CLEF 2011 - Semi-automated Artificial Intelligence to assist editing:
An opportunity for Wikimedia sites

To Aru Shiroi Neko
とある白い猫
What is Artificial Intelligence

- Artificial intelligence (AI)
  - Is a branch of computer science
  - Uses machines to process relationship of info
  - Without explicit programming
  - To predict how to handle future data
  - Grown in its use particularly in the past decade
  - Applications range from search engines to space exploration
  - You use it all the time probably without knowing
    - US Postal service
    - Online shopping suggestions
What do I mean by Semi-automated Artificial Intelligence

• Wikimedia sites have relied on “crowdsourcing”
  – Wikimedia is growing much faster than the “crowd”
  – Wikimedians can use the help

• Common “fear” in regards to automation
  – Skynet-phobia or bot takeover
  – Bot making mistakes
  – These points are fair

• Initially (until confidence forms)
  – Program will give suggestions ONLY
  – Suggestions may be restricted to limited people
CLEF: Cross-Language Evaluation Forum

- CLEF: Cross-Language Evaluation Forum
  - 19 - 22 September 2011, Amsterdam, Netherlands
  - 17 - 20 September 2012, Rome, Italy

- Artificial Intelligence classification tasks:
  - Text
  - Images
  - Text and Images
  - Audio

- [http://clef2012.org](http://clef2012.org)
Wikimedia projects

- Wikimedia wikis are Maintained by the Wikimedia Foundation
- Content hosted on Wikimedia projects include
  - Wikipedia, Wikimedia Commons,
  - Wikibooks, Wikisource,
  - Wikinews, Wikiquote,
  - Wiktionary, Wikispecies
- Content is used
  - Wikimedia-wide
  - Internet-wide
    - Mirrors, forum posts, social media
    - **Downloaded for offline use**
      - Wikireader (Commercial)
      - OLPC (Non-Profit)
      - Research (Scientific)
Free licenses

• Freely licensed
• Reusable, freely distributable
• Allows derivatives of any kind
• Commercially usable
• Used in very large projects like Wikipedia
• Used in research as demonstrated by ImageCLEF already
• Compels people to add to the existing repository of images
• Statistical information will be posted here
Wikipedia – Proposed task 1: Copyright/Plagiarism detection

• Classification task:
  – Identify content copied to and from Wikipedia
    • Since the start this has remained to be a problem
    • Currently, tools such as Copyscape is used
    • Incorrectly identifiable as copyvio by tools like Copyscape
    • Make use of page histories to help determine
  – Detect inclusion of previously oversigted content
    • Only users with oversight access would be able to read the reports

• CLEF
  – Similar track on PAN workshop of CLEF 2011
Wikipedia – Proposed task 2: Author identification

• **Classification task:**
  – **Sockpuppets**
    • Returning users that are previously blocked
  – **Meatpuppets & Lobbyists**
    • Forum driven effort to introduce bias
    • List of blocks in regards to CAMERA is in light years
  – **Political edits**
    • US Congressional staff edits to Wikipedia

• **CLEF**
  – Similar track on PAN workshop of CLEF 2011
Wikipedia – Proposed task 3: Vandalism detection

• Classification task:
  – Identify Vandalism
    • Blanking, Gibberish, Sneaky
      – RC feed is full of these
    • Slander/Defamation
      – In other words: fact checking
      – John Seigenthaler controversy
    • Spam
      – Links (page rank boost)
      – Bot generated spam articles

• CLEF
  – Similar track on PAN workshop of CLEF 2011
Wikipedia – Proposed task 4: Disambiguation & Category Identification

• Classification task:
  – People
  – Places
  – Abbreviations
  – Incorrectly linked pages

• Disambiguation pages can be...
  – Confusing
  – Less than helpful
  – Outright misleading

• CLEF
  – Nothing similar on CLEF, yet
Wikipedia – Proposed task 5: Correlate real life events

• Source?
  – News, Facebook, Google+, Twitter trends, Blogs
  – Wikipedia article on topic
    • Is it controversial? Sentiment analysis?

• Why?
  – Assist admins by directing them to what is popular
  – Help identify content for ITN section of main page
  – Identify pages that can be updated

• CLEF
  – Nothing similar on CLEF, yet
• **OCR - Object Character Recognition**
  – Technology already exists for raw text
  – OCR then correlate with Wikipedia
    • Break up, categorize, link and tag
  – Help establish copyright
    • Text known to have copyright?
    • Deleted before as a copyright violation?

• **CLEF**
  – Nothing on OCR
Wikipedia – Proposed task 1

• Copyright identification (detection of text copied from elsewhere)
• Copyright identification (deleted images)
• OCR for Wiki
• Vandalism detection
• Category identification
  – Automated content identification (what images and/or text contain)
  – Automated plant identification (images)
• Disambiguation
• Correlate real life events, twitter trends and etc. to determine potential articles of attention for administrators
Wikimedia Commons

- Wikimedia Commons (or simply Commons) is an online repository of free-use
  - Images,
  - Audio,
  - Video and
  - Other files

- http://commons.wikimedia.org
### Wikimedia Commons – Statistics 1

<table>
<thead>
<tr>
<th>Format</th>
<th>Count</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>midi</td>
<td>1,605</td>
<td>audio</td>
</tr>
<tr>
<td>wav</td>
<td>4</td>
<td>audio</td>
</tr>
<tr>
<td>ogg</td>
<td>142,193</td>
<td>audio/video</td>
</tr>
<tr>
<td>mp4</td>
<td>1</td>
<td>audio/video</td>
</tr>
<tr>
<td>gif</td>
<td>121,155</td>
<td>image/animation</td>
</tr>
<tr>
<td>jpeg</td>
<td>9,400,557</td>
<td>image</td>
</tr>
<tr>
<td>png</td>
<td>835,988</td>
<td>image</td>
</tr>
<tr>
<td>svg+xml</td>
<td>485,356</td>
<td>images, vector</td>
</tr>
<tr>
<td>tiff</td>
<td>64,121</td>
<td>image</td>
</tr>
<tr>
<td>vnd.djvu</td>
<td>18,558</td>
<td>image</td>
</tr>
<tr>
<td>x-xcf</td>
<td>263</td>
<td>image</td>
</tr>
<tr>
<td>pdf</td>
<td>17,204</td>
<td>mixed text &amp; images</td>
</tr>
</tbody>
</table>

Statistics by Benutzer:DaBPunkt
There are 103,112 galleries on Commons

There are 11,085,316 files on Commons

There are 1,259,001 deleted files on Commons

– %11.342 of the existing files are deleted
Examples of unwanted or controversial content

- Offensive/Vandalism
  - Genitalia (I will spare you the images)
  - Troll use of certain words or block of text within the images
  - Troll use of certain symbols within images: ☠ ☭ ☩

- Controversial
  - Nudity
  - Violence

- Copyright violations
  - Screenshots of movies
  - Photos of sculptures (non-FOP)
  - Special cases such as the Eiffel tower at night
  - Other stuff used without permission (frequently)
### Category: Birds

For the taxonomic classification of birds, see Category: Aves

This is a main category or subcategory requiring frequent diffusion and maybe maintenance. As many pictures and media files as possible should be moved into the right subcategories. (more information on categories)

#### Vernacular names

- **English**: Birds - Afrikaans: Pánu - Bahasa Indonesia: Burung - Bahasa Melayu: Burung

#### Subcategories

This category has the following 40 subcategories, out of 40 total.

A

- [+]: Abused birds (1 C, 1 F)
- [+]: Albinism in birds (4 C, 32 F)
- [+]: Bird anatomy (19 C, 20 F)
- [+]: Audio files of Aves (16 C, 6 F)

B

- [+]: Bird behavior (19 C, 23 F)
- [+]: Bird harness (6 F)
- [+]: Bird hunting (3 C, 14 F)
- [+]: Bird repellents (5 C, 26 F)
- [+]: Birdhouses (3 C, 1 P, 65 F)

C

- [+]: Bird conservation (9 C, 7 F)

D

- [+]: Bird distributions maps (23 C, 1 P, 41 F)

E

- [+]: Economics (17 C, 1 F)

F

- [+]: Bird faces (1 C, 33 F)
- [+]: Bird tracks (1 C, 36 F)

G

- [+]: Bird flight (2 C, 4 F)
- [+]: Flightless birds (27 C, 1 P, 3 F)

I

- [+]: Invasive birds (5 F)

L

- [x]: Bird reproduction (9 C)

P

- [+]: Bird parks (12 C, 7 F)
- [+]: Bird pellets (4 F)
- [+]: People with birds (12 C, 76 F)
- [+]: Pronunciation of words relating to birds (2 C, 29 F)

S

- [+]: Bird seed (9 F)
- [+]: Bird-related signs (2 C, 14 F)

T

- [+]: Bird toys (1 C, 7 F)
- [+]: Bird tracks (1 C, 36 F)

V

- [+]: Veterinary medicine for birds (1 C)
- [+]: Extinct birds (49 C, 8 F)
An **aircraft** is a vehicle which is able to fly by being supported by the air, or in general, the atmosphere, of a planet. Examples include balloons, airplanes and helicopters.

Editors please ensure that the subcategory hierarchy is similar to the corresponding Wikipedia category hierarchy.

### Subcategories

This category has the following 49 subcategories, out of 49 total.

- Aircraft by color (14 C)
- Aircraft by condition (6 C)
- Aircraft by country (119 C, 1 F)
- Aircraft by function (15 C)
- Aircraft by manufacturer (397 C, 1 F)

#### A cont.

- Aircraft in art (15 C, 102 F)
- Aircraft operators (6 C, 1 F)

#### C cont.

- Aircraft carriers (23 C, 3 F)
- Cutaways of aircraft (1 C, 9 F)

#### O cont.

- Aircraft operators maps (1 C, 81 F)
- Aircraft parts (34 C, 1 P, 132 F)
- People with aircraft (8 C, 85 F)
- Pronunciation of words relating to aircraft (16 F)

#### T cont.

- Transport of aircraft (4 C, 1 F)
- Videos of aircraft (24 C, 5 F)
- Views of aircraft (7 C, 1 F)

#### W cont.

- Water reflections of aircraft (6 F)
- Unidentified aircraft (1 C, 47 F)
An aircraft is a vehicle which is able to fly by being supported by the air, or in general, the atmosphere, of a planet. Examples include balloons, airplanes and helicopters.

Editors please ensure that the subcategory hierarchy is similar to the corresponding Wikipedia category hierarchy.
• Plant identification
  – Is it a plant
  – Type of image
    • Scan of leaf
    • Photo of individual leaf
    • Photo in nature
  – What type of leaf does it have?
    • Shape and Arrangement
    • Margin
    • Venation
  – Identification isn’t necessarily down to the species
• A human botanist can disambiguate further
Wikimedia Foundation – Proposed tasks

• Wikimedia servers
  – Performance
    • More efficient distribution of load
    • Detection of trends leading to failure
    • Help determine deadlines for upgrades
  – Cyber Defense
    • Anomaly detection
      – Detect botnets
      – Detect sneaky RATs
      – Consider Stuxnet and DuQu
    • Detection of manipulation to database
    • Does Wikipedia count as “critical infrastructure”?

• Sentiment analysis of social media and the web
  – Improve relations by not being deaf to the concerns of readers
  – Handle false information about the WMF
  – Identify impersonation attempts of WMF
Wikimedia Commons – Questions?
• Thank You!