PPGI

PROGRAMA
DE PÓS-GRADUAÇÃO
EM INFORMÁTICA

UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

Olá,
muito prazer!

RAFAEL ADNET PINHO

Formado em Sistemas de Informação pela PUC-RJ

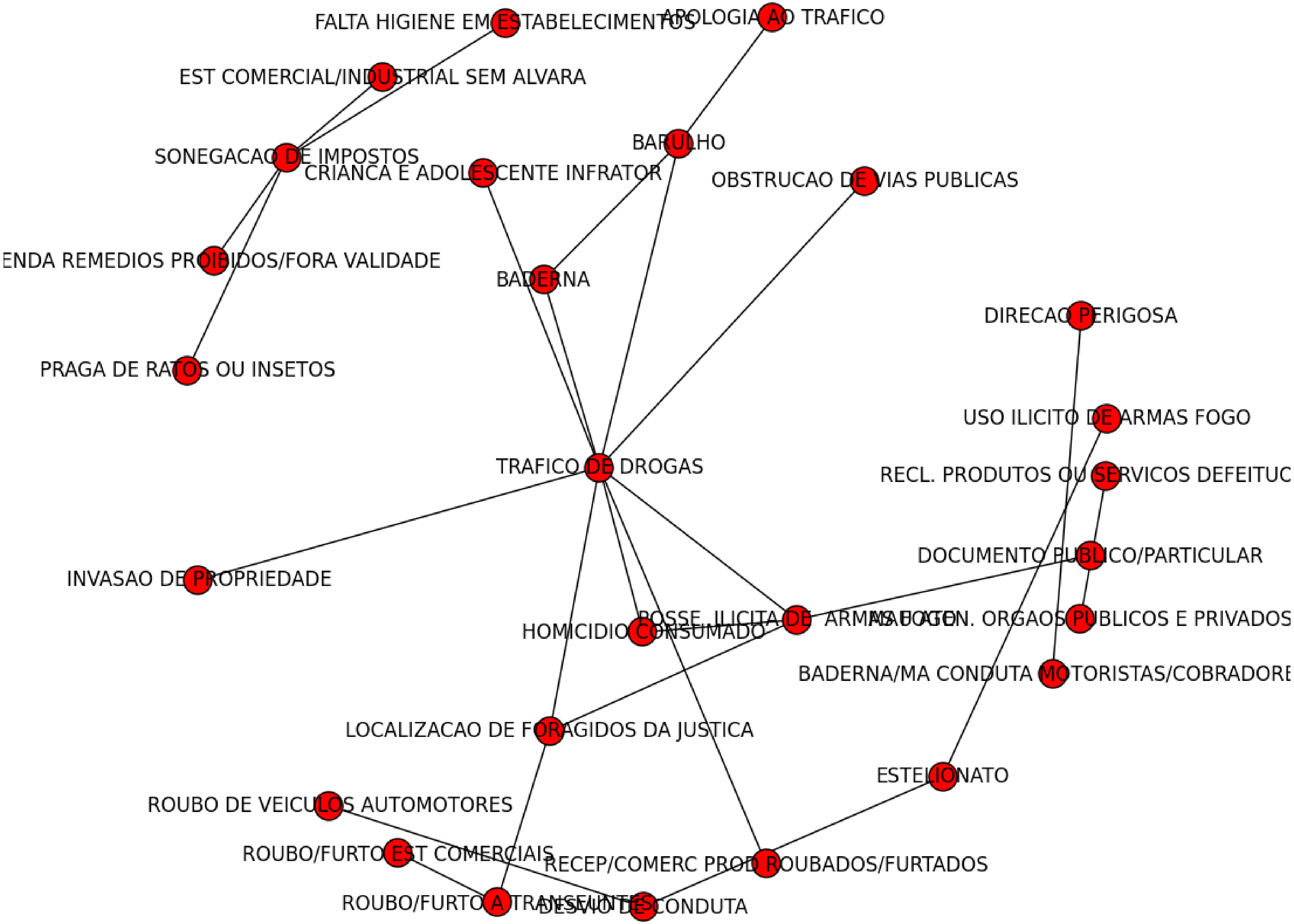
Mestrando em Informática pela UFRJ (PPGI - MASI)

Mais de 5 anos de experiência no mercado de tecnologia

Desde 2014 na BigData Corp - Product Manager

radnet

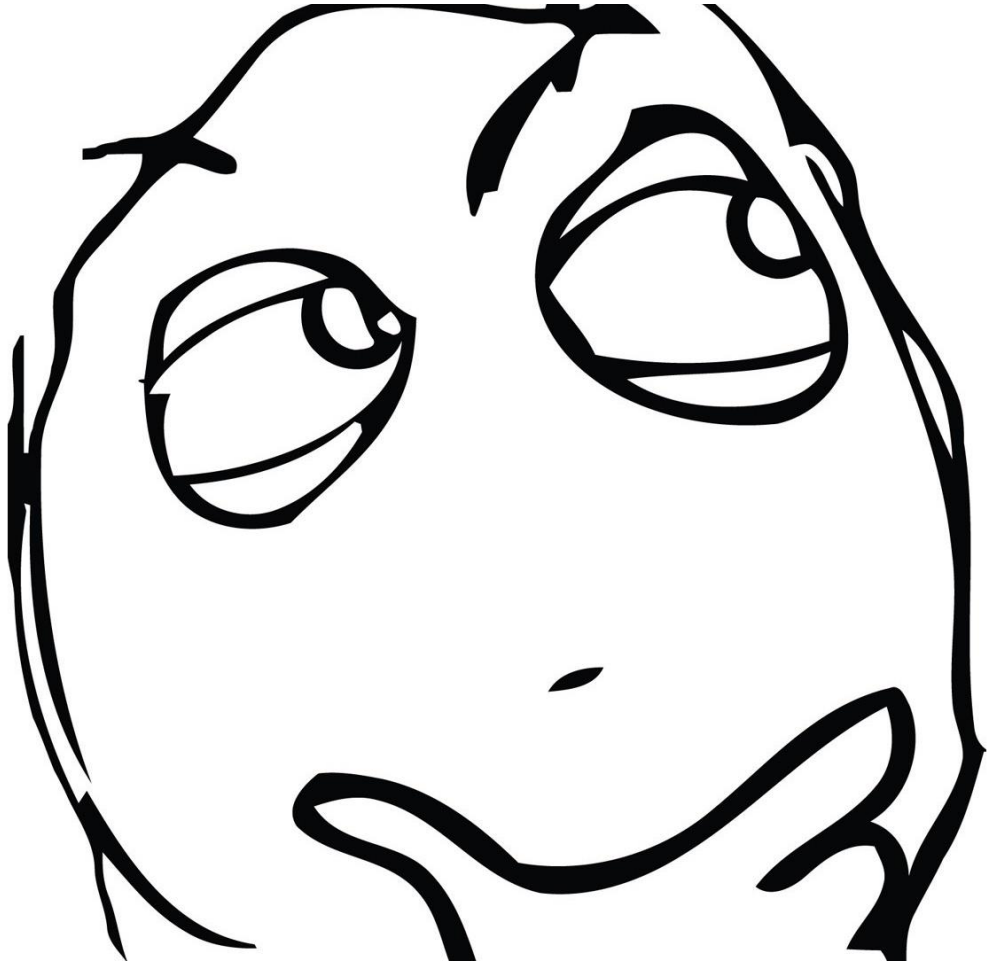
3 anos
de denúncias anônimas


FALTA HIGIENE EM ESTABELECIMENTOS
APOLOGIA AO TRAFICO
EST COMERCIAL/INDUSTRIAL SEM ALVARA
BARULHO
SONEGACAO DE IMPOSTOS
CRIANCA E ADOLESCENTE INFRATOR
OBSTRUCAO DE VIAS PUBLICAS
ENDA REMEDIOS PROIBIDOS/FORA VALIDADE
BADERNA
DIRECAO PERIGOSA
PRAGA DE RATOS OU INSETOS
USO ILICITO DE ARMAS FOGO
TRAFICO DE DROGAS
RECL. PRODUTOS OU SERVICOS DEFEITUC
DOCUMENTO PUBLICO/PARTICULAR
INVASAO DE PROPRIEDADE
POSSE ILICITA DE ARMAS FOGO
MAU ATEN. ORGAOS PUBLICOS E PRIVADOS
HOMICIDIO CONSUMADO
BADERNA/MA CONDUTA MOTORISTAS/COBRADORE
LOCALIZACAO DE FORAGIDOS DA JUSTICA
ESTELIONATO
ROUBO DE VEICULOS AUTOMOTORES
ROUBO/FURTO EST COMERCIAIS
RECEP/COMERC PROD ROUBADOS/FURTADOS
ROUBO/FURTO A TRANSEUNTES
DESVIO DE CONDUTA

Sistema de classificação de denúncias baseado em I.A.

- Auxilio ao operador
- Recomendação de classificações baseado nas características textuais da denúncia

Como é a estrutura de uma denúncia?



Classificação Principal

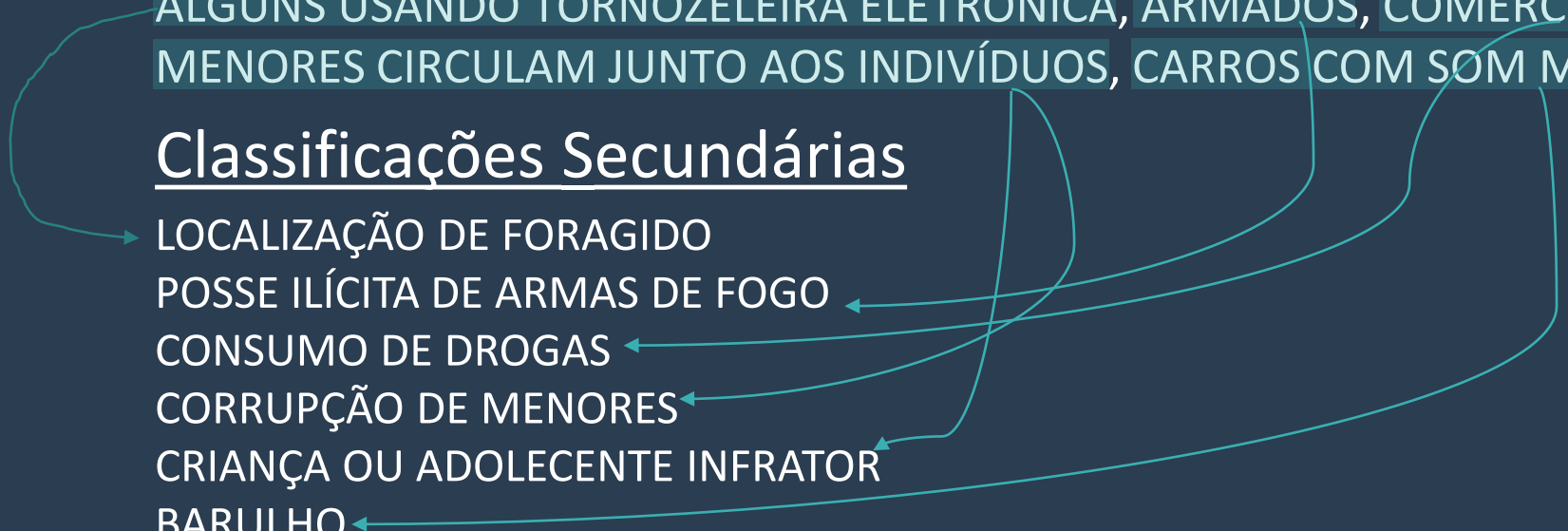
TRÁFICO DE DROGAS

Descrição

NA RUA XPTO, PRÓXIMO A UM COLÉGIO E A UMA PRAÇA, SOB UMA AMENDOEIRA, DIARIAMENTE, À PARTIR DAS 18H, INDIVÍDUOS DO SEXO MASCULINO (NÃO IDENTIFICADOS), ALGUNS USANDO TORNOZELEIRA ELETRÔNICA, ARMADOS, COMERCIALIZAM ENTORPECENTES, MENORES CIRCULAM JUNTO AOS INDIVÍDUOS, CARROS COM SOM MUITO ALTO TOCAM FUNKS.

Classificações Secundárias

LOCALIZAÇÃO DE FORAGIDO
POSSE ILÍCITA DE ARMAS DE FOGO
CONSUMO DE DROGAS
CORRUPÇÃO DE MENORES
CRIANÇA OU ADOLECENTE INFRATOR
BARULHO



1º

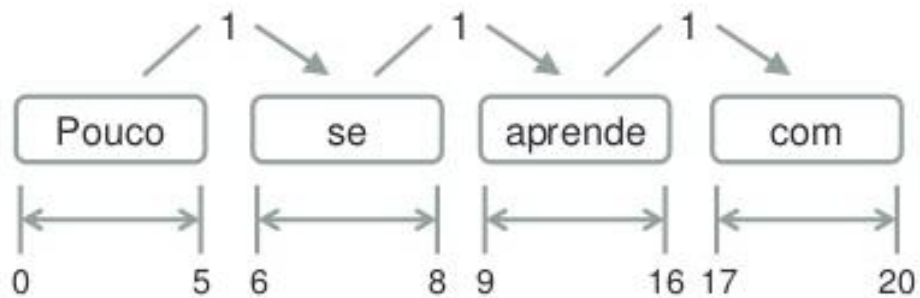
Data Mining

Pré-Processamento

1. Tokenização
2. Normalização
3. Stopwords
4. Stemming

- Interpreta o texto transformado em termos
- Exemplo

Texto: *Pouco se aprende com a vitória, mas muito com a derrota.*



Termos:

[“Pouco”, “se”, “aprende”, “com”, “a”, “vitória”, “mas”, “muito”, “com”, “a”, “derrota”]

1. Tokenização
2. Normalização
3. Stopwords
4. Stemming

- Diferentes formas de tokenização

***Pouco se aprende com a vitória**, mas muito com a derrota.*

Shingle n=4

Pouco	aprende com
Pouco se	aprende com a
Pouco se aprende	aprende com a vitória
Pouco se aprende com	com
se	com a
se aprende	com a vitória
se aprende com	a
se aprende com a	a vitória
aprende	vitória

1 Tokenização
2 Normalização
3 Stopwords
4 Stemming

Normalização

- Conversão do texto para letras minúsculas.
- Pode remover acentos, pontos, números, etc.

Texto: *Pouco se aprende com a vitória, mas muito com a derrota.*

[“**pouco**”, “se”, “aprende”, “com”, “a”, “**vitoria**”, “mas”, “muito”, “com”, “a”, “derrota”]

Remoção de Stopwords

- Remove as palavras comuns
  - Sem significado relevante
- Preposição, pronome, etc.
- Depende do idioma

Texto: *Pouco se aprende com a vitória, mas muito com a derrota.*

[“pouco”, “se”, “aprende”, “com”, “a”, “vitoria”, “mas”, “muito”, “com”, “a”, “derrota”]

[“pouco”, “aprende”, “vitoria”, “muito”, “derrota”]

Stemming

1. Tokenização
2. Normalização
3. Stopwords
4. Stemming

- Converte os termos em sua raiz gramatical
- Elimina plural

Pouco se aprende com a vitória, mas muito com a derrota.

[“pouco”, “se”, “aprende”, “com”, “a”, “vitoria”, “mas”, “muito”, “com”, “a”, “derrota”]

[“pouco”, “aprende”, “vitoria”, “muito”, “derrota”]

pouco	pouc
aprende	aprend
vitoria	vitor
muito	muit
derrota	derrot

2º

Indexação

- **Tratamento de termos que são muito usados em uma coleção de documentos**
- **Fator tf**
  - Quantidade de vezes que o termo i aparece no documento (Quão bem i descreve d)
- **Fator idf**
  - Inverso da frequência do termo i dentro da coleção de documentos.
  - Quanto menos usado for o termo, maior o idf

$$w_{i,d} = tf \times idf = tf_{i,d} \times \log(n / df_i)$$

Indexação

- **Tratamento de termos que são muito usados em uma coleção de documentos**

	a	an	apple	ate	banana	eat	i	today	will	yesterday
Doc 1		0.0811	0.0811	0.0811			0			0.0811
Doc 2		0.0676	0.0676			0.1831	0	0.1831	0.1831	
Doc 3	0.2197			0.0811	0.2197		0			0.0811

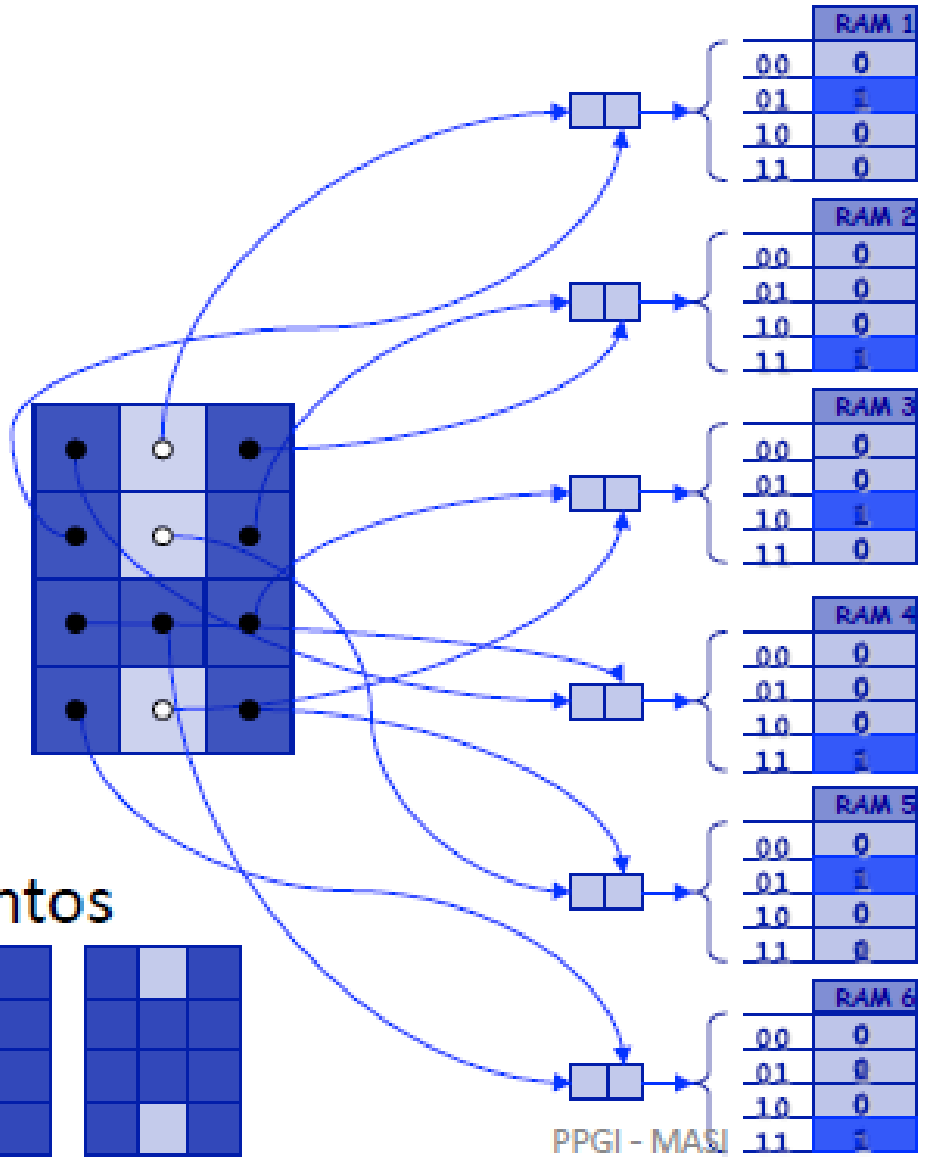
3º

Rede Neural sem Peso

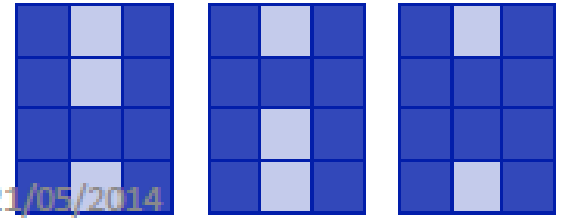
WISARD

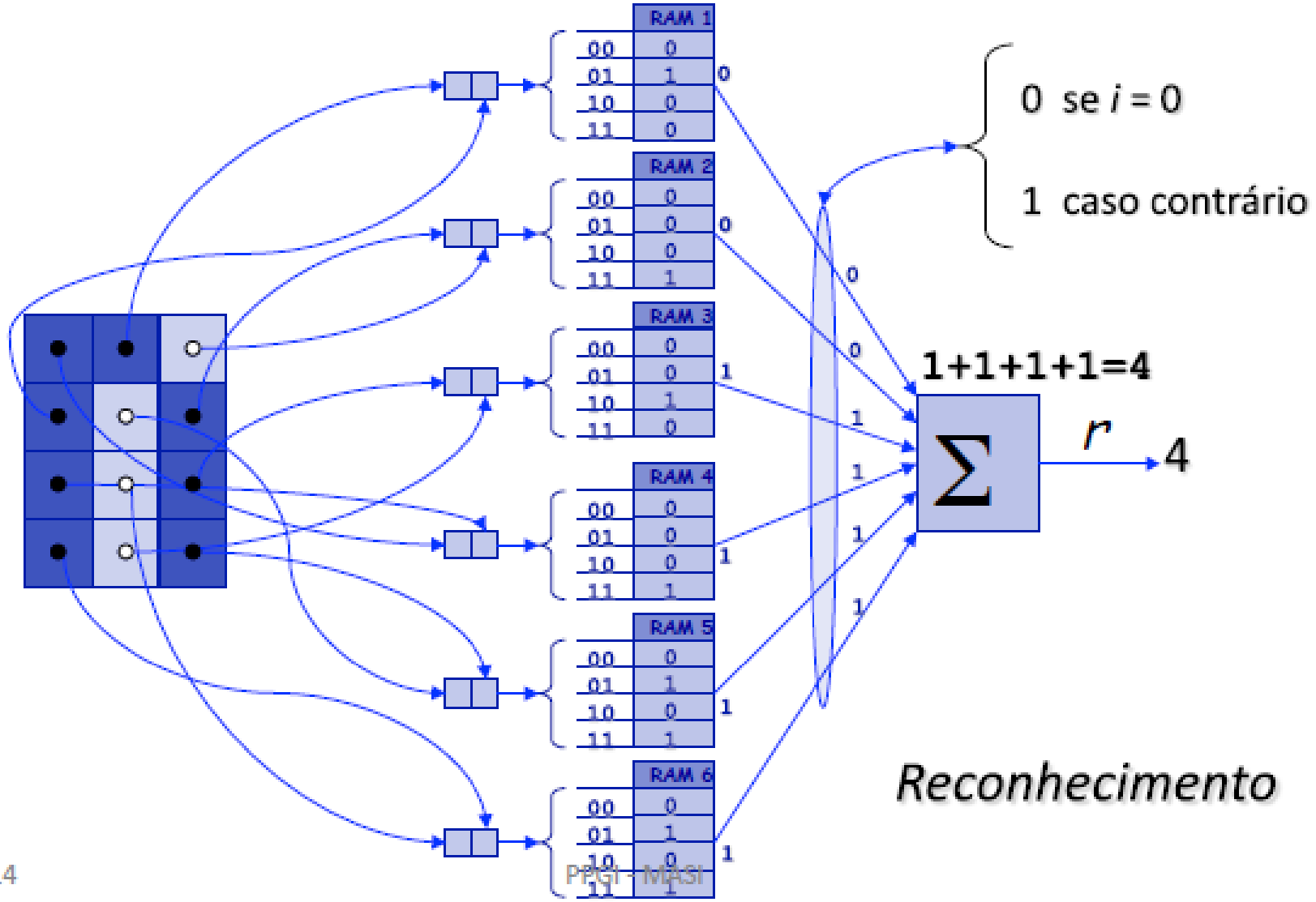
(WILKES,STONHAM, ALEKSANDER RECOGNITION DEVICE)

- PRIMEIRA MÁQUINA DE REDE NEURAL ARTIFICIAL A SER PRODUZIDA PARA COMERCIALIZAÇÃO.
- MAIS REPRESENTATIVO MODELO DE REDE NEURAL.



Treinamentos




RAM 1
00 0
01 1
10 0
11 0
RAM 2
00 0
01 0
10 0
11 1
RAM 3
00 0
01 0
10 1
11 0
RAM 4
00 0
01 0
10 0
11 1
RAM 5
00 0
01 1
10 0
11 1
RAM 6
00 0
01 1
10 0
11 1
0 se i = 0
1 caso contrário
1+1+1+1=4
r
4
Reconhecimento

Linguagem de Programação Alto Nível

Orientação a Objeto

Excelentes bibliotecas de apoio

Grande comunidade de desenvolvedores

```
        trainingSet3 = trainingSet3 + register[configs.description_index]

    ## Set biggest set
    #biggestSet = trainingSet1
    #if (len(biggestSet) < len(trainingSet2)):
    #    biggestSet = trainingSet2
    #if (len(biggestSet) < len(trainingSet3)):
    #    biggestSet = trainingSet3

    print "Vectorizing sets..."
    vectorizer = CountVectorizer (strip_accents='ascii', stop_words=None, binary=True, max_df = 0.80, min_df=0.05).fit([trainingSet1, trainingSet2, trainingSet3])
    print "Done!"

    v_1, v_2, v_3 = vectorizer.transform([trainingSet1, trainingSet2, trainingSet3])
    v_1 = v_1.toarray().ravel()
    v_2 = v_2.toarray().ravel()
    v_3 = v_3.toarray().ravel()

    print "Training Wisard"

    labels       = np.array (['ABANDONO','ROUBO DE CARGA','DESMONTE VEICULOS']) # labels
    wisard       = WISARD.WISARD (6, False)#,True, 3) # atribui a classe wisard para a
    training_set = np.array ([v_1,v_2,v_3])  # conjunto de treinamento
    wisard.fit (training_set, labels)  # associacao dos labels aos conjuntos de treinamento

    return wisard

def wisard_predict (wisard, ze):

    test_set = "ROUBO DE CARGAS REGISTRADO NA AVENIDA CAPIXABA ROUBADO BANDIDOS"

    global trainingSet1
    global trainingSet2
    global trainingSet3
    global labels

    vectorizer = CountVectorizer (strip_accents='ascii', stop_words=None, binary=True, min_df=0.00, max_df = 1.00).fit([trainingSet1, trainingSet2, trainingSet3, test_set])

    V_1, V_2, V_3, vetorized_validation_set = vectorizer.transform ([trainingSet1, trainingSet2, trainingSet3, test_set])
    vetorized_validation_set = vetorized_validation_set.toarray ().ravel ()

    X_test = np.array ([vetorized_validation_set]) # vetor de testes
    percentual = wisard.predict_proba (X_test) # calcula a probabilidade
    result     = wisard.predict (X_test) # Retorna o resultado

    print "\n----------------------------------------------"
    print 'Original Report : ' , "Roubo de Carga"
    print "Report recommendation:",  labels
    print "Classification percentage : ", (percentual)*100

__main__()
```

C:\Python27\python.exe

```
Total de registros =  845
Vectorizing sets...
Done!
Training Wisard

==========================================
Original Report : Roubo de Carga
Report recommendation: ['ABANDONO' 'ROUBO DE CARGA' 'DESMONTE VEICULOS']
Classification percentage :  [[ 11.5942029   13.04347826  75.36231884]]
Press any key to continue . . .
```

ARTIGO

3,5 Meses de trabalho

6 reuniões presenciais

(+ não presenciais)

550 linhas de código

2,565 palavras

Automatic Crime Report Classification through a Weightless Neural Network

Automatic Crime Report Classification through a Weightless Neural Network

Rafael Adnet Pinho, Walkir A. T. Brito, Claudia L. R. Motta
and Priscila Vieira Lima

Federal University of Rio de Janeiro (UFRJ)
Pos-Graduation Program in Informatics (PPGI), Rio de Janeiro, RJ - Brazil
(rafaadnet, walkir.brito)@gmail.com, (claudiam, priscila.lima)@nce.ufrj.br

Abstract. Anonymous crime reporting is a tool that helps to reduce and prevent crime occurrences. The classification of the crime reports received by the call center is necessary for the data organization and also to stipulate the importance of a particular report and its relation to others. The objective of this work is to develop a system that assists the call center's operator by recommending classification to new reports. The system uses a weightless neural network that automatically attribute a class to a report. At the end of this work it was possible to observe that automatic classifications of crime reports with high accuracy are possible using a weightless neural network.

ESANN

European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning

Bruges (Belgium), 26 - 28 April 2017

ACEITO


Today
GOLD 28°

Obrigado!

Perguntas?