Categorização de Texto

Bruno Brandelli, Renata Urbanski, Rodrigo Pacheco, Yago Vieira

I. Introdução

Este trabalho tem como foco a aprendizagem de máquina, onde o assunto será abordado através da categorização de textos. Seu principal objetivo é fornecer uma experiência prática com técnicas de categorização, passando por todas as etapas possíveis deste processo, e ao final analisar os resultados encontrados.

II. Pré-Processamento

Inicialmente foi necessário passar o corpus por uma ferramenta de anotação linguística e normalização morfológica, para realizar este processo utilizamos o parser **VISL** que havia sido indicado no enunciado do trabalho. Devido à dificuldades com o arquivo gerado pela ferramenta, foi utilizado um script em javascript para recuperar as informações no website.

Após esta primeira etapa de pré-processamento, foi necessário realizar a escolha da linguagem de programação que serviria de base para o restante da tarefa de pré-processamento. A linguagem escolhida foi **python**, na **versão 2.7**, visto que é ela é fácil de ser usada para manipulação de strings e arquivos. Com ela foram realizadas as etapas de extração de termos, seleção de termos mais relevantes e estruturação do arquivo utilizado para a aplicação da categorização. Para realizar os testes na etapa de classificação, cada corpus foi dividido em duas partes, uma de treino com *80 porcento* dos textos, e outra de teste com *20 porcento*.

A extração de termos utilizou *n-gramas*, com n variando de 1 a 3. Algumas classes gramáticais por não implicar tanto significado nos corpus foram deixadas de lado na construção dos *n-gramas*.

Ao realizar a seleção dos termos mais relevantes, definimos um número x de termos que seriam trazidos para a **Bag-of-Words**, estes termos foram definidos com base com a maior frequêcia de vezes em que apareciam nos textos do corpus. Nesta etapa foram realizados alguns testes com número de x termos selecionados, estes números foram *50, 100 e 150*, onde o primeiro número se mostrou trouxe bons resultados de termos e suas frequencias para todos *n-gramas*, porém os outros trouxeram termos que haviam aparecido apenas uma vez nos textos, o que pode não trazer bons resultados na etapa de categorização dos textos.

A ultima etapa de pré-processamento, foi a estruturação do arquivo que seria utilizado pela ferramenta de categorização. Tal etapa consistiu na verificação da **Bag-of-Words** geral, contra os textos já pré-processados.

III. Classificação dos Textos

Para realizar o processo de classificação, foi utilizada a ferramenta **WEKA**. Inicialmente foram realizados apenas testes utilizando o algoritmo *MultiLayer Perceptron*, onde foram fornecidos o arquivo de treino, e o arquivo de teste relativos ao *n-grama = 1*. Para $k = 50$, sendo k os termos mais relevantes de cada corpus, obtivemos um acerto de aproximadamente *75 porcento*, alterando k para 100 e 150, houve um aumento de acertos, elevando a taxa para quase *95 porcento*.