

Overview of the 4th International Competition on Plagiarism Detection

Martin Potthast
Tim Gollub
Matthias Hagen
Jan Graßegger
Johannes Kiesel
Maximilian Michel
Arnd Oberländer
Martin Tippmann
Benno Stein

Webis Group
Bauhaus-Universität Weimar
www.webis.de

Parth Gupta
Paolo Rosso

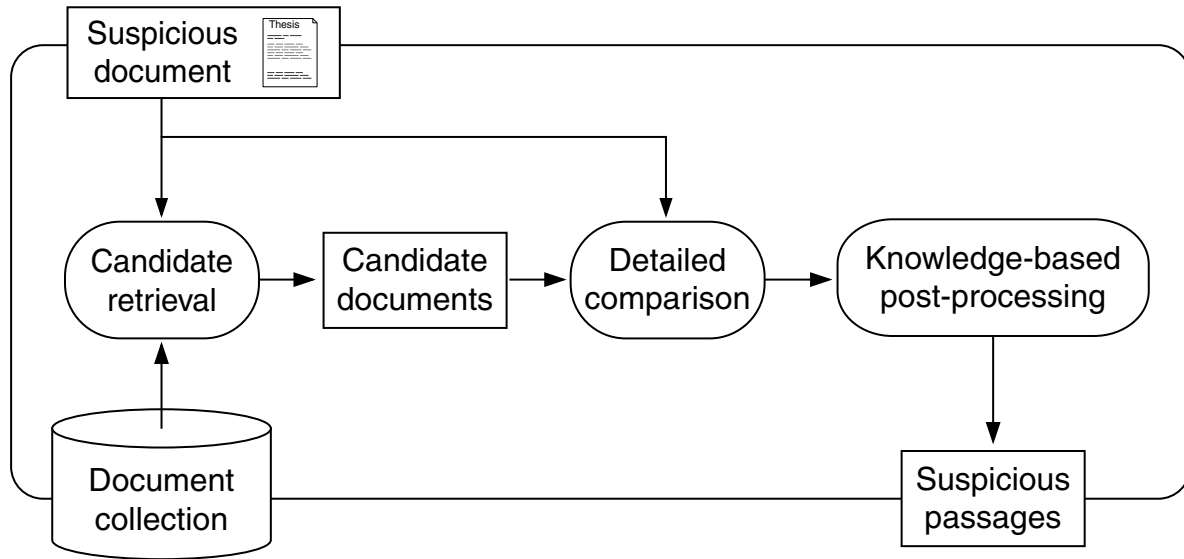
NLEL Group
Universitat Politècnica de València
www.dsic.upv.es/grupos/nle

Alberto Barrón-Cedeño

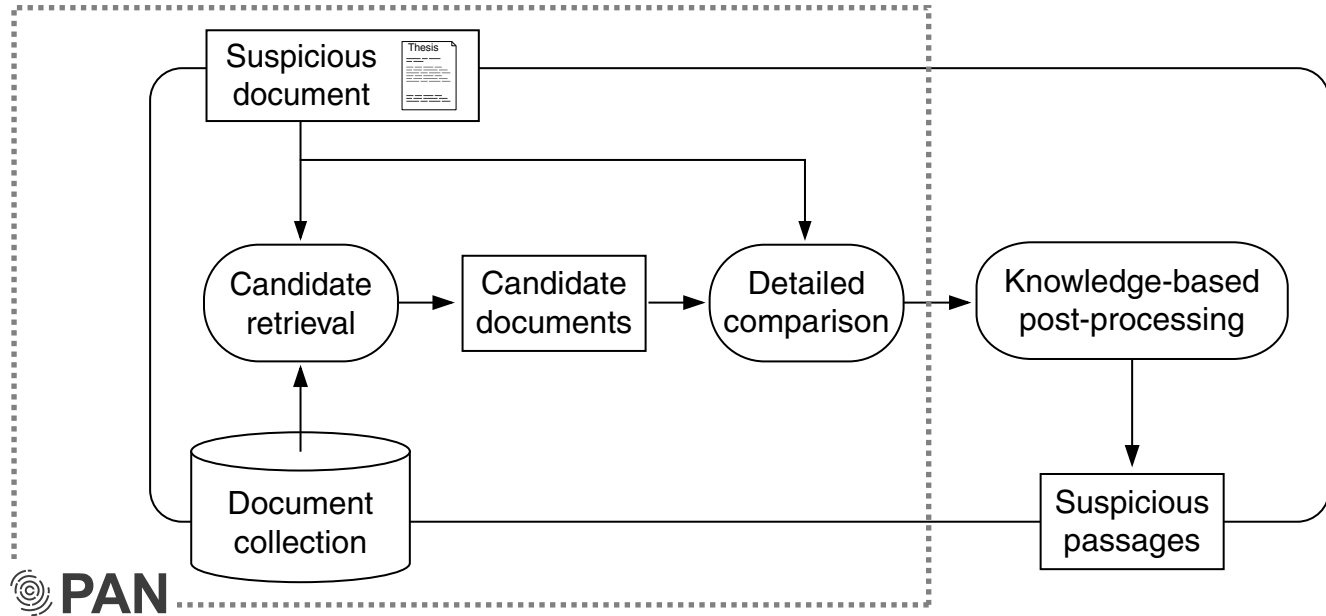
LSI Group
Universitat Politècnica de Catalunya
www.lsi.upc.edu

Introduction

Introduction



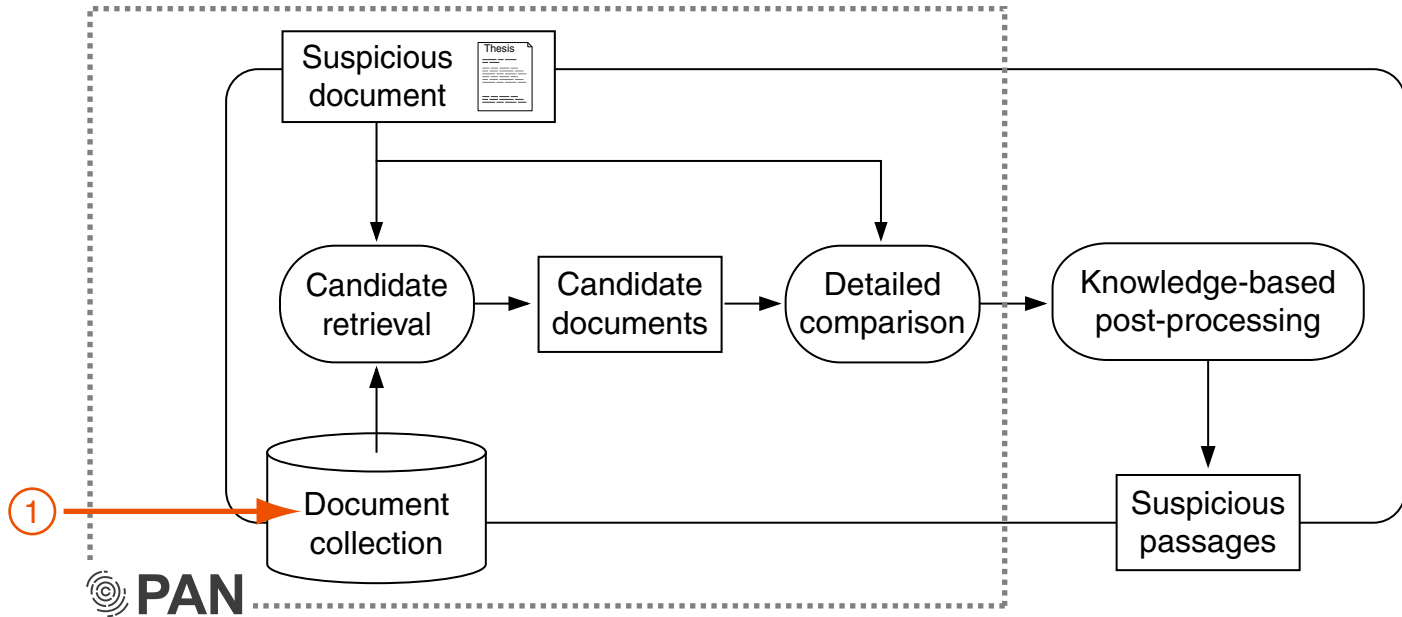
Introduction



Observations, problems:

1. Representativeness: the corpus consists of books, many of which are very old, whereas today the web is the predominant source for plagiarists.
2. Scale: the corpus is too small to enforce a true candidate retrieval situation; most participants did a complete detailed comparison on all $O(n^2)$ document pairs.
3. Realism: plagiarized passages consider not the surrounding document, paraphrasing mostly done by machines, the Web is not used as source.
4. Comparability: evaluation frameworks must be developed, too, and ours kept changing over the years, rendering the obtained results incomparable across years.

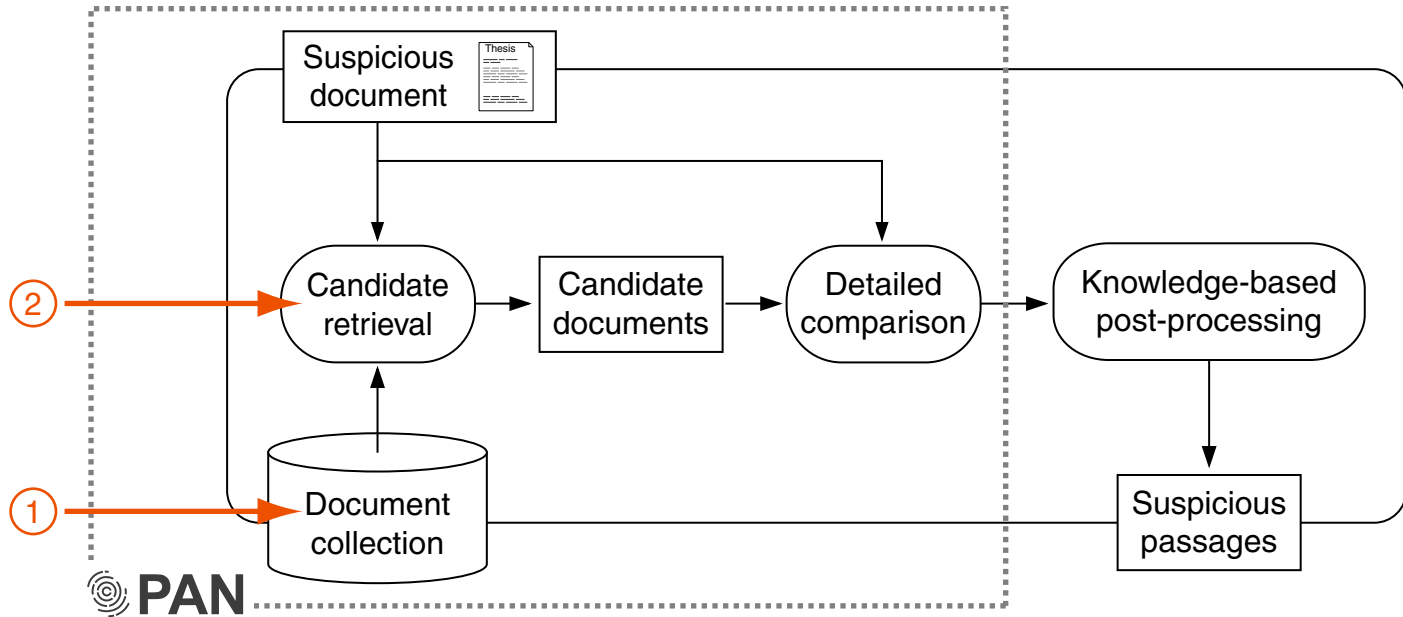
Introduction



Observations, problems:

1. Representativeness: the corpus consists of books, many of which are very old, whereas today the web is the predominant source for plagiarists.
2. Scale: the corpus is too small to enforce a true candidate retrieval situation; most participants did a complete detailed comparison on all $O(n^2)$ document pairs.
3. Realism: plagiarized passages consider not the surrounding document, paraphrasing mostly done by machines, the Web is not used as source.
4. Comparability: evaluation frameworks must be developed, too, and ours kept changing over the years, rendering the obtained results incomparable across years.

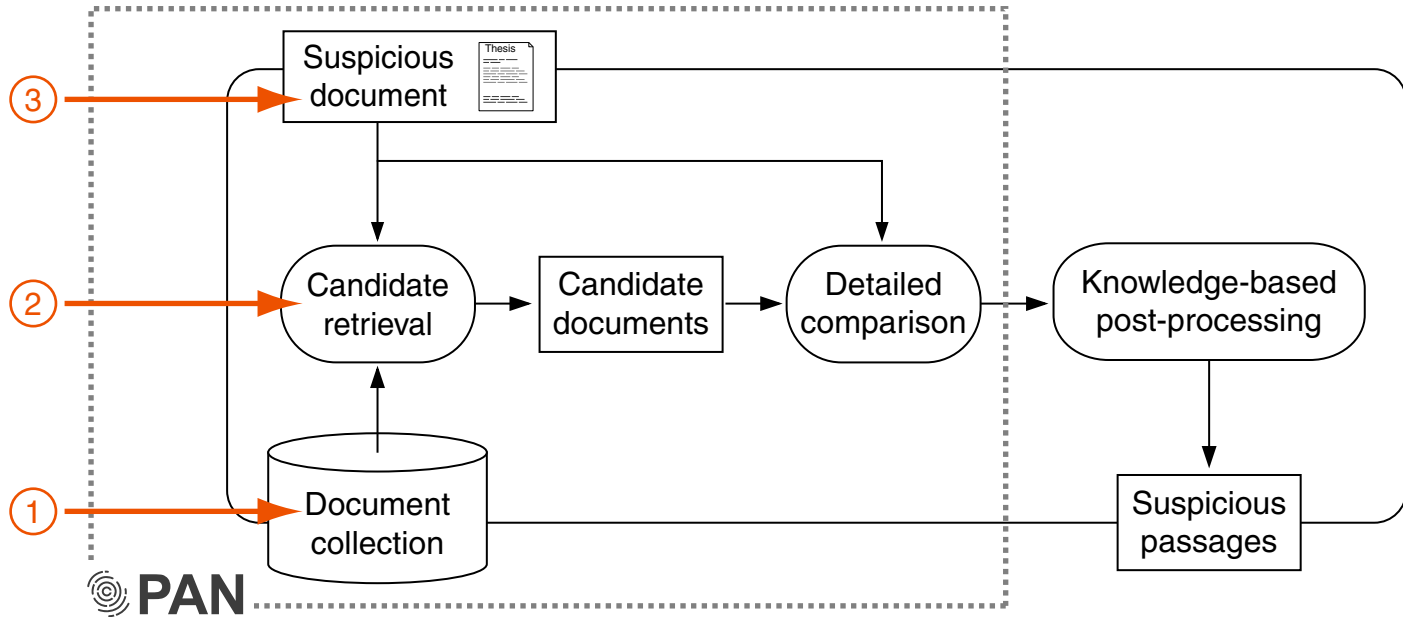
Introduction



Observations, problems:

1. Representativeness: the corpus consists of books, many of which are very old, whereas today the web is the predominant source for plagiarists.
2. Scale: the corpus is too small to enforce a true candidate retrieval situation; most participants did a complete detailed comparison on all $O(n^2)$ document pairs.
3. Realism: plagiarized passages consider not the surrounding document, paraphrasing mostly done by machines, the Web is not used as source.
4. Comparability: evaluation frameworks must be developed, too, and ours kept changing over the years, rendering the obtained results incomparable across years.

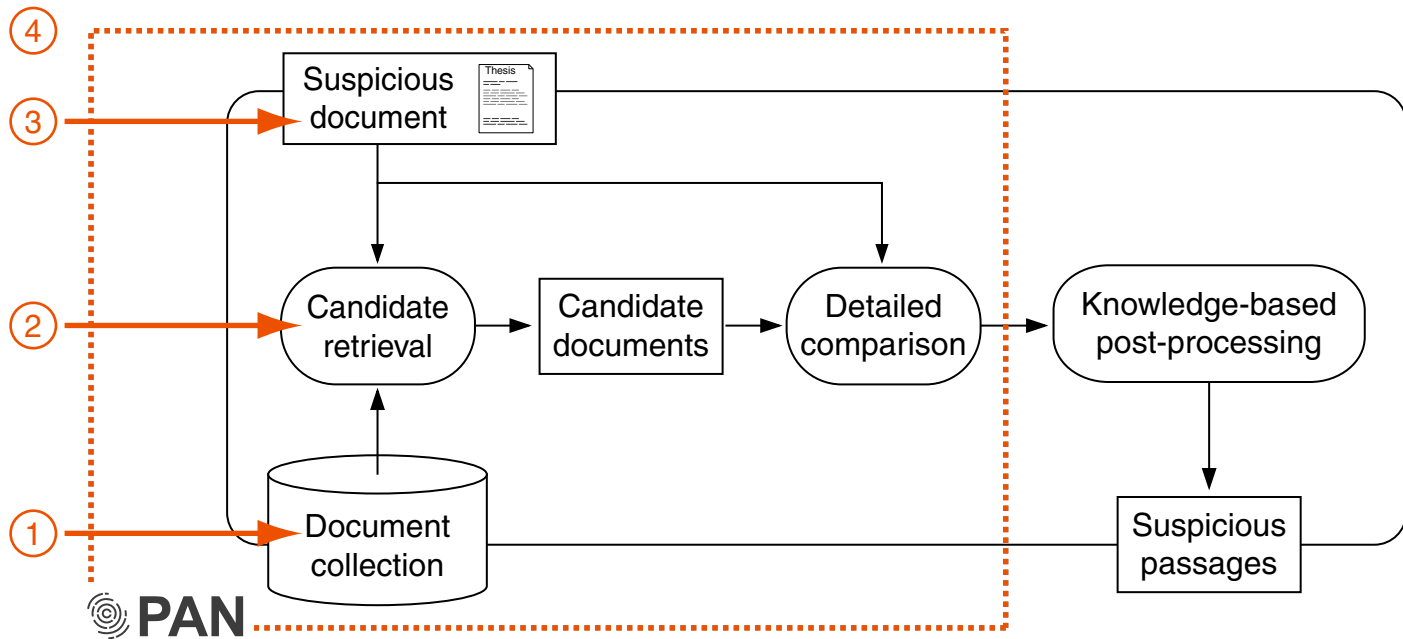
Introduction



Observations, problems:

1. Representativeness: the corpus consists of books, many of which are very old, whereas today the web is the predominant source for plagiarists.
2. Scale: the corpus is too small to enforce a true candidate retrieval situation; most participants did a complete detailed comparison on all $O(n^2)$ document pairs.
3. Realism: plagiarized passages consider not the surrounding document, paraphrasing mostly done by machines, the Web is not used as source.
4. Comparability: evaluation frameworks must be developed, too, and ours kept changing over the years, rendering the obtained results incomparable across years.

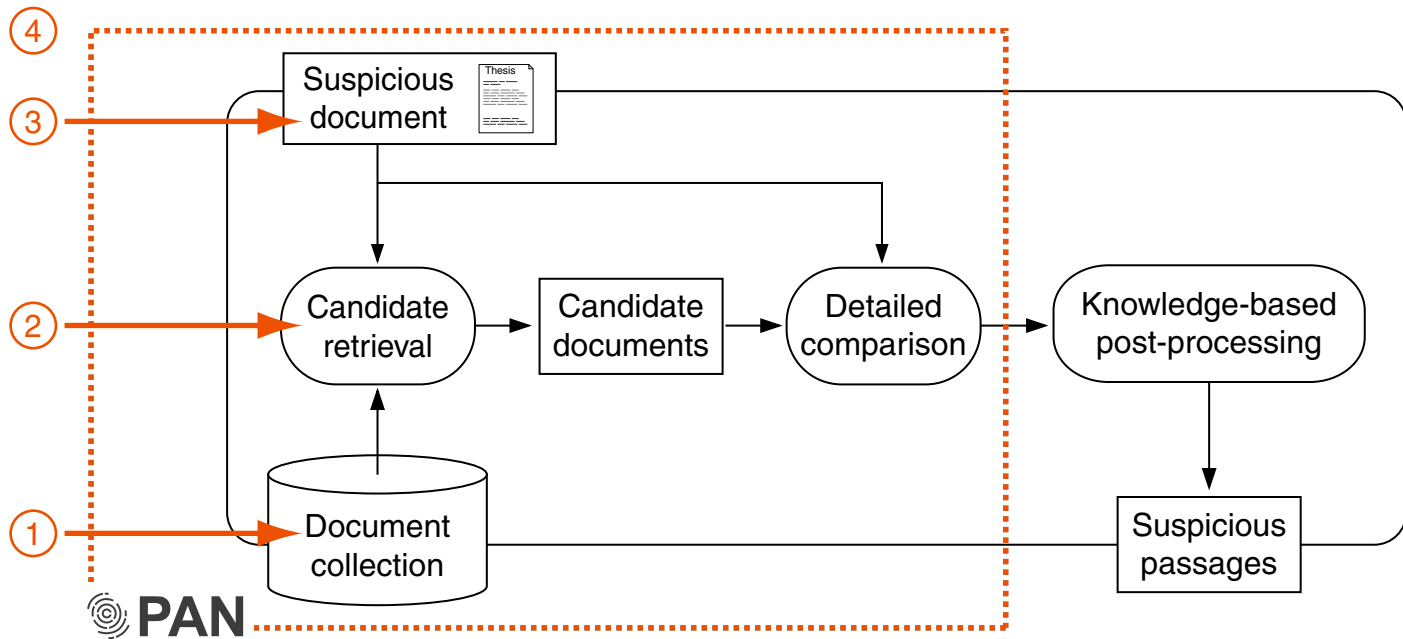
Introduction



Observations, problems:

1. Representativeness: the corpus consists of books, many of which are very old, whereas today the web is the predominant source for plagiarists.
2. Scale: the corpus is too small to enforce a true candidate retrieval situation; most participants did a complete detailed comparison on all $O(n^2)$ document pairs.
3. Realism: plagiarized passages consider not the surrounding document, paraphrasing mostly done by machines, the Web is not used as source.
4. Comparability: evaluation frameworks must be developed, too, and ours kept changing over the years, rendering the obtained results incomparable across years.

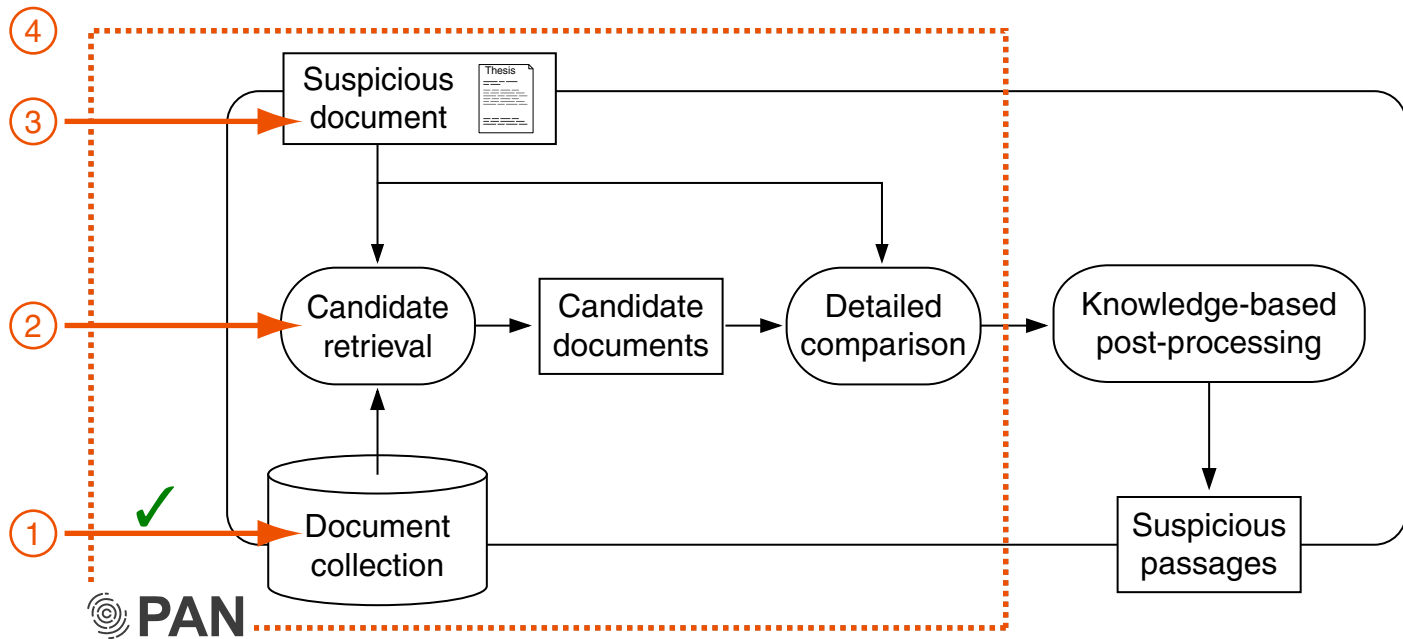
Candidate Retrieval



Considerations:

1. PAN'12 employed the English part of the ClueWeb09 corpus (used in TREC 2009-11 for several tracks) as a static Web snapshot. Size: 500 million web pages, 12.5TB
2. Participants was given efficient corpus access via the API of the ChatNoir search engine. ClueWeb and ChatNoir ensured experiment reproducibility and controllability.
3. The new corpus: manually written digestible texts, topically matching plagiarism cases, Web as source (for document synthesis and plagiarism detection).

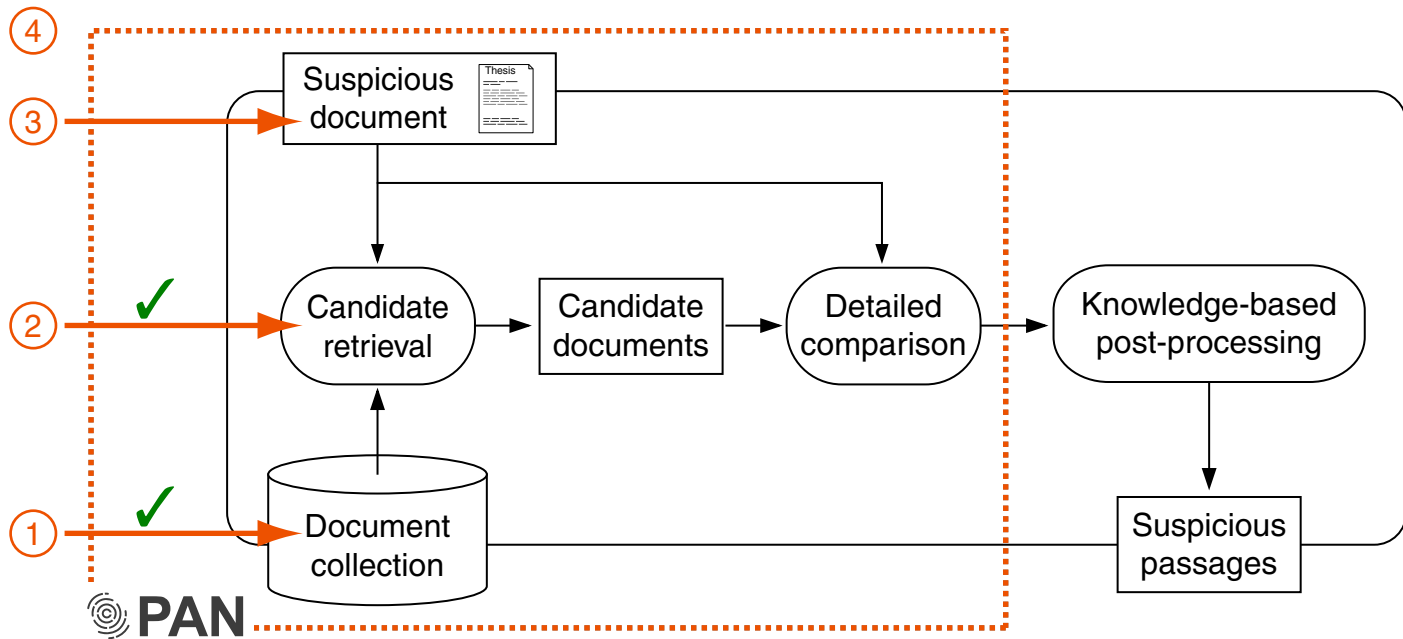
Candidate Retrieval



Considerations:

1. PAN'12 employed the English part of the ClueWeb09 corpus (used in TREC 2009-11 for several tracks) as a static Web snapshot. Size: 500 million web pages, 12.5TB
2. Participants was given efficient corpus access via the API of the ChatNoir search engine. ClueWeb and ChatNoir ensured experiment reproducibility and controllability.
3. The new corpus: manually written digestible texts, topically matching plagiarism cases, Web as source (for document synthesis and plagiarism detection).

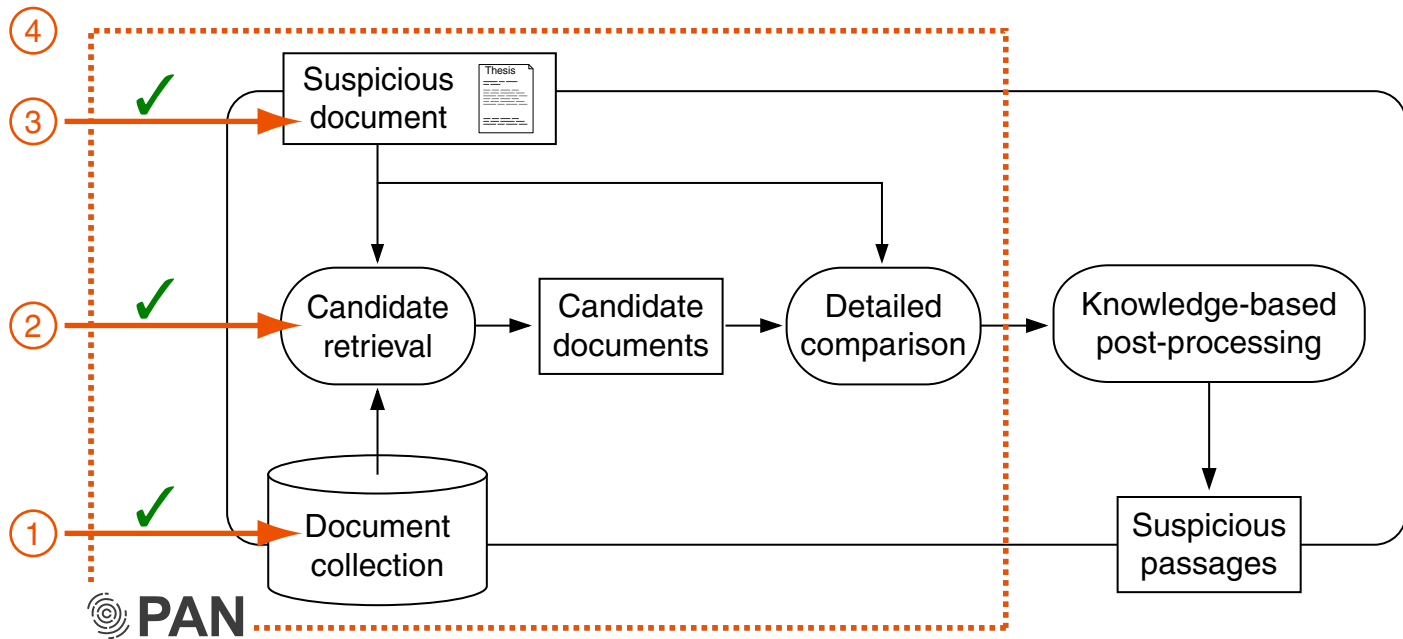
Candidate Retrieval



Considerations:

1. PAN'12 employed the English part of the ClueWeb09 corpus (used in TREC 2009-11 for several tracks) as a static Web snapshot. Size: 500 million web pages, 12.5TB
2. Participants was given efficient corpus access via the API of the ChatNoir search engine. ClueWeb and ChatNoir ensured experiment reproducibility and controllability.
3. The new corpus: manually written digestible texts, topically matching plagiarism cases, Web as source (for document synthesis and plagiarism detection).

Candidate Retrieval

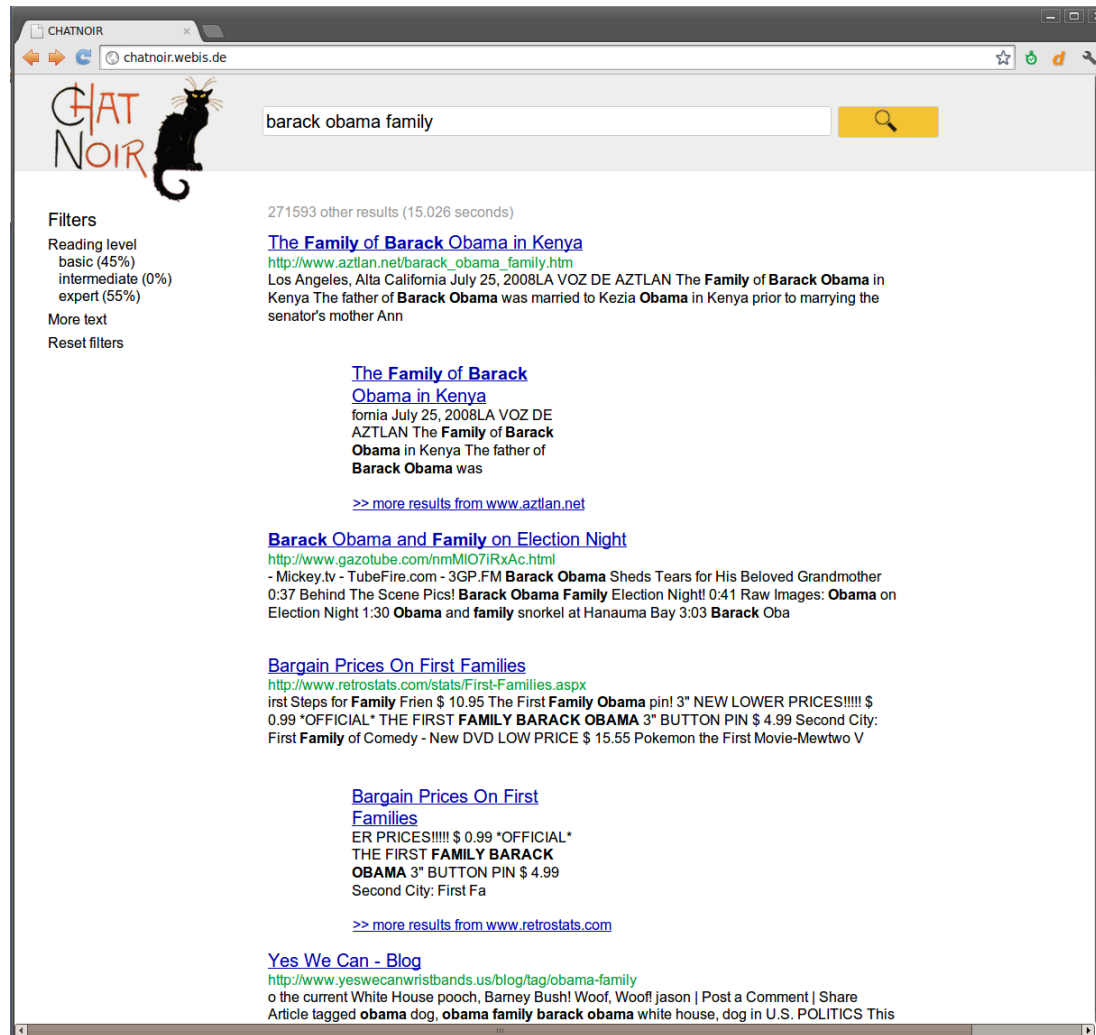


Candidate retrieval task:

- ❑ Humans write essays on given topics, plagiarizing from the ClueWeb, using the ChatNoir search engine for research.
- ❑ Detectors use ChatNoir to retrieve candidate documents from the ClueWeb.
- ❑ Detectors are expected to maximize recall, but use ChatNoir in a cost-effective way.

Candidate Retrieval

About ChatNoir [chatnoir.webis.de]



The screenshot shows a web browser window with the address bar displaying 'chatnoir.webis.de'. The ChatNoir logo, featuring a black cat silhouette, is in the top left. A search bar contains the text 'barack obama family' with a magnifying glass icon to its right. Below the search bar, a filter sidebar on the left lists 'Filters' with options for 'Reading level' (basic (45%), intermediate (0%), expert (55%)), 'More text', and 'Reset filters'. The main content area displays search results. At the top, it says '271593 other results (15.026 seconds)'. The first result is titled 'The Family of Barack Obama in Kenya' with a green URL 'http://www.aztlan.net/barack_obama_family.htm'. The text below the title describes Barack Obama's family background in Kenya. The second result is titled 'The Family of Barack Obama in Kenya' with a green URL 'http://www.gazolube.com/nmMIO7iRxAc.html'. The text below the title mentions Mickey TV and Barack Obama's election night. The third result is titled 'Bargain Prices On First Families' with a green URL 'http://www.retrostats.com/stats/First-Families.aspx'. The text below the title lists various items for sale, including a Family Friend, a Barack Obama button, and a DVD. The fourth result is titled 'Bargain Prices On First Families' with a green URL 'http://www.retrostats.com'. The text below the title lists various items for sale, including a Family Friend, a Barack Obama button, and a DVD. The fifth result is titled 'Yes We Can - Blog' with a green URL 'http://www.yeswecanwristbands.us/blog/tag/obama-family'. The text below the title mentions the current White House pooch, Barney the Dinosaur, and a link to post a comment or share the article.

CHATNOIR

chatnoir.webis.de

barack obama family

271593 other results (15.026 seconds)

[The Family of Barack Obama in Kenya](#)
http://www.aztlan.net/barack_obama_family.htm
Los Angeles, Alta California July 25, 2008LA VOZ DE AZTLAN The **Family of Barack Obama** in Kenya The father of **Barack Obama** was married to Kezia **Obama** in Kenya prior to marrying the senator's mother Ann

[The Family of Barack Obama in Kenya](#)
fornia July 25, 2008LA VOZ DE AZTLAN The **Family of Barack Obama** in Kenya The father of **Barack Obama** was

[>> more results from www.aztlan.net](#)

[Barack Obama and Family on Election Night](#)
<http://www.gazolube.com/nmMIO7iRxAc.html>
- Mickey.tv - TubeFire.com - 3GP.FM **Barack Obama** Sheds Tears for His Beloved Grandmother 0:37 Behind The Scene Picst **Barack Obama Family** Election Night! 0:41 Raw Images: **Obama** on Election Night 1:30 **Obama** and **family** snorkel at Hanauma Bay 3:03 **Barack Oba**

[Bargain Prices On First Families](#)
<http://www.retrostats.com/stats/First-Families.aspx>
irst Steps for **Family** Frien \$ 10.95 The First **Family Obama** pin! 3" NEW LOWER PRICES!!!! \$ 0.99 *OFFICIAL* THE FIRST **FAMILY BARACK OBAMA** 3" BUTTON PIN \$ 4.99 Second City: First **Family** of Comedy - New DVD LOW PRICE \$ 15.55 Pokemon the First Movie-Mewtwo V

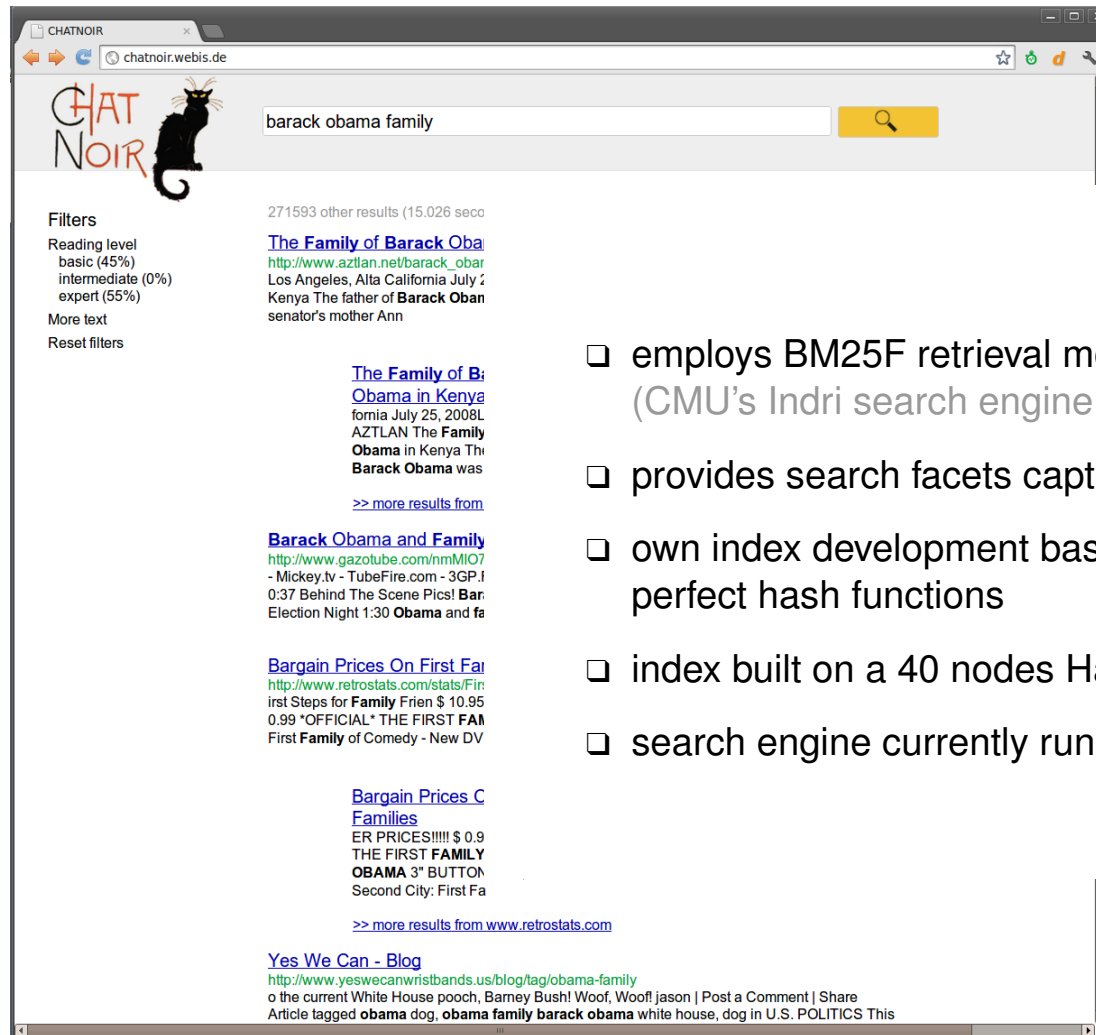
[Bargain Prices On First Families](#)
ER PRICES!!!! \$ 0.99 *OFFICIAL*
THE FIRST **FAMILY BARACK OBAMA** 3" BUTTON PIN \$ 4.99
Second City: First Fa

[>> more results from www.retrostats.com](#)

[Yes We Can - Blog](#)
<http://www.yeswecanwristbands.us/blog/tag/obama-family>
o the current White House pooch, Barney Bush! Woof, Woof! jason | Post a Comment | Share
Article tagged **obama** dog, **obama family** **barack obama** white house, dog in U.S. POLITICS This

Candidate Retrieval

About ChatNoir [chatnoir.webis.de]



The screenshot shows the ChatNoir web interface in a browser window. The search bar contains the text "barack obama family". Below the search bar, there are several search results. On the left side, there are filters for reading level (basic, intermediate, expert) and text type (More text, Reset filters). The search results include links to various articles and videos, such as "The Family of Barack Obama", "Barack Obama and Family", "Bargain Prices On First Family", and "Yes We Can - Blog".

Filters

Reading level
basic (45%)
intermediate (0%)
expert (55%)

More text
Reset filters

271593 other results (15.026 seco

[The Family of Barack Obama](#)
http://www.aztlan.net/barack_obar
Los Angeles, Alta California July 2
Kenya The father of **Barack Obam**
senator's mother Ann

[The Family of Barack Obama in Kenya](#)
fornia July 25, 2008L
AZTLAN The **Family**
Obama in Kenya Thi
Barack Obama was

[>> more results from](#)

[Barack Obama and Family](#)
<http://www.gazolube.com/nmMIO7>
- Mickey.tv - TubeFire.com - 3GP.f
0:37 Behind The Scene Picst Bar
Election Night 1:30 **Obama** and fa

[Bargain Prices On First Family](#)
<http://www.retrostats.com/stats/Fir>
irst Steps for **Family** Frien \$ 10.95
0.99 *OFFICIAL* THE FIRST **FAM**
First **Family** of Comedy - New DV

[Bargain Prices C Families](#)
ER PRICES!!!!!! \$ 0.9
THE FIRST **FAMILY**
OBAMA 3" BUTTON
Second City: First Fa

[>> more results from www.retrostats.com](#)

[Yes We Can - Blog](#)
<http://www.yeswecanwristbands.us/blog/tag/obama-family>
o the current White House pooch, Barney Bush! Woof, Woof! jason | Post a Comment | Share
Article tagged **obama** dog, **obama family** **barack obama** white house, dog in U.S. POLITICS This

- ❑ employs BM25F retrieval model
(CMU's Indri search engine is language-model-based)
- ❑ provides search facets capturing readability issues
- ❑ own index development based on externalized minimal perfect hash functions
- ❑ index built on a 40 nodes Hadoop cluster
- ❑ search engine currently running on 11 machines

About Corpus Construction

Plagiarism Editor

pcstein5.medien.uni-weimar.de:8080/pan-plagiarism-odesk-editor/#task=wt0911001-example

Barack Obama's Family

plagiariized by John Doe

The Family of Barack Obama is an extended clan of African American, English, Indonesian, and Kenyan heritage. They are best known through the writings and political career of Barack Obama, the current President of the United States of America. His immediate family is the First Family of the United States. The Obamas are the first First Family of African American descent in the United States and the youngest to enter the White House since the Kennedys. Obama's young, energetic family harks back todays of Camelot.

<http://webis15.medien.uni-weimar.de/chatnoir/clueweb?id=1000117099993&token=wt0911001-qrel>

In what follows, we give a detailed overview of Barack Obama's Family. We shed light on himself, his immediate and extended family, including maternal and paternal relations. Moreover, we give insights into the relations of Michelle Obama, Barack Obama's wife, as well as some distant relations of both.

Barack Obama

Barack Hussein Obama II is the 44th and current President of the United States. He is the first African American to hold the office. Obama was the Junior United States Senator from Illinois from 2005 until he resigned following his election to the presidency. Obama is a graduate of Columbia University and Harvard Law School. He worked as a community organizer in Chicago prior to earning his law degree, and practiced as a civil rights attorney in Chicago before serving three terms in the Illinois Senate from 1997 to 2004. He also taught Constitutional Law at the University of Chicago Law School from 1992 to 2004. Following an unsuccessful bid for a seat in the U.S. House of Representatives in 2000, Obama was elected to the Senate in November 2004.

Barack Obama was born on 4 August 1961 at the Kapʻolani Medical Center for Women & Children in Honolulu, Hawaii, to Ann Dunham of Wichita, Kansas and Barack Hussein Obama, Sr. of Nyangoma-Kogelo, Nyanza Province, Kenya. His parents met while both were attending the East-West Center of the University of Hawaii at Manoa, where his father was enrolled as a foreign student. The couple married only 6 months earlier on February 2, 1961.

Childhood and Youth

When Barack Obama was two years old, his parents divorced and his father moved to Connecticut to continue his education before returning to Kenya. He saw his son only once more before dying in an automobile accident in 1982.

When Obama was six, his mother married Lolo Soetoro, an Indonesian oil manager. In 1967, when Soetoro's student visa was revoked because of political unrest in Indonesia, Durham and Barack, then in first grade, accompanied him to Jakarta,

Instructions

Write a text about the topic specified below. The text shall contain passages which are plagiarized from different web pages.

- Search for sources matching the topic using the [ChatNoir search engine](#). Do not use any other search engine!
- Once you found a passage of text to plagiarize, copy it into your text.
- Change the background color of the copied passage. Also, add a link to the source web page with the same background color. This is so we can follow up on your work.
- Modify the plagiarizd passage so that an automatic plagiarism detector (like Turnitin) won't be able to detect it.
- Repeat these steps until your text is complete.

Remarks:

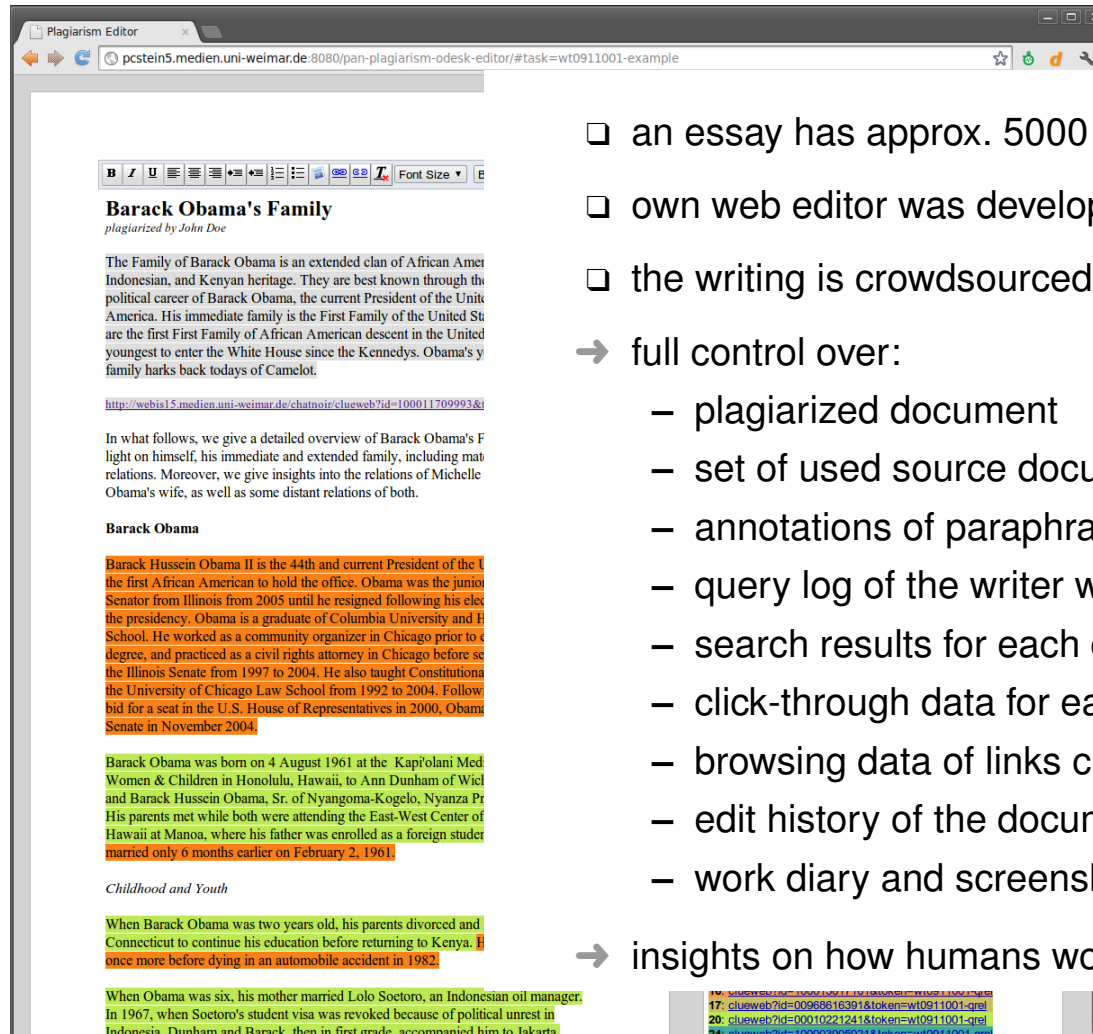
- The text shall be at least 5000 words long.
- It shall contain a couple of plagiarized passages.
- You shall also write some passages yourself.
- You may choose the text genre: an essay, a news article, a press release, a blog post, an advertisement etc.
- You may follow links on web pages found via the search engine.
- While modifying and rewriting a plagiarized passage, you may mix it with others, delete things, or add sentences.

Use the editor on the left to write your text. Do not use any other editor. Your text will be frequently saved on our servers. In case of errors, you will be notified in the status message below. Report errors back to us before you continue writing.

Topic	Links	Contact
Obama's family tree. Write about President Barack Obama's family history, including genealogy, national origins, places and dates of birth, etc. For instance: where did Barack Obama's parents and grandparents come from; what did his mother work; etc.	ChatNoir Search	pan@webis.de
Status	Word Count	
Document saved	7955	
Color Key		
4: clueweb?tid=1000117099993&token=wt0911001-qrel 5: clueweb?tid=100001018508&token=w0911001-qrel 6: clueweb?tid=00258906994&token=w0911001-qrel 8: clueweb?tid=100017805020&token=w0911001-qrel 9: clueweb?tid=1000077046098&token=w0911001-qrel 10: clueweb?tid=100012205862&token=w0911001-qrel 16: clueweb?tid=100013617161&token=w0911001-qrel 17: clueweb?tid=00968616391&token=w0911001-qrel 20: clueweb?tid=100010221241&token=w0911001-qrel		

Candidate Retrieval

About Corpus Construction



The screenshot shows a web browser window titled 'Plagiarism Editor' with the URL 'pcstein5.medien.uni-weimar.de:8080/pan-plagiarism-odesk-editor/#task=wt0911001-example'. The main content area displays a document titled 'Barack Obama's Family' by John Doe. The document text includes a paragraph about the Obama family's heritage, a link to a chatbot, and several paragraphs about Barack Obama's life and career. The text is annotated with various markers: orange highlights for specific sentences, green highlights for other sentences, and red underlines for certain words. A sidebar on the right contains a list of links, each with a unique identifier and a token.

Barack Obama's Family
plagiarized by John Doe

The Family of Barack Obama is an extended clan of African American, Indonesian, and Kenyan heritage. They are best known through the political career of Barack Obama, the current President of the United States. His immediate family is the First Family of the United States. They are the first First Family of African American descent in the United States to enter the White House since the Kennedys. Obama's family harks back to the days of Camelot.

<http://webis15.medien.uni-weimar.de/chatbot/clueweb?id=100011709993&>

In what follows, we give a detailed overview of Barack Obama's Family. We focus on his immediate and extended family, including maternal relations. Moreover, we give insights into the relations of Michelle Obama's wife, as well as some distant relations of both.

Barack Obama

Barack Hussein Obama II is the 44th and current President of the United States. He is the first African American to hold the office. Obama was the junior U.S. Senator from Illinois from 2005 until he resigned following his election to the presidency. Obama is a graduate of Columbia University and Harvard Law School. He worked as a community organizer in Chicago prior to his presidency. He practiced as a civil rights attorney in Chicago before his election to the Illinois Senate from 1997 to 2004. He also taught Constitutional Law at the University of Chicago Law School from 1992 to 2004. Following his election to the U.S. House of Representatives in 2000, Obama was elected to the U.S. Senate in November 2004.

Barack Obama was born on 4 August 1961 at the Kapi'olani Medical Center for Women & Children in Honolulu, Hawaii, to Ann Dunham of Wichita, Kansas, and Barack Hussein Obama, Sr. of Nyangoma-Kogelo, Nyanza Province, Kenya. His parents met while both were attending the East-West Center of the University of Hawaii at Manoa, where his father was enrolled as a foreign student. His father married only 6 months earlier on February 2, 1961.

Childhood and Youth

When Barack Obama was two years old, his parents divorced and he moved to Honolulu with his mother. He continued his education before returning to Kenya. He was once more before dying in an automobile accident in 1982.

When Obama was six, his mother married Lolo Soetoro, an Indonesian oil manager. In 1967, when Soetoro's student visa was revoked because of political unrest in Indonesia, Dunham and Barack then in first grade, accompanied him to Jakarta.

17: clueweb?id=00968616391&token=wt0911001-qrel
20: clueweb?id=00010221241&token=wt0911001-qrel
24: clueweb?id=10003300021&token=wt0911001-qrel

- ❑ an essay has approx. 5000 words which means 8-10 pages
- ❑ own web editor was developed for essay writing
- ❑ the writing is crowdsourced via oDesk
- ➔ full control over:
 - plagiarized document
 - set of used source documents
 - annotations of paraphrased passages
 - query log of the writer while researching the topic
 - search results for each query
 - click-through data for each query
 - browsing data of links clicked within ClueWeb
 - edit history of the document covering all keystrokes
 - work diary and screenshots as recorded by oDesk
- ➔ insights on how humans work when reusing text

Candidate Retrieval

Survey of Approaches

An analysis of the participants' notebooks reveals a candidate retrieval process:

1. Chunking

Given a suspicious document, it is divided into (possibly overlapping) passages of text. Each chunk of text is then processed individually.

2. Keyphrase Extraction

Given a chunk (or the entire suspicious document), keyphrases are extracted from it in order to formulate queries with them.

3. Query Formulation

Given sets of keywords extracted from chunks, queries are formulated which are tailored to the API of the search engine used.

4. Search Control

Given a set of queries, the search controller schedules their submission to the search engine and directs the download of search results.

5. Download Filtering

Given a set of downloaded documents, all documents are removed that are not worthwhile for detailed comparison to the suspicious document.

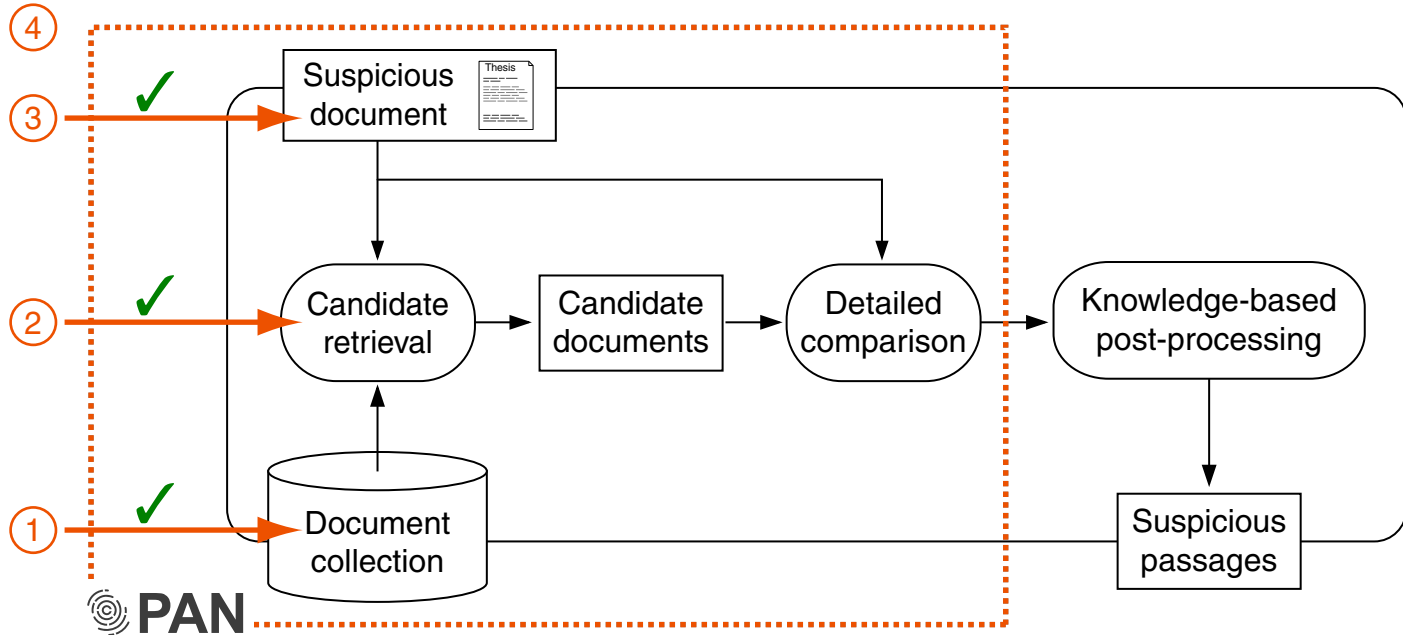
Candidate Retrieval

Evaluation Results

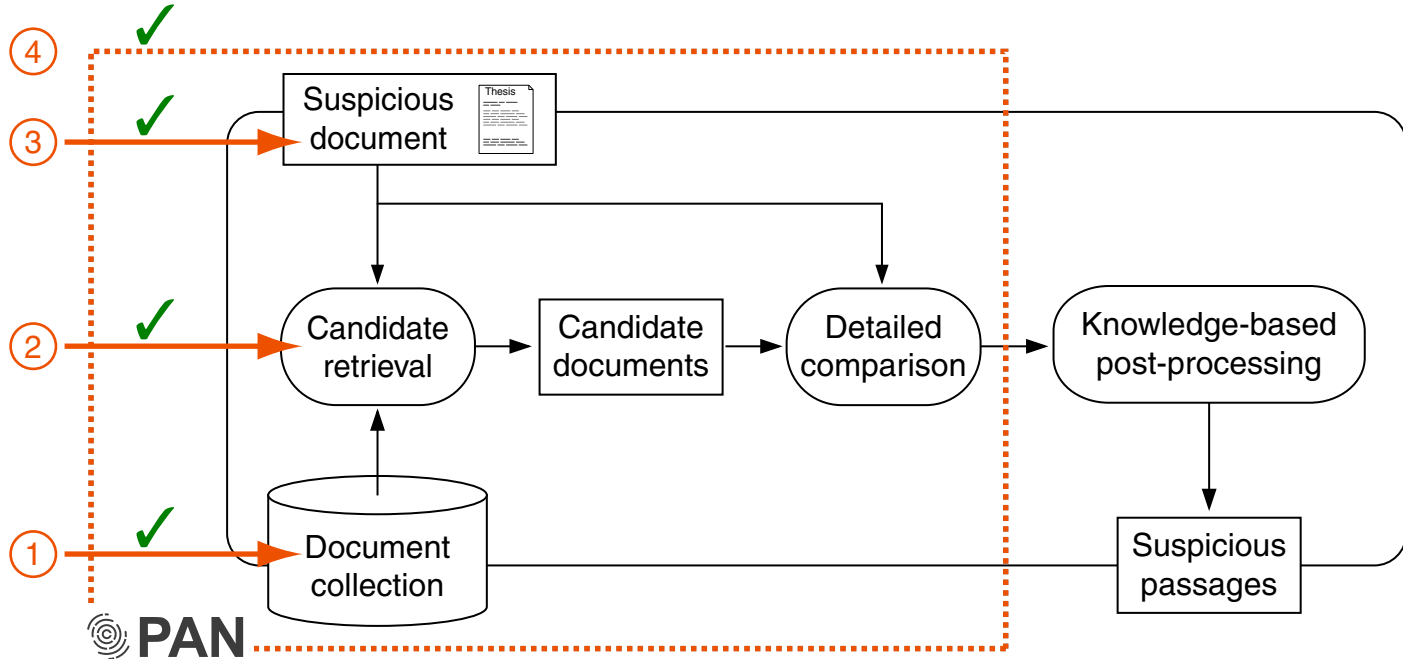
Team	Total Workload		Time to 1st Detection		Reported Sources		Downloaded Sources	
	Queries	Dwnlds	Queries	Dwnlds	Precision	Recall	Precision	Recall
Gillam	63	527	5	26	0.63	0.25	0.01	0.56
Jayapal	67	174	9	14	0.66	0.28	0.07	0.43
Kong	551	327	81	28	0.57	0.24	0.02	0.37
Palkovskii	63	1027	27	319	0.44	0.12	0.00	0.21
Suchomel	13	95	6	2	0.52	0.21	0.08	0.35

- ❑ Suchomel et al. implement the best tradeoff between cost and quality.
- ❑ Jayapal implements the best approach in terms of precision and recall.

Detailed Comparison



Detailed Comparison



Detailed comparison task:

- ❑ Detectors are presented with a suspicious and a candidate document, and are asked to extract the plagiarized passages.
- ❑ Developers submit their detection softwares instead of detection results.
- ❑ This allows for re-evaluating detectors, as well as to measure runtime and to use private corpora.

Detailed Comparison

Software Submissions and Runtime Analysis

- ❑ Eleven participants, about the average number from last years.
- ➔ Software submissions do not distract people from participating.

Team	Submission Size [MB]	Operating System	Programming Language	Average Runtime [sec/comparison]
Rodríguez Torrejón	1.80	Linux	sh, C/C++	0.19
Sánchez-Vega	0.04	Linux	C++	2.48
Oberreuter	0.19	Linux	Java	2.58
Palkovskii	68.20	Windows	C#	4.51
Grozea	1.90	Linux	Perl, Octave	4.82
Suchomel	0.02	Linux	Perl	5.36
Kong	2.60	Linux	Java	5.91
Jayapal	37.20	Linux	Java	8.43
Gillam	0.48	Linux	Python 2.7	9.40
Küppers	42.90	Linux	Java	27.64
Ghosh	554.50	Linux	sh, Java	—

- ➔ Congratulations to Rodríguez Torrejón et al. for submitting the most efficient detailed comparison program.

Detailed Comparison

Survey of Approaches

An analysis of the participants' notebooks reveals a detailed comparison process:

1. Seeding

Given a suspicious document and a source document, matches (also called „seeds”) between the two documents are identified using some seed heuristic. Seed heuristics either identify exact matches or *create* matches by changing the underlying texts in a domain-specific or linguistically motivated way.

2. Match Merging


Given seed matches identified between a suspicious document and a source document, they are merged into aligned text passages of maximal length between the two documents which are then reported as plagiarism detections.

3. Passage Filtering


Given a set of aligned passages, a passage filter removes all aligned passages that do not meet certain criteria.


Detailed Comparison

TIRA evaluation platform

 **τιρα** pan2012/evaluation-service

- testset +
- detector +
- detector -
- detector -


Windows7


Ubuntu12.04

Testset	Detector	Precision	Recall	Gran	Plagdet	Avg Time	Documents	Output Dir
08_all	kongleilei12	0.8249	0.6782	1.0109	0.7386	5.9187	3033	output-dir
08_all	kasprzak12	0.8931	0.5524	1.0000	0.6826	5.3679	3033	output-dir
08_all	torrejon12	0.8344	0.5004	1.0009	0.6252	0.1900	3033	output-dir

- ❑ TIRA takes locally executable programs and turns them into web services.
- ❑ TIRA assumes responsibility for storing and indexing of execution results.
- ❑ For the PAN evaluation, TIRA servers are provided for two operating systems, Windows and Ubuntu.
- ❑ Participants submit their plagiarism detection software for deployment on the appropriate TIRA server.
- ❑ A third TIRA server controls the overall evaluation of all deployed submissions on the private test set and provides the overall results.

[tira@localhost] [tira@buw]

Detailed Comparison

Evaluation Corpus Construction

- ❑ Like in last years based on books from Project Gutenberg.
- ❑ Divided into seven sub-corpora:

Evaluation Corpus Statistics		
Sub-Corpus	Number of Cases	Avg. Cosine Similarity
Real Cases	33	0.161
Simulated	500	0.364
Translation ($\{\text{de, es}\} \rightarrow \text{en}$)	500	0.018
Artificial (High)	500	0.392
Artificial (Low)	500	0.455
No Obfuscation	500	0.560
No Plagiarism	500	0.431
Overall	3033	0.369

- ❑ Similarity of document pairs was taken into account this year.
- ❑ Real Cases were taken from the Web. Cross-Language cases were constructed using the multi-lingual Europarl corpus.

Detailed Comparison

Evaluation Results: Overall Performance

Rank / Team	PlagDet	Precision	Recall	Granularity
1 Kong	0.738	0.824	0.678	1.01
2 Suchomel	0.682	0.893	0.552	1.00
3 Grozea	0.678	0.774	0.635	1.03
4 Oberreuter	0.673	0.867	0.555	1.00
5 Rodríguez Torrejón	0.625	0.834	0.500	1.00
6 Palkovskii	0.538	0.574	0.523	1.02
7 Küppers	0.349	0.776	0.282	1.26
8 Sánchez-Vega	0.309	0.537	0.349	1.57
9 Gillam	0.308	0.898	0.190	1.02
10 Jayapal	0.045	0.622	0.075	6.93

→ Congratulations to Kong et al. for submitting the most effective detailed comparison program.

Summary and Outlook

PAN 2012:

- ❑ Task-wise evaluation of plagiarism detectors.
 - ❑ Candidate document retrieval at Web scale using ChatNoir.
 - ❑ Software submissions for sustainable / repeatable evaluation using TIRA.
 - ❑ More realistic plagiarism corpus.
 - ❑ New performance measures in addition to the traditional ones.
- A lot of fun!

Thanks to everyone who volunteered to test our new setup!

PAN 2013 and beyond:

- ❑ Improvement and consolidation of the new tools.
 - ❑ Use of the plagiarism corpus for detailed comparison as well.
 - ❑ Community process to collect more plagiarism (real and manual).
- Fully automatic plagiarism detection evaluations.