INF1771 - INTELIGÊNCIA ARTIFICIAL

TRABALHO 3 – REDES BAYESIANAS

Descrição:

Todo mundo que possui um endereço de email sofre com problemas de Spam (mensagens não-solicitadas enviadas em massa). Existem duas principais técnicas utilizadas para detectar Spam:

1) Detecção com base na fonte (endereço, servidor, ip...) dos emails. A detecção a partir de fontes conhecidas é bastante efetiva, mas ao mesmo tempo tende a classificar mensagens legítimas como spam.
2) Detecção de padrões no conteúdo dos emails. Utiliza-se técnicas de análise de conteúdo que possibilitam averiguar padrões para identificação e caracterização de mensagem consideradas spam.

O trabalho 3 consiste em implementar um sistema **simples** de **Detecção de Spam** utilizando **Redes Bayesianas**.

Informações Adicionais:

- A abordagem mais simples para resolver esse problema é analisar um conjunto de emails para identificar as palavras mais utilizadas em spam. A rede bayesiana pode ser totalmente baseada na probabilidade de ocorrência (ou não ocorrência) de determinadas palavras. A estrutura da rede pode ser utilizada para definir uma hierarquia entre frases. Um pequeno exemplo de spam pode ser visto a seguir:

  "Assunto: Free Satellite T.V. System

  FREE SATELLITE T.V. SYSTEM Watch over 500 channels of Digital Broadcast quality television on your own FREE satellite television system. These new systems use the new 18 inch satellite dish antenna. For a limited time we'll give you this top of the line Digital Satellite System for FREE! This is a top of the line system with features like on screen graphics, dual LNB, stereo receiver and infrared remote. All you have to do is call us to arrange delivery and order the channels you want to receive. The monthly cost of satellite television is usually much less than cable T.V and satellite television offers over 500 channels of all digital broadcast video quality and CD audio. You even get local channels now. That's over 500 channels! Don't miss this offer…it's only available while supplies last. For your Free Satellite System call 626-568-0903 24 hours a day"

- O site http://untroubled.org/spam/ possui uma base de emails classificados como spam de 1997 até 2011. Esses emails podem ser utilizados como base para o planejamento da rede bayesiana. Para facilitar, no site do curso (http://www.inf.puc-rio.br/~elima/ia/) também existe uma seleção filtrada e sem formatação de alguns desses emails.
- É permitido remover manualmente a formação (elementos html e outros dados irrelevantes) dos emails que serão analisadas pelo sistema.
- O programa deve ser bem simples, possuir uma interface simples (podendo ser em console) somente para carregar um determinado email (arquivo de texto) e mostrar o resultado da classificação do email (spam ou não spam) juntamente com a probabilidade do email ser ou não ser um spam.

Requisitos:

- O programa pode ser implementado em qualquer linguagem (C, C++, C#, Java...).
- Juntamente com o trabalho deve ser entregue um esboço (digital ou em papel) da estrutura da rede bayesiana utilizada.

Bônus:

- O trabalho que conseguir classificar corretamente (maior numero de acertos) um conjunto de emails (que será selecionado no dia da entrega do trabalho) receberá 2 pontos extras na nota. Podendo tirar até 12 no trabalho. Em caso de empate ambos receberam a nota extra.

Forma de Avaliação:

Será avaliado se o trabalho atendeu a todos os requisitos especificados anteriormente. O trabalho que atender a todos os requisitos receberá nota 10.

Data de Entrega:

06/06

Forma de Entrega:

O programa deve ser apresentado na aula do dia 06/06 (segunda) e enviando até o dia 06/06 para o email edirlei.slima@gmail.com.