MINERAÇÃO SEMIAUTOMÁTICA DE OPINIÕES COM BIBLIOTECA NLTK (APOIO UNIP)

Aluno: Marcos Antonio Gonçalves Molter
Orientador: Prof. Ricardo Leandro Piantola da Silva
Curso: Ciência da Computação
***Campus*:** Alphaville

Na era da informação, quanto maior o acesso a dados, maior o poder; por esse motivo, extrair conhecimento de todos os elementos dos meios tecnológicos é essencial nos dias atuais. No entanto, processar a forma natural da linguagem é ainda um desafio e transformar textos livremente escritos em estruturas analíticas é uma tarefa complexa. Neste sentido, busca-se a mineração de opiniões por aprendizado de máquina para a extração de dados estruturados com o objetivo de obter conhecimento por qualquer meio textual escrito em linguagem natural. O estudo se deu com a utilização do *kit* de ferramentas NLTK, disponível para a linguagem de programação *Python*. Inicialmente foi realizado um *benchmark* entre os classificadores e se observa que, quanto menor o volume de documentos, menor a acurácia obtida na classificação. O algoritmo *Naive Bayes* apresenta melhor desempenho em categorias com baixa amostragem de documentos quando comparado com os algoritmos *Maxent* e *Decision Tree*. Os resultados foram significativamente positivos, com *F-Measure* de 70,5%, contudo, observa-se discrepância entre *F-Measure* e Acurácia, o que é explicado pela irregularidade na distribuição dos documentos de treinamento. A utilização das ferramentas do NLTK é apropriada para mineração de opiniões pela classificação de documentos. O volume do conjunto de treinamento influencia diretamente a obtenção dos resultados, fato observado tanto na classificação com algoritmo *Naive Bayes* como na utilização de outros dois algoritmos para essa função.