



ANT Pesquisa de Informação na Universidade do Porto

José Devezas
joseluisdevezas@gmail.com

INESC TEC & FEUP InfoLab
MAP-i 2016/2017



Universidade do Minho



universidade
de aveiro





Conteúdo

- Introdução
 - Pesquisa via correspondência de palavras-chave
 - Pesquisa via correspondência semântica
 - A relevância das entidades na pesquisa
 - Pesquisa orientada a entidades
- ANT
 - Arquitetura do motor de busca
 - Compreensão da consulta
 - Score Hypergraph
- Conclusões
 - Observações finais
 - Projetos relacionados
 - Uma visão do futuro

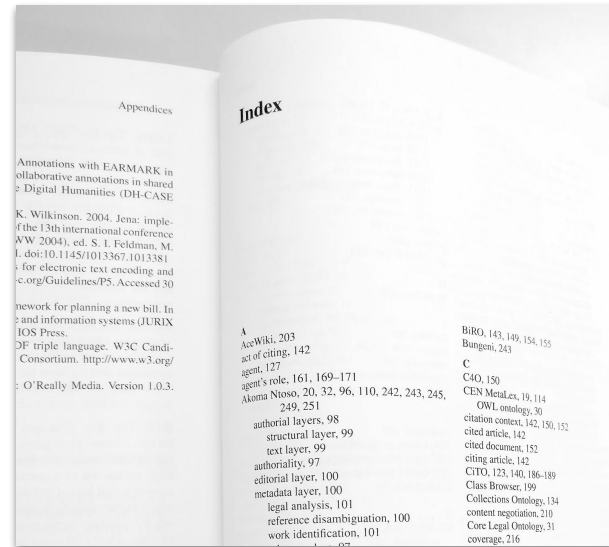
Introdução

O que é pesquisa orientada a entidades, e qual o interesse?



Pesquisa via correspondência de palavras-chave

- Modelada com base no índice remissivo dos livros.
- Encontrar conteúdo relevante envolve:
 1. Selecionar uma ou várias palavras-chave;
 2. Saltar para as páginas indicadas;
 3. Ler passagens e usar conhecimento, quer interno quer externo ao livro, para avaliar a relevância.



Pesquisa via correspondência semântica

- Mais próximo da necessidade de informação do utilizador.
- Requer a interpretação do sentido da consulta e da semântica dos documentos.
- Integra dados não-estruturados do texto e dados estruturados das bases de conhecimentos.

The screenshot shows a Google search for "sci-fi movies from 1985". The search results include a carousel of movie posters for "Enemy Mine", "Explorers", "Cocoon", "Back to the Future", "Lifeforce", "Weird Science", "Real Genius", and "Re-Animator". Below the carousel, there are search results for "IMDb: Top 10 Sci Fi Movies of 1985 - a list by jeffshibley" and "List of science fiction films of the 1980s - Wikipedia".

Pesquisa via correspondência semântica

- Torna-se possível responder, de forma mais adequada, a consultas como:
 - [president of portugal 2018]
 - [sci-fi movies from 1985]
- Em vez de documentos que contêm estas palavras-chave;
- É fornecida diretamente uma lista de entidades com a resposta.
- Deixa de ser necessária a leitura dos documentos para encontrar a resposta.

The screenshot shows a Google search for "sci-fi movies from 1985". The search results include a carousel of movie posters for "Enemy Mine", "Explorers", "Cocoon", "Back to the Future", "Lifeorce", "Weird Science", "Real Genius", and "Re-Animator". Below the carousel, there are search results from IMDb and MovieWeb, and a Wikipedia link for "List of science fiction films of the 1980s".

Google sci-fi movies from 1985

All Images Videos News More Settings Tools

Movies Science Fiction < 1985 >

Enemy Mine 1985 Explorers 1985 Cocoon 1985 Back to the Future 1985 Lifeorce 1985 Weird Science 1985 Real Genius 1985 Re-Ani 1985

IMDb: Top 10 Sci Fi Movies of 1985 - a list by jeffshibley
www.imdb.com/list/ls056100695/ ▼
 Aug 5, 2013 - A boy obsessed with 50s sci-fi movies about aliens has a recurring dream about a blueprint of some kind, which he draws for his inventor friend ...

1985 Sci-Fi Movies
movieweb.com/movies/1985/sci-fi/ ▼
 1985 Sci-Fi Movies, 1985 Sci-Fi Movies In Theaters, Upcoming Sci-Fi Movies 1985, New Sci-Fi 1985 Movies.

List of science fiction films of the 1980s - Wikipedia
https://en.wikipedia.org/wiki/List_of_science_fiction_films_of_the_1980s ▼
 A list of science fiction films released in the 1980s. These films include core elements of ... Years [hide] 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 ... 1980 1981 1982 1983



A relevância das entidades na pesquisa

Nas consultas:

- O estudo do AOL Query Log mostrou que:
 - 18-39% das consultas se referem diretamente a entidades;
 - 73-87% das consultas contêm pelo menos uma entidade.

Nos documentos:

- O conjunto de dados de treino, em inglês, anotado para o CoNLL 2003 contém:
 - 14,987 frases;
 - 23,499 entidades;
 - 1.6 entidades por frase.



Pesquisa orientada a entidades

Consoante a consulta, os resultados podem ser:

- Documentos
 - Recuperados com recurso à informação semântica (entidades e suas relações).
- Entidades
 - Uma específica, uma lista, ou ambos.
 - Recuperadas por nome, tipo ou outra descrição.
 - Representando atributos ou relações.

No ANT, as entidades podem ser:

- Estudantes
- Pessoal
- Departamentos
- Salas
- Cadeiras
- Cursos
- Notícias

ANT

Arquitetura do motor de busca e método de compreensão da consulta.





ANT



departamento de engenharia informática

Todos Estudantes Notícias Funcionários Cadeiras Salas Cursos Departamentos Ferramentas de Pesquisa

91049 resultados (0.17 segundos)

Departamento de Engenharia Informática
Departamento https://sigarra.up.pt/feup/pt/uni_geral.unidade_view?pv_unidade=151
Faculdade de Engenharia da Universidade do Porto (FEUP)
Responsável: Raul Fernando de Almeida Moreira Vidal

Programa Doutoral em Engenharia Informática
Curso https://sigarra.up.pt/feup/pt/cur_geral.cur_view?pv_ano_lectivo=2016...
Faculdade de Engenharia da Universidade do Porto (FEUP)
Áreas Científicas Predominantes: Engenharia Informática
Diretores: Eugénio da Costa Oliveira

Ad hoc search of entities and Text.

- Suporta cinco categorias de consulta definidas por Pound et al. (2010):
 - Entidade;
 - Tipo;
 - Atributo;
 - Relação;
 - Palavras-chave.
- Requer dois índices Lucene:
 - Índice de análise da consulta;
 - Índice de entidades.
- E uma quad store no Virtuoso:
 - Útil para consultas de atributo e relações.

Logo: 

Search bar:

Navigation: [Todos](#) | [Notícias](#) | [Estudantes](#) | [Salas](#) | [Pessoal](#) | [Cadeiras](#) | [Cursos](#) | [Departamentos](#) | [Ferramentas de Pesquisa](#)

26340 resultados (16.51 segundos) 

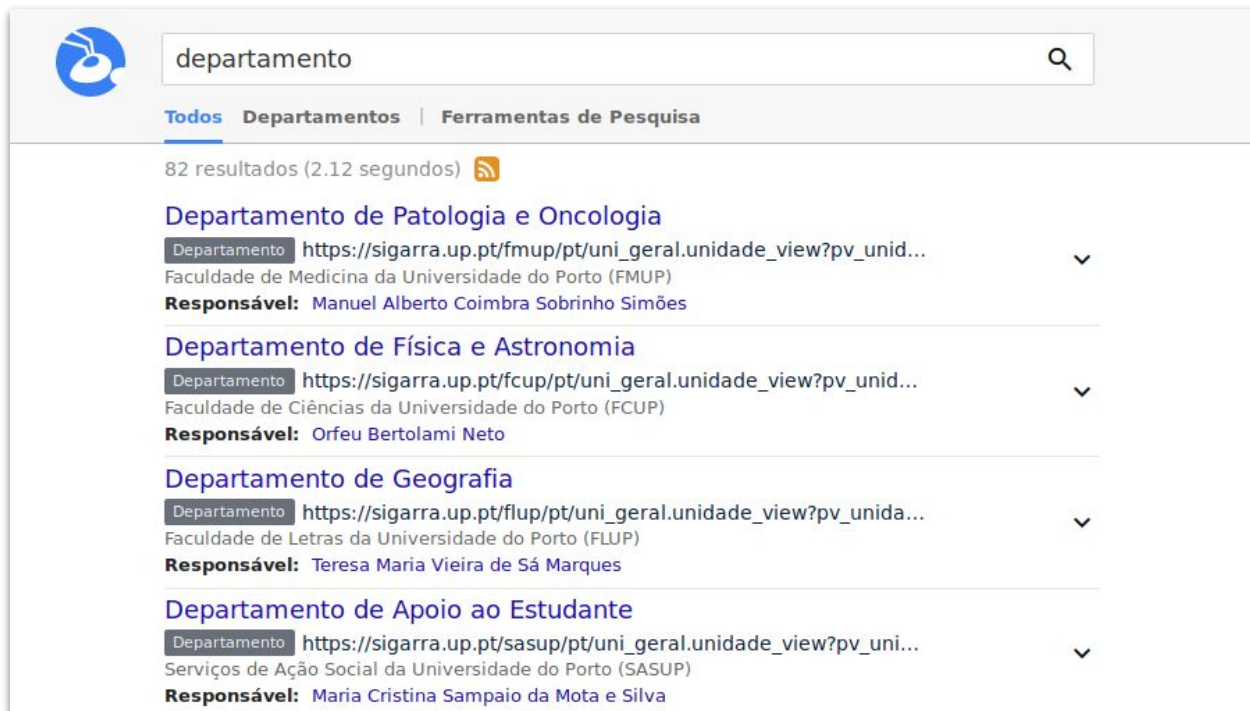
- 

José Luís da Silva Devezas
Estudante https://sigarra.up.pt/flup/pt/vld_entidades_geral.entidade_pa...
Faculdade de Belas Artes da Universidade do Porto (FBAUP) (mais 4)
Curso: [Doutoramento em Informática](#)
Código: 200303288
- 

Vítor Bruno dos Santos Devezas
Estudante https://sigarra.up.pt/icbas/pt/vld_entidades_geral.entidade_p...
Instituto de Ciências Biomédicas Abel Salazar (ICBAS)
Curso: [Mestrado Integrado em Medicina](#)
Código: 200706081
- 

Diana Maria Rodrigues Alves Devezas
Estudante https://sigarra.up.pt/flup/pt/vld_entidades_geral.entidade_pa...
Faculdade de Letras da Universidade do Porto (FLUP) (mais 1)
Curso: [Direito](#)
Código: 200300607


Consulta entidade. A intenção da consulta é encontrar uma entidade.





The screenshot shows a search interface with a search bar containing the text "departamento". Below the search bar, there are navigation tabs: "Todos", "Departamentos" (which is selected), and "Ferramentas de Pesquisa". The search results are displayed as a list of four department entries, each with a title, a URL, the faculty name, and the responsible person. The results are: 1. Departamento de Patologia e Oncologia (FMUP), responsible: Manuel Alberto Coimbra Sobrinho Simões. 2. Departamento de Física e Astronomia (FCUP), responsible: Orfeu Bertolami Neto. 3. Departamento de Geografia (FLUP), responsible: Teresa Maria Vieira de Sá Marques. 4. Departamento de Apoio ao Estudante (SASUP), responsible: Maria Cristina Sampaio da Mota e Silva. Each entry has a small downward arrow icon on the right side.


departamento


Todos | **Departamentos** | Ferramentas de Pesquisa

82 resultados (2.12 segundos) 


Departamento de Patologia e Oncologia
Departamento https://sigarra.up.pt/fmup/pt/uni_geral.unidade_view?pv_unid... 
Faculdade de Medicina da Universidade do Porto (FMUP)
Responsável: Manuel Alberto Coimbra Sobrinho Simões

Departamento de Física e Astronomia
Departamento https://sigarra.up.pt/fcup/pt/uni_geral.unidade_view?pv_unid... 
Faculdade de Ciências da Universidade do Porto (FCUP)
Responsável: Orfeu Bertolami Neto

Departamento de Geografia
Departamento https://sigarra.up.pt/flup/pt/uni_geral.unidade_view?pv_unida... 
Faculdade de Letras da Universidade do Porto (FLUP)
Responsável: Teresa Maria Vieira de Sá Marques


Departamento de Apoio ao Estudante
Departamento https://sigarra.up.pt/sasup/pt/uni_geral.unidade_view?pv_uni... 
Serviços de Ação Social da Universidade do Porto (SASUP)
Responsável: Maria Cristina Sampaio da Mota e Silva

Consulta tipo. A intenção da consulta é encontrar entidades de um tipo ou classe em particular.



🔍


[Todos](#) [Salas](#) [Pessoal](#) [Estudantes](#) [Notícias](#) | [Ferramentas de Pesquisa](#)

4 resultados (3.81 segundos) 

José Luís da Silva Devezas

Sala

I123


 **José Luís da Silva Devezas**


Estudante https://sigarra.up.pt/flup/pt/vld_entidades_geral.entidade_pa...

Faculdade de Belas Artes da Universidade do Porto (FBAUP) (mais 4)

Curso: Doutoramento em Informática


Código: 200303288

 **José Luís da Silva Devezas**


Pessoal  https://sigarra.up.pt/feup/pt/func_geral.FormView?P_CODIGO...

Faculdade de Engenharia da Universidade do Porto (FEUP) (mais 1)

Código: 493720 **Sigla:** JLSD

 1317

I123

Sala  https://sigarra.up.pt/feup/pt/instal_geral.espaco_view?pv_id=...

Laboratório de I&D de Sistemas de Informação, Faculdade de Engenharia da Universidade do Porto (FEUP)

Responsáveis: [Carla Alexandra Teixeira Lopes](#)

Edifício: Electrotecnia (I) **Andar:** 1

Consulta atributo. A intenção da consulta é encontrar valores para um atributo particular de uma entidade ou tipo.



6 resultados obtidos em 2,88 segundos

José Luís da Silva Devezas

Faculdade de Belas Artes da Universidade do Porto (FBAUP), Faculdade de Letras da Universidade do Porto (FLUP), Faculdade de Engenharia da Universidade do Porto (FEUP), Faculdade de Economia da Universidade do Porto (FEP), Faculdade de Ciências da Universidade do Porto (FCUP)

Estudante

score = 0.002238

**Tiago Nuno Mesquita Folgado leitão Devezas**

Faculdade de Engenharia da Universidade do Porto (FEUP) Domingo, 07 Outubro 2018, 00h32

Pessoal

score = 0.001971

**José Luís da Silva Devezas**

Faculdade de Engenharia da Universidade do Porto (FEUP), Faculdade de Ciências da Universidade do Porto (FCUP) Domingo, 02 Setembro 2018, 00h26

Pessoal

score = 0.001106

**Tiago Nuno Mesquita Folgado Leitão Devezas**

Faculdade de Belas Artes da Universidade do Porto (FBAUP), Faculdade de Letras da Universidade do Porto (FLUP), Faculdade de Engenharia da Universidade do Porto (FEUP), Faculdade de Economia da Universidade do Porto (FEP), Faculdade de Ciências da Universidade do Porto (FCUP)

Estudante

score = 0.001789



I123

Faculdade de Engenharia da Universidade do Porto (FEUP)

Sala

score = 0.000431

**Projetos no laboratório SAPO/U.Porto**

Faculdade de Engenharia da Universidade do Porto (FEUP) Sexta-feira, 28 Junho 2013, 00h00

Notícia

score = 0.0

[Anterior](#)

Página 1 de 1

[Seguinte](#)

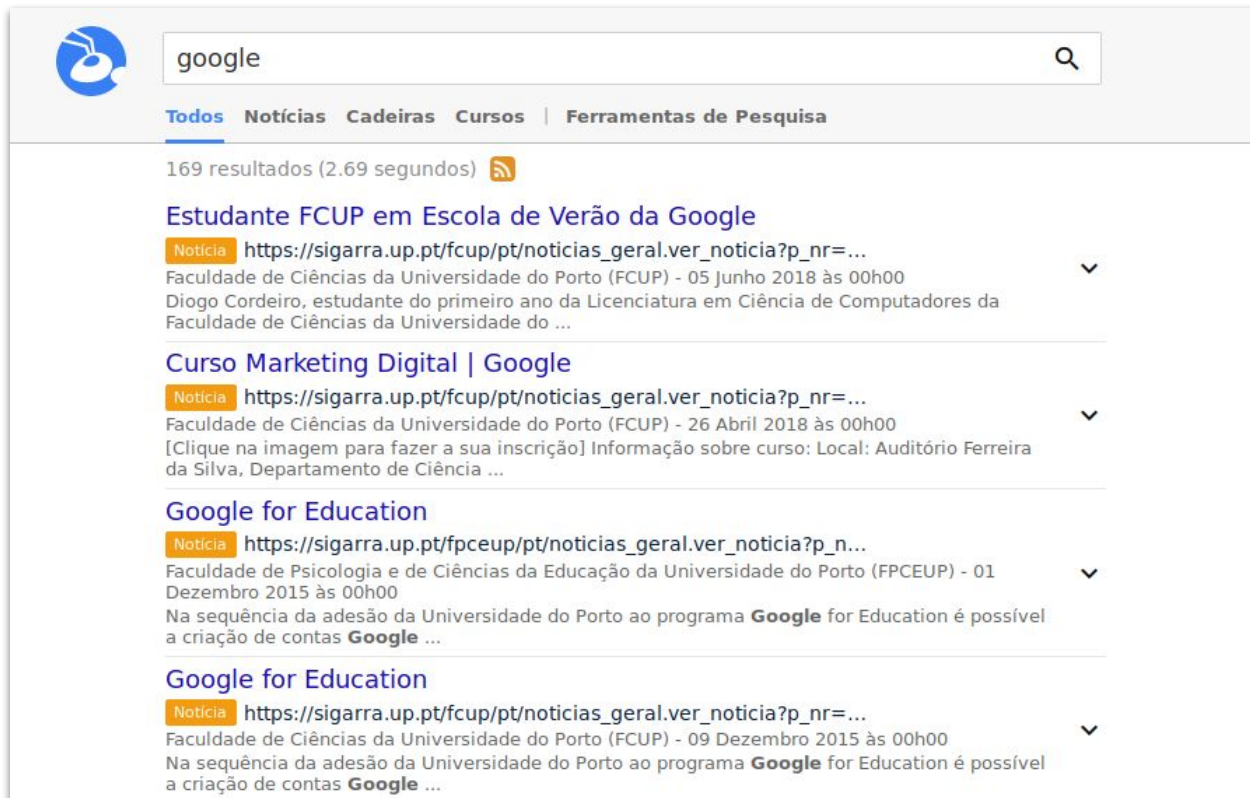
Entidades

**Tiago Nuno Mesquita Folgado leitão Devezas****José Luís da Silva Devezas**

Ligações

Faculdade: Faculdade de Engenharia, Faculdade de Engenharia da Universidade do Porto**Sala:** I123**Projetor** Não**Edifício** Electrotecnia (I)**Mapa** https://sigarra.up.pt/feup/pt/instal_geral?nei_mana?nv_id=77486[Ver mais](#)

Consulta relação. A intenção da consulta é descobrir como duas ou mais entidades ou tipos se relacionam.



The image shows a search engine interface with the word "google" in the search bar. Below the search bar, there are navigation tabs: "Todos", "Notícias", "Cadeiras", "Cursos", and "Ferramentas de Pesquisa". The search results are displayed below, showing 169 results in 2.69 seconds. The first four results are news items from the University of Porto (FCUP) website, each with a title, a URL, a date, and a brief description. The titles are "Estudante FCUP em Escola de Verão da Google", "Curso Marketing Digital | Google", "Google for Education", and "Google for Education".

google

Todos Notícias Cadeiras Cursos | Ferramentas de Pesquisa

169 resultados (2.69 segundos)

Estudante FCUP em Escola de Verão da Google
Notícia https://sigarra.up.pt/fcup/pt/noticias_geral.ver_noticia?p_nr=...
Faculdade de Ciências da Universidade do Porto (FCUP) - 05 Junho 2018 às 00h00
Diogo Cordeiro, estudante do primeiro ano da Licenciatura em Ciência de Computadores da Faculdade de Ciências da Universidade do ...

Curso Marketing Digital | Google
Notícia https://sigarra.up.pt/fcup/pt/noticias_geral.ver_noticia?p_nr=...
Faculdade de Ciências da Universidade do Porto (FCUP) - 26 Abril 2018 às 00h00
[Clique na imagem para fazer a sua inscrição] Informação sobre curso: Local: Auditório Ferreira da Silva, Departamento de Ciência ...

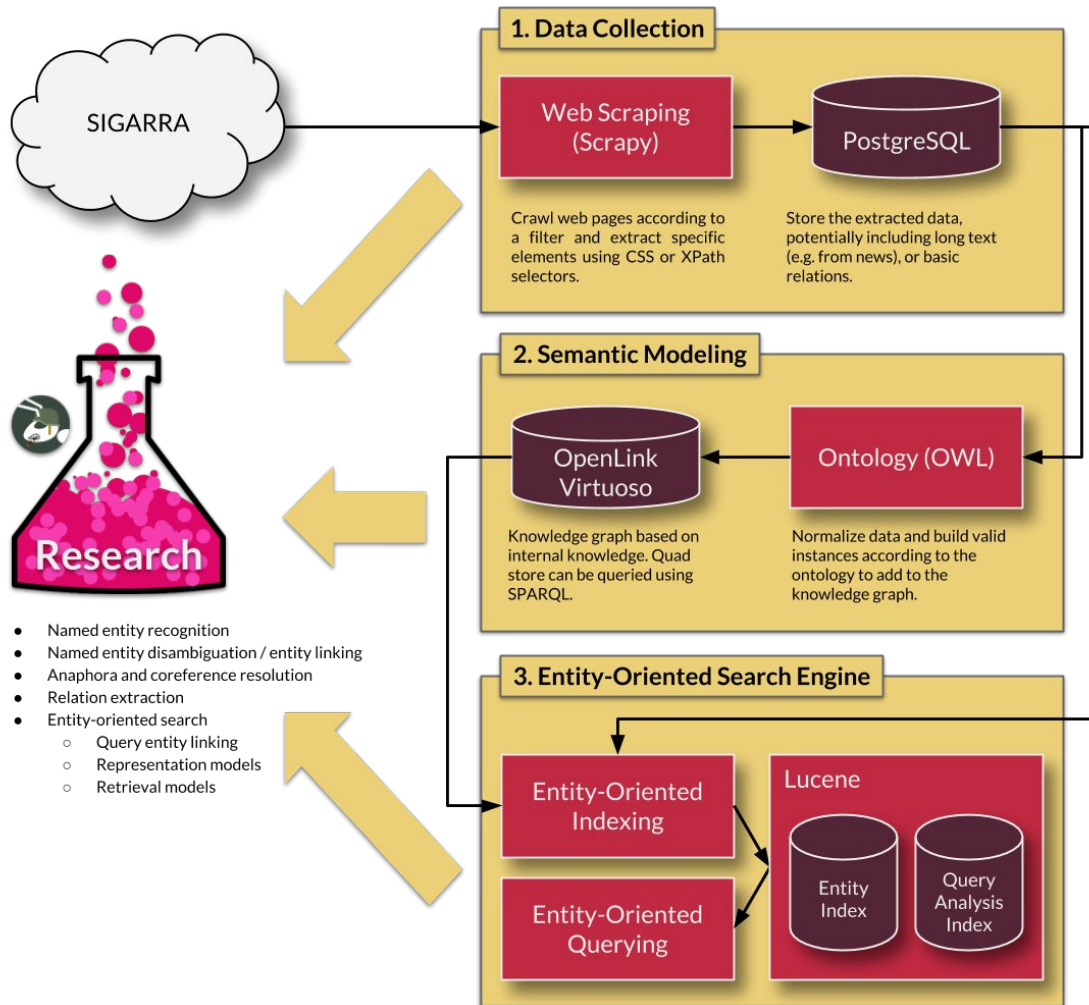
Google for Education
Notícia https://sigarra.up.pt/fpceup/pt/noticias_geral.ver_noticia?p_n...
Faculdade de Psicologia e de Ciências da Educação da Universidade do Porto (FPCEUP) - 01 Dezembro 2015 às 00h00
Na sequência da adesão da Universidade do Porto ao programa **Google** for Education é possível a criação de contas **Google** ...

Google for Education
Notícia https://sigarra.up.pt/fcup/pt/noticias_geral.ver_noticia?p_nr=...
Faculdade de Ciências da Universidade do Porto (FCUP) - 09 Dezembro 2015 às 00h00
Na sequência da adesão da Universidade do Porto ao programa **Google** for Education é possível a criação de contas **Google** ...

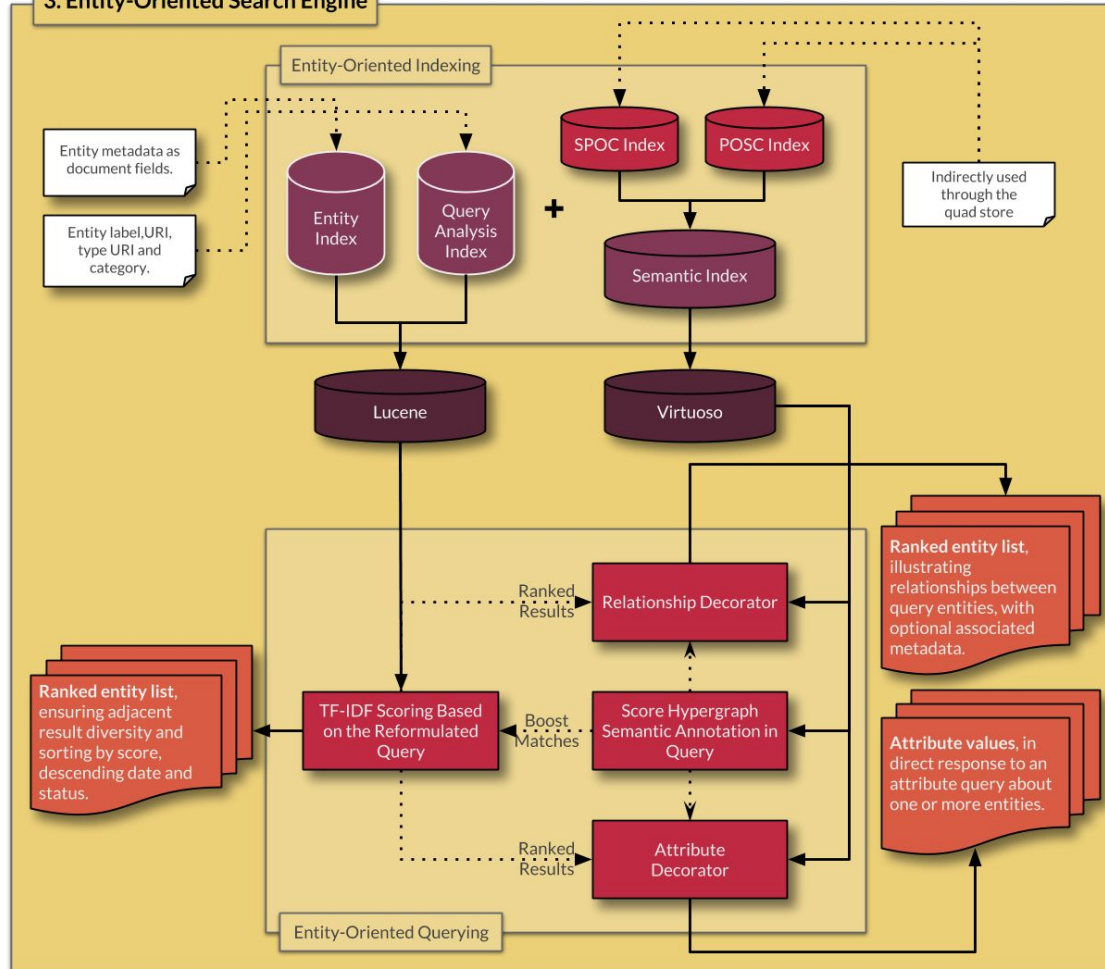
Consulta palavras-chave. A intenção da consulta é descrita por tudo aquilo que não cabe nas categorias anteriores.

Arquitetura do motor de busca

Componentes do ANT, da recolha à pesquisa.



3. Entity-Oriented Search Engine





API REST

- Os vários serviços relacionados com a pesquisa são fornecidos via API REST.
- A descrição da API é feita no formato OpenAPI 2.0 (Swagger).
 - <http://ant.fe.up.pt/api/>
 - <https://swagger.io/specification/>
- O que permite disponibilizar uma consola para a exploração da API.
 - <http://ant.fe.up.pt/api-console/>
- Os serviços suportados classificam-se como:
 - Analytics
 - Autocomplete
 - Decorators*
 - JavaScript
 - Log
 - Search*

* Serviços mais relevantes.

Compreensão da consulta

O que significa compreender e qual a abordagem do ANT?



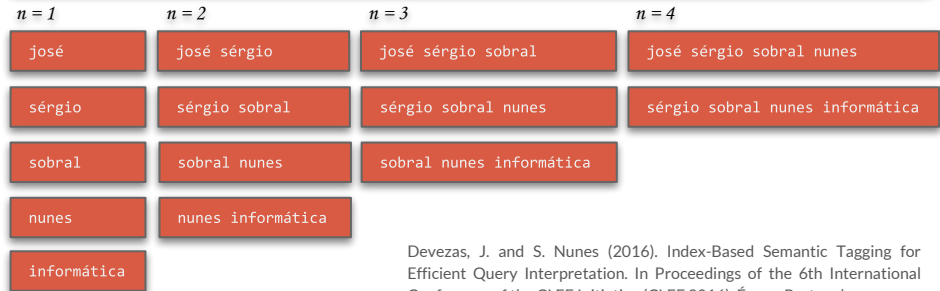
Compreensão da consulta

- Segmentar a consulta.
 - De acordo com o conhecimento no corpus indexado e na base de conhecimento, qual a divisão mais provável da consulta?
- Anotar com classes semânticas.
 - Para cada segmento, atribuir a classe semântica mais provável (ou nenhuma).
- Classificar a consulta.
 - Com base nos segmentos anotados, atribuir à consulta uma das cinco categorias de Pound et al. (2010).
- Vemos a consulta como um conjunto de palavras-chave.
- Mas poderíamos vê-la como linguagem natural:
 - Frases bem formadas;
 - Questões bem estruturadas.
- Os métodos de processamento da consulta podem variar com o tipo de consulta ou até de interface.

Como é que o ANT compreende as consultas?

- Segmentação da consulta baseada na correspondência com entidades, para todos os n -gramas da consulta até um valor máximo de n .
- Marcação semântica dos segmentos da consulta baseado na probabilidade de associar um dado tipo de entidade ao n -grama.

Query: josé sérgio sobral nunes informática



Devezas, J. and S. Nunes (2016). Index-Based Semantic Tagging for Efficient Query Interpretation. In Proceedings of the 6th International Conference of the CLEF Initiative (CLEF 2016), Évora, Portugal.



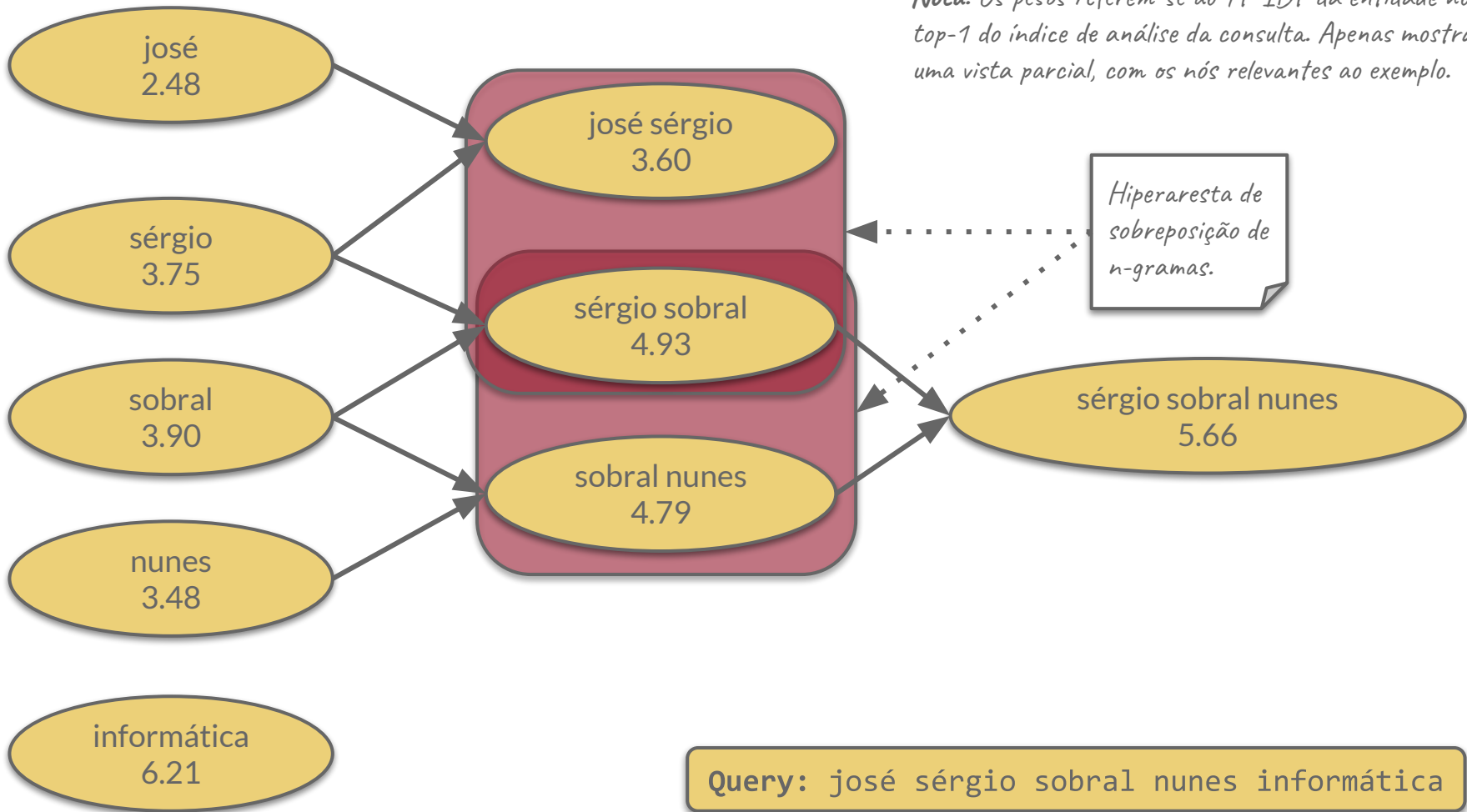
Como é que o ANT compreende as consultas?

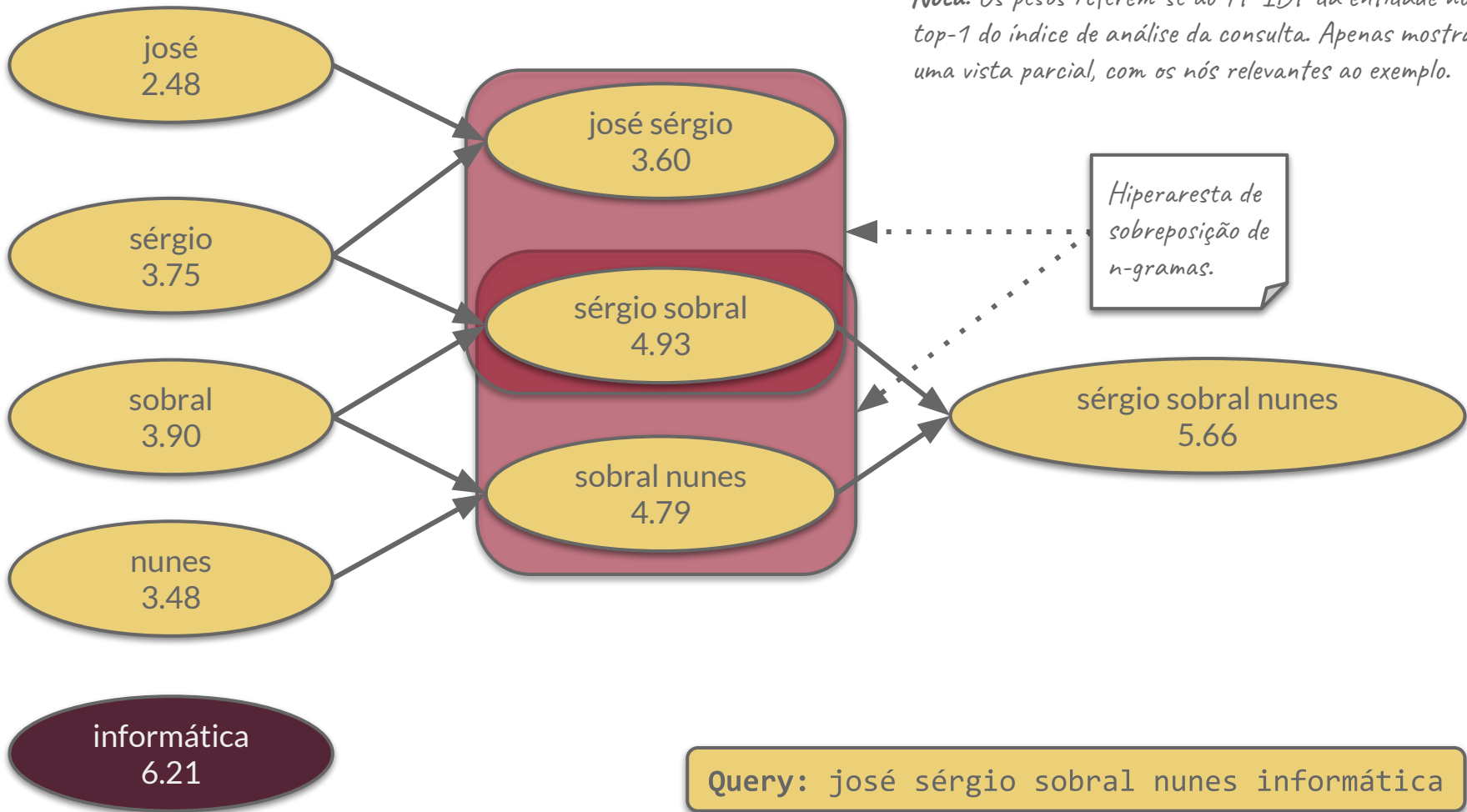
- Na realidade, o método que acabamos por utilizar é uma variação do método anterior, à qual chamamos “**Score Hypergraph**”.
- Usamos pontuações TF-IDF em vez de probabilidades.
- Temos um índice dedicado à análise da consulta, para pesquisar entidades que fazem correspondência com os n -gramas.
- Hipergrafo* de n -gramas para resolver sobreposições de segmentos da consulta e para corrigir bugs do método anterior.

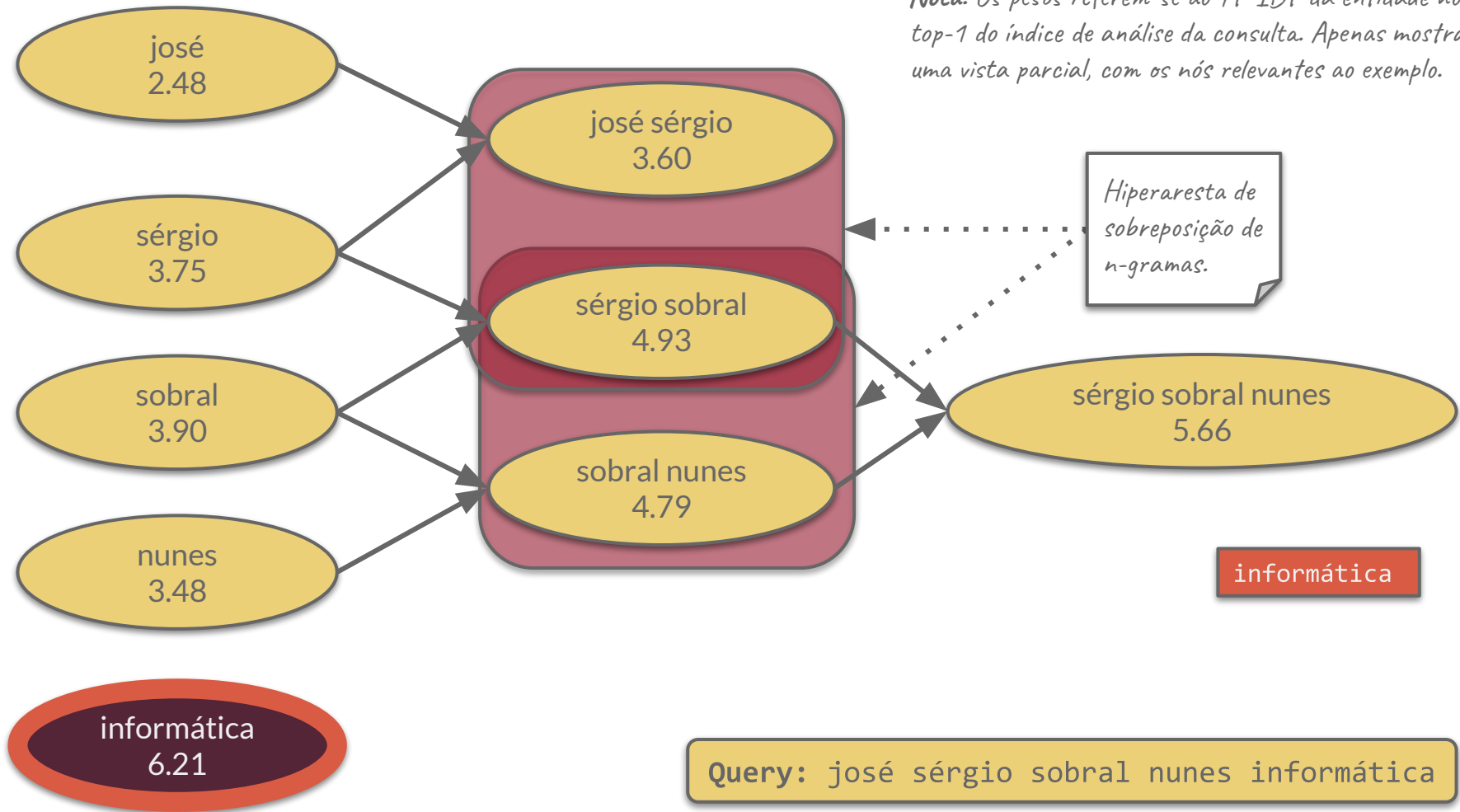
* Um hipergrafo é uma generalização de um grafo, onde as arestas podem ter um número arbitrário de nós.

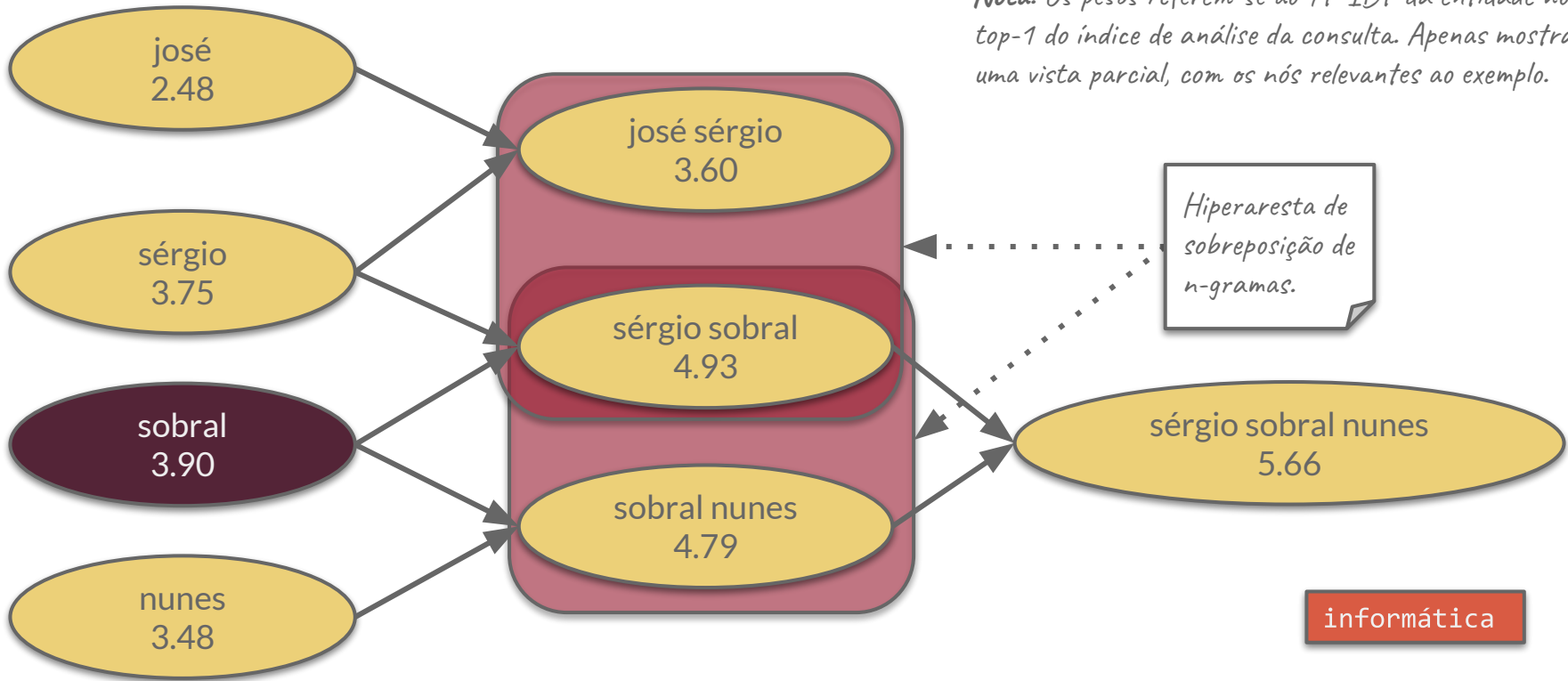
Score Hypergraph

Segmentação da consulta e marcação semântica no ANT.

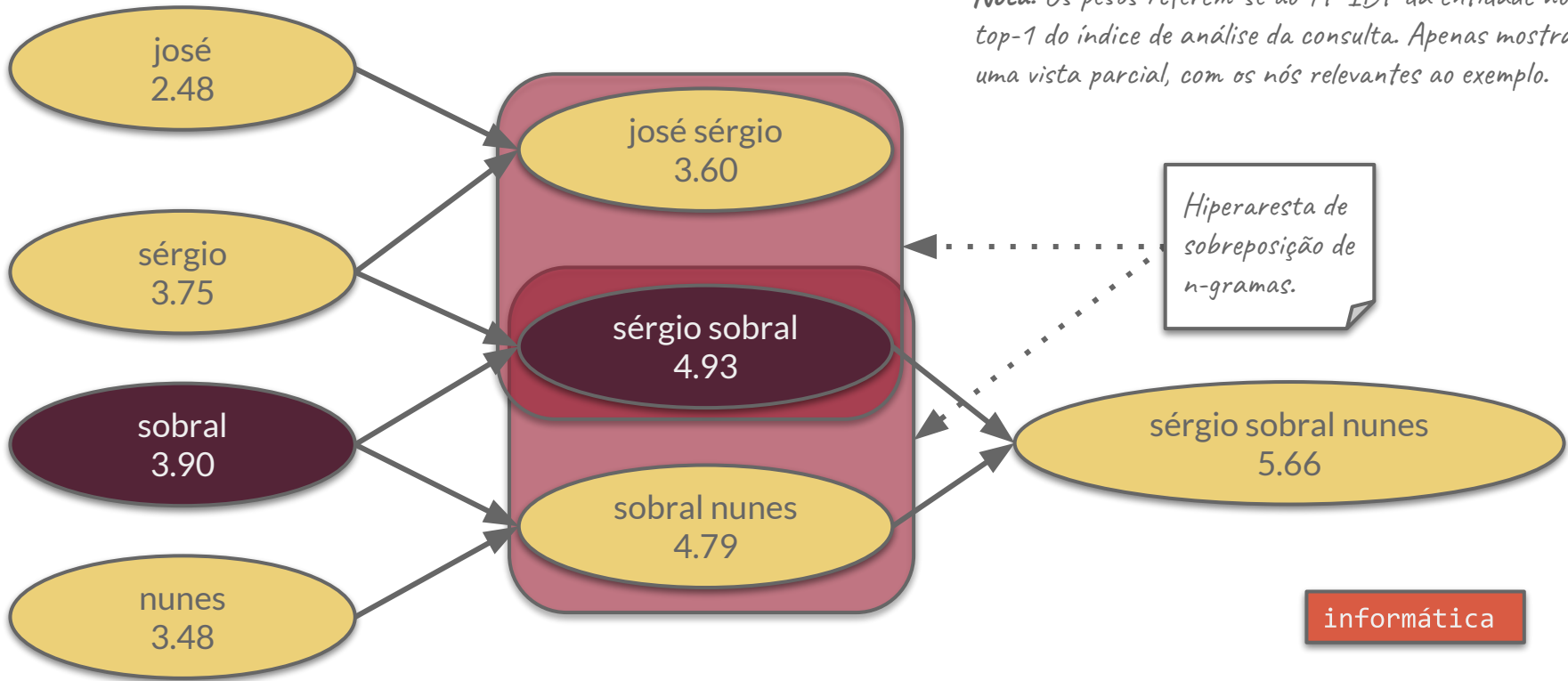




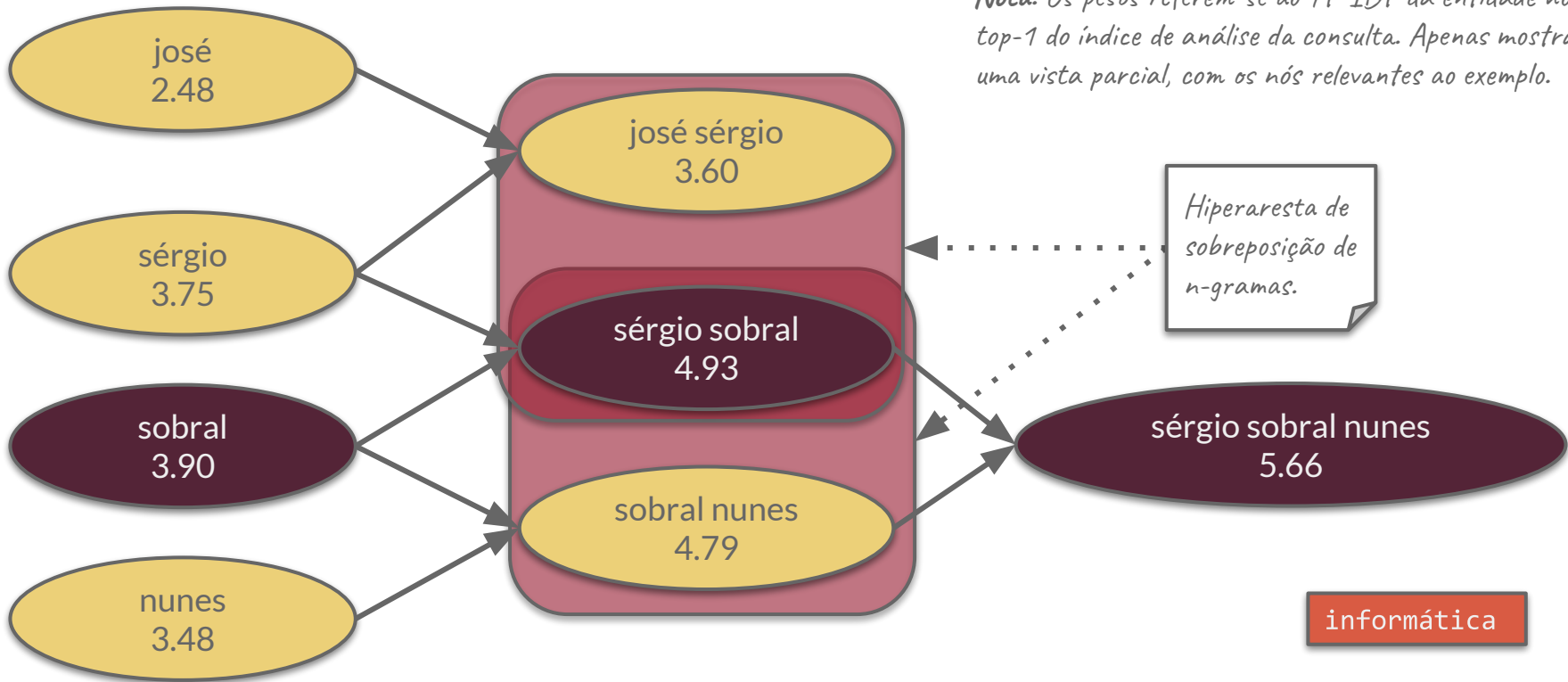




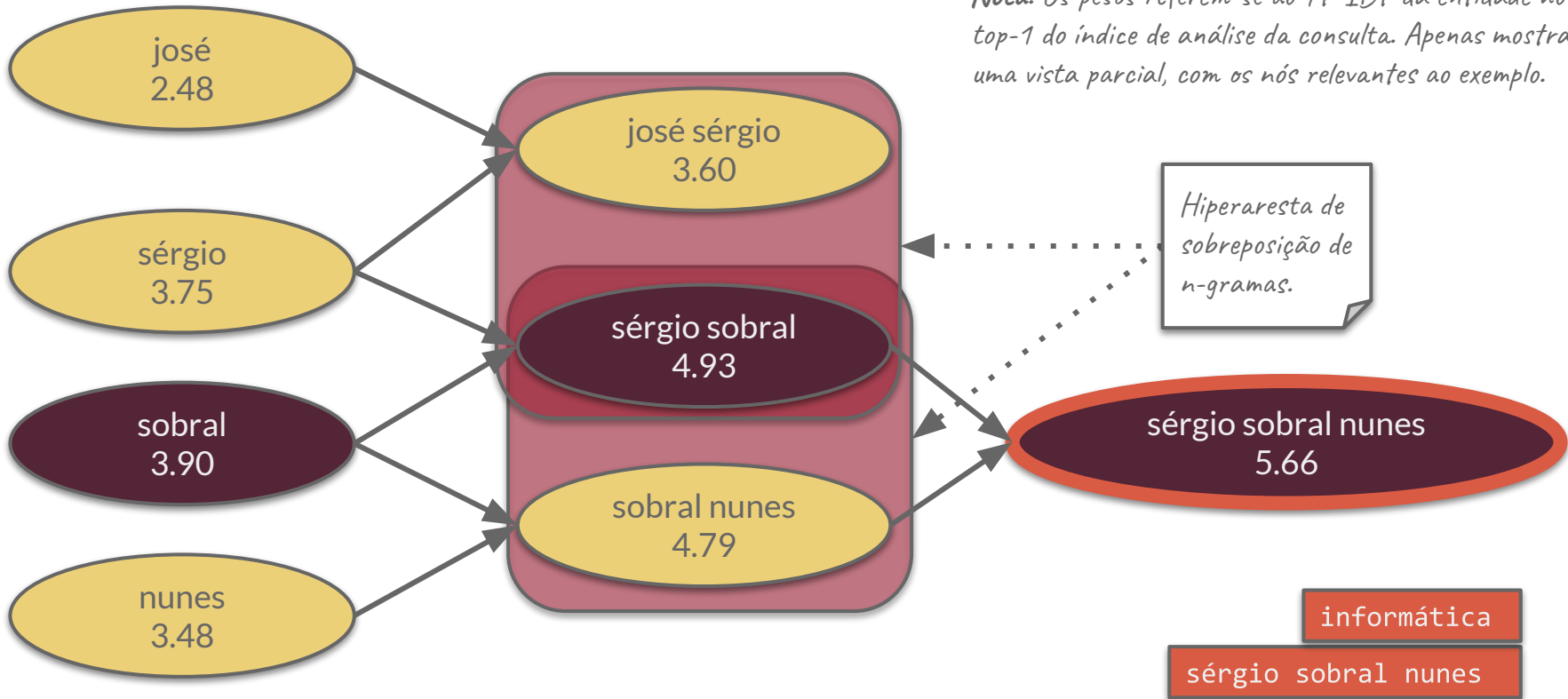
Query: jósé sérgio sobral nunes informática



Query: jósé sérgio sobral nunes informática



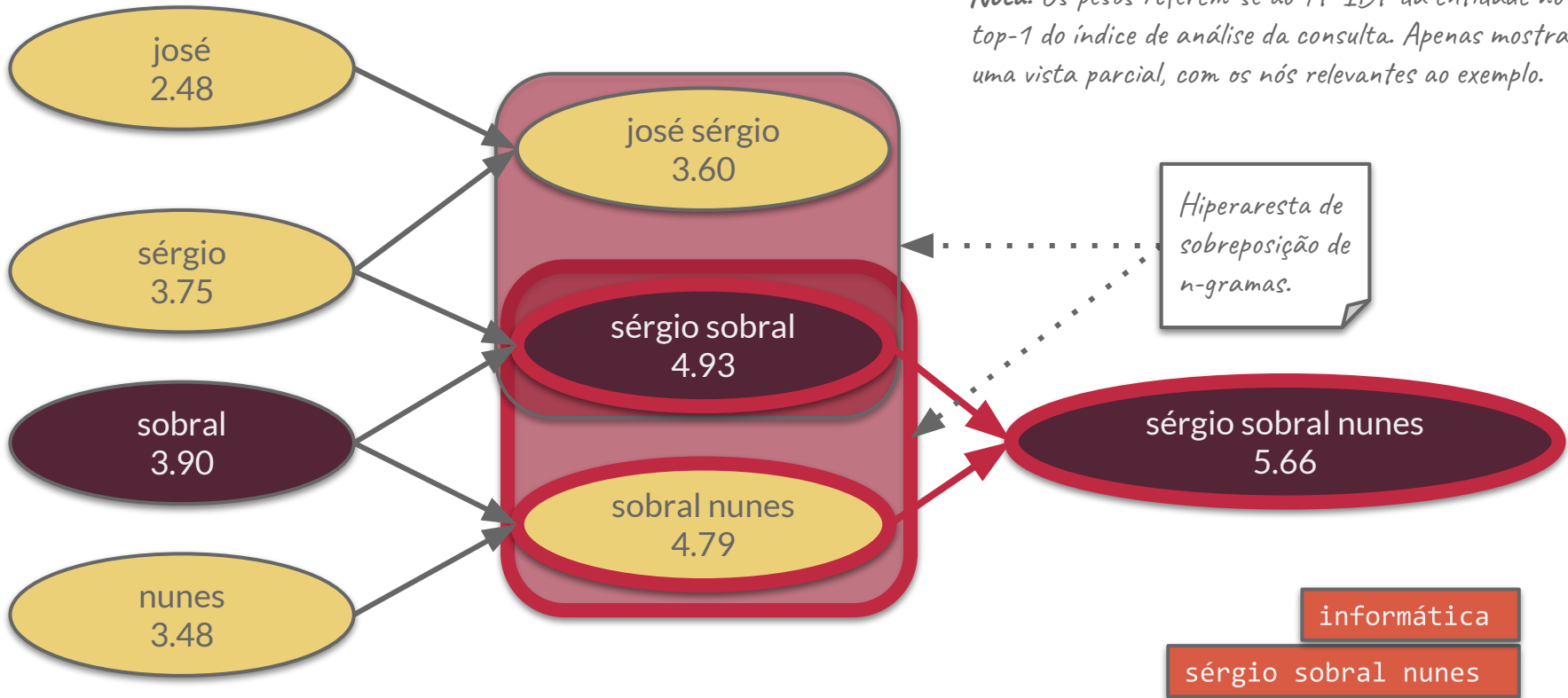
Query: jósé sér gio sobral Nunes informática



Nota: Os pesos referem-se ao TF-IDF da entidade no top-1 do índice de análise da consulta. Apenas mostramos uma vista parcial, com os nós relevantes ao exemplo.

Hiperaresta de sobreposiç3o de n-gramas.

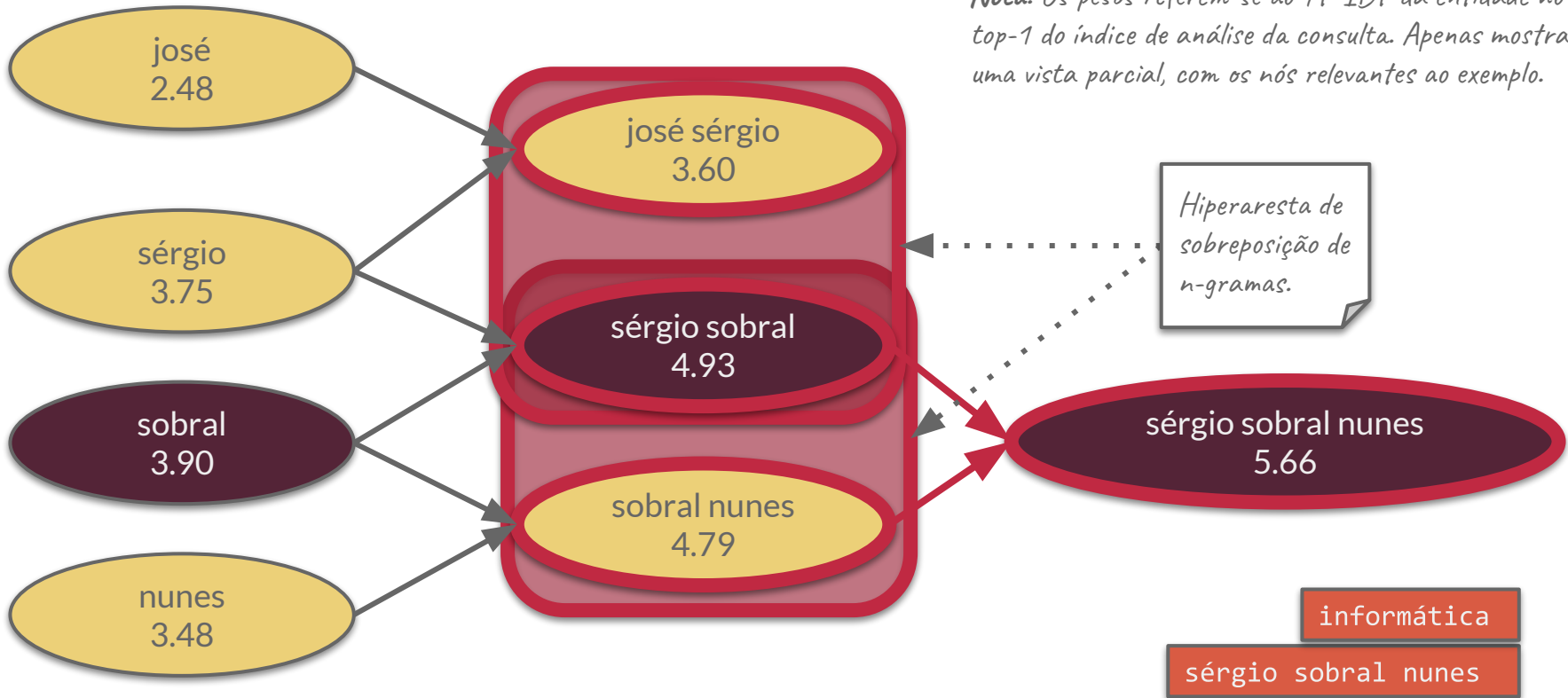
Query: jósé s3rgio sobral nunes inform3tica



Nota: Os pesos referem-se ao TF-IDF da entidade no top-1 do índice de análise da consulta. Apenas mostramos uma vista parcial, com os nós relevantes ao exemplo.

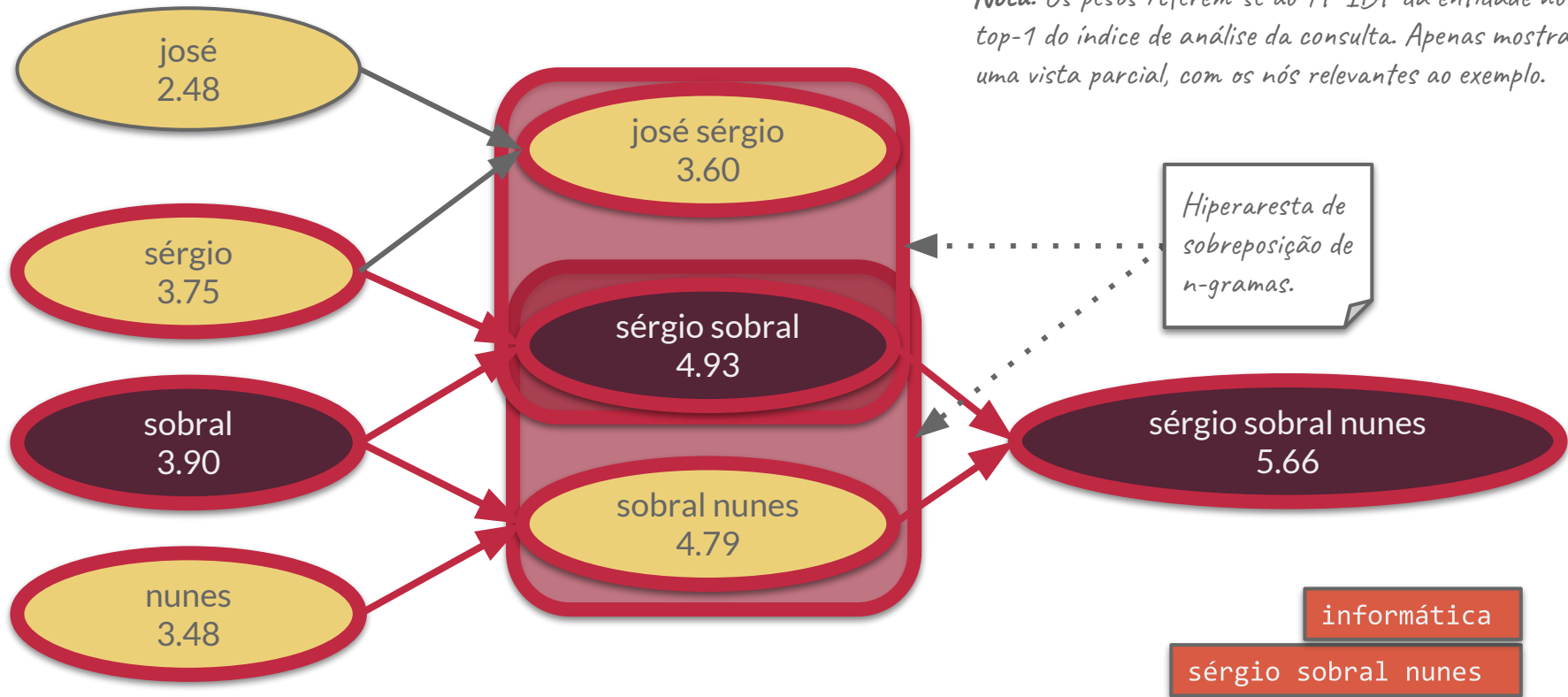
Hiperaresta de sobreposição de n-gramas.

Query: j osé s érgio s obral n unes informática



Nota: Os pesos referem-se ao TF-IDF da entidade no top-1 do índice de análise da consulta. Apenas mostramos uma vista parcial, com os nós relevantes ao exemplo.

Query: j osé s érgio s obral n unes informática



Query: osé sérgio sobral nunes informática

josé
2.48

Nota: Os pesos referem-se ao TF-IDF da entidade no top-1 do índice de análise da consulta. Apenas mostramos uma vista parcial, com os nós relevantes ao exemplo.

informática

sérgio sobral nunes

Query: josé sérgio sobral nunes informática

josé
2.48

Nota: Os pesos referem-se ao TF-IDF da entidade no top-1 do índice de análise da consulta. Apenas mostramos uma vista parcial, com os nós relevantes ao exemplo.

informática

sérgio sobral nunes

Query: josé sérgio sobral nunes informática

josé
2.48

Nota: Os pesos referem-se ao TF-IDF da entidade no top-1 do índice de análise da consulta. Apenas mostramos uma vista parcial, com os nós relevantes ao exemplo.

informática

sérgio sobral nunes

josé

Query: josé sérgio sobral nunes informática

informática

sérgio sobral nunes

josé

Query: josé sérgio sobral nunes informática

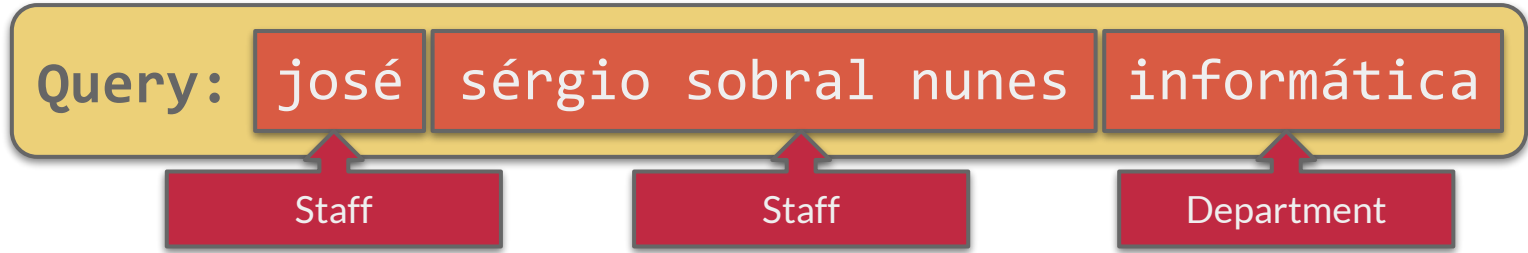
Query:

josé

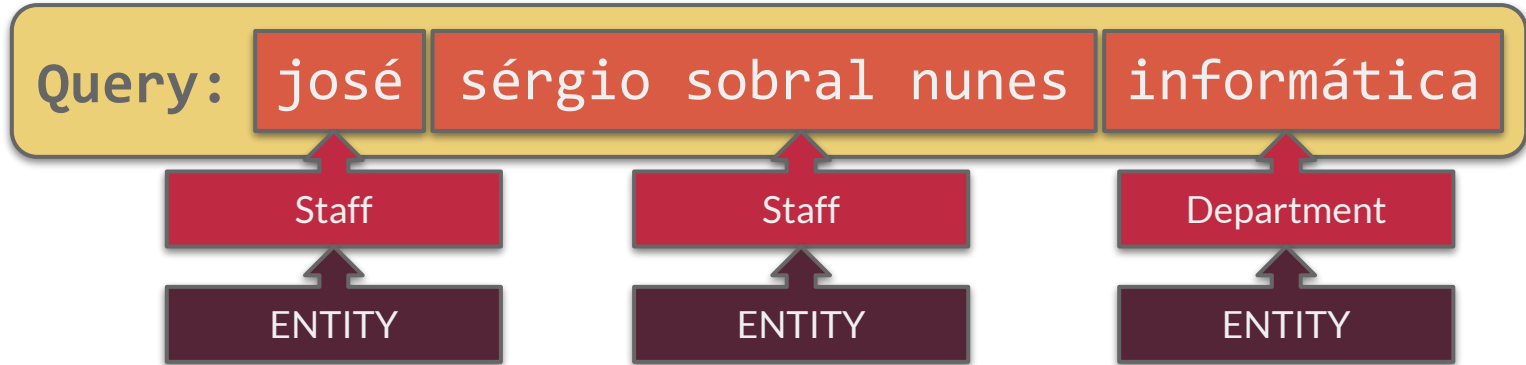
sérgio sobral nunes

informática

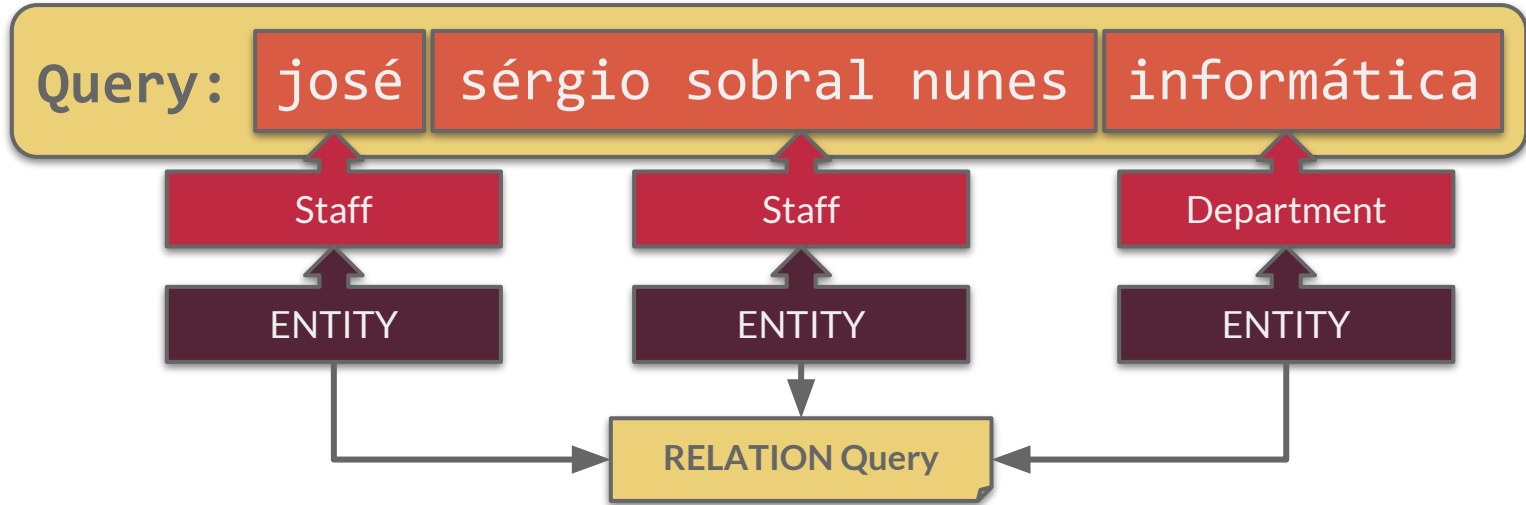
1. The query was segmented based on the n -grams with the highest-scoring entities.



2. The query was assigned semantic tags based on the type of the highest-scoring entity.



3. From the semantic tag, we directly derived a higher level tag that could either be ENTITY (e.g., instance of Staff class), ATTRIBUTE (e.g., property) or TYPE (e.g., Staff class).



4. Based on the combination of higher level tags, we conditionally obtained the query category.

Conclusões

Observações finais, projetos em curso, e uma visão do futuro.





Observações finais

- O motor de busca ANT está a servir a comunidade académica local, embora com algumas limitações de infraestrutura e recursos humanos (é um protótipo).
- Ao mesmo tempo, permite recolher feedback de relevância implícita, com base em cliques, para os resultados das consultas efetuadas.
- O ANT tem vindo a proporcionar a oportunidades de colaboração entre diversas áreas de investigação:
 - Web Design (colaboração com MM para o desenvolvimento do front-end);
 - Experiência de Utilizador (1 dissertação do MM em interfaces de pesquisa orientada a entidades);
 - Extração de Informação (1 dissertação do MIEIC em reconhecimento de entidades mencionadas em PT).



Projetos relacionados

Army ANT

- Serve as necessidades de investigação na área de pesquisa orientada a entidades.
- Permite explorar ideias inovadoras de pesquisa, fornecendo ferramentas de exploração e avaliação.

Tese de doutoramento

- “Graph-Based Entity-Oriented Search”
 - Representação conjunta de texto, entidades e suas relações.
 - Generalização das tarefas de pesquisa orientada a entidades.
 - Melhoramento da eficácia da pesquisa?
- Exploração de random walks em grafos e hipergrafos.

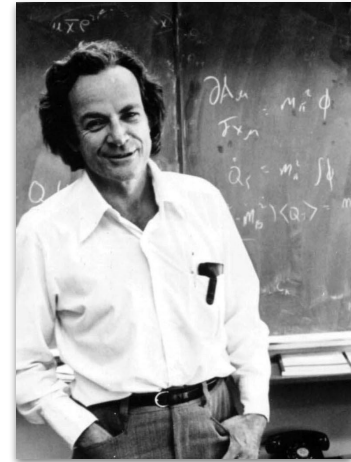


Uma visão do futuro...

...pode facilmente começar num passado semi-esquecido.



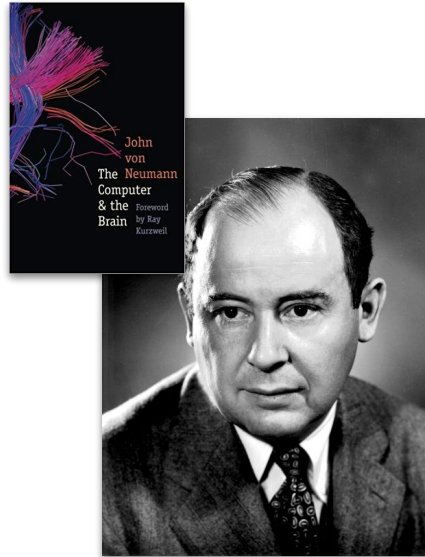
Von Neumann



Richard Feynman

Uma visão do futuro...

...pode facilmente começar num passado semi-esquecido.



Von Neumann

Antes da sua morte, em 1957, John Von Neumann estava a trabalhar nas aulas que iria lecionar nas Yale Silliman lectures. A sua ideia era comparar o cérebro e o computador, estudando os órgãos de ambos (neurónio/transístor).

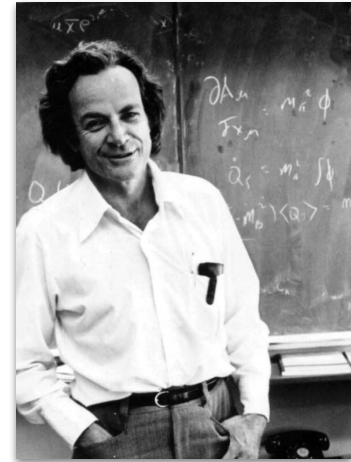
Hoje em dia, temos a Ciência Cognitiva que já une, entre outros, a Neurociência e a Inteligência Artificial. No entanto, ainda há caminhos por explorar nas intersecções. Há respostas e ideias perdidas no passado.

Uma visão do futuro...

...pode facilmente começar num passado semi-esquecido.

Físico teórico com contribuições relevantes para a Mecânica Quântica.

Durante a Segunda Guerra Mundial, administrou o grupo de computação humana na divisão teórica do Laboratório de Los Alamos. Computação humana não é mais do que resolver manualmente e em grupo um conjunto de cálculos longos e complexos, seguindo estritamente as regras estabelecidas.



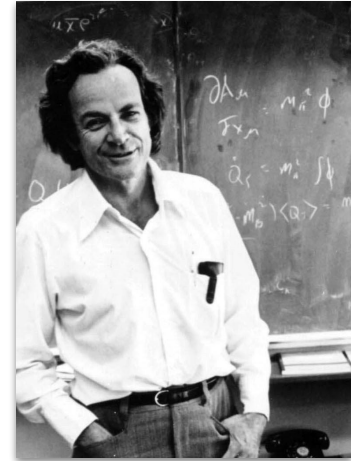
Richard Feynman

Uma visão do futuro...

...pode facilmente começar num passado semi-esquecido.

Talvez por isso, tenha colaborado com o Stanley Frankel e o Nicholas Metropolis, no desenvolvimento de um sistema para utilizar cartões perfurados da IBM para a computação.

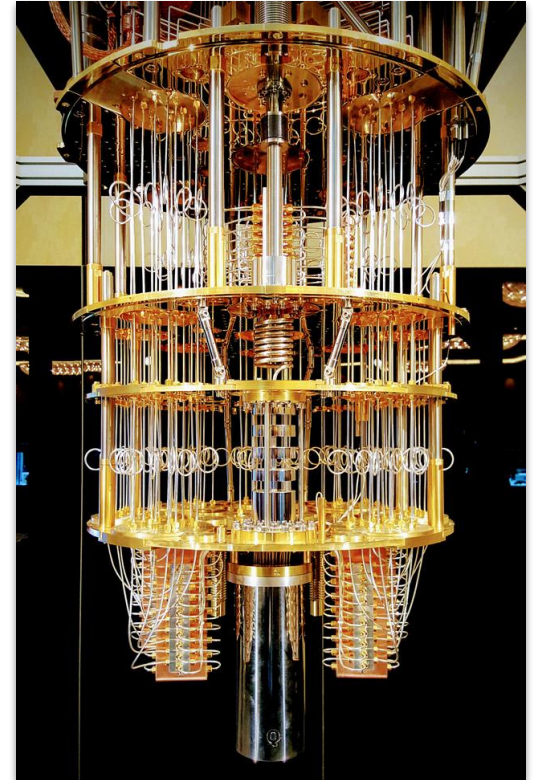
Inventou ainda um novo método de computar logaritmos que mais tarde utilizou na Connection Machine, uma arquitetura diferente da do Von Neumann, proposta pelo Daniel Hillis.



Richard Feynman

Uma visão do futuro...

...em direção aos motores de busca quânticos?

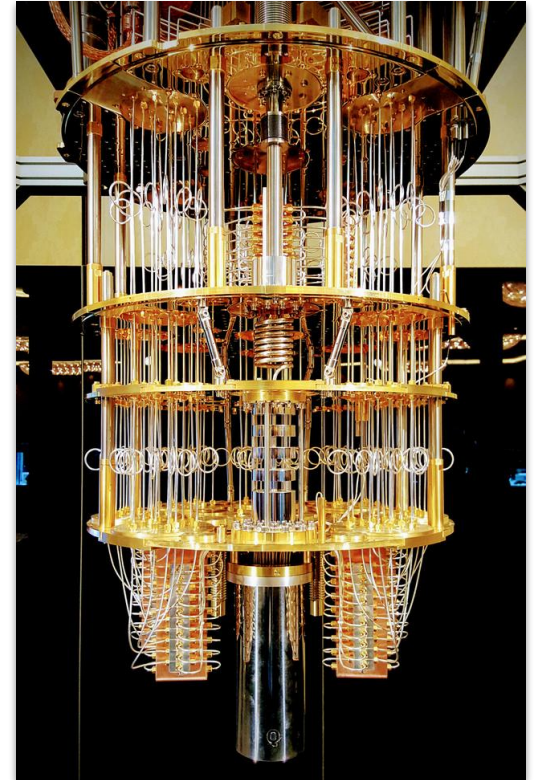


Uma visão do futuro...

...com muitas questões por responder.

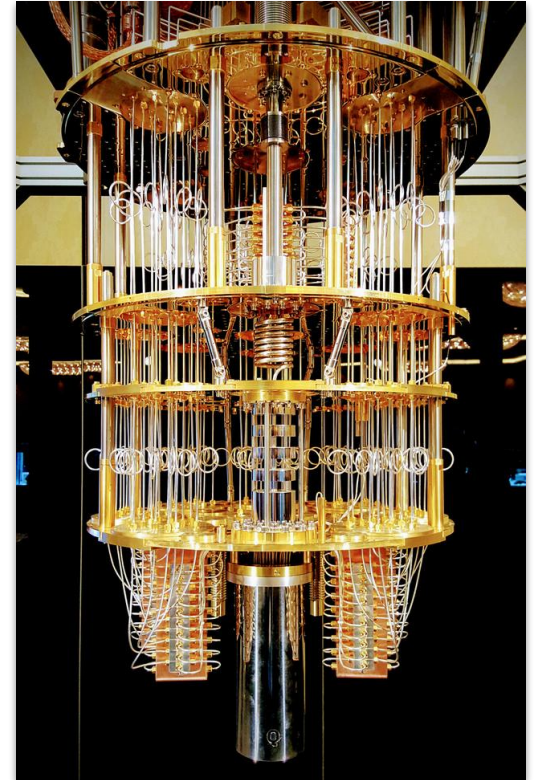
“Major Quantum Computing Advance Made Obsolete by Teenager”

<https://www.quantamagazine.org/teenager-finds-classical-alternative-to-quantum-recommendation-algorithm-20180731/>



Uma visão do futuro...

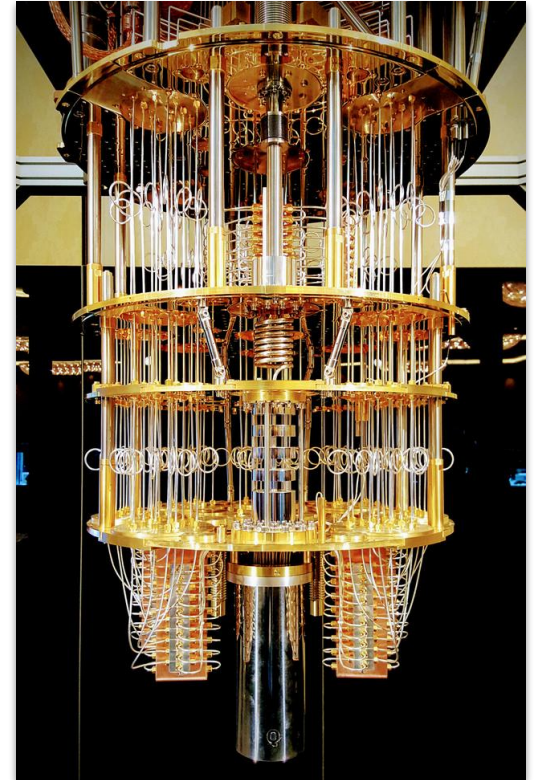
...onde o binário dá lugar à incerteza e às probabilidades.



Uma visão do futuro...

...onde a pesquisa baseada em incerteza se torna cada vez mais relevante.

No FEUP InfoLab, estamos a trabalhar com computadores clássicos, mas a explorar ideias que podem ter aplicações quânticas.



Uma visão do futuro...

...mais feliz, mais criativa e mais exploratória.

A altura de usar as ferramentas que a ciência nos deu é agora. Evoluir (também) significa integrar!



Obrigado!

<https://ant.fe.up.pt>

José Devezas is supported by research grant PD/BD/128160/2016, provided by the Portuguese national funding agency for science, research and technology, Fundação para a Ciência e a Tecnologia (FCT), within the scope of Operational Program Human Capital (POCH), supported by the European Social Fund and by national funds from MCTES.

