FACULDADE·DE·CIÊNCIAS UNIVERSIDADE·DE·LISBOA

REMBRANDT

Reconhecimento de Entidades Mencionadas Baseado em Relações e ANálise Detalhada do Texto

Nuno Cardoso

Universidade de Lisboa, Faculdade de Ciências, Laboratório LASIGE

ncardoso@xldb.di.fc.ul.pt

Introdução

- **REMBRANDT** - sistema de reconhecimento de entidades mencionadas e de detecção de relações entre entidades em textos portugueses.

"Rembrandt Harmenszoon van Rijn nasceu em 15 de julho de 1606 (tradicionalmente) mas provavelmente em 1607 em Leiden, Países Baixos. (...) Em 1634, Rembrandt casou com a prima de Hendricks, Saskia van Uylenburgh. Em 1639, Rembrandt e Saskia mudaram-se para uma casa maior em Jodenbreestraat."

Motivação

Contexto: reformulação de consultas para motores de busca com âmbitos geográficos.

- Gerar assinaturas geográficas completas dos documentos, para determinar âmbitos geográficos de cada um.
- Reformular linhas de consultas dos utilizadores usando as anotações do REMBRANDT.
- Usar o conhecimento humano sobre os tópicos abordados na pesquisa, para realizar uma reformulação inteligente das consultas.

Objectivos

1. Reconhecer todas as entidades mencionadas no texto e extrair a sua "geograficidade" (Santos e Chaves, 2006).

2. Processar grandes quantidades de texto de uma forma eficaz e eficiente.

3. Explorar a Wikipédia como fonte de conhecimento em bruto, aplicado na desambiguação e classificação de EM.

Resumo

- Principais características
- Receita do REMBRANDT
- Funcionamento da SASKIA
- Exemplos de utilização
- Detector de relações
- Conclusões e trabalho futuro

Principais características

- Sistema dependente da língua, inspirado no **PALAVRAS_NER** (Bick, 2006).
- Desenvolvido para português, adaptado para inglês.
- Procura evidências internas e externas de EM com um sistema de regras e cláusulas.
- Classificação de EM: módulo SASKIA + regras com sistema de resolução de conflitos (um "tribunal" de EM).

Receita do REMBRANDT (1/2)

1. Reconhecimento de EM numéricas (números, expressões temporais e valores).
2. Geração de candidatos de EM (sequências de termos com maiúscula + “d[aeo]s?|e”).
3. Classificar cada EM candidata com a Saskia + regras para todas as categorias.
4. Segunda ronda de regras externas, aproveitando as classificações existentes.
   (não foi a tempo para o HAREM...)

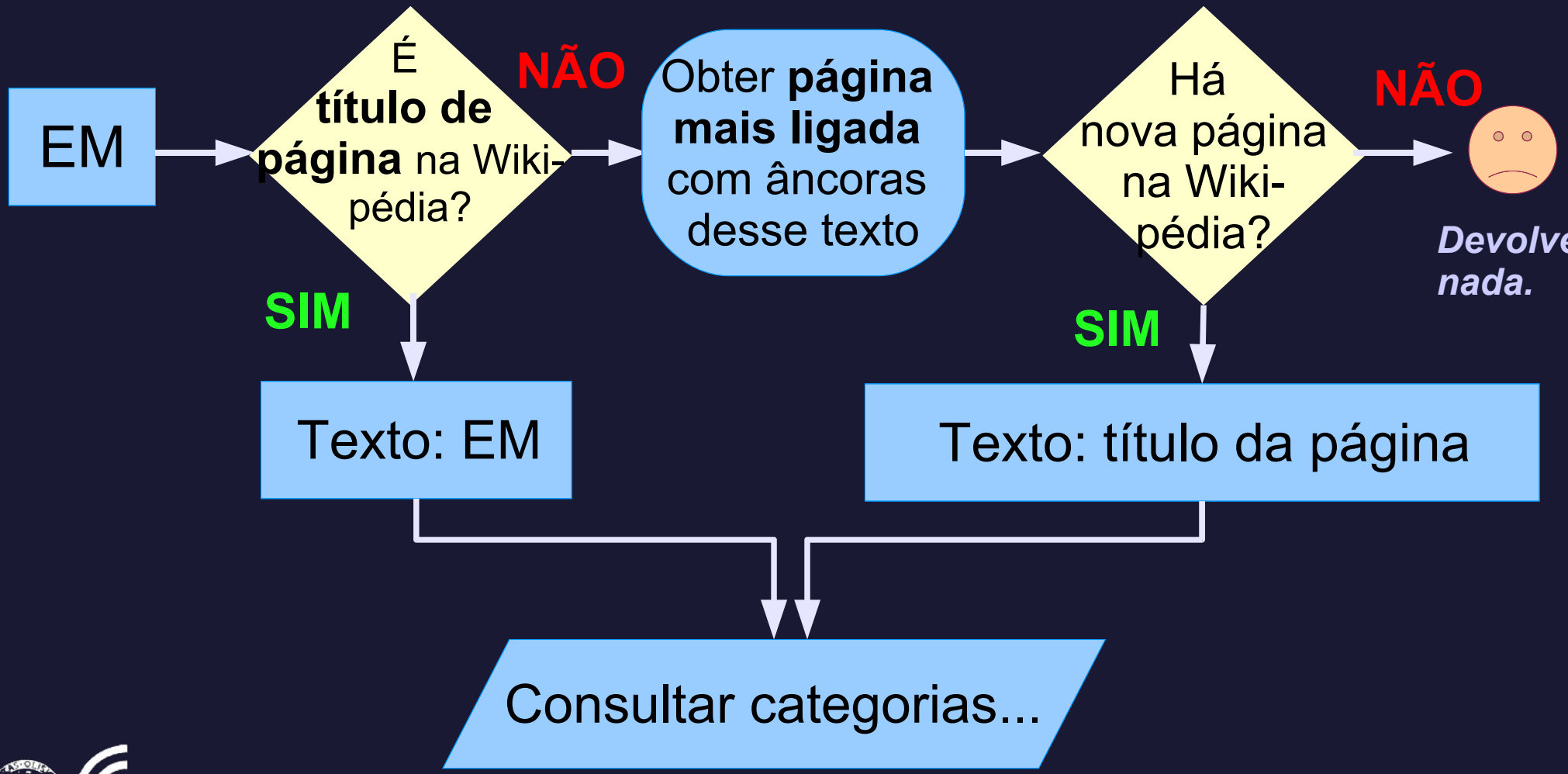
   (5.) Previsto: etapa “contextualizadora” de EM.

Receita do REMBRANDT (2/2)

6. Detecção de relações entre EM.
   - Regras baseadas na classificação das EM.
     (Ex: <ORGANIZACAO> de <LOCAL>)
   - Propaga as classificações conhecidas.
     (Ex: cidade de XPTO → XPTO)
   - Usa ligações da Wikipédia para inferir relações.
     (Ex: Armstrong → Neil Armstrong → NASA)
7. "Repescagem" de EM.
   - Procura de nomes de pessoas. EM sem categorias e que comecem frases são eliminados.

Funcionamento da SASKIA* (1/3)

1 – Emparelhar a EM a uma página da Wikipédia.

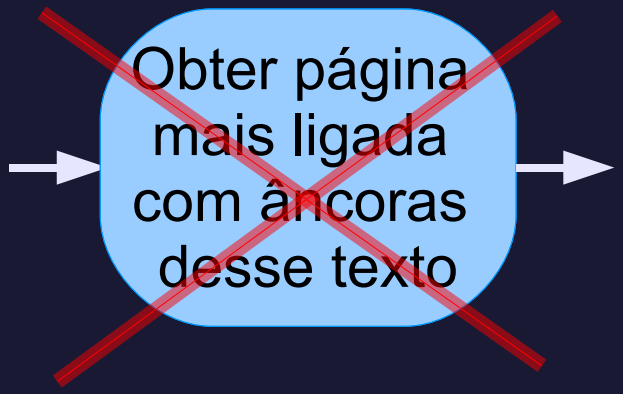


*segundo a versão REMBRANDT 0.7 (usada no HAREM)

Funcionamento da SASKIA** (1/3)

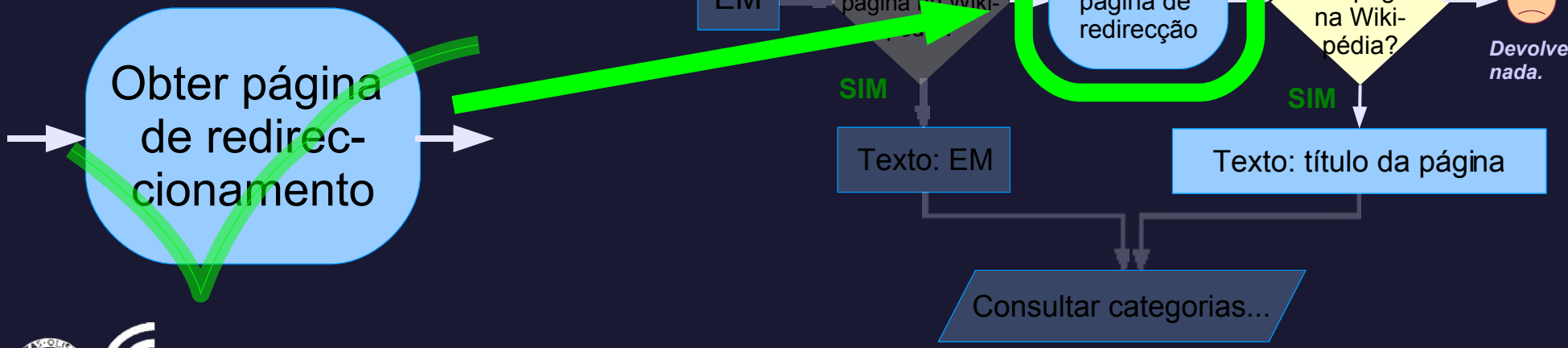
1 – Emparelhar a EM a uma página da Wikipédia.

Versão 0.7:



- Wikipédia em SQL não tem âncoras.
- Wikipédia EN em XML com 28 GB(!)
- Redireccionamentos são comuns e exactos.

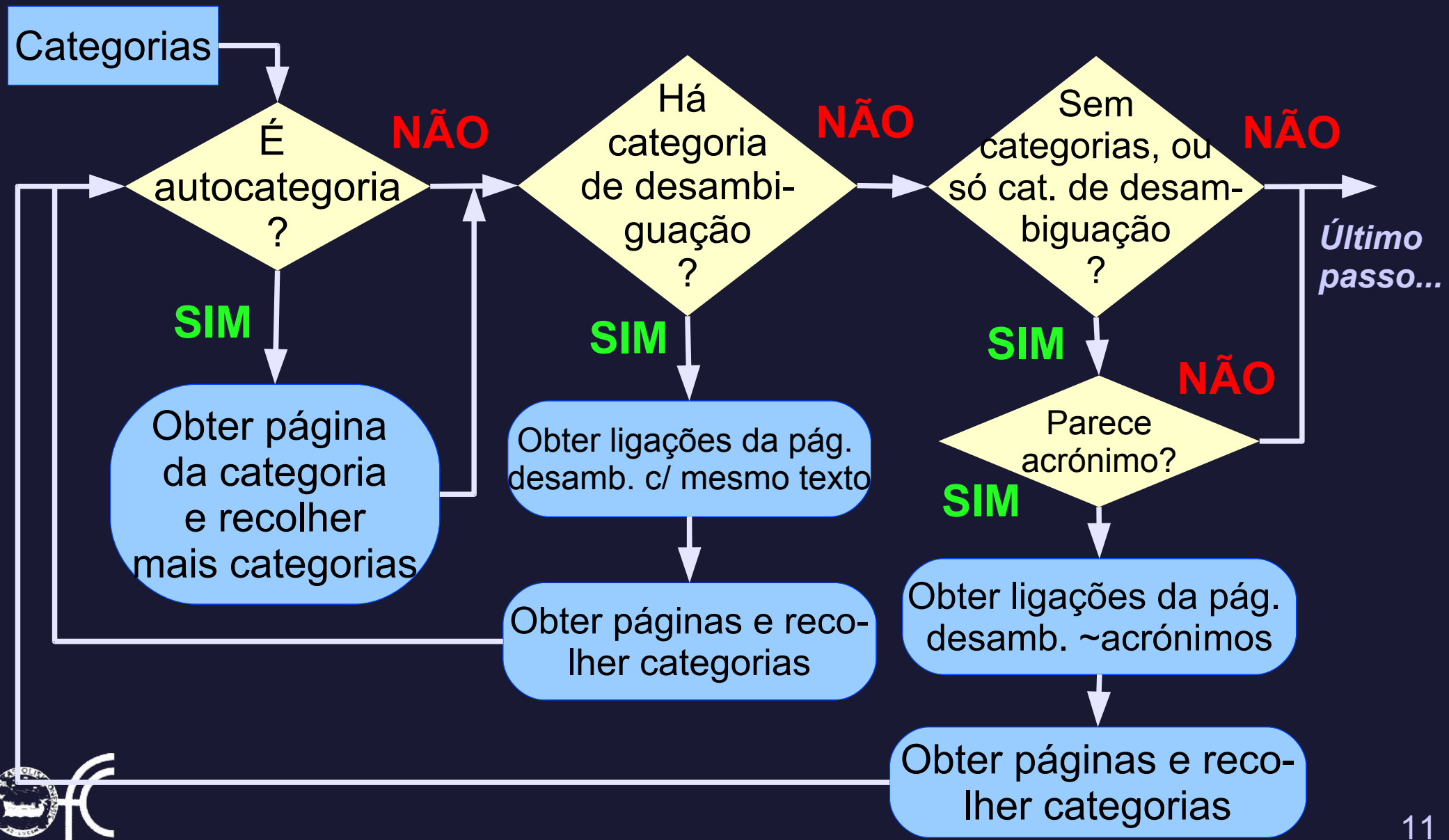
Versão 0.8:



**segundo a versão REMBRANDT 0.8 (actual)

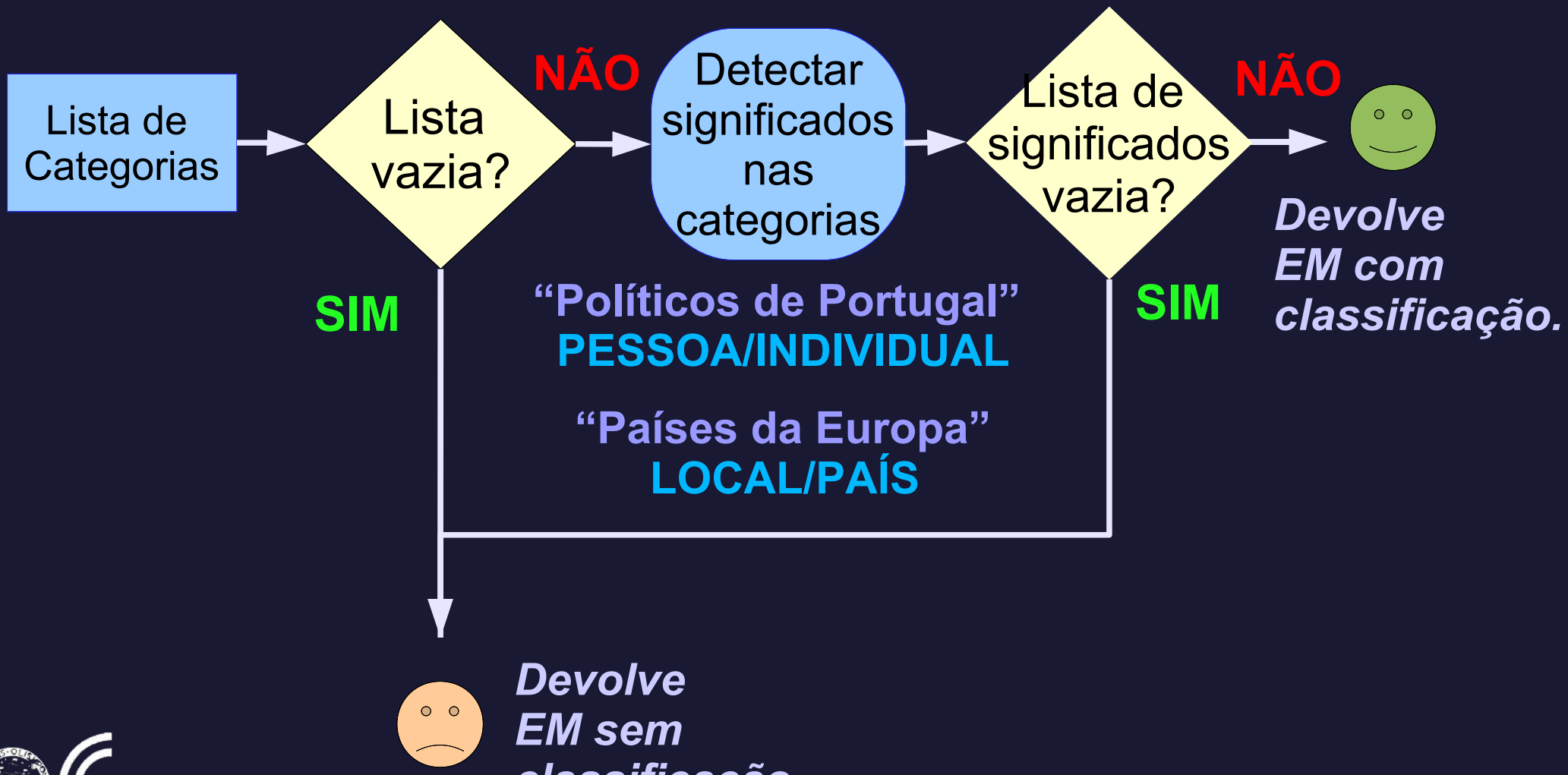
Funcionamento da SASKIA (2/3)

2 – Recolher categorias e analisar ligações.



Funcionamento da SASKIA (3/3)

3 – Corresponder categorias a classificações.



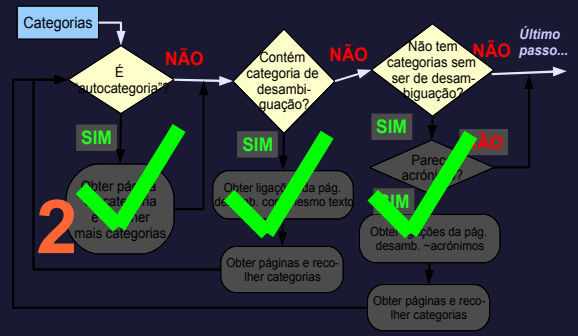
Exemplos (1/6)

1. O caso “ideal”: Simples e directo.

1. “Fernando Ribeiro”?

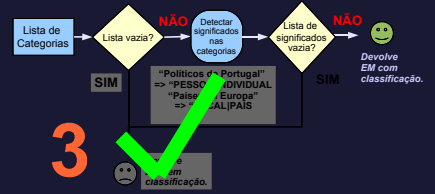
2. Categorias:

- “Esboços de biografia...”
- “Cantores de Portugal”
- “Moonspell”
- “Cantores de heavy metal”

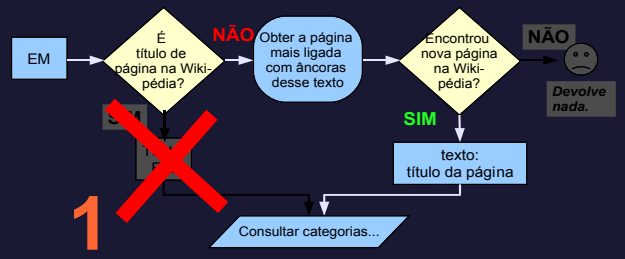


3. Classificação:

“Cantores de Portugal”
→ **PESSOA/INDIVIDUAL**



Exemplos (2/6)



2. O caso "indirecto": (REMBRANDT v0.7)
EUA, Estados Unidos, U.S., U.S.A., ...

EUA

Estados Unidos da América	3325
Billboard Hot 100	34
Selecção de Futebol dos Estados Unidos da América	24
Billboard 200	22
(...)	...

U.S.

Estados Unidos da América	7
Billboard 200	4

Estados Unidos

Estados Unidos da América	6750
Selecção de Futebol dos Estados Unidos da América	31
Selecção Norte-Americana de Futebol Feminino	9
(...)	...

E.U.A.

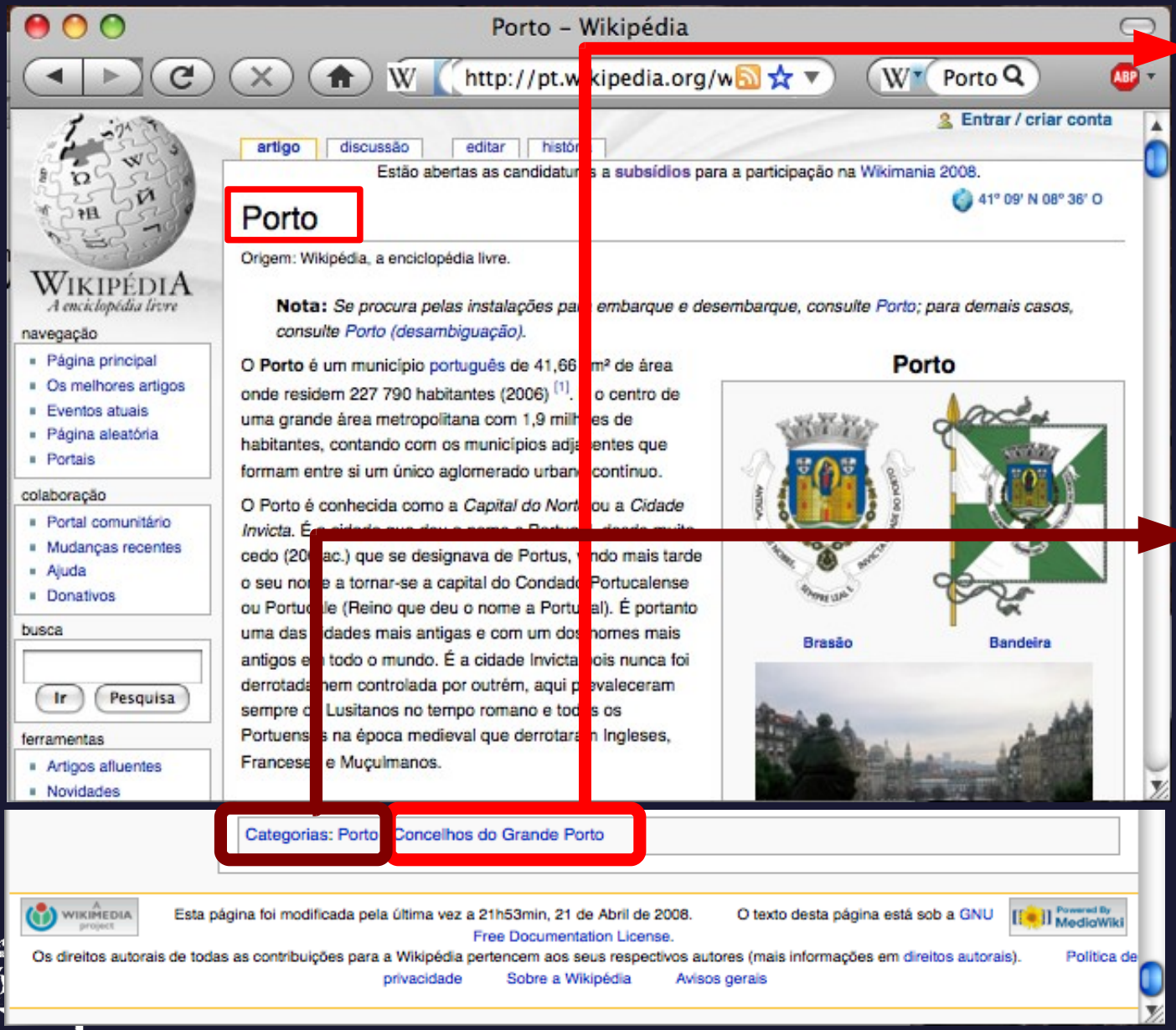
Estados Unidos da América	127
Billboard Hot 100	1

USA

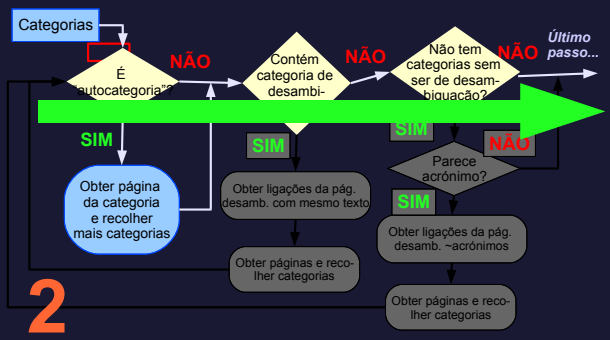
Estados Unidos da América	163
Estados Unidos da América nos Jogos Pan-americanos de 2007	25
Grande Prémio dos Estados Unidos de 2007 (Fórmula 1)	14
(...)	...

Exemplos (3/6)

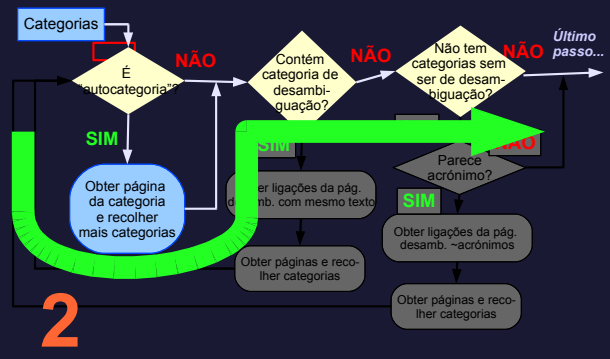
3. O caso de “auto-categorias”. (1/2)



1. “Concelhos do Grande Porto”:



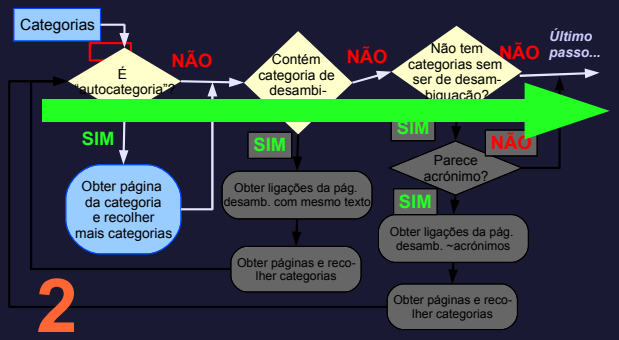
2. “Categoria:Porto”:



Exemplos (4/6)

3. O caso de “auto-categorias”. (2/2)

“Cidades de Portugal”,
“Municípios de Portugal”:



Lista de categorias final:

1. “Conselhos do Grande Porto”
2. “Cidades de Portugal”
3. “Municípios de Portugal”

3 significados:
- Concelho
- Cidade
- Município

Exemplos (5/6)

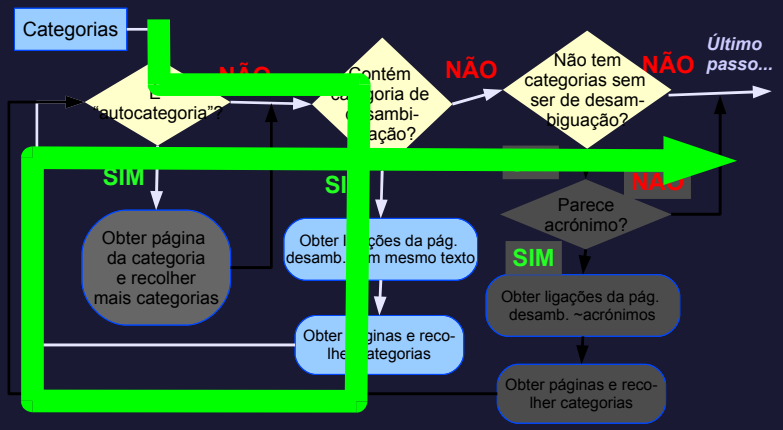
4. O caso da página de desambiguação.

Ligações de saída da página também inclui:

basquete, 1971, 1810, 1900, Argentina, Lua, ...

Solução:

- Usar apenas ligações com "Armstrong" na âncora



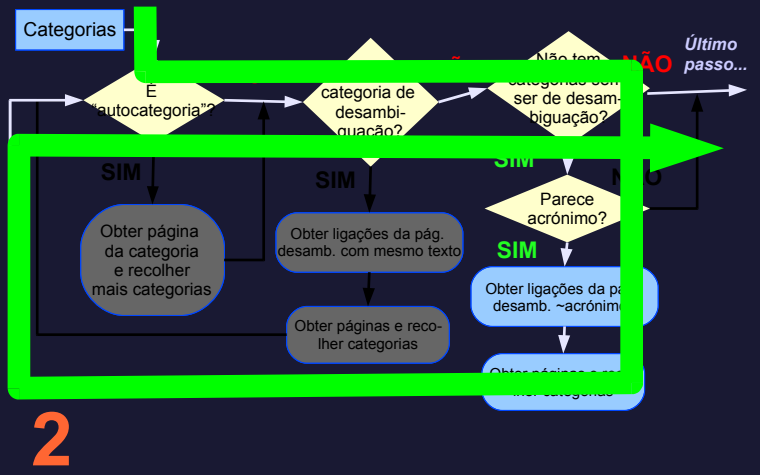
2

Exemplos (6/6)

5. O caso de acrónimos.

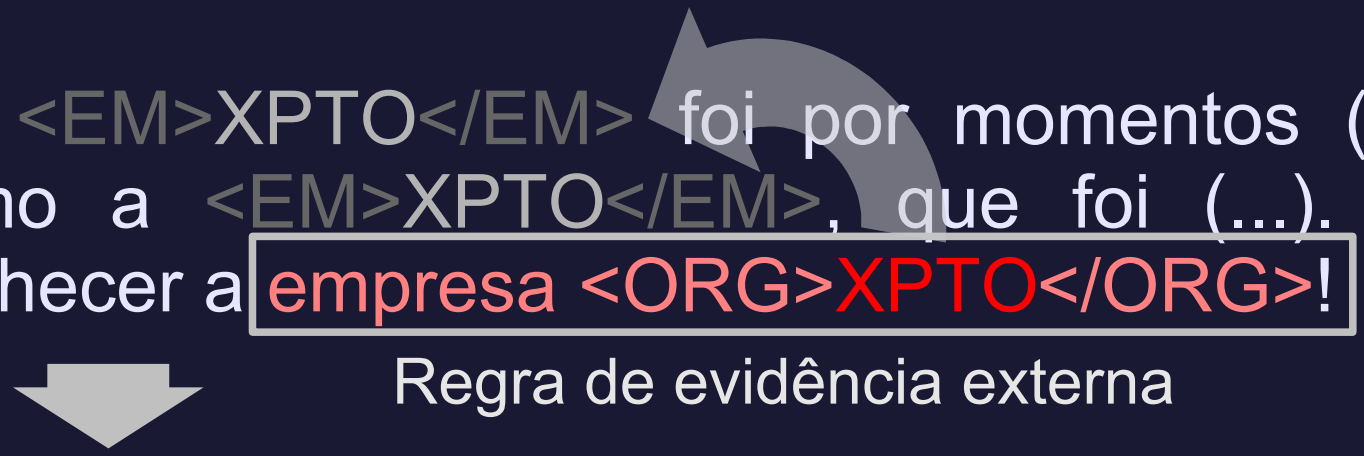
Raramente usada...

- Usar apenas ligações que possuem palavras com as letras que compõem o acrónimo.



Detector de relações (1/3)

- Regras sobre as classificações de EM e sobre semelhança de termos.
  - Ajuda na “repescagem” de algumas EM.



Detector de relações (2/3)

- Uso de ligações na Wikipédia.

" <INDIVIDUAL(?)/DIVISAO(?)>Armstrong</> participou activamente no projecto <PROJECTO>Gemini</>.

Armstrong: ligações para 10 pessoas e 4 cidades.

É provável que as EM vizinhas (i.e., na mesma frase) possam constar no documento de um deles, desambiguando o sentido da EM.

Detector de relações (3/3)

Usado no HAREM, mas ainda não sei ao certo a eficácia desta estratégia.

<INDIVIDUAL(!)/ ~~DIVISAO(X)~~>
Armstrong</> participou activamente no projecto <PROJECTO>Gemini</>.

acompanhamento da fabricação dos motores, foguetes e avião que se destinariam aos projetos Gemini e Apollo. Em março de 1966, ele realizou seu primeiro

Trabalho futuro

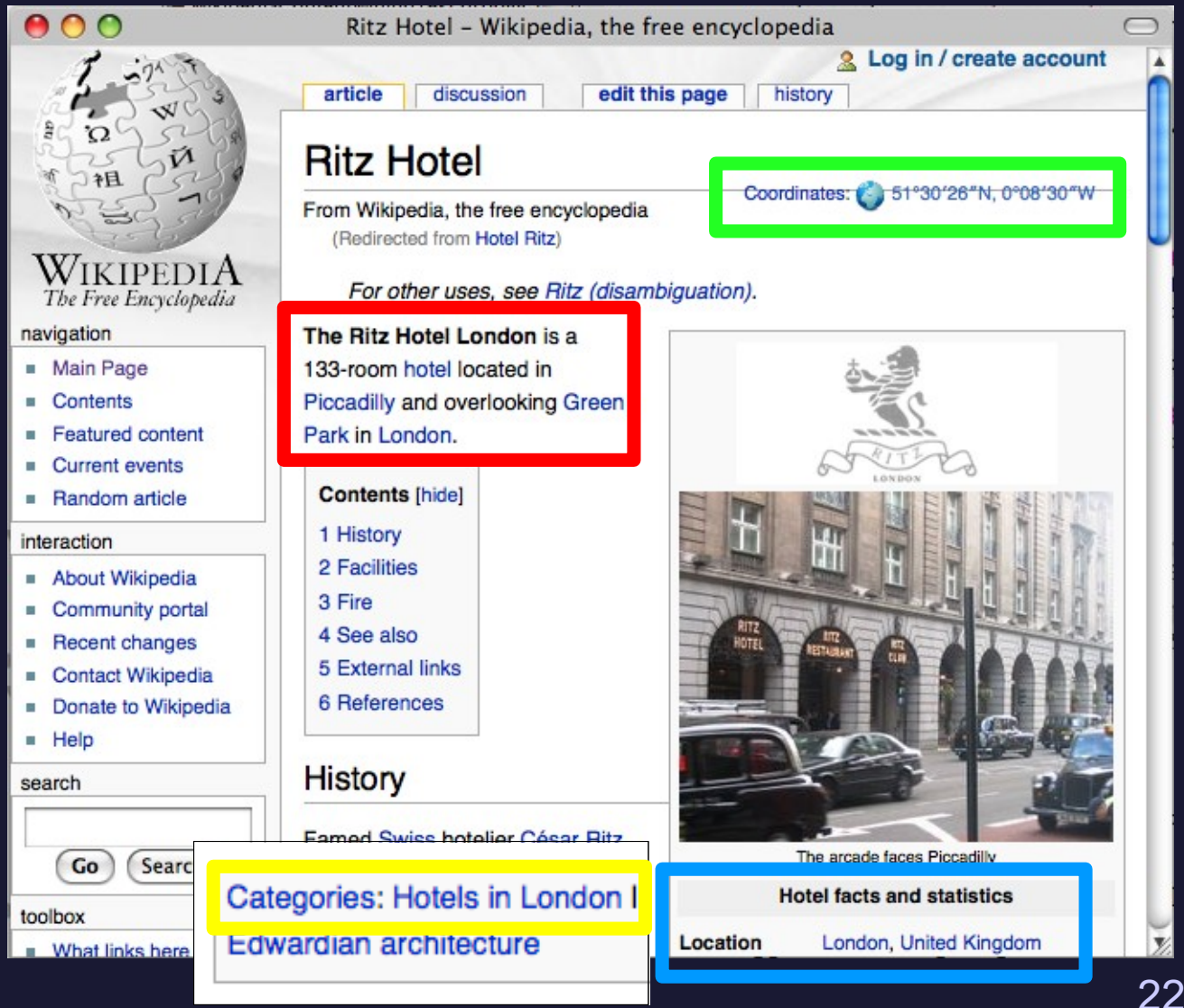
- Explorar melhor a informação na página.

Segundo HAREM:

categorias:
"Hotels in London"

Mas falta explorar:

- coordenadas
- caixa de informação
- parágrafo inicial



Trabalho futuro

- Melhorar o desempenho para Inglês.
- Melhorar o mecanismo de inferência de geograficidade das EM.
- Optimizar, sobretudp o detector de relações (actualmente, é o ponto de constrangimento do sistema).
- Melhorar a SASKIA.
- Documentar o REMBRANDT e promover a sua utilização em várias aplicações.

Fim.

REMBRANDT
Reconhecimento de
Entidades **M**encionadas
Baseado em **R**elações
e **AN**álise **D**etalhada do
Texto

Nuno Cardoso
Faculdade de Ciências, Universidade de Lisboa
Laboratório LASIGE

ncardoso@xldb.di.fc.ul.pt