FACULDADE·DE·CIÊNCIAS UNIVERSIDADE·DE·LISBOA

# REMBRANDT

## Named-entity recognition based on Wikipedia and manual rules.


Nuno Cardoso

Faculty of Sciences, University of Lisbon
LaSIGE Laboratory, XLDB Team

ncardoso@xldb.di.fc.ul.pt


# Introduction

- **REMBRANDT** – named entity recognition and entity relation detection software for Portuguese.
  (English is also supported, other languages will follow).

*"Rembrandt Harmenszoon van Rijn was born in June 15, 1606 (tradicionally) but probably in 1607, in Leiden, Netherlands. (...) In 1634, Rembrandt married to Hendricks' cousin, Saskia van Uylenburgh. In 1639, Rembrandt and Saskia moved to Jodenbreestraat."*

# Goals

1. Identify & classify all kinds of named entities (NE), and extract their "geograficness" (Santos and Chaves, 2006);
2. Process large collections of documents.
3. Explore knowledge resources (Wikipedia, DBpedia, web collections, ontologies, …) for NE desambiguation and classification.

# Features

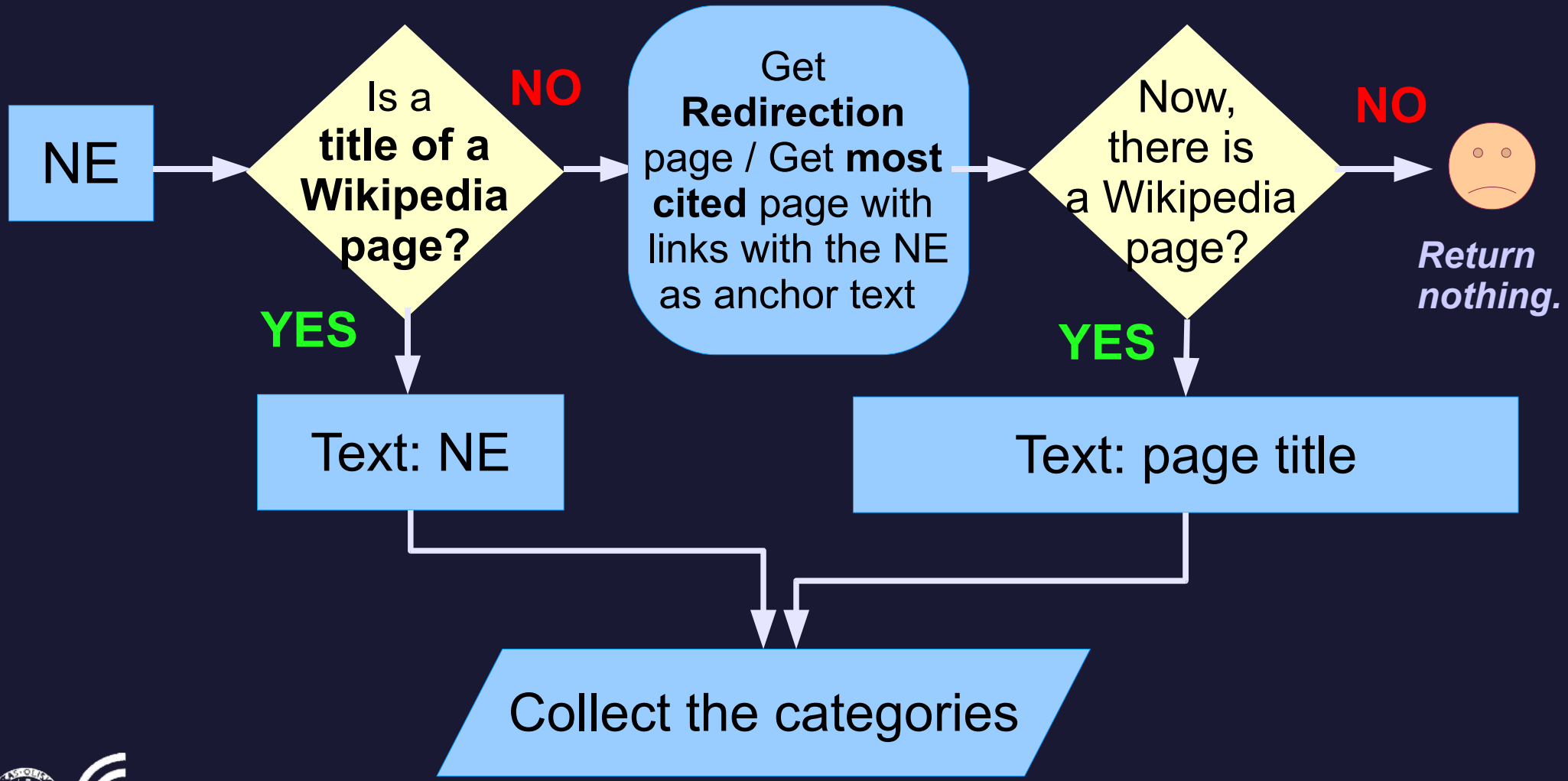
- Inspired on the NER system **Palavras_NER** (Bick, 2006):
  - No machine learning (yet);
  - Developed for Portuguese, adapted for English;
  - Explores internal & external evidences of NE for each language, through manually- crafted rules.
- Knowledge source (so far): **Wikipedia**
  - Uses Wikipedia categories for NE classification;
  - Links & disambiguation pages used for difficult NE.
  - **SASKIA** is REMBRANT's knowledge miner.

# Receipt

1. Initial numeric, timestamp and value pattern matching;
2. Generation of *candidate NE* (clusters of capitalized words + a few stopwords);
3. For each candidate NE, lauch SASKIA + external & evidence rules;
4. Second round of rules, using the first tier of classifications;
5. Entity relation detection;
6. Last minute recall on some NE leftovers.

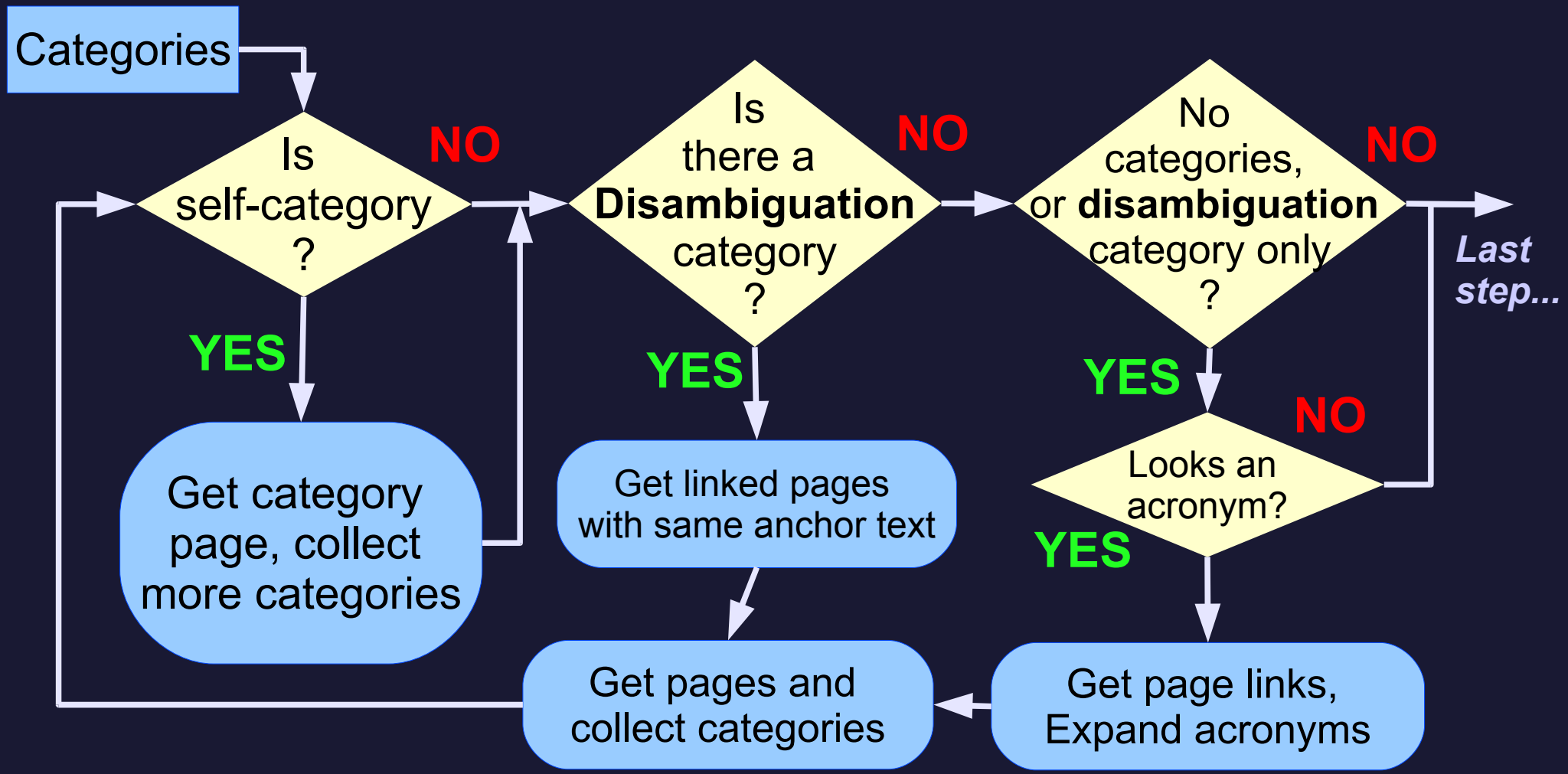
# SASKIA (1/3)

## 1 – Match each NE to a Wikipedia page



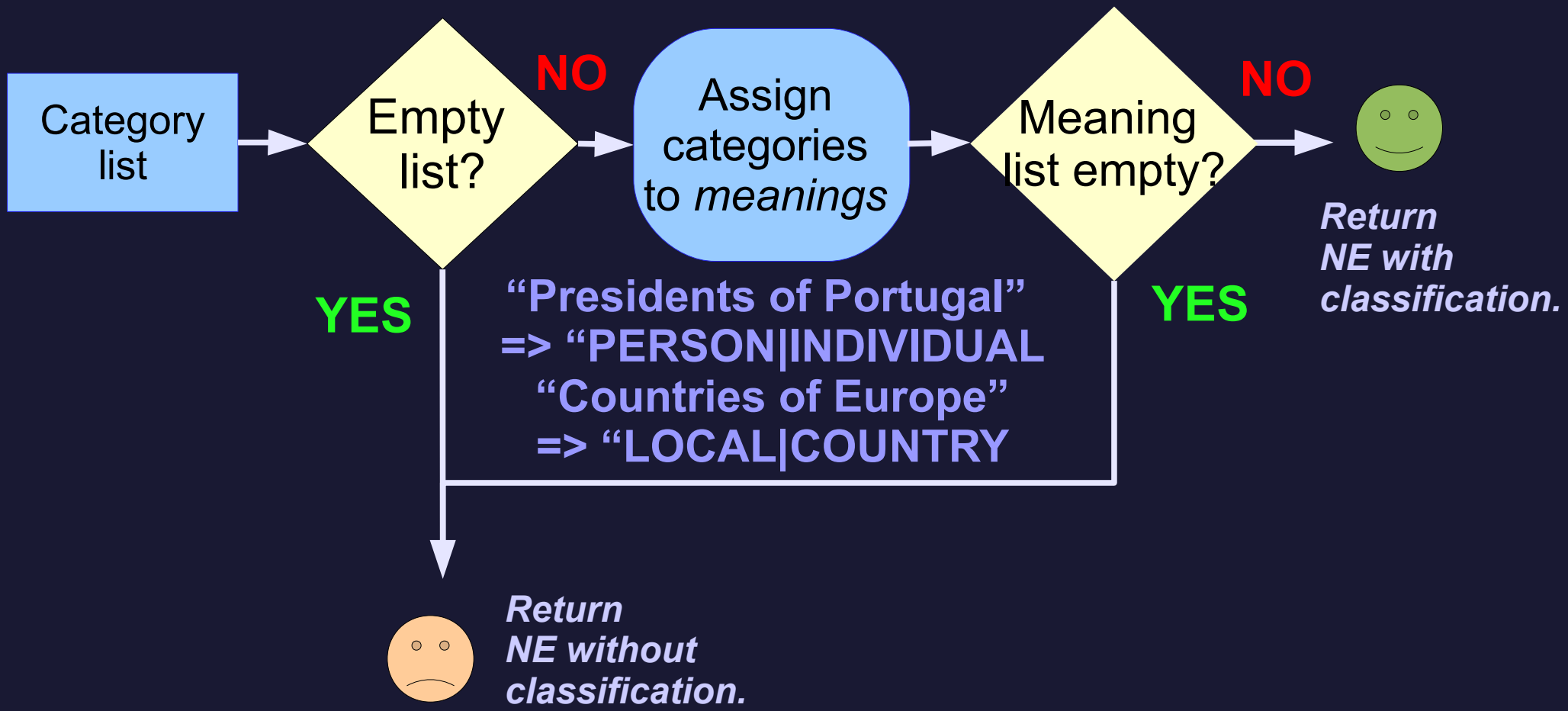
# Saskia (2/3)

## 2 – Collect categories



# Saskia (3/3)

## 3 – Match Wikipedia categories to NE classifications

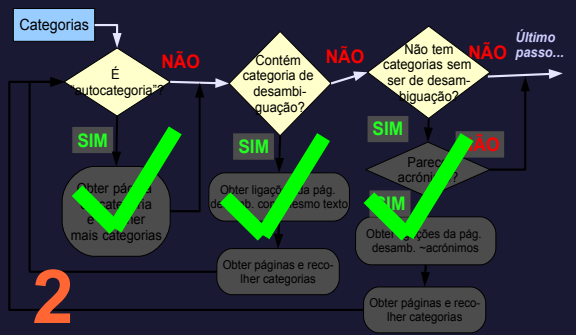


# Examples

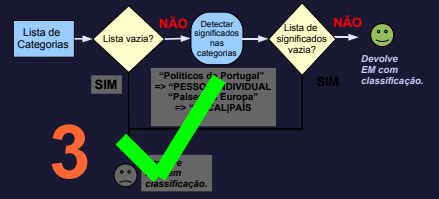
1. The 'simple & straightforward' case:


"Fernando Ribeiro"


"Esboços de biografia..."
"Cantores de Portugal"
"Moonspell"
"Cantores de heavy metal"

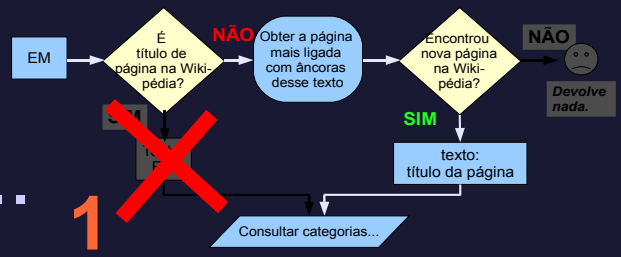



**Cantores de X -> PERSON**



# Examples

## 2. The “indirect” case: 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.*

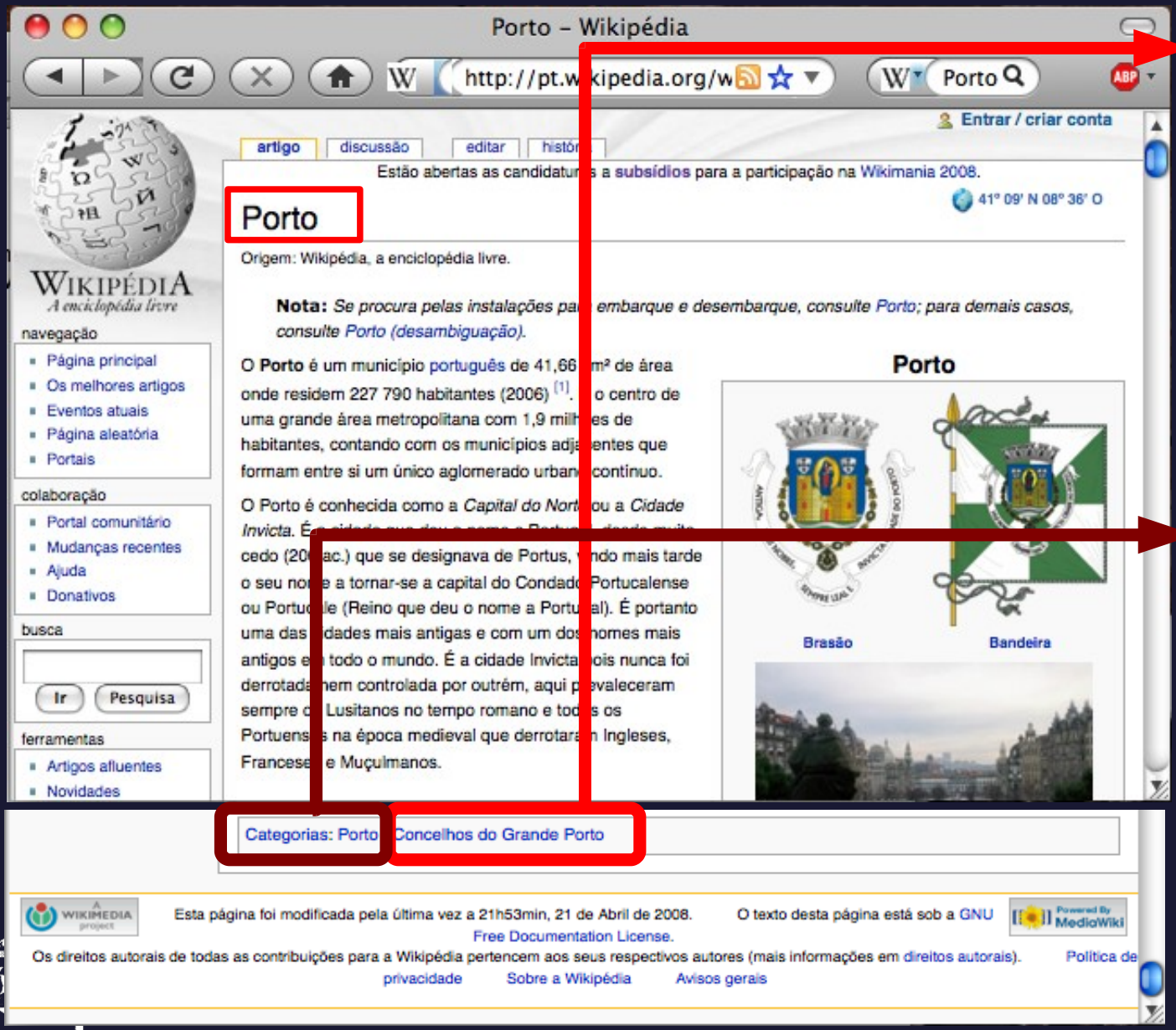
| | |
|---|---|
| Estados Unidos da América | 127 |
| Billboard Hot 100 | 1 |

*USA*

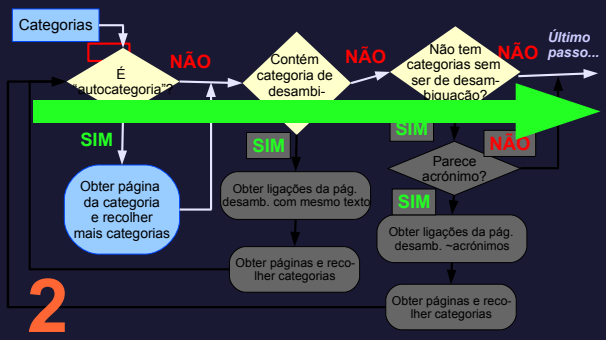
| | |
|---|---|
| Estados Unidos da América | 163 |
| Estados Unidos da América nos Jogos Pan-americanos de 2007 | 25 |
| Grande Prémio dos Estados Unidos de 2007 (Fórmula 1) | 14 |
| (...) | ... |

# Examples

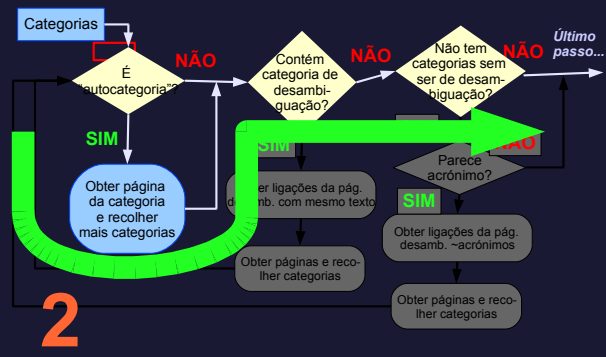
## 3. The “self-category” case (1/2):




1. “Concelhos do Grande Porto”:



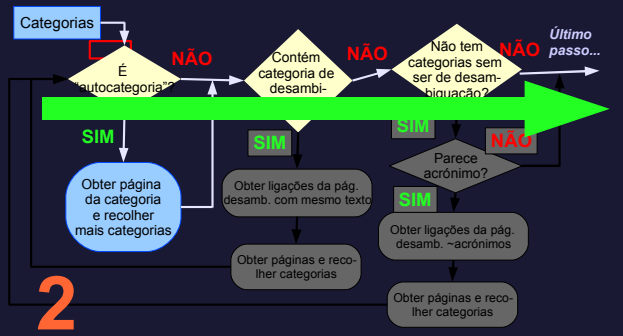

2. “Categoria:Porto”:



# Examples

## 3. The “self-category” case (2/2):


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




Final category list:

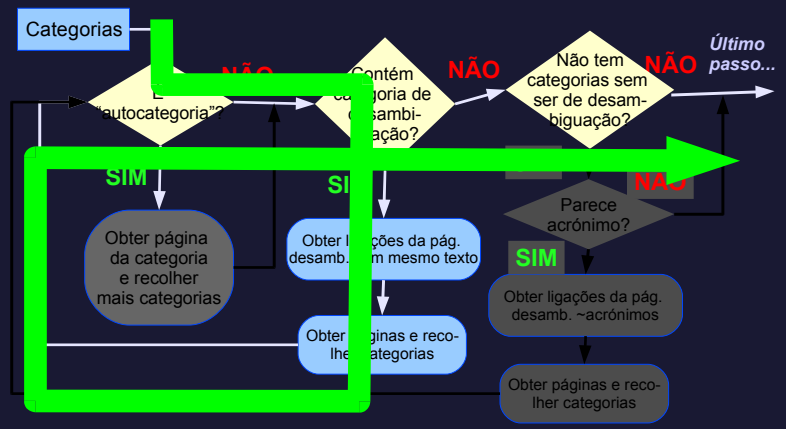
1. “Conselhos do Grande Porto”
2. “Cidades de Portugal”
3. “Municípios de Portugal”

# Examples

## 4. The “disambiguation page” case:


Problem;
Links include:
*basquete, 1971, 1810, 1900, Argentina, Lua, ...*

Solution:
- Use only links with “Armstrong” as anchor text

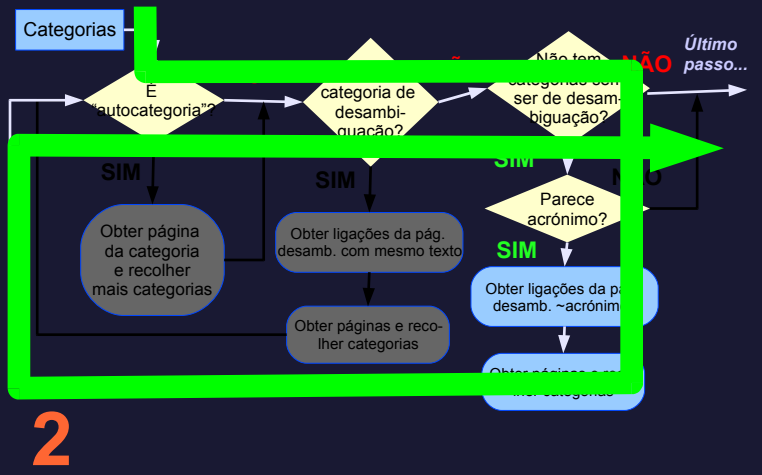



2

# Examples

## 5. The “acronym” case:

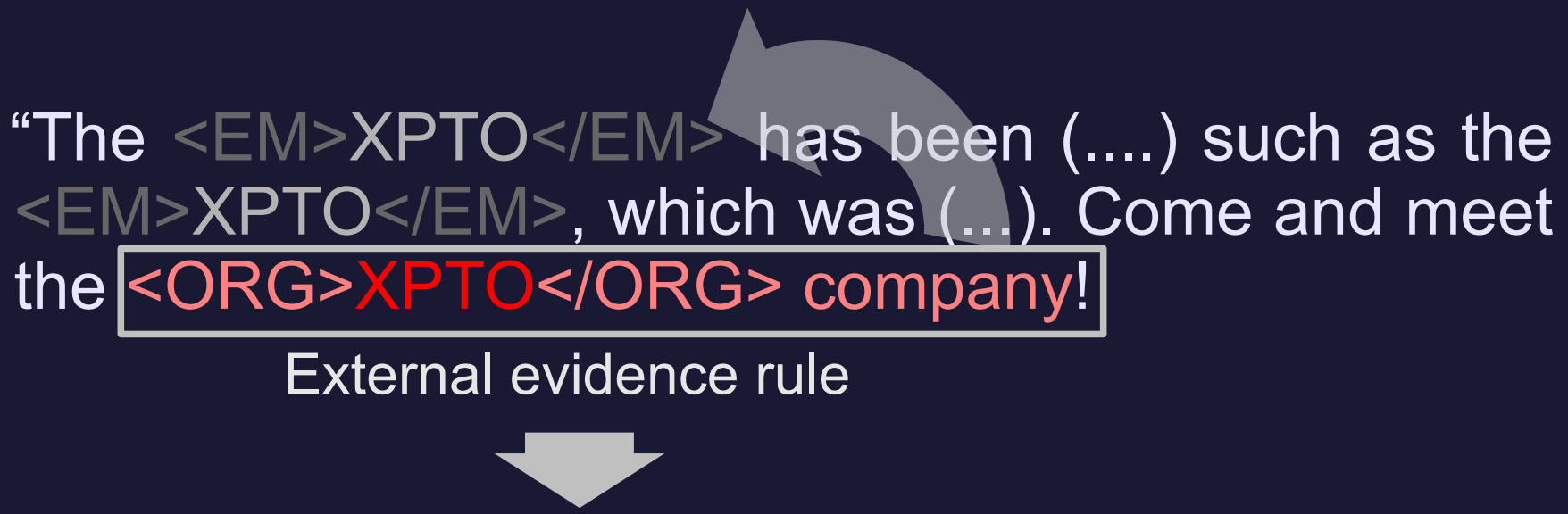

- Use only links with anchor text containing words that might recreate the acronym.




'Acronym' category is rarely used, unfortunately...

# Entity relation detection (1/3)

- Rules over NE classification and term similarity.
  - Helps to classify some NE leftovers.



# Entity relation detection (2/3)

- Exploring Wikipedia links.

" <PERSON(?)/LOCAL(?)>Armstrong</> participated on the <PROJECT>Gemini</> project.


Armstrong: links to 10 persons and 4 cities.

**It is likely that neighbor NE (i.e., on the same sentence) may occur on one of these documents, helping the disambiguation process.**

# Entity relation detection (3/3)

<INDIVIDUAL(!)/
~~DIVISAO(X)~~>
Armstrong</> participated
on <PROJECT>Gemini</>
project.

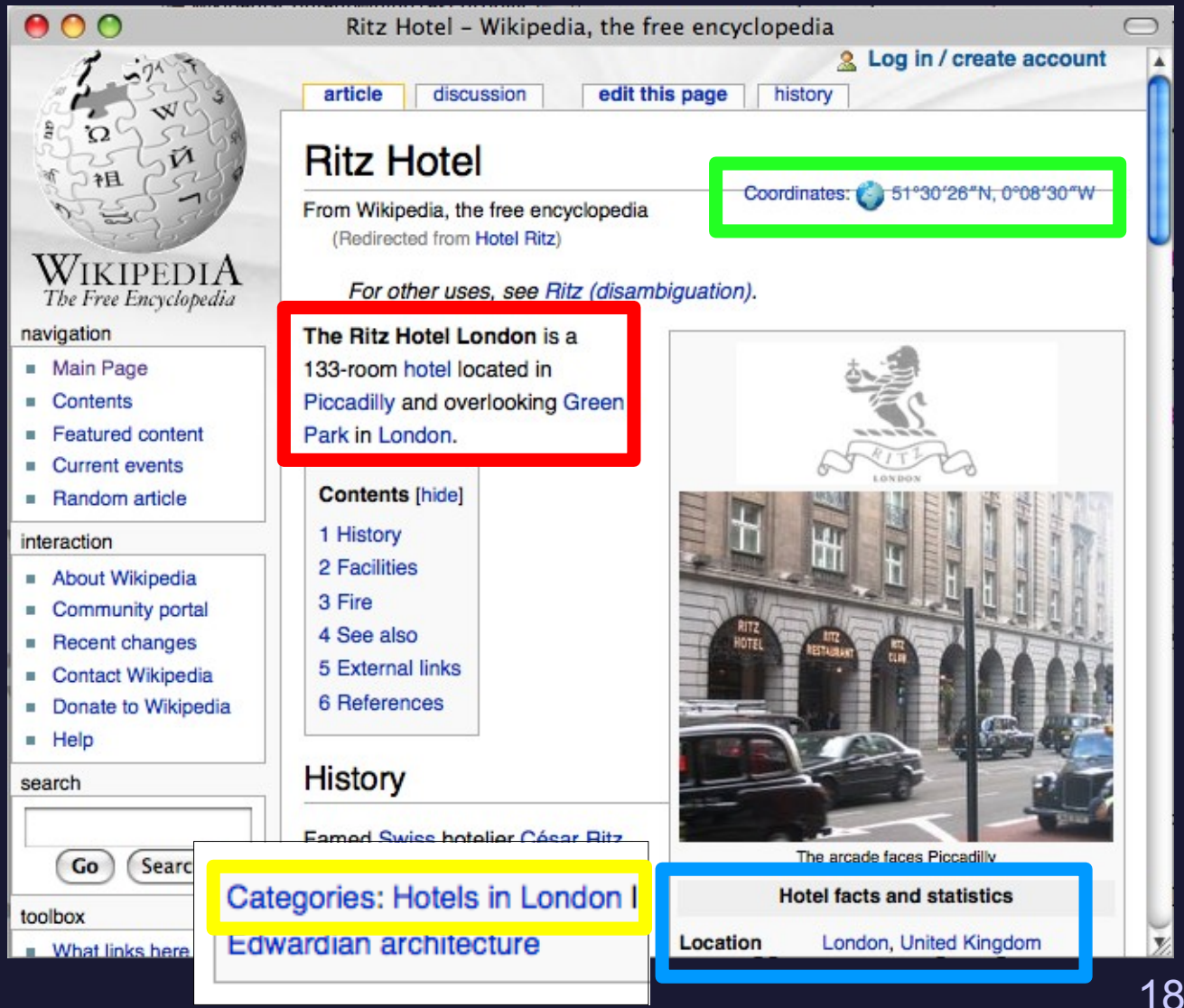
# Future work

- Better Wikipedia mining.

Explore:

- coordinates
- infoboxes
- Initial paragraph

Use DBpedia info.



# End.

## Rembrandt

Nuno Cardoso
Faculdade de Ciências, Universidade de Lisboa
Laboratório LASIGE, XLDB
xldb.di.fc.ul.pt
xldb.di.fc.ul.pt/wiki/Nuno_Cardoso

ncardoso@xldb.di.fc.ul.pt