redes neurais) e Brown clustering (agrupamento baseado na vizinhança de palavras).

O conceito de Redução de Dimensionalidade é extremamente importante em diversos meios. A ideia é utilizar um vetor de menor dimensão, mas que represente o vetor maior. Porém, como identificar um conjunto menor de dimensões? Existem técnicas específicas e bibliotecas prontas para realizar esse procedimento dentro do próprio sklearn do Python. Uma delas, talvez a mais conhecida é a PCA (Principal Componente Analysis) que aplica a redução de dimensionalidade baseada na análise de um componente principal. Existem técnicas e motivos para as melhores escolhas do algoritmo voltadas a estatística, como por exemplo, a escolha do sub-espaço amostral que mantém a maior variância do conjunto de dados originais quando projetados, entre outros. Uma comparação com diversos métodos de redução de dimensionalidade foi mostrada pelo professor onde foi possível observar o modo de entrada dos dados, o tipo de algoritmo, linearidade e complexidade. Em uma segunda tabela, o professor mostrou os respectivos métodos, em R e Python, para trabalhar com os algoritmos de redução de dimensionalidade.

Continuando com o tema, entramos em maiores detalhes sobre o SVD (Singular Value Decomposition) e nos foi mostrado o “output” dessa decomposição. Em primeiro lugar, temos uma matriz U como sendo uma matriz unitária m x m. Temos também uma matriz sigma, sendo definida como uma matriz retangular diagonal m x n com números reais não-negativos na diagonal. Por fim, temos a matriz V\* como sendo a conjugada transpostas de U unitária de dimensões n x n. O conceito de SVD é muito utilizado para analisar sistemas multivariados. Exemplos foram mostrados em aula mostrando matrizes reais e imagens reconstruídas.