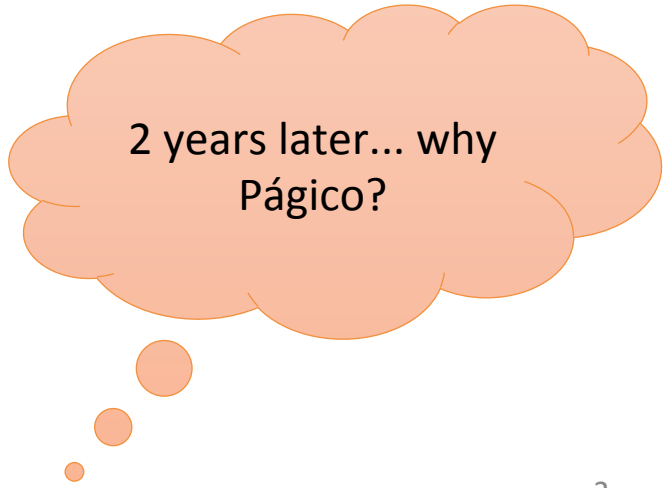
Págico: Evaluating Wikipedia-based information retrieval in Portuguese

Cristina Motta (Linguateca/FCCN)
Alberto Simões (CEHUM/UM)
Cláudia Freitas (PUC-Rio / Linguateca)
Luis Costa (Linguateca/FCCN)
Diana Santos (Univ. of Oslo/Linguateca)

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



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%
$]862, \infty[$	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%
$]33, \infty[$	6	0.001%

Table 8: Number of categories per document.

Topic creation

- Topics should be interesting to the Portuguese-speaking comunity
- The answers
  - should not be ovbious
    - ...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 #	%
Letras	**69**	**46,00**
- história	50	33,33
- literatura	15	10,00
- linguística	6	4,00
- jornalismo	3	2,00
- filosofia	2	1,33
Artes	**36**	**24,00**
- música	19	12,67
- cinema	10	6,67
- televisão	4	2,67
- artes plásticas	2	1,33
- artes	2	1,33
Geografia	**34**	**22,67**
- geografia	26	17,33
- arquitetura/urbanismo	7	4,67
- demografia	4	2,67
- geologia	2	1,33
Cultura	**27**	**18,00**
- antropologia/folclore	12	8,00
- religião	7	4,67
- culinária	5	3,33
- cultura	3	2,00
- ensino	2	1,33
Política	**19**	**12,67**
Desporto/Esportes	**18**	**12,00**
Ciência	**14**	**9,33**
- saúde	4	2,67
- zoologia	3	2,00
- ciência	2	1,33
- botânica	2	1,33
- geologia	2	1,33
- matemática	1	0,67
Economia	**6**	**4,00**

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…

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

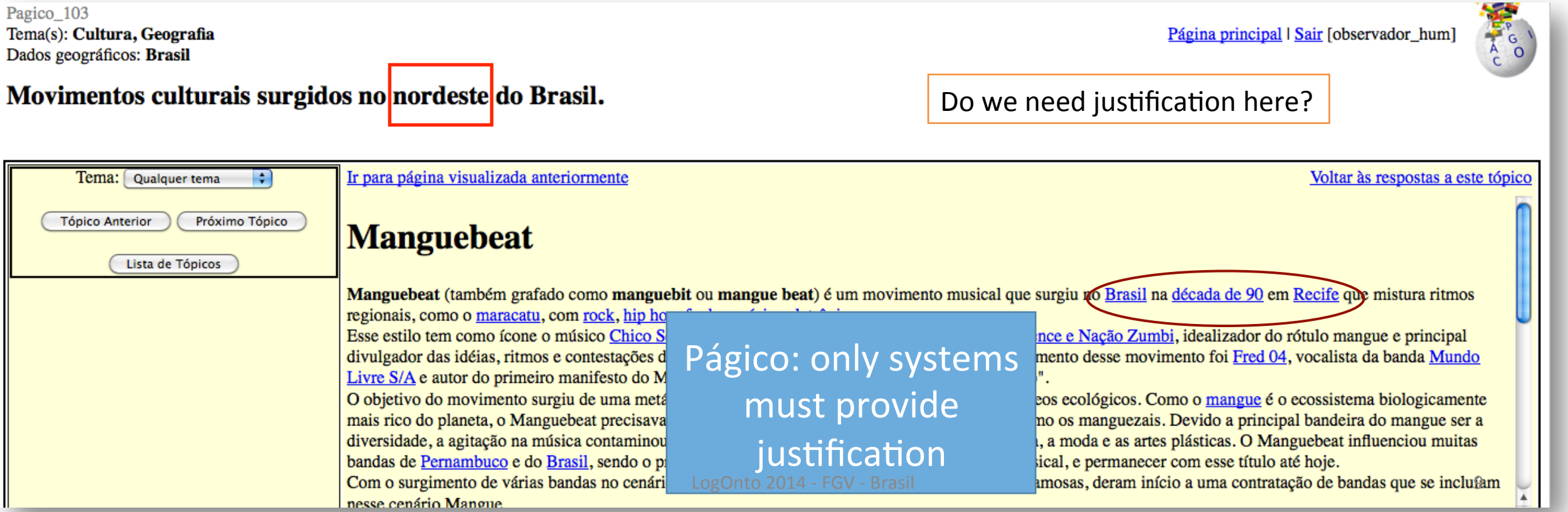
Correct answer
X
Useful answer

Bad answer from a logic perspective

…taking logic into consideration facilitates the evaluation task

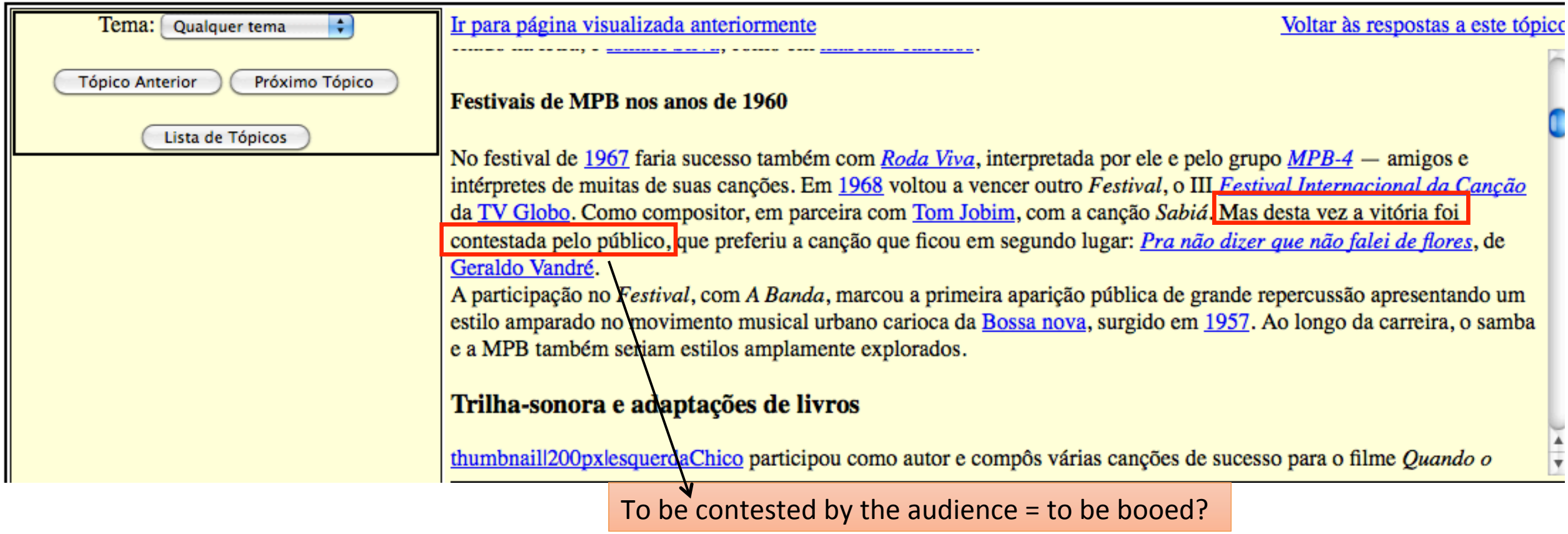
Answer documents – justification pages

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



Answer documents – justification pages

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



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
Viceroys 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

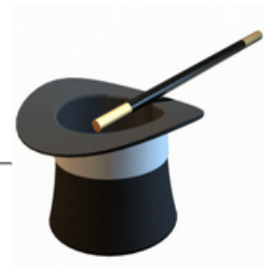
The most correct topics

	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, Linguateca

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

UMIC Agência para a Sociedade do Conhecimento
FCCN Fundação para a Computação Científica Nacional
MCTES MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR
FCT Fundação para a Ciência e a Tecnologia

Mas continuaram a ser apoiados pelas seguintes instituições:

PUC RIO · U · C · Universidade de Coimbra

#[16 Feb 2012] Este ficheiro contÃ©m todos as respostas corretas e justificadas corretamente. NÃ£o inclui as justificaÃ§Ãµes.

```
Pagico_001      pt/1/9/7/1972__filme_.965afb.xml
Pagico_001      pt/a/r/a/Araguaya_-_ConspiraÃ§Ã£o_do_SilÃªncio.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/DedÃ©_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/MemÃ³rias_do_CÃ¡rcere__filme_.d248b4.xml
Pagico_001      pt/o/_/a/O_Ano_em_que_Meus_Pais_SaÃram_de_FÃ©rias.b58539.xml
Pagico_001      pt/o/_/b/O_Bom_BurguÃªs.4c5c73.xml
Pagico_001      pt/o/_/q/O_Que_Ã%_Isso,_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_IrmÃ£os.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_CoraÃ§Ã£o.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_CoraÃ§Ã£o.045ff2.xml
```

Pagico_001: FILMES SOBRE A DITADURA OU SOBRE O GOLPE MILITAR NO BRASIL

Pagico_002: TELENOVELAS BRASILEIRAS DO GÊNERO REALISMO FANTÁSTICO

```
#[16 Feb 2012] Este ficheiro contÃ©m todas as respostas corretas e justificadas corretamente, bem como as suas justificaÃ§Ãµes.

Pagico_001      pt/1/9/7/1972__filme_.965afb.xml pt/a/n/e/Anexo_Lista_de_filmes_sobre_ditaduras_militares_na_AmÃ©rica_Latina.605516.xml
Pagico_001      pt/a/r/a/Araguaya_-_ConspiraÃ§Ã£o_do_SilÃªncio.a41d5e.xml pt/a/n/e/Anexo_Lista_de_filmes_sobre_ditaduras_militares_na_AmÃ©rica_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_AmÃ©rica_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_AmÃ©rica_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_AmÃ©rica_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_AmÃ©rica_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/DedÃ©_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_AmÃ©rica_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_AmÃ©rica_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_AmÃ©rica_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_AmÃ©rica_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/MemÃ³rias_do_CÃ¡rcere__filme_.d248b4.xml
Pagico_001      pt/o/_/a/O_Ano_em_que_Meus_Pais_SaÃram_de_FÃ©rias.b58539.xml
Pagico_001      pt/o/_/a/O_Ano_em_que_Meus_Pais_SaÃram_de_FÃ©rias.b58539.xml pt/a/n/e/Anexo_Lista_de_filmes_sobre_ditaduras_militares_na_AmÃ©rica_Lat
Pagico_001      pt/o/_/b/O_Bom_BurguÃªs.4c5c73.xml
Pagico_001      pt/o/_/q/O_Que_Ã%_Isso,_Companheiro___filme_.ec72f9.xml
Pagico_001      pt/o/_/q/O_Que_Ã%_Isso,_Companheiro___filme_.ec72f9.xml pt/a/n/e/Anexo_Lista_de_filmes_sobre_ditaduras_militares_na_AmÃ©rica_Latina.6
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_AmÃ©rica_Latina.605516.xml
Pagico_001      pt/q/u/a/Quase_Dois_IrmÃ£os.c80d70.xml pt/a/n/e/Anexo_Lista_de_filmes_sobre_ditaduras_militares_na_AmÃ©rica_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_AmÃ©rica_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_AmÃ©rica_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_AmÃ©rica_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 Págico foram financiados pelas seguintes entidades até 31 de Dezembro de 2011.

UMIC Agência para a Sociedade do Conhecimento

FCCN Fundação para a Computação Científica Nacional

MCTES MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

FCT Fundação para a Ciência e a Tecnologia

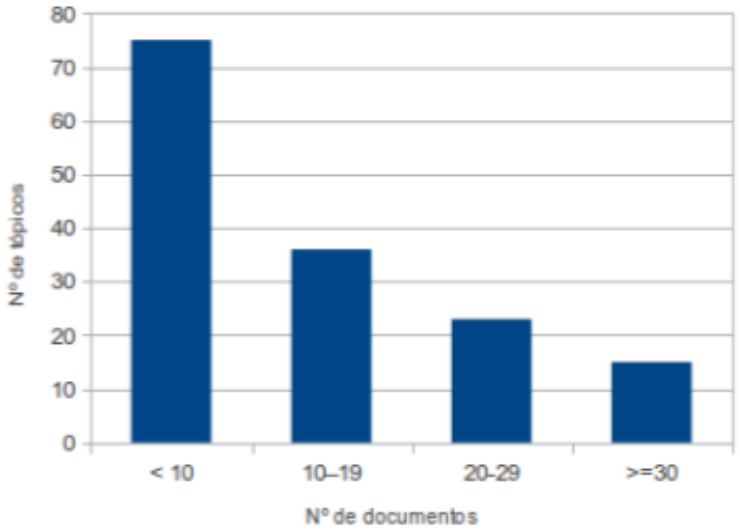
Mas continuaram a ser apoiados pelas seguintes instituições:

PUC RIO · U · C · Universidade de Coimbra

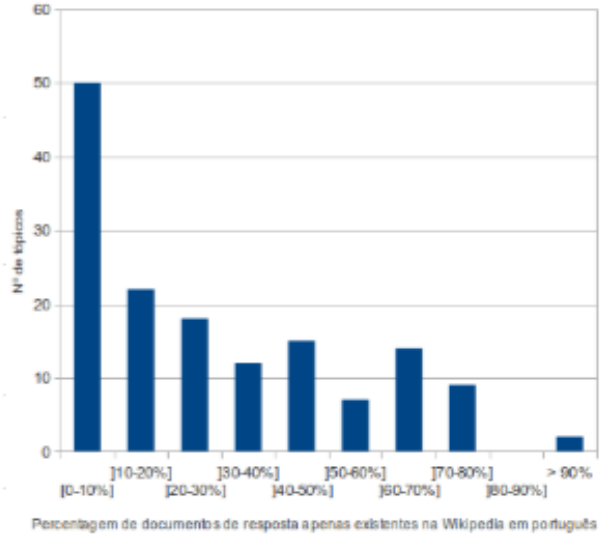
CARTOLA http://www.linguateca.pt/Cartola

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

Tópico Anterior | Próximo Tópico

Lista de Tópicos

Resposta	Justificação	Correta?	Justificada?	
Sonhos_e_Desejos	Sem Justificações	Sim	Sim	Corrigir avaliação
Lamarca	Sem Justificações	Sim	Sim	Corrigir avaliação
Quase_Dois_Irmãos	Sem Justificações	Sim	Sim	Corrigir avaliação
As_Meninas_filme_	Sem Justificações	Sim	Sim	Corrigir avaliação
Cabra_Marcado_para_Morrer	Sem Justificações	Sim	Sim	Corrigir avaliação
Brazil,_a_Report_on_Torture	Sem Justificações	Sim	Sim	Corrigir avaliação
Lamarca	Sem Justificações	Sim	Sim	Corrigir avaliação
Dedé_Mamata	Sem Justificações	Sim	Sim	Corrigir avaliação
O_Ano_em_que_Meus_Pais_Saíram_de_Férias	Sem Justificações	Sim	Sim	Corrigir avaliação
Cabra-cega_filme_	Sem Justificações	Sim	Sim	Corrigir avaliação
O_Bom_Burguês	Sem Justificações	Sim	Sim	Corrigir avaliação
Marighella_-_Retrato_Falado_do_Guerrilheiro	Sem Justificações	Sim	Sim	Corrigir avaliação
Os_Herdeiros	Sem Justificações	Sim	Sim	Corrigir avaliação
Zuzu_Angel_filme_	Sem Justificações	Sim	Sim	Corrigir avaliação
O_Que_É_Isso,_Companheiro_filme_	Sem Justificações	Sim	Sim	Corrigir avaliação
Jango_filme_	Sem Justificações	Sim	Sim	Corrigir avaliação
Batismo_de_Sangue	Sem Justificações	Sim	Sim	Corrigir avaliação

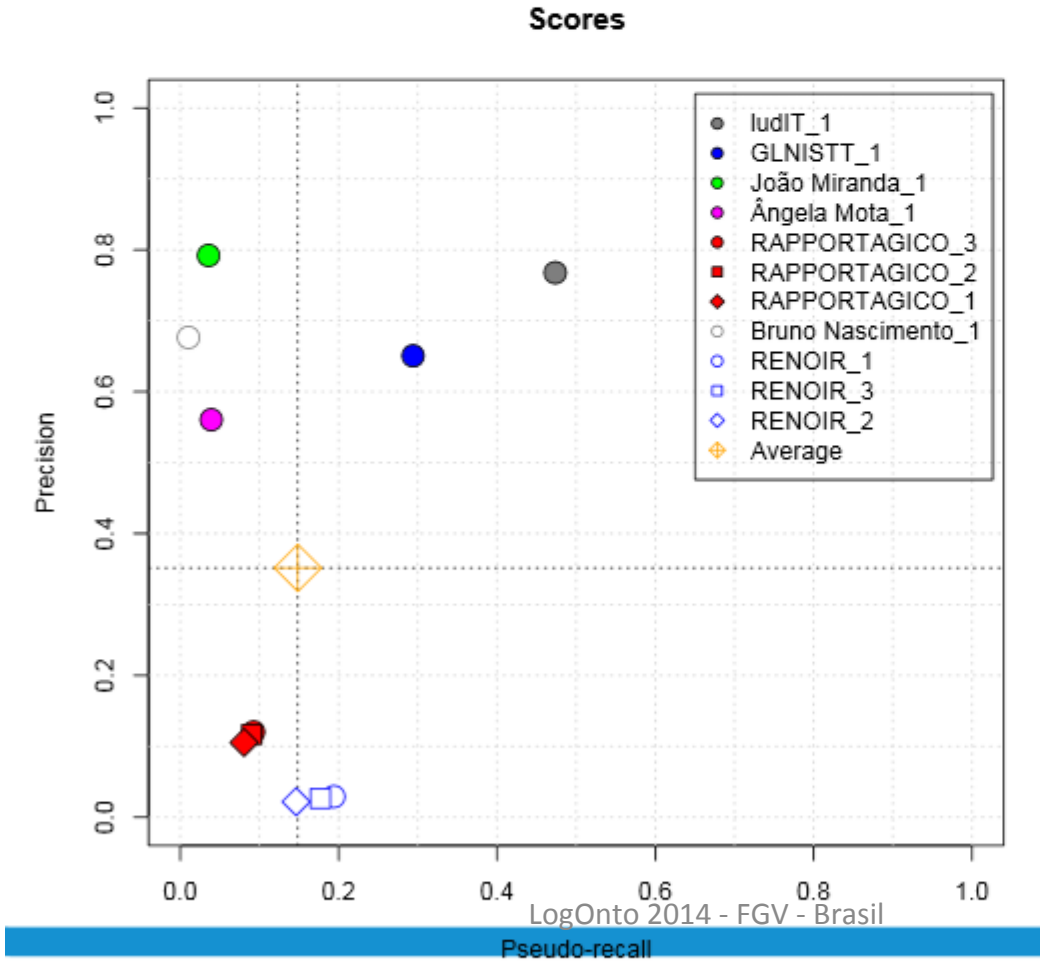
PÁGICO COLLECTION

- 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;

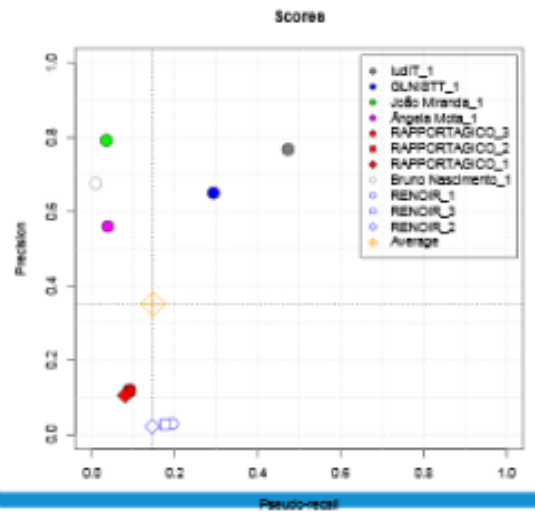
Collection constitution:

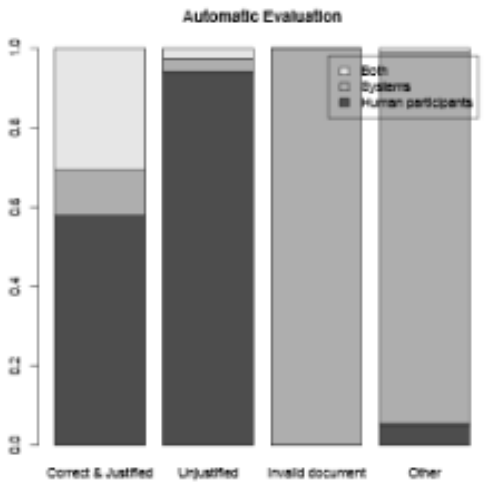
Page type	Total docs
Template pages	32 900
Disambiguation pages	5 006
Redirection pages	574 077
Multimedia pages	9 678
Article pages	856 005

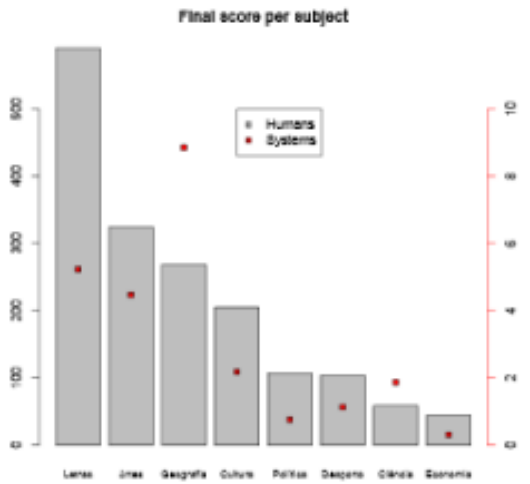
HUMANS VS. SYSTEMS

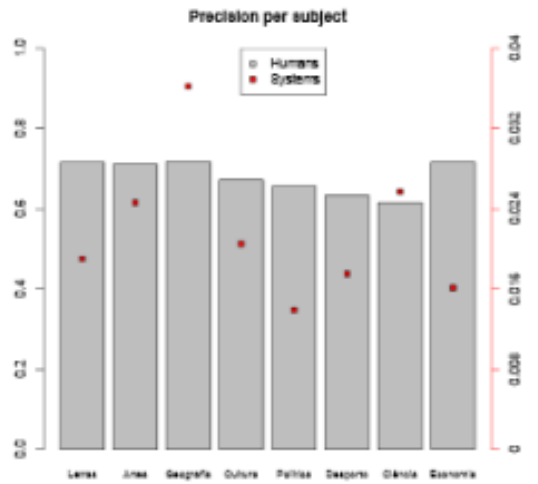


HUMANS VS. SYSTEMS

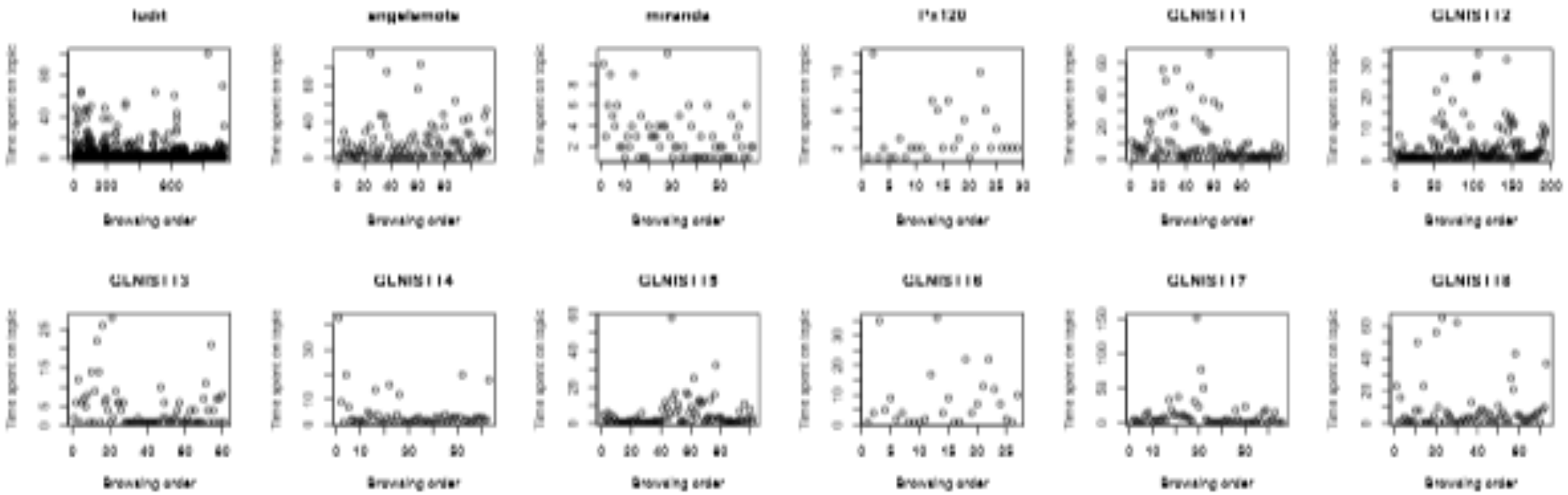


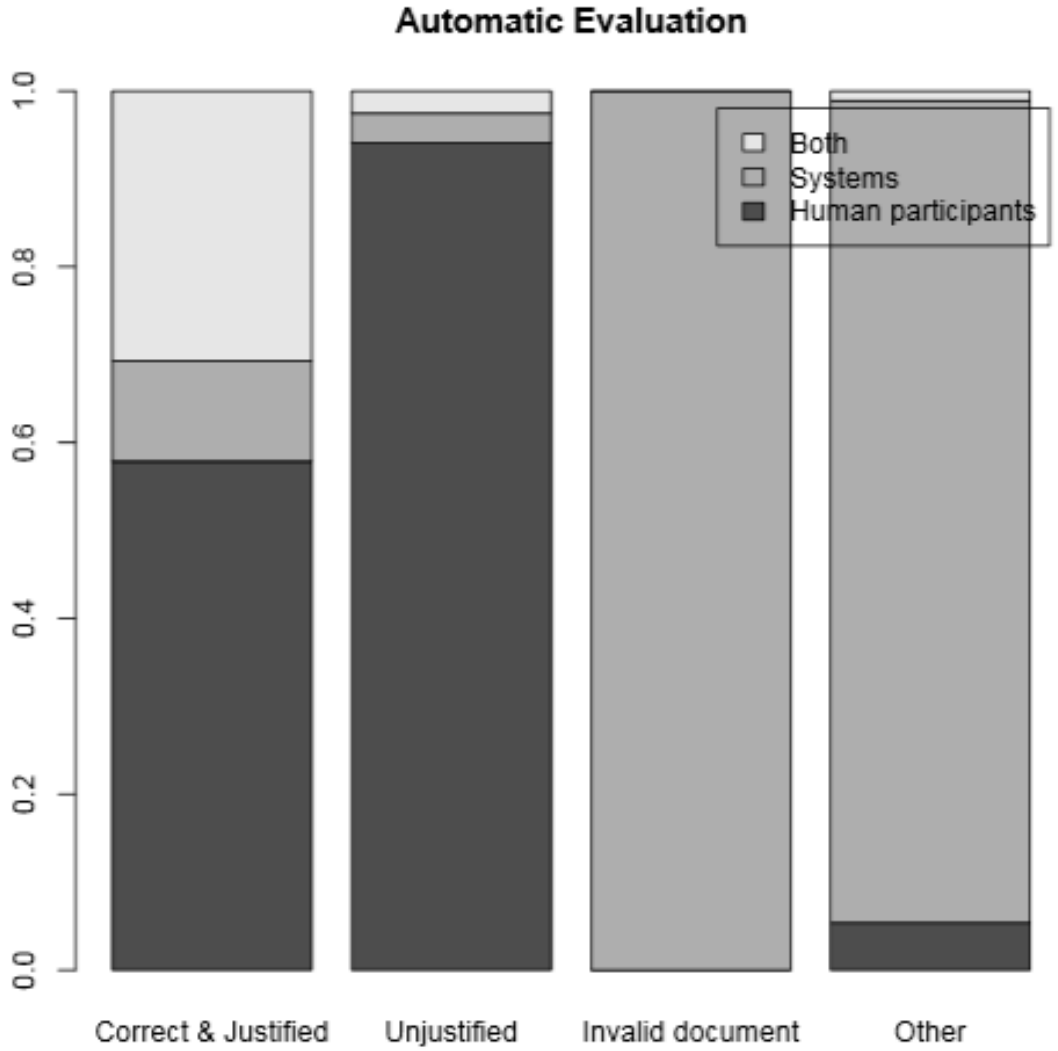


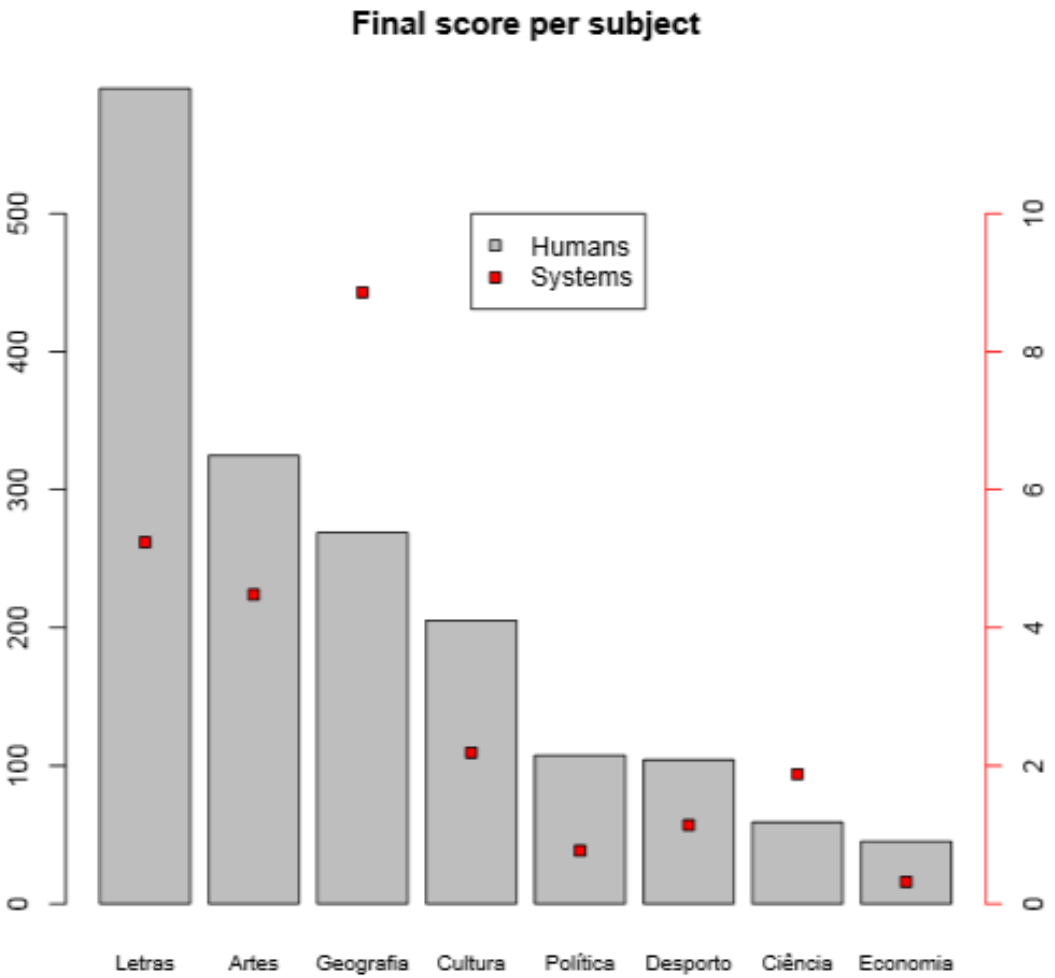


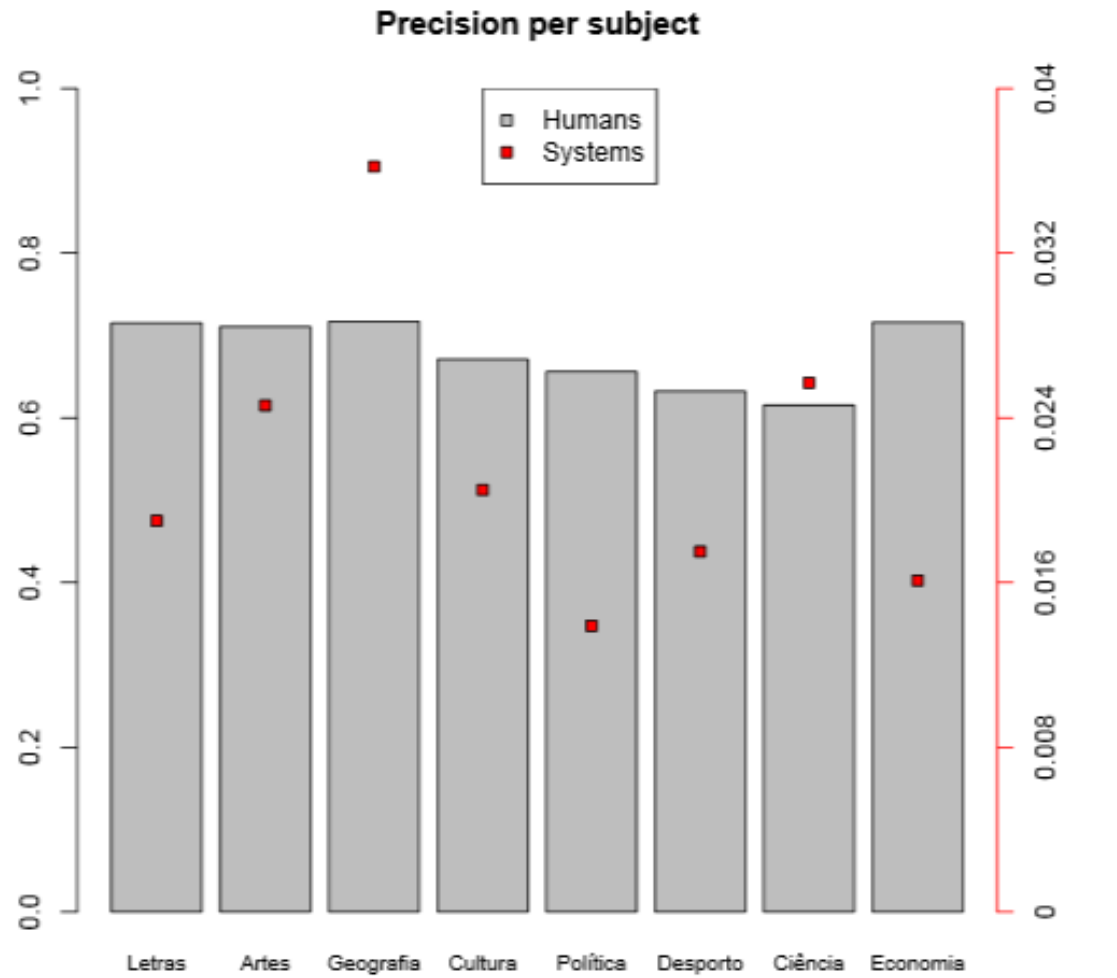


USER BROWSER BEHAVIOUR




Automatic Evaluation
1.0
0.8
0.6
0.4
0.2
0.0
Both
Systems
Human participants
Correct & Justified
Unjustified
Invalid document
Other


Final score per subject
Humans
Systems
0
100
200
300
400
500
0
2
4
6
8
10
Letras
Artes
Geografia
Cultura
Política
Desporto
Ciência
Economia


Precision per subject
Humans
Systems
0.0
0.2
0.4
0.6
0.8
1.0
0
0.008
0.016
0.024
0.032
0.04
Letras
Artes
Geografia
Cultura
Política
Desporto
Ciência
Economia

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

Answer evaluation process

- 5 “judges”
- Judgmenet : 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?

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