Extração de Informação

Luiz Carlos d´Oleron – lcadb@cin.ufpe.br
Vítor Braga – vtb@cin.ufpe.br

Roteiro

- Introdução
- Conceitos Básicos
- Classificação de Sistemas de EI
- Aplicações
- Conclusão

Introdução

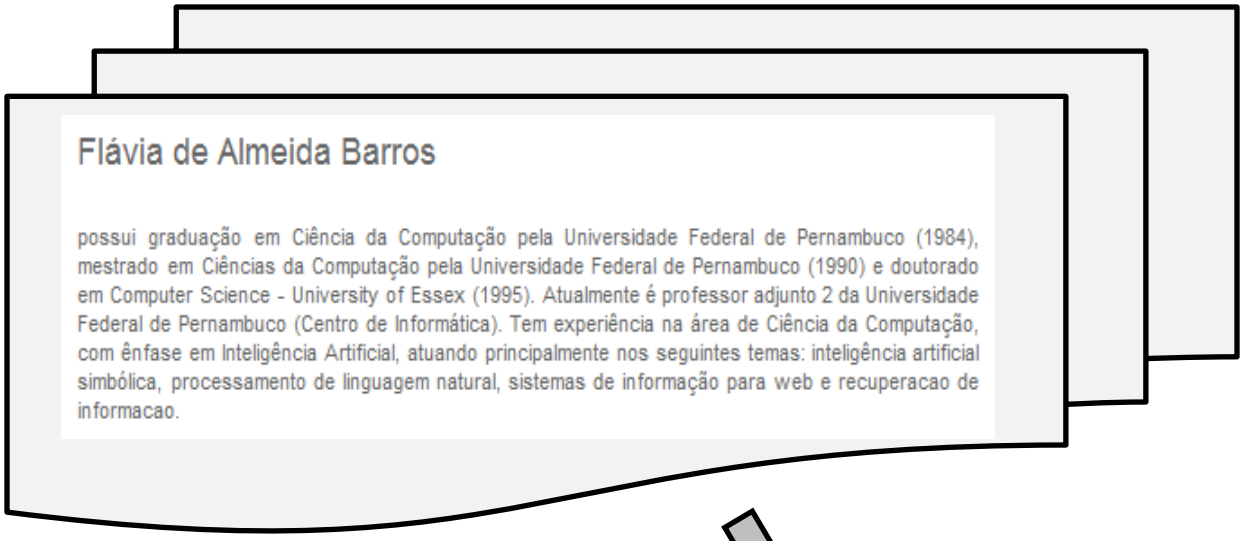
Motivação

- Problemas
  - Maior parte da informação está em forma de texto livre

- Questões importantes:
  - Como localizar informação relevante?
  - Como extrair a informação relevante?
  - Como gerar BDs ou bases de conhecimento automaticamente?

Extração de Informação (EI)

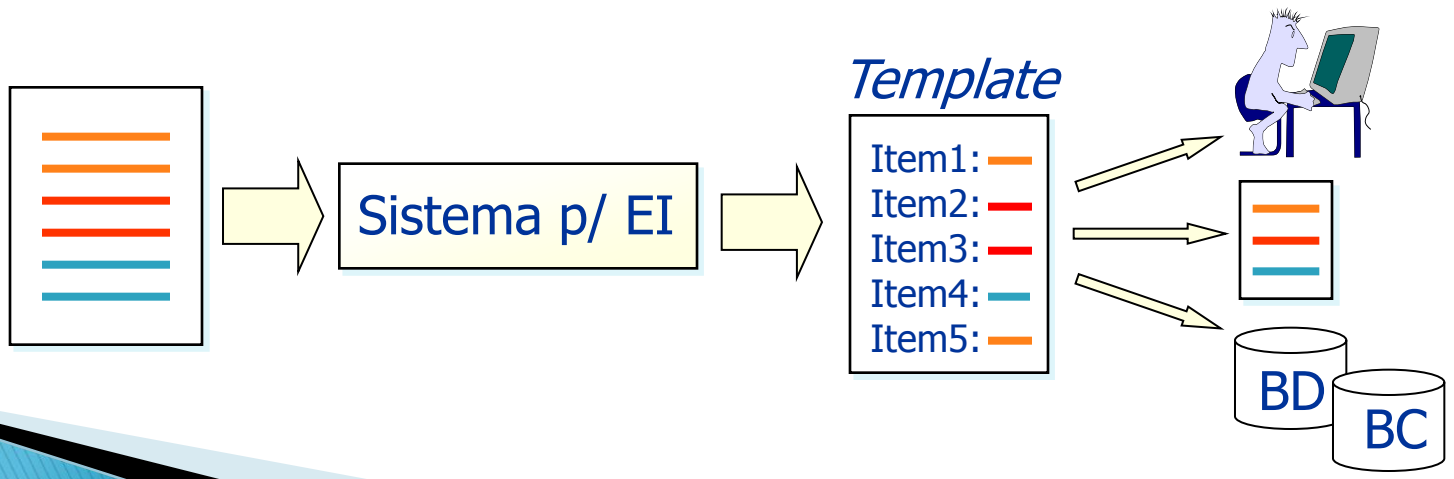
- Extração de Informação pode ajudar...
  - Trata o problema da extração de dados relevantes a partir de uma coleção de documentos [Mus99]
    - Blah blah blah *trecho relevante* blah blah blah



Nome	Titulação	Instituição de Ensino Atual	Cargo	Interesses
Flávia de Almeida Barros	Doutor	Universidade Federal de Pernambuco	Professor Adjunto 2	Inteligência artificial simbólica, Processamento de linguagem natural, Sistemas de informação para web e Recuperação de informação
Ricardo Bastos Cavalcante Prudêncio	Doutor	Universidade Federal de Pernambuco	Professor Adjunto	Redes neurais artificiais, Aprendizado de máquina, Sistemas inteligentes híbridos, Bibliotecas digitais e Recuperação de informação.

Extração de Informação (EI)

- Os dados a serem extraídos são previamente definidos em um *template* (formulário)
- Os dados extraídos podem
  - ser diretamente apresentados na tela
  - ser usados para preencher um BD ou uma BC



Extração de Informação (EI)

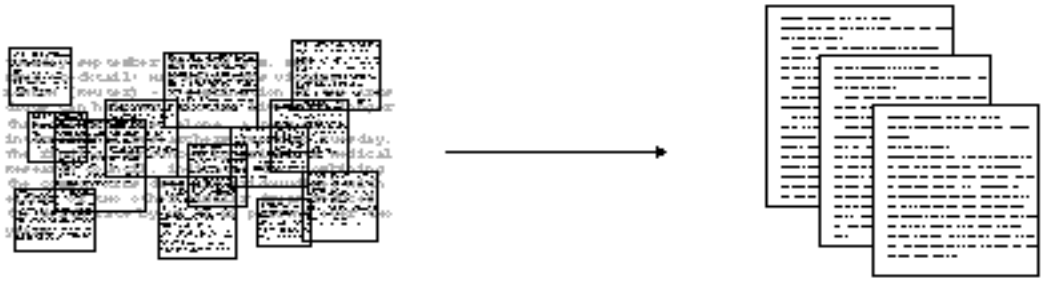
- Técnica pode ser aplicada a diferentes tipos de textos:
  - Artigos de Jornais
  - Web pages
  - Artigos Científicos
  - Mensagens de Newsgroup
  - Classificados
  - Anotações Médicas

Extração de Informação (EI)

- História
  - Década de 60
    - Processamento de Linguagem Natural
  - Década de 90
    - MUC – Message Understanding Conference
  - Após década de 90 …
    - Internet
      - Wrappers (extratores)

EI vs. Recuperação de Informação

- Recuperação de Informação:
  - Entrega documentos para o usuário



- Extração de Informação:
  - Entrega fatos para o usuário/aplicações

Por que EI é difícil?

- Línguagem Natural é difícil de tratar automaticamente
  - é muito flexível
  - várias formas para expressar uma única informação
    - *Frodo Baggins succeeds Bilbo Baggins as chairperson of Bank of America.*
    - *Bank of America named Frodo Baggins as its new chair-person after Bilbo Baggins.*
    - *Bilbo Baggins was succeeded by Frodo Baggins as chair-person of Bank of America*.

    …

Conceitos Básicos

Texto Estruturado

- Formato pré-definido e rígido
- Facilita a extração através de regras simples
  - Baseadas na ordem de apresentação
  - Rótulo das informações

Texto Estruturado

PREVISÃO		SATÉLITE - PE		MAPA DE PREVISÃO - PE		
	TEMPO	TEMP.	UMIDADE	VENTO	SOL	IUV
Quinta, 25/09	1mm	26°C 30°C	67% 82%	SE 23km/h	05h09m 17h12m	12 12
Sol, alternando com chuva em forma de pancada rápida e isolada						DETALHAR
Sexta, 26/09	0mm	26°C 30°C	64% 79%	SE 24km/h	05h08m 17h12m	11 11
Predomínio de sol, apenas com pouca variação de nuvens						DETALHAR
Sábado, 27/09	1mm	26°C 30°C	67% 81%	ESE 19km/h	05h08m 17h12m	12 12
Sol, alternando com chuva em forma de pancada rápida e isolada						DETALHAR

Previsão do Tempo

Texto extraído do Tempo Agora (UOL)

Texto Estruturado

```
<?xml version="1.0"?>
<!DOCTYPE mail system “http://infowest.com/DTDS/email.dtd”>
<email>
        <de>Autor</de>
        <para>Alguém</para>
        <data>Terça-feira – 7 de outubro de 2008</data>
        <assunto>Extração de Informação</assunto>
        <texto>Aviso:
                <p alinhamento=“justificado”>Apresentação do seminário “Extração de Informação” hoje às 14:00h </p>
                <p>Esperamos sua presença</p>
        </texto>
</email>
```

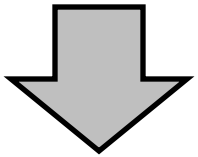
Exemplo de um documento em XML

Texto Não-Estruturado

- Sentenças escritas em alguma linguagem natural

- Requer pré-processamento linguístico

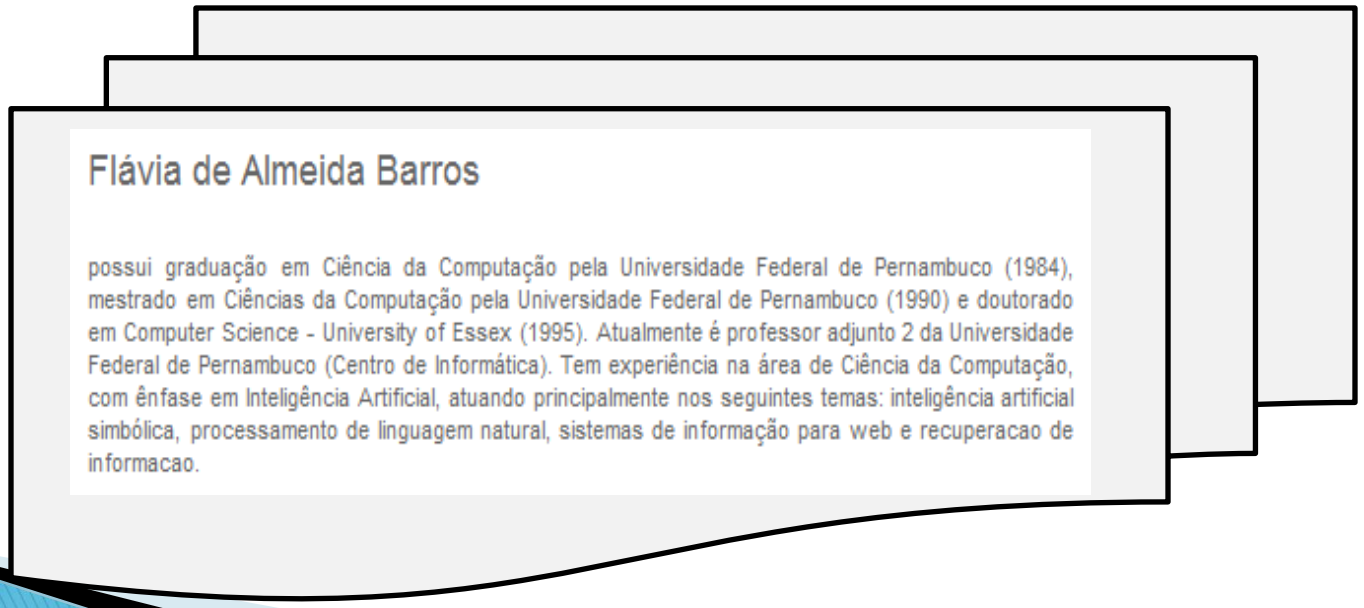
Análise sintática e semântica



Padrões de relacionamentos sintáticos e/ou semânticos

Texto Não-Estruturado

- Exemplos:
  - artigos de jornais e revistas
  - textos literários
  - cartas, etc



Texto Semi-Estruturado

- Formatação não segue regras rígidas
  - Ex: Estilo telegráfico

- Algum grau de estruturação
  - Campos ausentes
  - Variações de layout
  - Variação na ordem de apresentação dos dados

Texto Semi-Estruturado

L.J. Cahill, R. Gaizauskas, and R. Evans. (1992) POETIC: A Fully-Implemented NL System for Understanding Traffic Reports In Fully-Implemented Natural Language Understanding Systems: Proceedings of the Trento Workshop, March 30, 1992, pp. 86-99, IWBS Report No. 236, IBM Institute for Knowledge Based Systems, Heidelberg, 1992.

Yorick Wilks, Jim Cowie, Ted Dunning, and Louise Guthrie. (in press) Text processing using multi-lingual resources at CRL. In N. Ostler (ed.) Special Issue of the Journal of Literary and Linguistic Computing: Text Processing In Europe and the USA. Vol 8. Oxford: Oxford University Press. 1994

Retirado de [1]

Tipos de Texto

- Perspectivas
  - Comunidade de Inteligência Artificial (PIA)
    - Estruturados
    - Semi-estruturados
    - Não-estruturados (texto livres)
  - Comunidade de Banco de Dados (PBD)
    - Estruturados
    - Semi-estruturados

Tipos de Texto

- PIA

Estruturados	Semi-estruturados	Não-estruturados

- PBD

Estruturados	Semi-estruturados

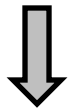
Tipo de Extração

- Obtenção das informações e relacionamentos
  - Single-slot
  - Multi-slot
- Forma de obtenção de informações complexas
  - Top-down
  - Bottom-up

Obtenção das informações e relacionamentos

- Single-Slot
  - Isola as informações em campos (slots) separados, não relacionados entre si.

Cidade Universitária. excelente 3 - qts suíte, varandão, sala 2 ambientes, dependências, nascente, garagem, guarita, R$ 750,00. novo 2 qts, sala, varanda, garagem, R$ 500,00. Próximo Bompreço. 9999-9999

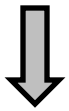


Bairro: Cidade Universitária
Bairro: Cidade Universitária
Quartos: 3
Quartos: 2
Preço: R$ 750,00
Preço: R$ 500,00

Obtenção das informações e relacionamentos

- Multi-Slot
  - Agrupa informações relacionadas em estruturas de múltiplos campos.

Cidade Universitária. excelente 3 - qts suíte, varandão, sala 2 ambientes, dependências, nascente, garagem, guarita, R$ 750,00. novo 2 qts, sala, varanda, garagem, R$ 500,00. Próximo Bompreço. 9999-9999



Bairro: Cidade Universitária Quartos: 3 Preço: R$ 750,00	Bairro: Cidade Universitária Quartos: 2 Preço: R$ 500,00

Reestruturação de informações complexas

- Top-down
  - Identificação de objetos complexos no texto.
  - Extração das informações mais simples contidas nesses objetos.
- Bottom-up
  - Identificação de todas as informações mais simples contidas no documento.
  - Agrupamentos dessas informações em estruturas mais complexas.
  -

Problemas de Extração de Informação

- Campos ausentes
  - Campos presentes em um documento e ausente em outro.
- Campos multivalorados
  - Campos relacionados a vários valores.
- Múltiplas ordens de campos
  - Variação da ordem em que campos e delimitadores aparecem em documentos do mesmo domínio.

Problemas de Extração de Informação

- Delimitadores disjuntivos
  - Um mesmo campo pode apresentar vários delimitadores diferentes.
- Delimitadores ausentes
  - Campos podem não ter delimitadores.
- Exceções e erros tipográficos
  - Erros de escrita podem inviabilizar a extração devido a variações.

Métricas de Avaliação

- Informações extraídas X Informações desejadas

		Saída	
		Presente	Ausente
Entrada	Presente	Verdadeiro positivo (TP)	Falso negativo (FN)
	Ausente	Falso positivo (FP)	Verdadeiro negativo (TN)

Métricas de Avaliação

- Precisão

$$Precisão = \frac{TP}{TP + TN}$$

- Cobertura

$$Cobertura = \frac{TP}{TP + FN}$$

- F-Measure

$$F_{Measure} = \frac{((\beta^2 + 1) * Cobertura * Precisão)}{((\beta^2) * Precisão + Cobertura)}$$

Classificação de Sistemas de EI

Tipos de Sistemas para EI

Wrappers

- Principalmente para textos estruturados e semi-estruturados
- Formatação do texto, marcadores, freqüência estatística das palavras
- Construção
  - Manual X Aprendizagem

Baseados em PLN

- Extrair informações de textos em linguagem natural (livre)
- Padrões lingüísticos

Wrappers

- Construção Manual
  - Baseada em engenharia do conhecimento
    - Construção manual de regras de extração
    - Padrões de extração são descobertos por especialistas após examinarem o corpus de treinamento
  - Vantagens
    - Boa performance dos Sistemas
  - Desvantagens
    - Processo de desenvolvimento trabalhoso
    - Escalabilidade
    - Especialista pode não estar disponível

Wrappers

- Construção Automática
  - Aprendizagem de máquina
    - Aprender sistemas de EI a partir de um conjunto de treinamento
  - Vantagens
    - Mais fácil marcar um corpus do que criar regras de extração
    - Menor esforço do especialista
    - Escalabilidade
  - Desvantagens
    - Esforço de marcação do corpus de treinamento

Wrappers

- Técnicas de Extração
  - Autômatos Finitos
  - Casamento de Padrões
  - Classificação de Textos
  - Modelos de Markov Escondidos

Autômatos Finitos

- Regras de extração na forma de autômatos finitos

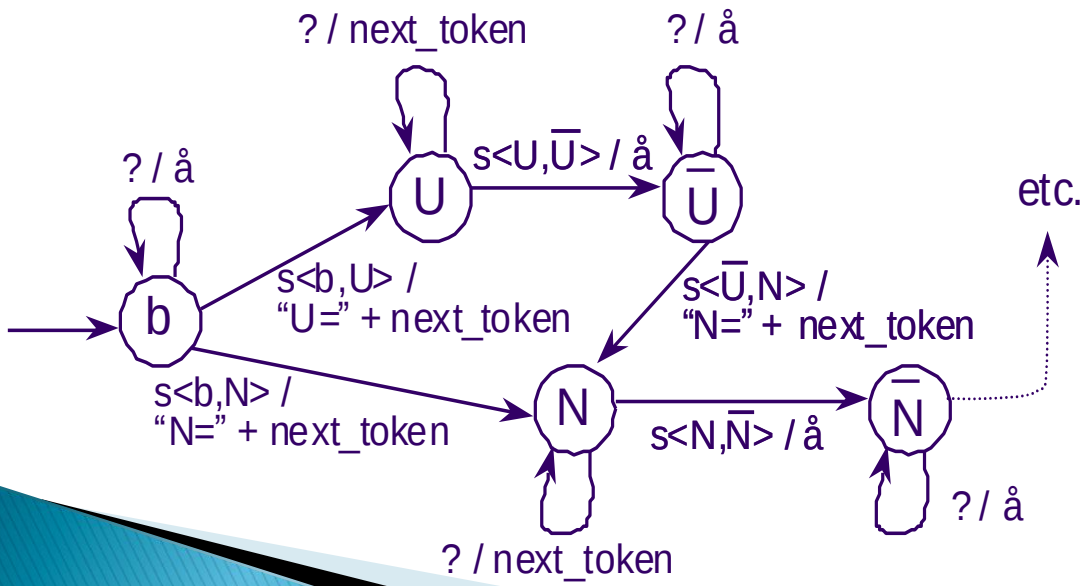
- Definidos por:
  - (1) estados que “aceitam” os símbolos do texto que preenchem algum campo do formulário de saída,
  - (2) os estados que apenas consomem os símbolos irrelevantes encontrados no texto, e
  - (3) os símbolos que provocam as transições de estado

- Textos estruturados e semi-estruturados
  - Delimitadores, ordem dos elementos

Autômatos finitos

- Exemplo

```
<LI> <A HREF="…"> Mani Chandy </A>, <I>Professor of Computer
    Science</I> and <I>Executive Officer for Computer Science</I>
…
<LI> Fred Thompson, <I>Professor Emeritus of Applied Philosophy and
    Computer Science</I>
```



Key

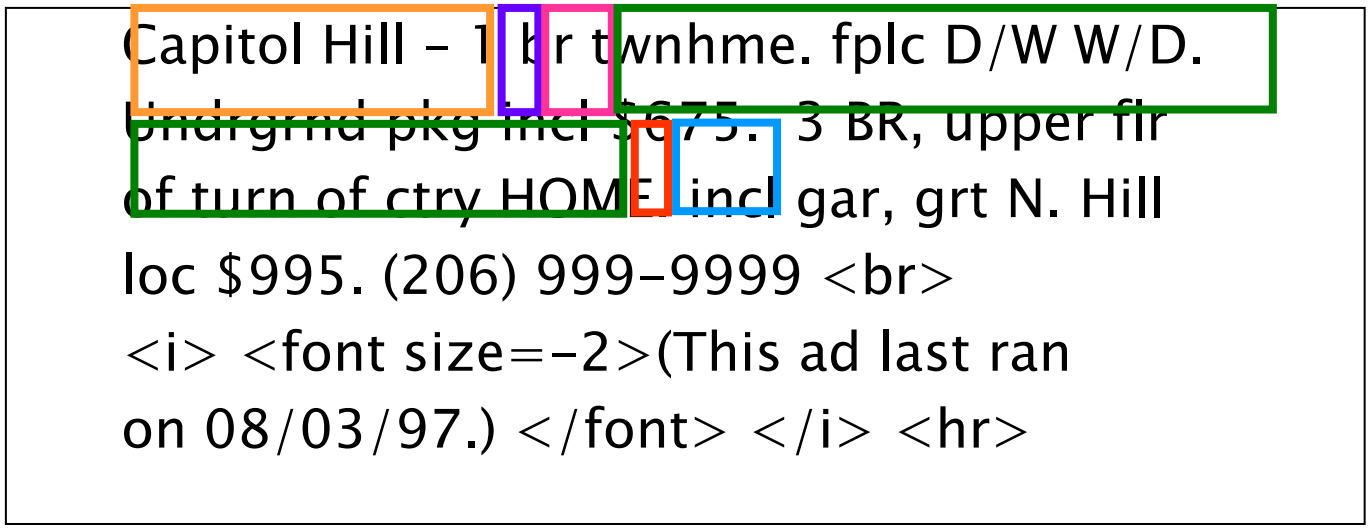
- ? : wildcard
- U : state to extract URL
- Ū : state to skip over tokens until we reach N
- N : state to extract Name
- N̄ : state to skip over tokens until we reach A
- s<X,Y> : separator rule for the separator of states X and Y
- etc.

Casamento de Padrões

- Aprendem regras na forma de expressões regulares
- Expressões regulares que “casam” com o texto para extrair as informações
- Textos livres, estruturados e semi-estruturados
  - Delimitadores, padrões regulares (Ex. data, CEP)

Casamento de Padrões

Padrão :: * (*Digit*) ‘BR’ * ‘$’ (*Number*)
Formulário:: Aluguel {Quartos $1}{Preço $2}



Classificação de Textos

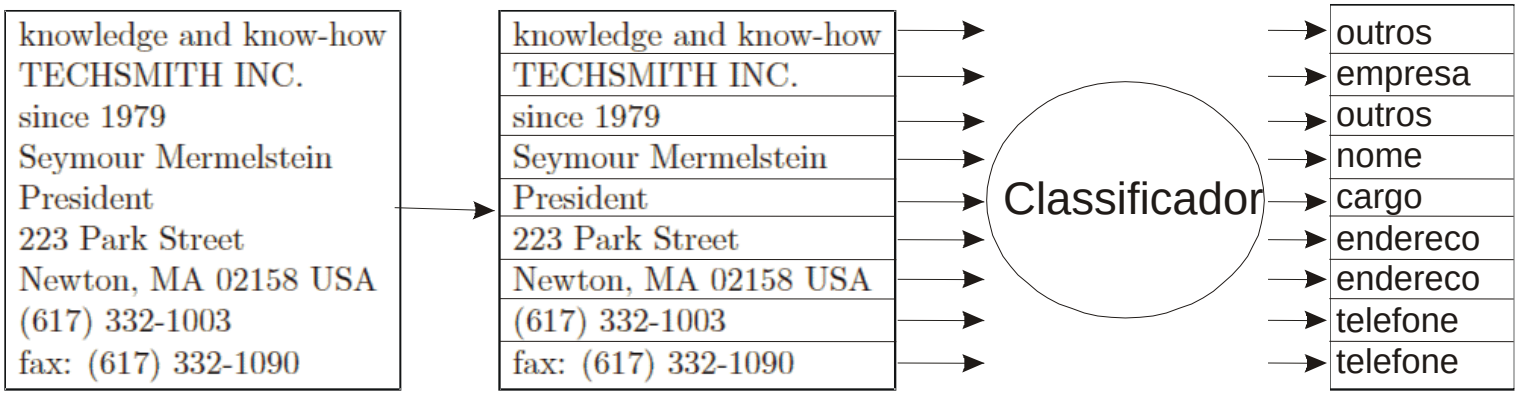
- Dividem o texto de entrada em fragmentos candidatos a preencher algum campo do formulário de saída.

- Classificam os fragmentos com base em suas características
  - posição
  - número de palavras
  - presença de palavras específicas
  - letras capitalizadas

- Textos semi-estruturados

Classificação de Textos

- Classificam fragmentos do documento para determinar que campo do formulário eles devem preencher



Desvantagem

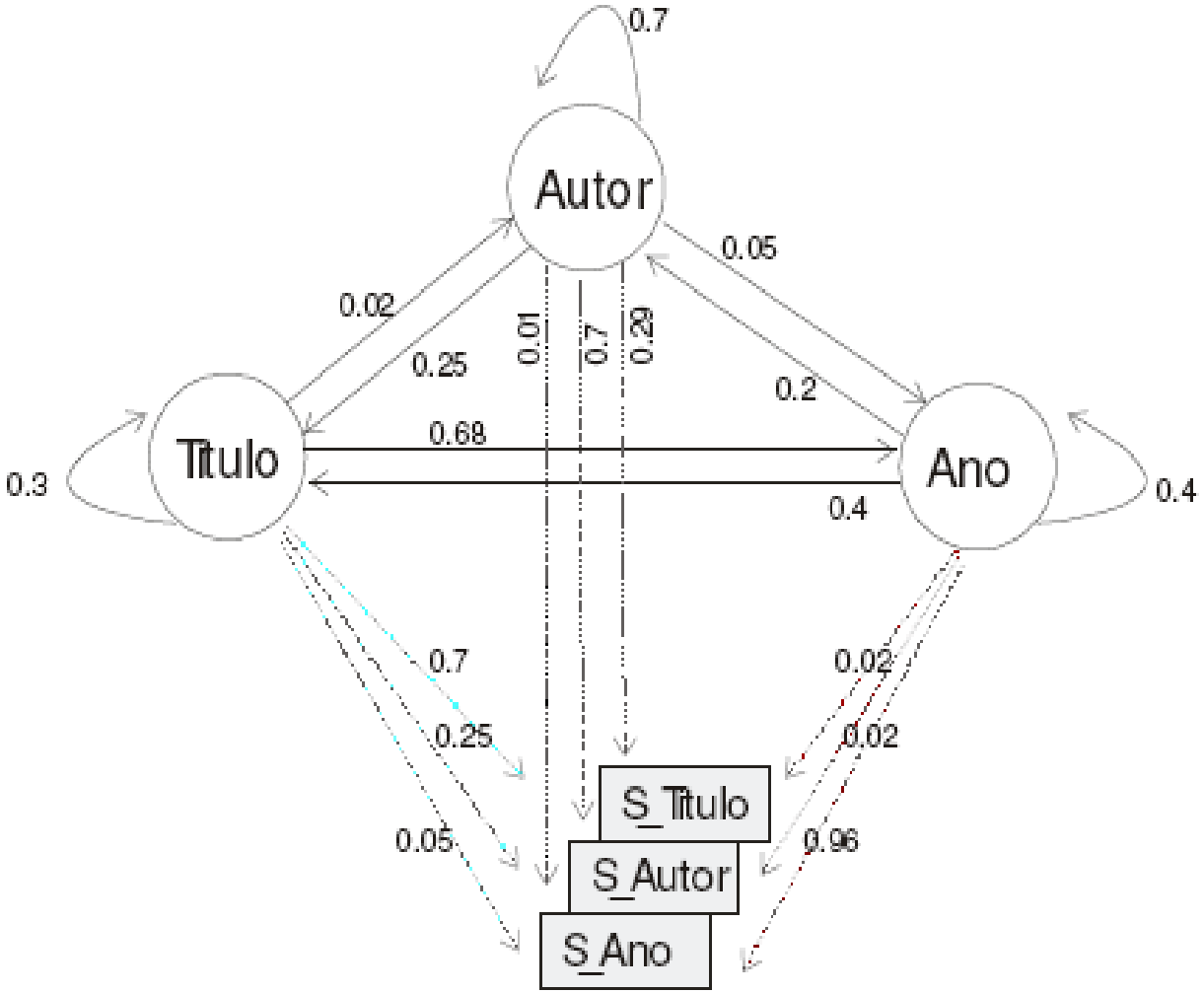
Classificação local independente para cada fragmento

Modelos de Markov Escondidos (HMM)

- Um HMM é um autômato finito probabilístico que classifica seqüências de entrada
- Processo de classificação
  - Retorna a seqüência de campos com maior probabilidade para uma sequencia de fragmentos de entrada
- Vantagem
  - Realizar uma classificação ótima para a seqüência completa de entrada.

Modelos de Markov Escondidos (HMM)

- Exemplo:



Exemplos

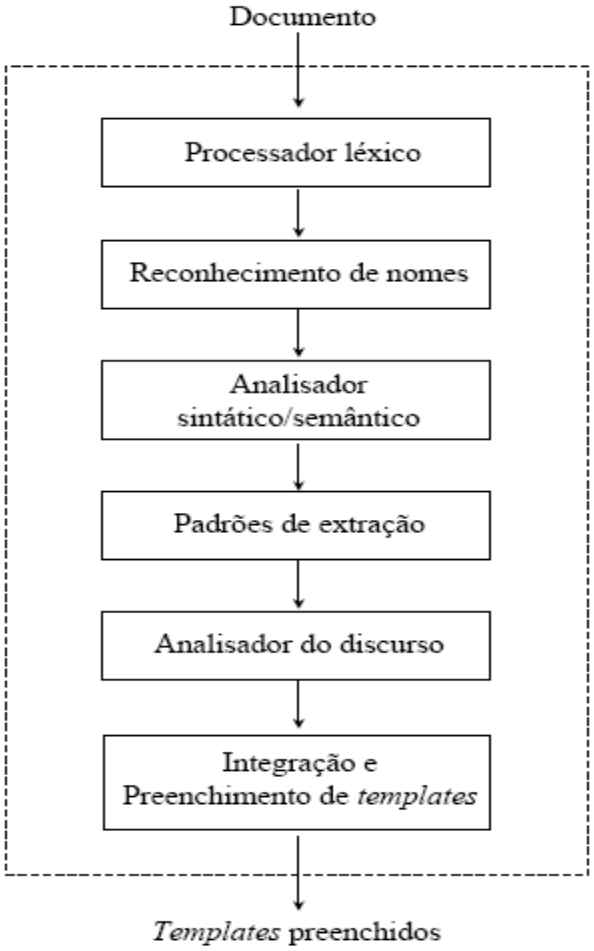
- Autômatos Finitos
  - Stalker
  - WIEN
  - SoftMealy
- Casamento de Padrões
  - Whisk
  - Rapier
- Classificação de Textos
  - SRV
- Modelos de Markov Escondidos
  - DATAMOLD

Processamento de Linguagem Natural

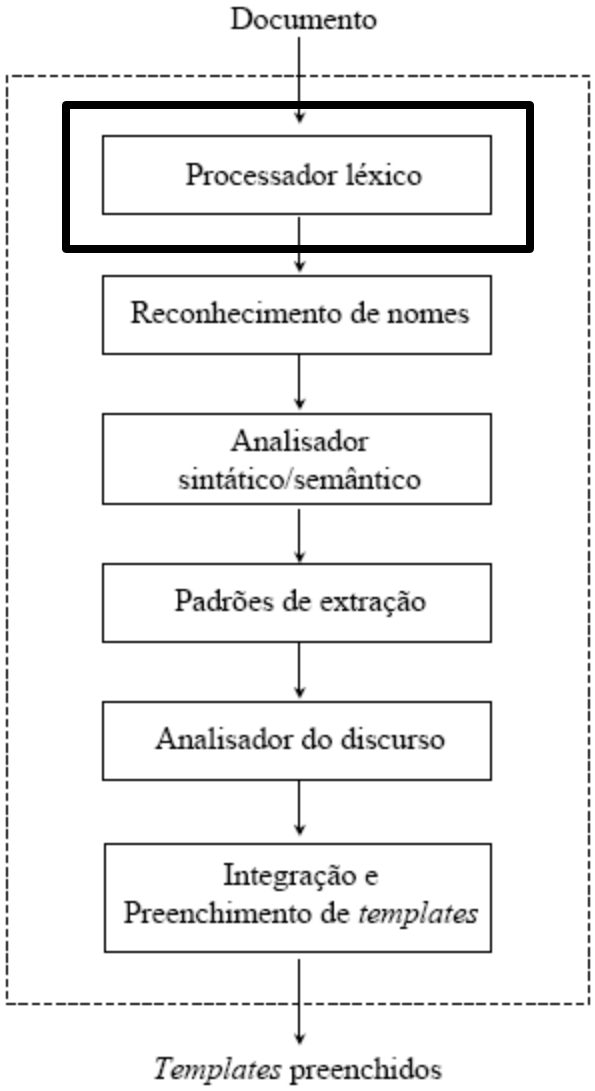
- **Processo de extração**
  - Extração de fatos (unidades de informação)
    - Através da análise local do texto
  - Integração e combinação de fatos
    - Produção de fatos maiores ou novos fatos
  - Estruturação de fatos relevantes
    - Padrão de saída

Processamento de Linguagem Natural

- **Arquitetura**

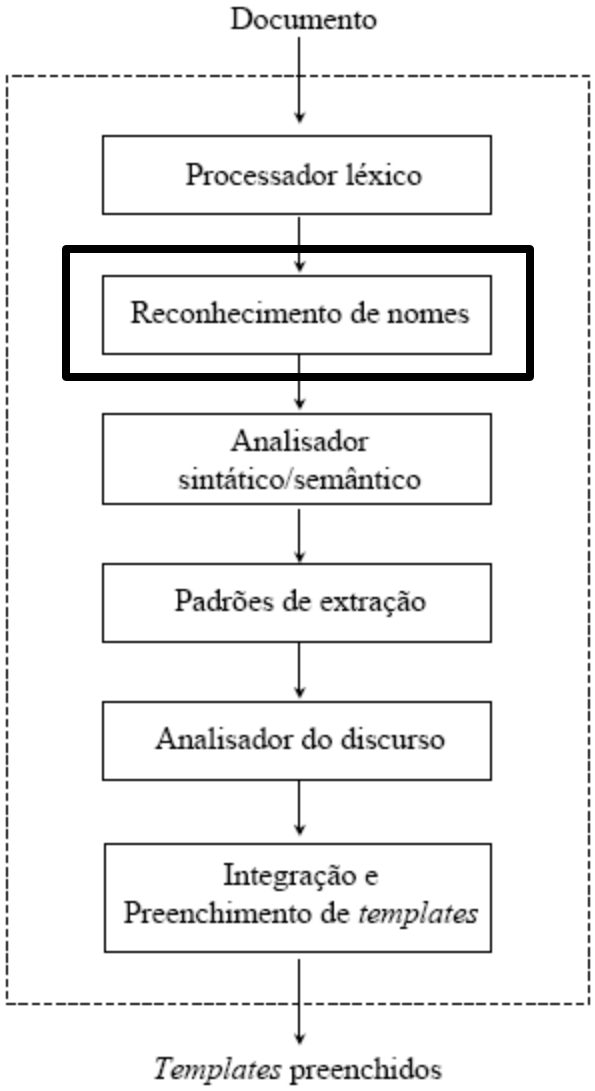


Processador Léxico



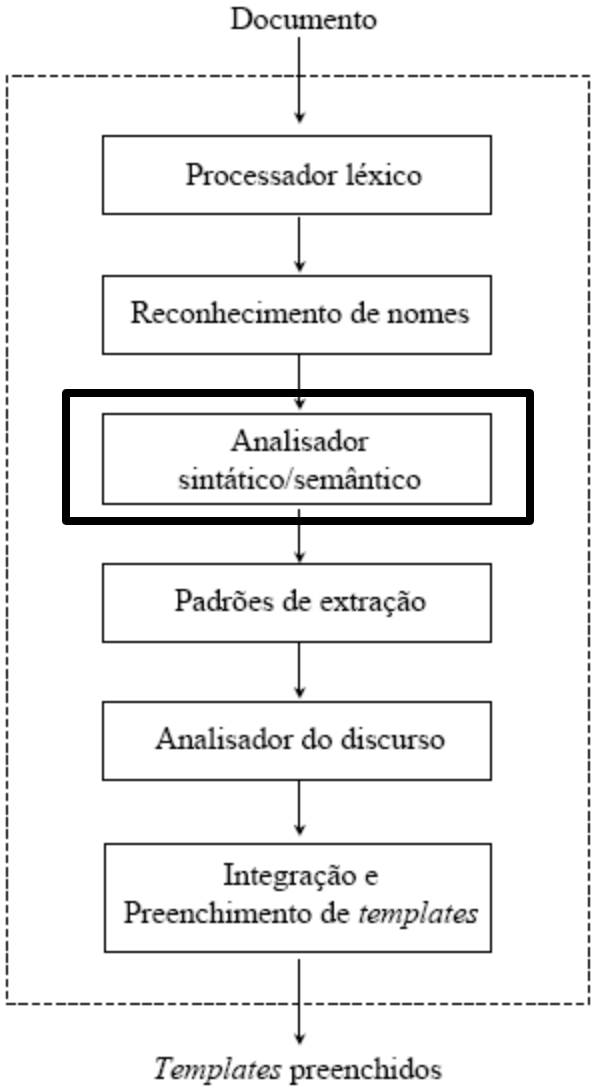
- Separação dos termos (*tokenization*) pelo reconhecimento de espaços em branco e sinais de pontuação que delimitam o texto;
- Análise léxica e morfológica dos termos para determinar suas possíveis classes (substantivo, verbo, etc.) e outras características (masculino, feminino);
- É comum o uso de autômatos finitos para o reconhecimento das informações

Reconhecimento de Nomes



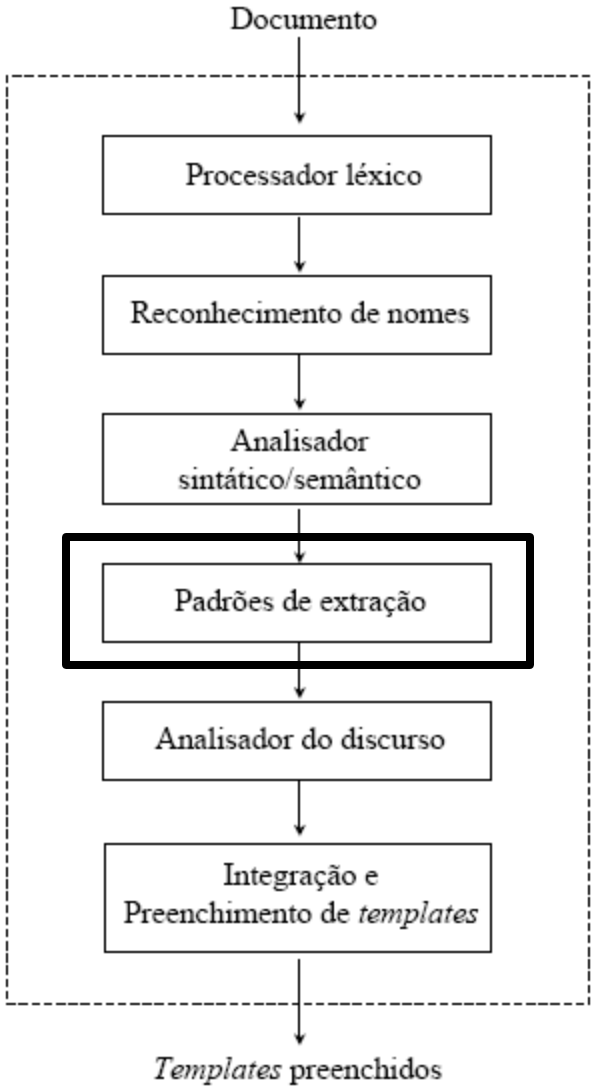
- Identifica nomes próprios;
- Itens que têm estrutura interna como da data e hora;
- Nomes são identificados por expressões regulares expressos em função das classes morfossintáticas (part-of-speech) e características sintáticas e ortográficas (letras maiúsculas) presentes nos termos.

Analisador Sintático/Semântico



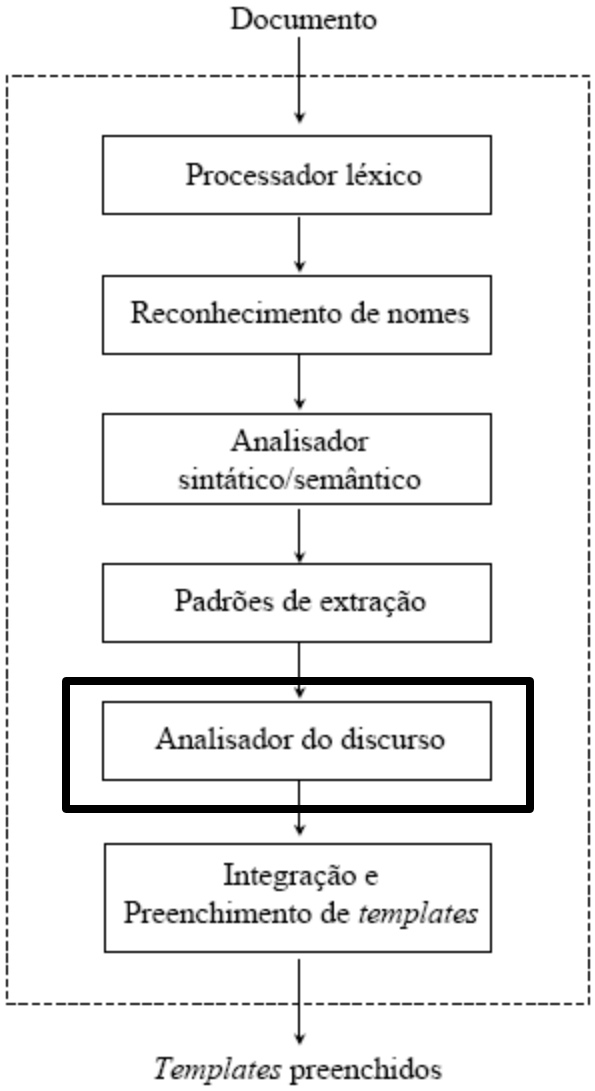
- Recebe uma seqüência de itens léxicos e tenta construir uma estrutura sintática junto com alguma semântica;
- Identifica os segmentos de texto e para cada um associa alguma característica que podem ser combinadas na fase seguinte.

Padrões de Extração



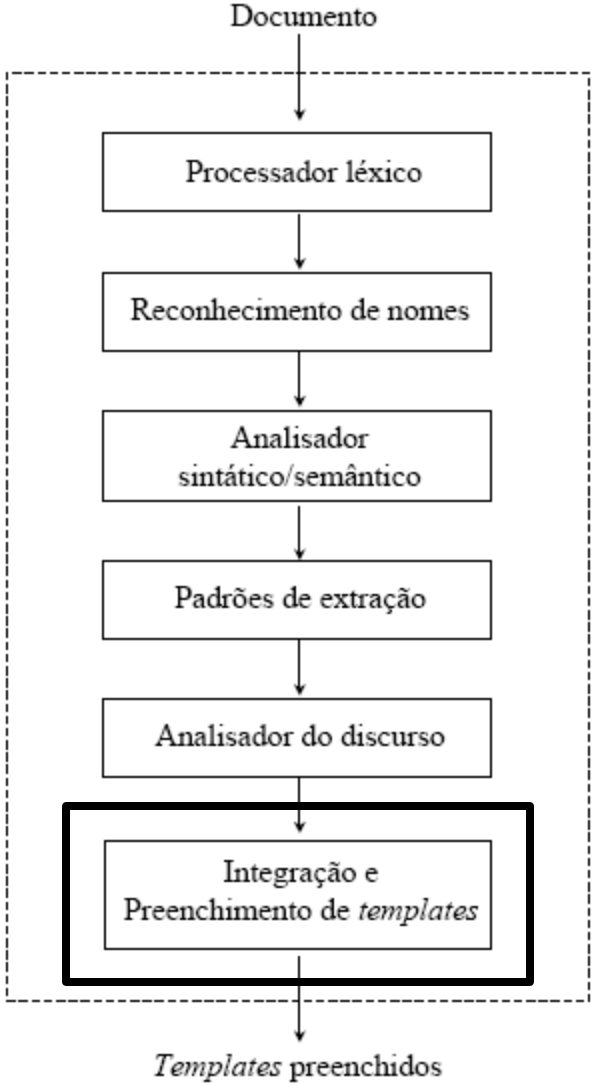
- Consiste na indução de um conjunto de regras de extração para o domínio tratado;
- Esses padrões baseiam-se em restrições sintáticas e semânticas aplicadas as sentenças.

Analisador do Discurso



- Relaciona diferentes elementos do texto;
- Análise de frases nominais, reconhece apostos e outros grupos nominais complexos;
- Resolução de conferência, identifica quando uma frase nominal se refere a outra já citada;
- Descoberta de relacionamento entre as partes do texto, para estruturar palavras do texto em uma rede associativa.

Interpretação e Preenchimento de *Templates*



- As informações são combinadas
- Os templates são preenchidos com as informações relevantes ao domínio

Resumo dos tipos de documentos e técnicas de extração

Tipos de Textos	Características dos Textos	Bases das Técnicas	Aprendizagem de Regras	Exemplos
Texto Livre	Gramaticalmente inflexível	Ontologia	Manual	BYU
		PLN (sintaxe e semântica)	Semi-automática	SRV, RAPIER e WHISK
Texto Semi-estruturado	Flexibilidade gramatical e estrutural	Linguagem para *Wrappers*	Manual	Minerva, SIMMIS e WebOQL
		Probabilidade	Semi-automática	Redes Bayesianas e HMM
		Indução (lógica indutiva)	Semi-automática	WIEN, SoftMealy e STALKER
		Modelagem (tuplas, listas, etc.)	Semi-automática (tendendo para Automática)	NoDoSE e DEByE
Texto Estruturado	Estrutura tabular rigidamente estruturada	HTML-ware (dependência com HTML)	Automática	W4F, Xwrap e RoadRunner

Algumas Aplicações

Open Amplify™

- API para análise semântica

Input:
Text content from any source

Amplify™
The meaning platform

Output:
13 different 'Signals', as XML

XML

```
</DomainResult>
</Domains>
- <TopTopics>
  - <TopicResult>
    - <Topic>
        <Name>Obama</Name>
        <Value>18.000000</Value>
      </Topic>
    - <Polarity>
      - <Min>
          <Name>Negative</Name>
          <Value>-0.753258</Value>
        </Min>
      - <Mean>
          <Name>Positive</Name>
          <Value>0.215225</Value>
        </Mean>
      - <Max>
          <Name>Positive</Name>
          <Value>1.000000</Value>
        </Max>
      </Polarity>
```

Demo: Vídeo


Open Amplify
OpenAmplify Demonstration:
Online Advertising
Amplify
www.openamplify.com
0:05 / 3:31
HQ

Algumas aplicações

Gmail Add-on

TweetyPants

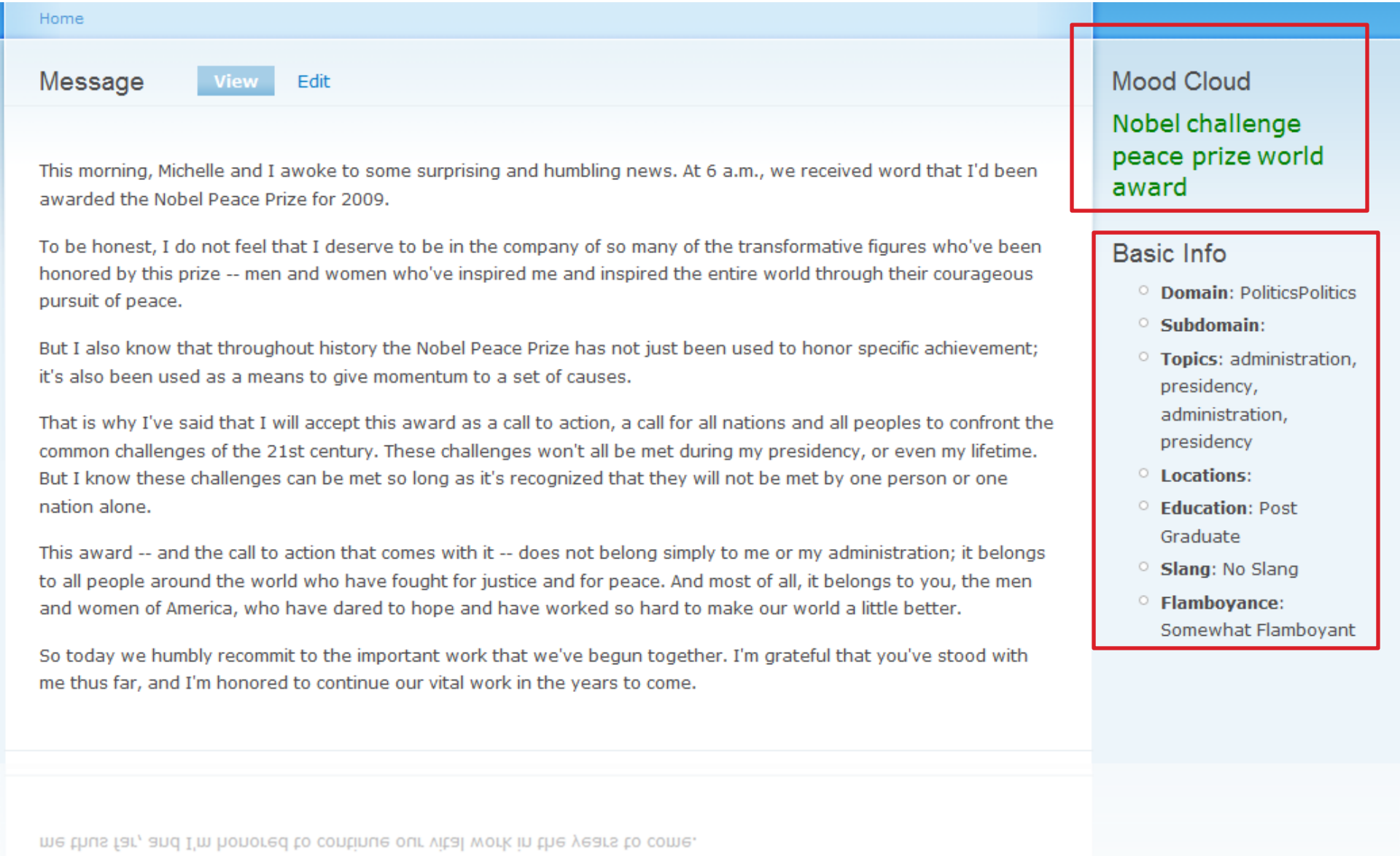
Are your Twitter friends smarty-pantses?

Twitter Friend vitorvtb Go

vitorvtb talks about:

de Veloxultra Estelionato Programmable Web RT Andrelmaraujo RT Bob_fernandes RT Brunobpe RT Commentlab RT Farleymillano RT Igormoura RT Johandenhaan RT Leozeba RT Mesh RT Modeldriven RT Napipoca RT Neto RT Recifetequer RT Rodrigoteoria Recife SOA Tem Wolfram Alpha Easter Zapatero Adobemax Andrelmaraujo Bmfej Copa Criana Espanha Faa Fiz Fofinha Johandenhaan Legal Lula Marciobezerra

Teste : Módulo para Drupal



PERSONAS

HOW DOES THE INTERNET SEE YOU?

LAUNCH PERSONAS

MUSIC

SPORTS

EDUCATION

ART

Feedback

WHAT IS PERSONAS?

Personas is a component of the Metropath(ologies) exhibit, recently on display at the MIT Museum by the Sociable Media Group from the MIT Media Lab (Please contact us if you want to show it next!). It uses sophisticated natural language processing and the Internet to create a data portrait of one's aggregated online identity. In short, Personas shows you how the Internet sees you.

HOW DOES IT WORK?

Enter your name, and Personas scours the web for information and attempts to characterize the person - to fit them to a predetermined set of categories that an algorithmic process created from a massive corpus of data. The computational process is visualized with each stage of the analysis, finally resulting in the presentation of a seemingly authoritative personal profile.

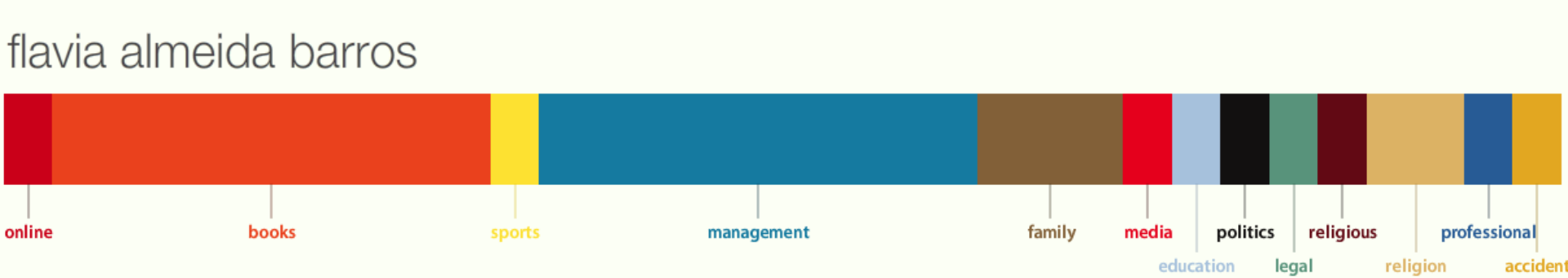
PHILOSOPHY

In a world where fortunes are sought through data-mining vast information repositories, the computer is our indispensable but far from infallible assistant. Personas demonstrates the computer's uncanny insights and its inadvertent errors, such as the mischaracterizations caused by the inability to separate data from multiple owners of the same name. It is meant for the viewer to reflect on our current and future world, where digital histories are as important if not more important than oral histories, and computational methods of condensing our digital traces are opaque and socially ignorant.

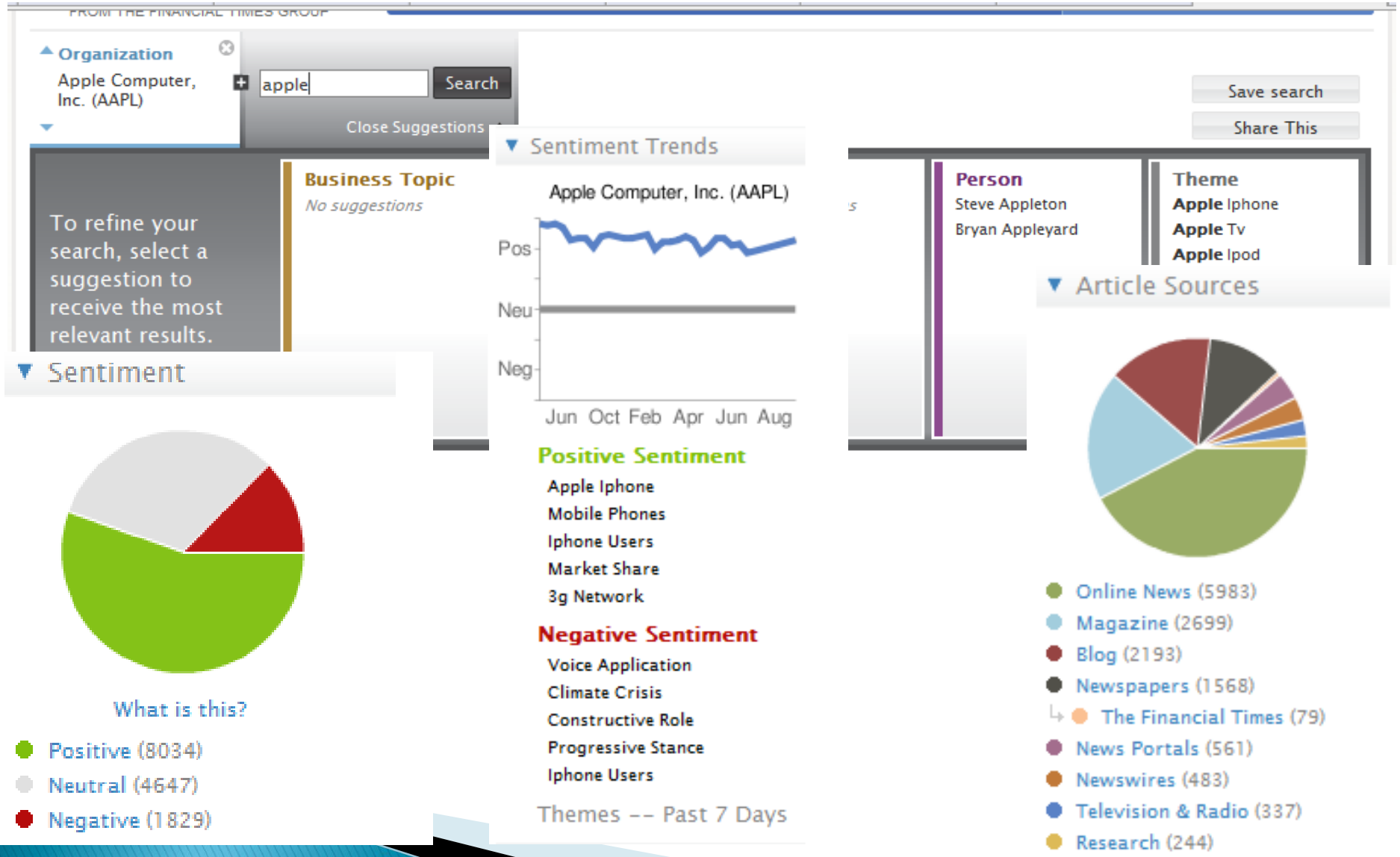
CREDITS

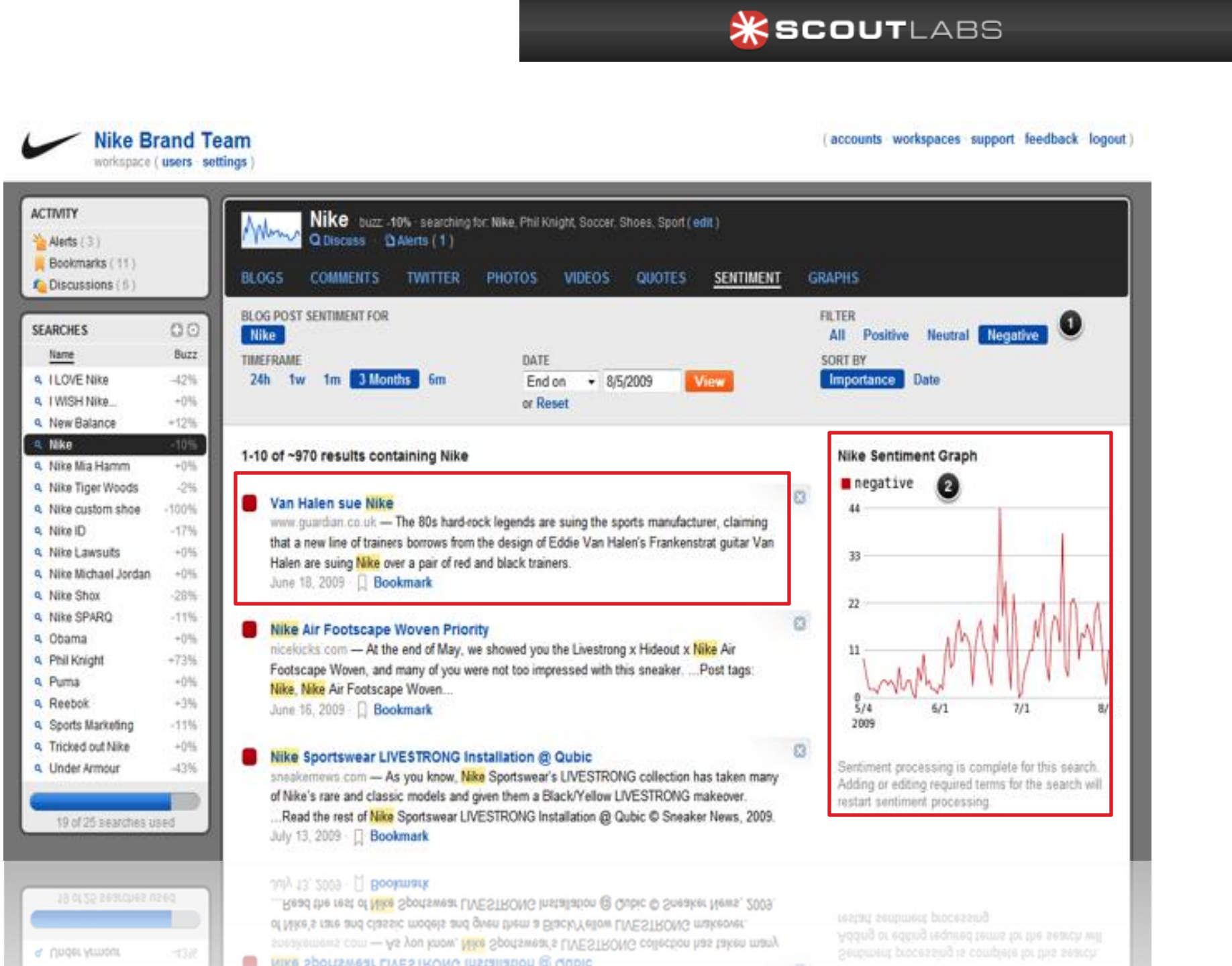
SCREENSHOTS

Personas


flavia almeida barros
online
books
sports
management
family
media
politics
religious
professional
education
legal
religion
accident
start over

Análise de “sentimentos”


Newssift
BETA
FROM THE FINANCIAL TIMES GROUP
Organization
Apple Computer, Inc. (AAPL)
apple
Search
Close Suggestions
Save search
Share This
To refine your search, select a suggestion to receive the most relevant results.
Business Topic
No suggestions
Sentiment Trends
Apple Computer, Inc. (AAPL)
Pos
Neu
Neg
Jun Oct Feb Apr Jun Aug
Person
Steve Appleton
Bryan Appleyard
Theme
Apple Iphone
Apple Tv
Apple Ipod
Sentiment
What is this?
Positive (8034)
Neutral (4647)
Negative (1829)
Positive Sentiment
Apple Iphone
Mobile Phones
Iphone Users
Market Share
3g Network
Negative Sentiment
Voice Application
Climate Crisis
Constructive Role
Progressive Stance
Iphone Users
Themes -- Past 7 Days
Article Sources
Online News (5983)
Magazine (2699)
Blog (2193)
Newspapers (1568)
The Financial Times (79)
News Portals (561)
Newswires (483)
Television & Radio (337)
Research (244)


SCOUTLABS
Nike Brand Team
workspace (users settings)
(accounts workspaces support feedback logout)
ACTIVITY
Alerts (3)
Bookmarks (11)
Discussions (6)
SEARCHES
Name
Buzz
I LOVE Nike -42%
I WISH Nike... +0%
New Balance +12%
Nike -10%
Nike Mia Hamm +0%
Nike Tiger Woods -2%
Nike custom shoe -100%
Nike ID -17%
Nike Lawsuits +0%
Nike Michael Jordan +0%
Nike Shox -28%
Nike SPARQ -11%
Obama +0%
Phil Knight +73%
Puma +0%
Reebok +3%
Sports Marketing -11%
Tricked out Nike +0%
Under Armour -43%
19 of 25 searches used
Nike buzz -10% · searching for: Nike, Phil Knight, Soccer, Shoes, Sport (edit)
Discuss · Alerts (1)
BLOGS COMMENTS TWITTER PHOTOS VIDEOS QUOTES SENTIMENT GRAPHS
BLOG POST SENTIMENT FOR
Nike
TIMEFRAME
24h 1w 1m 3 Months 6m
DATE
End on 8/5/2009 View
or Reset
FILTER
All Positive Neutral Negative
1
SORT BY
Importance Date
1-10 of ~970 results containing Nike
Van Halen sue Nike
www.guardian.co.uk — The 80s hard-rock legends are suing the sports manufacturer, claiming that a new line of trainers borrows from the design of Eddie Van Halen's Frankenstrat guitar Van Halen are suing Nike over a pair of red and black trainers.
June 18, 2009 · Bookmark
Nike Air Footscape Woven Priority
nicekicks.com — At the end of May, we showed you the Livestrong x Hideout x Nike Air Footscape Woven, and many of you were not too impressed with this sneaker. ...Post tags: Nike, Nike Air Footscape Woven...
June 16, 2009 · Bookmark
Nike Sportswear LIVESTRONG Installation @ Qubic
sneakernews.com — As you know, Nike Sportswear's LIVESTRONG collection has taken many of Nike's rare and classic models and given them a Black/Yellow LIVESTRONG makeover. ...Read the rest of Nike Sportswear LIVESTRONG Installation @ Qubic © Sneaker News, 2009.
July 13, 2009 · Bookmark
Nike Sentiment Graph
negative
2
44
33
22
11
0
5/4
6/1
7/1
2009
Sentiment processing is complete for this search. Adding or editing required terms for the search will restart sentiment processing.

e.life
we deliver insights ™

Monitorar e Analisar: BuzzMonitor

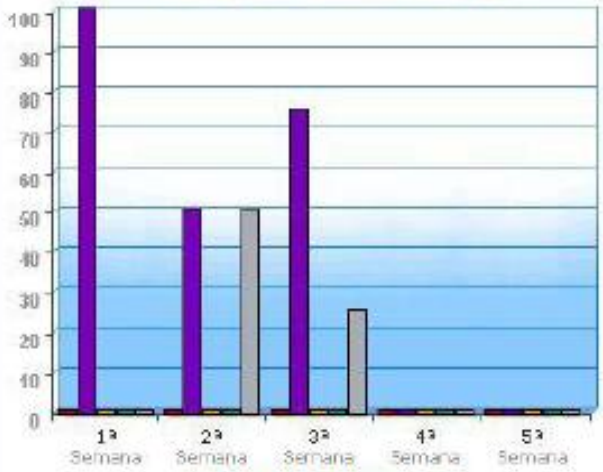
English | Español | Português

e.life BUZZ monitor 27 postings e 12 comentários monitorados desde 24/01/2008
27 postings e 12 comentários no mês de outubro/2008

MONSANTO imagine

Exibir resultados: Hoje | Desta semana | Deste mês | Todos Buscar: Logout
Por data de: Coleta | Publicação SAC Busca avançada Gerar Relatório

EVOLUÇÃO: OUTUBRO - 2008 (%)



MÉDIA NO PERÍODO: 1,25

TAGS	INTENSIDADE
AÇÃO IMEDIATA - 0	[5] MUITO IMPORTANTE
NEGATIVO - 1	[4] IMPORTANTE
MIXED - 2	[3] REGULAR
NEUTRO - 3	[2] FRACO
POSITIVO - 4	[1] IRRELEVANTE

Relatório de contatos

OUTUBRO - 2008

DOM	SEG	TER	QUA	QUI	SEX	SAB
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

DESTAQUE
13/10, 00:22
"Soja transgênica: qual o problema? ... As experiências " (...) [4]

NUVEM DE TERMOS
biotecnologia compra produção mst campesina agricultores mercado experimentos **sementes** **brasil** via **milho** países forma grande **monsanto** reforma **área** plantas cultivo anos ano agrária produtos **transgênica** 2007 **empresas**

Ranking AÇÃO IMEDIATA
Postings
Sábado, 31/05/2008
"mais uma denúncia contra a monsanto e akatugente, a mon" (...) [5]
tags: Negativo , Ação Imediata [reclassificar]
Comentários: Sábado, 31/05/2008
"ajudajá solicitei a ajuda do greenpeace para tenta" (...)
tags: Negativo , Ação Imediata [reclassificar]

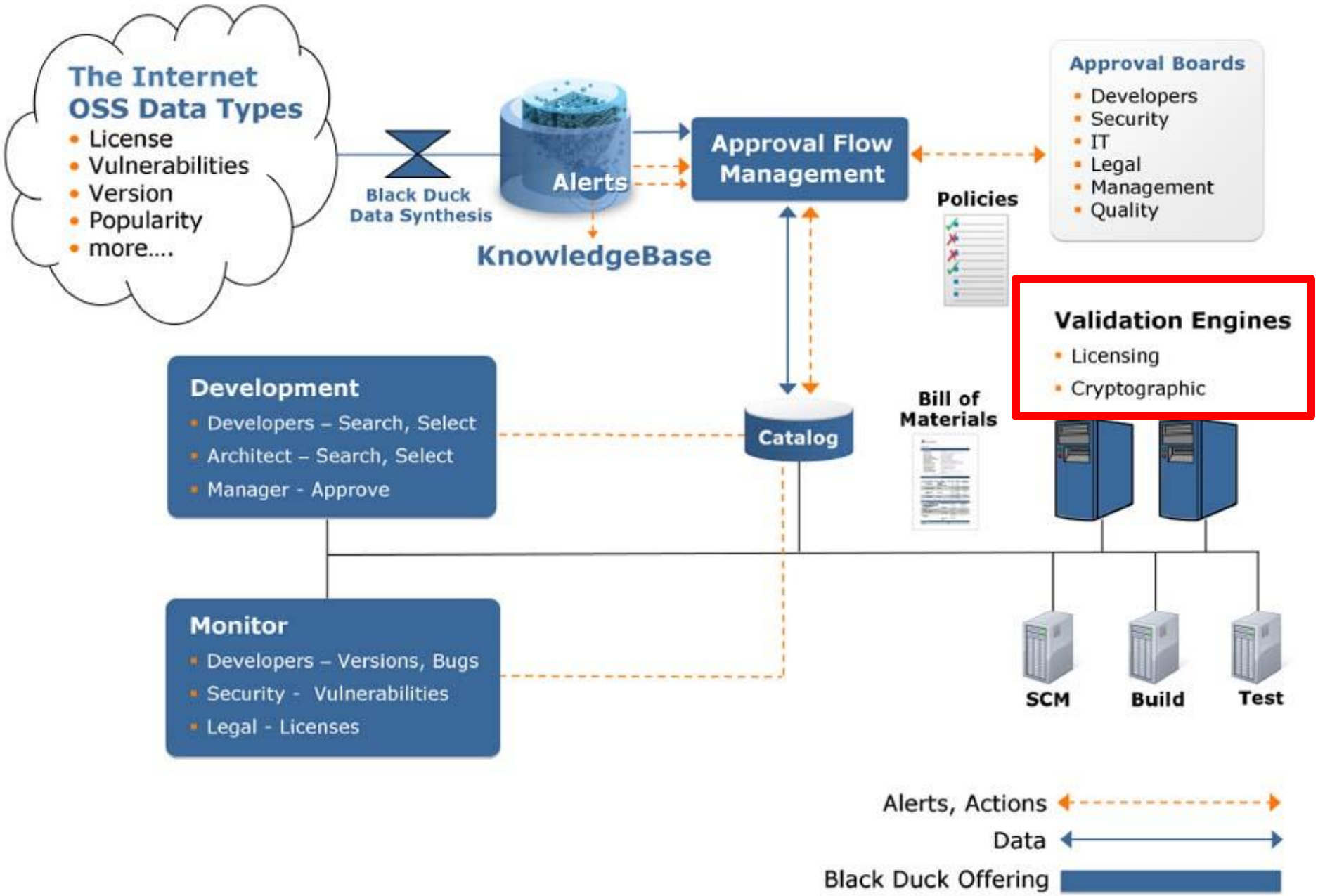
Ranking NEGATIVOS [Mais +]
Postings
Terça-feira, 14/10/2008
"óleo de canola - a perfeição humana da canola ... o óle" (...) [4]
tags: Negativo , Destaque [reclassificar]
Terça-feira, 14/10/2008
"monsanto: ameaça dos transgênicosvídeo que expõe o inde" (...) [5]
tags: Negativo , Destaque [reclassificar]
Segunda-feira, 13/10/2008
"os 7 pecados capitais dos transgênicos/conheça os princ" (...) [4]
tags: Negativo , Destaque [reclassificar]
Sábado, 11/10/2008
"tensão entre colômbia e equador esconde crise humanitá" (...) [4]
tags: Negativo , Destaque [reclassificar]
Segunda-feira, 06/10/2008
"transgênicos - os 7 pecados capitais. ... no brasil, in" (...) [5]

Outras Aplicações

blackduck™

- Análise do Código Fonte de Aplicações
  - base de dados gigantesca
  - Procura de componentes
  - Rastreamento

The Black Duck Suite



a.m.i.g.o.s.

oro·aro
AGORA SÓ FALTA VOCÊ!

HOME | HISTÓRIAS | CONTATOS | COMUNIDADES | OBJETOS | PROJETOS | MERCADO

vitor Braga

Online

Criar história

Adicionar Objeto

Mensagens (5 novas)

MENSAGENS

Exibir Todas

Todas
Contatos que me Adicionaram
Mensagens do Administrador
Apenas Recomendações
Recomendações de Histórias
Recomendações de Comunidades
Recomendações de Tópicos
Recomendações de Objetos
Recomendações de Usuários
Não Lidas
Mensagens gerais

1 - 5 de 5 resultados

Assunto		
Parada para manutenção do oro-aro.com		
Ambiente fora do ar para manutenção		
Manutenção no link de Internet		
você já criou seu usuário no twitter e está SEGUINDO in953?	Administrador	19/03/2009
Ambiente fora do ar para manutenção	Administrador	26/11/2008

1 - 5 de 5 resultados

Marcar Todas

Excluir

Outros (Extração de Informações Estratégicas)

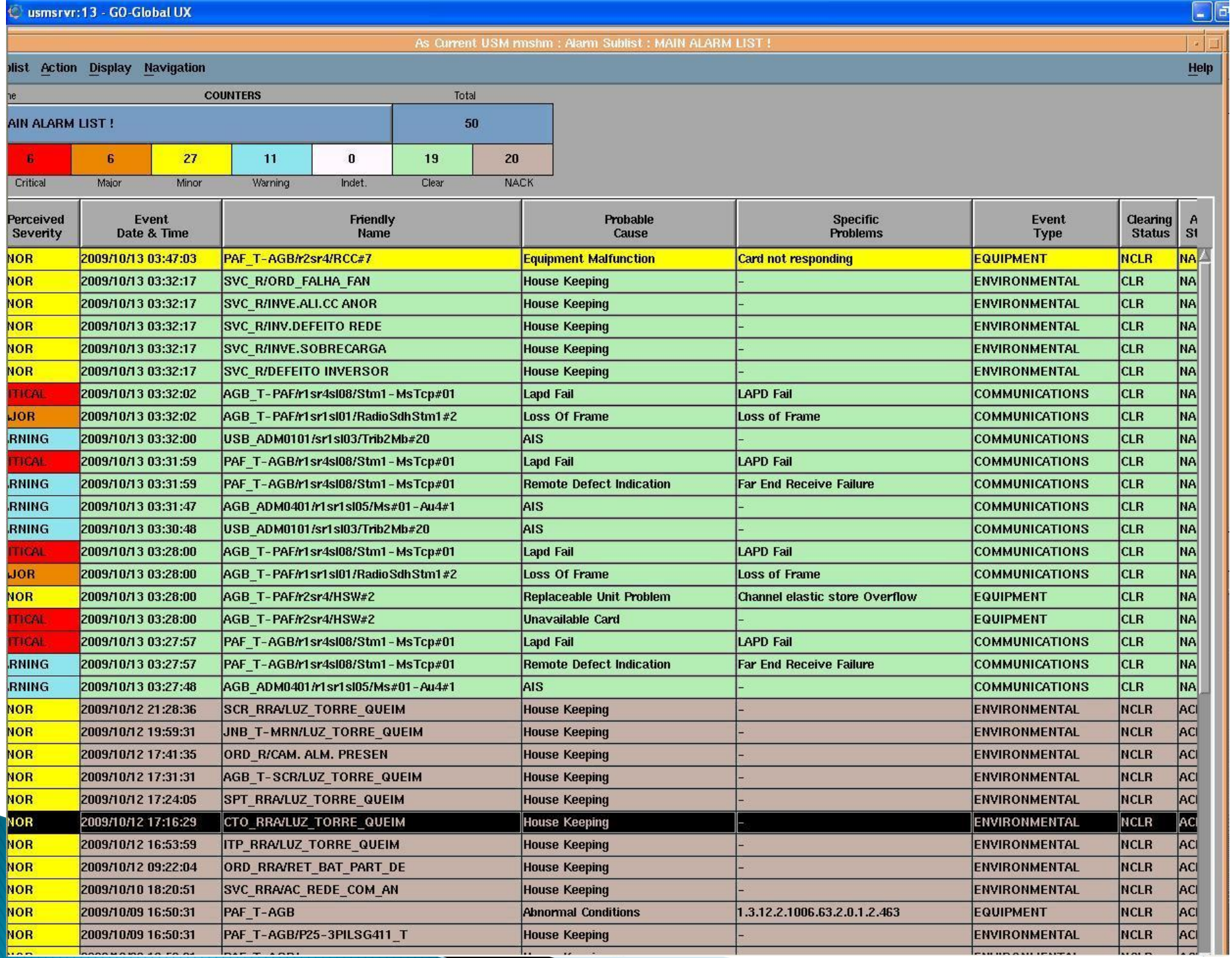
- Business Intelligence
  - Análise de Mercado
  - Melhoria de Processos
  - Gerenciamento Eletrônico de Documentos

- Análises de Arquivos de LOG
  - Logs de Erro
  - Logs de Acesso

- Extração de Informação na WEB
  - Filtragem de Fóruns
    - Controle do Conteúdo
    - Assunto dos Diálogos
    - Empresa de São Paulo com mais de 20 anos de mercado. Oferece soluções para e-learning.

Sistema de Extração de Informações de Gerências de Telecomunicações da Chesf

Luiz Carlos d´Oleron
lcadb@cin.ufpe.br luizdb@chesf.gov.br


usmsrvr:13 - GO-Global UX
As Current USM mnshm : Alarm Sublist : MAIN ALARM LIST !
blist Action Display Navigation
Help
ne
COUNTERS
Total
AIN ALARM LIST !
50
6
6
27
11
0
19
20
Critical
Major
Minor
Warning
Indet.
Clear
NACK
Perceived Severity
Event Date & Time
Friendly Name
Probable Cause
Specific Problems
Event Type
Clearing Status
A St
NOR 2009/10/13 03:47:03 PAF_T-AGB/r2sr4/RCC#7 Equipment Malfunction Card not responding EQUIPMENT NCLR NA
NOR 2009/10/13 03:32:17 SVC_R/ORD_FALHA_FAN House Keeping - ENVIRONMENTAL CLR NA
NOR 2009/10/13 03:32:17 SVC_R/INVE.ALI.CC ANOR House Keeping - ENVIRONMENTAL CLR NA
NOR 2009/10/13 03:32:17 SVC_R/INV.DEFEITO REDE House Keeping - ENVIRONMENTAL CLR NA
NOR 2009/10/13 03:32:17 SVC_R/INVE.SOBRECARGA House Keeping - ENVIRONMENTAL CLR NA
NOR 2009/10/13 03:32:17 SVC_R/DEFEITO INVERSOR House Keeping - ENVIRONMENTAL CLR NA
TICAL 2009/10/13 03:32:02 AGB_T-PAF/r1sr4sl08/Stm1-MsTcp#01 Lapd Fail LAPD Fail COMMUNICATIONS CLR NA
JOR 2009/10/13 03:32:02 AGB_T-PAF/r1sr1sl01/RadioSdhStm1#2 Loss Of Frame Loss of Frame COMMUNICATIONS CLR NA
RNING 2009/10/13 03:32:00 USB_ADM0101/sr1sl03/Trib2Mb#20 AIS - COMMUNICATIONS CLR NA
TICAL 2009/10/13 03:31:59 PAF_T-AGB/r1sr4sl08/Stm1-MsTcp#01 Lapd Fail LAPD Fail COMMUNICATIONS CLR NA
RNING 2009/10/13 03:31:59 PAF_T-AGB/r1sr4sl08/Stm1-MsTcp#01 Remote Defect Indication Far End Receive Failure COMMUNICATIONS CLR NA
RNING 2009/10/13 03:31:47 AGB_ADM0401/r1sr1sl05/Ms#01-Au4#1 AIS - COMMUNICATIONS CLR NA
RNING 2009/10/13 03:30:48 USB_ADM0101/sr1sl03/Trib2Mb#20 AIS - COMMUNICATIONS CLR NA
TICAL 2009/10/13 03:28:00 AGB_T-PAF/r1sr4sl08/Stm1-MsTcp#01 Lapd Fail LAPD Fail COMMUNICATIONS CLR NA
JOR 2009/10/13 03:28:00 AGB_T-PAF/r1sr1sl01/RadioSdhStm1#2 Loss Of Frame Loss of Frame COMMUNICATIONS CLR NA
NOR 2009/10/13 03:28:00 AGB_T-PAF/r2sr4/HSW#2 Replaceable Unit Problem Channel elastic store Overflow EQUIPMENT CLR NA
TICAL 2009/10/13 03:28:00 AGB_T-PAF/r2sr4/HSW#2 Unavailable Card - EQUIPMENT CLR NA
TICAL 2009/10/13 03:27:57 PAF_T-AGB/r1sr4sl08/Stm1-MsTcp#01 Lapd Fail LAPD Fail COMMUNICATIONS CLR NA
RNING 2009/10/13 03:27:57 PAF_T-AGB/r1sr4sl08/Stm1-MsTcp#01 Remote Defect Indication Far End Receive Failure COMMUNICATIONS CLR NA
RNING 2009/10/13 03:27:48 AGB_ADM0401/r1sr1sl05/Ms#01-Au4#1 AIS - COMMUNICATIONS CLR NA
NOR 2009/10/12 21:28:36 SCR_RRA/LUZ_TORRE_QUEIM House Keeping - ENVIRONMENTAL NCLR AC
NOR 2009/10/12 19:59:31 JNB_T-MRN/LUZ_TORRE_QUEIM House Keeping - ENVIRONMENTAL NCLR AC
NOR 2009/10/12 17:41:35 ORD_R/CAM. ALM. PRESEN House Keeping - ENVIRONMENTAL NCLR AC
NOR 2009/10/12 17:31:31 AGB_T-SCR/LUZ_TORRE_QUEIM House Keeping - ENVIRONMENTAL NCLR AC
NOR 2009/10/12 17:24:05 SPT_RRA/LUZ_TORRE_QUEIM House Keeping - ENVIRONMENTAL NCLR AC
NOR 2009/10/12 17:16:29 CTO_RRA/LUZ_TORRE_QUEIM House Keeping - ENVIRONMENTAL NCLR AC
NOR 2009/10/12 16:53:59 ITP_RRA/LUZ_TORRE_QUEIM House Keeping - ENVIRONMENTAL NCLR AC
NOR 2009/10/12 09:22:04 ORD_RRA/RET_BAT_PART_DE House Keeping - ENVIRONMENTAL NCLR AC
NOR 2009/10/10 18:20:51 SVC_RRA/AC_REDE_COM_AN House Keeping - ENVIRONMENTAL NCLR AC
NOR 2009/10/09 16:50:31 PAF_T-AGB Abnormal Conditions 1.3.12.2.1006.63.2.0.1.2.463 EQUIPMENT NCLR AC
NOR 2009/10/09 16:50:31 PAF_T-AGB/P25-3PILSG411_T House Keeping - ENVIRONMENTAL NCLR AC

Motivação

- Equipes de telecomunicação de tempo real utilizam gerências para supervisionar redes fibra ótica, comutação, OPLAT, teleproteção, WAN e outros
- Cada gerência gera centenas de eventos (alarmes e informações) a cada minuto
- Informações aparecem em campos de texto de tamanho curto, com difícil compreensão
- Eventos de diferentes naturezas, equipamentos e localidades

Solução

- Sistema de extração de informações
- Capaz de oferecer uma visão condensada dos eventos
- Descartando informações desnecessárias
- Apresentando informações relevantes
- Totalizando eventos
- Enriquecendo dados através de outras fontes de dados

Fonte de Dados

	2009/10/13 03:32:17	SVC_R/ORD_FALHA_FAN	House Keeping	-
	2009/10/13 03:32:17	SVC_R/INVE.ALI.CC ANOR	House Keeping	-
	2009/10/13 03:32:17	SVC_R/INV.DEFEITO REDE	House Keeping	-
	2009/10/13 03:32:17	SVC_R/INVE.SOBRECARGA	House Keeping	-
	2009/10/13 03:32:17	SVC_R/DEFEITO INVERSOR	House Keeping	-
	2009/10/13 03:32:02	AGB_T-PAF/r1sr4sl08/Stm1-MsTcp#01	Lapd Fail	LAPD Fail
	2009/10/13 03:32:02	AGB_T-PAF/r1sr1sl01/RadioSdhStm1#2	Loss Of Frame	Loss of Frame
	2009/10/13 03:32:00	USB_ADM0101/sr1sl03/Trib2Mb#20	AIS	-
	2009/10/13 03:31:59	PAF_T-AGB/r1sr4sl08/Stm1-MsTcp#01	Lapd Fail	LAPD Fail
	2009/10/13 03:31:59	PAF_T-AGB/r1sr4sl08/Stm1-MsTcp#01	Remote Defect Indication	Far End Receive Failure
	2009/10/13 03:31:47	AGB_ADM0401/r1sr1sl05/Ms#01-Au4#1	AIS	-
	2009/10/13 03:30:48	USB_ADM0101/sr1sl03/Trib2Mb#20	AIS	-
	2009/10/13 03:28:00	AGB_T-PAF/r1sr4sl08/Stm1-MsTcp#01	Lapd Fail	LAPD Fail
	2009/10/13 03:28:00	AGB_T-PAF/r1sr1sl01/RadioSdhStm1#2	Loss Of Frame	Loss of Frame

Extração

2009/10/13 03:32:02	AGB_T-PAF/r1sr1sl01/RadioSdhStm1#2	Loss Of Frame	COMMUNICATIONS
2009/10/13 03:32:00	USB_ADM0101/sr1sl03/Trib2Mb#20	AIS	COMMUNICATIONS

AGB_T-PAF/r1sr1sl01/RadioSdhStm1#2

AGB_T-PAF (equipamento)		
Rádio	Origem: Repetidora Água Branca (AGB)	Destino: SE Paulo Afonso (PAF)

R1sr1sl01 (localização)		
Bastidor 1 (R1)	Sub Bastidor 1 (sr1)	Slot 01 (sl01)

RadioSdhStm1#2 (canalização)		
Protocolo SDH (Sdh)	Pacote de 155,52 Mbit/s (Stm1)	Porta 2 (#2)

Extração

2009/10/13 03:32:02	AGB_T-PAF/r1sr1sl01/RadioSdhStm1#2	Loss Of Frame	COMMUNICATIONS
2009/10/13 03:32:00	USB_ADM0101/sr1sl03/Trib2Mb#20	AIS	COMMUNICATIONS

USB_ADM0101/sr1sl03/Trib2Mb#20

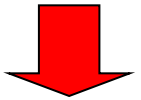
USB_ADM0101 (equipamento)	
Add-drop Multiplex	Local: Usina Sobradinho

sr1sl03 (localização)	
Sub Bastidor 1 (sr1)	Slot 03 (sl03)

Trib2Mb#2	
Tirbutário de 2 Mbits/s	Porta 20 (#20)

Extração e enriquecimento

Friendly Name	Location	Type	Rack	Subrack	Slot	Port	Board	ToDirection	
ITP_R/ PAF_FALHA_FAN	ITP_R	FAN	N/A	N/A	N/A	N/A	N/A	N/A	N
AGB_T-PAF/r1sr1sl01/RadioSdhStm1#2	AGB	Radio	1	1	1	2	N/A	PAF	N
USB_ADM0101/sr1sl03/Trib2Mb#2	USB	Mux	N/A	1	3	2	N/A	N/A	N
PAF_T-AGB/r1sr4sl04/Stm1-MsTcp#01	PAF	Board	1	4	4	1	MSTCP	AGB	N
SPT_R/INV.DEFEITO REDE	SPT_R	Inversor	N/A	N/A	N/A	N/A	N/A	N/A	D
CTO_RRA/USCC_FLUTU_AN	CTO_R	USCC	N/A	N/A	N/A	N/A	N/A	N/A	F
ITP_R/r1sr2sl04/RadioSdhStm1#1W	ITP_R	Radio	1	1	4	1	N/A	W	N



enriquecimento

Location	Type	Frequence	Problem Cause	ToDirection	Rack	Subrack	S
Repetidora Itaparica	Cooler Painel Rádio	117	Falha cooler	N/A	N/A	N/A	N
Repetidora Itaparica	Rádio Digital SDH	273	Falha demodulador	SE Paulo Afonso	1	1	
Repetidora Água Branca	Rádio Digital SDH	389	Perda de quadro	SE Paulo Afonso	1	1	
Usina Sobradinho	Multiplex Digital	566	Sinal de defeito	N/A	N/A	1	
SE Paulo Afonso	Placa	45	Falha protocolo supervisão	Repetidora Água Branca	1	4	
Repetidora Serra da Prata	Inversor	1107	Defeito inversor	N/A	N/A	N/A	N
Repetidora Cabeça de Touro	Unidade supervisão de corrente contínua	8	Flutuação de tensão anormal	N/A	N/A	N/A	N

Geração de Relatório

Companhia Hidro Elétrica do São Francisco - CHESF
Departamento de Telecomunicações - DTL

CSTL – Centro de Supervisão de telecomunicações

Relatório Operacional

Período:
Das [redacted] de 2009 às [redacted] de 2009

Localidades envolvidas:

Repetidora Água Branca, Repetidora Serra da Barriga, Xingó Margem Esquerda, Repetidora Cabeça de Touro, SE Paulo Afonso, Usina Sobradinho, SE Recife II, Repetidora Moreno, Usina Xingó.

Eventos:

1. Em **Repetidora Água Branca**:

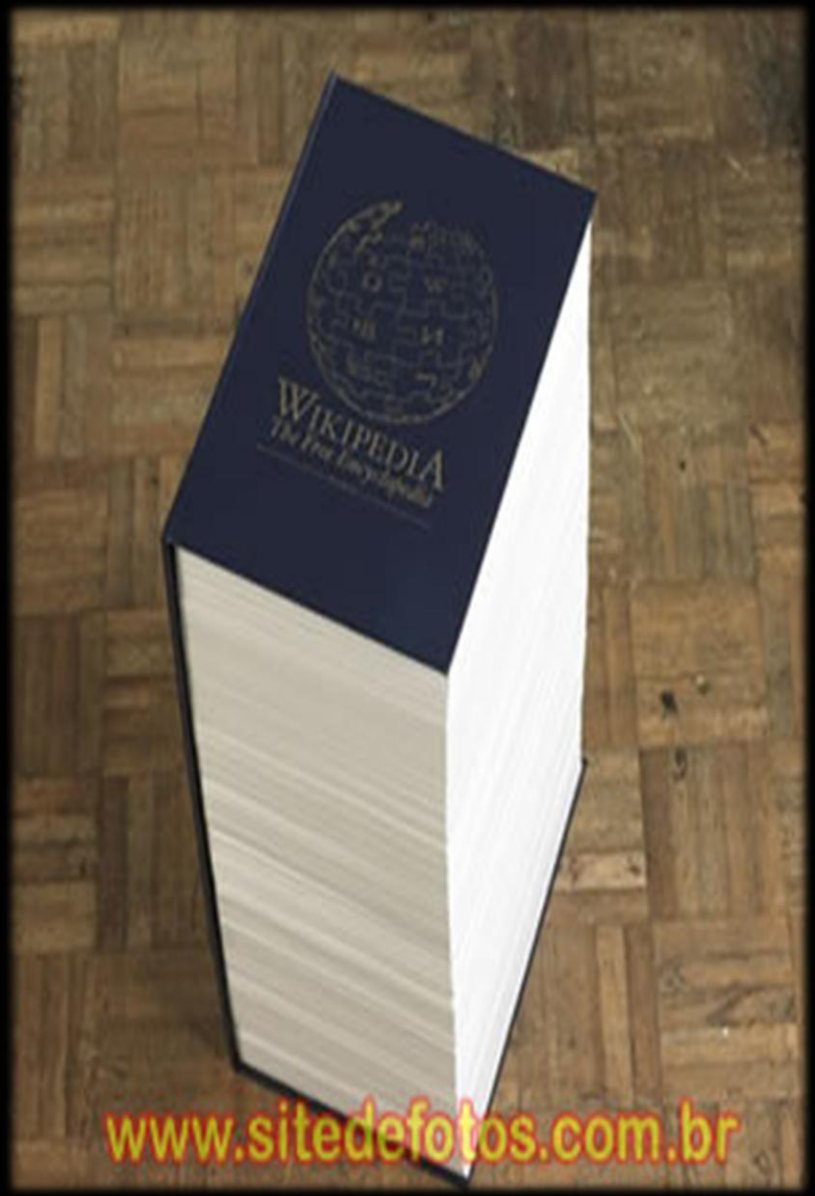
Rádio Digital SDH apresentando MODERADA ocorrência de *Perda de quadro* na porta 2, bastidor 1, sub-bastidor 1, slot 1, na direção de SE Paulo Afonso. Partida GGE no dia [redacted] e parada GGE dia [redacted]. Perda de alimentação comercial dia [redacted] e normalização dia [redacted].

2. Em **Repetidora Itaparica**:

Rádio Digital SDH apresentando MODERADA ocorrência de *Falha demodulador* na porta 1, bastidor 1, sub-bastidor 1, slot 4, na direção de SE Paulo Afonso, Cooler Painel Rádio apresentando MODERADA ocorrência de Falha cooler e Placa MSTCP apresentando PREOCUPANTE ocorrência de

Futuro

Futuro



- Síntese de informação

Respostas as nossas perguntas

Will extraction tools automatically assemble our own personal 7 o’clock television news from all sources available?

Sentiment analysis (Real)

Processos + ferramentas + domínio

Referências Bibliográficas

[1] Cabral, Davi Medeiros. **Um Framework para Extração de Informações: Uma Abordagem Baseada em XML.** Dissertação de Mestrado – UFPE (Cin), Recife, 2005.

[2] ÁLVARES, Alberto Cáceres. **Extração de informação de artigos científicos: uma abordagem baseada em indução de regras de etiquetagem.** Dissertação de Mestrado – USP (ICMC), São Carlos, 2007.

[3] SILVA, Eduardo F.A; BARROS, Flávia A; PRODÊNCIO, Ricardo B. C. **Uma Abordagem de Aprendizagem Híbrida para Extração de Informação em Textos Semi-Estruturados.**

[4] SILVA, Eduardo Fraga do Amaral. **Sistema de extração de informação em referências bibliográficas baseadas em aprendizagem de máquina.** Dissertação de Mestrado – UFPE (CIn), Recife, 2004.

Referências Bibliográficas

[5] A.M.I.G.O.S, http://www.oro-aro.com/

[6] E.Life, http://www.elife.com.br/

[7]Open Amplify, http://www.openamplify.com/

[8] Personas, http://personas.media.mit.edu/

[9] Scout Labs, http://www.scoutlabs.com/

[10] Newssit, http://www.newssift.com/

[11] Marie-Francine Moens. Information Extraction: Algorithms and Prospects in a Retrieval Context. Springer, 2006.

Extração de Informação

Luiz Carlos d´Oleron – lcadb@cin.ufpe.br
Vítor Braga – vtb@cin.ufpe.br