

# Págico: Evaluating Wikipedia-based information retrieval in Portuguese

Cristina Motta (Linguatca/FCCN)

Alberto Simões (CEHUM/UM)

**Cláudia Freitas** (PUC-Rio / Linguatca)

Luis Costa (Linguatca/FCCN)

Diana Santos (Univ. of Oslo/Linguatca)



# Págico – Português Mágico



- Evaluation contest
- Organized by Linguateca, a distributed language resource center for Portuguese
  - <http://www.linguateca.pt>
- Announced in June 2011
- Results delivered in the beginning of 2012

2 years later... why  
Págico?



# Págico – Português Mágico

- Evaluation contest
  - Systems (and humans) must find non-trivial answers to complex information needs in Portuguese Wikipedia (150 topics were created):
    - Which Portuguese-speaking football players played professionally in more than three different countries?
    - Jornais que circularam no Rio de Janeiro entre 1910 e 1960
    - Produtos agrícolas com os quais se pode produzir combustível em escala comercial
    - Nomes ligados à luta contra o racismo no sec XX no Brasil
    - Filmes brasileiros premiados na categoria Montagem
    - Pratos brasileiros de origem ou influência indígena

answer to information needs that would require browsing hundreds of pages

“we are **not proposing** evaluation of toy systems just to check if **computers can be as good as humans**. Rather, we are **interested in systems that really help humans** in non trivial information finding”



# Págico – Português Mágico

- The task:
  - Find (PT)Wikipedia pages that answer the topics. The answer must be **the name of the page**.
    - Additional pages must be provided when the answer page is incomplete
  - Pages (answers and justifications) must be selected from a static version of the portuguese Wikipedia.
    - **Pagico Collection:** a Wikipedia snapshot of April 25th 2011
      - Documents converted in XHTML files
      - Without *Infoboxes*



# The Collection: Wikipedia folksonomy

- Distribution of categories across PT Wikipedia pages
  - 99.446 categories classify 681.058 documents

# Documents	# Categories	Percent
]0, 1]	32 652	34.21%
]1, 66]	59 775	62.63%
]66, 130]	1 789	1.87%
]130, 194]	507	0.53%
]194, 260]	231	0.24%
]260, 345]	166	0.17%
]345, 442]	108	0.11%
]442, 592]	84	0.09%
]592, 862]	68	0.07%
]862, ∞[	65	0.07%

Table 7: Number of documents per category number.

# Categories	# Documents	Percent
0	8 771	1.271%
]0, 8]	676 705	98.097%
]8, 15]	4 008	0.581%
]15, 23]	314	0.046%
]23, 33]	25	0.004%
]33, ∞[	6	0.001%

Table 8: Number of categories per document.



# Topic creation

- Topics should be interesting to the Portuguese-speaking community
- The answers
  - should not be obvious
    - ...and not obvious to Google(!), in order to justify a wikipedia search
  - should be spread among different pages in PT Wikipedia
- Topic: *Músicos do início da Bossa Nova / Musicians associated with the development of Bossa Nova*
  - Browse Bossa\_Nova page..

It doesn't involve complex search...  
...from the standpoint of humans.

# Topic creation

- PT Wikipedia in 150 topics
  - 8 Super-themes
  - 25 subthemes
  - One or more areas/subareas were assigned to each topic:
    - *“Documentary films about brazilian politicians”*
    - Theme: Artes; Letras; Política
    - Subtheme: cinema; política; história
    - *“Comidas de santo (comidas rituais do Candomblé ou Umbanda) que também fazem parte da culinária brasileira”// religious food that are also brazilian food*
    - Theme: Cultura
    - Subtheme: culinária; religião; antropologia; folclore

Tema -subtema	Tópicos	
	#	%
<b>Letras</b>	<b>69</b>	<b>46,00</b>
- história	50	33,33
- literatura	15	10,00
- linguística	6	4,00
- jornalismo	3	2,00
- filosofia	2	1,33
<b>Artes</b>	<b>36</b>	<b>24,00</b>
- música	19	12,67
- cinema	10	6,67
- televisão	4	2,67
- artes plásticas	2	1,33
- artes	2	1,33
<b>Geografia</b>	<b>34</b>	<b>22,67</b>
- geografia	26	17,33
- arquitetura/urbanismo	7	4,67
- demografia	4	2,67
- geologia	2	1,33
<b>Cultura</b>	<b>27</b>	<b>18,00</b>
- antropologia/folclore	12	8,00
- religião	7	4,67
- culinária	5	3,33
- cultura	3	2,00
- ensino	2	1,33
<b>Política</b>	<b>19</b>	<b>12,67</b>
<b>Desporto/Esportes</b>	<b>18</b>	<b>12,00</b>
<b>Ciência</b>	<b>14</b>	<b>9,33</b>
- saúde	4	2,67
- zoologia	3	2,00
- ciência	2	1,33
- botânica	2	1,33
- geologia	2	1,33
- matemática	1	0,67
<b>Economia</b>	<b>6</b>	<b>4,00</b>





# Answer documents

- A correct answer , according to Págico:
  - Wiki-PT page title → expected semantic class

*“Gêneros musicais que misturam samba e gêneros norte americanos”*

→ “Turma\_da\_Pilantragem” ❌

*“Que países têm amarelo na bandeira?”*

→ “Bandeira\_do\_Brasil” ❌

Good answer from a practical perspective...

Bad answer from a logic perspective

.. how to decide/define what should be considered a “good answer” ?

Correct answer  
X  
Useful answer

LogOnto 2014 – FGV - Brasil

...taking logic into consideration facilitates the evaluation task





# Answer documents – justification pages

- Additionally, participants need to provide the wikipedia pages that support that the chosen answer is, indeed, the correct one

Pagico\_103  
Tema(s): **Cultura, Geografia**  
Dados geográficos: **Brasil**

[Página principal](#) | [Sair](#) [observador\_hum]

**Movimentos culturais surgidos no nordeste do Brasil.**

Do we need justification here?

Tema: Qualquer tema

[Tópico Anterior](#) [Próximo Tópico](#)

[Lista de Tópicos](#)

[Ir para página visualizada anteriormente](#) [Voltar às respostas a este tópico](#)

## Manguebeat

Manguebeat (também grafado como **manguebit** ou **mangue beat**) é um movimento musical que surgiu no **Brasil** na **década de 90** em **Recife** que mistura ritmos regionais, como o **maracatu**, com **rock**, **hip hop** e **reggae**. Esse estilo tem como ícone o músico **Chico Science e Nação Zumbi**, idealizador do rótulo mangue e principal divulgador das idéias, ritmos e contestações do movimento desse movimento foi **Fred 04**, vocalista da banda **Mundo Livre S/A** e autor do primeiro manifesto do Manguebeat. O objetivo do movimento surgiu de uma meta mais rica do planeta, o Manguebeat precisava de diversidade, a agitação na música contaminou bandas de **Pernambuco** e do **Brasil**, sendo o principal cenário. Com o surgimento de várias bandas no cenário nesse cenário Manguebeat.

Págico: only systems must provide justification

LogOnto 2014 - FGV - Brasil



# Answer documents – justification pages

Cantores **vaiados** nos festivais de música brasileira na década de 60.

Tema: Qualquer tema

Tópico Anterior Próxímo Tópico

Lista de Tópicos

[Ir para página visualizada anteriormente](#) [Voltar às respostas a este tópico](#)

### Festivais de MPB nos anos de 1960

No festival de [1967](#) faria sucesso também com [Roda Viva](#), interpretada por ele e pelo grupo [MPB-4](#) — amigos e intérpretes de muitas de suas canções. Em [1968](#) voltou a vencer outro *Festival*, o III [Festival Internacional da Canção da TV Globo](#). Como compositor, em parceria com [Tom Jobim](#), com a canção [Sabiá](#). **Mas desta vez a vitória foi contestada pelo público**, que preferiu a canção que ficou em segundo lugar: [Pra não dizer que não falei de flores](#), de [Geraldo Vandré](#).

A participação no *Festival*, com *A Banda*, marcou a primeira aparição pública de grande repercussão apresentando um estilo amparado no movimento musical urbano carioca da [Bossa nova](#), surgido em [1957](#). Ao longo da carreira, o samba e a MPB também seriam estilos amplamente explorados.

### Trilha-sonora e adaptações de livros

[thumbnail/200px/esquerda/Chico](#) participou como autor e compôs várias canções de sucesso para o filme *Quando o*

To be contested by the audience = to be booed?

# Answers in Págico

Topics with most and least correct answer documents

Topic	#
Indigenous tribes living in the Amazon Rainforest	95
Museums in capitals of Lusophone countries	62
Locations mentioned in "Os Lusíadas"	51
Indigenous Brazilian peoples considered extinct.	50
Viceroy of the Portuguese India	48
...	
Politicians from Portuguese speaking African countries who studied in the Soviet Union	2
Churches in Rio de Janeiro constructed by Afro-Brazilian confraternities.	1
Members of Parliament from FRELIMO	1
Mozambican writers who received Prémio Camões	1
Foreign writers who visited Portugal in the 19th Century and published descriptions of their travels	1

## The most correct topics

ID	Topic	Total	Hum	Sys	H & S	
H	135	Aves de Angola	54	10	44	0
	19	Tribos indígenas que vivem na Amazônia.	115	56	35	24
	90	Filmes brasileiros premiados na categoria Montagem.	34	8	19	7
	13	Dinossauros carnívoros que habitaram o Brasil.	23	6	12	5
S	19	Tribos indígenas que vivem na Amazônia.	115	56	35	24
	62	Praias de Portugal boas para a prática de surf	30	5	6	19
	7	Guitarristas portugueses que também foram compositores.	34	17	0	17
	11	Filmes sobre o cangaço.	41	20	4	17



# Págico – by-products

- CARTOLA – <http://www.linguateca.pt/Cartola>
  - Freely available package
    - Págico Collection: 681.058 documents from wikipedia portuguesa de 25 de abril de 2011;
    - Págico topics
    - Evaluated answers
    - ..
- Special Edition of the Linguamática journal of April 2012 (Santos et al., 2012) – <http://www.linguamatica.com>

## Cartola: Pacote de recursos do Págico

[Págico](#), [Linguatca](#)

O Cartola, que foi disponibilizado no dia 17 de abril de 2012, inclui todos os recursos criados no âmbito do Págico, nomeadamente:

- [coleção do Págico](#), de 681.058 documentos da wikipedia portuguesa de 25 de abril de 2011;
- coleção de tópicos do Págico ([xml](#), [txt](#))
- [monte das respostas avaliadas](#)
- [subcoleção do monte do Págico](#)

Além disso, e para facilitar estudos de aspetos particulares, também disponibilizamos

- [lista de respostas corretas e justificadas, sem as respetivas justificações](#)
- [lista de respostas corretas e justificadas, com as respetivas justificações](#)
- [lista de respostas consideradas corretas independentemente de estarem bem justificadas](#)

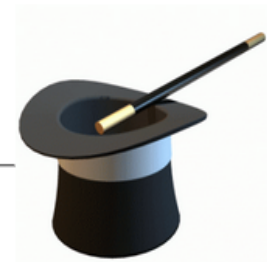
E a nova versão do SIGA encontra-se disponível em formato tar.gz: [SIGA\\_2012.tar.gz](#)

### Financiamento

A Linguatca e o Págico foram financiados pelas seguintes entidades até 31 de Dezembro de 2011.



Mas continuaram a ser apoiados pelas seguintes instituições:





Pagico\_001 pt/1/9/7/1972\_\_filme\_.965afb.xml  
Pagico\_001 pt/a/r/a/Araguaya\_-\_Conspiração\_do\_Silêncio.a41d5e.xml  
Pagico\_001 pt/a/s/\_/As\_Meninas\_\_filme\_.7acb87.xml  
Pagico\_001 pt/b/a/t/Batismo\_de\_Sangue.26611b.xml  
Pagico\_001 pt/b/e/y/Beyond\_Citizen\_Kane.d35325.xml  
Pagico\_001 pt/b/r/a/Brazil,\_a\_Report\_on\_Torture.b7dae1.xml  
Pagico\_001 pt/b/r/a/Brazil\_\_Cinema,\_Sex\_and\_the\_Generals.f2bdd0.xml  
Pagico\_001 pt/c/a/b/Cabra-cega\_\_filme\_.e29b13.xml  
Pagico\_001 pt/c/a/b/Cabra\_Marcado\_para\_Morrer.f5574e.xml  
Pagico\_001 pt/c/o/n/Condor\_\_filme\_.c34b5d.xml  
Pagico\_001 pt/c/o/r/Corpo\_em\_Delito.072df4.xml  
Pagico\_001 pt/d/e/d/Dedão\_Mamata.214f2f.xml  
Pagico\_001 pt/f/e/l/Feliz\_Ano\_Velho\_\_filme\_.9f7035.xml  
Pagico\_001 pt/j/a/n/Jango\_\_filme\_.2b9fe4.xml  
Pagico\_001 pt/l/a/m/Lamarca.259536.xml  
Pagico\_001 pt/l/u/l/Lula,\_o\_Filho\_do\_Brasil\_\_filme\_.b218b5.xml  
Pagico\_001 pt/l/u/t/Lutas\_\_filme\_.01bc03.xml  
Pagico\_001 pt/m/a/r/Marighella\_-\_Retrato\_Falado\_do\_Guerrilheiro.c697c3.xml  
Pagico\_001 pt/m/e/m/Memórias\_do\_Câncer\_\_filme\_.d248b4.xml  
Pagico\_001 pt/o/\_/a/O\_Ano\_em\_que\_Meus\_Pais\_Saíram\_de\_Férias.b58539.xml  
Pagico\_001 pt/o/\_/b/O\_Bom\_Burguês.4c5c73.xml  
Pagico\_001 pt/o/\_/q/O\_Que\_É\_Iso,\_Companheiro\_\_filme\_.ec72f9.xml  
Pagico\_001 pt/o/s/\_/Os\_Herdeiros.a0f25f.xml  
Pagico\_001 pt/p/i/n/Pindorama\_\_filme\_.58777d.xml  
Pagico\_001 pt/p/r/a/Pra\_frente,\_Brasil.634849.xml  
Pagico\_001 pt/q/u/a/Quase\_Dois\_Irmãos.c80d70.xml  
Pagico\_001 pt/s/o/n/Sonhos\_e\_Desejos.e40399.xml  
Pagico\_001 pt/t/e/r/Terra\_em\_Transe.d67729.xml  
Pagico\_001 pt/z/u/z/Zuzu\_Angel\_\_filme\_.b6838a.xml  
Pagico\_002 pt/a/\_/i/A\_Indomada.de74a7.xml  
Pagico\_002 pt/a/\_/v/A\_Viagem\_1975\_.ea152d.xml  
Pagico\_002 pt/a/\_/v/A\_Viagem\_1994\_.3b816d.xml  
Pagico\_002 pt/c/a/m/Caminhos\_do\_Coração.c40597.xml  
Pagico\_002 pt/f/e/r/Fera\_Ferida.c92cc4.xml  
Pagico\_002 pt/k/u/b/Kubanacan.9ff6cd.xml  
Pagico\_002 pt/o/\_/b/O\_Beijo\_do\_Vampiro.5bddf8.xml  
Pagico\_002 pt/o/\_/b/O\_Bem-Amado\_\_telenovela\_.41a30d.xml  
Pagico\_002 pt/o/\_/c/O\_Clone.12457d.xml  
Pagico\_002 pt/o/\_/f/O\_Fim\_do\_Mundo.5d4e8e.xml  
Pagico\_002 pt/o/s/\_/Os\_Mutantes\_\_Caminhos\_do\_Coração.c40597.xml

## Pagico\_001: FILMES SOBRE A DITADURA OU SOBRE O GOLPE MILITAR NO BRASIL

## Pagico\_002: TELENÓVELAS BRASILEIRAS DO GÊNERO REALISMO FANTÁSTICO

#[16 Feb 2012] Este ficheiro contém todas as respostas corretas e justificadas corretamente, bem como as suas justificações.



Pagico\_001 pt/1/9/7/1972\_filme\_.965afb.xml pt/a/n/e/Anexo\_Lista\_de\_filmes\_sobre\_ditaduras\_militares\_na\_América\_Latina.605516.xml  
Pagico\_001 pt/a/r/a/Araguaya\_-\_Conspiración\_do\_Silêncio.a41d5e.xml pt/a/n/e/Anexo\_Lista\_de\_filmes\_sobre\_ditaduras\_militares\_na\_América\_La  
Pagico\_001 pt/a/s/\_/As\_Meninas\_filme\_.7acb87.xml  
Pagico\_001 pt/b/a/t/Batismo\_de\_Sangue.26611b.xml  
Pagico\_001 pt/b/a/t/Batismo\_de\_Sangue.26611b.xml pt/a/n/e/Anexo\_Lista\_de\_filmes\_sobre\_ditaduras\_militares\_na\_América\_Latina.605516.xml  
Pagico\_001 pt/b/e/y/Beyond\_Citizen\_Kane.d35325.xml pt/a/n/e/Anexo\_Lista\_de\_filmes\_sobre\_ditaduras\_militares\_na\_América\_Latina.605516.xml  
Pagico\_001 pt/b/r/a/Brazil,\_a\_Report\_on\_Torture.b7dae1.xml  
Pagico\_001 pt/b/r/a/Brazil\_Cinema,\_Sex\_and\_the\_Generals.f2bdd0.xml  
Pagico\_001 pt/c/a/b/Cabra-cega\_filme\_.e29b13.xml  
Pagico\_001 pt/c/a/b/Cabra-cega\_filme\_.e29b13.xml pt/a/n/e/Anexo\_Lista\_de\_filmes\_sobre\_ditaduras\_militares\_na\_América\_Latina.605516.xml  
Pagico\_001 pt/c/a/b/Cabra\_Marcado\_para\_Morrer.f5574e.xml  
Pagico\_001 pt/c/a/b/Cabra\_Marcado\_para\_Morrer.f5574e.xml pt/a/n/e/Anexo\_Lista\_de\_filmes\_sobre\_ditaduras\_militares\_na\_América\_Latina.605516.xml  
Pagico\_001 pt/c/o/n/Condor\_filme\_.c34b5d.xml  
Pagico\_001 pt/c/o/r/Corpo\_em\_Delito.072df4.xml  
Pagico\_001 pt/d/e/d/Dedão\_Mamata.214f2f.xml  
Pagico\_001 pt/f/e/l/Feliz\_Ano\_Velho\_filme\_.9f7035.xml pt/a/n/e/Anexo\_Lista\_de\_filmes\_sobre\_ditaduras\_militares\_na\_América\_Latina.605516.xml pt  
Pagico\_001 pt/j/a/n/Jango\_filme\_.2b9fe4.xml  
Pagico\_001 pt/j/a/n/Jango\_filme\_.2b9fe4.xml pt/a/n/e/Anexo\_Lista\_de\_filmes\_sobre\_ditaduras\_militares\_na\_América\_Latina.605516.xml pt/c/i/n/Cir  
Pagico\_001 pt/l/a/m/Lamarca.259536.xml  
Pagico\_001 pt/l/a/m/Lamarca.259536.xml pt/a/n/e/Anexo\_Lista\_de\_filmes\_sobre\_ditaduras\_militares\_na\_América\_Latina.605516.xml  
Pagico\_001 pt/l/u/l/Lula,\_o\_Filho\_do\_Brasil\_filme\_.b218b5.xml pt/a/n/e/Anexo\_Lista\_de\_filmes\_sobre\_ditaduras\_militares\_na\_América\_Latina.60551  
Pagico\_001 pt/l/u/t/Lutas\_filme\_.01bc03.xml  
Pagico\_001 pt/m/a/r/Marighella\_-\_Retrato\_Falado\_do\_Guerrilheiro.c697c3.xml  
Pagico\_001 pt/m/e/m/Memórias\_do\_Cárcere\_filme\_.d248b4.xml  
Pagico\_001 pt/o/\_/a/O\_Ano\_em\_que\_Meus\_Pais\_Saíram\_de\_Férias.b58539.xml  
Pagico\_001 pt/o/\_/a/O\_Ano\_em\_que\_Meus\_Pais\_Saíram\_de\_Férias.b58539.xml pt/a/n/e/Anexo\_Lista\_de\_filmes\_sobre\_ditaduras\_militares\_na\_América\_Lat  
Pagico\_001 pt/o/\_/b/O\_Bom\_Burguês.4c5c73.xml  
Pagico\_001 pt/o/\_/q/O\_Que\_É\_Issó,\_Companheiro\_filme\_.ec72f9.xml  
Pagico\_001 pt/o/\_/q/O\_Que\_É\_Issó,\_Companheiro\_filme\_.ec72f9.xml pt/a/n/e/Anexo\_Lista\_de\_filmes\_sobre\_ditaduras\_militares\_na\_América\_Latina.(  
Pagico\_001 pt/o/s/\_/Os\_Herdeiros.a0f25f.xml  
Pagico\_001 pt/o/s/\_/Os\_Herdeiros.a0f25f.xml pt/c/i/n/Cinema\_do\_Brasil.a12cc6.xml  
Pagico\_001 pt/p/i/n/Pindorama\_filme\_.58777d.xml  
Pagico\_001 pt/p/r/a/Pra\_frente,\_Brasil.634849.xml pt/a/n/e/Anexo\_Lista\_de\_filmes\_sobre\_ditaduras\_militares\_na\_América\_Latina.605516.xml  
Pagico\_001 pt/q/u/a/Quase\_Dois\_Irmãos.c80d70.xml pt/a/n/e/Anexo\_Lista\_de\_filmes\_sobre\_ditaduras\_militares\_na\_América\_Latina.605516.xml  
Pagico\_001 pt/s/o/n/Sonhos\_e\_Desejos.e40399.xml pt/a/n/e/Anexo\_Lista\_de\_filmes\_sobre\_ditaduras\_militares\_na\_América\_Latina.605516.xml  
Pagico\_001 pt/t/e/r/Terra\_em\_Transe.d67729.xml pt/a/n/e/Anexo\_Lista\_de\_filmes\_sobre\_ditaduras\_militares\_na\_América\_Latina.605516.xml  
Pagico\_001 pt/z/u/z/Zuzu\_Angel\_filme\_.b6838a.xml  
Pagico\_001 pt/z/u/z/Zuzu\_Angel\_filme\_.b6838a.xml pt/a/n/e/Anexo\_Lista\_de\_filmes\_sobre\_ditaduras\_militares\_na\_América\_Latina.605516.xml



# Págico and LogOnto..

- Págico
  - Can it be useful to reasoning systems?
  - Bootstrap a KB from Págico?
  - Test reasoning systems?
- More about Pagico:
  - <http://www.linguateca.pt/Pagico/>

Page type	# Documents
Templates	32 900
Disambiguation	5 006
Redirection	574 077
Media	9 678
Articles	856 005

Table 6: Document distribution per page type.





# References

- Diana Santos, Cristina Mota, Cláudia Freitas e Luís Costa (eds.)  
*Linguamática* 4 (1). Abril, 2012

- Acknowledgments

## Financiamento

A Linguateca e o Páxico foram financiados pelas seguintes entidades até 31 de Dezembro de 2011.



MCTES MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR



Mas continuaram a ser apoiados pelas seguintes instituições:

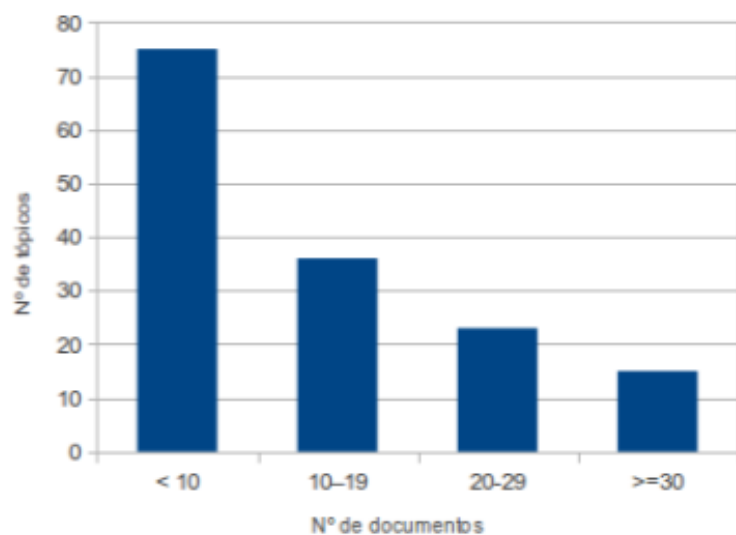




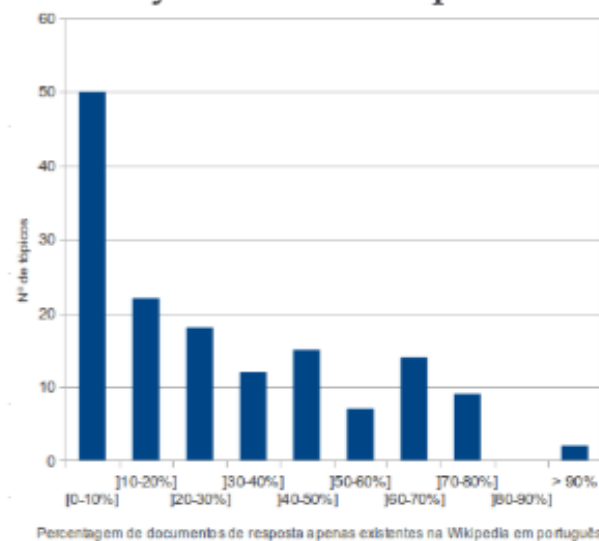
# CARTOLA <http://www.linguateca.pt/Cartola>

## Págico answers pool

Number of answers and justification documents



Percentage of answers and justifications only in the PT wikipedia



# Págico – Português Mágico



## MOTIVATION AND TASK

Is it possible to develop better systems to answer realistic user needs, searching for answers to a particular topic in Wikipedia? Is Wikipedia in Portuguese good enough to provide information on lusophone topics? Can we learn from watching people trying to answer them? Is competition or cooperation between human and automatic participants worth indulging in?



Pagico

Tópicos

Temas

Dados geográficos I

[Filmes sobre a ditadura ou sobre o golpe militar no Brasil](#)

Artes, Letras

Brasil

Pagico\_001

Tema(s): Artes, Letras

Dados geográficos: Brasil

[Página principal](#) | [Sair](#) [observador\_hum]



## Filmes sobre a ditadura ou sobre o golpe militar no Brasil

Tema: Qualquer tema	Resposta	Justificação	Correta?	Justificada?	
<a href="#">Tópico Anterior</a> <a href="#">Próximo Tópico</a>	<a href="#">Sonhos e Desejos</a>	Sem Justificações	Sim	Sim	<a href="#">Corrigir avaliação</a>
<a href="#">Lista de Tópicos</a>	<a href="#">Lamarca</a>	Sem Justificações	Sim	Sim	<a href="#">Corrigir avaliação</a>
	<a href="#">Quase Dois Irmãos</a>	Sem Justificações	Sim	Sim	<a href="#">Corrigir avaliação</a>
	<a href="#">As Meninas filme</a>	Sem Justificações	Sim	Sim	<a href="#">Corrigir avaliação</a>
	<a href="#">Cabra Marcado para Morrer</a>	Sem Justificações	Sim	Sim	<a href="#">Corrigir avaliação</a>
	<a href="#">Brazil, a Report on Torture</a>	Sem Justificações	Sim	Sim	<a href="#">Corrigir avaliação</a>
	<a href="#">Lamarca</a>	Sem Justificações	Sim	Sim	<a href="#">Corrigir avaliação</a>
	<a href="#">Dedé Mamata</a>	Sem Justificações	Sim	Sim	<a href="#">Corrigir avaliação</a>
	<a href="#">O Ano em que Meus Pais Saíram de Férias</a>	Sem Justificações	Sim	Sim	<a href="#">Corrigir avaliação</a>
	<a href="#">Cabra-cega filme</a>	Sem Justificações	Sim	Sim	<a href="#">Corrigir avaliação</a>
	<a href="#">O Bom Burguês</a>	Sem Justificações	Sim	Sim	<a href="#">Corrigir avaliação</a>
	<a href="#">Marighella - Retrato Falado do Guerrilheiro</a>	Sem Justificações	Sim	Sim	<a href="#">Corrigir avaliação</a>
	<a href="#">Os Herdeiros</a>	Sem Justificações	Sim	Sim	<a href="#">Corrigir avaliação</a>
	<a href="#">Zuzu Angel filme</a>	Sem Justificações	Sim	Sim	<a href="#">Corrigir avaliação</a>
	<a href="#">O Que É Isso, Companheiro filme</a>	Sem Justificações	Sim	Sim	<a href="#">Corrigir avaliação</a>
	<a href="#">Jango filme</a>	Sem Justificações	Sim	Sim	<a href="#">Corrigir avaliação</a>
	<a href="#">Batismo de Sangue</a>	Sem Justificações	Sim	Sim	<a href="#">Corrigir avaliação</a>

# PÁGICO COLLECTION

- based on the 25 April 2011 wikipedia snapshot;
- converted to XHTML using:
  - `mwlib` for the markup conversion;
  - `MediaWiki::DumpFile` to control the snapshot parsing;
  - in-house tools to manage macro expansion;

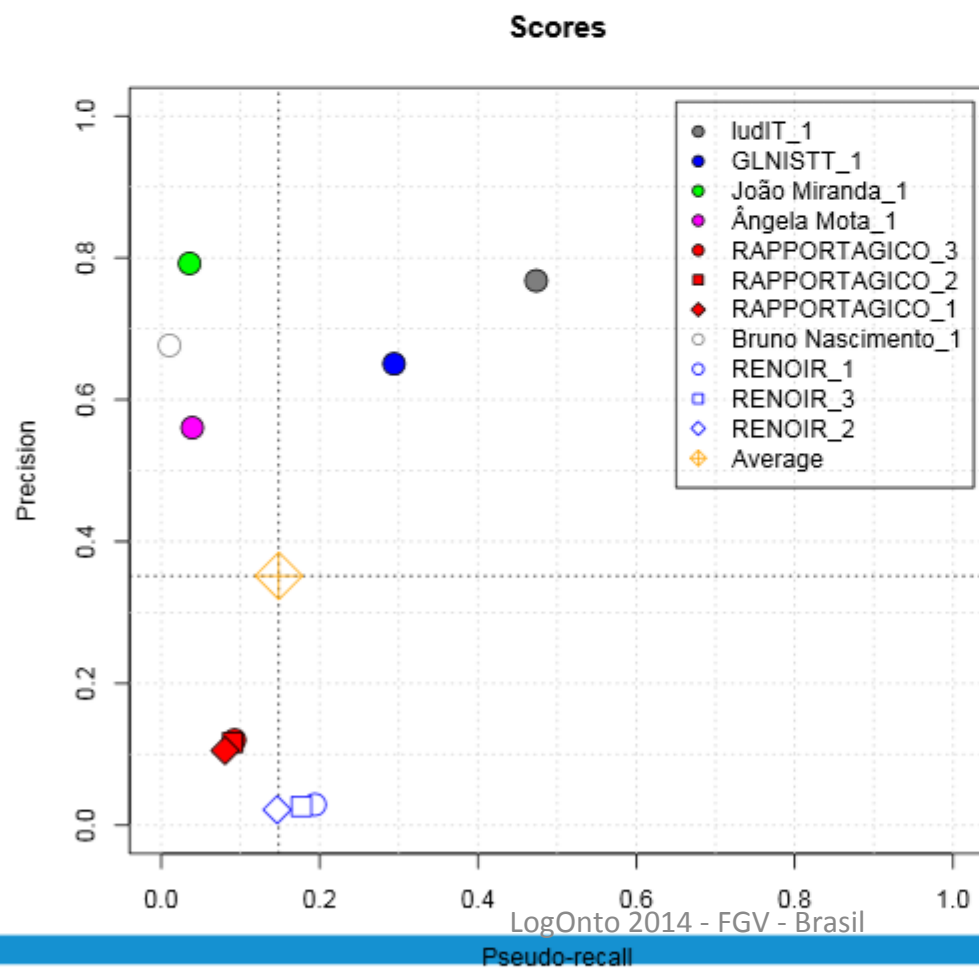
## Collection constitution:

Page type	Total docs
Template pages	32 900
Disambiguation pages	5 006
Redirection pages	574 077
Multimedia pages	9 678
Article pages	856 005



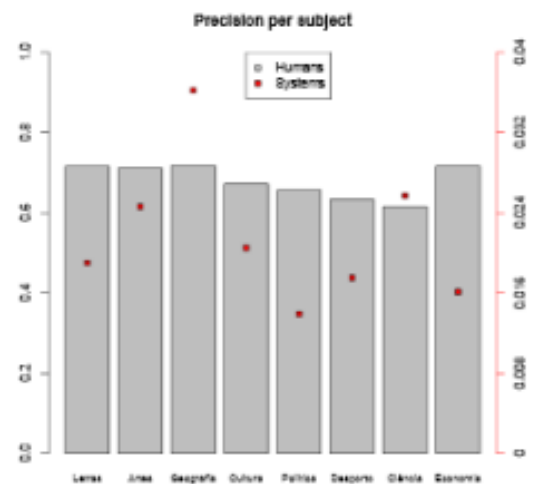
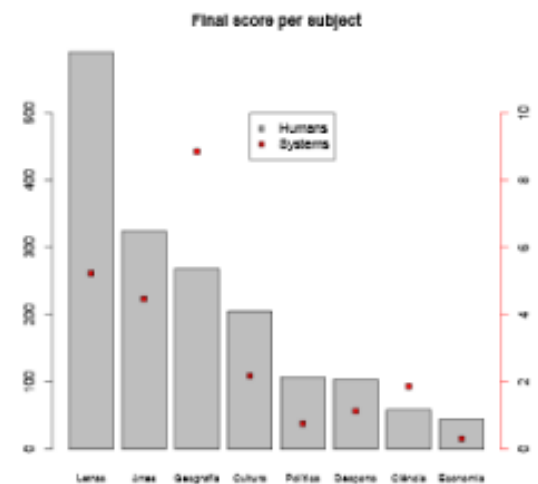
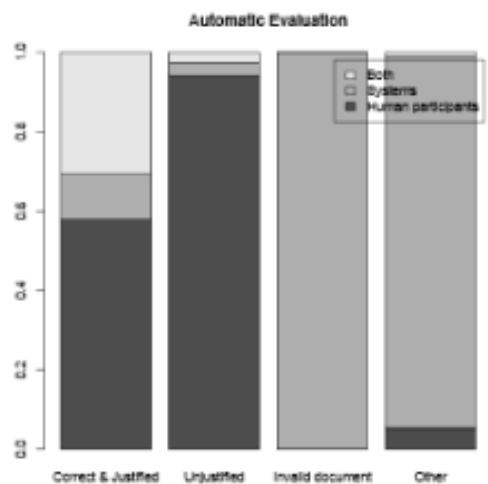
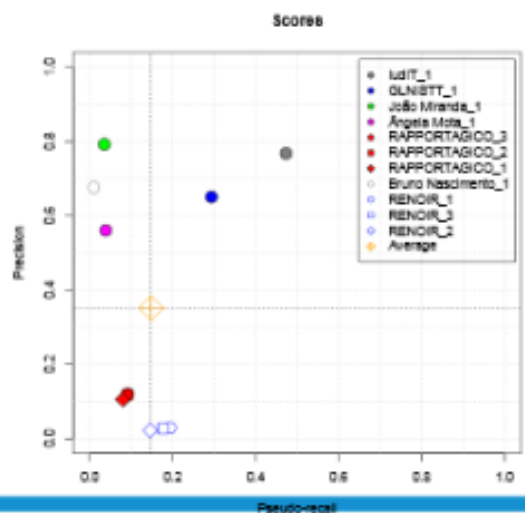


# HUMANS VS. SYSTEMS



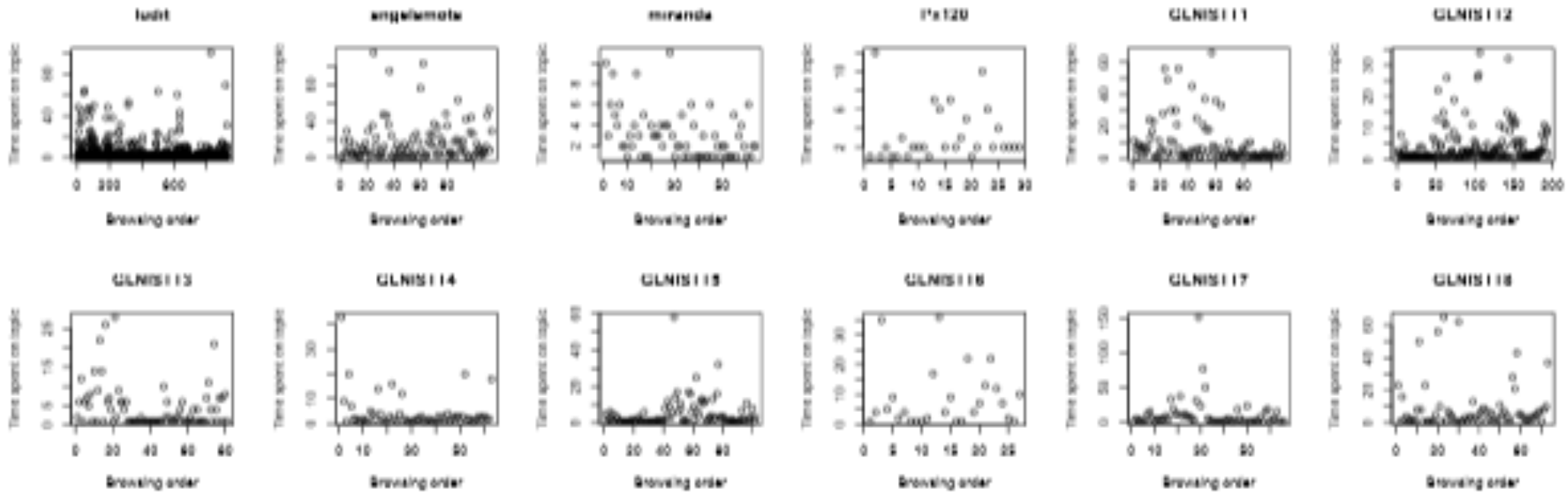


# HUMANS VS. SYSTEMS

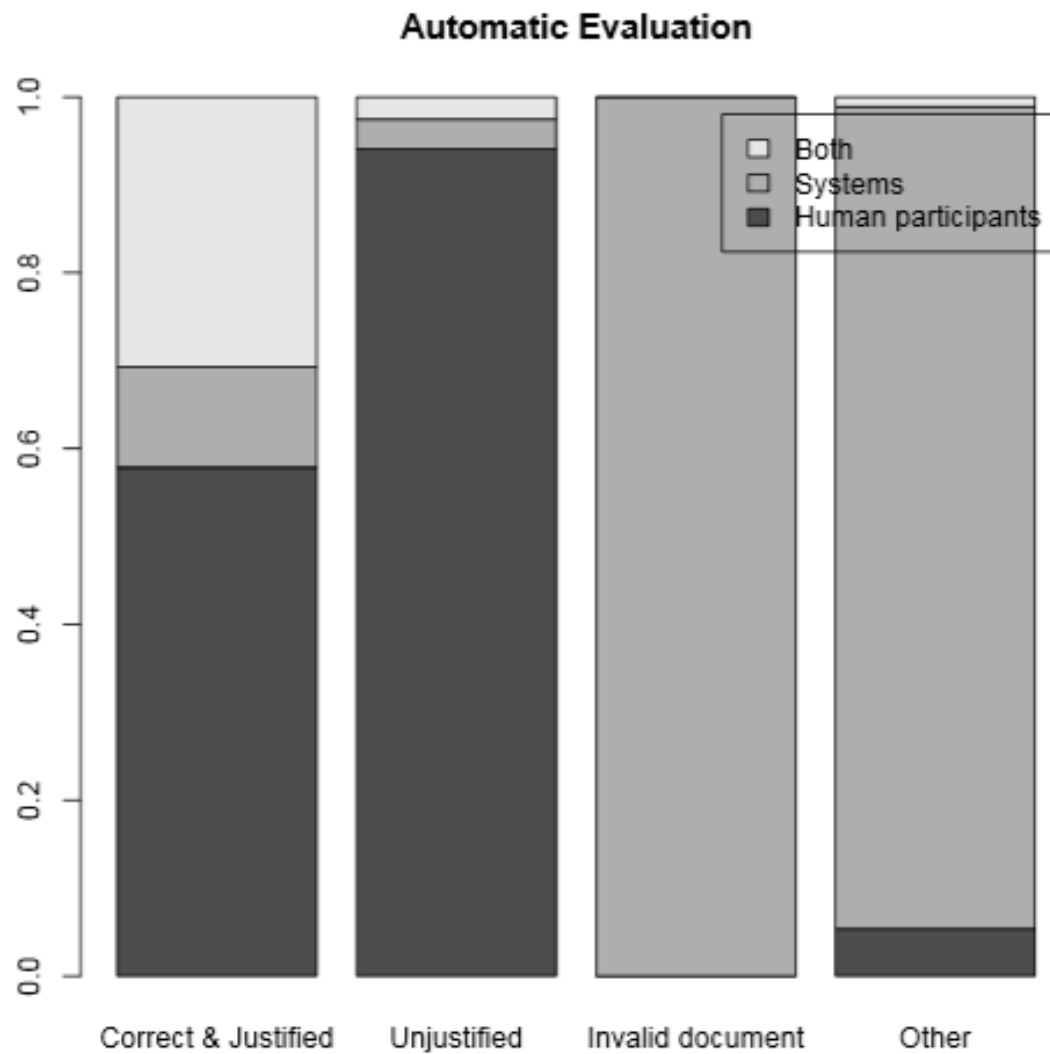




# USER BROWSER BEHAVIOUR

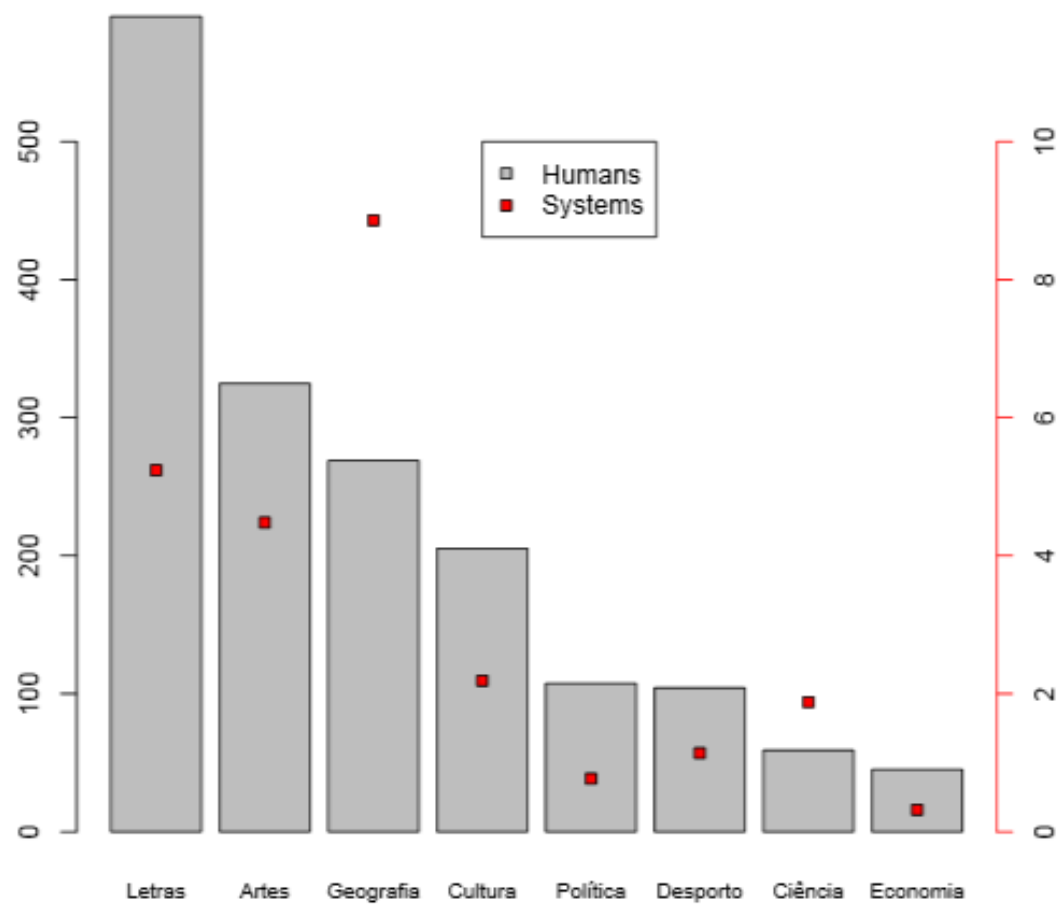




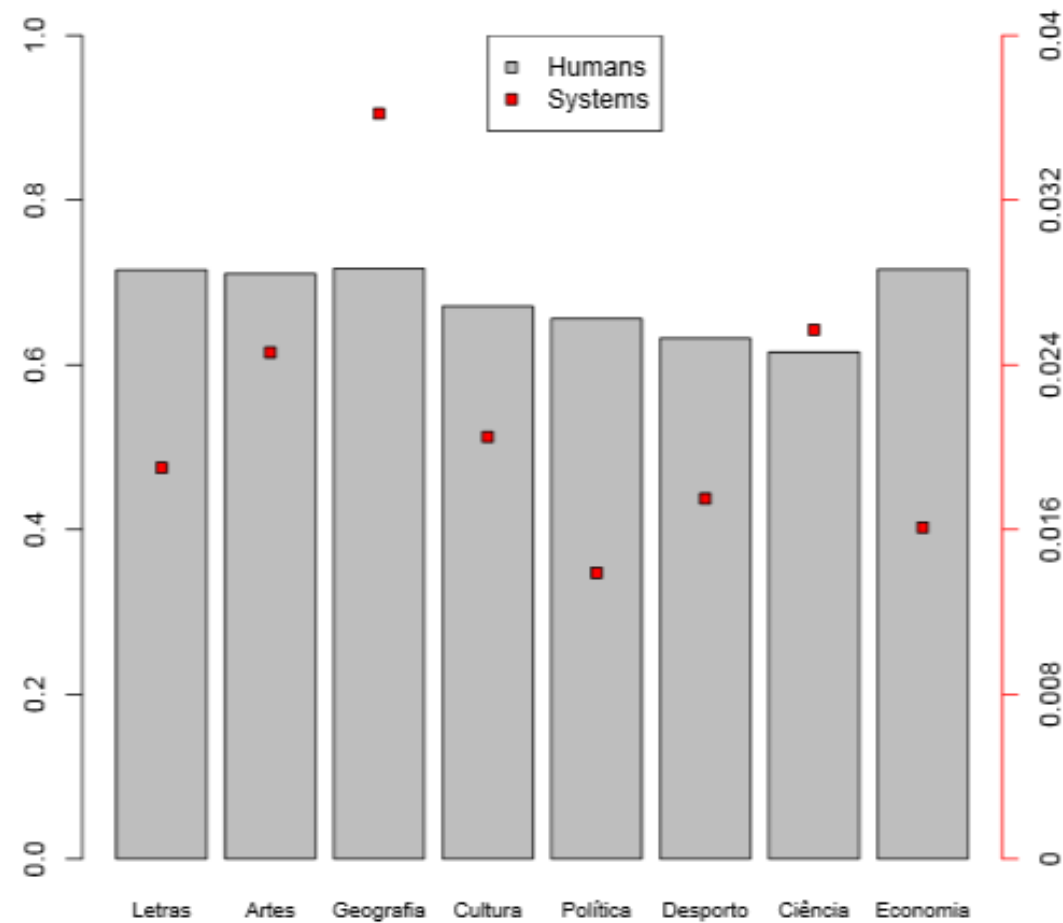




Final score per subject



Precision per subject



# EVAL MEASURES

**Precision:**  $P_{p,c} = \frac{|C_{p,c}|}{|R_{p,c}|}$

**Pseudo-recall:**

$$\alpha_{p,c} = \frac{|C_{p,c}|}{|C_{Pagico}| + |C_{aval}|}$$

**Pseudo-F-measure:**

$$\phi_{p,c} = 2 \times \frac{P_{p,c} \times \alpha_{p,c}}{P_{p,c} + \alpha_{p,c}}$$

**Originality:**

$$O_{p,c} = \sum_i^T \sum_j^{R_{p,c,i}} o(r_{p,c,i,j})$$

**Creativity:**

$$K_{p,c} = \sum_i^T \sum_j^{R_{p,c,i}} k(r_{p,c,i,j})$$

**Final score:**  $M_{p,j} = |C_{p,c}| \times P_{c,j}$

In addition to the measures used in GikiP and GikiCLEF, we chose to investigate originality and creativity, by weighing differently answers according to the number of participants who found them.





# SIGA

- Topic creation
- System run submission and testing
- Human participation interface
- Assessment interface
- Conflict resolution
- Pool browsing
- Scoring

ID do tópico: **Página\_004**  
 Descrição do tópico: **Descrição do tópico**  
 Classificação e exemplo de uso: **Classificação e exemplo de uso**  
 Respostas: **Respostas**  
 Ver comentários: **Ver comentários**  
 Adicionar: **Adicionar**

ID do tópico: **Página\_004**  
 Nome do tópico: **DianaAr**  
 Tópico: **Mulheres violencelinas de língua portuguesa**  
 Classificação adicional e exemplo de uso: **Uma pessoa pode estar interessada na vida e na carreira de mulheres que se distinguiram na música, ou especialmente a tocar violoncelo. Tema: música**  
 Temas: **Música**  
 Supertemas: **Artes**  
 Dados geográficos: **Lausfémia**

**Adicionar ou remover respostas**

Documento	Tamanho	Avaliar	Justificações (0)	Auto-justificado	
<a href="#">página/Carmen_Monacha_19419b.xml</a>	5,95 KB	DianaAr	Justificações (0)	Sim	Remover
<a href="#">página/Denise_Emmer_6b270a.xml</a>	22,64 KB	DianaAr	Justificações (0)	Sim	Remover
<a href="#">página/Ovelhemina_Suggis_1a7652.xml</a>	29,18 KB	DianaAr	Justificações (0)	Sim	Remover

Para encontrar um documento na coleção basta:  
 - ou URL da Wikipédia (por exemplo, "[http://pt.wikipedia.org/wiki/PC\\_Porto](#)")  
 - ou um título da Wikipédia (por exemplo, "[Pedro Nunes \(matemático\)](#)")  
 - ou uma lista de termos que façam parte do título (por exemplo, "[Pedro Nunes matemático](#)").

Página: 001  
 Tema(s): **Artes, Letras**  
 Dados geográficos: **Brasil**  
[Página principal](#) | [Sair](#) (ajuda)

## Filmes sobre a ditadura ou sobre o golpe militar no Brasil

Página: 001  
 Não avaliado | Não avaliado  
 Tópico | Tópico

Procurar na página:

**Resposta: Dois Córregos\_film\_cae806**  
 Curioso | Incompleto | Quebrado

**Justificação:**  
 Dois Córregos\_film\_cae806  
 Não foram dadas justificações adicionais.

**Comentários:**

Comentários do avaliador:

**Respostas e justificações pré-determinadas**  
[Resposta de Sérgio266123](#) (Auto-justificado)  
[Cada Menina tem Seus Segredos](#) (Auto-justificado)  
[Cinder\\_film\\_c20624](#) (Auto-justificado)  
[Luzerna220230](#) (Auto-justificado)  
[Margarida\\_-\\_Retrato\\_Filme\\_de\\_Octávio\\_07712](#) (Auto-justificado)  
[Monstros\\_de\\_Cásmo\\_film\\_c24804](#) (Auto-justificado)  
[Os\\_Homens\\_0102](#) (Auto-justificado)  
[O\\_Ano\\_em\\_que\\_Meu\\_Dois\\_Sobres\\_de\\_Filmes\\_010730](#) (Auto-justificado)  
[O\\_Bem\\_Burguês\\_65472](#) (Auto-justificado)  
[O\\_Don\\_e\\_Seu\\_Compertem\\_film\\_c1720](#) (Auto-justificado)  
[Ezra\\_Segul\\_film\\_168284](#) (Auto-justificado)

**Filmes sobre a ditadura ou sobre o golpe militar no Brasil**  
 Os filmes podem ser como pano de fundo o período da ditadura.  
 Documento visualizado: [página/Dois\\_Córregos\\_film\\_cae806](#)

### Dois Córregos (filme)

Dois Córregos é um filme de 1999, dirigido por [Carlos Reichardt](#) e ambientado no município brasileiro de [Dois Córregos](#).

### Sinopse

No início do filme, [Bóli Goulart](#) lembra de sua adolescência quando vai ver sua propriedade em [Dois Córregos](#). Reichardt usa três mulheres diferentes que convivem com um homem que está escondido por perseguição no período da ditadura. As três mulheres são as atrizes [Inga Liberato](#), [Yamara Goulart](#) e [Luciana Brand](#), que também é pianista e interpreta canções na trilha sonora do filme, encabeçada pelo músico e compositor [Ivan Lins](#). O homem é [Carlos Alberto Riccardi](#).

### Premiações

- [Festival Luso Brasileiro: Santa Maria de Feiras](#): Melhor Atriz, [Inga Liberato](#).
- [Festival de Natal de 1999](#): Melhor Filme (Júri Oficial), Melhor Trilha Sonora ([Ivan Lins](#)), Melhor Atriz Condiçante para [Luciana Brand](#) e Melhor Fotografia.
- [VII Festival de Cinema e Vídeo de Curitiba](#): Melhor Filme (Júri Popular), Melhor Direção e Melhor Atriz para [Inga Liberato](#).
- [Prêmio SESAC - Os Melhores De Ano](#): Melhor Filme (Júri Popular) e Melhor Diretor (Prêmio dos Críticos).

Categoria: [Filmes de Carlos Reichardt](#)
 Categoria: [Filmes de 1999](#)
 Categoria: [Filmes do Brasil](#)
 Categoria: [Filmes de drama](#)
 Categoria: [Filmes em língua portuguesa](#)

Voltar às respostas a este tópico

Voltar para página visualizada anteriormente

Tema:

## Sonhos e Desejos

**Sonhos e Desejos** é um filme brasileiro de 2006, do gênero drama, dirigido por [Marcelo Santiago](#). O roteiro é baseado no romance [Astrô de Utopia](#), de [Adriano Caldas](#). O filme foi produzido por [Luiz Carlos Barreto](#) Produções Cinematográficas e distribuído por [Paramount](#) Filmes do Brasil.

### Sinopse

Três milítenes são obrigados a ficar confinados dentro de um apartamento nos anos 70, em [Belo Horizonte](#). Um deles, um bailarino, acaba de chegar ferido e mantém o rosto coberto por um capuz. Ele é recebido pela estudante [Cristiana](#) e seu professor de literatura [Saulão](#), com quem a jovem tem um romance. Durante a convivência, eles discutem e conversam sobre suas vidas, a política e a realidade da época. Com o tempo, a garota se envolve e se apaixoa por [Vaclav](#), o revolucionário, causando sérios problemas com o namorado.

Artigo: [Discussão](#)

**WIKIPÉDIA**  
 A enciclopédia livre.

Página principal  
 Conteúdo destacado

**Sonhos e Desejos**  
 Origem: Wikipédia, a enciclopédia livre.

**Sonhos e Desejos** é um filme brasileiro de 2006, do gênero drama, dirigido por [Marcelo Santiago](#). O roteiro é baseado no romance [Astrô de Utopia](#) de [Adriano Caldas](#). O filme foi produzido por [Luiz Carlos Barreto](#).



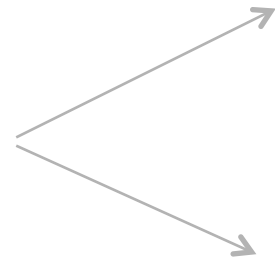
# Answer evaluation process

- 5 “judges”
- Judgment : **correct** / **incorrect** / **doubtful**

- The answer seemed correct, but it wasn't 100% justified
- The page classification wasn't ideal



resposta CORRETA




apropriada


útil

Perspectivas diferentes

Resposta CORRETA e não-apropriada

Onde fica o Taj Mahal?  
- Na 5a Avenida...   
(Voorhees e Tice, 2000)

Resposta CORRETA e não-útil

Quem é o autor de Ivanhoé?  
- O autor de Rob Roy...   
(Spark Jones, 2003)

São respostas “corretas”?

Depende ...  
↓  
NÃO HÁ UMA RESPOSTA ÚNICA CORRETA



# ...mas nem tudo está perdido!

as perguntas/tópicos podem ser decompostas em pedaços

Dinossauros carnívoros que habitaram o Brasil

Resposta: X

- X é dinossauro ✓
- X é carnívoro ✓
- X habitou o Brasil ✓

Se algum dos pedaços não está na página resposta, então deve estar em alguma outra página → **JUSTIFICATIVA**

... e as respostas podem ser avaliadas em pedaços

Filmes sobre as relações entre Portugal e suas colônias

Resposta: X

- X é filme ✓
- X é sobre as relações entre Portugal e suas colônias ✓

Devemos exigir?

Como definir?

- Tema principal?
- Romance?
- Comédia?

Os pedaços também podem ser decompostos em outros pedaços... onde parar?

LogOnto 2014 - FGV - Brasil

Brasil é colônia  
Angola é colônia  
...



# Topic creation

- PT Wikipedia in 150 topics

## PT WIKIPEDIA IN 150 TOPICS

Information needs related to Portuguese-speaking countries and their history, with enough coverage in Wikipedia, and not easily browsable through simple categories or infoboxes, spanning areas from History (50) through Geography (26) and Music (19) to Mathematics (1) and Geology (2). How to assess the answers and their justifications was often quite difficult.



# Tarefa da organização

- Criar um sistema (SIGA) e preparar a coleção
- Identificar tipos de necessidades de informação com resposta na coleção
- Escolher as respostas
- Avaliar as enviadas pelos participantes
- Definir critérios de avaliação
- Definir medidas
- Criar e disponibilizar recursos