Resumo 7 – PLN

A sétima aula de Processamento de Linguagem Natural teve início a partir de comentários do professor Jesús a respeito do projeto final da disciplina. Algumas orientações foram dadas como forma de guiar os alunos no desenvolvimento adequado do projeto em todas as suas etapas. Assuntos como forma de busca dos artigos, fontes de consulta e suas respectivas confiabilidades, implementações do projeto com base nos artigos, entre outros temas.

Em seguida, dando continuidade aos tópicos da aula, utilizaremos um algoritmo de classificação de texto para análise de sentimentos. Para introduzir o assunto, o professor comentou sobre a importância universal da classificação, seja esta aplicada em qualquer área do conhecimento. Entretanto, trazendo este conceito para o contexto da disciplina, como classificamos texto? De certa forma, a categorização de texto é a tarefa de classificar textos inteiros atribuindo um rótulo (de um conjunto finito de rótulos). Dada uma base de conhecimento, um sistema de categorização permite a recepção de um conjunto de elementos e, no fim, irá associar um rótulo para cada uma das entradas. Algumas possibilidades: análise de sentimento, identificação de idiomas, intenção do texto (positiva ou negativa), entre outras aplicações.

Como exemplo prático, temos um sistema de classificação de e-mails entre spam ou não spam. Neste caso, estamos falando de um classificador binário, ou seja, com apenas duas possibilidades na variável target. Várias técnicas podem ser aplicadas e, entre elas, o processamento de linguagem natural. Ainda sem entrar a fundo em critérios de classificação, podemos associar algumas frases como mal intencionadas como, por exemplo, “Prezado ganhador”, “Super Desconto”, “Prêmio a ser retirado”.

Formalmente, o objetivo é determinar, para cada entrada “x”, uma classe “y” que melhor se associa a esta entrada. Quando falamos de texto, “x” se transforma em “d” (documento) em um conjunto finito de classes “C”, sendo o objetivo determinar a classe “c” para cada entrada “d’.

Em uma primeira abordagem, é possível realizar essa classificação sem utilizar técnicas de Processamento de Linguagem Natural, definindo regras para serem aplicadas aos documentos. Entretanto, a manutenção dessas regras pode ser extremamente custosa. Assim, o grande desafio é construir um classificador que seja capaz de mapear o novo documento “d”, dada uma base de treinamento contendo diversos documentos e suas respectivas classes. Dessa forma, o objetivo é construir uma função capaz de determinar a classe associada ao novo documento de entrada. Para realizar esse procedimento, temos alguns algoritmos: Naive Bayes, Regressão Logística, Support Vector Machines (SVM), K-Neares Neighbors (KNN) entre outros, fazendo deste um problema de aprendizagem supervisionada.

De alguma forma, precisamos realizar transformações nos textos para podermos inserir esses dados ao classificador para que este possa interpretar as informações passadas. Os caracteres precisam ser transformados em números e, para tal, existe uma técnica chamada Bag of Words. Nessa abordagem, expressões regulares, lematizadores e diversas outras técnicas de PLN podem ser aplicadas para que, no final, seja obtida uma saída que traduza as palavras em um conjunto numérico. Algumas desvantagens: a dimensão da saída é muito alta e o algoritmo ignora totalmente o significado semântico das palavras.

De fato, um documento pode ser representado em uma bag of words. É possível obter uma matriz para representação dos documentos onde os índices representam as palavras contidas no bag of words e as demais colunas contém um flag binário para identificar, em cada um dos N documentos (representados em cada coluna), se determinada palavra do índice está (flag = 1) ou não está (flag = 0) no documento.

Em seguida, a técnica de Naive Bayes foi exemplificada e traduzida no contexto de classificação de textos. Em um problema desse tipo, temos um documento “d” como entrada para que o modelo possa calcular, de forma individual, a probabilidade condicional de cada uma das features, sendo estas representadas pelas palavras contidas no bag of word. O treinamento de um classificador de texto com Naive Bayes foi exemplificado pelo professor e, logo após isso, uma atividade foi passada em sala para fixação dos conhecimentos.

Na atividade aplicada pelo professor, duas questões foram dadas com exemplos sobre classificações de texto utilizando a probabilidade de Bayes. Na primeira questão, foi necessário analisar o conteúdo dos vocabulários em duas classes diferentes (positiva e negativa) através da construção de um conjunto de palavras e a classificação de um trecho textual.