Uncovering Plagiarism, Authorship, and Social Software Misuse

PAN 2011

[pan.webis.de]

CLEF 2011 Amsterdam

YAHOO! RESEARCH

The PAN Team

Teresa Holfeld

Andreas Eiselt

Martin Potthast

Alberto Barrón-Cedeño

Efstathios Stamatatos

Moshe Koppel

Patrick Juola

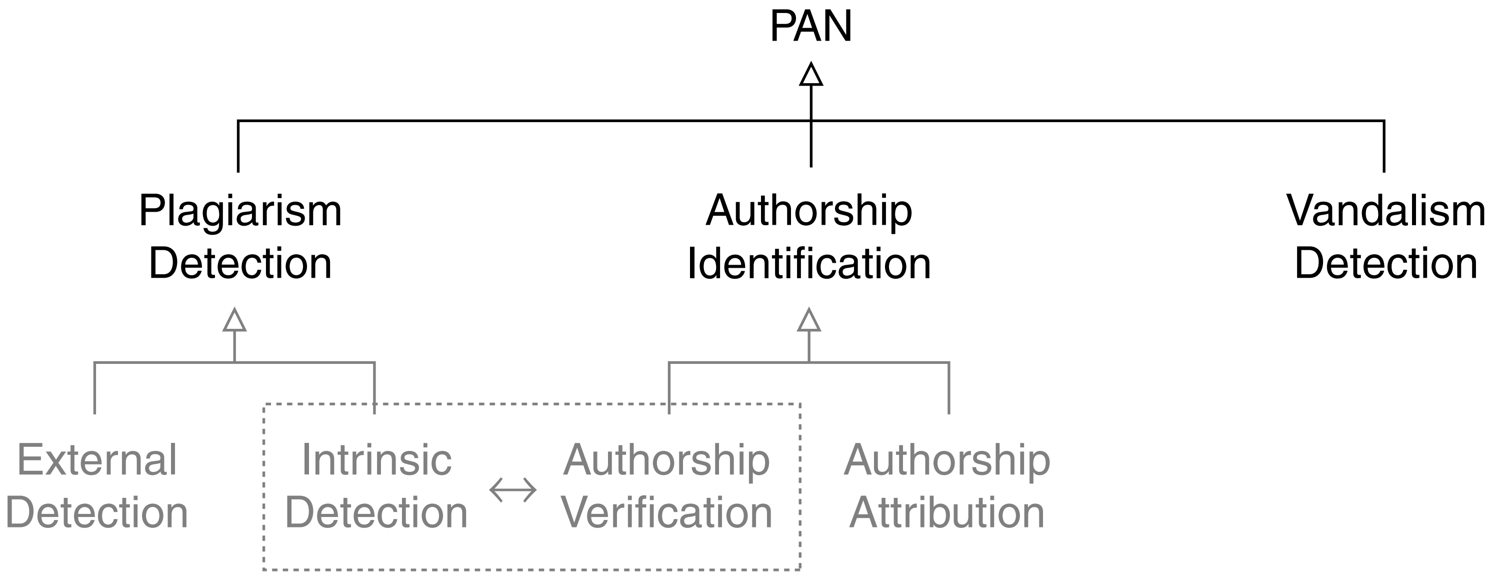
Shlomo Argamon

Paolo Rosso

Benno Stein

Bauhaus-Universität Weimar	Martin Potthast, Benno Stein, Andreas Eiselt, Teresa Holfeld
Universidad Politécnica de Valencia	Alberto Barrón-Cedeño, Paolo Rosso
University of the Aegean	Efstathios Stamatatos
Bar-Ilan University	Moshe Koppel
Illinois Institute of Technology	Shlomo Argamon
Duquesne University	Patrick Juola

PAN Overview



PAN Overview

Mission

- Plagiarism Detection
  - text plagiarism within and across languages, multimedia plagiarism
  - text reuse, paraphrasing, information flow, meme tracking
  - near-duplicates, high-similarity search, fingerprinting, hash-based search

PAN Overview

Mission

- Plagiarism Detection
  - text plagiarism within and across languages, multimedia plagiarism
  - text reuse, paraphrasing, information flow, meme tracking
  - near-duplicates, high-similarity search, fingerprinting, hash-based search

- Authorship Identification
  - models for authorship verification, authorship attribution, and writing style
  - models to capture personal traits and sentiment
  - text forensics, ghostwriting, intrinsic plagiarism detection

PAN Overview

Mission

- Plagiarism Detection
  - text plagiarism within and across languages, multimedia plagiarism
  - text reuse, paraphrasing, information flow, meme tracking
  - near-duplicates, high-similarity search, fingerprinting, hash-based search

- Authorship Identification
  - models for authorship verification, authorship attribution, and writing style
  - models to capture personal traits and sentiment
  - text forensics, ghostwriting, intrinsic plagiarism detection

- Social Software Misuse
  - serial sharing, lobbyism, spam
  - trolling, stalking, Wikipedia vandalism
  - social trust, anonymity and de-anonymization

PAN Overview

Mission & Tasks

- Plagiarism Detection
  - text plagiarism within and across languages, multimedia plagiarism
  - text reuse, paraphrasing, information flow, meme tracking
  - near-duplicates, high-similarity search, fingerprinting, hash-based search

- Authorship Identification
  - models for authorship verification, authorship attribution, and writing style
  - models to capture personal traits and sentiment
  - text forensics, ghostwriting, intrinsic plagiarism detection

- Social Software Misuse
  - serial sharing, lobbyism, spam
  - trolling, stalking, Wikipedia vandalism
  - social trust, anonymity and de-anonymization

Plagiarism Detection

Plagiarism Detection

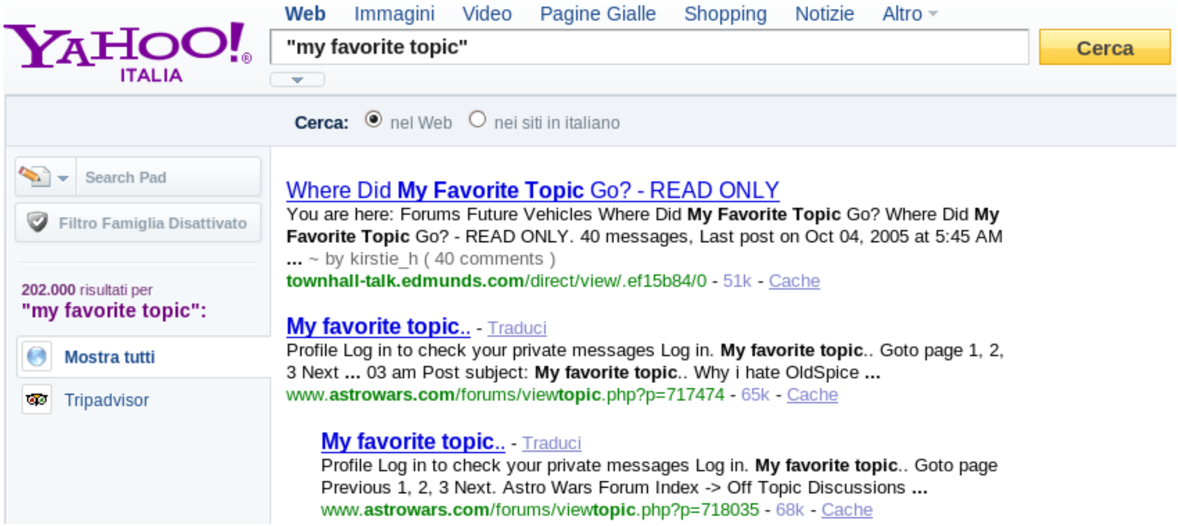
Plagiarism is the practice of claiming, or implying, original authorship of someone else's written or creative work, in whole or in part, into one's own without adequate acknowledgment.

Plagiarism Detection

Plagiarism is the practice of claiming, or implying, original authorship of someone else's written or creative work, in whole or in part, into one's own without adequate acknowledgment.

[Wikipedia: Plagiarism, 2009]

. . . better technology nowadays ;–)



"my favorite topic"

Cerca: ● nel Web ○ nei siti in italiano

202.000 risultati per "my favorite topic":

Where Did My Favorite Topic Go? - READ ONLY
You are here: Forums Future Vehicles Where Did **My Favorite Topic** Go? Where Did **My Favorite Topic** Go? - READ ONLY. 40 messages, Last post on Oct 04, 2005 at 5:45 AM **...** ~ by kirstie_h (40 comments)
townhall-talk.edmunds.com/direct/view/.ef15b84/0 - 51k - Cache

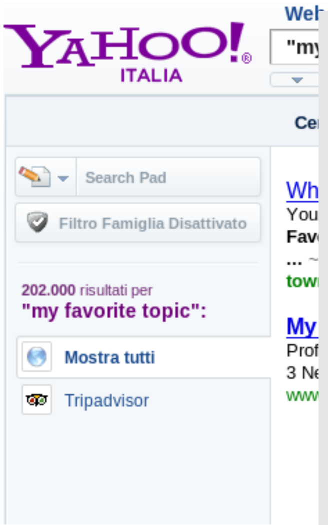
My favorite topic.. - Traduci
Profile Log in to check your private messages Log in. **My favorite topic**.. Goto page 1, 2, 3 Next **...** 03 am Post subject: **My favorite topic**.. Why i hate OldSpice **...**
www.astrowars.com/forums/viewtopic.php?p=717474 - 65k - Cache

My favorite topic.. - Traduci
Profile Log in to check your private messages Log in. **My favorite topic**.. Goto page Previous 1, 2, 3 Next. Astro Wars Forum Index -> Off Topic Discussions **...**
www.astrowars.com/forums/viewtopic.php?p=718035 - 68k - Cache

+

WIKIPEDIA
The Free Encyclopedia

. . . better technology nowadays ;–)


Web Immagini Video Pagine Gialle Shopping Notizie Altro
YAHOO! ITALIA
"my
Ce
Search Pad
Filtro Famiglia Disattivato

202.000 risultati per "my favorite topic":

Mostra tutti

Tripadvisor

Wh
You
Fav
... ~
tow

My
Prof
3 Ne
www

"Nice essay, Tom, your cut and paste skills are beyond reproach."

Authorship Identification

Authorship Identification

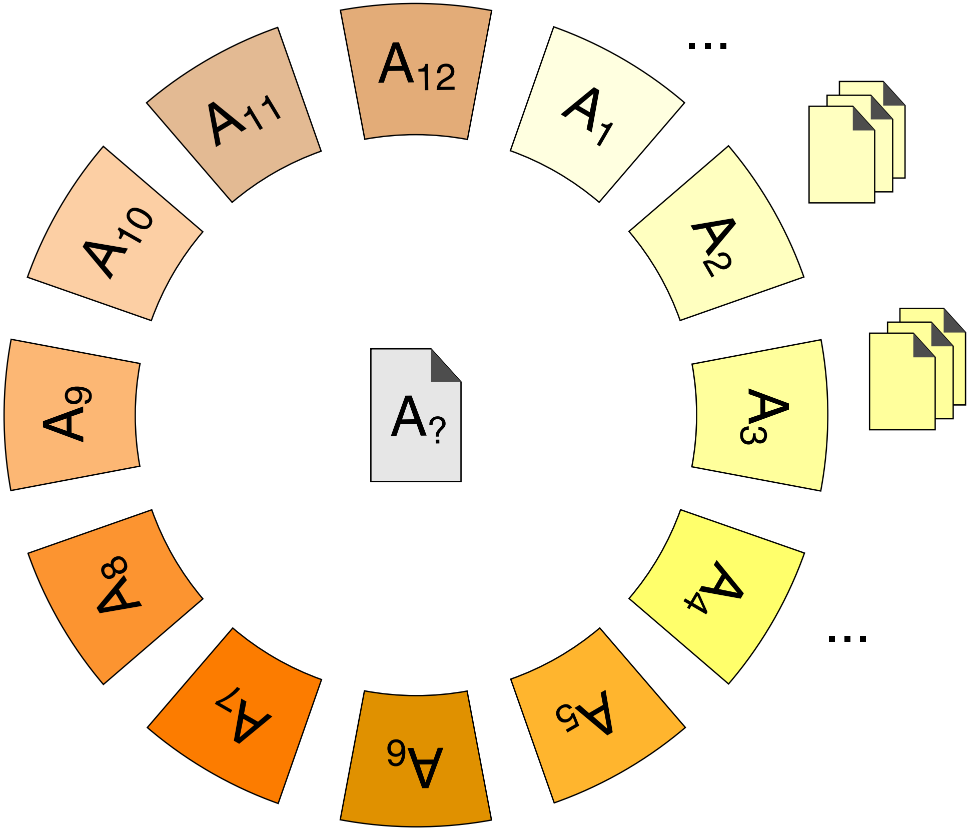
Sub-task: Authorship Attribution

Given a text of uncertain authorship and texts from a set of candidate authors, the task is to map the uncertain text onto the true author among the candidates.

Authorship Identification

Sub-task: Authorship Attribution

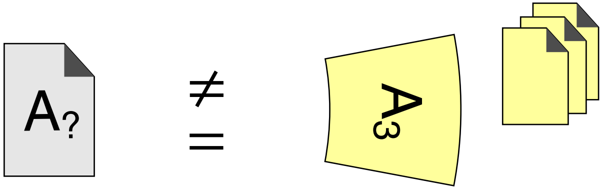
Given a text of uncertain authorship and texts from a set of candidate authors, the task is to map the uncertain text onto the true author among the candidates.



Authorship Identification

Sub-task: Authorship Verification

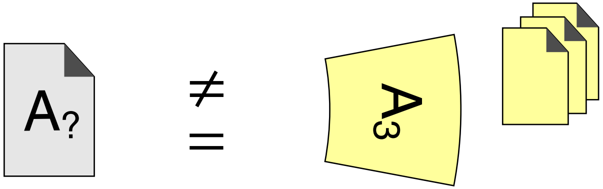
Given a text of uncertain authorship and text from a specific author,
the task is to determine whether the given text has been written by that author.



Authorship Identification

Sub-task: Authorship Verification

Given a text of uncertain authorship and text from a specific author,
the task is to determine whether the given text has been written by that author.



The problem can be considered as a one-class classification problem.

Vandalism Detection

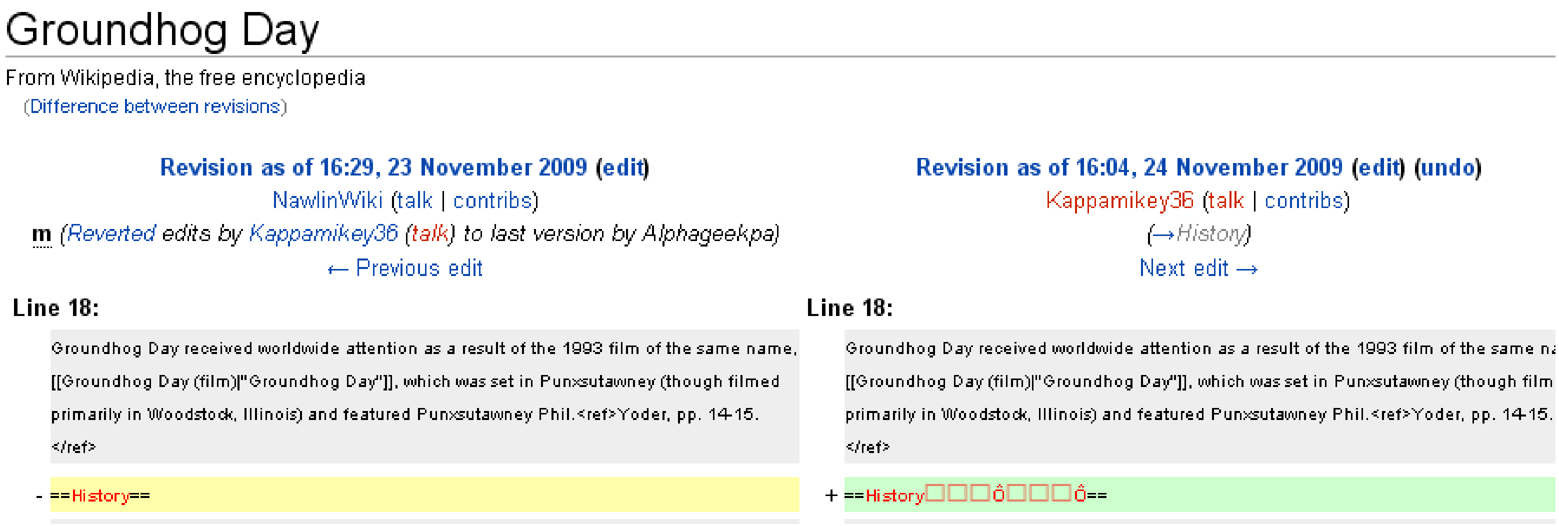
Vandalism Detection in Amsterdam

Vandalism Detection in Amsterdam

Vandalism Detection in Wikipedia

Vandalism Detection in Wikipedia

Example: special chars, spacing



Groundhog Day

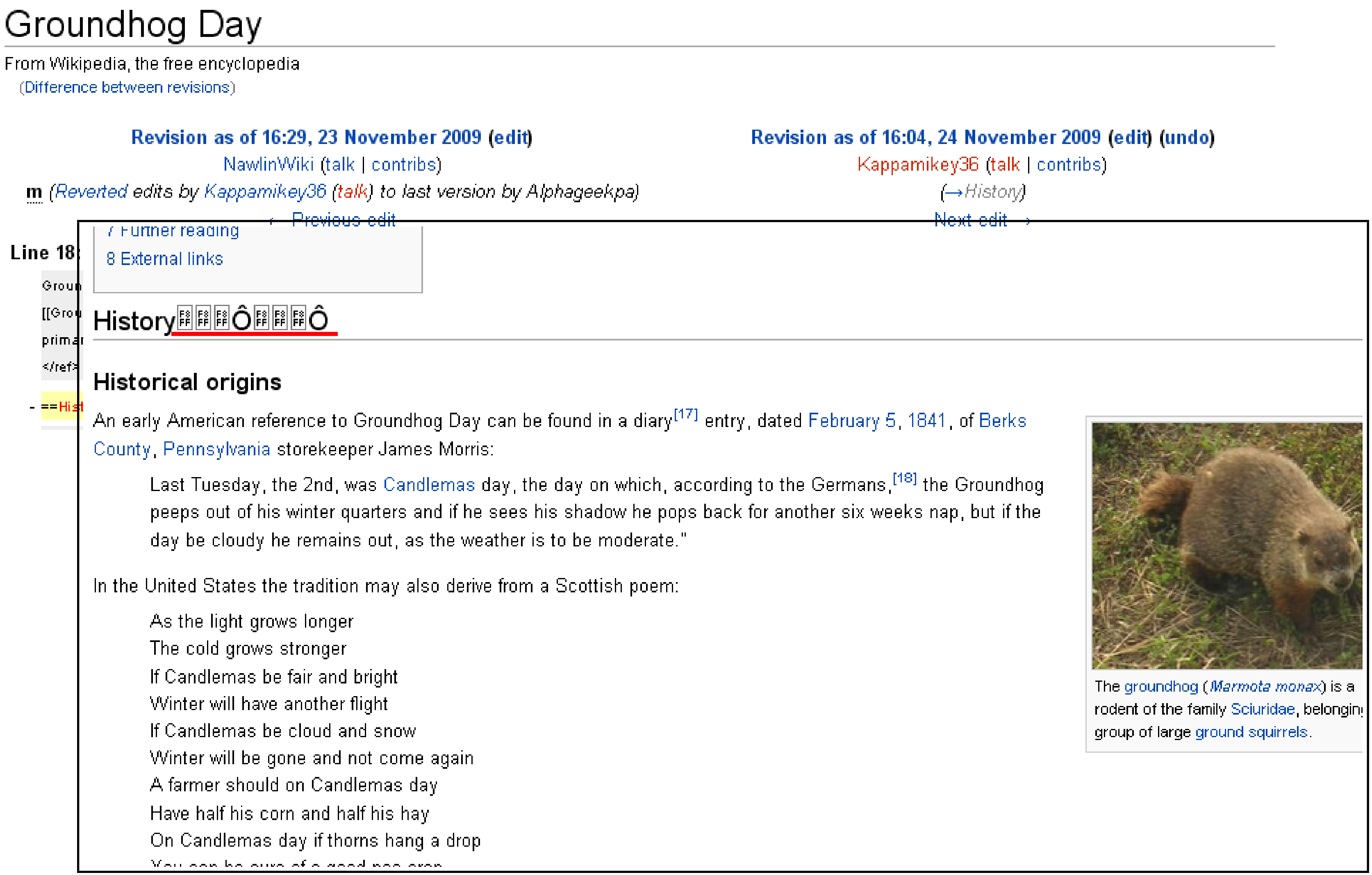
From Wikipedia, the free encyclopedia

(Difference between revisions)

Revision as of 16:29, 23 November 2009 (edit) NawlinWiki (talk | contribs) m (Reverted edits by Kappamikey36 (talk) to last version by Alphageekpa) ← Previous edit	Revision as of 16:04, 24 November 2009 (edit) (undo) Kappamikey36 (talk | contribs) (→History) Next edit →
Line 18:	**Line 18:**
Groundhog Day received worldwide attention as a result of the 1993 film of the same name, [[Groundhog Day (film)|"Groundhog Day"]], which was set in Punxsutawney (though filmed primarily in Woodstock, Illinois) and featured Punxsutawney Phil.<ref>Yoder, pp. 14-15.</ref>	Groundhog Day received worldwide attention as a result of the 1993 film of the same na [[Groundhog Day (film)|"Groundhog Day"]], which was set in Punxsutawney (though film primarily in Woodstock, Illinois) and featured Punxsutawney Phil.<ref>Yoder, pp. 14-15. </ref>
- ==History==	+ ==History□□□ô□□□ô==

Vandalism Detection in Wikipedia

Example: special chars, spacing



Groundhog Day

From Wikipedia, the free encyclopedia

(Difference between revisions)

Revision as of 16:29, 23 November 2009 (edit)	Revision as of 16:04, 24 November 2009 (edit) (undo)
NawlinWiki (talk | contribs)	Kappamikey36 (talk | contribs)
m (*Reverted edits by Kappamikey36 (talk) to last version by Alphageekpa*)	(→*History*)
← Previous edit	Next edit →

Line 18:

Groun

[[Grou

primar

</ref>

- ==Hist

7 Further reading

8 External links

History

Historical origins

An early American reference to Groundhog Day can be found in a diary[17] entry, dated February 5, 1841, of Berks County, Pennsylvania storekeeper James Morris:

> Last Tuesday, the 2nd, was Candlemas day, the day on which, according to the Germans,[18] the Groundhog peeps out of his winter quarters and if he sees his shadow he pops back for another six weeks nap, but if the day be cloudy he remains out, as the weather is to be moderate."

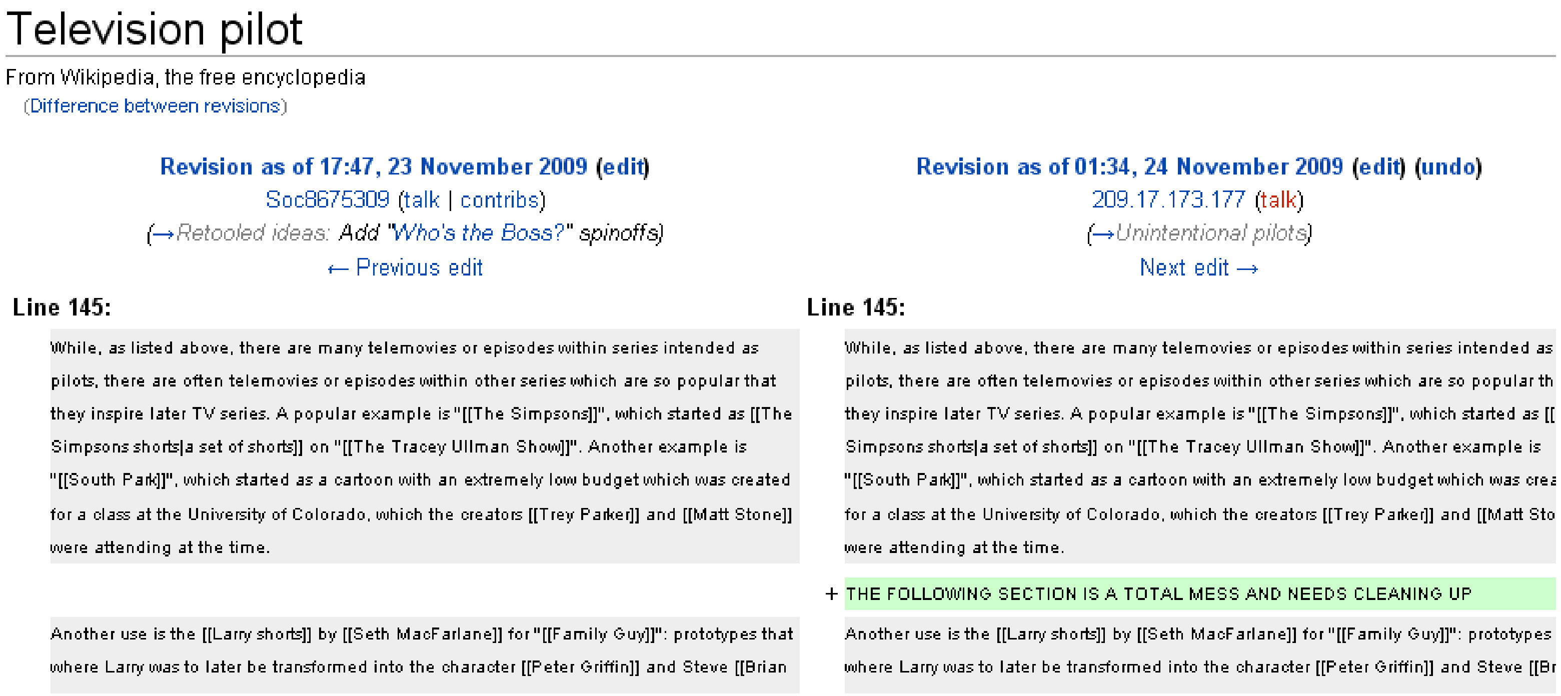
In the United States the tradition may also derive from a Scottish poem:

> As the light grows longer
> The cold grows stronger
> If Candlemas be fair and bright
> Winter will have another flight
> If Candlemas be cloud and snow
> Winter will be gone and not come again
> A farmer should on Candlemas day
> Have half his corn and half his hay
> On Candlemas day if thorns hang a drop
> You can be sure of a good pea crop

The groundhog (*Marmota monax*) is a rodent of the family Sciuridae, belongin group of large ground squirrels.

Vandalism Detection in Wikipedia

Example: misguided helping



Television pilot

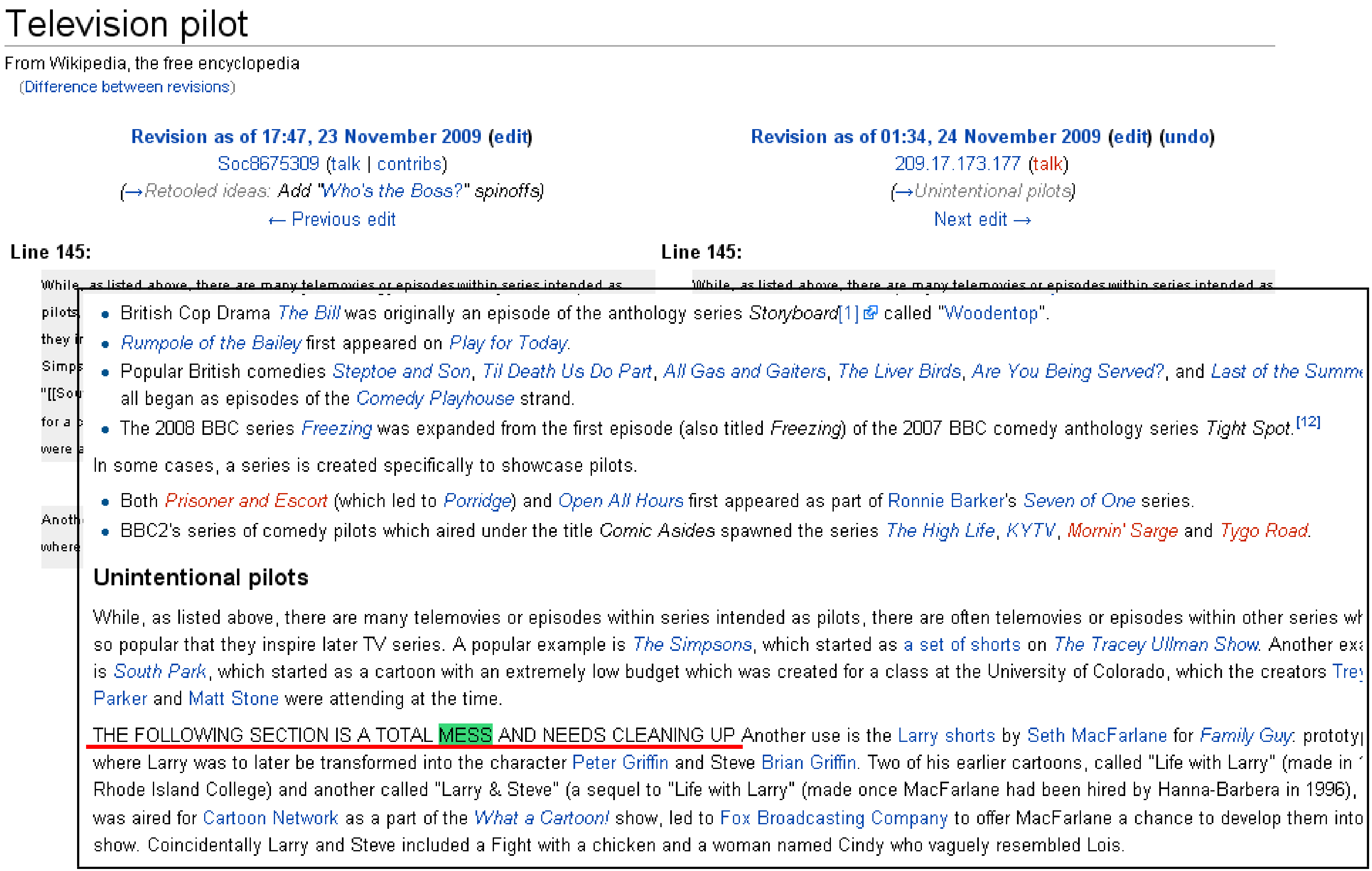
From Wikipedia, the free encyclopedia

(Difference between revisions)

Revision as of 17:47, 23 November 2009 (edit) Soc8675309 (talk | contribs) *(→Retooled ideas: Add "Who's the Boss?" spinoffs)* ← Previous edit	**Revision as of 01:34, 24 November 2009 (edit) (undo)** 209.17.173.177 (talk) *(→Unintentional pilots)* Next edit →
Line 145:	**Line 145:**
While, as listed above, there are many telemovies or episodes within series intended as pilots, there are often telemovies or episodes within other series which are so popular that they inspire later TV series. A popular example is "[[The Simpsons]]", which started as [[The Simpsons shorts|a set of shorts]] on "[[The Tracey Ullman Show]]". Another example is "[[South Park]]", which started as a cartoon with an extremely low budget which was created for a class at the University of Colorado, which the creators [[Trey Parker]] and [[Matt Stone]] were attending at the time.	While, as listed above, there are many telemovies or episodes within series intended as pilots, there are often telemovies or episodes within other series which are so popular th they inspire later TV series. A popular example is "[[The Simpsons]]", which started as [[Simpsons shorts|a set of shorts]] on "[[The Tracey Ullman Show]]". Another example is "[[South Park]]", which started as a cartoon with an extremely low budget which was crea for a class at the University of Colorado, which the creators [[Trey Parker]] and [[Matt Sto were attending at the time.
	+ THE FOLLOWING SECTION IS A TOTAL MESS AND NEEDS CLEANING UP
Another use is the [[Larry shorts]] by [[Seth MacFarlane]] for "[[Family Guy]]": prototypes that where Larry was to later be transformed into the character [[Peter Griffin]] and Steve [[Brian	Another use is the [[Larry shorts]] by [[Seth MacFarlane]] for "[[Family Guy]]": prototypes where Larry was to later be transformed into the character [[Peter Griffin]] and Steve [[Br

Vandalism Detection in Wikipedia

Example: misguided helping



Television pilot

From Wikipedia, the free encyclopedia

(Difference between revisions)

Revision as of 17:47, 23 November 2009 (edit)	**Revision as of 01:34, 24 November 2009 (edit) (undo)**
Soc8675309 (talk | contribs)	209.17.173.177 (talk)
(→*Retooled ideas: Add "Who's the Boss?" spinoffs*)	(→*Unintentional pilots*)
← Previous edit	Next edit →

Line 145: **Line 145:**

While, as listed above, there are many telemovies or episodes within series intended as

pilots

they i

Simp

"[[So

for a

were

Anoth

where

While, as listed above, there are many telemovies or episodes within series intended as

- British Cop Drama *The Bill* was originally an episode of the anthology series *Storyboard*[1] called "Woodentop".
- *Rumpole of the Bailey* first appeared on *Play for Today*.
- Popular British comedies *Steptoe and Son*, *Til Death Us Do Part*, *All Gas and Gaiters*, *The Liver Birds*, *Are You Being Served?*, and *Last of the Summe* all began as episodes of the *Comedy Playhouse* strand.
- The 2008 BBC series *Freezing* was expanded from the first episode (also titled *Freezing*) of the 2007 BBC comedy anthology series *Tight Spot*.[12]

In some cases, a series is created specifically to showcase pilots.

- Both *Prisoner and Escort* (which led to *Porridge*) and *Open All Hours* first appeared as part of Ronnie Barker's *Seven of One* series.
- BBC2's series of comedy pilots which aired under the title *Comic Asides* spawned the series *The High Life*, *KYTV*, *Mornin' Sarge* and *Tygo Road*.

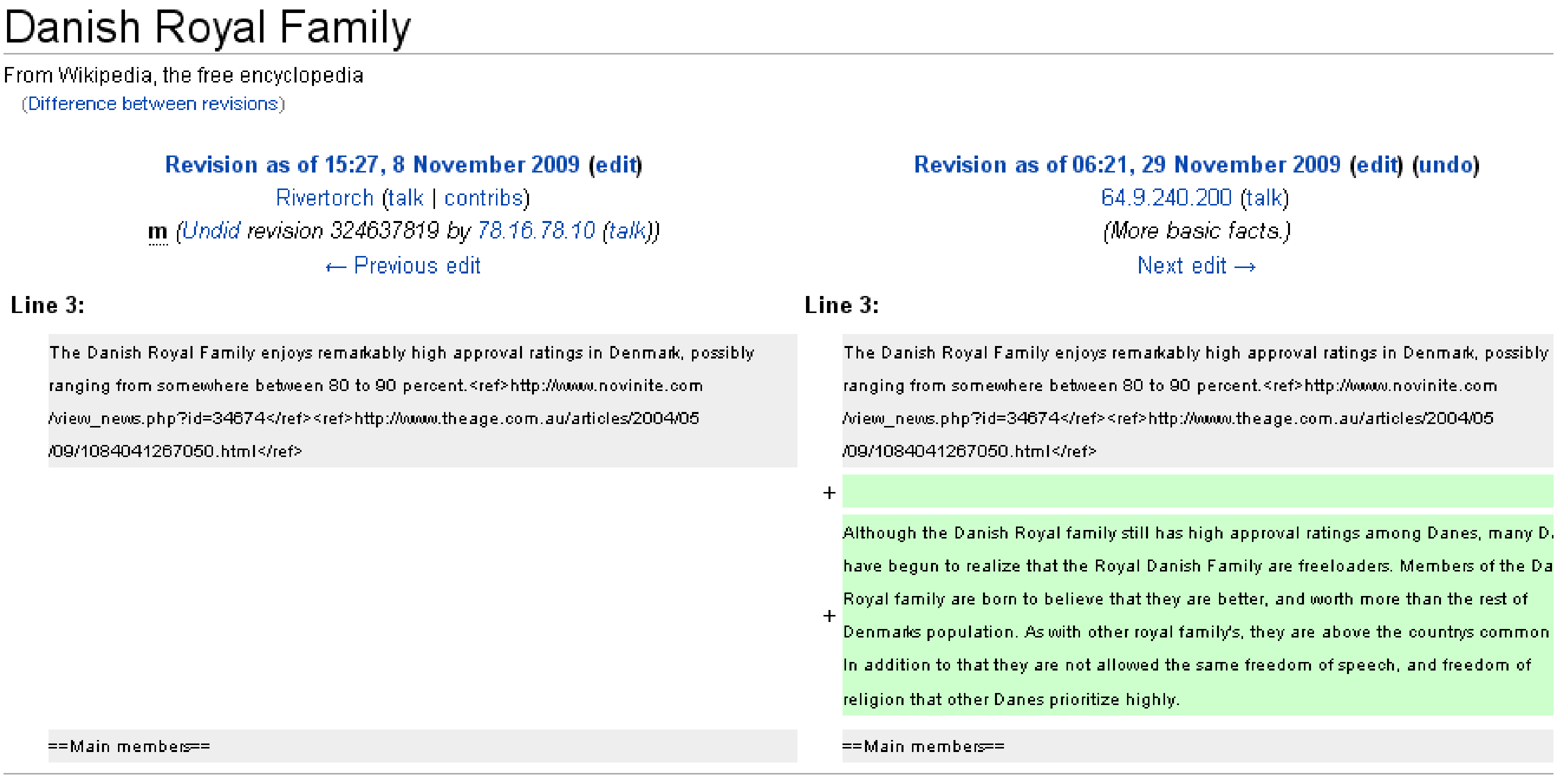
Unintentional pilots

While, as listed above, there are many telemovies or episodes within series intended as pilots, there are often telemovies or episodes within other series wh so popular that they inspire later TV series. A popular example is *The Simpsons*, which started as a set of shorts on *The Tracey Ullman Show*. Another exa is *South Park*, which started as a cartoon with an extremely low budget which was created for a class at the University of Colorado, which the creators Trey Parker and Matt Stone were attending at the time.

THE FOLLOWING SECTION IS A TOTAL MESS AND NEEDS CLEANING UP Another use is the Larry shorts by Seth MacFarlane for *Family Guy*: prototy where Larry was to later be transformed into the character Peter Griffin and Steve Brian Griffin. Two of his earlier cartoons, called "Life with Larry" (made in ' Rhode Island College) and another called "Larry & Steve" (a sequel to "Life with Larry" (made once MacFarlane had been hired by Hanna-Barbera in 1996), was aired for Cartoon Network as a part of the *What a Cartoon!* show, led to Fox Broadcasting Company to offer MacFarlane a chance to develop them into show. Coincidentally Larry and Steve included a Fight with a chicken and a woman named Cindy who vaguely resembled Lois.

Vandalism Detection in Wikipedia

Example: wrong facts, opinionated, nonsense



Danish Royal Family

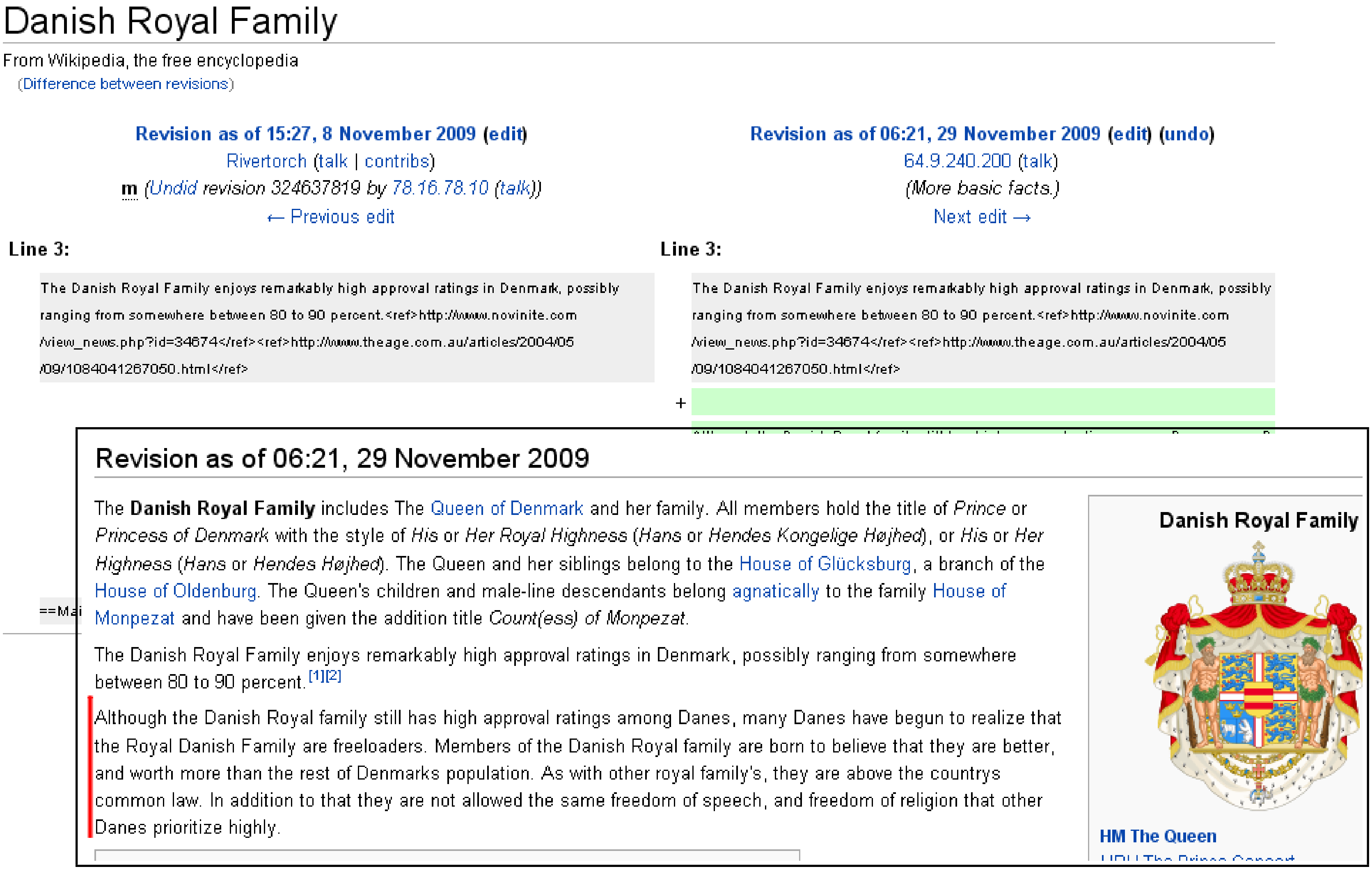
From Wikipedia, the free encyclopedia

(Difference between revisions)

Revision as of 15:27, 8 November 2009 (edit)	**Revision as of 06:21, 29 November 2009 (edit) (undo)**
Rivertorch (talk | contribs)	64.9.240.200 (talk)
m *(Undid revision 324637819 by 78.16.78.10 (talk))*	*(More basic facts.)*
← Previous edit	Next edit →
Line 3:	**Line 3:**
The Danish Royal Family enjoys remarkably high approval ratings in Denmark, possibly ranging from somewhere between 80 to 90 percent.<ref>http://www.novinite.com /view_news.php?id=34674</ref><ref>http://www.theage.com.au/articles/2004/05 /09/1084041267050.html</ref>	The Danish Royal Family enjoys remarkably high approval ratings in Denmark, possibly ranging from somewhere between 80 to 90 percent.<ref>http://www.novinite.com /view_news.php?id=34674</ref><ref>http://www.theage.com.au/articles/2004/05 /09/1084041267050.html</ref>
	+
	+ Although the Danish Royal family still has high approval ratings among Danes, many D. have begun to realize that the Royal Danish Family are freeloaders. Members of the Da Royal family are born to believe that they are better, and worth more than the rest of Denmarks population. As with other royal family's, they are above the countrys common In addition to that they are not allowed the same freedom of speech, and freedom of religion that other Danes prioritize highly.
==Main members==	==Main members==

Vandalism Detection in Wikipedia

Example: wrong facts, opinionated, nonsense



Danish Royal Family

From Wikipedia, the free encyclopedia

(Difference between revisions)

Revision as of 15:27, 8 November 2009 (edit)	**Revision as of 06:21, 29 November 2009 (edit) (undo)**
Rivertorch (talk | contribs)	64.9.240.200 (talk)
m *(Undid revision 324637819 by 78.16.78.10 (talk))*	*(More basic facts.)*
← Previous edit	Next edit →

Line 3:

The Danish Royal Family enjoys remarkably high approval ratings in Denmark, possibly ranging from somewhere between 80 to 90 percent.<ref>http://www.novinite.com/view_news.php?id=34674</ref><ref>http://www.theage.com.au/articles/2004/05/09/1084041267050.html</ref>

Line 3:

The Danish Royal Family enjoys remarkably high approval ratings in Denmark, possibly ranging from somewhere between 80 to 90 percent.<ref>http://www.novinite.com/view_news.php?id=34674</ref><ref>http://www.theage.com.au/articles/2004/05/09/1084041267050.html</ref>

+

==Mai

Revision as of 06:21, 29 November 2009

The **Danish Royal Family** includes The Queen of Denmark and her family. All members hold the title of *Prince* or *Princess of Denmark* with the style of *His* or *Her Royal Highness* (*Hans* or *Hendes Kongelige Højhed*), or *His* or *Her Highness* (*Hans* or *Hendes Højhed*). The Queen and her siblings belong to the House of Glücksburg, a branch of the House of Oldenburg. The Queen's children and male-line descendants belong agnatically to the family House of Monpezat and have been given the addition title *Count(ess) of Monpezat*.

The Danish Royal Family enjoys remarkably high approval ratings in Denmark, possibly ranging from somewhere between 80 to 90 percent.[1][2]

Although the Danish Royal family still has high approval ratings among Danes, many Danes have begun to realize that the Royal Danish Family are freeloaders. Members of the Danish Royal family are born to believe that they are better, and worth more than the rest of Denmarks population. As with other royal family's, they are above the countrys common law. In addition to that they are not allowed the same freedom of speech, and freedom of religion that other Danes prioritize highly.

Danish Royal Family

HM The Queen

More about PAN

More about PAN

History [pan.webis.de]

2007

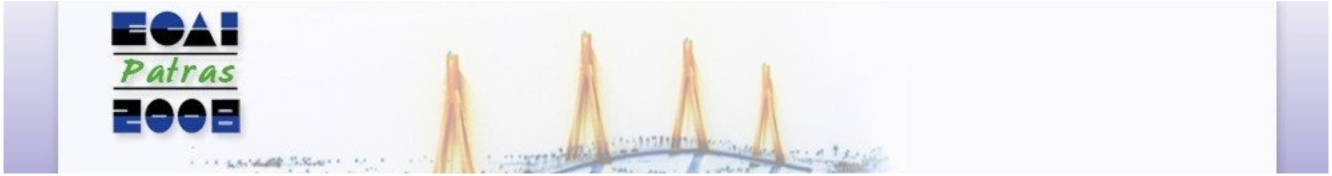
Workshop: PAN'07
Call for Papers
Important Dates
Submission
Program Committee
Program / Slides
Proceedings / [PDF]
Contact

International Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN)

held in conjunction with

The 30th Annual International ACM SIGIR Conference
23-27 July 2007, Amsterdam

2008



2009

3rd & PAN Workshop
1st & Competition
on Plagiarism Detection

2010

PAN 2010 LAB
Uncovering Plagiarism, Authorship, and Social Software Misuse

2011

PAN 2011 Lab
Uncovering Plagiarism, Authorship, and Social Software Misuse
held in conjunction with the CLEF 2011 Conference on Multilingual and Multimodal Information Access Evaluation
19-22 September 2011, Amsterdam

More about PAN

Key Figures 2011

	2009	2010		2011		
Task(s)	plagiarism	plagiarism	vandalism	plagiarism	authorship	vandalism
Corpus size	5GB	3.4GB	8.2GB	4.6GB	3MB	8.4GB
Corpus size (cases)	94 000	68 000	32 000	61 000	4 100	64 000
Languages	3	3	1	3	1	3

Sponsorship by YAHOO! Research.

Media coverage on German and Spanish television, among others.

More about PAN

Key Figures 2011

	2009	2010		2011		
Task(s)	plagiarism	plagiarism	vandalism	plagiarism	authorship	vandalism
Corpus size	5GB	3.4GB	8.2GB	4.6GB	3MB	8.4GB
Corpus size (cases)	94 000	68 000	32 000	61 000	4 100	64 000
Languages	3	3	1	3	1	3
Registrations	21	38	15	30	31	18
Countries	17	24	11	21	23	14
Run submissions	14	18	9	11	13	3
Notebook submissions	11	17	5	11	8	3
Followers (mailing list)	78	151		181		

Sponsorship by YAHOO! Research.

Media coverage on German and Spanish television, among others.

More about PAN

Program 2011

Today	16:30	Poster Session
Wednesday	10:30	Vandalism Detection
	11:00	Authorship Identification
	14:30	Keynote: *Linguists' Achievements and Analysis Challenges*
		María Teresa Turell and Malcolm Coulthard
	15:10	Panel Discussion
Thursday	9:00	Plagiarism Detection
	11:30	Reports from the Labs

Quo Vadis PAN?

Quo Vadis PAN?

Ideas for Future Editions

- ❑ Hide plagiarism cases in a really large corpus such as ClueWeb.
- ❑ Provide a unified experimentation platform for all participants.
  - ➜ Simplify participation.
  - ➜ Equalize implementation / hardware issues.
- ❑ Add “semantic” challenges.
  - ➜ Distinguish improper text reuse from correct citations.
  - ➜ Find “excuse” citations.
- ❑ Scale up evaluation corpora for authorship identification.
  - ➜ Different genres, languages, and time periods.
  - ➜ Focus on specific task variants.
- ❑ Compile significantly more training data for vandalism detection.

Thank you!

Visit us at pan.webis.de.

Mail us at pan@webis.de.