WikiCAPTCHA a ReCAPTCHA-like solution for Wikisource

Cristian Consonni
(CristianCantoro)
Wikimedia Italia

Background

- In Wikisource digitized books in Djvu format are transcribed by volunteers
- Djvu → OCR → Commons →Wikisource
- Djvu have layers (img, text, …) you can add to Djvu as many layers as you like.

The Idea

- In Feb 2011 it.ws sysop Alex brollo noticed that unrecognised characters where marked in the OCR layer with a caret symbol “^”
- He wrote a basic script to find words containing carets and producing images
- We had the idea that such a system could use to produce a ReCAPTCHA-like system to use on Wikisource

What's ReCAPTCHA

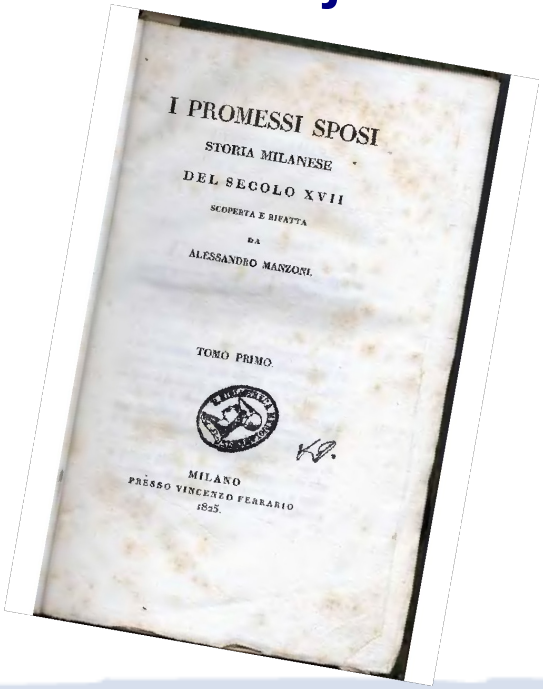
- CAPTCHA system: **C**ompletely **A**utomated **P**ublic **T**uring test to tell **C**omputers and **H**umans **A**part
- Louis von Ahn (then at CMU) – 2008
- CAPTCHA (also by LvA) = meaningless words → ReCAPTCHA = “stop SPAM, read books”
- ReCAPTCHA has been acquired by Google: recaptcha.net
- Transcribing old books in libraries

How it works

- ReCAPTCHA challenges: user presented with two words, one recognized (RW) and the other not (UW). The user is asked to transcribe the words.
- Correct transcription of the RW → human
- Transcription of UWs are collected → rules → transcription accepted as correct → new RWs
- Words can be refused → unrecognizable words

Application

- WikiCAPTCHA could be used as a replacement of or with the current system with has some known limitations
- Better transcription of Djvu in Commons



Where we are

- WikiCAPTCHA is a POC
- Process a Djvu and extract:
  - UW
  - Images of them
  - Store them in a DB + filesystem
- Submit them to the user and collect answers (using Django)

Applicazioni Risorse -2 °C gio 19 gen, 0.01 cristian

Mozilla Firefox

File Modifica Visualizza Cronologia Segnalibri Strumenti Aiuto

http://localhost:8000/captchas/

localhost:8000/captchas/

Google

Più visitati Latest Headlines Wikipedia - Osserv... Wikimedia Italia Cristian Consonni's... UBUNTUWIN BOINC.Italy Haiku Italia

Captchas for everybody!

Capthca:

Word:

Submit

Trova: increment Precedente Successivo Evidenzia Maiuscole/minuscole Fine della pagina raggiunta; si continua dall'inizio

Mozilla Firefox views.py (~/Scrivani...

What's missing

- How to produce the RW for the challenge
- Define rules to accept words
- Write the result back in the Djvu (we need to modify the way we store Djvu in Commons)
- Make the system scalable (!)
- Make the system secure (!!)

Contact

- Download code at github:
  http://github.com/wikicaptcha
- I'm CristianCantoro @ it.wiki
- I'm also deputy chair of Wikimedia Italia:
  cristian.consonni@wikimedia.it
- Skype: cristiancantoro

The End