Proceedings of the V Brazilian Conference on Neural Networks - V Congresso Brasileiro de Redes Neurais
pp. 451-456, April 2-5, 2001 - Rio de Janeiro - RJ - Brazil

Reconhecimento de Idiomas em Texto Escrito utilizando Redes Neuronais

Marcelo Costa P. e Santos[2], Manoel Gomes de Pinho[1], José Manoel Seixas[1], Luiz Pereira Calôba[1]
[1]Laboratório de Processamento de Sinais
[2]Laboratório de Redes de Alta Velocidade - RAVEL
COPPE - Universidade Federal do Rio de Janeiro
CP 68504, Rio de Janeiro, 21945-970
E-mails: marcsan@ravel.ufrj.br, pinho@lps.ufrj.br, seixas@lps.ufrj.br, caloba@lps.ufrj.br

Abstract

This paper studies the feasibility determining the language used in a written text by means of neural processing. A variety of texts collected from newspapers written in six of the main western languages - Portuguese, English, Spanish, French, Italian and German- were analyzed and the correct language identification was above 95%. This ability of language recognition has applications on information systems, notably on World Wide Web application programs and data compression.

1. Introdução

O presente trabalho visa o estudo do uso de redes neuronais para a determinação do idioma em que foi escrito um texto de assunto geral.

A maioria das pessoas que já teve um contato, mesmo que pequeno, com uma língua estrangeira consegue, com relativa facilidade, reconhecer o idioma em que foi escrito um texto. Portanto, é de se esperar que uma rede neuronal artificial (afinal, inspirada biologicamente) também consiga fazer tal reconhecimento.

Este reconhecimento automático poderia ser usado em diversas aplicações como, por exemplo:

- **Determinação de código de compressão ótimo**: Sabemos da Teoria da Informação que se desejarmos um código de compressão de dados ótimo (código de Huffman), isto é, próximo à entropia para seqüências longas de caracteres, devemos conhecer a distribuição probabilística do alfabeto de entrada [1]. Como esta distribuição varia com o idioma, se pudermos identificá-lo automaticamente, certamente conseguiremos taxas de compressão melhores para textos escritos em idiomas variados.
- **Tradução on-line**: Existem atualmente no mercado softwares para consulta a Bancos de Dados com expressões em diversas línguas (dicionários), e softwares de tradução de um dado texto. A maioria destes programas pergunta ao usuário qual o idioma de origem do texto e para qual idioma se deseja traduzir. Com uma identificação automática eficiente poderíamos dispensar pelo menos uma destas perguntas (qual o idioma de origem), o que tornaria mais fácil e versátil a sua utilização.
- **Mecanismos de busca da INTERNET**: Muitos sítios de busca da INTERNET utilizam programas automáticos para alimentação de seus Bancos de Dados. Tais programas, denominados *spiders*, navegam na *World Wide Web* colhendo amostras de páginas que são enviadas ao cadastro e que passam a integrá-lo automaticamente, sem a interferência humana. Uma das informações constantes da maioria destes Bancos de Dados é a língua em que está escrita a página, permitindo restringir a consulta a páginas de uma determinada língua. Uma forma automática de reconhecimento de língua certamente é de grande interesse para este tipo de aplicação.

Restringimos nosso estudo às principais línguas ocidentais - Português, Inglês, Espanhol, Francês, Italiano e Alemão - porém consideramos fácil a extensão do trabalho a outras línguas, se necessário. Eventualmente, idiomas ou versões regionais (dialetos) podem vir a ser reconhecidos. Idiomas não ocidentais, como o Japonês, Árabe, Russo e Hebraico, se tiverem que ser reconhecidos diante de línguas ocidentais, são mais fáceis de serem determinados, por terem alfabetos próprios, não havendo a necessidade do uso de redes neuronais para esta tarefa.

2. Levantamento de dados

Consideramos que, quanto maior a variedade, tanto de autor quanto de assunto, melhor seria o nosso estudo, pois tal variedade impediria que a rede neuronal perdesse o seu poder de generalização. Por exemplo, poderíamos buscar livros inteiros, já disponíveis para download via INTERNET no Projeto Gutemberg [2], porém, certamente um autor usa, devido ao seu estilo literário, determinadas palavras com maior freqüência do que outras. Sendo assim, correríamos o risco da rede neuronal aprender a reconhecer o autor em vez da língua. O mesmo poderia ocorrer com o assunto. Se escolhêssemos somente textos sobre Informática, a freqüência de termos técnicos seria anormal, não atendendo aos nossos objetivos.

No entanto, seria interessante que tivéssemos os mesmos textos em diversas línguas, de forma a evitar que nomes próprios muito freqüentes determinassem a escolha da língua pela rede. Se os textos forem sobre o mesmo assunto, a possibilidade dos nomes próprios também se repetirem, e não influenciarem o resultado, é maior.

Em face destes problemas, decidimos que as melhores fontes de dados para nosso estudo seriam artigos de jornal, porque são escritos por autores diversos e, normalmente, são ecléticos quanto aos assuntos abordados. Determinamos, então, o noticiário internacional de jornais de diversas línguas como fonte de dados pois estes, com grande probabilidade, conteriam textos sobre os mesmos assuntos, em um dado período do ano.

Decisão tomada, deparamo-nos com o problema de conseguir jornais internacionais e digitar seus conteúdos. Tal dificuldade foi contornada com o uso da INTERNET, que possui páginas online dos principais jornais do mundo.

Passamos então para a primeira fase do trabalho, visitando diariamente os sítios de diversos jornais estrangeiros e nacionais e colecionando seus noticiários internacionais. Entre eles destacaram-se:

- Português: O GLOBO (Brasileiro, www.oglobo.com.br);
- Inglês: NEW YORK TIMES (Americano, www.nytimes.com);
- Espanhol: LA NACIÓN (Argentino, www.lanacion.com); EL NACIONAL (Venezuela, www.el-nacional.com); EL PAÍS (Espanhol, www.diarioelpais.com);
- Francês: LE MONDE (Francês, www.lemonde.fr);
- Italiano: CORRIERE DELLA SIERA (Itália, www.corriere.it);
- Alemão: SÜDDEUTSCHE ZEITUNG (Alemanha, http://www.diz-muenchen.de/).

Conseguimos colecionar, em algumas semanas, aproximadamente 200 Kbytes em artigos de assuntos variados nas seis línguas escolhidas, o que nos permitiu avançar para a próxima fase do trabalho.

3. Preparação dos dados

3.1. Processamento Preliminar

Os artigos coletados se encontravam no formato HTML, que possui diversos comandos e marcadores derivados do inglês e, por isto, todos os artigos foram convertidos para texto puro, utilizando-se o conversor incluso no navegador Internet Explorer 5 (opção "Salvar como", escolhendo o formato "texto" para o arquivo).

Como os artigos de uma língua foram praticamente todos retirados de uma mesma fonte, continham ainda características de formatação própria daquela publicação eletrônica como: linhas horizontais convertidas em longas seqüências do caracter "-", formatos padrão para informação de autor e data de publicação, etc. Portanto, foi necessário o desenvolvimento de um programa que filtrasse apenas caracteres ASCII imprimíveis e eliminasse longas repetições de um mesmo caracter. Ainda após esta filtragem, procedemos a uma inspeção visual nos dados e retiramos, manualmente, trechos que não eram realmente texto, como notas sobre direitos autorais, cabeçalhos padrão, etc.

3.2. Divisão em Janelas

Como os dados devem ser apresentados à rede neuronal em conjuntos de tamanho fixo, notamos a necessidade de quebrar os longos arquivos texto resultantes do processamento anterior em outros menores.

Optamos por apresentar os dados à rede de duas formas diferentes, considerando-se uma dada janela de texto adquirida:

1. FREQÜÊNCIAS: Apresentando à rede a freqüência relativa dos caracteres em um determinado trecho do texto;
2. CARACTERES: Apresentando à rede os códigos ASCII dos caracteres que compõem o texto.

Os casos possuem características distintas. No primeiro (apresentando FREQÜÊNCIAS), espera-se que o desempenho cresça com o tamanho da janela pois, quanto maior a janela, a freqüência relativa dos caracteres mais se aproxima da distribuição de freqüências naturais da língua e, consequentemente, maior a facilidade no trabalho de discriminação da rede. Além disto, a dimensão da entrada da rede é determinada pelo número de caracteres contados, e não pelo tamanho da janela. Portanto, o aumento da janela não implica em qualquer aumento de complexidade computacional para a rede.

No segundo caso (apresentando CARACTERES), o tamanho da janela será a dimensão dos vetores de entrada da rede e, portanto, janelas muito grandes implicarão em redes grandes, com conseqüente aumento nos tempos de processamento.

Foram escritas rotinas que dividem os dados preparados no item anterior para cada um dos seis idiomas, em janelas de 10, 30, 50, 100, 200 e 500 caracteres, com intenção de utilizar as janelas de 10, 30 e 50 para apresentação como CARACTERES e as janelas de 50, 100, 200 e 500 para FREQÜENCIAS. Para o caso de CARACTERES, testaremos as mesmas janelas com sobreposição de 50 por cento. Para o caso de FREQÜÊNCIAS, contamos o número de ocorrências dos 192 caracteres imprimíveis da tabela ASCII, conforme a Tabela 1.

Geraram-se também vetores alvo para treinamento supervisionado, com 6 linhas (uma para cada língua). Assim, cada nó de saída da rede neuronal passou a ser associado a um idioma. Convencionou-se que a linha 1 seria a correspondente ao Português, a 2 ao Inglês, 3 Espanhol, 4 Francês, 5 Italiano e 6 Alemão. A codificação binária dos vetores-alvo fez com que a matriz contivesse 1 na linha correspondente à língua correta e -1 em todas as outras. Por exemplo, um texto em Italiano foi codificado com 1 na linha 5 e -1 em todas as outras na matriz alvo (veja Tabelas 1 e 2).

Tabela 1: Aspecto da matriz de FREQÜÊNCIAS "p", de 192 linhas , uma para cada caracter imprimível, para qualquer tamanho de janela

Nº de ocorrências do caracter:						
65 – A	6	10	14	5	13	...
66 – B	5	2	19	2	9	...
66 – C	1	2	1	0	5	...
67 – D	0	0	2	4	9	...
68 – E	5	9	8	13	15	
...	...	...	...	...	...	...
...	...	...	...	...	...	...
97 – a	23	13	31	69	26	...
98 – b	19	24	22	18	32	...
...	...	...	...	...	...	...
...	...	...	...	...	...	...
253 – ý	0	1	0	1	0	...
254 – þ	0	0	0	1	0	...
255 – ÿ	1	0	0	0	0	...

Tabela 2: Aspecto da matriz de CARACTERES "p" de treinamento e teste "t" para alvos do treinamento supervisionado

matriz p										matriz t				
...	b	p	o	e	a	i		...		...	1	1	1	...
...	a	ó	n	x	p	s	p	...		...	-1	-1	-1	...
...	r	l	t	p	e		l	...		...	-1	-1	-1	...
...	r	v	o	l	n	d	e	...		...	-1	-1	-1	...
...	i	o		o	a	i	b	...		...	-1	-1	-1	...
...	l	r	p	d	s	a	i	...		...	-1	-1	-1	...
...		a	a	i		s	s	...						
...	d		r	r	d		c	...						
...	e	p	a		o	d	i	...						
...		r		a	i	o	t	...						

3.3. Concatenação, Normalização e Escolha de Componentes Principais

Neste ponto desejamos preparar as matrizes que serão apresentadas à rede neuronal para treinamento e teste, definindo a metodologia de normalização dos sinais e usando uma técnica de compactação (análise de componentes principais).

Decidimos, para as experiências iniciais, limitar a massa de dados em 150 registros por língua, sendo 100 elementos para treinamento e 50 para teste, somando um total de 900 elementos, sendo 600 para treinamento e 300 para teste. Foram feitos alguns testes também com um número maior de vetores de treinamento de teste (4000 vetores de treinamento e 2000 vetores de teste) para descobrirmos o comportamento da rede neuronal com base em um conjunto maior de treinamento.

Foi implementada uma rotina que concatena, em ordem aleatória, 150 registros de cada língua formando as matrizes "p" e "t" (treinamento e alvos) únicas. O conteúdo da matriz "p" foi então normalizado para média 0 e desvio padrão 1.

Para o caso das matrizes com freqüências relativas, utilizamos o procedimento estatístico de determinação de Componentes Principais (PCA - *Principal Component Analysis*), onde selecionamos apenas os caracteres que contribuíssem com mais que 2% para a variância da amostra. Desta forma, a dimensão da matriz de entrada caiu de 192 (número de caracteres imprimíveis) para aproximadamente 10, dependendo do tamanho da janela. A vantagem dessa abordagem de compactação de sinais por meio da análise de componentes principais é que ela diminui enormemente a complexidade computacional da rede e seleciona automaticamente os caracteres mais importantes para a fase de treinamento. Outra vantagem é ortogonalizar as componentes dos vetores de entrada, de modo que ficam descorrelacionadas entre si [3].

4. Rede Neuronal

O método *Backpropagation* mostrou convergência e resultados bem melhores em testes preliminares e, portanto, foi escolhida para a base de nossos trabalhos.

O método *Backpropagation* demonstrou eficiência satisfatória para o problema a partir de 10 neurônios na camada intermediária, quando os dados de freqüência lhe eram apresentados. Resultados não tão promissores foram conseguidos com a matriz de caracteres, apresentada à mesma rede. Neste caso, o aumento de neurônios na camada intermediária não demonstrou ter influência significativa na performance da rede, influindo sim, é claro, de forma decisiva no tempo de treinamento necessário.

Portanto variamos entre 10, 25 e 50 o número de neurônios da camada intermediária e utilizamos 6 neurônios, um para cada língua, na última camada, a mesma dimensão dos vetores alvos construídos.

Foram então criadas redes *backpropagation* com quantidade variável de neurônios, do tipo tangente hiperbólica. Convencionamos que o neurônio com maior valor de saída determinaria a língua do texto apresentado à rede (critério de máxima probabilidade) [4].

Dentre os diversos métodos de treinamento disponíveis, o trainrp (*Resilient Backpropagation*) foi escolhido, pela rapidez na convergência e uso modesto de memória.

O método trainrp elimina o efeito da pouca inclinação da sigmóide nos pontos mais afastados da origem, levando em consideração somente o sinal da derivada, não tendo seu módulo qualquer efeito sobre a atualização dos pesos.

O módulo da atualização dos pesos parte de uma constante e é aumentado sempre que uma atualização não alterar o sinal do gradiente e diminuído sempre que o sinal do gradiente se inverte. Desta forma, caminhamos para o ponto de mínimo sempre com velocidade crescente e, se o ponto de mínimo é ultrapassado, retornamos uma distância cada vez menor, até atingi-lo. Tal método fornece uma convergência rápida, apesar de apresentar alguma oscilação em torno do mínimo no final do processo.

5. Resultados Obtidos

Variou-se a forma de apresentação dos dados de entrada (CARACTERES ou FREQÜÊNCIAS), o tamanho das janelas de caracteres e a existência ou não superposição entre janelas consecutivas.

Notou-se, de modo geral, que a variação da taxa de aprendizado não afetou significativamente os resultados obtidos, nem a convergência da rede. Portanto, foi usada a taxa de aprendizado de 0,05 na maioria dos experimentos.

Foi observado também, nos experimentos que obtiveram piores resultados, como nos que foram apresentados CARACTERES ou FREQÜÊNCIAS com janelas de aquisição muito estreitas, que o erro médio quadrático (MSE) diminuía muito lentamente à medida em que a rede era treinada, ficando longe do objetivo de 0,01 mesmo depois de um grande número de épocas (mais de 30.000). Nesses casos, treinando a rede ao longo de um número elevado de épocas, uma pequena melhoria do erro percentual médio no conjunto de treinamento era observada, porém o erro percentual no conjunto de teste crescia, indicando que a rede não conseguia generalizar o problema. Dois destes gráficos, um com convergência lenta e outro com convergência mais rápida são mostrados nas figuras 1 e 2.

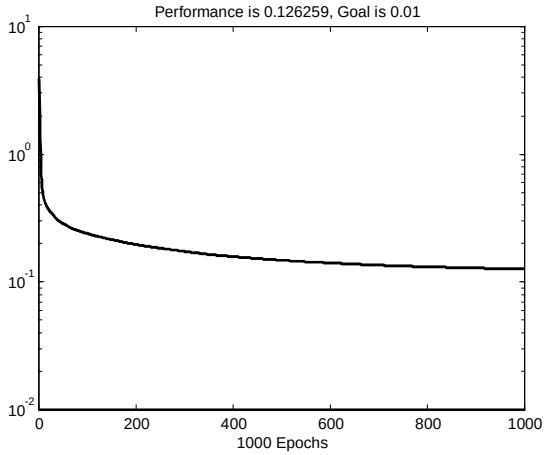

Figura 1: Exemplo de erro médio quadrático na fase de treinamento com convergência lenta

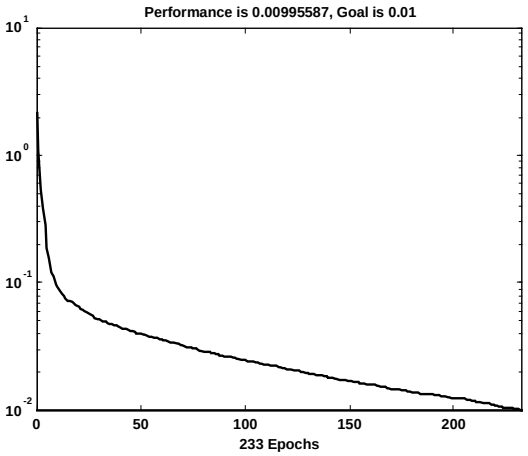

Figura 2: Exemplo de erro médio quadrático na fase de treinamento com convergência rápida

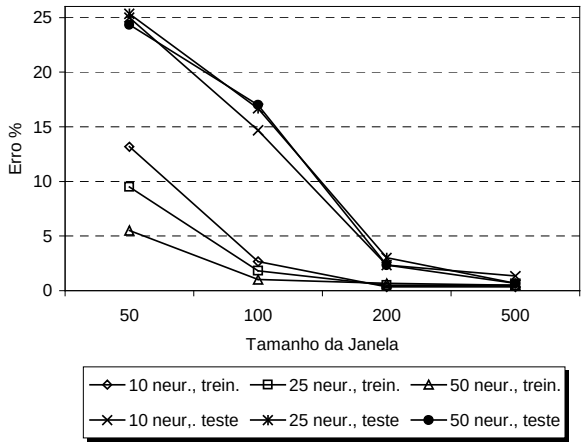

Figura 3: Desempenho do classificador neuronal usando a FREQÜÊNCIA de caracteres

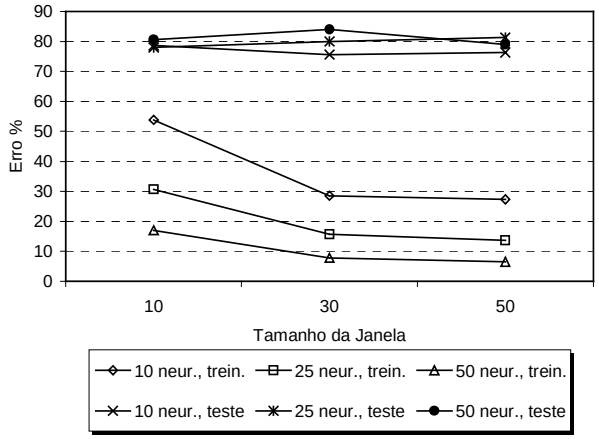

Figura 4: Desempenho do classificador neuronal usando janelas de CARACTERES sem superposição

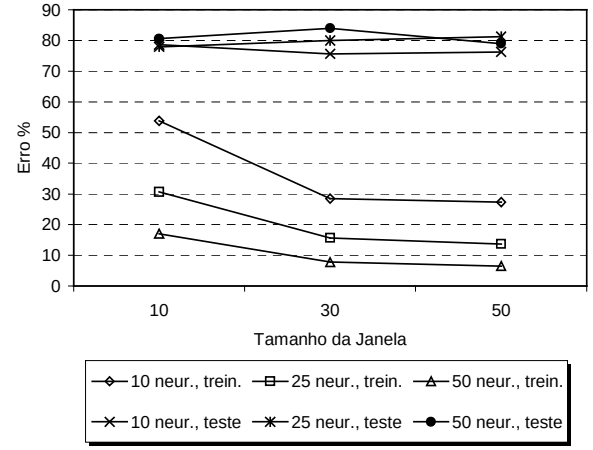

Figura 5: Desempenho do classificador neuronal usando janelas de CARACTERES com superposição

Como se pode notar das Figuras 3, 4 e 5, os resultados obtidos, apresentando como entrada a FREQÜÊNCIA relativa de caracteres, foram muito superiores àqueles obtidos quando foram apresentados à rede os CARACTERES de janelas com ou sem superposição. Os resultados mostram que é possível obter resultados muito bons (com erros percentuais menores do que 5%) através de janelas a partir de 200 caracteres e redes *backpropagation* simples, com duas camadas e 10 neurônios na camada intermediária. O treinamento requerido também foi relativamente pequeno (600 padrões de entrada).

Os maus resultados obtidos apresentando-se CARACTERES diretamente à rede, como era de se esperar, mostram que a identificação de idiomas usando pequenas janelas de texto, com ou sem superposição, é um problema de alta complexidade. Embora isto possa parecer simples para um ser humano, a rede neuronal não consegue generalizar bem. Foram tentados também treinamentos com um número muito maior de vetores de entrada (4000 vetores no conjunto de treinamento e 2000 vetores no conjunto de teste) e, embora haja uma pequena melhoria no erro percentual do conjunto de teste, a diferença é muito pequena. Isso sugere que a rede não terá um desempenho aceitável mesmo com um treinamento exaustivo. Observou-se também que o aumento do número de neurônios na camada intermediária, embora diminua o erro percentual no conjunto de treinamento, não modifica os maus resultados no conjunto de teste.

Pode-se notar, ainda, que os resultados obtidos apresentando-se CARACTERES diretamente à rede foram praticamente iguais usando ou não superposição na janela de texto apresentada para treinamento, o que indica a pouca importância da superposição de janelas para este enfoque.

6. Conclusões

A freqüência relativa dos caracteres em uma janela de texto demonstrou ser um padrão viável para reconhecimento da língua utilizada em um texto e sua eficiência, conforme esperado, mostrou-se proporcional ao comprimento da janela de texto utilizada. Janelas com tamanho superior a 200 caracteres mostraram uma eficiência de classificação acima de 95%.

Um tema de pesquisa interessante seria a utilização da freqüência relativa de seqüências curtas de caracteres de cada língua (provavelmente artigos). Para a língua inglesa, conforme [1], existe um estudo conhecido como **Brown corpus**, que contém mais de um milhão de palavras de diferentes assuntos e autores e a estatística dos caracteres, digramas (conjuntos de 2 letras), trigramas, tetragramas e freqüência das palavras e frases mais usadas no Inglês. O nosso problema é mais difícil pois envolve a classificação de textos de seis línguas. Alguns testes preliminares, contando os caracteres da Tabela 3, apresentaram resultados muito promissores com redes bem mais simples do que as utilizadas anteriormente.

Tabela 3: Caracteres utilizados em rede *backpropagation* com 6 neurônios na camada intermediária, com resultados promissores (o símbolo _ indica espaço em branco)

Português	ç, ã
Inglês	_The_, _the_
Espanhol	_ Las_, _Los_ , _las_, _los_
Francês	è
Italiano	_ il_, _per_, _i_
Alemão	Ü, ü (tremado)

Existem em todo o mundo diversos projetos e pesquisas em andamento visando o reconhecimento automático de linguagem escrita, com diferentes modelos e abordagens [5]. A maioria deles, contudo, visa também o aspecto semântico do texto e não apenas a identificação do idioma.

Referências

[1] Bell, T. C., Cleary, J. G., Witten, I. H., "Text compression", Prentice-Hall, New Jersey, 1990.

[2] Endereço na Internet: http://www. gutenberg.net

[3] Demuth, H., Beale, M., "Neural Network Toolbox for use with Matlab - User's Guide Version 3.0", 1999.

[4] Haykin, S., "Neural Networks - A Comprehensive Foundation", 2nd ed., Prentice-Hall, New jersey, 1999.

[5] Endereços na Internet:

http://www.mcs.vuw.ac.nz/comp/Publications/CS-TR-94-8.abs.html

http://www.dna.lth.se/Education/Exarbete/98.04.Eric/Babel.html

http://vismod.www.media.mit.edu/people/starner/asl/asl-tr466/index.html

http://mirach.cs.buffalo.edu/Publications/TechReps/WLR/wlr.html

http://www.unm.edu/~hdls/hdls-2/christianvogler.html

http://speech.inesc.pt/~dcaseiro/html/bibliografia.html

http://www.jguru.com/ANTLR/book/index.html

Agradecimentos

Os autores gostariam de agradecer o apoio do CNPq, CAPES e da FAPERJ a este projeto.