

USO DA WIKIPÉDIA NA INVESTIGAÇÃO EM INFORMÁTICA

Sérgio Nunes

Faculdade de Engenharia
Universidade do Porto

Academia Wikipédia
16 Abril 2010 | Porto, Portugal

SUMÁRIO

1. Investigação sobre/com a Wikipédia.
2. Tópicos e áreas de investigação.
3. Projecto WikiChanges.
4. Recursos e ferramentas existentes.

INVESTIGAÇÃO / WIKIPÉDIA

ESTUDOS SOBRE A WIKIPÉDIA

Estudos **focados na compreensão da Wikipédia**, agrupados em dois tipos: (1) produção e qualidade do conteúdo, (2) aspectos sociais (ex. uso, edição).

ESTUDOS COM BASE NA WIKIPÉDIA

Estudos que **fazem uso dos recursos disponibilizados** para resolver problemas gerais, não limitados à Wikipédia. Alguns exemplos: classificação de textos, sumariação, tradução automática, etc.

QUALIDADES DA WIKIPÉDIA (1)

DIMENSÃO E DIVERSIDADE

Mais de 15 milhões de artigos em mais de 250 idiomas (3,2 em Inglês; 0,5 em Português). Mais de 270 mil sub-categorias organizadas sob 11 categorias principais.

QUALIDADE

Grande diversidade na qualidade dos artigos. Pode ser estimada com base no número e conteúdo das revisões, e nas categorias a que um artigo pertence.

QUALIDADES DA WIKIPÉDIA (2)

IMPORTÂNCIA E IMPACTO

É um fenómeno de importância mundial, justificando, por si só, a investigação. Os dados disponíveis permitem não só estudos ao nível do conteúdo dos artigos mas também ao nível das interacções humanas associadas às revisões.

ACESSIBILIDADE

Fácil acesso aos dados através de arquivos disponibilizados pela Wikipédia ou terceiros (ex. DBpedia). Acesso programático a dados e funções com a API do MediaWiki.

QUALIDADES DA WIKIPÉDIA (3)

REPRODUTIBILIDADE

A reprodutibilidade é um aspecto crucial do método científico. O uso da Wikipédia permite a repetição das experiências por outras equipas e a comparação dos resultados.

ÁREAS DE INVESTIGAÇÃO

PROCESSAMENTO DE LINGUAGEM NATURAL

Uso da Wikipédia para resolver tarefas associadas ao processamento automático de linguagem natural: proximidade semântica de termos, desambiguação semântica, ...

RECUPERAÇÃO DE INFORMAÇÃO

Uso da Wikipédia para melhorar a recuperação de informação noutros contextos: expansão de interrogações, sumariação de documentos, resposta a perguntas, categorização, ...

EXTRACÇÃO DE INFORMAÇÃO

Uso da Wikipédia para extracção de informação estruturada a partir de dados não estruturados: identificação de entidades mencionadas, ...

ALGUNS EXEMPLOS

TRADUÇÃO AUTOMÁTICA

A existência de **coleções paralelas** em diferentes idiomas torna possível desenvolver sistemas de tradução automática.

The screenshot shows the Portuguese Wikipedia page for "Maracujá". The article describes the fruit, its botanical classification, and its uses. It includes a photograph of the fruit and a list of curiosities. The text is in Portuguese and provides detailed information about the plant and its fruit.

Maracujá

The screenshot shows the Japanese Wikipedia page for "パッションフルーツ". The article provides information about the fruit, including its botanical classification and its uses. It includes a photograph of the fruit and a list of curiosities. The text is in Japanese and provides detailed information about the plant and its fruit.

パッションフルーツ

DESAMBIGUAÇÃO SEMÂNTICA AUTOMÁTICA

As páginas de **redirecionamento** e **desambiguação** da Wikipédia são recursos importantes na resolução automática de problemas de **polissemia** e **sinonímia**.

The screenshot shows the Wikipedia page for "Amsterdam (disambiguation)". The page title is "Amsterdam (disambiguation)" and it is categorized as a disambiguation page. The main text states: "Amsterdam is the largest city in, and titular capital of the Kingdom of the Netherlands. Amsterdam may also refer to:" followed by a list of geographical locations. The list includes: "New Amsterdam (disambiguation) or New Amsterdam" (with a note to look up Amsterdam in Wiktionary, the free dictionary); "In the United States:" with sub-items for Amsterdam, California; Amsterdam, Missouri; Amsterdam, New York; Amsterdam (town), New York; Amsterdam (train station); Amsterdam (Ohio); Amsterdam, Texas; and Amsterdam Avenue or Tenth Avenue (Manhattan). It also lists "Amsterdam, Mpumalanga" in South Africa; "Amsterdam, Saskatchewan" in Canada; "Amsterdamsøya", a Norwegian island; "Amsterdam, a French island in the Indian Ocean"; and "Tongatapu, an island of Tonga, once named Amsterdam". There are also sections for "Ships:" listing the MS Amsterdam, the Dutch East India Company ship (1682), and the USS Amsterdam (LST-1151).

The screenshot shows the Wikipedia page for "Marco". The page title is "Marco" and it is categorized as a disambiguation page. The main text states: "Esta é uma página de desambiguação, a qual lista artigos associados a um mesmo título. Se uma ligação interna a consulta não está sugerida que a consulte para obter o detalhamento do artigo adequado." Below this, there is a section "Outros artigos pode estar em os seguintes artigos de Wikipédia:" followed by a list of categories: "Localidades" (Marco (paróquia), 1 Pessoas, 2 Localidades, 3 Eventos locais, 4 Outros), "Pessoas" (Marco André, Marco Polo), "Localidades" (Marco (município brasileiro do estado do Ceará), Marco de Conaravão - concelho de Portugal, Pa do Marco - um nome antigo de Pa do Conro (Japão), Marco - povoação ruínas entre Alentejo e Extremadura, Marco - Museu de Arte Contemporânea, localizada em Campo Grande - MG), and "Moedas" (Marco andaluz, Marco Sforza).

EXTRACÇÃO DE INFORMAÇÃO

A Wikipédia é rica em **elementos estruturados** que possibilitam a prospecção automática de conceitos, relações e factos.

- ▶ Artigos
- ▶ Páginas de desambiguação
- ▶ Redireccionamentos
- ▶ Hiperligações *internas e externas*
- ▶ Estrutura de categorias
- ▶ *Predefinições e Infocaixas*
- ▶ Páginas de discussão
- ▶ Historial de revisões

EXEMPLO: *infocaixas*

As predefinições existentes nas páginas da Wikipédia, como as *infocaixas*, permitem a identificação e extracção automática de factos.

The screenshot shows the Wikipedia article for the UEFA Euro 2004. The article text is partially visible, describing the tournament as the twelfth European Football Championship, held in Portugal from June 12 to July 4, 2004. It mentions that Spain and Austria-Hungary were the other bidding nations for the hosting of the event. The article also notes that sixteen teams contested the final tournament after going through a qualification round which began in 2002. The tournament took place in ten venues located in eight cities — Avintes, Braga, Coimbra, Guimarães, Faro/Loulé, Leiria, Porto and Lisbon.

During the tournament there were several surprises: Germany, Italy and Spain were knocked out during the group stage; the title-holders France were eliminated in the quarter-finals by unranked Greece, and the Portuguese hosts recovered from their opening defeat to reach the final, eliminating Spain, England and Netherlands along the way. For the first time, the final featured the same teams as the opening match, with the hosts losing both of them also for the first time, as Portugal were beaten by Greece on both occasions. Greece's triumph was even more outstanding considering that they had only qualified for two other major tournaments, in 1980 and 1994 and their win in the opening match in 2004 was the first time they had ever won a game in a major tournament.

During the opening ceremony, the Portuguese portrayed a ship, symbolizing the voyages of the Portuguese explorers, sailing through a sea which gave place to the flags of all competing countries. It such was the enthusiasm that overtook the Greek fans that the ship became the symbol of the Greek victory, as Greece chartered for the "Private Ship" (εμπειρικό). Also, Portuguese-Canadian pop singer Nelly Furtado performed her single "Pierce", which represents Portuguese culture.

Comments (hide)

- 1 Qualifying
- 2 Teams
- 3 Venues
- 4 Match officials
- 5 Medal
- 6 Awards

UEFA Euro 2004

Campeonato de Europa de Futebol 2004

UEFA Euro 2004 official logo

Tournament details

Host country	 Portugal
Dates	12 June – 4 July
Teams	16
Venue(s)	10 (in 8 host cities)

Final positions

Champions	 Greece (1st title)
Runner-up	 Portugal

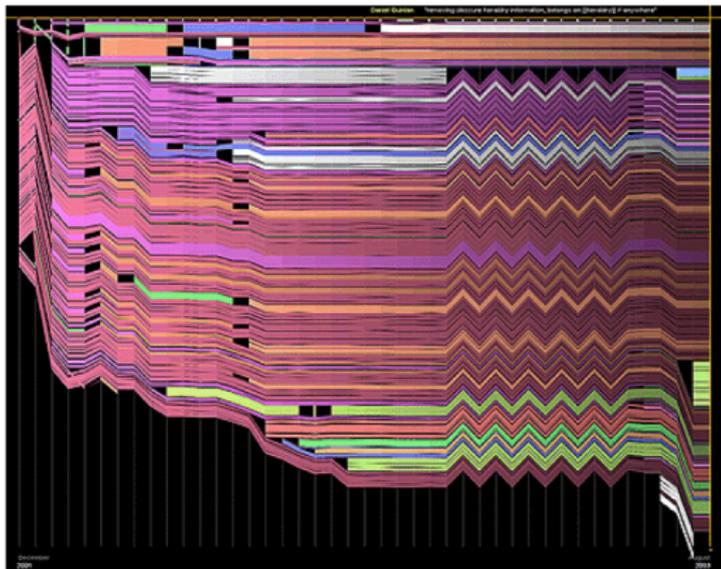
Tournament statistics

Matches played	31
Goals scored	77 (2.48 per match)
Attendance	1,156,473 (37,306 per match)
Top scorer(s)	 Milan Baroš (5 goals)
Best player	 Theodoros Zagorakis

← 2000 2008 →

VISUALIZAÇÃO DE INFORMAÇÃO

Técnicas de visualização têm sido exploradas para compreender a dinâmica interna da Wikipédia.



Viégas, F., Wattenberg, M., Dave, K. (2004) – *Studying Cooperation and Conflict between Authors with history flow Visualizations*. ACM CHI'04.

PROJECTO WIKICHANGES

PROJECTO WIKICHANGES

IDEIA BASE

Explorar as propriedades temporais para melhorar a recuperação de informação em colecções (ex. pesquisa).

QUESTÕES INICIAIS

- ▶ Como evolui o conteúdo de um documento?
- ▶ As edições a um documento seguem um padrão *neutro*?
- ▶ É possível extrair informação do perfil de edições?

HISTORIAL DE EDIÇÕES A UM ARTIGO



WIKIPÉDIA
The Free Encyclopedia

navigation

- [Main page](#)
- [Contents](#)
- [Featured content](#)
- [Current events](#)
- [Random article](#)

search

interaction

- [About Wikipedia](#)
- [Community portal](#)
- [Recent changes](#)
- [Contact Wikipedia](#)
- [Donate to Wikipedia](#)
- [Help](#)

toolbox

- [What links here](#)
- [Related changes](#)
- [RSS](#) [Atom](#)
- [Upload file](#)
- [Special pages](#)

[article](#) [discussion](#) [edit this page](#) [history](#)

[Try Beta](#) [Log in](#) / [create account](#)

Revision history of Porto

From Wikipedia, the free encyclopedia

[View logs for this page](#)

Browse history

From year (and earlier): From month (and earlier): all Tag filter:

For any version listed below, click on its date to view it. For more help, see [Help:Page history](#) and [Help:Edit summary](#).

External tools: [Revision history statistics](#) [Revision history search](#) [Number of watchers](#) [Page view statistics](#)

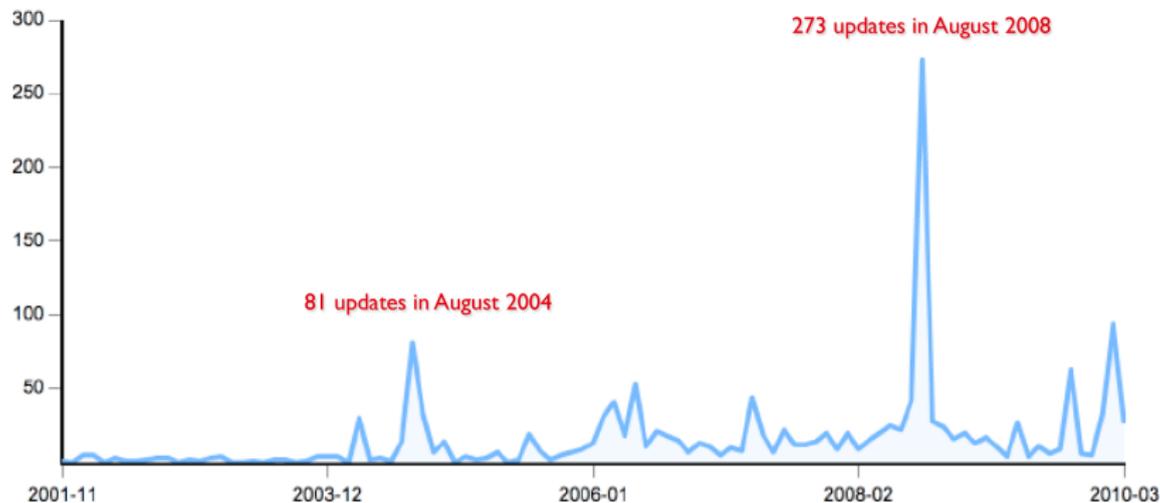
(cur) = difference from current version, (prev) = difference from preceding version, m = minor edit, → = section edit, ↔ = automatic edit summary

(latest | **earliest**) [View](#) ([newer 50](#) | [older 50](#)) ([20](#) | [50](#) | [100](#) | [250](#) | [500](#))

Compare selected revisions

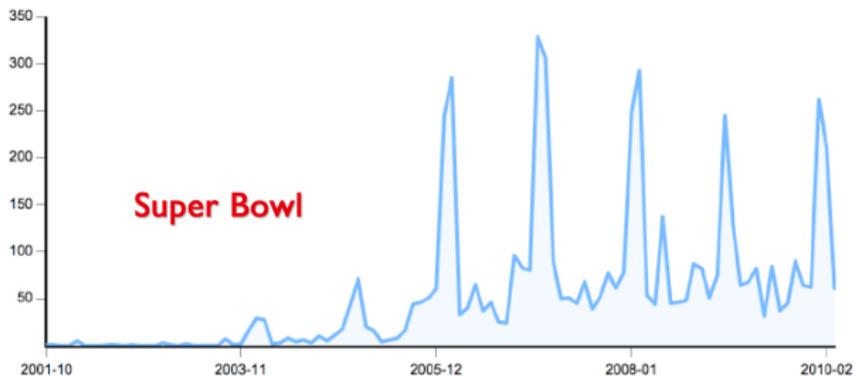
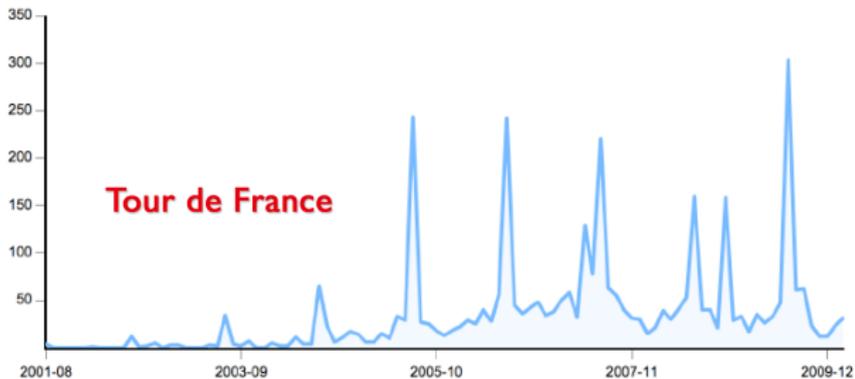
- (cur) (prev) 01:14, 29 March 2010 [Wilson44691](#) (talk | contribs) [m](#) (47,401 bytes) (*Reverted edits by 76.24.236.147 (talk) to last version by The Ogre*) (undo)
- (cur) (prev) 01:13, 29 March 2010 [76.24.236.147](#) (talk) (46,789 bytes) (↔*Public health*) (undo) (*Tag: references removed*)
- (cur) (prev) 14:06, 21 March 2010 [The Ogre](#) (talk | contribs) (47,401 bytes) (↔*Famous inhabitants: rm red links*) (undo)
- (cur) (prev) 00:46, 21 March 2010 [151.54.240.236](#) (talk) (47,469 bytes) (↔*Famous inhabitants*) (undo)
- (cur) (prev) 03:20, 14 March 2010 [Marek69](#) (talk | contribs) (47,436 bytes) (↔*Roads and bridges: clean up and general fixes using AWB*) (undo)
- (cur) (prev) 01:22, 13 March 2010 [Etan J. Tal](#) (talk | contribs) (47,437 bytes) (*additional photo*) (undo)
- (cur) (prev) 04:22, 10 March 2010 [Marek69](#) (talk | contribs) (47,354 bytes) (↔*Entertainment: clean up and general fixes using AWB*) (undo)
- (cur) (prev) 22:24, 1 March 2010 [79.112.88.161](#) (talk) (47,352 bytes) (↔*Sports*) (undo)
- (cur) (prev) 22:23, 1 March 2010 [79.112.88.161](#) (talk) (47,344 bytes) (↔*Sports*) (undo)
- (cur) (prev) 22:20, 1 March 2010 [79.112.88.161](#) (talk) (47,350 bytes) (↔*Sports*) (undo)
- (cur) (prev) 16:03, 28 February 2010 [The Ogre](#) (talk | contribs) [m](#) (47,239 bytes) (↔*Arts: cped*) (undo)
- (cur) (prev) 14:31, 28 February 2010 [Wonder64](#) (talk | contribs) (47,234 bytes) (↔*Culture*) (undo)
- (cur) (prev) 00:14, 28 February 2010 [FrescoBot](#) (talk | contribs) [m](#) (47,186 bytes) (*Bot: links syntax*) (undo)
- (cur) (prev) 18:15, 26 February 2010 [Wonder64](#) (talk | contribs) (47,221 bytes) (↔*Culture: right-adjusted image to clean up design*) (undo)
- (cur) (prev) 18:12, 26 February 2010 [Wonder64](#) (talk | contribs) (47,226 bytes) (↔*Arts: added image of She Changes*) (undo)
- (cur) (prev) 13:58, 19 February 2010 [XPTO](#) (talk | contribs) (47,102 bytes) (*rv*) (undo)
- (cur) (prev) 10:19, 19 February 2010 [89.151.115.162](#) (talk) (47,540 bytes) (↔*Parishes*) (undo)

PERFIL DE EDIÇÕES AO LONGO DO TEMPO

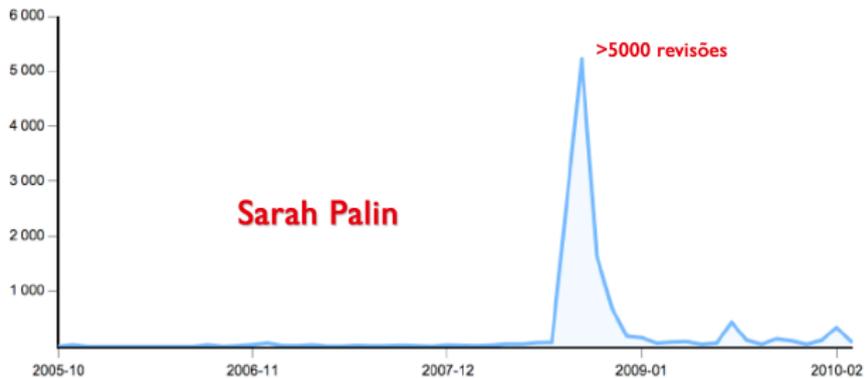


Número de edições mensais ao artigo **Summer Olympic Games**.

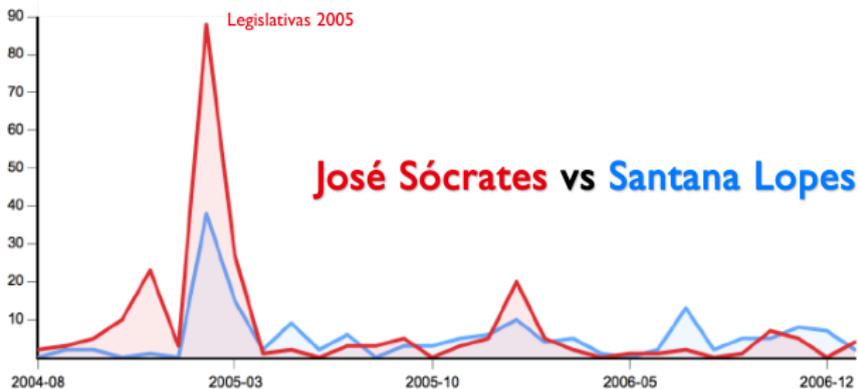
ACONTECIMENTOS RECORRENTES



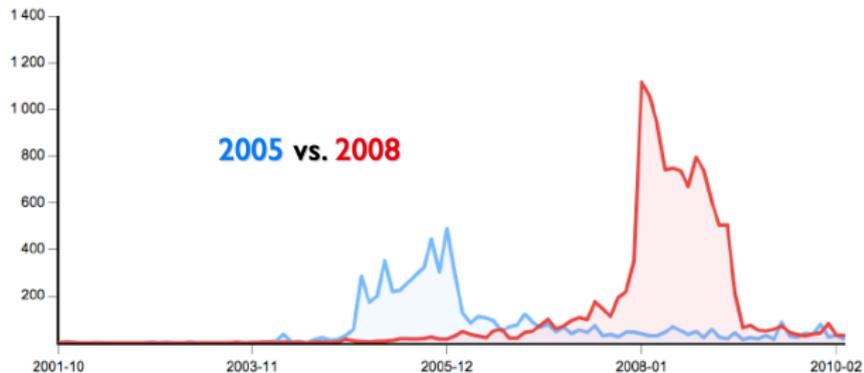
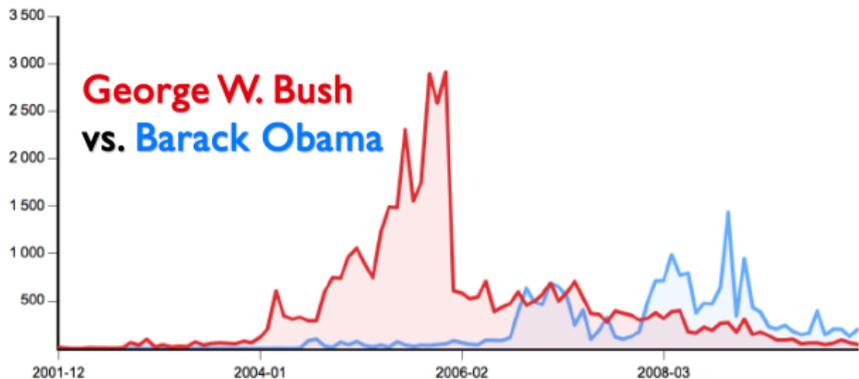
ACONTECIMENTOS INESPERADOS



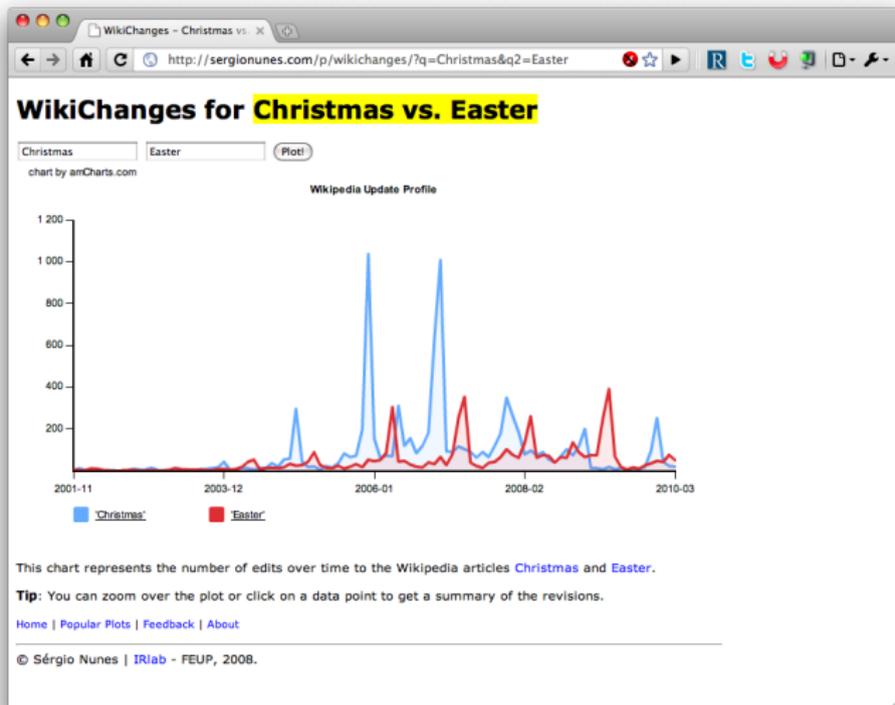
ACONTECIMENTOS REGIONAIS



COMPARAÇÕES LADO A LADO

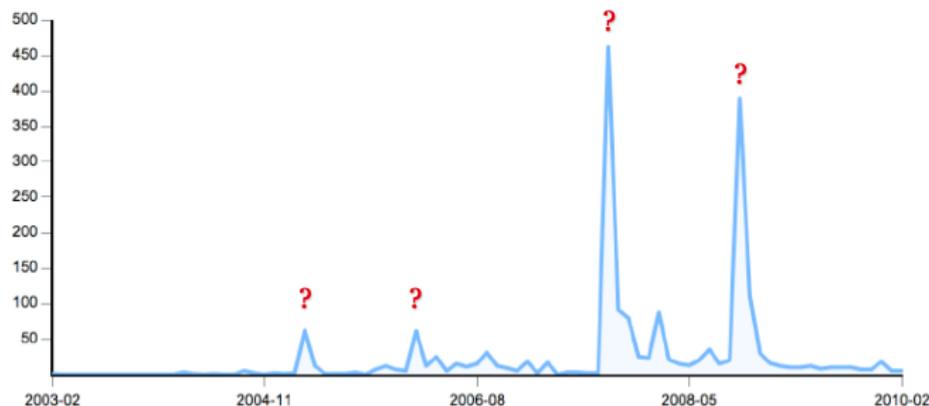


WIKICHANGES



<http://irlab.fe.up.pt/p/wikichanges>

COMO SUMARIAR EDIÇÕES?



O que significam os picos?
Como produzir sumários automáticos?

FERRAMENTAS E RECURSOS

ACESSO AOS DADOS

ARQUIVOS DE TEXTO

<http://download.wikimedia.org>

Cópia dos dados disponíveis na Wikipédia (*download* ou *cd*).
Artigos da Wikipédia Inglesa (>100 GB) + historial (5.6 TB!).

API

<http://en.wikipedia.org/w/api.php>

A plataforma MediaWiki inclui uma API que permite aceder programaticamente aos dados de forma parcelar.

OUTROS

Colecções preparadas e processadas por terceiros. Por exemplo: DBpedia (factos), YAGO (ontologia) ...

PROJECTOS DE INVESTIGAÇÃO

WIKIFY

<http://wikifyer.com>

Anota automaticamente o texto de qualquer página web com ligações para artigos da Wikipédia.

KORU

<http://www.nzdl.org/koru>

Motor de pesquisa que incorpora conhecimento extraído da Wikipédia (ex. tópicos, relações, etc).

WIKIDASHBOARD

<http://wikidashboard.parc.com>

Ferramenta de visualização e análise da dinâmica editorial dos artigos da Wikipédia.

WIKIDASHBOARD

Browser window: José Sócrates - Wikipedia, 11 x

URL: http://wikidashboard.parc.com/wiki/José_Sócrates#

Navigation: article | discussion | edit this page | history

Try Beta | Log in / create account

José Sócrates

From Wikipedia, the free encyclopedia

Most Recently - User:Rothorpe on 20100414

User	Page Up	Percentage
User:Page Up	59	7.7%
User:Rothorpe	36	4.7%
User:Tugaworld	31	4.0%
User:Husond	23	3.0%
User:G.-M. Cuperlino	21	2.7%
User:Miguelzinho	15	1.9%
User:Jomig	14	1.8%
User:Pularoid	12	1.6%
User:87.196.65.164	12	1.6%
User:Lcor	10	1.3%

Total edits: 865 (Since 2006: 771)

Max 45
Max 13
Max 14
Max 12
Max 4
Max 5
Max 5
Max 5
Max 4
Max 2
Max 6

Click bars to browse through time

Full Page Edit History

Reddit WikiDashboard | Submit this page | Preferences | Guide | Original Copy | Disclaimer | WikiDashboard, ©2008 PARC

This is a Portuguese name; the first family name is Carvalho and the second is Pinto de Sousa.

José Sócrates Carvalho Pinto de Sousa, GCIH (Alijó, 6 September 1957), commonly known simply as José Sócrates (Portuguese pronunciation: [ʒuʒɐ sɔkɾɐˈtɨ]) is the Prime Minister of Portugal and Secretary-General of the Socialist Party. Sócrates became Prime Minister on 12 March 2005. For the second half of 2007, he acted as the President-in-Office of the Council of the European Union. In addition to these posts, José Sócrates was Portugal's Minister for Youth and Sports and one of the organisers of the UEFA Euro 2004 football championship in Portugal, as well as being a former Environment Minister in the governments of António Guterres.

Contents

1 Biography

- 1.1 Early years
- 1.2 Education
- 1.3 Political career
- 1.4 Personal life
- 1.4.1 Family and residence
- 1.4.2 Health and well-being

2 Prime Minister of Portugal

José Sócrates GCIH



Prime Minister of Portugal

Incumbent

Assumed office
12 March 2005

President Jorge Sampaio

QUESTÕES?

REFERÊNCIAS

- ▶ **Academic studies about Wikipedia**

http://en.wikipedia.org/wiki/Academic_studies_about_Wikipedia

- ▶ **Academic studies of Wikipedia**

http://en.wikipedia.org/wiki/Wikipedia:Academic_studies_of_Wikipedia

- ▶ **Mining meaning from Wikipedia**

Medelyan, O., Milne, D., Legg, C., Witten, I. H. (2008)

<http://arxiv.org/abs/0809.4530>