Evaluation

Anisio Lacerda

Evaluation

- Evaluation is key to building effective and efficient search engines
  - measurement usually carried out in controlled laboratory experiments
  - online testing can also be done

Evaluation Corpus

- *Test collections* consisting of documents, queries, and relevance judgments, e.g.
  - **CACM**: Titles and abstracts from the Communications of the ACM from 1958-1979. Queries and relevance judgments generated by computer scientists.
  - **AP**: Associated Press newswire documents from 1988-1990 (from TREC disks 1-3). Queries are the title fields from TREC topics 51-150. Topics and relevance judgments generated by government information analysts.
  - **GOV2**: Web pages crawled from websites in the .gov domain during early 2004. Queries are the title fields from TREC topics 701-850. Topics and relevance judgments generated by government analysts.

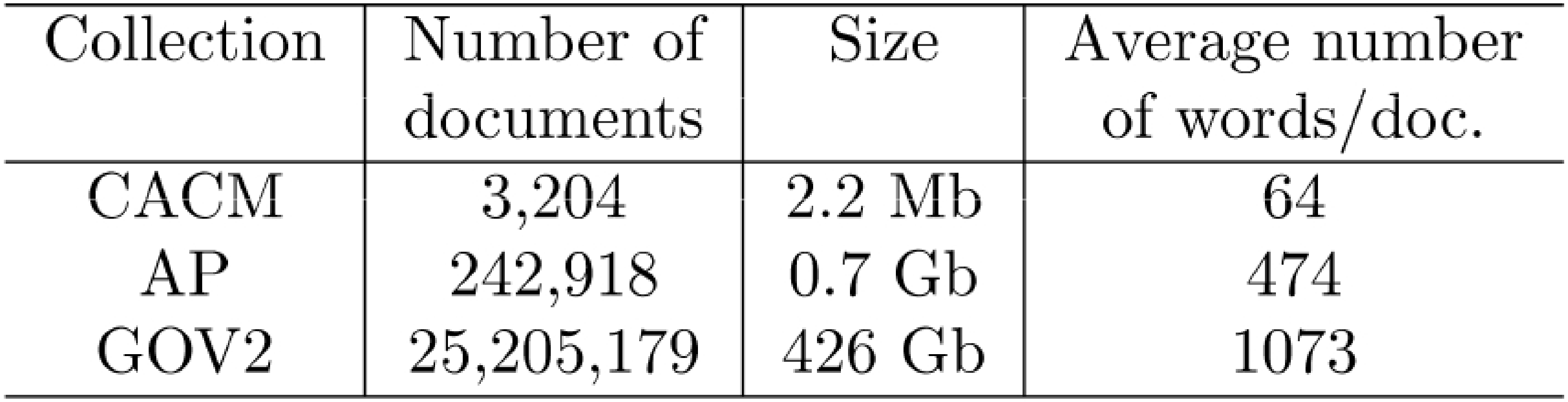
Evaluation Corpus

- *Test collections* consisting of documents, queries, and relevance judgments, e.g.
  - **CACM**: Titles and abstracts from the Communications of the ACM from 1958-1979. Queries and relevance judgments generated by computer scientists.
  - **AP**: Associated Press newswire documents from 1988-1990 (from TREC disks 1-3). Queries are the title fields from TREC topics 51-150. Topics and relevance judgments generated by government information analysts.
  - **GOV2**: Web pages crawled from websites in the .gov domain during early 2004. Queries are the title fields from TREC topics 701-850. Topics and relevance judgments generated by government analysts.

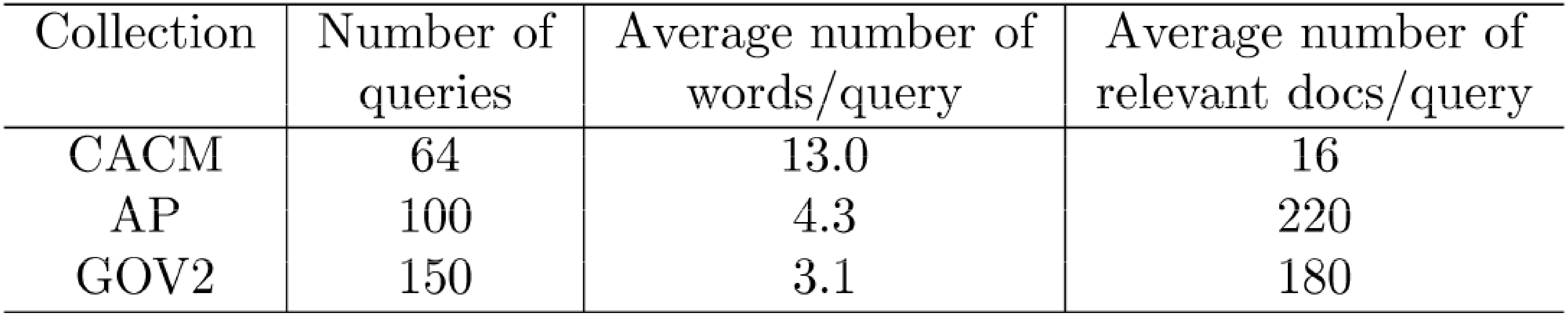
Evaluation Corpus

- *Test collections* consisting of documents, queries, and relevance judgments, e.g.
  - **CACM**: Titles and abstracts from the Communications of the ACM from 1958-1979. Queries and relevance judgments generated by computer scientists.
  - **AP**: Associated Press newswire documents from 1988-1990 (from TREC disks 1-3). Queries are the title fields from TREC topics 51-150. Topics and relevance judgments generated by government information analysts.
  - **GOV2**: Web pages crawled from websites in the .gov domain during early 2004. Queries are the title fields from TREC topics 701-850. Topics and relevance judgments generated by government analysts.

Test Collections

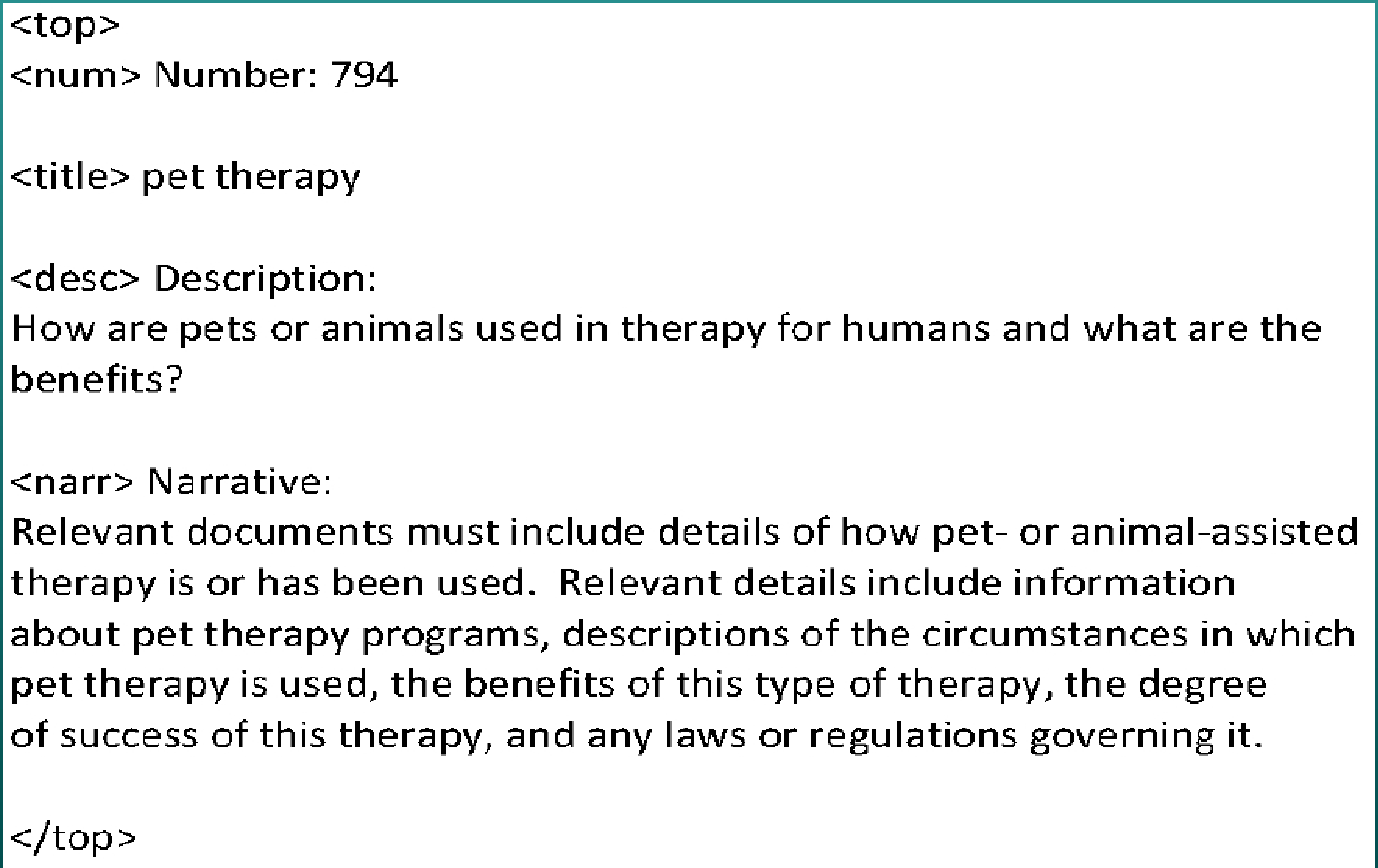


Collection	Number of documents	Size	Average number of words/doc.
CACM	3,204	2.2 Mb	64
AP	242,918	0.7 Gb	474
GOV2	25,205,179	426 Gb	1073



Collection	Number of queries	Average number of words/query	Average number of relevant docs/query
CACM	64	13.0	16
AP	100	4.3	220
GOV2	150	3.1	180

TREC Topic Example



```
<top>
<num> Number: 794

<title> pet therapy

<desc> Description:
How are pets or animals used in therapy for humans and what are the
benefits?

<narr> Narrative:
Relevant documents must include details of how pet- or animal-assisted
therapy is or has been used.  Relevant details include information
about pet therapy programs, descriptions of the circumstances in which
pet therapy is used, the benefits of this type of therapy, the degree
of success of this therapy, and any laws or regulations governing it.

</top>
```

Relevance Judgments

- Obtaining relevance judgments is an expensive, time-consuming process
  - who does it?
  - what are the instructions?
  - what is the level of agreement?
- TREC judgments
  - generally binary
  - agreement good because of “narrative”

Relevance Judgments

- Obtaining relevance judgments is an expensive, time-consuming process
  - who does it?
  - what are the instructions?
  - what is the level of agreement?
- TREC judgments
  - generally binary
  - agreement good because of “narrative”

Pooling

- Exhaustive judgments for all documents in a collection is not practical
- Pooling technique is used in TREC
  - top *k* (for TREC, k varied between 50 and 200) from the rankings obtained by different search engines (or retrieval algorithms) are merged into a pool
  - duplicates are removed
  - documents are presented in some random order to the relevance judges
- Produces a large number of relevance judgments for each query, although still incomplete

Pooling

- Exhaustive judgments for all documents in a collection is not practical
- Pooling technique is used in TREC
  - top *k* (for TREC, k varied between 50 and 200) from the rankings obtained by different search engines (or retrieval algorithms) are merged into a pool
  - duplicates are removed
  - documents are presented in some random order to the relevance judges
- Produces a large number of relevance judgments for each query, although still incomplete

Pooling

- Exhaustive judgments for all documents in a collection is not practical
- Pooling technique is used in TREC
  - top *k* (for TREC, k varied between 50 and 200) from the rankings obtained by different search engines (or retrieval algorithms) are merged into a pool
  - duplicates are removed
  - documents are presented in some random order to the relevance judges
- Produces a large number of relevance judgments for each query, although still incomplete

Can we avoid human judgments?

- No
- Makes experimental work hard
  - Especially on a large scale

Evaluating at large search engines

- Search engines have test collections of queries and hand-ranked results
- Search engines often use precision at top k, e.g., k = 10
- Seach engines also use non-relevance-based measures.
  - Clickthrough on first result
    - Not very reliable if you look at a single clickthrough... but pretty reliable in the aggregate.
  - Studies on user behavior in the lab

Evaluating at large search engines

- Search engines have test collections of queries and hand-ranked results
- Search engines often use precision at top k, e.g., k = 10
- Seach engines also use non-relevance-based measures.
  - Clickthrough on first result
    - Not very reliable if you look at a single clickthrough... but pretty reliable in the aggregate.
  - Studies on user behavior in the lab

Evaluating at large search engines

- Search engines have test collections of queries and hand-ranked results
- Search engines often use precision at top k, e.g., k = 10
- Seach engines also use non-relevance-based measures.
  - Clickthrough on first result
    - Not very reliable if you look at a single clickthrough... but pretty reliable in the aggregate.
  - Studies on user behavior in the lab

A/B testing

- Purpose: Test a single innovation
- Prerequisite: You have a large search engine up and running.
- Have most users use old system

A/B testing

- Divert a small proportion of traffic (e.g., 1%) to the new system that includes the innovation
- Evaluate with an “automatic” measure like clickthrough on first result
- Now we can directly see if the innovation does improve user happiness.
- Probably the evaluation methodology that large search engines trust most

A/B testing

- Divert a small proportion of traffic (e.g., 1%) to the new system that includes the innovation
- Evaluate with an “automatic” measure like clickthrough on first result
- Now we can directly see if the innovation does improve user happiness.
- Probably the evaluation methodology that large search engines trust most

A/B testing

- Divert a small proportion of traffic (e.g., 1%) to the new system that includes the innovation
- Evaluate with an “automatic” measure like clickthrough on first result
- Now we can directly see if the innovation does improve user happiness.
- Probably the evaluation methodology that large search engines trust most

A/B testing

- Divert a small proportion of traffic (e.g., 1%) to the new system that includes the innovation
- Evaluate with an “automatic” measure like clickthrough on first result
- Now we can directly see if the innovation does improve user happiness.
- Probably the evaluation methodology that large search engines trust most

Query Logs

- Used for both tuning and evaluating search engines
  - also for various techniques such as query suggestion
- Typical contents
  - User identifier or user session identifier
  - Query terms – stored exactly as user entered
  - List of URLs of results, their ranks on the result list, and whether they were clicked on
  - Timestamp(s) – records the time of user events such as query submission, clicks

Query Logs

- Used for both tuning and evaluating search engines
  - also for various techniques such as query suggestion
- Typical contents
  - User identifier or user session identifier
  - Query terms – stored exactly as user entered
  - List of URLs of results, their ranks on the result list, and whether they were clicked on
  - Timestamp(s) – records the time of user events such as query submission, clicks

Query Logs

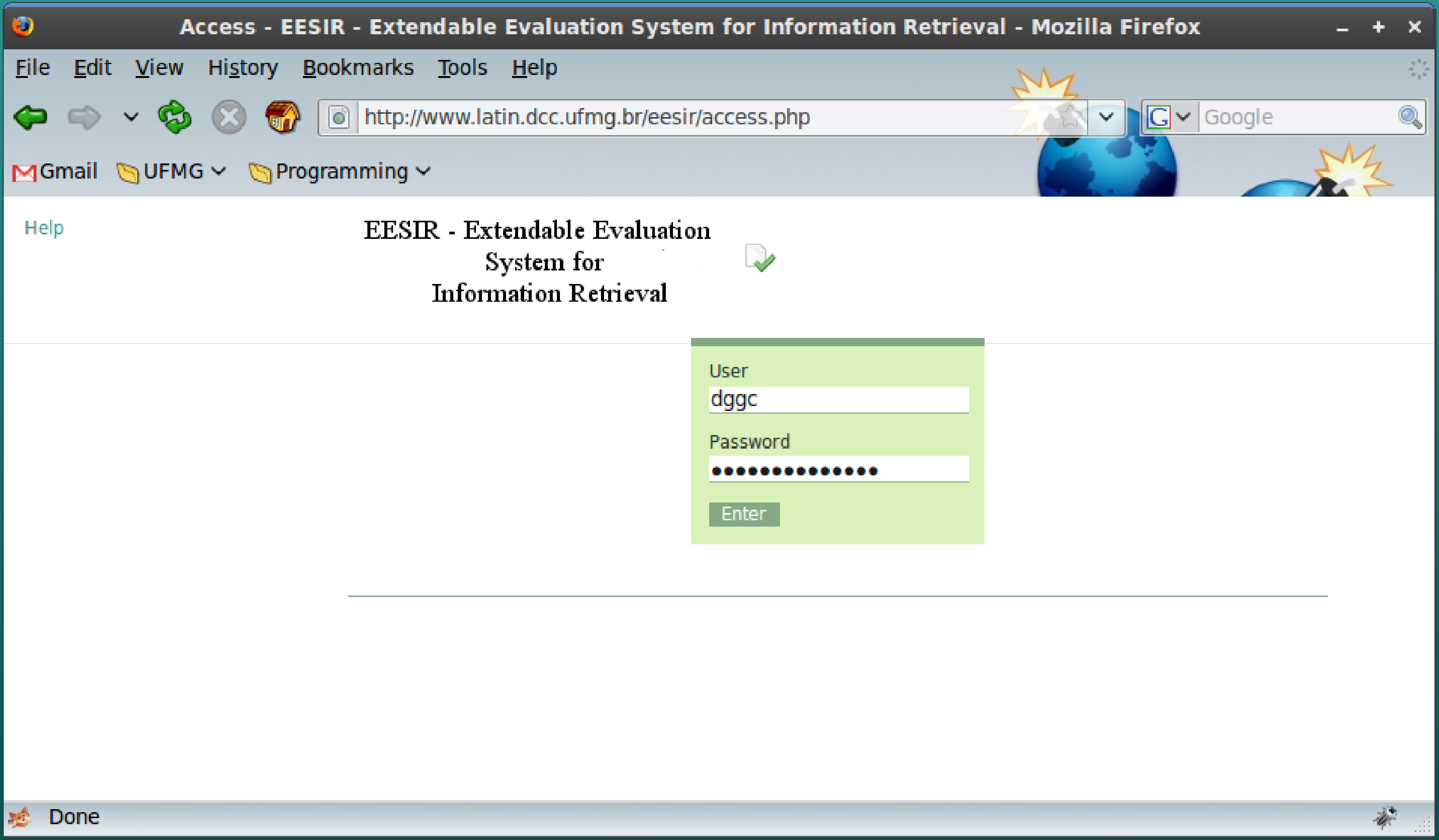
- Clicks are not relevance judgments
  - although they are correlated
  - biased by a number of factors such as rank on result list
- Can use clickthough data to predict *preferences* between pairs of documents
  - appropriate for tasks with multiple levels of relevance, focused on user relevance
  - various “policies” used to generate preferences

Query Logs

- Clicks are not relevance judgments
  - although they are correlated
  - biased by a number of factors such as rank on result list
- Can use clickthough data to predict *preferences* between pairs of documents
  - appropriate for tasks with multiple levels of relevance, focused on user relevance
  - various “policies” used to generate preferences

EESIR – Extendable Evaluation System for Information Retrieval

Login



Access - EESIR - Extendable Evaluation System for Information Retrieval - Mozilla Firefox

EESIR - Extendable Evaluation System for Information Retrieval

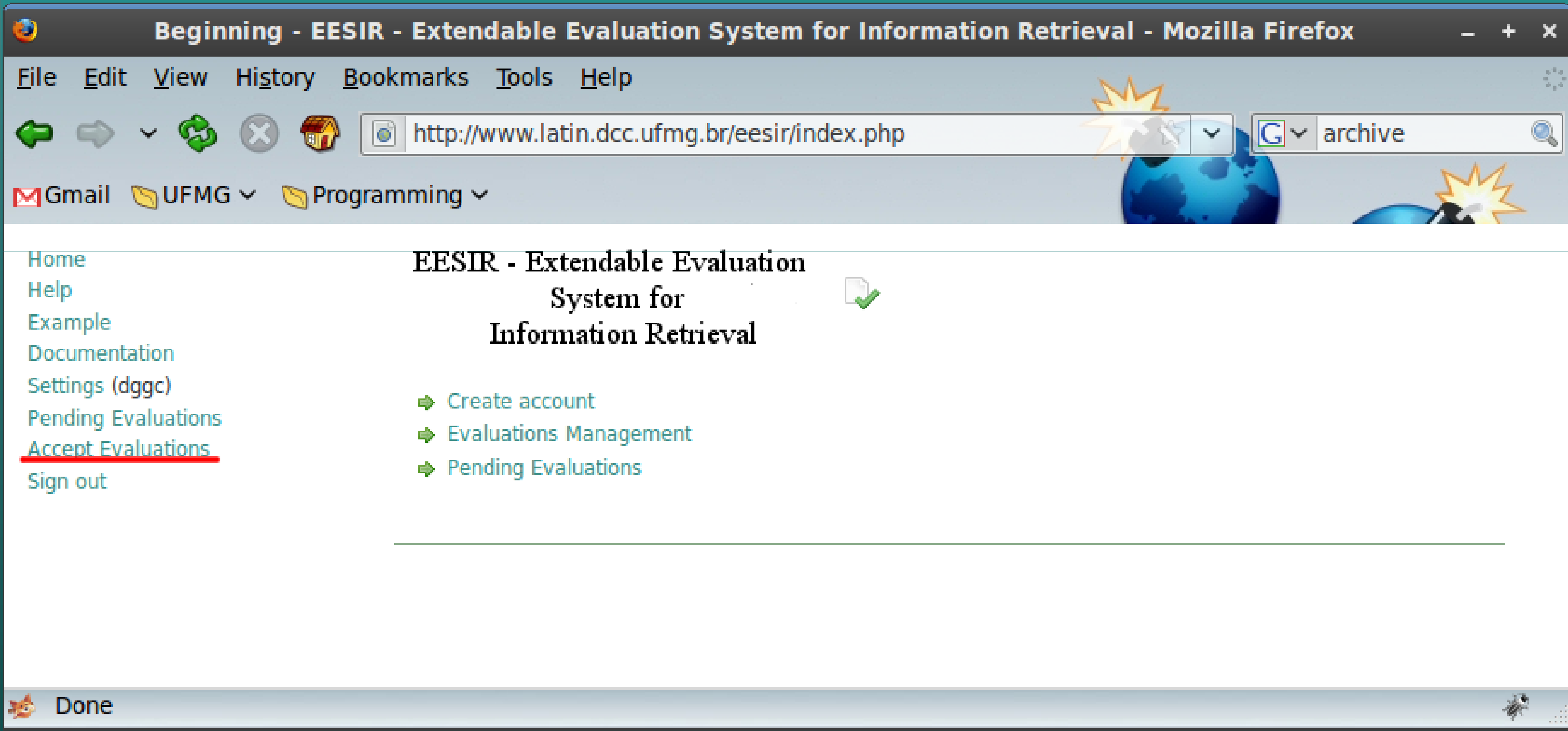
User

dggc

Password

●●●●●●●●●●●●●●

Accept



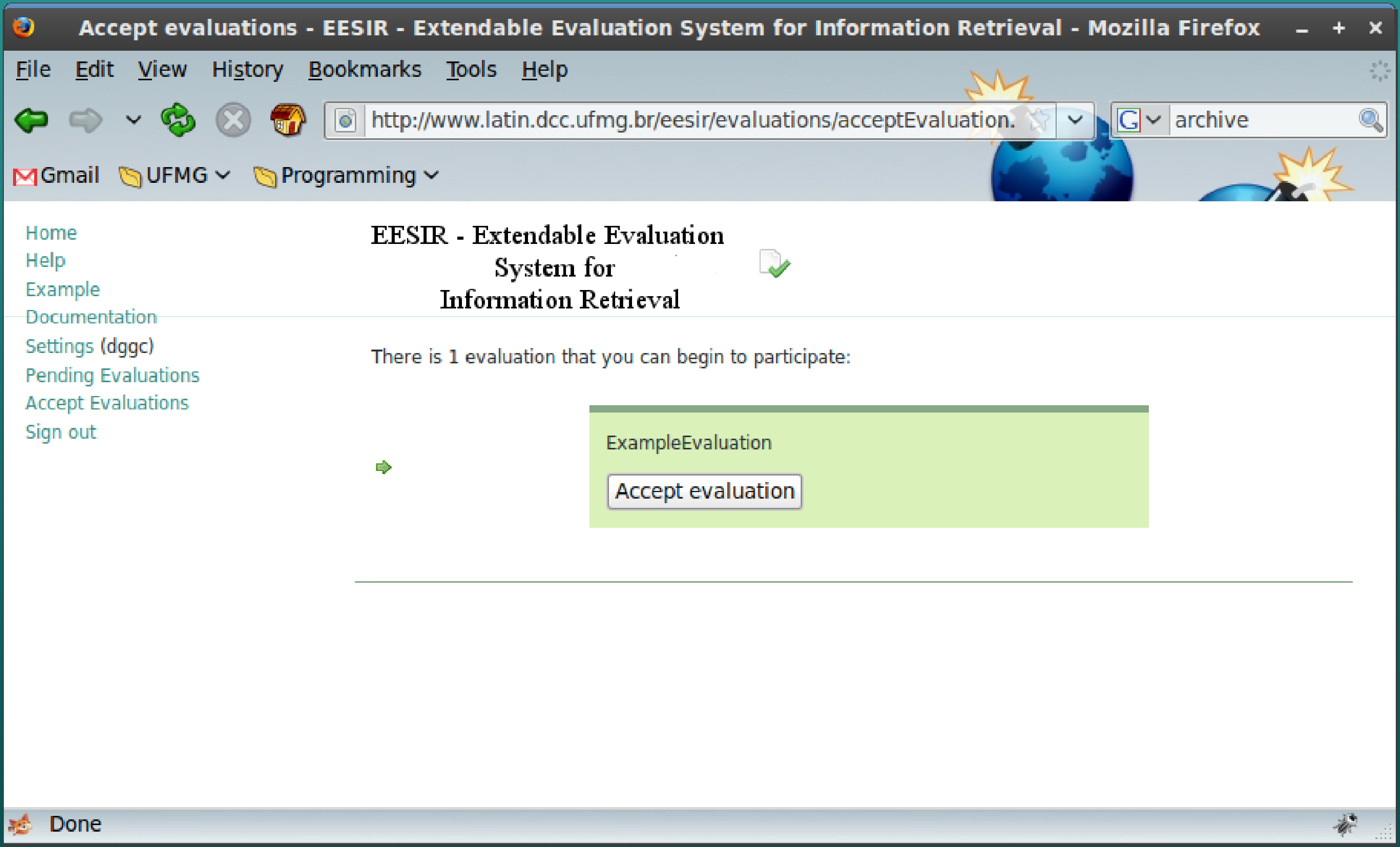
Beginning - EESIR - Extendable Evaluation System for Information Retrieval - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.latin.dcc.ufmg.br/eesir/index.php
archive
Gmail
UFMG
Programming
Done

Home
Help
Example
Documentation
Settings (dggc)
Pending Evaluations
Accept Evaluations
Sign out

EESIR - Extendable Evaluation System for Information Retrieval

- Create account
- Evaluations Management
- Pending Evaluations

Accept



Accept evaluations - EESIR - Extendable Evaluation System for Information Retrieval - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.latin.dcc.ufmg.br/eesir/evaluations/acceptEvaluation.
archive
Gmail
UFMG
Programming
Home
Help
Example
Documentation
Settings (dggc)
Pending Evaluations
Accept Evaluations
Sign out
Done

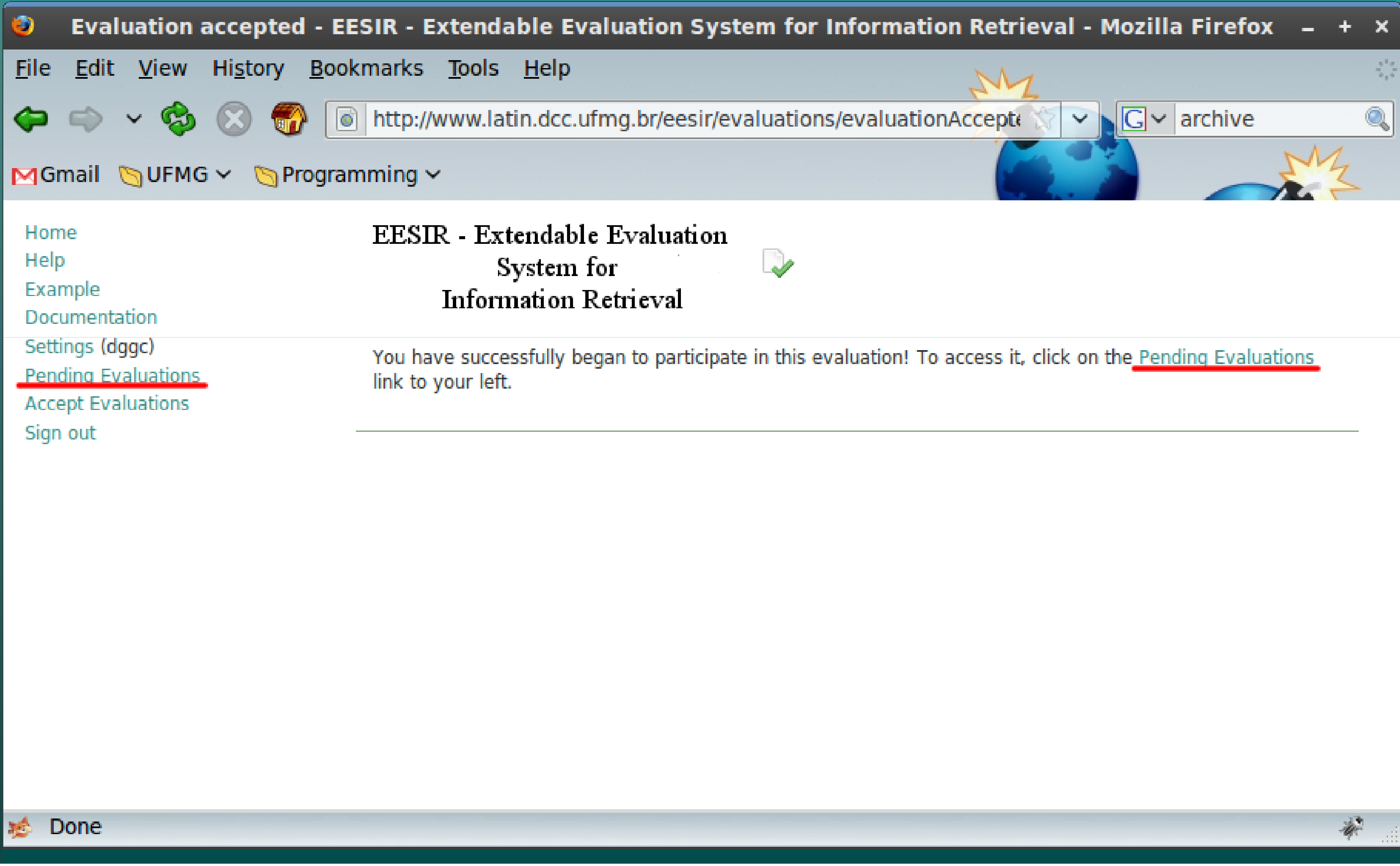
EESIR - Extendable Evaluation System for Information Retrieval

There is 1 evaluation that you can begin to participate:

ExampleEvaluation

Accept evaluation

Pending evaluations

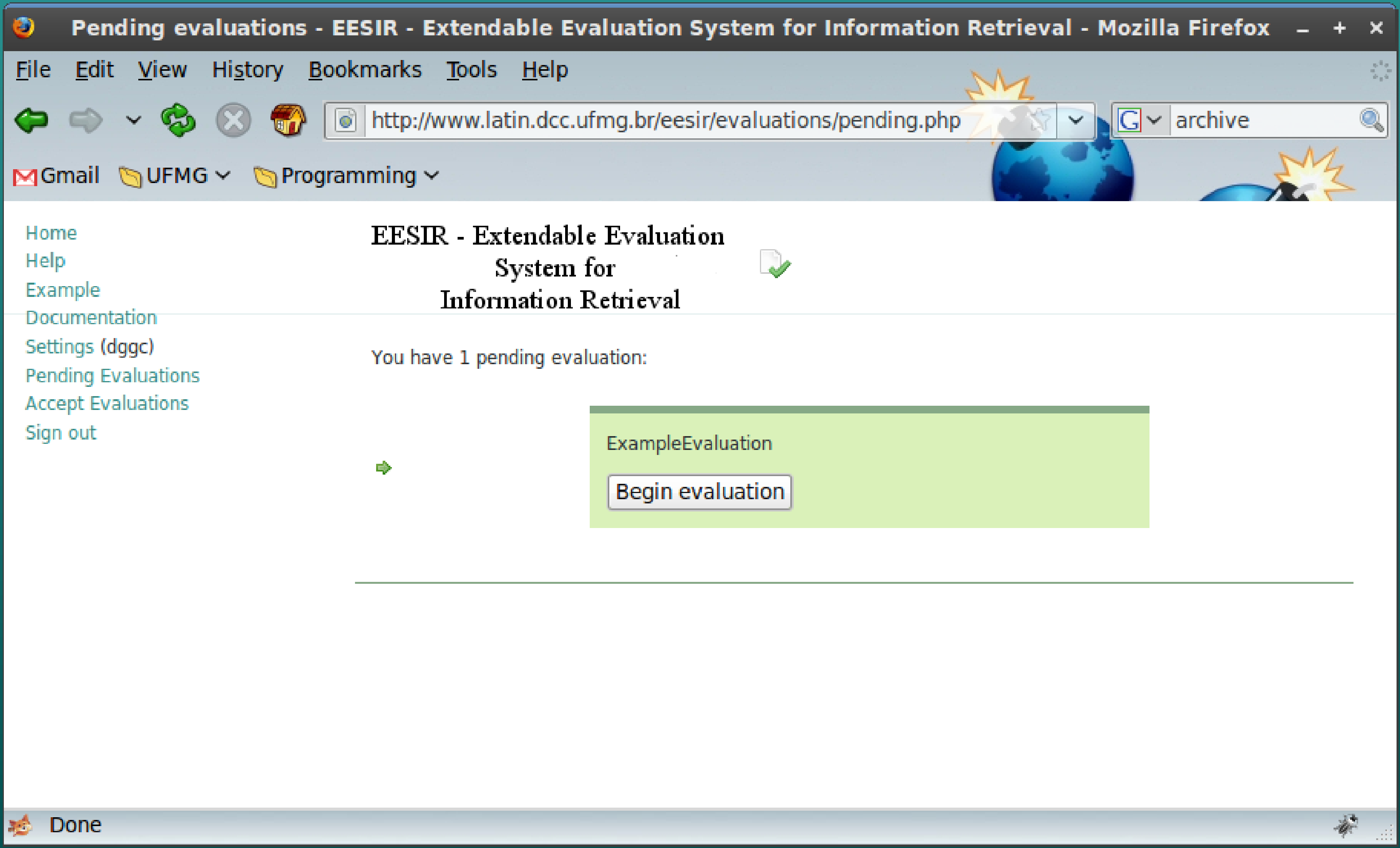


Evaluation accepted - EESIR - Extendable Evaluation System for Information Retrieval - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.latin.dcc.ufmg.br/eesir/evaluations/evaluationAccepte
archive
Gmail
UFMG
Programming
Home
Help
Example
Documentation
Settings (dggc)
Pending Evaluations
Accept Evaluations
Sign out
Done

EESIR - Extendable Evaluation System for Information Retrieval

You have successfully began to participate in this evaluation! To access it, click on the Pending Evaluations link to your left.

Begin



Pending evaluations - EESIR - Extendable Evaluation System for Information Retrieval - Mozilla Firefox

File Edit View History Bookmarks Tools Help
http://www.latin.dcc.ufmg.br/eesir/evaluations/pending.php
archive
Gmail
UFMG
Programming

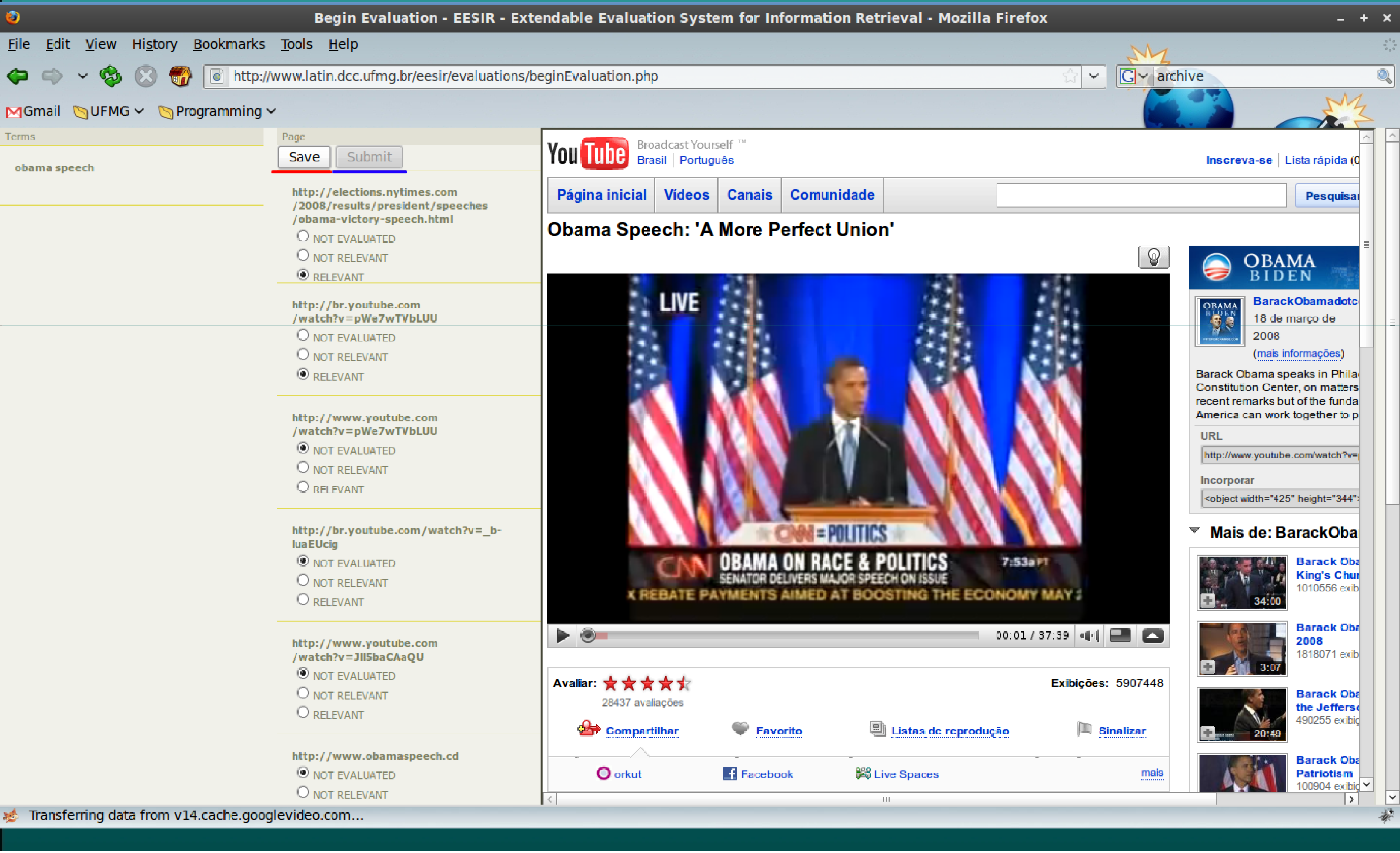
- Home
- Help
- Example
- Documentation
- Settings (dggc)
- Pending Evaluations
- Accept Evaluations
- Sign out

EESIR - Extendable Evaluation System for Information Retrieval

You have 1 pending evaluation:

ExampleEvaluation

Let's



Begin Evaluation - EESIR - Extendable Evaluation System for Information Retrieval - Mozilla Firefox

File Edit View History Bookmarks Tools Help
http://www.latin.dcc.ufmg.br/eesir/evaluations/beginEvaluation.php
archive
Gmail UFMG Programming

Terms

obama speech

Page

Save Submit

http://elections.nytimes.com/2008/results/president/speeches/obama-victory-speech.html
○ NOT EVALUATED
○ NOT RELEVANT
● RELEVANT

http://br.youtube.com/watch?v=pWe7wTVbLUU
○ NOT EVALUATED
○ NOT RELEVANT
● RELEVANT

http://www.youtube.com/watch?v=pWe7wTVbLUU
● NOT EVALUATED
○ NOT RELEVANT
○ RELEVANT

http://br.youtube.com/watch?v=_b-luaEUcig
● NOT EVALUATED
○ NOT RELEVANT
○ RELEVANT

http://www.youtube.com/watch?v=Jll5baCAaQU
● NOT EVALUATED
○ NOT RELEVANT
○ RELEVANT

http://www.obamaspeech.cd
● NOT EVALUATED
○ NOT RELEVANT

Obama Speech: 'A More Perfect Union'

LIVE
CNN POLITICS
CNN OBAMA ON RACE & POLITICS
SENATOR DELIVERS MAJOR SPEECH ON ISSUE
7:53a PT
X REBATE PAYMENTS AIMED AT BOOSTING THE ECONOMY MAY
00:01 / 37:39
OBAMA BIDEN

BarackObamadotc
18 de março de 2008
(mais informações)

Barack Obama speaks in Phila
Constitution Center, on matters
recent remarks but of the funda
America can work together to p

URL
http://www.youtube.com/watch?v=

Incorporar
<object width="425" height="344":

Mais de: BarackOba

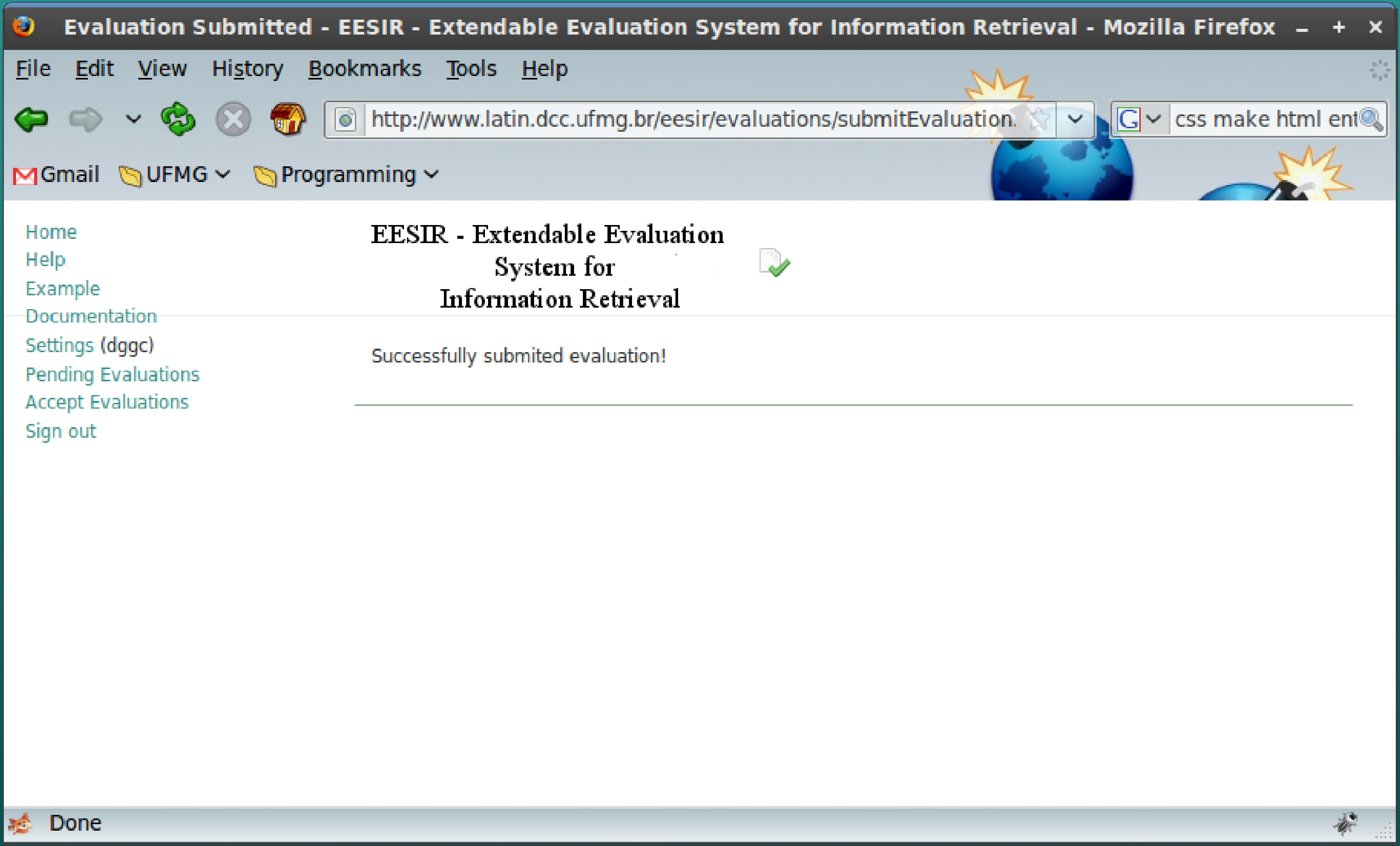
Barack Oba King's Chur 34:00 1010556 exib
Barack Oba 2008 3:07 1818071 exib
Barack Oba the Jeffers 20:49 490255 exibiç
Barack Oba Patriotism 100904 exibiç

Avaliar: 28437 avaliações

Exibições: 5907448

Transferring data from v14.cache.googlevideo.com...

Done



Evaluation Submitted - EESIR - Extendable Evaluation System for Information Retrieval - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.latin.dcc.ufmg.br/eesir/evaluations/submitEvaluation.
css make html ent
Gmail
UFMG
Programming
Home
Help
Example
Documentation
Settings (dggc)
Pending Evaluations
Accept Evaluations
Sign out
Done

EESIR - Extendable Evaluation System for Information Retrieval

Successfully submited evaluation!

Interface

- **Thrift** is a software framework for scalable cross-language services development.
  - It combines a powerful software stack with a code generation engine to build services that work efficiently and seamlessly between C++, Java, Python, PHP, and Ruby. Thrift was developed at Facebook and released as open source.

http://incubator.apache.org/thrift