

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.

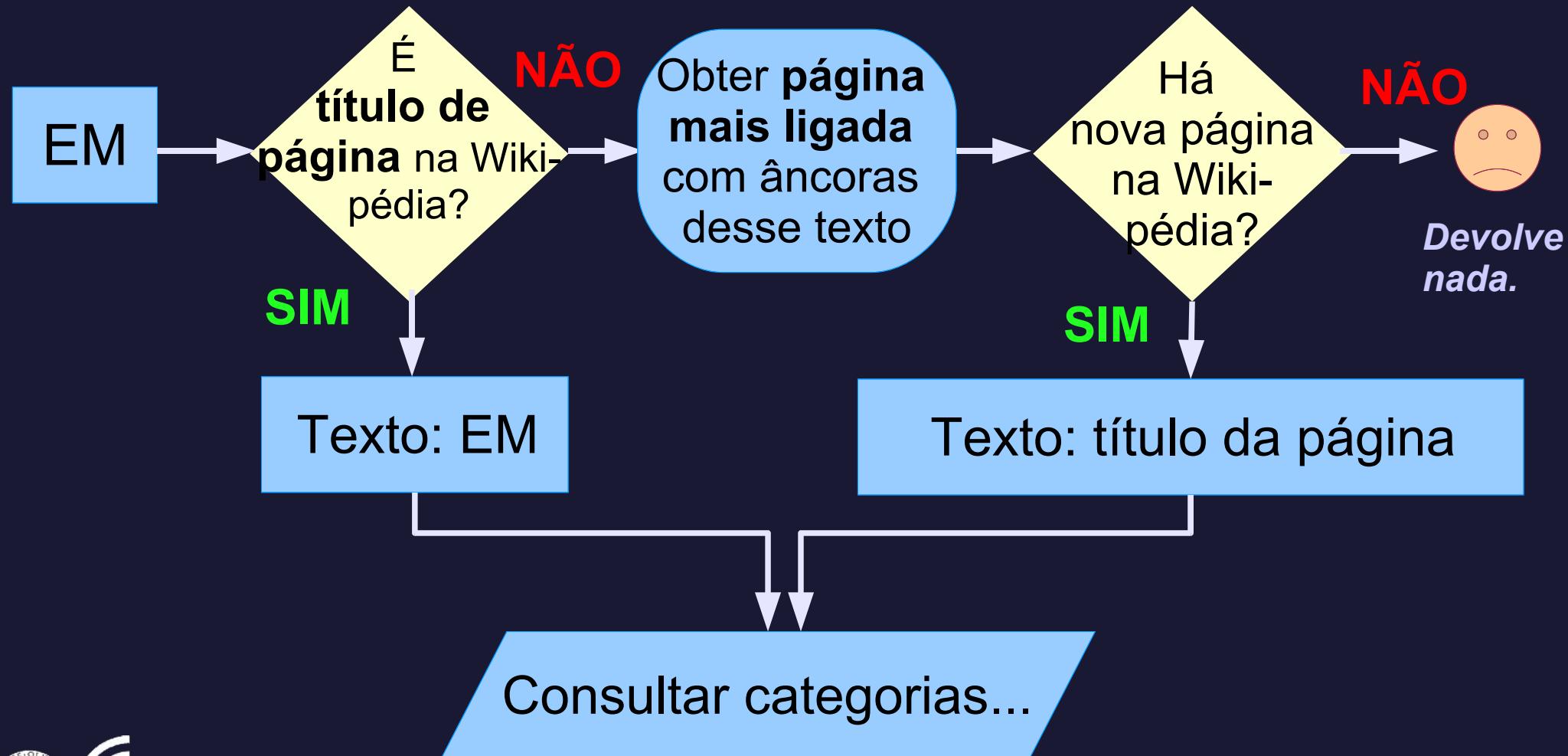
- Regras baseadas na classificação das EM.
(Ex: <ORGANIZACAO> de <LOCAL>)
- Propaga as classificações conhecidas.
(Ex: cidade de XPTO → XPTO)
- Usa ligações da Wikipédia para inferir relações.
(Ex: Armstrong → Neil Armstrong → NASA)

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.

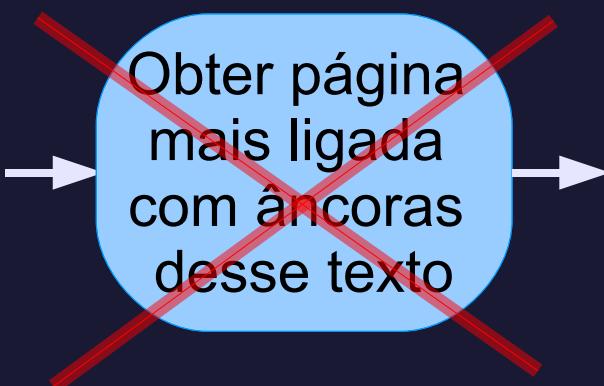


*segundo a versão REMBRANDT 0.7 (usada no HAREM)

Funcionamento da SASKIA** (1/3)

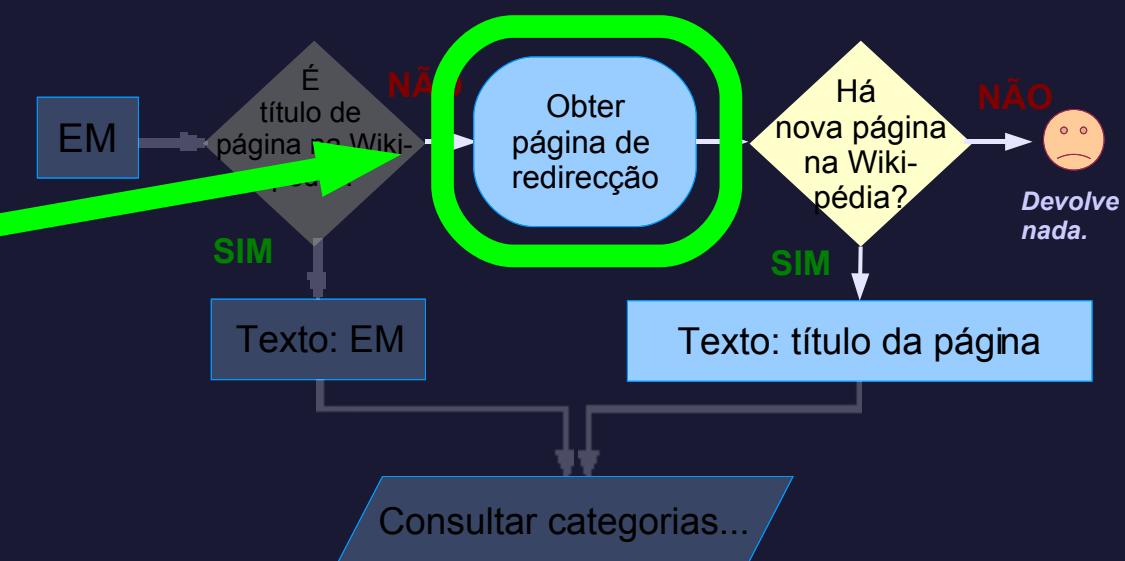
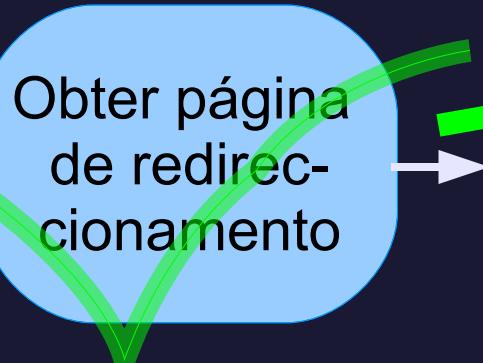
1 – Emparelhar a EM a uma página da Wikipédia.

Versão 0.7:



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

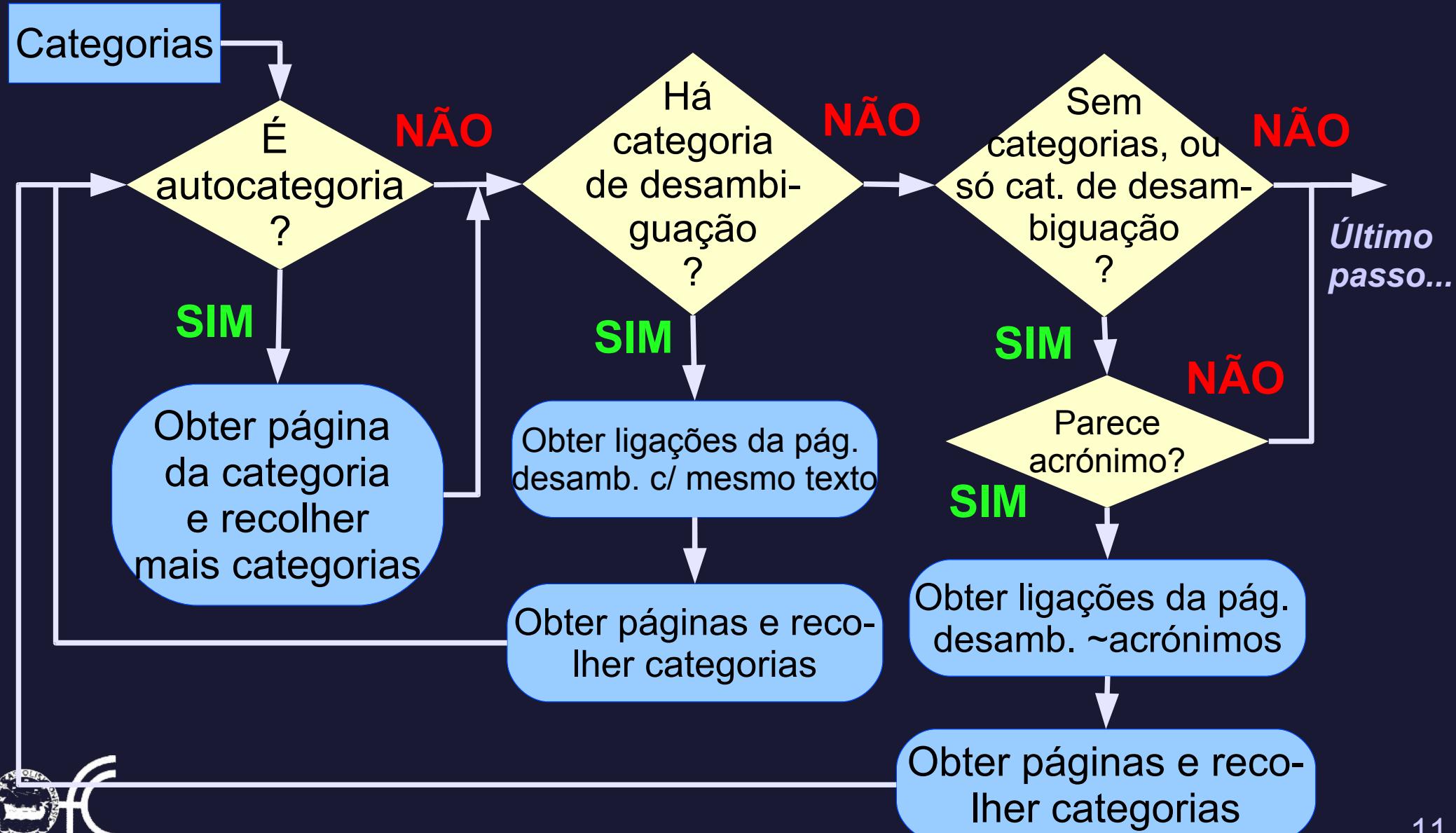
Versão 0.8:



**segundo a versão REMBRANDT 0.8 (actual)

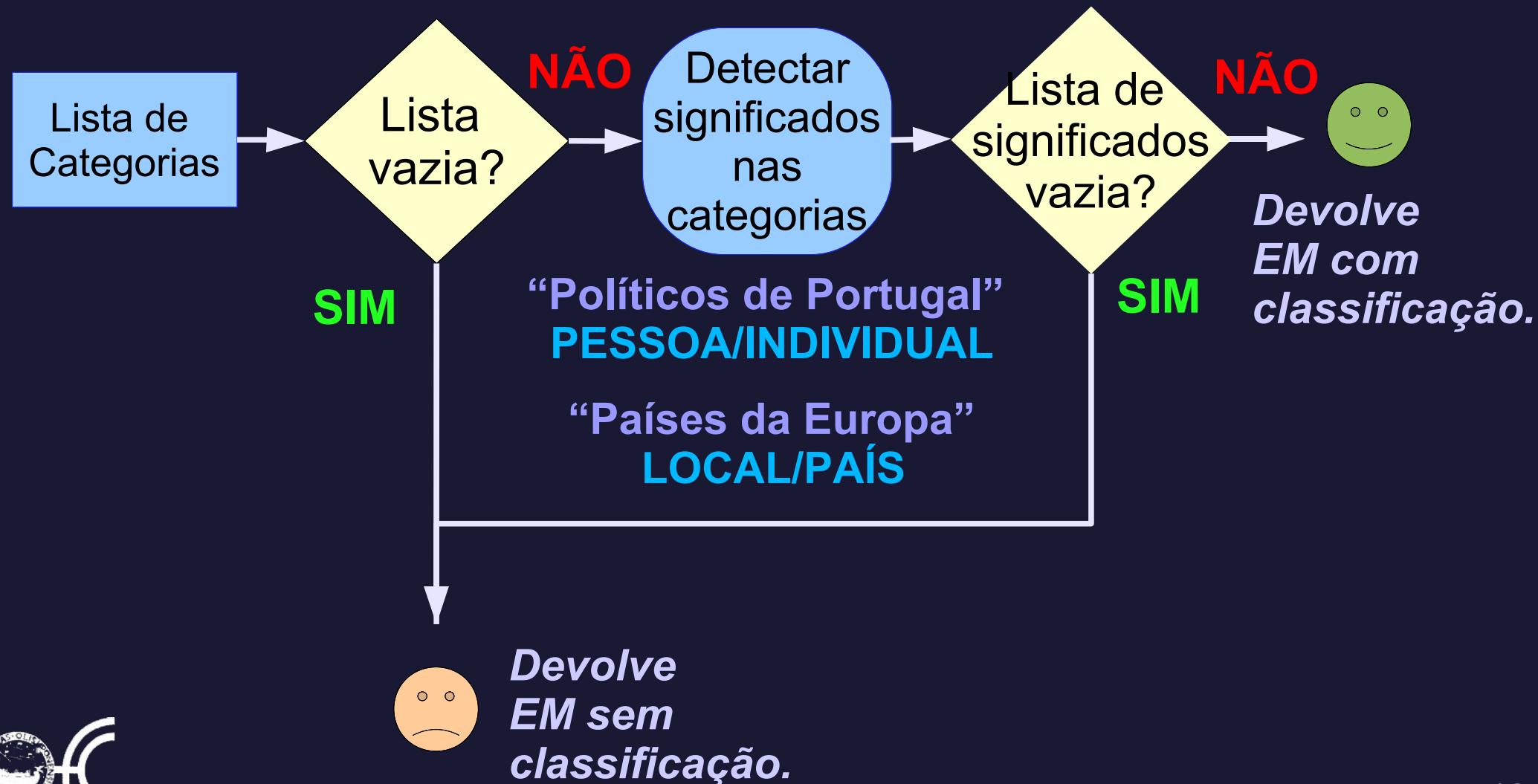
Funcionamento da SASKIA (2/3)

2 – Recolher categorias e analisar ligações.



Funcionamento da SASKIA (3/3)

3 – Correspondar categorias a classificações.

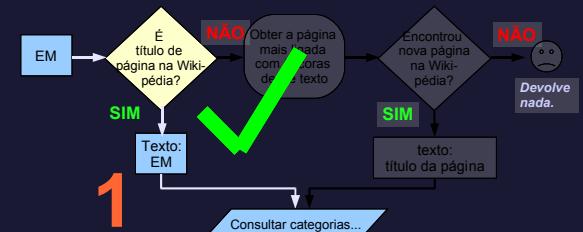


Exemplos (1/6)

1. O caso “ideal”: Simples e directo.

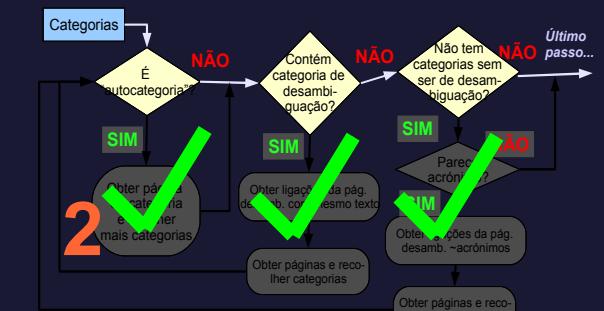
The screenshot shows the Wikipedia article for Fernando Ribeiro. The title 'Fernando Ribeiro' is highlighted with a red box. Below the title, it says 'Origem: Wikipédia, a encyclopédia livre.' The main text describes Fernando Miguel Santos Ribeiro's life and career, mentioning his work with Moonspell and his poetry. A photograph of Fernando Ribeiro singing into a microphone is displayed. The 'Ligações externas' section lists an interview from 'Correio da Manhã'. At the bottom, a note says 'Este artigo é um esboço sobre um músico. Você pode ajudar a Wikipédia expandindo-o.' The 'Categorias' section at the bottom is also highlighted with a red box, showing links to 'Esboços de biografia de músicos', 'Cantores de Portugal', 'Moonspell', and 'Cantores de heavy metal'.

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

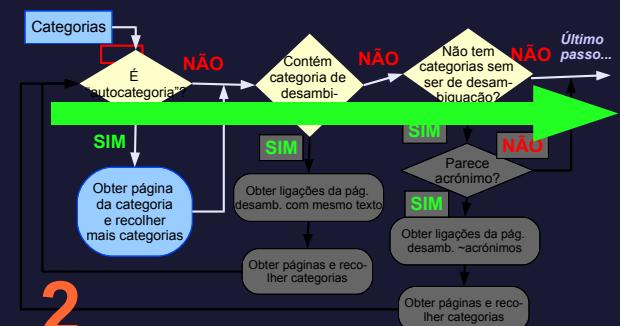
Estados Unidos da América	163
Estados Unidos da América nos Jogos Pan-americanos de 2007	25
Grande Prémio dos Estados Unidos de 2007 (Fórmula 1)	14
(...)	...

Exemplos (3/6)

3. O caso de “auto-categorias”. (1/2)

The screenshot shows the Wikipedia article for "Porto". The title "Porto" is highlighted with a red box. Below the title, the text "Origem: Wikipédia, a encyclopédia livre." is visible. A note at the top of the page reads: "Estão abertas as candidaturas a subsídios para a participação na Wikimania 2008." The main text describes Porto as a municipality in Portugal with 41,66 km² of area and 227,790 inhabitants in 2006. It highlights its status as the "Capital do Norte" and "Cidade Invicta". The text also mentions its history, including its former name "Portus" and its role as the capital of the Condado Portucalense. Below the text, there are images of the Porto coat of arms (Brasão) and flag (Bandeira). At the bottom of the page, under the heading "Categorias:", the links "Porto" and "Concelhos do Grande Porto" are shown, both highlighted with red boxes.

1. “Concelhos do Grande Porto”:



2. “Categoria:Porto”:



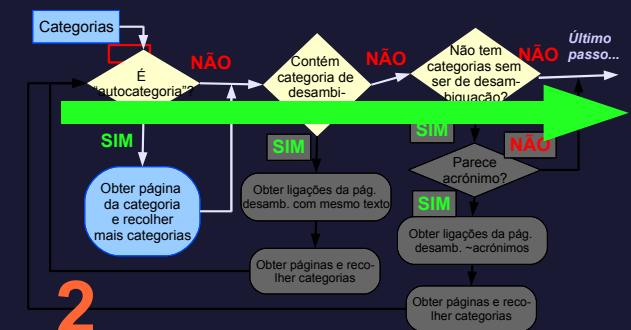
Exemplos (4/6)

3. O caso de “auto-categorias”. (2/2)



The screenshot shows the 'Categoria:Porto' page on Wikipedia. A red box highlights the 'Categorias' section at the bottom of the page, which contains the text 'Categorias: Cidades de Portugal | Municípios de Portugal'. The rest of the page displays information about the Porto category, including its sub-categories (Bairros do Porto, Freguesias do Porto, Futebol Clube do Porto, Metro do Porto, Museus do Porto, Património edificado no Porto, Teatros do Porto, and Universidades do Porto), and a list of articles in the 'Artigos na categoria "Porto"' section.

“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.

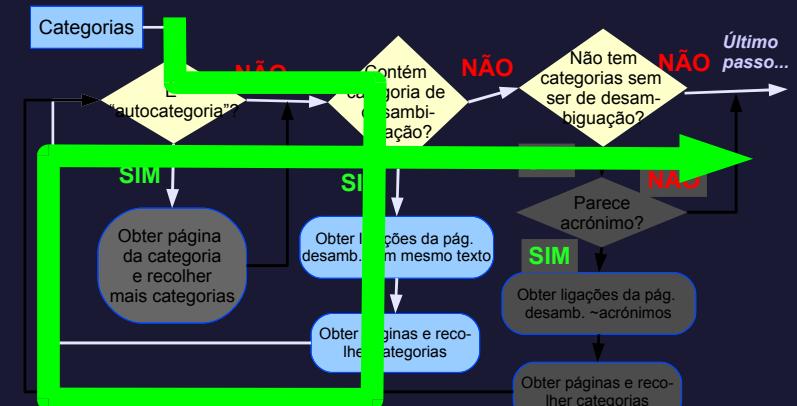
The screenshot shows the Wikipedia disambiguation page for the surname "Armstrong". The title bar says "Armstrong - Wikipédia". The main content area starts with a notice about being a disambiguation page. Below it, two sections are highlighted with red boxes: "Pessoas" (People) and "Locais" (Places). The "Pessoas" section lists famous people with their names in blue links. The "Locais" section lists Armstrong as a place name with its variants in blue links. A sidebar on the left contains links for "navegação", "colaboração", "busca", "ferramentas", and "outras línguas".

Ligações de saída da página também inclui:

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

Solução:

- Usar apenas ligações com “Armstrong” na âncora



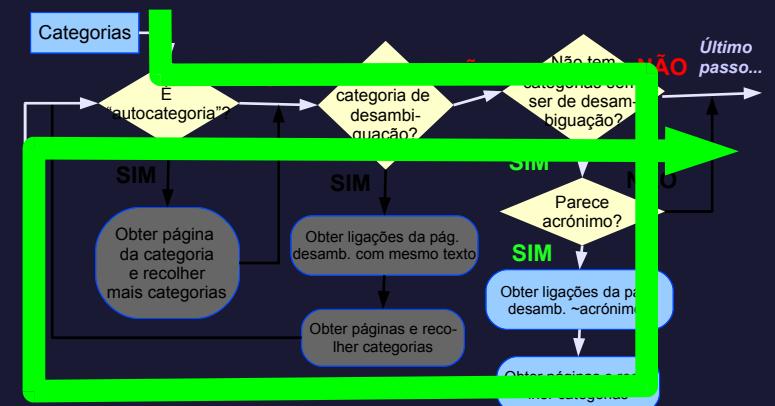
Exemplos (6/6)

5. O caso de acrónimos.

The screenshot shows the Portuguese Wikipedia page for the acronym 'PSP'. The page title is 'PSP' and the subtitle reads 'Origem: Wikipédia, a encyclopédia livre.'. Below the title, there is a note: 'Esta é uma página de desambiguação, que lista artigos associados a um mesmo título.' A red box highlights the section 'PSP pode referir-se à sigla de:' which lists five potential meanings. Another red box highlights the category link 'Categorias: Desambiguação | Acrônimos' at the bottom of the page. A large red arrow points from the bottom of the page up towards the category link.

Raramente usada...

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



2

Detector de relações (1/3)

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

“A XPTO foi por momentos (...) tal como a XPTO, que foi (...). Venha conhecer a **empresa <ORG>XPTO</ORG>!**”



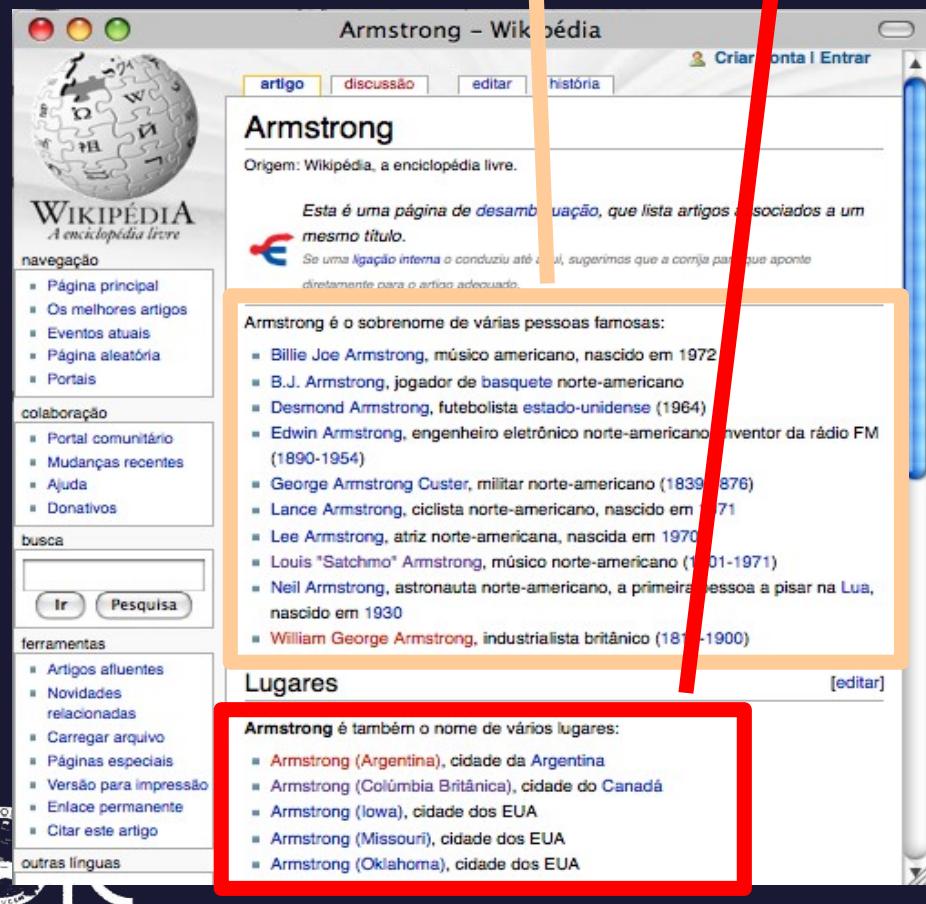
Regra de evidência externa

“A <ORG>XPTO</ORG> foi por momentos (...) tal como a <ORG>XPTO</ORG>, que foi (...). Venha conhecer a empresa <ORG>XPTO</ORG>!”

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

The screenshot shows the Wikipedia article for the Ritz Hotel in London. The page includes the following elements:

- Header:** "Ritz Hotel – Wikipedia, the free encyclopedia".
- Top menu:** article, discussion, edit this page, history.
- Log in / create account:** button.
- Image:** Wikipedia logo.
- Section: Ritz Hotel**
 - From Wikipedia, the free encyclopedia (Redirected from Hotel Ritz)
 - For other uses, see [Ritz \(disambiguation\)](#).
- Text box (highlighted in red):** "The Ritz Hotel London is a 133-room hotel located in Piccadilly and overlooking Green Park in London."
- Section: Contents [hide]**
 - 1 History
 - 2 Facilities
 - 3 Fire
 - 4 See also
 - 5 External links
 - 6 References
- Section: History**
- Section: Famous Swiss hotelier César Ritz**
- Footnote:** "Categories: Hotels in London | Edwardian architecture"
- Image:** Logo of the Ritz Hotel.
- Image:** Exterior view of the Ritz Hotel building.
- Text:** "The arcade faces Piccadilly".
- Section: Hotel facts and statistics**
 - Location: London, United Kingdom
- Toolbox:** Go, Search, What links here.

Trabalho futuro

- Melhorar o desempenho para Inglês.
- Melhorar o mecanismo de inferência de geograficidade das EM.
- Optimizar, sobretudo 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 Mencionadas
Baseado em Relações
e ANálise Detalhada do
Texto

Nuno Cardoso
Faculdade de Ciências, Universidade de Lisboa
Laboratório LASIGE

ncardoso@xldb.di.fc.ul.pt

