



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 (traditionally) 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.



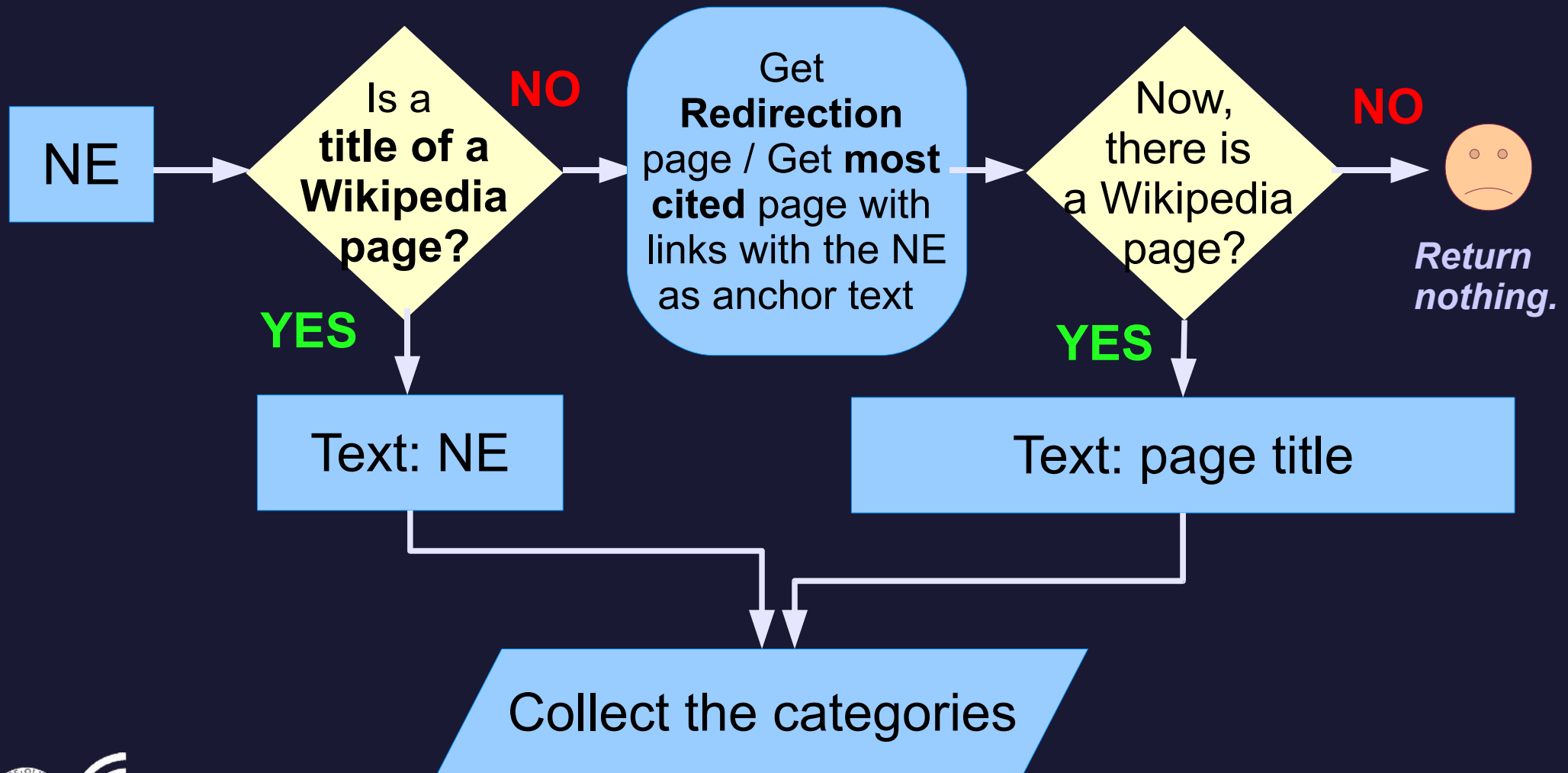
Receipt

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, launch 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.



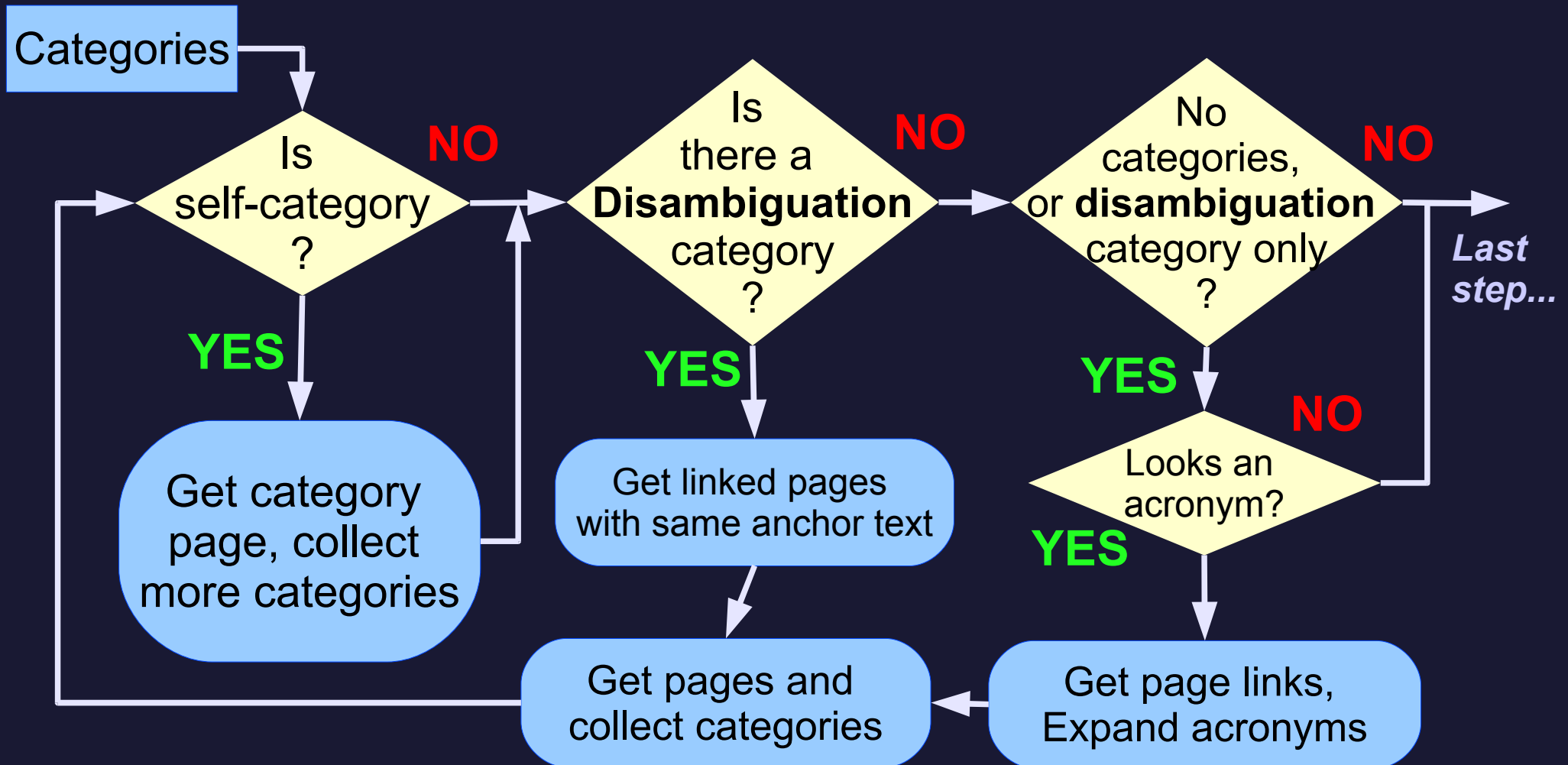
SASKIA (1/3)

1 – Match each NE to a Wikipedia page



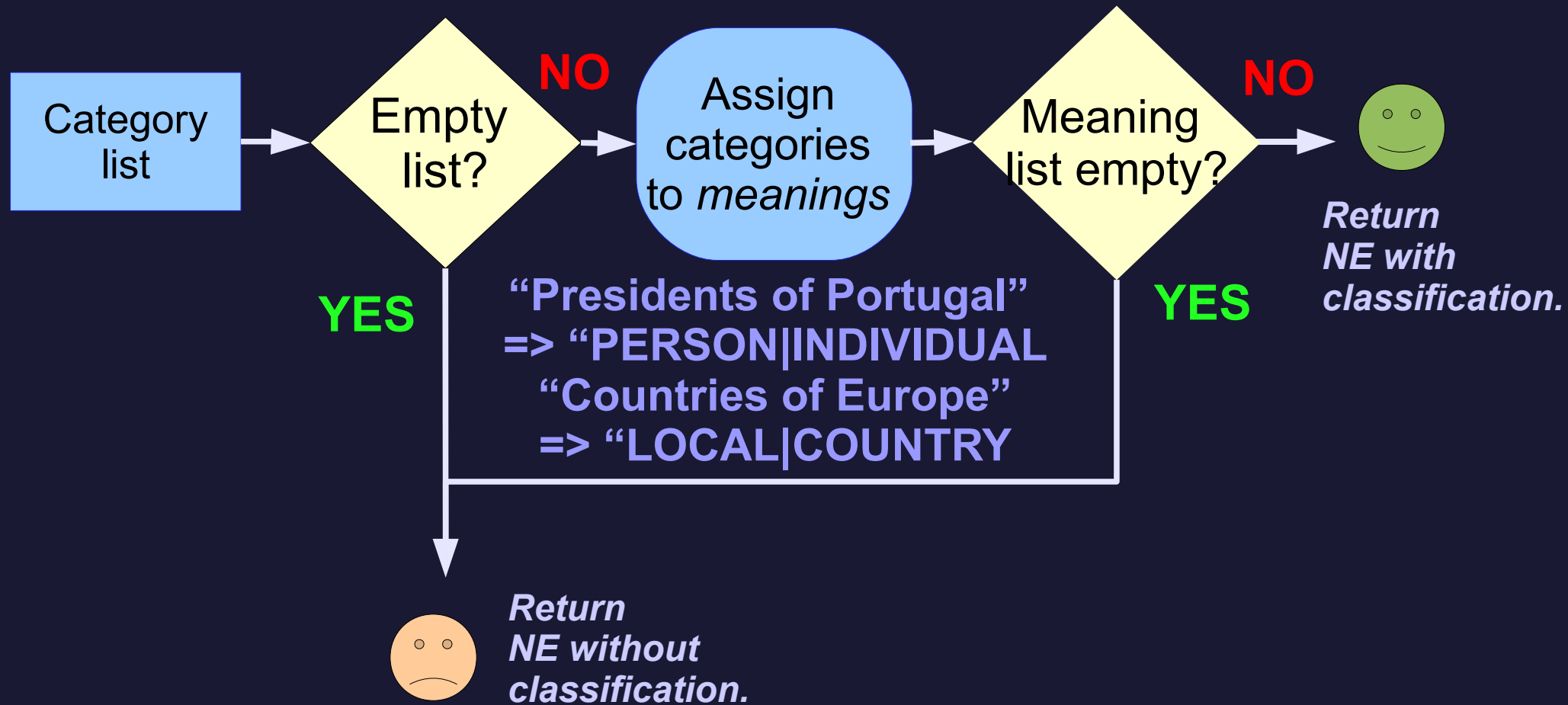
Saskia (2/3)

2 – Collect categories



Saskia (3/3)

3 – Match Wikipedia categories to NE classifications

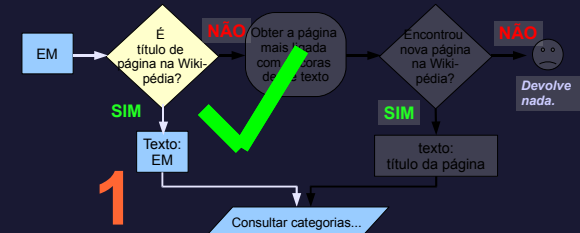


Examples

1. The 'simple & straightforward' case:



“Fernando Ribeiro”

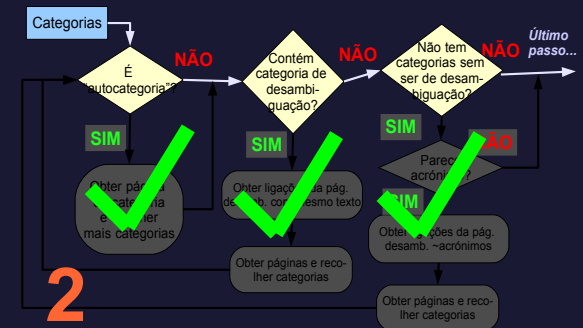


“Esboços de biografia...”

“Cantores de Portugal”

“Moonspell”

“Cantores de heavy metal”

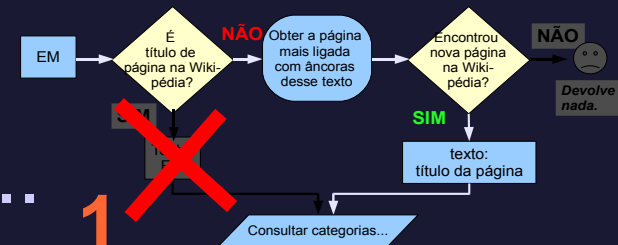


Cantores de X -> PERSON



Examples

2. The “indirect” case: EUA, Estados Unidos, U.S., U.S.A., ...



EUA

Estados Unidos da América	3325
Billboard Hot 100	34
Seleccção 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
Seleccção de Futebol dos Estados Unidos da América	31
Seleccção 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
(...)	...

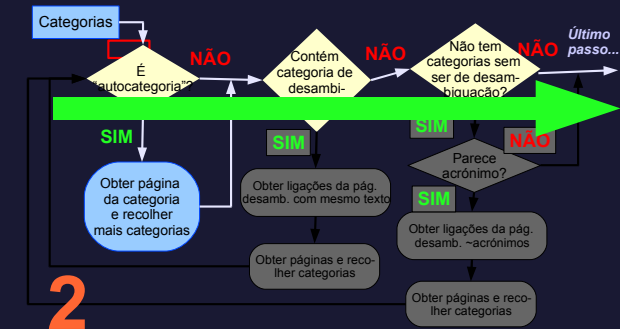


Examples

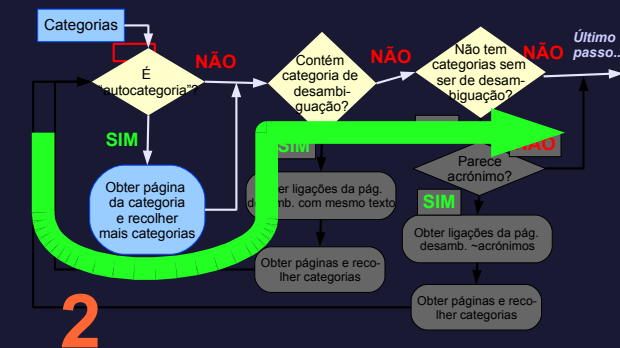
3. The “self-category” case (1/2):

The screenshot shows the Portuguese Wikipedia page for 'Porto'. The title 'Porto' is highlighted with a red box. Below the title, there is a note about the Wikimania 2008 subsidies. The main text describes Porto as a municipality in Portugal. To the right, there are images of the coat of arms (Brasão) and the flag (Bandeira) of Porto. At the bottom of the page, the 'Categories' (Categorias) section is highlighted with a red box, showing 'Porto' and 'Concelhos do Grande Porto'. A red arrow points from the 'Porto' title to the 'Categorias' section.

1. “Concelhos do Grande Porto”:



2. “Categoria:Porto”:

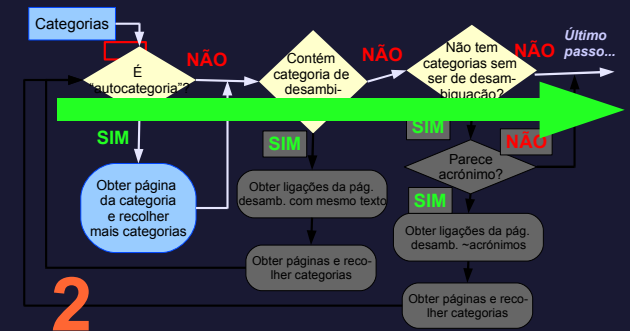


Examples

3. The “self-category” case (2/2):

The screenshot shows the Wikipedia page for the category "Categoria:Porto". The title "Categoria:Porto" is highlighted with a red box. Below the title, it states "Origem: Wikipédia, a enciclopédia livre." and "Esta categoria reúne artigos sobre Porto." The "Subcategorias" section lists 8 sub-categories: "Bairros do Porto", "Futebol Clube do Porto", "Património edificado no Porto", "Freguesias do Porto", "Metro do Porto", "Teatros do Porto", "Museus do Porto", and "Universidades do Porto". At the bottom, a red box highlights the text "Categorias: Cidades de Portugal | Municípios de Portugal".

“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:



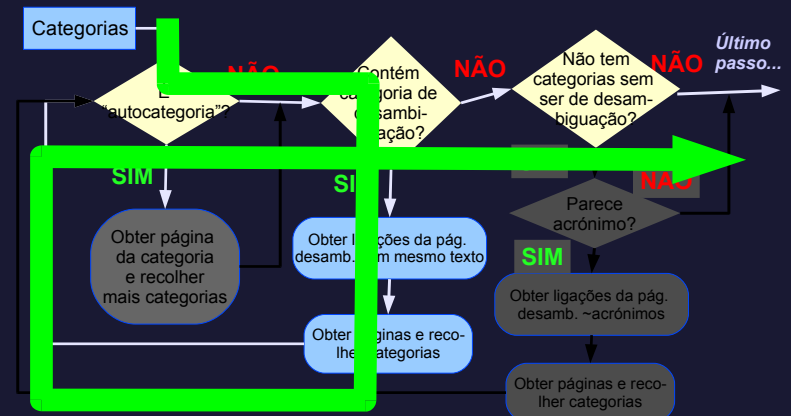
Problem;

Links include:

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

Solution:

- Use only links with “Armstrong” as anchor text

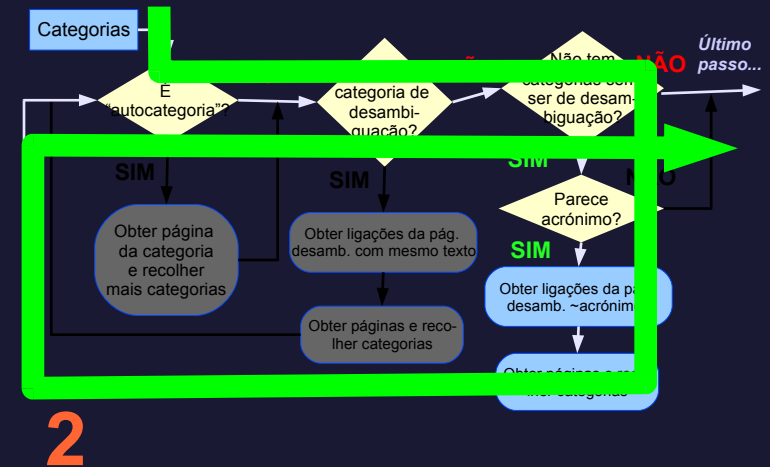


Examples

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.

“The `XPTO` has been (...) such as the `XPTO`, which was (...). Come and meet the `<ORG>XPTO</ORG>` company!”

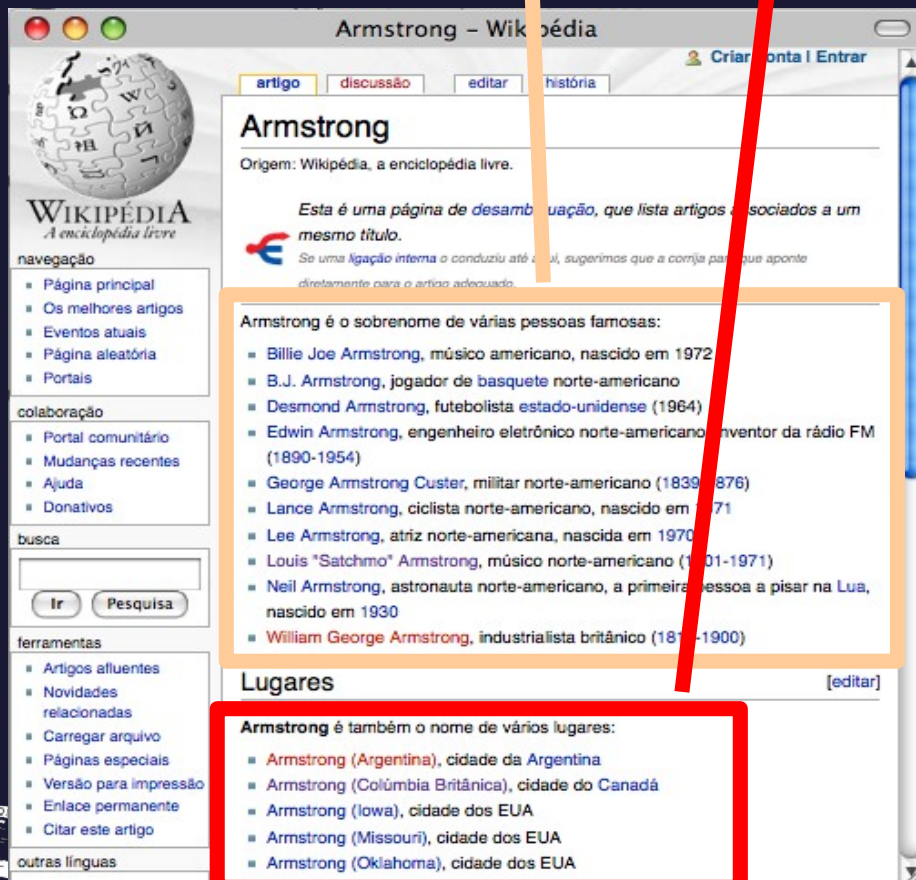
External evidence rule

“The `<ORG>XPTO</ORG>` has been (...) such as the `<ORG>XPTO</ORG>`, which was (...). Come and meet the `<ORG>XPTO</ORG>` company!”

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.



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

Future work

- Better Wikipedia mining.

Explore:

- coordinates
- infoboxes
- Initial paragraph

Use DBpedia info.

The screenshot shows the Wikipedia page for the Ritz Hotel. Several elements are highlighted with colored boxes to illustrate data extraction points:

- Coordinates:** A green box highlights the text "Coordinates: 51°30'26"N, 0°08'30"W".
- Initial Paragraph:** A red box highlights the first sentence of the main text: "The Ritz Hotel London is a 133-room hotel located in Piccadilly and overlooking Green Park in London."
- Categories:** A yellow box highlights the "Categories" section, listing "Hotels in London" and "Edwardian architecture".
- Location:** A blue box highlights the "Location" field in the "Hotel facts and statistics" table, showing "London, United Kingdom".

Other visible elements include the Wikipedia logo, navigation links (Main Page, Contents, etc.), a table of contents, a history section, and a photograph of the hotel's facade.

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

