Resumo 9 – PLN

Na oitava aula de Processamento de Linguagem Natural, o professor Jesús iniciou as discussões a partir de um tópico já introduzido na aula passada: semelhança de documentos. Como curiosidade, foi comentado na aula passada um artigo criado artificialmente utilizando técnicas de PLN que, além de ser aprovado pela academia, gerou um convite a seus criadores para apresentação das ideias.

O professor demonstrou a implementação de um algoritmo, em Python, capaz de identificar, dados trechos de documentos, a similaridade entre estes. Como exemplo prático, informações sobre algumas matérias do Bacharelado de Ciência e Computação foram coletadas para analisar, através de grafos, a semelhança entre estas. Como conclusão, foi possível perceber que a matéria de Processamento de Linguagem Natural possui um alto índice de semelhança com Algorítimos e Estruturas de Dados I.

Nesta análise, não foram realizados procedimentos de preparação ou transformação nos dados como, por exemplo, stop words, tratamento de maiúsculas e minúsculas, stemmer, entre outras técnicas. Em um outro exemplo utilizando a relação entre algumas obras de Machado de Assis, como Dom Casmurro, Memórias Póstumas de Brás Cubas, Quincas Borba e Esaú e Jacó, foi identificado, a partir da análise de similaridade entre textos, que as obras Dom Casmurro e Memórias Póstumas são as que mais se assemelham. Este tipo de análise é extremamente útil em sistemas de recomendação, uma vez que é possível recomendar ao usuário obras, livros, filmes, séries e qualquer outro produto baseado no índice de similaridade entre produtos que causaram satisfação a este cliente.

Para calcular a distância entre as palavras presentes em um trecho textual, foram utilizadas a distância Euclidiana e a distância de Cossenos. Para os exemplos, o professor demonstrou os algoritmos criados em Python com a biblioteca NumPy. Atualmente, existem módulos de bibliotecas capazes de realizar este tipo de procedimento de uma maneira mais rápida e simples, como aqueles contidos nas bibliotecas Scikit-learn e Pandas.

Assim como observamos, até aqui, as colunas, podemos também observar as linhas, analisando assim a similaridade entre palavras e não entre documentos inteiros. Nas abordagens anteriores, a dimensão do vetor é o número de documentos. Ao analisar palavras, a matriz de distância é igual ao tamanho do vocabulário, possuindo dimensões extremamente grandes em alguns casos. Essa matriz também é conhecida como matriz esparsa.

As mudanças entre os códigos anteriores que serviam para comparação de documentos foram mostradas pelo professor. Algumas medidas de cálculo de distância foram consideradas, dada que a dimensão dos vetores, para a comparação de palavras, foi alterada consideravelmente.

Como exemplo prático, o professor utilizou todas as ementas do curso de BCC, possuindo, cada uma, um total de 190 palavras diferentes, totalizando então 190 vértices de um grafo. Após a aplicação da normalização, um exemplo de similaridade entre palavras foi aplicado (lineares, complexidade e tempo são um exemplo).

Um outro exemplo interessante demonstrado pelo professor via código e testado com as obras de Machado de Assis, um algoritmo indica todas as palavras similares em um contexto, dado um input de uma palavra qualquer. Este mesmo exemplo foi aplicado a um livro de Processamento de Linguagem Natural e, ao digitar a palavra “datasets”, foram obtidos resultados como “thesaurus” entre outras palavras.

Em um interlúdio da aula, o professor comentou a respeito de atividades bônus que podem somar notas às médias finais. Através do incentivo à pesquisa, o aluno deve resumir uma tese ou dissertação defendida ano passado e responder a pergunta: “Por que o trabaho vale um mestrado/doutorado”?