BRUNO PAZ

CAETANO ALMEIDA

MARLON FONTOURA

PEDRO IVO FRAGA

RODRIGO SILVEIRA

CAREGORIZAÇÃO DE TEXTO

Trabalho apresentado no Curso de Ciências da Computação, Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS).

Professora: Silvia Moraes

PORTO ALEGRE, RIO GRANDE DO SUL

2017

1. INTRODUÇÃO

Este relatório visa demonstrar como foi o processo de desenvolvimento de um programa de categorização de texto, no escopo da disciplina de Inteligência Artificial. Para o desenvolvimento do mesmo, utilizamos a linguagem Python 3.6.

2. AS CATEGORIAS

Recebemos um conjunto com 336 textos distintos, todos extraídos do *Diário Gaúcho Online*, no ano de 2010. Cada um destes textos pode ser classificado em 4 seções diferentes, a seção de Esportes (95 textos, a maioria sobre futebol), Polícia (89 textos), Espaço do Trabalhador (textos sobre oportunidades de emprego, 78 textos) e Seu Problema é Nosso (textos com relatos de problemas ocorridos com os leitores, 74 textos). Sendo assim, o programa deveria "ler" cada um dos textos e classifica-los, demonstrando qual a categoria de cada texto.

3. ETAPAS DE DESENVOLVIMENTO

O trabalho foi dividido em duas etapas, a etapa de Pré-processamento e a etapa de Categorização e Análise dos Resultados.

3.1. Pré-Processamento

O pré-processamento significa "arrumar" os textos de entrada para não serem processados ruídos durante o processo, dentre as técnicas utilizadas estão a normalização morfológica, anotação linguística, extração dos termos, seleção dos termos mais relevantes e estruturação.

A implementação foi feita da seguinte forma. Os arquivos são lidos e os textos são extraídos, desconsiderando o conteúdo irrelevante como as tags HTML, por exemplo. Após isto, é feita a lematização do mesmo, utilizando o Software CoGrOO.
Foi obtido no GitHub uma biblioteca para Python, que possibilita trabalhar com ele em aplicações pessoais.
(https://github.com/gpassero/cogroo4py)

Depois da etapa de lematização, é feita a anotação linguística, que serve para obtermos informações sobre o texto para que possamos, em um segundo momento, escolher os termos mais

relevantes de um texto. Neste trabalho, utilizamos a anotação Part-Of-Speech (POS), as tags da POS indicam as classes gramaticais das palavras: verbo (V), substantivo (N, PROP), adjetivo (ADJ) e advérbio (ADV).
Com isto, são extraídos os termos mais relevantes do texto. Os termos obtidos são conhecidos como a Bag-Of-Words (BoW).

3.2. Categorização e Análise dos Resultados

A implementar.

4. CONSIDERAÇÕES FINAIS

A proposta do trabalho foi interessante, mas infelizmente, devido ao cronograma apertado, o grupo não conseguiu progredir muito na sua implementação. Entretanto, a parte implementada foi um bom exercício para o que foi apresentado durante as aulas da disciplina. Gostaríamos de ter nos dedicado mais ao trabalho para fazer algo mais funcional e completo.