Detection of Text Plagiarism and Wikipedia Vandalism

Benno Stein
Bauhaus-Universität Weimar
www.webis.de

Keynote at SEPLN, Valencia, 8. Sep. 2010

The webis Group

```
  *****
 *     *
|-o-o-|
|  ^  |
 | - /
  --
```

Benno Stein

```
 """"""
" -OO-"
|   } |
 \  - /
   ---
```

Dennis Hoppe

```
  ****
 * o o*
**  ^ *
 |  _ |
  \___|
```

Maik Anderka

```
 #####
#    ##
|Ö-Ö--|
| '_ /
 \__/
```

Martin Potthast

```
 ******
****  *
|0  0 |
|  Y  |
 \ = /
```

Matthias Hagen

```
 *****
*  ** *
|-o-o-|
\  ^  /
 \ - /
```

Nedim Lipka

```
  ___
 ///.\
 //  >
,|  \_
```

Peter Prettenhofer

```
  ___
 "   '
 |'  '|
 | &  |
 |\_/|
  ---
```

Tim Gollub

```
   :::
  :::::
 :: ää :.
 ::  . ::
 ::\ -/::
.:      :.
```

Christin Gläser

The webis Group

Benno Stein

Dennis Hoppe

Maik Anderka

Martin Potthast

Matthias Hagen

```
  *****
 *  **  *
 |-o-o-|
 \   ^   /
  \  -  /
```

Nedim Lipka

Peter Prettenhofer

Tim Gollub

Christin Gläser

The webis Group

```
  *****
 *     *
|-o-o-|
|  ^  |
 | - /
   --
```

Benno Stein

```
 """"""
"  -OO-"
|    }  |
 \   -  /
   ---
```

Dennis Hoppe

```
   ****
  * o o*
 **   ^  *
 |    _  |
  \____|
```

Maik Anderka

Martin Potthast

```
 ******
*****   *
|O   O|
|  Y  |
 \ = /
```

Matthias Hagen

```
  *****
 *  ** *
|-o-o-|
\  ^  /
 \ - /
```

Nedim Lipka

```
  ___
 ///.\
 //  >
,|  \_
```

Peter Prettenhofer

```
  ____
 "    '
 |'  '|
 |  & |
 |\_/|
  ---
```

Tim Gollub

```
   :::
  :::::
 :: ää :.
 ::  . ::
 ::\ -/::
.:      :.
```

Christin Gläser

Outline

- External Plagiarism Detection
- The PAN Competition
- Intrinsic Plagiarism Detection

- Vandalism Detection in Wikipedia
- The PAN Competition Continued

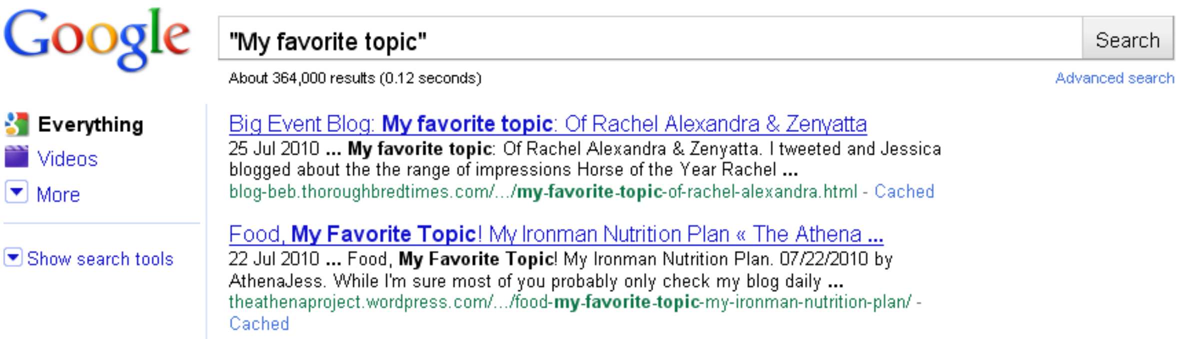
-  + 

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 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]

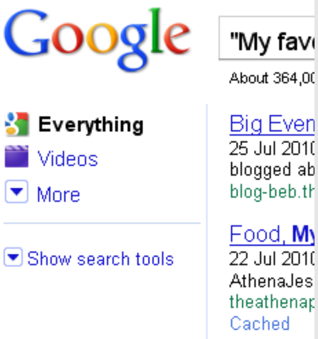
... better technology nowadays ;–)



+

WIKIPEDIA
The Free Encyclopedia

... better technology nowadays ;–)



?

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

- Is plagiarism a problem with respect to education?

- Is there a misunderstanding wrt. an evolving cultural technique?

  (Netspeak—a service that exploits the unacknowledged wisdom of the crowd.)

- Can plagiarsim be detected by humans?

- Can plagiarsim be detected by machines?

- Should automatic plagiarism detection algorithms become standard?

- Is plagiarism a problem with respect to education?

- Is there a misunderstanding wrt. an evolving cultural technique?

  (Netspeak—a service that exploits the unacknowledged wisdom of the crowd.)

- Can plagiarsim be detected by humans?

- Can plagiarsim be detected by machines?

- Should automatic plagiarism detection algorithms become standard?

- Is plagiarism a problem with respect to education?

- Is there a misunderstanding wrt. an evolving cultural technique?

  (Netspeak—a service that exploits the unacknowledged wisdom of the crowd.)

- Can plagiarsim be detected by humans?

- Can plagiarsim be detected by machines?

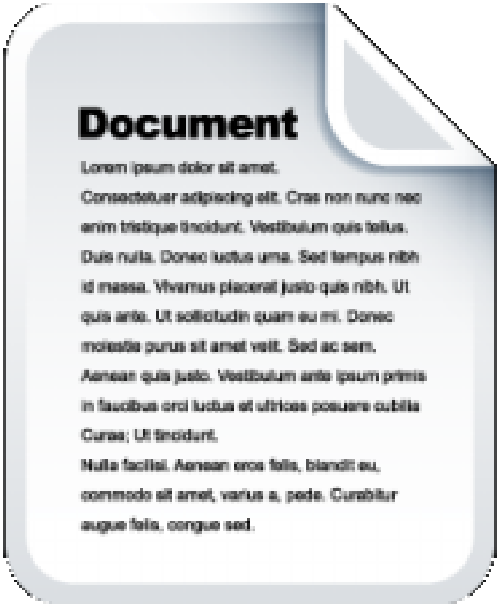
- Should automatic plagiarism detection algorithms become standard?

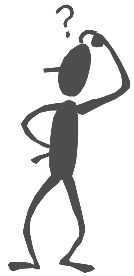
For several reasons we should say “text reuse” rather than “plagiarism”.

External Plagiarism Detection

External Plagiarism Detection

How Humans Spot Plagiarism





External Plagiarism Detection

How Humans Spot Plagiarism

External Plagiarism Detection

How Humans Spot Plagiarism

+ Exploits human intuition for peculiar passages.

+ Exploits human experience to analyze the search engine results.

+ Is applied easily and in an ad-hoc manner.

External Plagiarism Detection

How Humans Spot Plagiarism

+ Exploits human intuition for peculiar passages.
+ Exploits human experience to analyze the search engine results.
+ Is applied easily and in an ad-hoc manner.

– Cannot be done on large scale.
– Depends on (commercial) third-party services.
– Fails in the case of obfuscated / modified text.
– Cannot be used to find the reuse of structure or argumentation lines.

External Plagiarism Detection

Algorithms for Machines

External Plagiarism Detection

Algorithms for Machines

External Plagiarism Detection

Algorithms for Machines

Keyword Extraction from Document

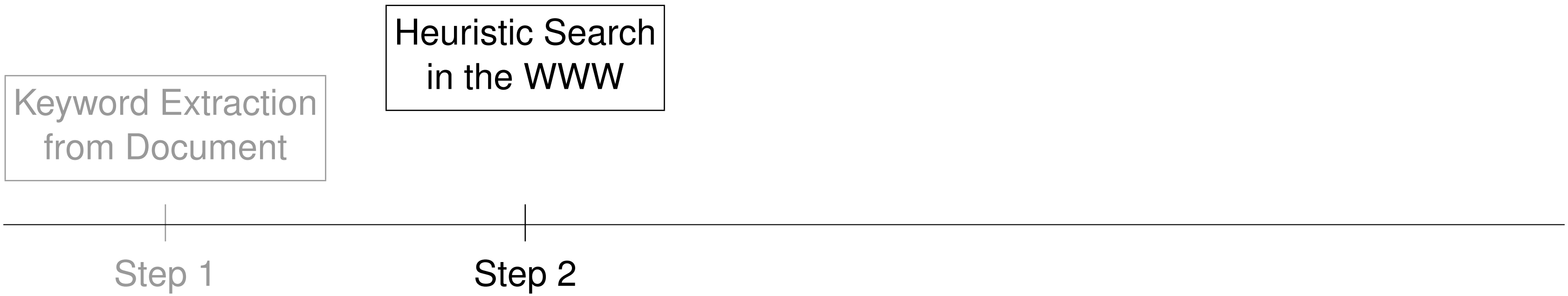
Step 1

Where are the crucial keywords?

- Check for noun phrases.
- Find orthographic mistakes.
- Consider word frequency classes.
- But, don't look in titles, captions, or headings.

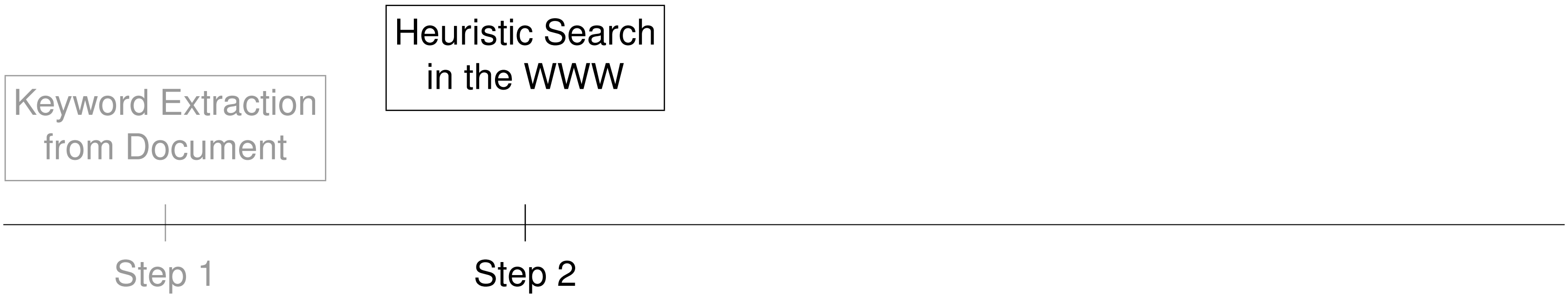
External Plagiarism Detection

Algorithms for Machines



External Plagiarism Detection

Algorithms for Machines



Query keywords: “information retrieval”, “query formulation”, “search session”, “user support”

External Plagiarism Detection

Algorithms for Machines

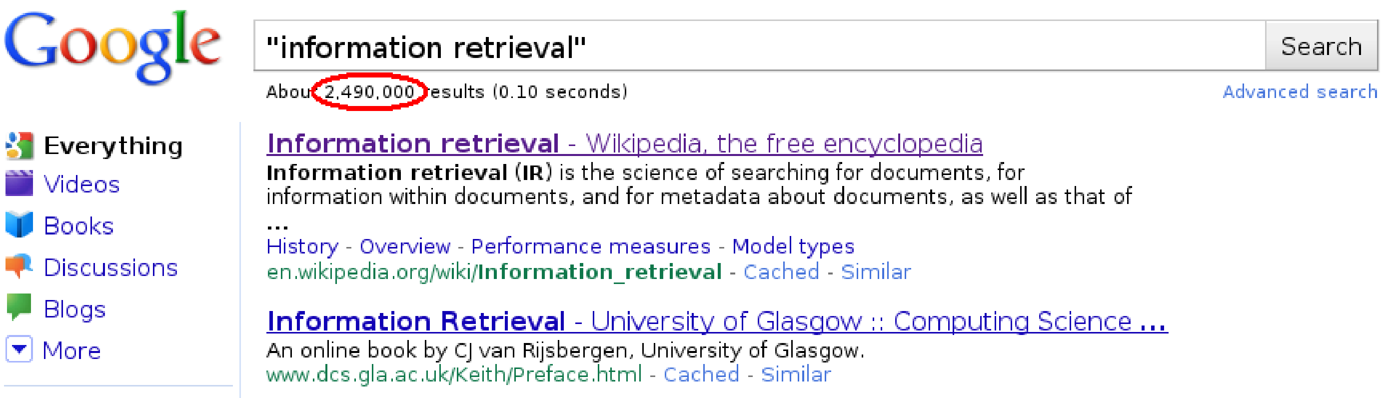
Heuristic Search in the WWW

Keyword Extraction from Document

Step 1

Step 2

Query keywords: "information retrieval", "query formulation", "search session", "user support"



External Plagiarism Detection

Algorithms for Machines

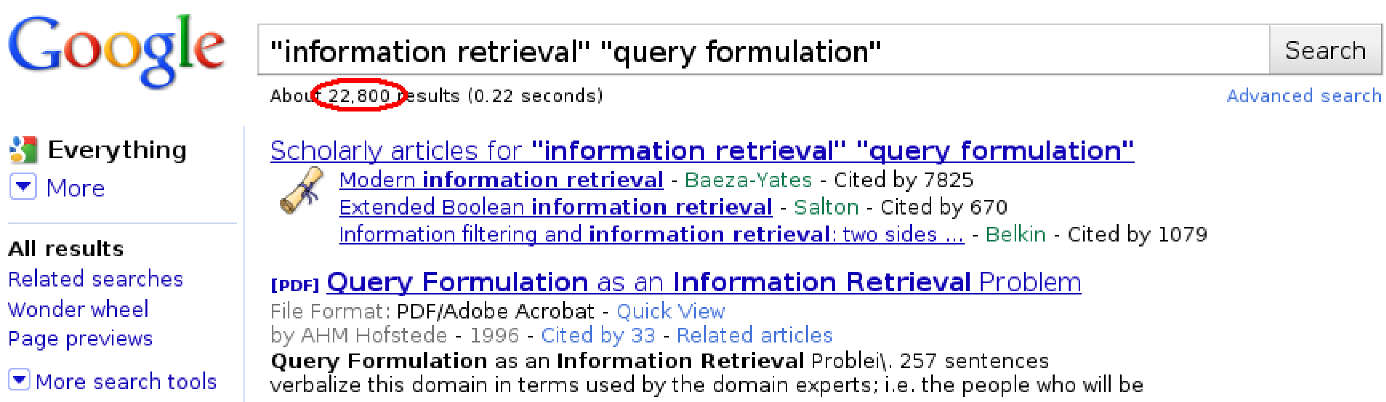
Heuristic Search in the WWW

Keyword Extraction from Document

Step 1

Step 2

Query keywords: “information retrieval”, “query formulation”, “search session”, “user support”



External Plagiarism Detection

Algorithms for Machines

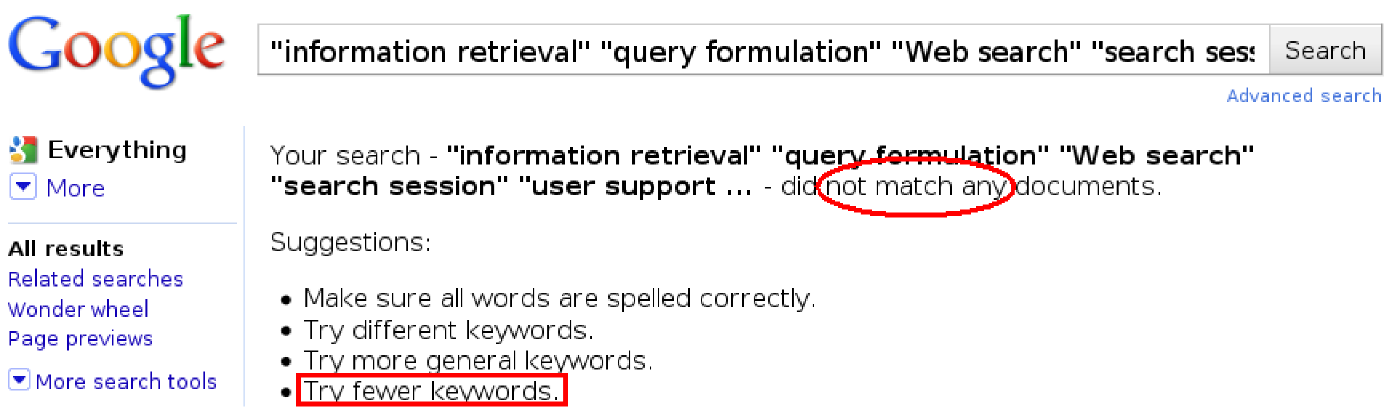
Heuristic Search in the WWW

Keyword Extraction from Document

Step 1

Step 2

Query keywords: “information retrieval”, “query formulation”, “search session”, “user support”



External Plagiarism Detection

Algorithms for Machines

Keyword Extraction from Document

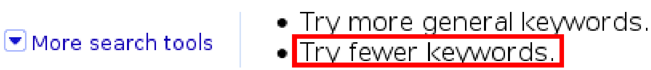
Heuristic Search in the WWW

Step 1

Step 2

Query keywords: "information retrieval", "query formulation", "search session", "user support"

The Query Cover Problem.



External Plagiarism Detection

Algorithms for Machines

Keyword Extraction from Document — Step 1

Heuristic Search in the WWW — Step 2

Document Selection — Step 3

Given:

1. A set W of keywords.
2. A query interface for a Web search engine $\mathcal{S}$.
3. An upper bound k on the result list length.

Todo:

- Find a family of queries $\mathcal{Q}$ covering W yielding at most k Web results.

External Plagiarism Detection

Algorithms for Machines

Document Selection

Heuristic Search in the WWW

Keyword Extraction from Document

Step 1 Step 2 Step 3

Given:

1. A set W of keywords.
2. A query interface for a Web search engine $\mathcal{S}$.
3. An upper bound k on the result list length. ➜ “User over Ranking”

Todo:

- Find a family of queries $\mathcal{Q}$ covering W yielding at most k Web results.

External Plagiarism Detection

Algorithms for Machines

External Plagiarism Detection

Algorithms for Machines

basic

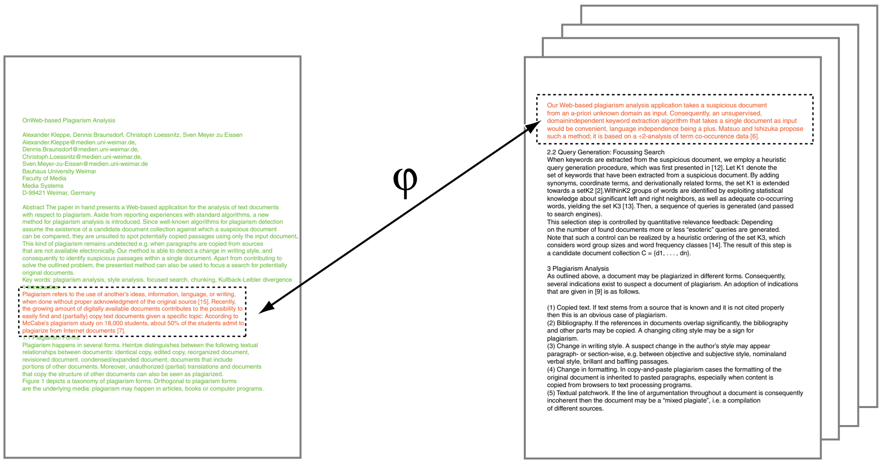
complex

Technology	What can be detected
MD5 hashing	Identity analysis for paragraphs
Hashed breakpoint chunking	Synchronized identity analysis for paragraphs
Fuzzy-fingerprinting	Tolerant similarity analysis for paragraphs
Dot plotting	Sequences of word n-grams

External Plagiarism Detection

Algorithms for Machines: Pairwise Comparison

1. Partition each document in meaningful sections, also called “chunks”.
2. Do a pairwise comparison using a similarity function φ.



suspicious document   corpus documents

Complexity:

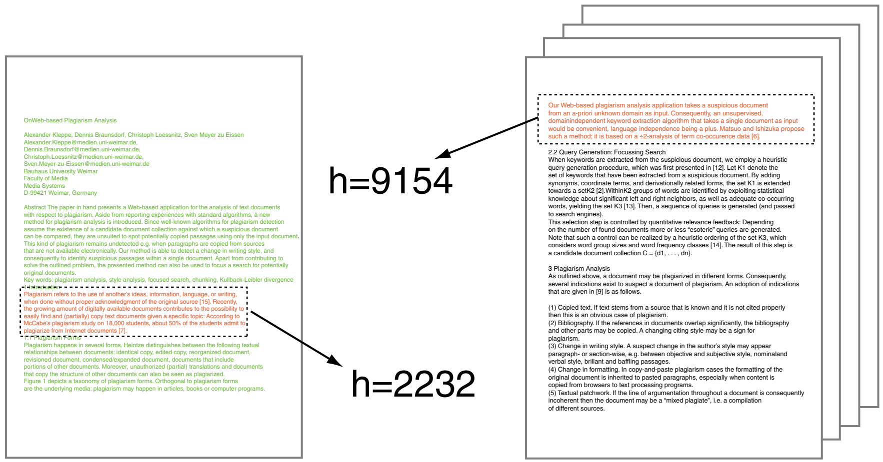
n documents in corpus, c chunks per document on average

➜ $O(n \cdot c^2)$ comparisons

External Plagiarism Detection

Algorithms for Machines: MD5 Hashing

1. Partition each document into equidistant sections.
2. Compute hash values of the chunks using a hash function h.
3. Put all hashes into a hash table. A collision indicates matching chunks.



suspicious document

corpus documents

Complexity:

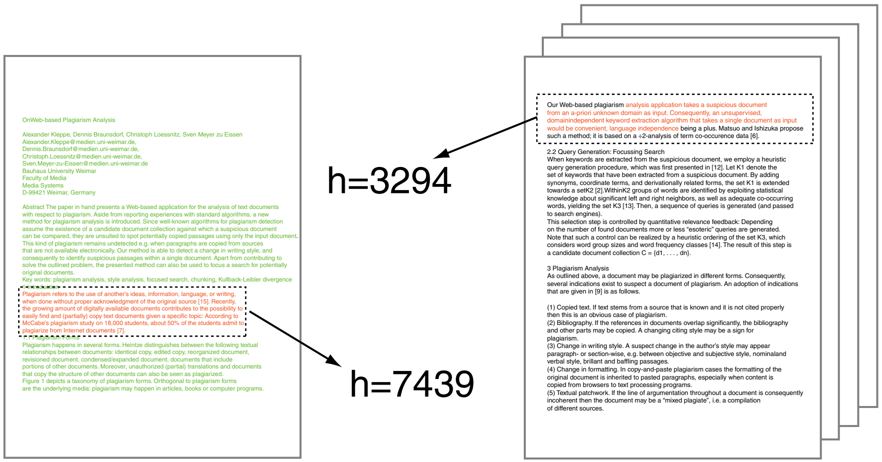
n documents in corpus, c chunks per document on average

➜ $O(n \cdot c)$ operations (fingerprint generation, hash table operations)

External Plagiarism Detection

Algorithms for Machines: Hashed Breakpoint Chunking

1. Partition each document into synchronized sections.
2. Compute hash values of the chunks using a hash function h.
3. Put all hashes into a hash table. A collision indicates matching chunks.



suspicious document   corpus documents

Complexity:

n documents in corpus, c chunks per document on average

➜ $O(n \cdot c)$ operations (fingerprint generation, hash table operations)

External Plagiarism Detection

Algorithms for Machines: Fuzzy-fingerprinting

Standard hashing:

- Equal chunks yield the same hash key:

  $h(c_1) = h(c_2) \quad \Rightarrow \quad c_1, c_2$ are equal.

- Problem: sensitive to smallest changes.

External Plagiarism Detection

Algorithms for Machines: Fuzzy-fingerprinting

Standard hashing:

- Equal chunks yield the same hash key:

  $h(c_1) = h(c_2) \quad \Rightarrow \quad c_1, c_2$ are equal.

- Problem: sensitive to smallest changes.

Fuzzy-fingerprinting:

- Different but similar chunks yield the same hash key:

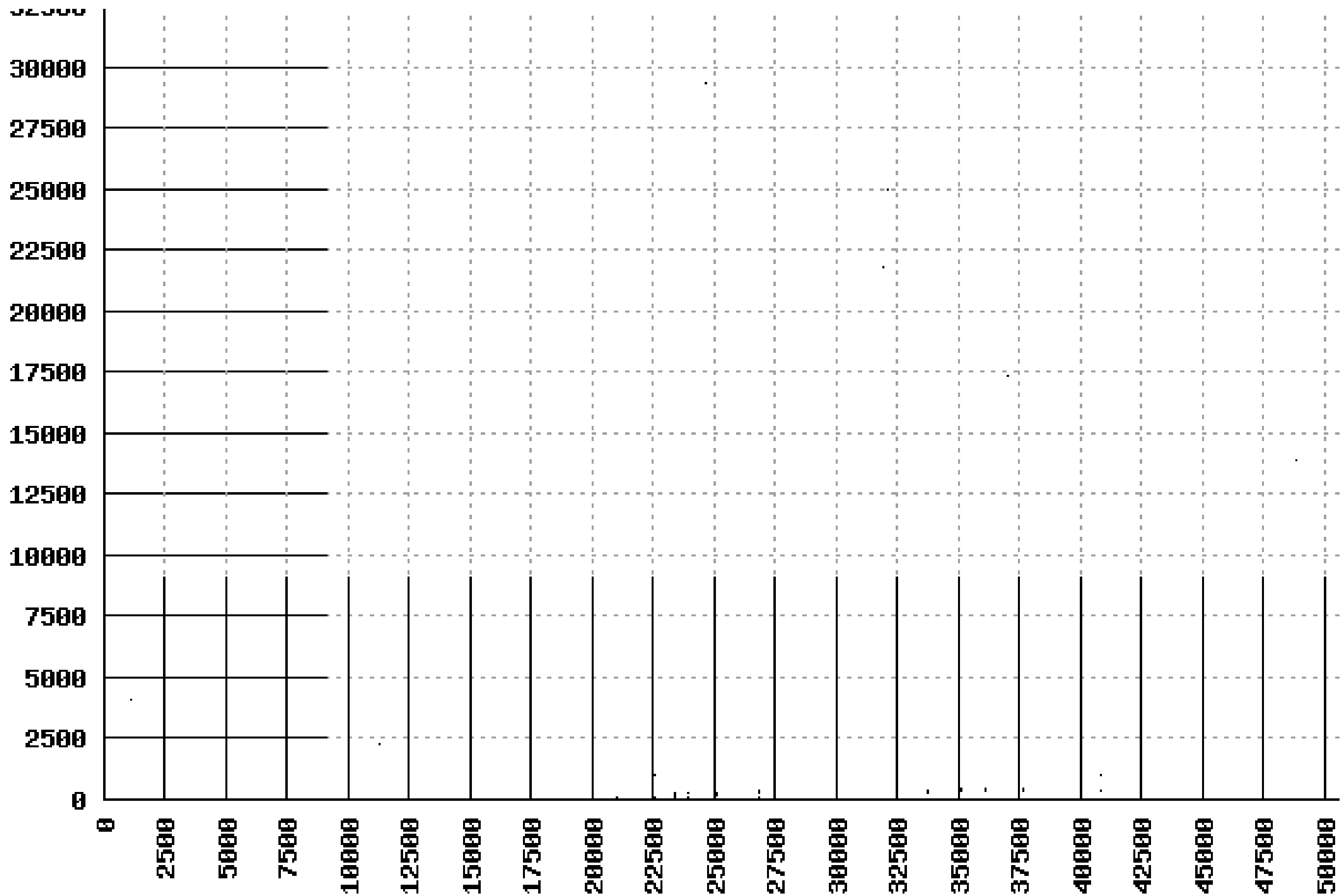
  $h_\varphi(c_1) = h_\varphi(c_2) \quad \Rightarrow \quad c_1, c_2$ are similar with high probability.

- Approach: abstraction by reducing the alphabet, neglecting word order.

- Problem: similarity-sensitive hash functions suffer from a low recall.

External Plagiarism Detection

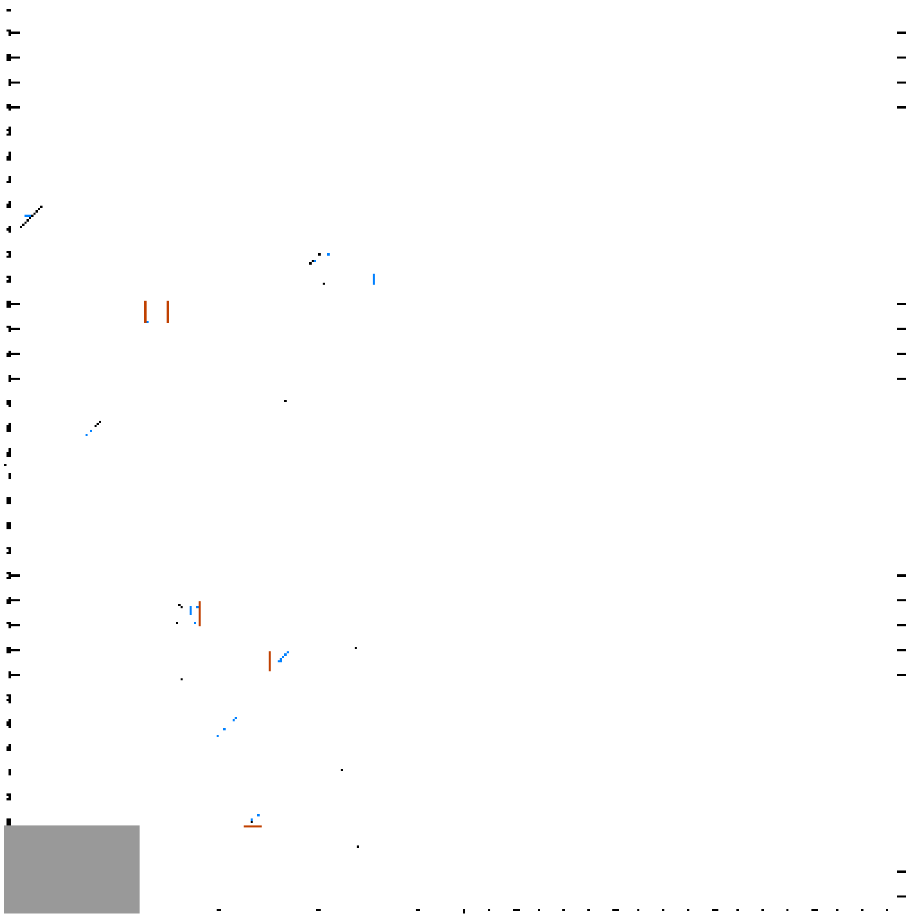
Algorithms for Machines: Dot Plotting



Geometric sequence analysis of all word 4-grams of two interesting documents.

External Plagiarism Detection

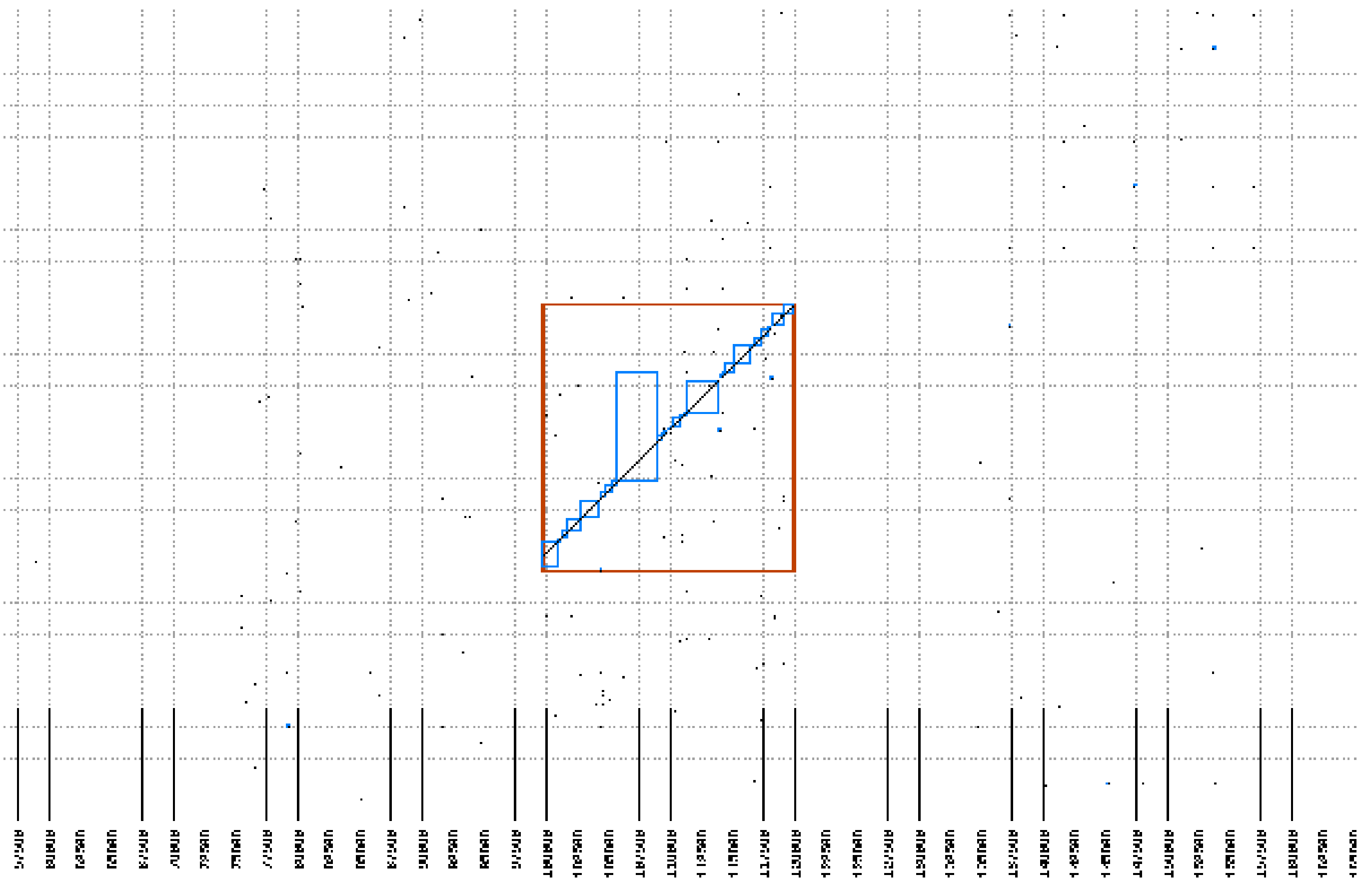
Algorithms for Machines: Dot Plotting



Geometric sequence analysis of all word 4-grams of two interesting documents.

External Plagiarism Detection

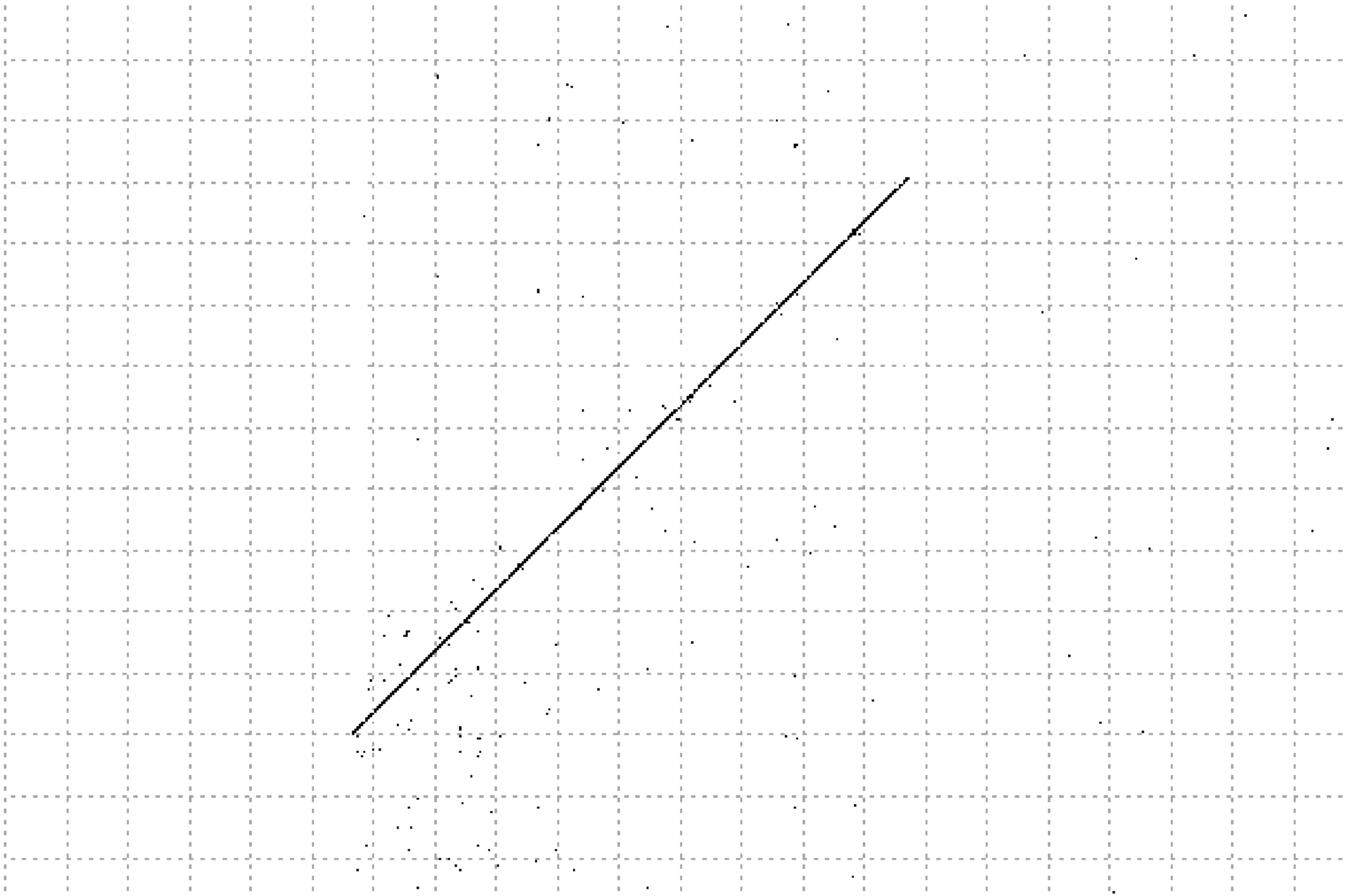
Algorithms for Machines: Dot Plotting



Geometric sequence analysis of all word 4-grams of two interesting documents.

External Plagiarism Detection

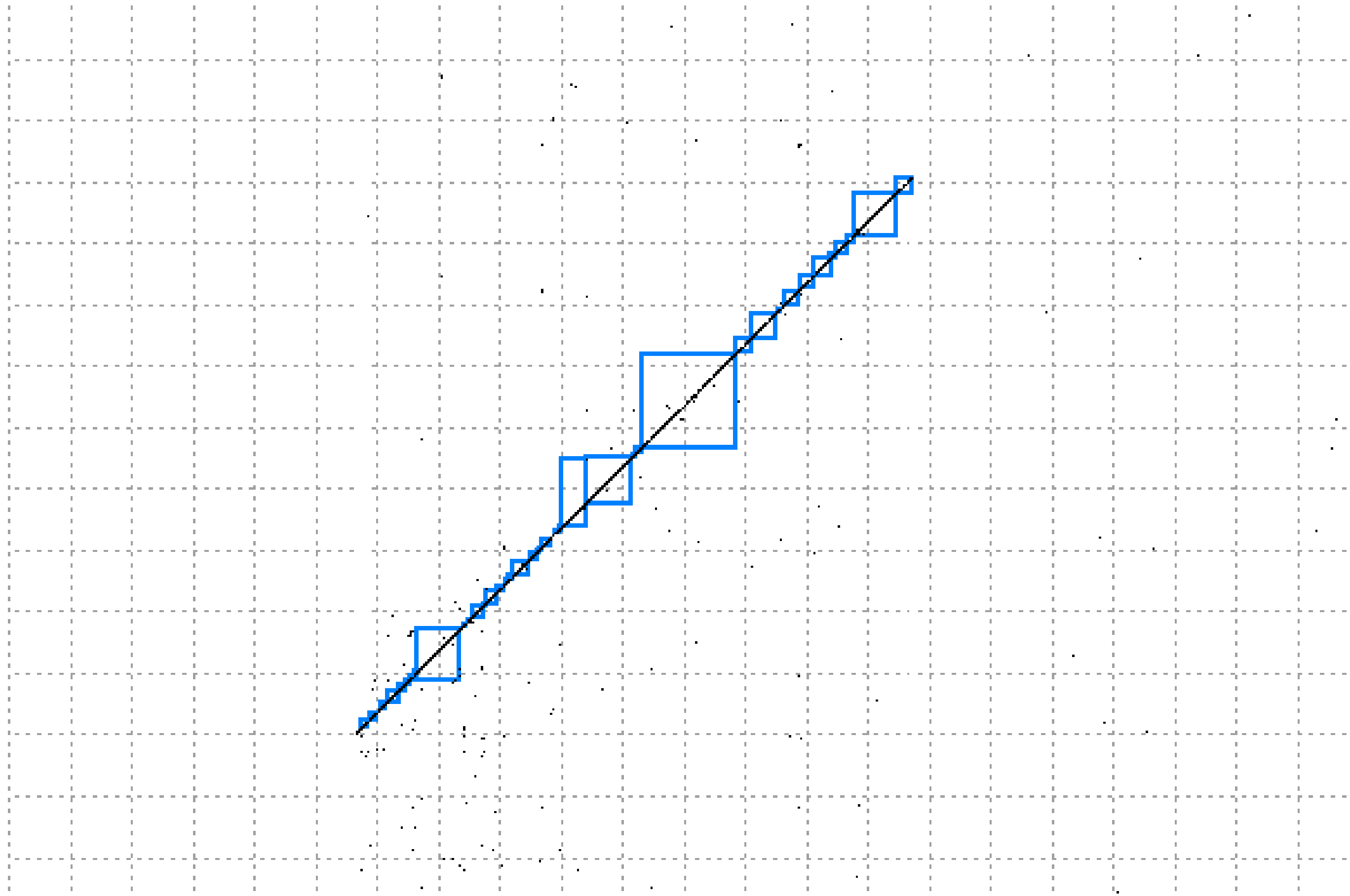
Algorithms for Machines: Dot Plotting



Level 1 (black): each dot indicates a common word 4-gram (hash collision).

External Plagiarism Detection

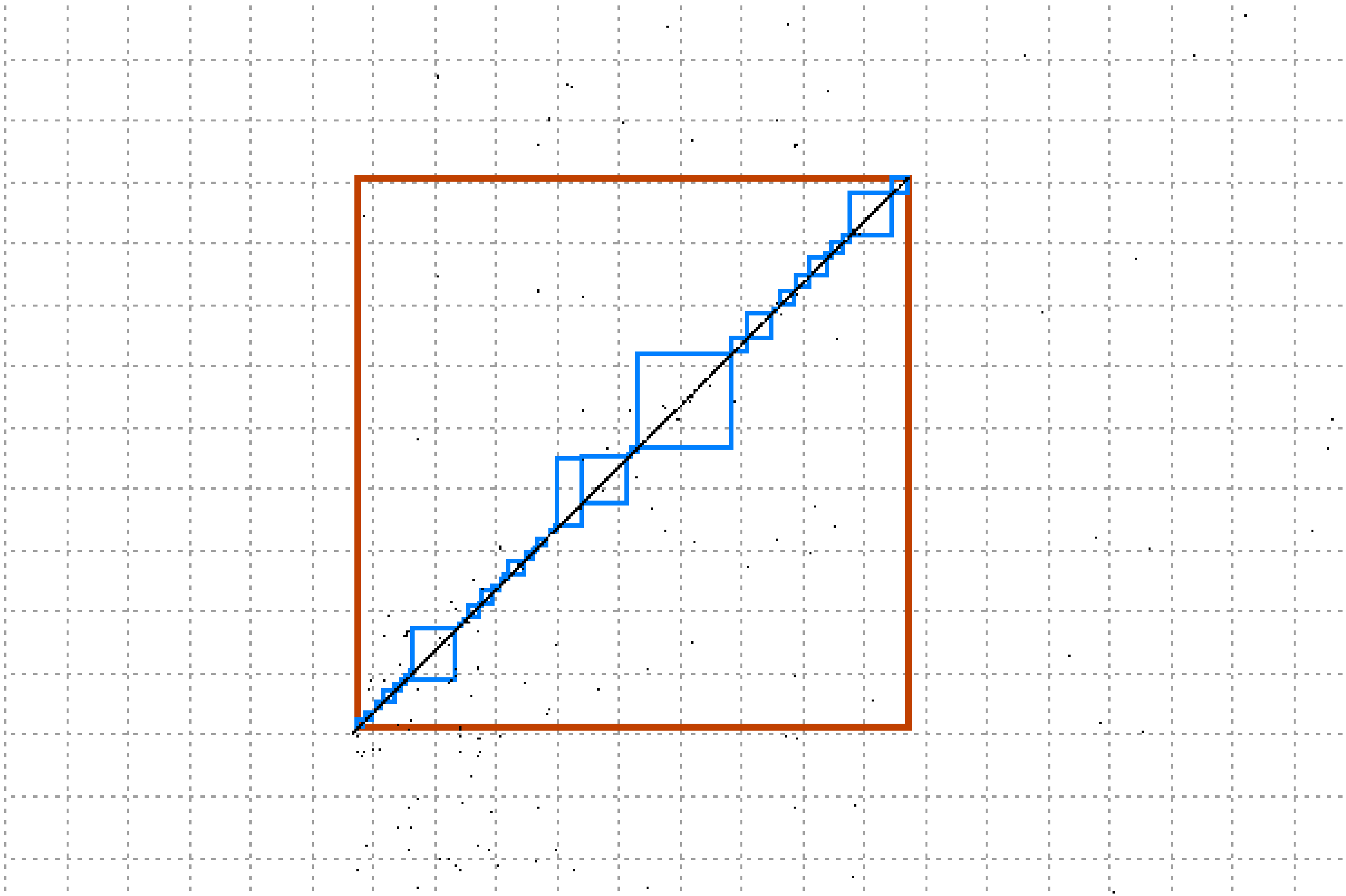
Algorithms for Machines: Dot Plotting



Level 2 (blue): neighbored common 4-grams are heuristically comprised.

External Plagiarism Detection

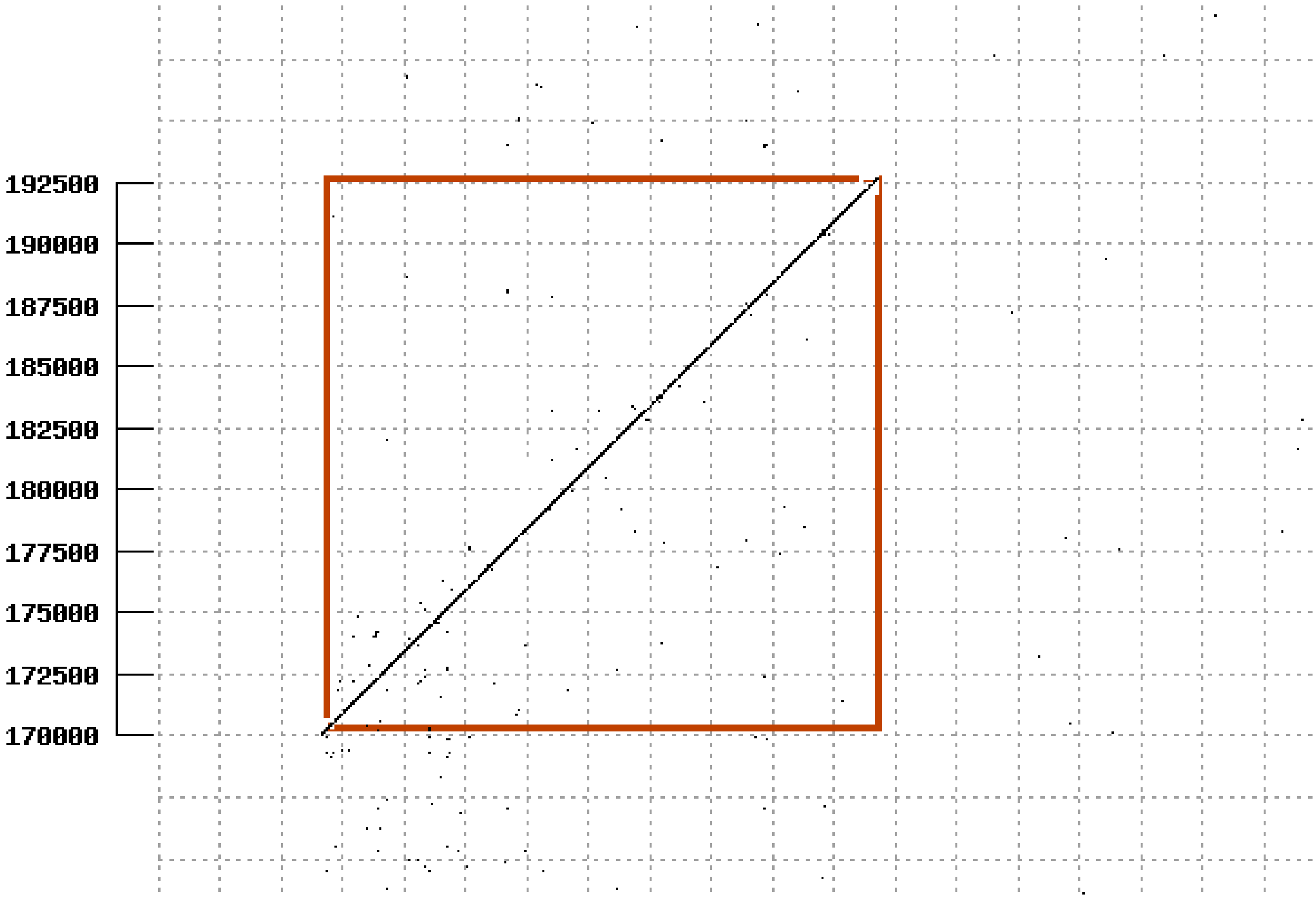
Algorithms for Machines: Dot Plotting



Level 3 (red): blue groups are merged by a cluster analysis (DBscan).

External Plagiarism Detection

Algorithms for Machines: Dot Plotting



The involved text of a cluster forms a plagiarism candidate.

External Plagiarism Detection

Algorithms for Machines

External Plagiarism Detection

Algorithms for Machines

Check for problematic decisions:

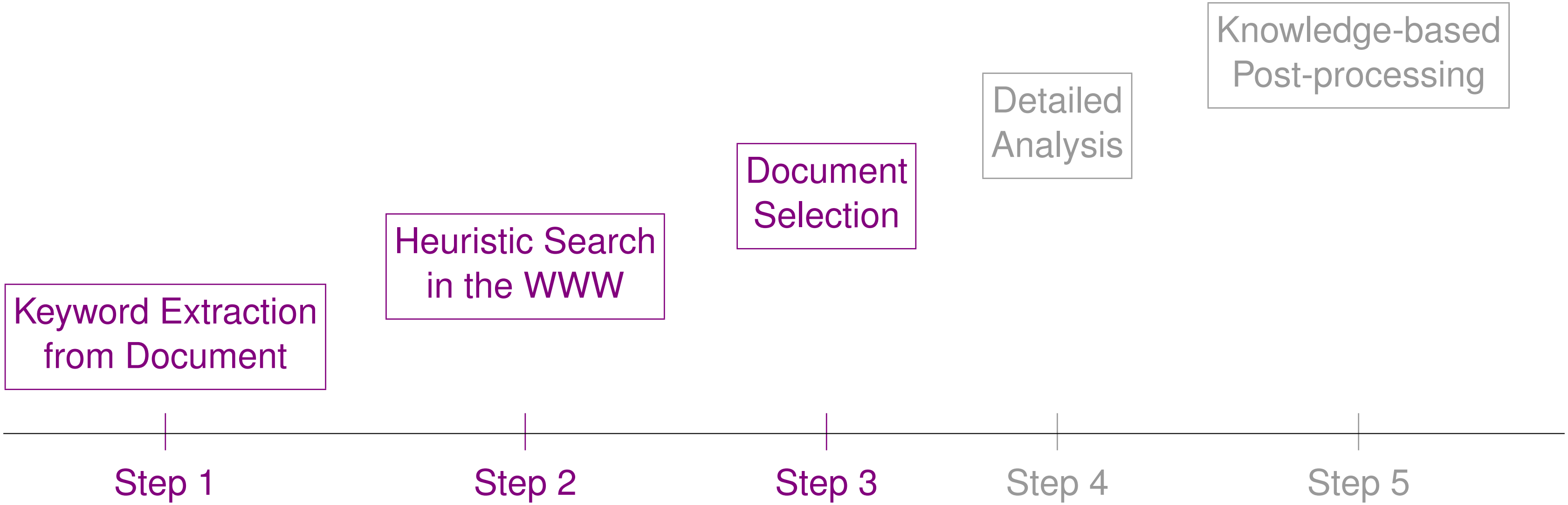
- Citation analysis

  (can be problematic: consider an “excuse citation” in a footnote along with a completely reused text)

- Comparison of authors and co-authors

External Plagiarism Detection

Algorithms for Machines



How to overcome the language barrier:

- Machine translation services
- Mapping into a concept space (ESA, CL-ESA)

The PAN Competition

The PAN Competition

2nd International Competition on Plagiarism Detection, PAN 2010

Facts:

- organized as CLEF 2010 Lab
- 18 groups from 12 countries participated
- 15 weeks of training and testing (March – June)
- training corpus was the corpus PAN-PC-09
- test corpus was the PAN-PC-10, a new version of last year's corpus.
- incidentally, the 1st competition was held at SEPLN'09.

The PAN Competition

2nd International Competition on Plagiarism Detection, PAN 2010

Facts:

- organized as CLEF 2010 Lab
- 18 groups from 12 countries participated
- 15 weeks of training and testing (March – June)
- training corpus was the corpus PAN-PC-09
- test corpus was the PAN-PC-10, a new version of last year's corpus.
- incidentally, the 1st competition was held at SEPLN'09.

Task:

Given a set of suspicious documents and a set of source documents, find all plagiarized sections in the suspicious documents and, if available, the corresponding source sections.

The PAN Competition

Plagiarism Corpus PAN-PC-10[1]

Large-scale resource for the controlled evaluation of detection algorithms:

- 27 073 documents (obtained from 22 874 books from the Project Gutenberg[2])
- 68 558 plagiarism cases (about 0-10 cases per document)

[1] www.uni-weimar.de/cms/medien/webis/research/corpora/pan-pc-10.html
[2] www.gutenberg.org

The PAN Competition

Plagiarism Corpus PAN-PC-10[1]

Large-scale resource for the controlled evaluation of detection algorithms:

- 27 073 documents (obtained from 22 874 books from the Project Gutenberg[2])
- 68 558 plagiarism cases (about 0-10 cases per document)

[1] www.uni-weimar.de/cms/medien/webis/research/corpora/pan-pc-10.html
[2] www.gutenberg.org

PAN-PC-10 addresses a broad range of plagiarism situations by varying reasonably within the following parameters:

1. document length
2. document language
3. detection task
4. plagiarism case length
5. plagiarism case obfuscation
6. plagiarism case topic alignment

The PAN Competition

PAN-PC-10 Document Statistics

100% 27 073 documents

The PAN Competition

PAN-PC-10 Document Statistics

100% 27 073 documents

Document length:

50% short (1-10 pages)	35% medium (10-100 pages)	15% long (100-1 000 pp.)

The PAN Competition

PAN-PC-10 Document Statistics

100% 27 073 documents

Document length:

50% short (1-10 pages)	35% medium (10-100 pages)	15% long (100-1 000 pp.)

Document language:

80% English	10% de	10% es

The PAN Competition

PAN-PC-10 Document Statistics

100% 27 073 documents

Document length:

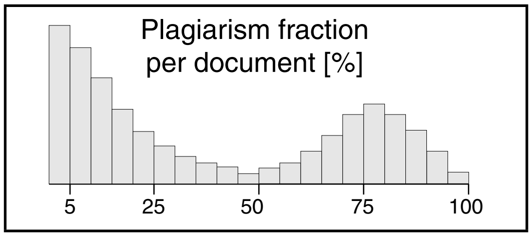
50% short (1-10 pages)	35% medium (10-100 pages)	15% long (100-1 000 pp.)

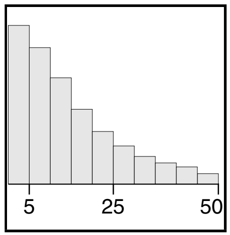
Document language:

80% English	10% de	10% es

Detection task:

70% external analysis		30% intrinsic analysis	
plagiarized	unmodified (plagiarism source)	plagiarized	unmodified





The PAN Competition

PAN-PC-10 Plagiarism Case Statistics

100% 68 558 plagiarism cases

The PAN Competition

PAN-PC-10 Plagiarism Case Statistics

100% 68 558 plagiarism cases

Plagiarism case length:

34% short (50-150 words)	33% medium (300-500 words)	33% long (3 000-5 000 words)

The PAN Competition

PAN-PC-10 Plagiarism Case Statistics

100% 68 558 plagiarism cases

Plagiarism case length:

34% short (50-150 words)	33% medium (300-500 words)	33% long (3 000-5 000 words)

Plagiarism case obfuscation:

40% none	40% artificial[3]		6%[4]	14%[5]	
	low obfuscation	high obfuscation	AMT	de	es

[3] Artificial plagiarism: algorithmic obfuscation.
[4] Simulated plagiarism: obfuscation via Amazon Mechanical Turk.
[5] Cross-language plagiarism: obfuscation due to machine translation de→en and es→en.

The PAN Competition

PAN-PC-10 Plagiarism Case Statistics

100% 68 558 plagiarism cases

Plagiarism case length:

34% short (50-150 words)	33% medium (300-500 words)	33% long (3 000-5 000 words)

Plagiarism case obfuscation:

40% none	40% artificial[3]		6%[4]	14%[5]	
	low obfuscation	high obfuscation	AMT	de	es

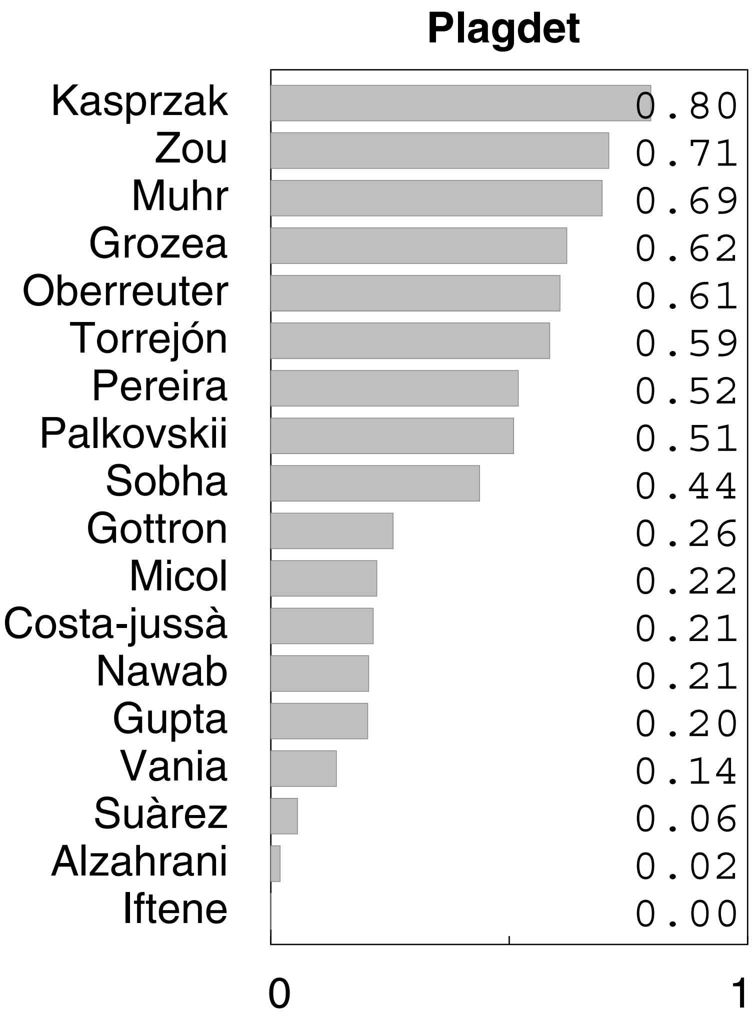
[3] Artificial plagiarism: algorithmic obfuscation.
[4] Simulated plagiarism: obfuscation via Amazon Mechanical Turk.
[5] Cross-language plagiarism: obfuscation due to machine translation de→en and es→en.

Plagiarism case topic alignment:

50% intra-topic	50% inter-topic

The PAN Competition

Plagiarism Detection Results

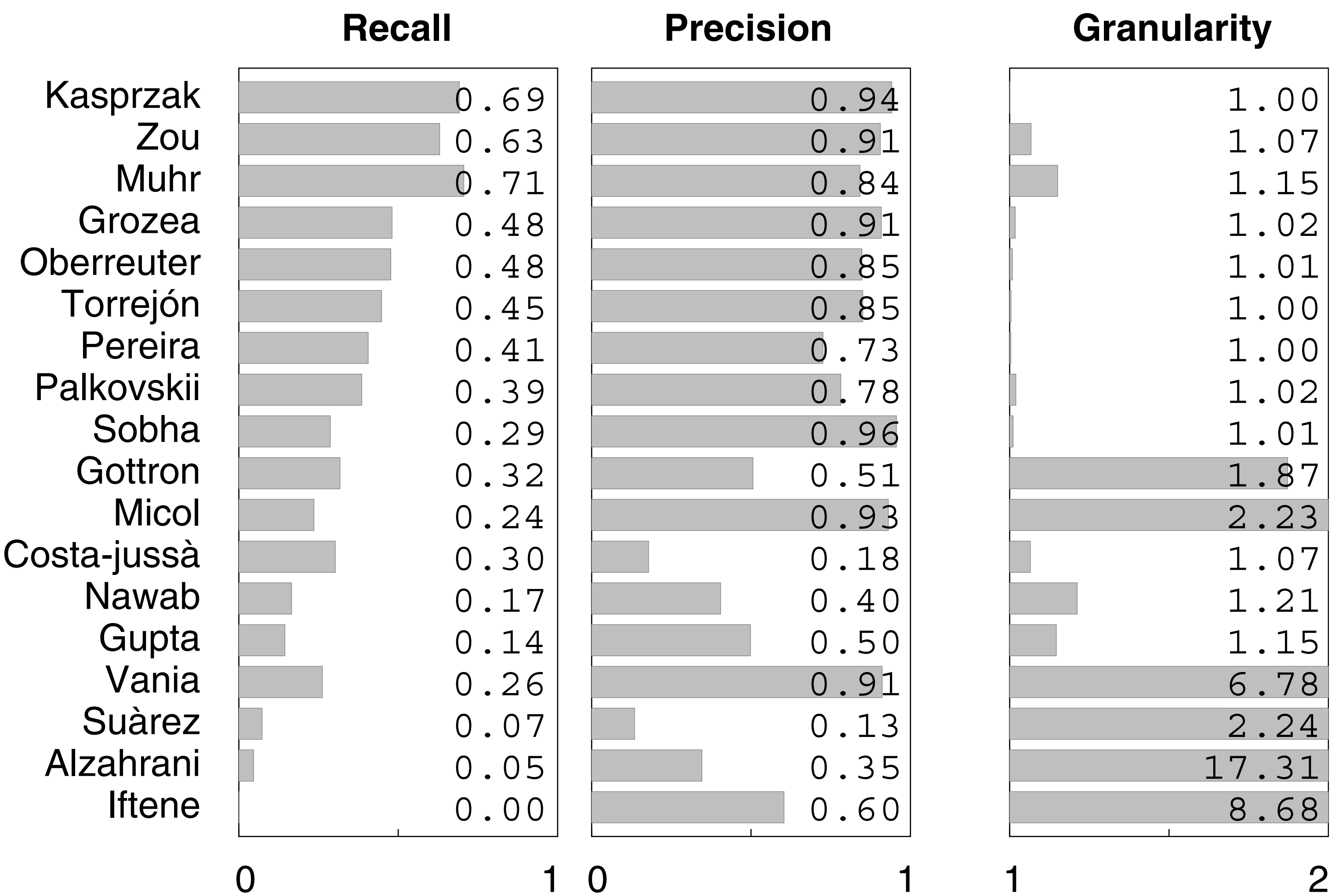


- Plagdet combines precision, recall, and granularity.
- Precision and recall are well-known, yet not well-defined.
- Granularity measures the number of times a single plagiarism case has been detected.

[Potthast et al., 2010]

The PAN Competition

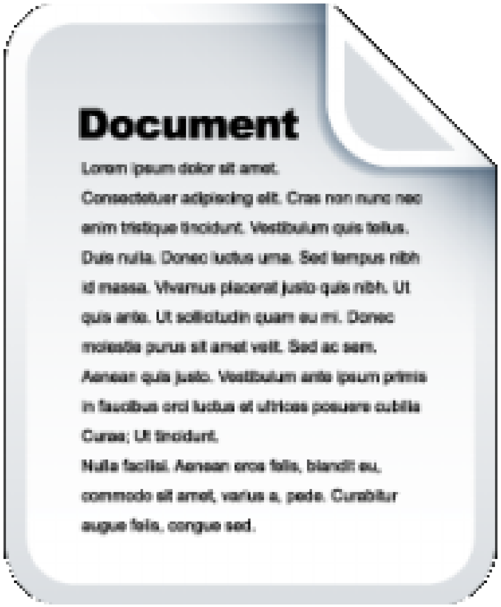
Plagiarism Detection Results

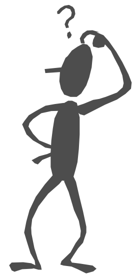


Intrinsic Plagiarism Detection

Intrinsic Plagiarism Detection

How Humans Spot Plagiarism

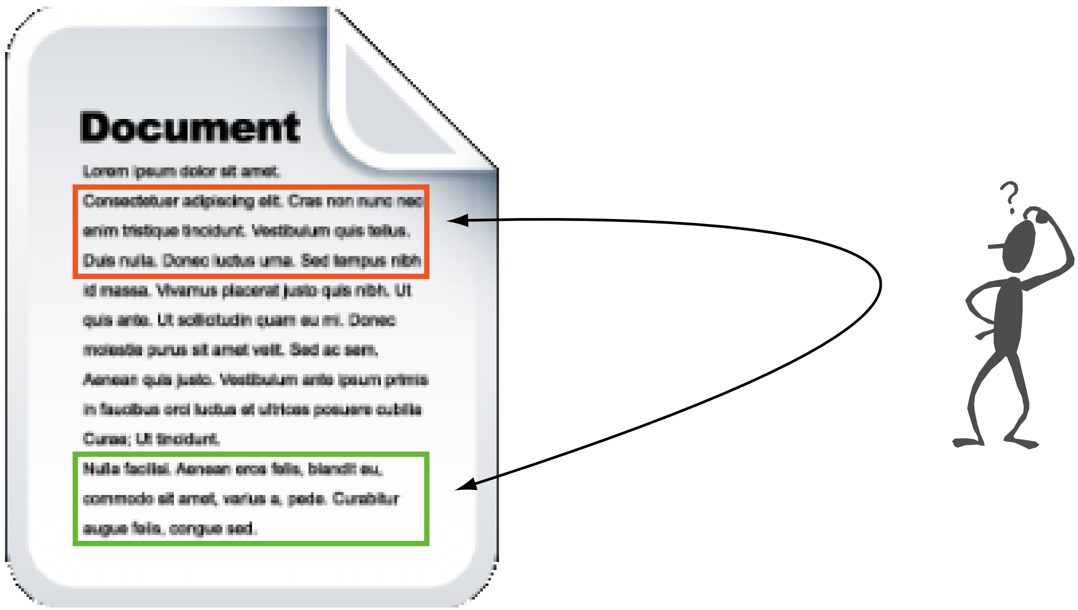




When no corpus—along with a powerful search engine—is at hand . . .

Intrinsic Plagiarism Detection

How Humans Spot Plagiarism



When no corpus—along with a powerful search engine—is at hand . . .

- look for style changes
- check for peculiarities (orthographic mistakes, typographical habits)
- listen to the instincts (perhaps the most powerful “technology”)

Intrinsic Plagiarism Detection

Algorithms for Machines

Intrinsic Plagiarism Detection

Algorithms for Machines

Intrinsic Plagiarism Detection

Algorithms for Machines

Impurity Assessment

Step 1

How large is the fraction θ of plagiarized text?

- document length analysis
- genre analysis (e.g. scientific article versus editorial)
- analysis of issuing institution

Intrinsic Plagiarism Detection

Algorithms for Machines

Intrinsic Plagiarism Detection

Algorithms for Machines

Chunking Strategy

Impurity Assessment

Step 1

Step 2

How to find text positions where plagiarism starts or ends?

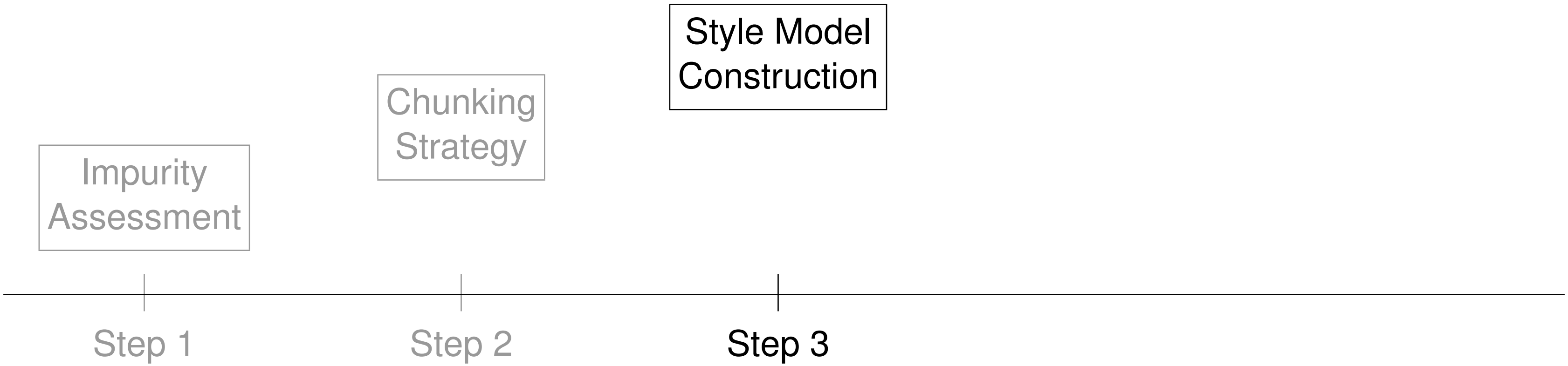
basic

complex

- uniform length chunking (simple but naive)
- structural boundaries (chapters, paragraphs, tables, captions)
- topical boundaries (difficult but powerful)
- stylistic boundaries (best, but usually intractable)

Intrinsic Plagiarism Detection

Algorithms for Machines



Impurity Assessment

Chunking Strategy

Style Model Construction

Step 1

Step 2

Step 3

The question of Stylometry: How to quantify writing style?

basic

- structural features (paragraph lengths, use of tables, signatures)
- character-based lexical features (n-gram frequency, compression rate)
- word-based lexical features (readability, writing complexity)
- syntactic features (part of speech, function words)
- dialectic power and argumentation

complex

Intrinsic Plagiarism Detection

Algorithms for Machines: Style Features that Work

Stylometric feature	F Measure
Flesch Reading Ease Score	0.208
Average number of syllables per word	0.205
Frequency of term: of	0.192
Noun-Verb-Nountri-gram	0.189
Noun-Noun-Verbtri-gram	0.182
Verb-Noun-Nountri-gram	0.179
Gunning Fog index	0.179
Yule’s K measure	0.176
Flesch Kincaid grade level	0.175
Average word length	0.173
Noun-Preposition-ProperNountri-gram	0.173
Honore’s R measure	0.165
Average word length	0.165
Average word frequency class	0.162
Consonant-Vowel-Consonanttri-gram	0.154
Frequency of term: is	0.151
Noun-Noun-CoordinatingConjunctiontri-gram	0.150
NounPlural-Preposition-Determinertri-gram	0.149
Determiner-NounPlural-Prepositiontri-gram	0.148
Consonant-Vowel-Voweltri-gram	0.146

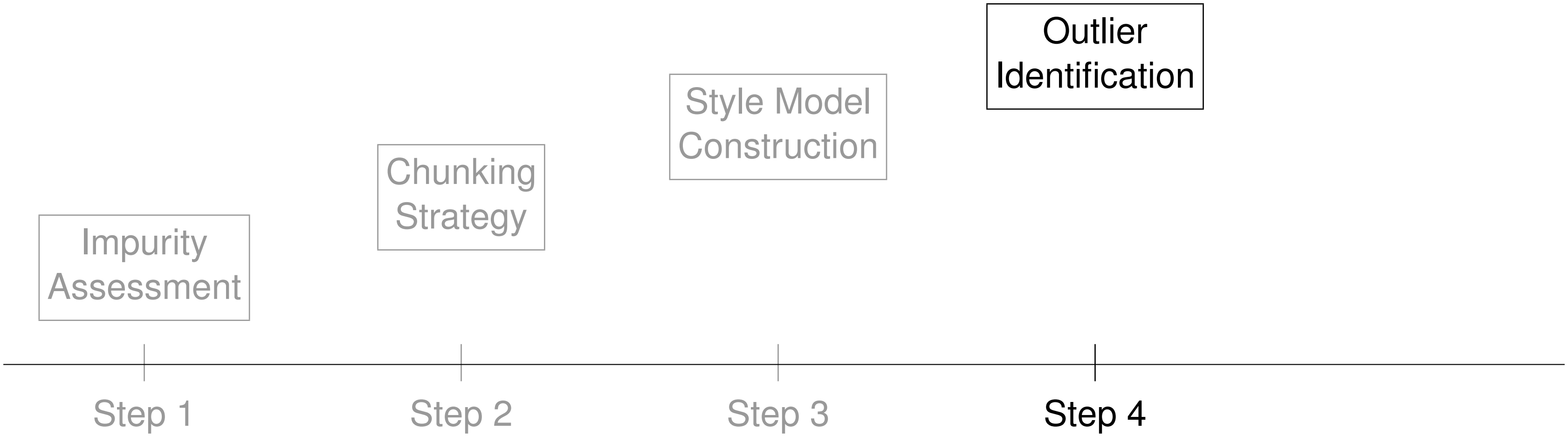
Intrinsic Plagiarism Detection

Algorithms for Machines: Style Features that Work

Stylometric feature	F Measure
Flesch Reading Ease Score	0.208
Average number of syllables per word	0.205
Frequency of term: of	0.192
Noun-Verb-Nountri-gram	0.189
Noun-Noun-Verbtri-gram	0.182
Verb-Noun-Nountri-gram	0.179
Gunning Fog index	0.179
Yule's K measure	0.176
Flesch Kincaid grade level	0.175
Average word length	0.173
Noun-Preposition-ProperNountri-gram	0.173
Honore's R measure	0.165
Average word length	0.165
Average word frequency class	0.162
Consonant-Vowel-Consonanttri-gram	0.154
Frequency of term: is	0.151
Noun-Noun-CoordinatingConjunctiontri-gram	0.150
NounPlural-Preposition-Determinertri-gram	0.149
Determiner-NounPlural-Prepositiontri-gram	0.148
Consonant-Vowel-Voweltri-gram	0.146

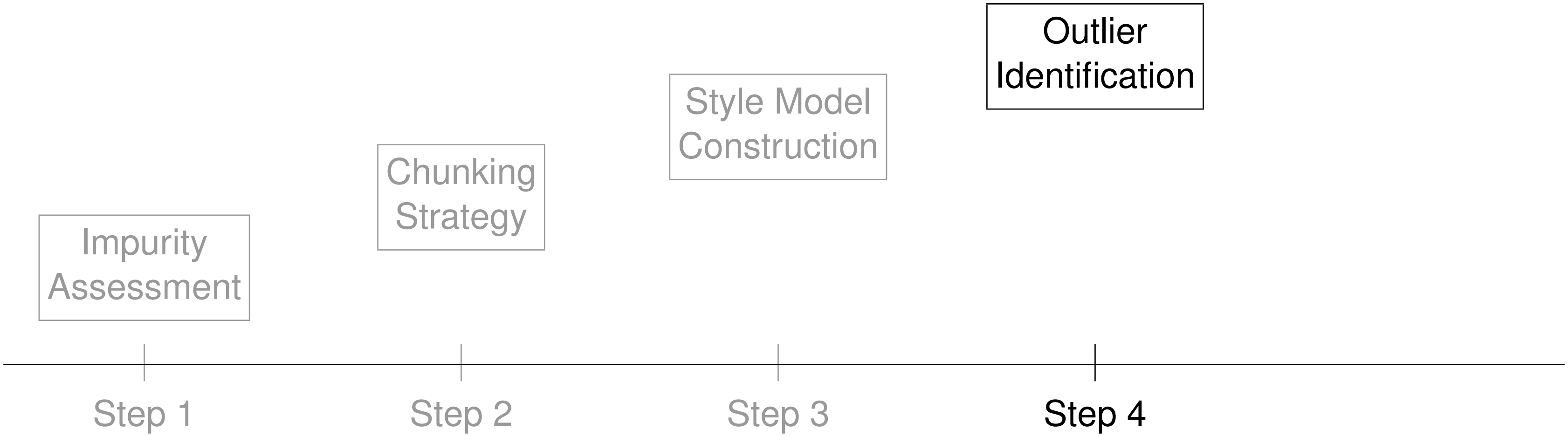
Intrinsic Plagiarism Detection

Algorithms for Machines



Intrinsic Plagiarism Detection

Algorithms for Machines



One-class classification: we have an idea about positive examples only. : (

- Density methods try to model a style feature's distribution.
- Boundary methods try to cluster text portions of similar style.
- Reconstruction methods quantify the style generation error under the average style model.

Intrinsic Plagiarism Detection

Algorithms for Machines: Outlier Identification

Alexander Kleppe, Dennis Braunsdorf, Christoph Loessnitz, Sven Meyer zu Eissen
Alexander.Kleppe@medien.uni-weimar.de,
Dennis.Braunsdorf@medien.uni-weimar.de,
Christoph.Loessnitz@medien.uni-weimar.de,
Sven.Meyer-zu-Eissen@medien.uni-weimar.de
Bauhaus University Weimar
Faculty of Media
Media Systems
D-99421 Weimar, Germany

Abstract The paper in hand presents a Web-based application for the analysis of text documents with respect to plagiarism. Aside from reporting experiences with standard algorithms, a new method for plagiarism analysis is introduced. Since well-known algorithms for plagiarism detection assume the existence of a candidate document collection against which a suspicious document can be compared, they are unsuited to spot potentially copied passages using only the input document. This kind of plagiarism remains undetected e.g. when paragraphs are copied from sources that are not available electronically. Our method is able to detect a change in writing style, and consequently to identify suspicious passages within a single document. Apart from contributing to solve the outlined problem, the presented method can also be used to focus a search for potentially original documents.

Key words: plagiarism analysis, style analysis, focused search, chunking, Kullback-Leibler divergence

1 Introduction

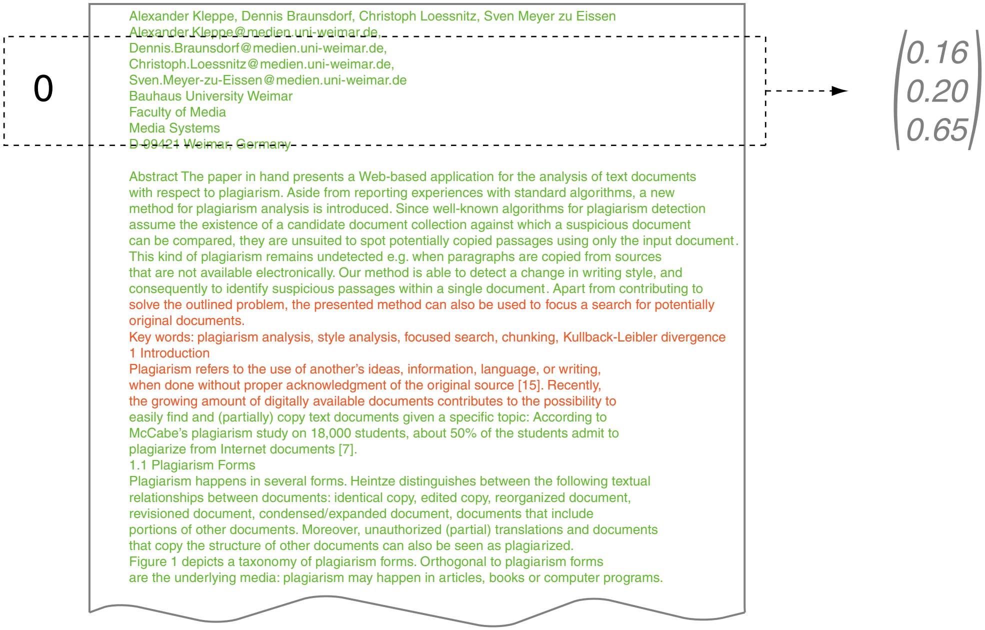
Plagiarism refers to the use of another's ideas, information, language, or writing, when done without proper acknowledgment of the original source [15]. Recently, the growing amount of digitally available documents contributes to the possibility to easily find and (partially) copy text documents given a specific topic: According to McCabe's plagiarism study on 18,000 students, about 50% of the students admit to plagiarize from Internet documents [7].

1.1 Plagiarism Forms

Plagiarism happens in several forms. Heintze distinguishes between the following textual relationships between documents: identical copy, edited copy, reorganized document, revisioned document, condensed/expanded document, documents that include portions of other documents. Moreover, unauthorized (partial) translations and documents that copy the structure of other documents can also be seen as plagiarized. Figure 1 depicts a taxonomy of plagiarism forms. Orthogonal to plagiarism forms are the underlying media: plagiarism may happen in articles, books or computer programs.

Intrinsic Plagiarism Detection

Algorithms for Machines: Outlier Identification



Alexander Kleppe, Dennis Braunsdorf, Christoph Loessnitz, Sven Meyer zu Eissen
Alexander.Kleppe@medien.uni-weimar.de,
Dennis.Braunsdorf@medien.uni-weimar.de,
Christoph.Loessnitz@medien.uni-weimar.de,
Sven.Meyer-zu-Eissen@medien.uni-weimar.de
Bauhaus University Weimar
Faculty of Media
Media Systems
D-99421 Weimar, Germany

Abstract The paper in hand presents a Web-based application for the analysis of text documents with respect to plagiarism. Aside from reporting experiences with standard algorithms, a new method for plagiarism analysis is introduced. Since well-known algorithms for plagiarism detection assume the existence of a candidate document collection against which a suspicious document can be compared, they are unsuited to spot potentially copied passages using only the input document. This kind of plagiarism remains undetected e.g. when paragraphs are copied from sources that are not available electronically. Our method is able to detect a change in writing style, and consequently to identify suspicious passages within a single document. Apart from contributing to solve the outlined problem, the presented method can also be used to focus a search for potentially original documents.

Key words: plagiarism analysis, style analysis, focused search, chunking, Kullback-Leibler divergence

1 Introduction

Plagiarism refers to the use of another's ideas, information, language, or writing, when done without proper acknowledgment of the original source [15]. Recently, the growing amount of digitally available documents contributes to the possibility to easily find and (partially) copy text documents given a specific topic: According to McCabe's plagiarism study on 18,000 students, about 50% of the students admit to plagiarize from Internet documents [7].

1.1 Plagiarism Forms

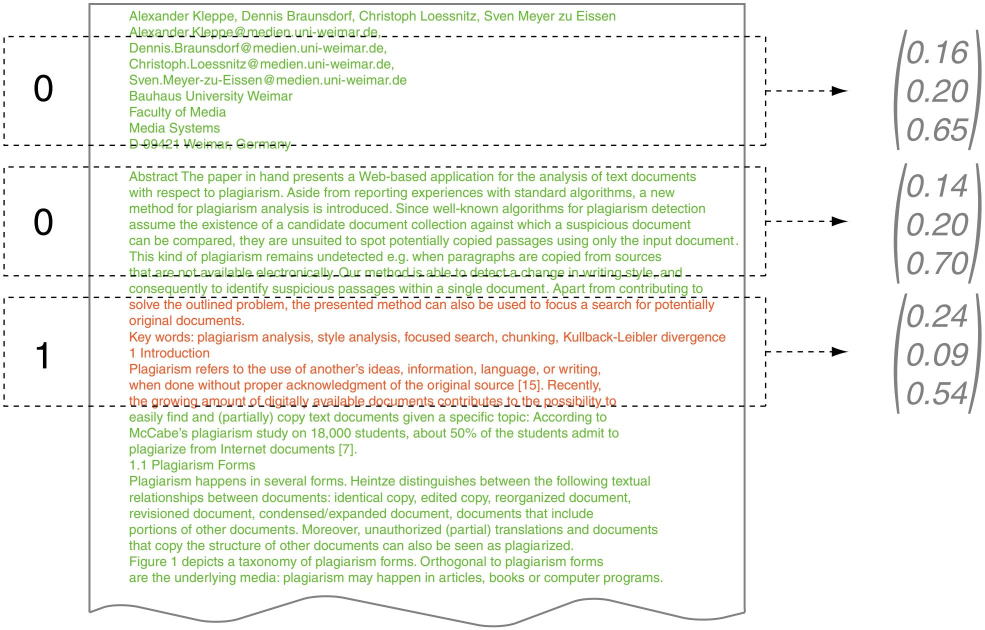
Plagiarism happens in several forms. Heintze distinguishes between the following textual relationships between documents: identical copy, edited copy, reorganized document, revisioned document, condensed/expanded document, documents that include portions of other documents. Moreover, unauthorized (partial) translations and documents that copy the structure of other documents can also be seen as plagiarized. Figure 1 depicts a taxonomy of plagiarism forms. Orthogonal to plagiarism forms are the underlying media: plagiarism may happen in articles, books or computer programs.

Intrinsic Plagiarism Detection

Algorithms for Machines: Outlier Identification

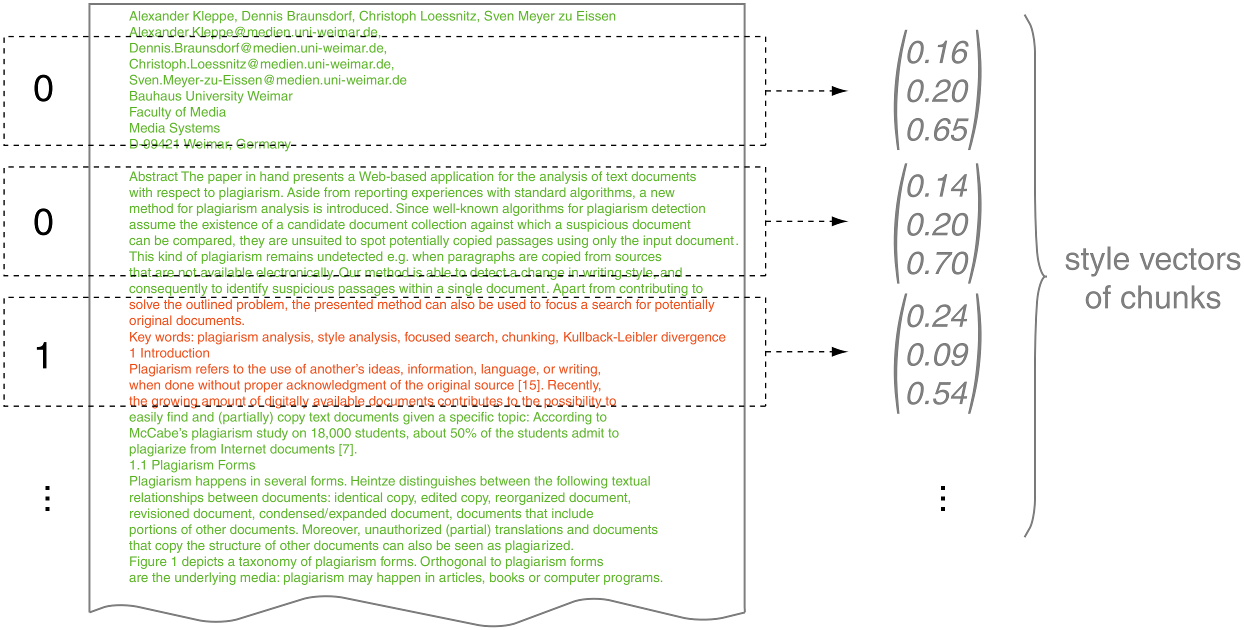
Intrinsic Plagiarism Detection

Algorithms for Machines: Outlier Identification



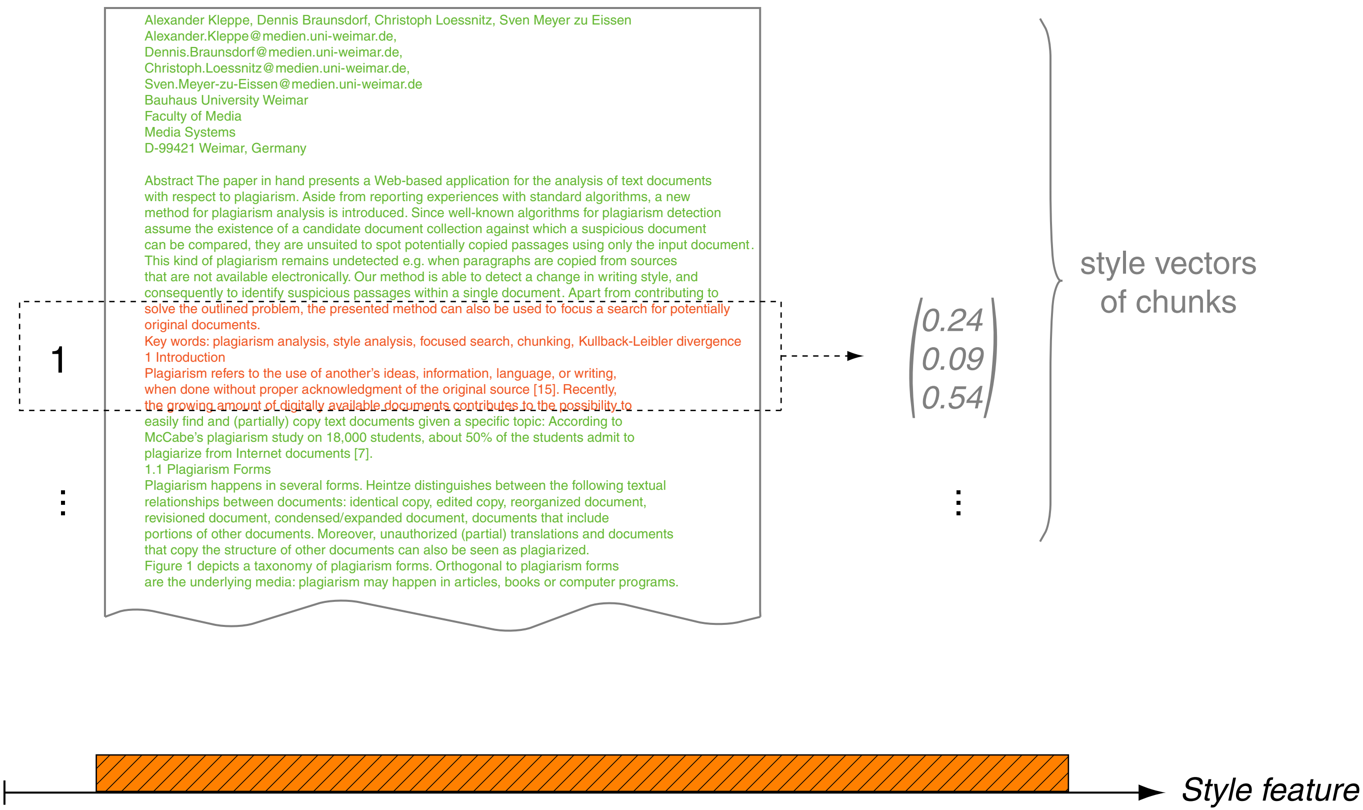
Intrinsic Plagiarism Detection

Algorithms for Machines: Outlier Identification



Intrinsic Plagiarism Detection

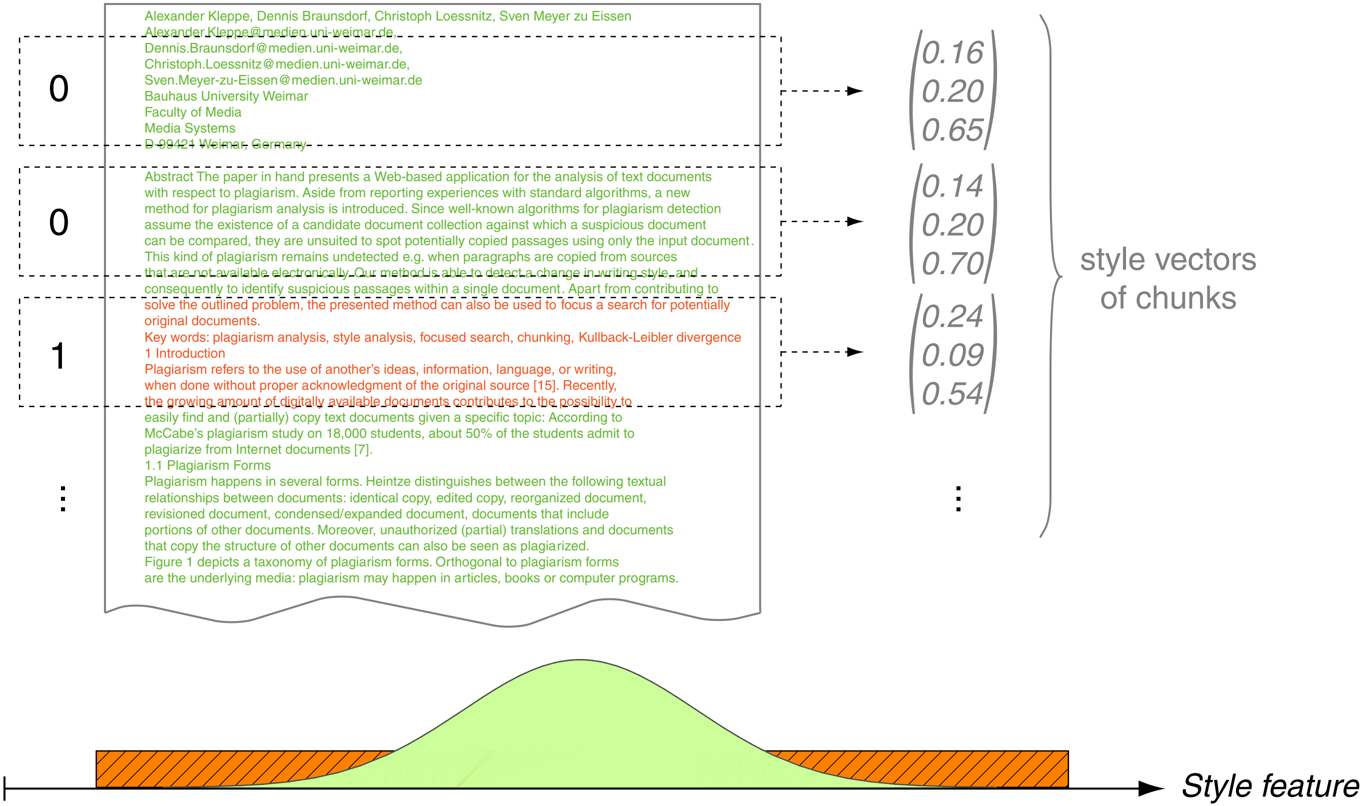
Algorithms for Machines: Outlier Identification



Assume that style features of outliers (plagiarized text) are uniformly distributed.

Intrinsic Plagiarism Detection

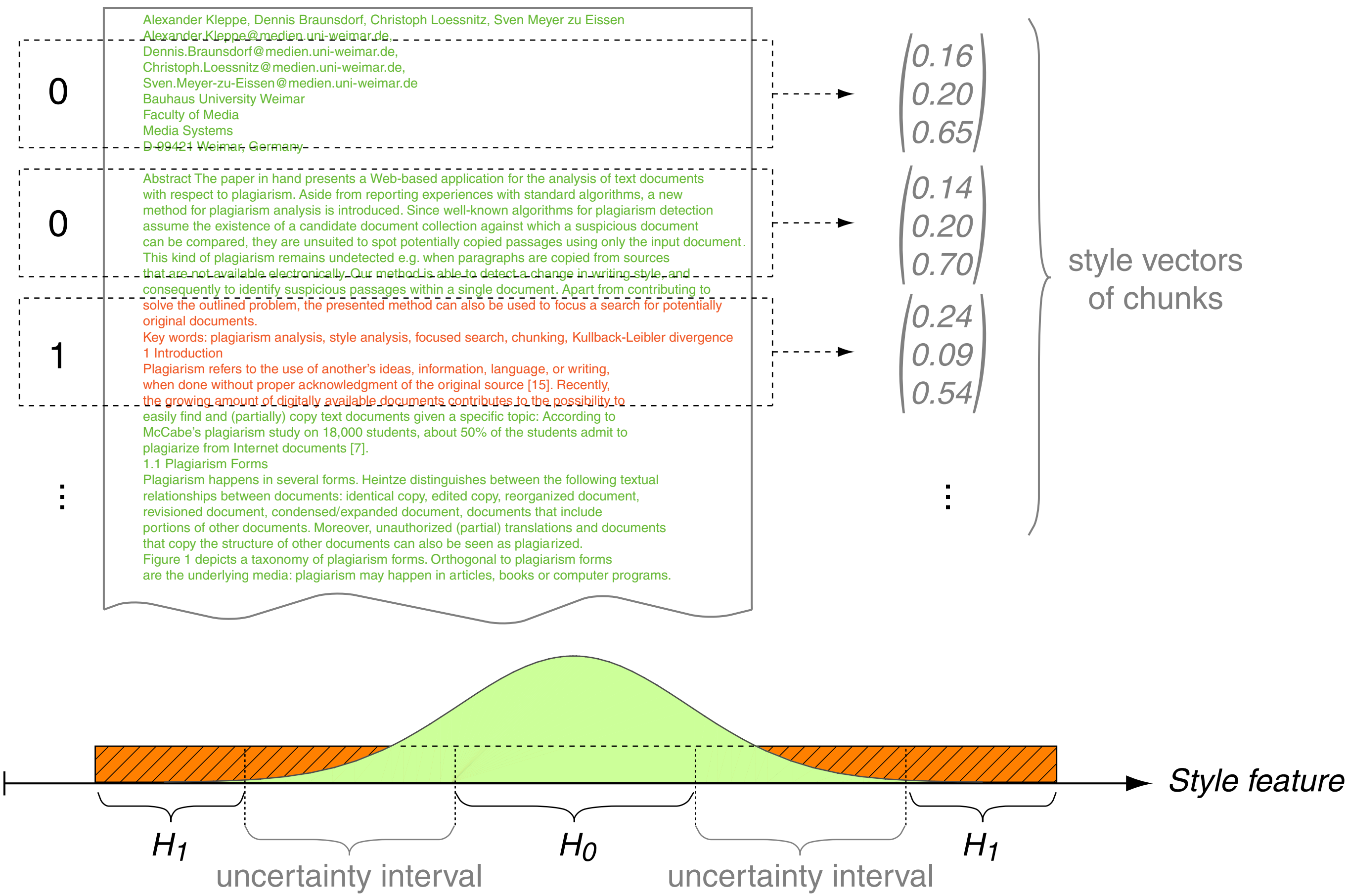
Algorithms for Machines: Outlier Identification



Assume that style features of original text are Gaussian distributed.

Intrinsic Plagiarism Detection

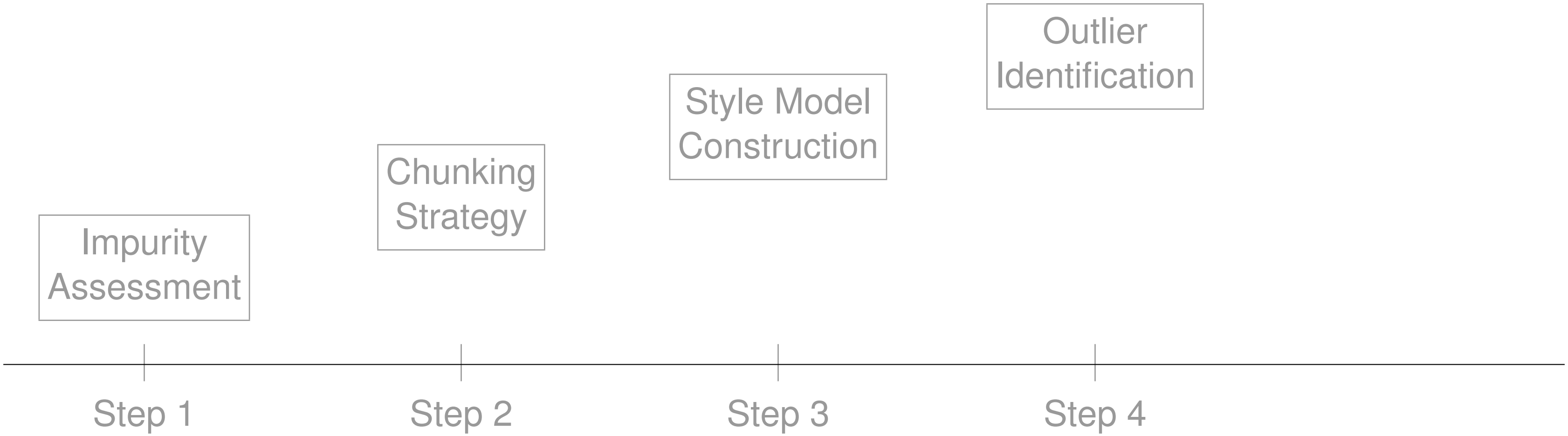
Algorithms for Machines: Outlier Identification



Compute maximum a-posteriori hypothesis under Naive Bayes.

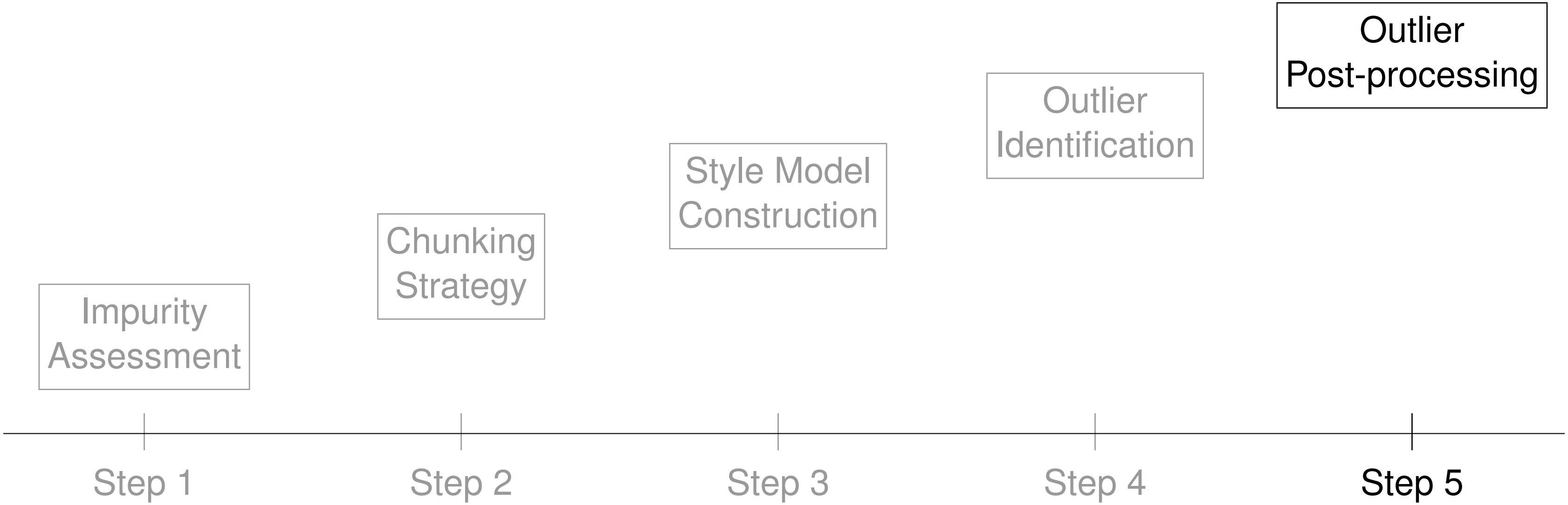
Intrinsic Plagiarism Detection

Algorithms for Machines



Intrinsic Plagiarism Detection

Algorithms for Machines



Since we are still unsecure . . .

How to obtain additional evidence about authorship?

- Raise precision at the expense of recall. (analyze ROC characteristic)
- If sufficient text is available, apply authorship verification technology. (unmasking)

Vandalism Detection in Wikipedia

Vandalism Detection in Wikipedia

Example: size, capitalization, punctuation, word existence

Bushido (rapper)

From Wikipedia, the free encyclopedia

(Difference between revisions)

Revision as of 07:33, 29 November 2009 (edit)	Revision as of 07:33, 29 November 2009 (edit) (undo)
71.193.157.91 (talk)	71.193.157.91 (talk)
(→Relationship with Bill Kaulitz)	*(→Lawsuits for assault and copyright infringement)*
← Previous edit	Next edit →
Line 36:	**Line 36:**
get rick rolld	get rick rolld
- ===Lawsuits for assault and copyright infringement===	+ get rick rolld
-	
More legal trouble found its way into his life following a party in Linz, Austria on July 30, 2005. Upon discovery that his tires had been slashed, he and two of his bodyguards were	

Vandalism Detection in Wikipedia

Example: size, capitalization, punctuation, word existence

Bushido (rapper)

From Wikipedia, the free encyclopedia

(Difference between revisions)

Revision as of 07:33, 29 November 2009 (edit)	Revision as of 07:33, 29 November 2009 (edit) (undo)
71.193.157.91 (talk)	71.193.157.91 (talk)
(→Relationship with Bill Kaulitz)	(→Lawsuits for assault and copyright infringement)
← Previous edit	Next edit →

Line 36:

get rick

- ===Law

-

More le

2005. U

Contents [hide]

Background information	
Birth name	Anis Mohamed Youss Ferchichi
Born	September 28, 1978 (a
Origin	Tempelhof ,Berlin, Ger
Genres	Hip-Hop, Gangsta Rap
Occupations	Rapper & Producer
Years active	1998-present
Labels	I Luv Money Records (1999-2001) Aggro Berlin (2001-20 ersguterjunge (2004-p
Associated acts	Baba Saad, Chakuza, Fler
Website	www.kingbushido.de

Musical career

get rick rolld

get rick rolld

get rick rolld

get rick rolld

Vandalism Detection in Wikipedia

Example: size, capitalization, punctuation, repetition

Henry Wadsworth Longfellow

From Wikipedia, the free encyclopedia

(Difference between revisions)

Revision as of 17:33, 1 December 2009 (edit)	Revision as of 22:02, 1 December 2009 (edit) (undo)
Snigbrook (talk | contribs)	141.154.179.145 (talk)
m *(Reverted edits by 70.57.239.4 to last revision by Snigbrook (HG))*	*(→Critical response)*
← Previous edit	Next edit →
Line 89:	**Line 89:**
===Critical response===	===Critical response===
[[Image:Sumner-Longfellow.jpg|thumb|right|Longfellow and his good friend [[United States Senate|Senator]] [[Charles Sumner]]]]	[[Image:Sumner-Longfellow.jpg|thumb|right|Longfellow and his good friend [[United Sta Senate|Senator]] [[Charles Sumner]]]]
Longfellow's early collections, "Voices of the Night" and "Ballads and Other Poems", made him instantly popular. The "New-Yorker" called him "one of the very few in our time who has	Longfellow's early collections, ugly"Voices of the Night" and "Ballads and Other Poems" made him instantly popular. The "New-Yorker" called him "one of the very few in our tim

Vandalism Detection in Wikipedia

Example: size, capitalization, punctuation, repetition

Henry Wadsworth Longfellow

From Wikipedia, the free encyclopedia

(Difference between revisions)

Revision as of 17:33, 1 December 2009 (edit)	Revision as of 22:02, 1 December 2009 (edit) (undo)
Snigbrook (talk | contribs)	141.154.179.145 (talk)
m *(Reverted edits by 70.57.239.4 to last revision by Snigbrook (HG))*	*(→Critical response)*
← Previous edit	Next edit →

Line 89:

===Cri

[[Imag

Senat

Longfe

him in

want a national literature altogether shaggy and unshorn, that shall shake the earth, like a herd of buffaloes thundering over the prairies.[99]

He was also important as a translator; his translation of Dante became a required possession for those who wanted to be a part of high culture.[100] He als encouraged and supported other translators. In 1845, he published *The Poets and Poetry of Europe*, an 800-page compilation of translations made by othe writers, including many by his friend and colleague Cornelius Conway Felton. Longfellow intended the anthology "to bring together, into a compact and cor form, as large an amount as possible of those English translations which are scattered through many volumes, and are not accessible to the general read In honor of Longfellow's role with translations, Harvard established the Longfellow Institute in 1994, dedicated to literature written in the United States in lar other than English.[102]

In 1874, Longfellow oversaw a 31-volume anthology called *Poems of Places*, which collected poems representing several geographical locations, including European, Asian, and Arabian countries.[103] Emerson was disappointed and reportedly told Longfellow: "The world is expecting better things of you than t You are wasting time that should be bestowed upon original production".[104] In preparing the volume, Longfellow hired Katherine Sherwood Bonner as an amanuensis.[105]

Critical response

Longfellow's early collections, ugly *Voices of the Night* and *Ballads and Other Poems*, made him instantly popular. The *New-Yorker* called him "one of the very few in our time who has successfully aimed in putting poetry to its best and sweetest uses".[45] The *Southern Literary Messenger* immediately put Longfellow "among the first of our American poets".[45] Poet John Greenleaf Whittier said that Longfellow's poetry illustrated "the careful moulding by which art attains the graceful ease and chaste simplicity of nature".[106] Longfellow's friend Oliver Wendell Holmes, Sr. wrote of him as "our chief singer" and one who "wins and warms... kindles, softens, cheers [and] calms the wildest woe and stays the bitterest tears!"[107]

The rapidity with which American readers embraced Longfellow was unparalleled in publishing history in the United States;[108] by 1874, he was earning $3,000 per poem.[109] His popularity spread throughout Europe as well and his

Vandalism Detection in Wikipedia

Example: vulgarism, sentiment

Jingle Bells/U Can't Touch This

From Wikipedia, the free encyclopedia

(Difference between revisions)

Revision as of 12:41, 28 August 2009 (edit) 85.132.47.9 (talk) (→Single track listing) ← Previous edit	Revision as of 09:50, 2 December 2009 (edit) (undo) Ragger256 (talk | contribs) (→Jingle Bells) Next edit →
Line 55:	**Line 55:**
===Jingle Bells===	===Jingle Bells===
- In some clips, his genitals are censored.	+ In some clips, his genitals are HUGE.
The video starts with the Crazy Frog playing in the snow with the bounty hunter robot from previous clips. It then shows flashbacks from clips of "[[Axel F#Crazy Frog version|Axel F]]," a trip to a carnival, and the Crazy DJ clip, then more of the "Axel F" clip. The flashbacks end,	The video starts with the Crazy Frog playing in the snow with the bounty hunter robot fro previous clips. It then shows flashbacks from clips of "[[Axel F#Crazy Frog version|Axel F]] trip to a carnival, and the Crazy DJ clip, then more of the "Axel F" clip. The flashbacks e

Vandalism Detection in Wikipedia

Example: vulgarism, sentiment

Jingle Bells/U Can't Touch This

From Wikipedia, the free encyclopedia
(Difference between revisions)

Revision as of 12:41, 28 August 2009 (edit)
85.132.47.9 (talk)
(→Single track listing)

Revision as of 09:50, 2 December 2009 (edit) (undo)
Ragger256 (talk | contribs)
(→Jingle Bells)

Line 55:

Music videos

"U Can't Touch This"

The video depicts Crazy Frog causes chaos at the underwater sealab of "The Boss".

Jingle Bells

In some clips, his genitals are HUGE.

The video starts with the Crazy Frog playing in the snow with the bounty hunter robot from previous clips. It then shows flashbacks from clips of "Axel F," a "Axel F" clip. The flashbacks end, and the bounty hunter robot begins to throw a snowball at the frog. But instead he kisses the bounty hunter robot, and h then make snow angels and a message reads "Have a ding dong Christmas, everyone!"

Certifications

Country	Certification	Date	Sales certified
Australia[3]	Gold	2005	35,000

Charts

Chart (2006)	Peak position

End of year chart (2005)	Position

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:	**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:

Grou

[[Grou

prima

</ref>

- ==Hist

History▯▯▯Ô▯▯▯Ô

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."

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

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

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, 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

- 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, defamation

Mainframe computer

From Wikipedia, the free encyclopedia

(Difference between revisions)

Revision as of 15:57, 17 November 2009 (edit)	Revision as of 14:48, 24 November 2009 (edit) (undo)
TXiKiBoT (talk | contribs)	74.202.102.94 (talk)
m *(robot Adding: bg:Мейнфрейм компютър)*	Next edit →
← Previous edit	

Line 12:		**Line 12:**	
	Today in practice, the term usually refers{{Citation needed|date=November 2009}} to computers compatible with the [[IBM System/360]] line, first introduced in 1965. ([[IBM System z10]] is the latest incarnation.) Otherwise, large systems that are not based on the System/360 but are used for similar tasks are usually referred to as [[computer server|servers]] or even [[supercomputer]]s. However, "[[server]]", "[[supercomputer]]" and "mainframe" are not synonymous (see [[client-server]]).		Today in practice, the term usually refers{{Citation needed|date=November 2009}} to computers compatible with the [[IBM System/360]] line, first introduced in 1965. ([[IBM System z10]] is the latest incarnation.) Otherwise, large systems that are not based on th System/360 but are used for similar tasks are usually referred to as [[computer server|serv or even [[supercomputer]]s. However, "[[server]]", "[[supercomputer]]" and "mainframe" a not synonymous (see [[client-server]]).
-	Some non-System/360-compatible systems derived from or compatible with older (pre-Web) server technology may also be considered mainframes. These include the [[Burroughs large systems]], the [[UNIVAC 1100/2200 series]] systems, and the pre-System/360 [[IBM 700/7000 series]]. Most large-scale computer system architectures were firmly established in the 1960s and most large computers were based on architecture established during that era up until the advent of Web servers in the 1990s. (Interestingly, the first Web server running anywhere outside Switzerland ran on an IBM mainframe at Stanford University as early as 1990. See [[History of the World Wide Web]] for details.)	+	Some non-System/360-compatible systems derived from or compatible with older (pre-W server technology may also be considered mainframes. These include the [[Burroughs la systems]], the [[UNIVAC 1100/2200 series]] systems, and the pre-System/360 [[IBM 700/7 series]]. Most large-scale computer system architectures were firmly established in the 19 and most large computers were based on architecture established during that era up unt the advent of Web servers in the 1990s. (Interestingly, the first Web server running anywh outside Switzerland ran on an IBM mainframe at Stanford University as early as 1990. Harris HAcks the Mainframe all the time. See [[History of the World Wide Web]] for details.)

Vandalism Detection in Wikipedia

Example: wrong facts, defamation

Mainframe computer

From Wikipedia, the free encyclopedia
(Difference between revisions)

Revision as of 15:57, 17 November 2009 (edit)	Revision as of 14:48, 24 November 2009 (edit) (undo)
TXiKiBoT (talk | contribs)	74.202.102.94 (talk)
m *(robot Adding: bg:Мейнфрейм компютър)*	Next edit →
← Previous edit	

Line 12:	Line 12:
Today in practice, the term usually refers{{Citation needed|date=November 2009}} to computers compatible with the [[IBM System/360]] line, first introduced in 1965. ([[IBM System z10]] is the latest incarnation.) Otherwise, large systems that are not based on the System/360 but are used for similar tasks are usually referred to as [[computer server|servers]] or eve… not sy…	Today in practice, the term usually refers{{Citation needed|date=November 2009}} to computers compatible with the [[IBM System/360]] line, first introduced in 1965. ([[IBM System z10]] is the latest incarnation.) Otherwise, large systems that are not based on th… System/360 but are used for similar tasks are usually referred to as [[computer server|serv…
Some… serve… system… series]… and m… the a… outsid… [[Hist…	

Mainframes (often colloquially referred to as **Big Iron**[1]) are computers used mainly by large organizations for critical applications, typically bulk data processing such as census, industry and consumer statistics, enterprise resource planning, and financial transaction processing.

An IBM 704 mainframe

The term probably had originated from the early mainframes, as they were housed in enormous, room-sized metal boxes or frames.[2] Later the term was used to distinguish high-end commercial machines from less powerful units.

Today in practice, the term usually refers[citation needed] to computers compatible with the IBM System/360 line, first introduced in 1965. (IBM System z10 is the latest incarnation.) Otherwise, large systems that are not based on the System/360 but are used for similar tasks are usually referred to as servers or even supercomputers. However, "server", "supercomputer" and "mainframe" are not synonymous (see client-server).

Some non-System/360-compatible systems derived from or compatible with older (pre-Web) server technology may also be considered mainframes. These the Burroughs large systems, the UNIVAC 1100/2200 series systems, and the pre-System/360 IBM 700/7000 series. Most large-scale computer system architectures were firmly established in the 1960s and most large computers were based on architecture established during that era up until the advent of W… servers in the 1990s. (Interestingly, the first Web server running anywhere outside Switzerland ran on an IBM mainframe at Stanford University as early as 1… Harris HAcks the Mainframe all the time. See History of the World Wide Web for details.)

Vandalism Detection in Wikipedia

Example: opinionated

Olga Kurylenko

From Wikipedia, the free encyclopedia

(Difference between revisions)

Revision as of 06:00, 19 November 2009 (edit) 95.179.86.49 (talk) (→Personal life) ← Previous edit	Revision as of 21:05, 19 November 2009 (edit) (undo) 86.158.13.40 (talk) Next edit →
Line 18:	**Line 18:**
| homepage =	| homepage =
}}	}}
'''Olga Kostyantynivna Kurylenko''' ({{lang-uk|Ольга Костянтинівна Куриленко}}; born November 14, 1979) is an [[actress]] and [[Model (person)|model]]. She is perhaps best known as the [[Bond girl]], [[Camille Montes]], in the 22nd [[James Bond]] film, ''[[Quantum	'''Olga Kostyantynivna Kurylenko''' ({{lang-uk|Ольга Костянтинівна Куриленко}}; born November 14, 1979) is fit and an [[actress]] and [[Model (person)|model]]. She is perhap best known as the [[Bond girl]], [[Camille Montes]], in the 22nd [[James Bond]] film,

Vandalism Detection in Wikipedia

Example: opinionated

Olga Kurylenko

From Wikipedia, the free encyclopedia

(Difference between revisions)

Revision as of 06:00, 19 November 2009 (edit)	Revision as of 21:05, 19 November 2009 (edit) (undo)
95.179.86.49 (talk)	86.158.13.40 (talk)
(→Personal life)	Next edit →
← Previous edit	
Line 18:	**Line 18:**
| homepage =	| homepage =
}}	}}
'''Olga Kostyantynivna Kurylenko''' ({{lang-uk|Ольга Костянтинівна Куриленко}}; born	'''Olga Kostyantynivna Kurylenko''' ({{lang-uk|Ольга Костянтинівна Куриленко}}; born
Nove	
known	

Revision as of 21:05, 19 November 2009

Olga Kostyantynivna Kurylenko (Ukrainian: Ольга Костянтинівна Куриленко; born November 14, 1979) is fit and an actress and model. She is perhaps best known as the Bond girl, Camille Montes, in the 22nd James Bond film, *Quantum of Solace*. She also portrayed Nika Boronina in the movie adaptation of the video game *Hitman*. Born in Ukraine, she became a French citizen in 2001.[3]

Olga Kurylenko
Ольга Куриленко

Contents [hide]

Early life and background

Olga Kurylenko was born in Berdyansk, Ukraine. Her father, Kostyantyn Kurylenko, is Ukrainian and her mother,

Vandalism Detection in Wikipedia

Example: wrong facts, 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, 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>

\+

==Mai

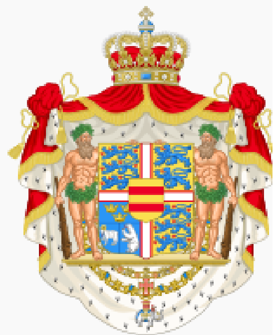
Revision as of 06:21, 29 November 2009

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

Vandalism Detection in Wikipedia

Example: wrong facts, article history may suggest otherwise

Jerome Is the New Black

From Wikipedia, the free encyclopedia
(Difference between revisions)

Revision as of 01:29, 22 November 2009 (edit) GageSkidmore (talk | contribs) ← Previous edit		Revision as of 00:49, 23 November 2009 (edit) (undo) 173.79.146.174 (talk) Next edit →	
Line 6:		**Line 6:**	
	| Season = 8		| Season = 8
	| Episode = 7		| Episode = 7
-	| Airdate = November 22, 2009	+	| Airdate = November 22, 1990
	| Production = 7ACX08		| Production = 7ACX08
	| Writer = TBA		| Writer = TBA

Revision as of 00:49, 23 November 2009

"Jerome is the New Black" is the seventh episode of the eighth season of *Family Guy*. It is scheduled to air on November 22, 2009 on Fox.

"Jerome is the New Black"

Family Guy **episode**

Plot

Jerome is a candidate when Peter and his friends interview potential friends to fill the vacancy left by Cleveland. It is soon discovered that Quagmire hates him.[1]

References

1. ^ [a] [b] http://www.foxflash.com/div.php/main/page?aID=1z4&mo=11&d=15

Vandalism Detection in Wikipedia

The Machine Learning Perspective

The achievements of ML enfold their full power in discrimination situations.

Vandalism Detection in Wikipedia

The Machine Learning Perspective

The achievements of ML enfold their full power in discrimination situations.

The tasks

- intrinsic plagiarism analysis
- authorship verification
- vandalism detection

share a particular characteristic: they are one-class classification problems.

→ Feature engineering plays an outstanding role.

Vandalism Detection in Wikipedia

Two Types of Edit Features

Vandalism Detection in Wikipedia

Two Types of Edit Features: Content-based

Feature	Description
Character-level Features	
Capitalization	Ratio of upper case chars to lower case chars (all chars)
Distribution	Kullback-Leibler divergence of the char distribution from the expectation
Compressibility	Compression rate of the edit differences
Markup	Ratio of new (changed) wikitext chars to all wikitext chars
Word-level Features	
Vulgarism	Frequency of vulgar words
Pronouns	Frequency of personal pronouns
Sentiment	Frequency of sentiment words
Spelling and Grammar Features	
Word Existence	Ratio of words that occur in an English dictionary
Spelling	Frequency (impact) of spelling errors
Grammar	Number of grammatical errors
Edit Type Features	
Edit Type	The edit is an insertion, deletion, modification, or a combination
Replacement	The article (a paragraph) is completely replaced, excluding its title

Vandalism Detection in Wikipedia

Two Types of Edit Features: Context-based

Feature	Description
Edit Comment Features	
Existence	A comment was given
Length	Length of the comment
Edit Time Features	
Edit time	Hour of the day the edit was made
Successiveness	Logarithm of the time difference to the previous edit
Article Revision History Features	
Revisions	Number of revisions
Regular	Number of regular edits
Article Trustworthiness Features	
Suspect Topic	The article is on the list of often vandalized articles
WikiTrust	Values from the WikiTrust trust histogram
Editor Reputation Features	
Anonymous	Anonymous editor
Reputation	Scores that compute a user's reputation based on previous edits
Registration	Time the editor was registered with Wikipedia

The PAN Competition Continued

The PAN Competition Continued

1st International Competition on Wikipedia Vandalism Detection, PAN 2010

Facts:

- organized as CLEF 2010 Lab
- 9 groups from 5 countries participated, 5 groups from the USA
- 15 weeks of training and testing (March – June)
- the corpus was newly created for the purpose of the competition

The PAN Competition Continued

1st International Competition on Wikipedia Vandalism Detection, PAN 2010

Facts:

- organized as CLEF 2010 Lab
- 9 groups from 5 countries participated, 5 groups from the USA
- 15 weeks of training and testing (March – June)
- the corpus was newly created for the purpose of the competition

Task:

Given a set of edits on Wikipedia articles,
distinguish ill-intentioned edits from well-intentioned edits.

The PAN Competition Continued

Vandalism Corpus PAN-WVC-10

Large-scale resource for the controlled evaluation of detection algorithms:

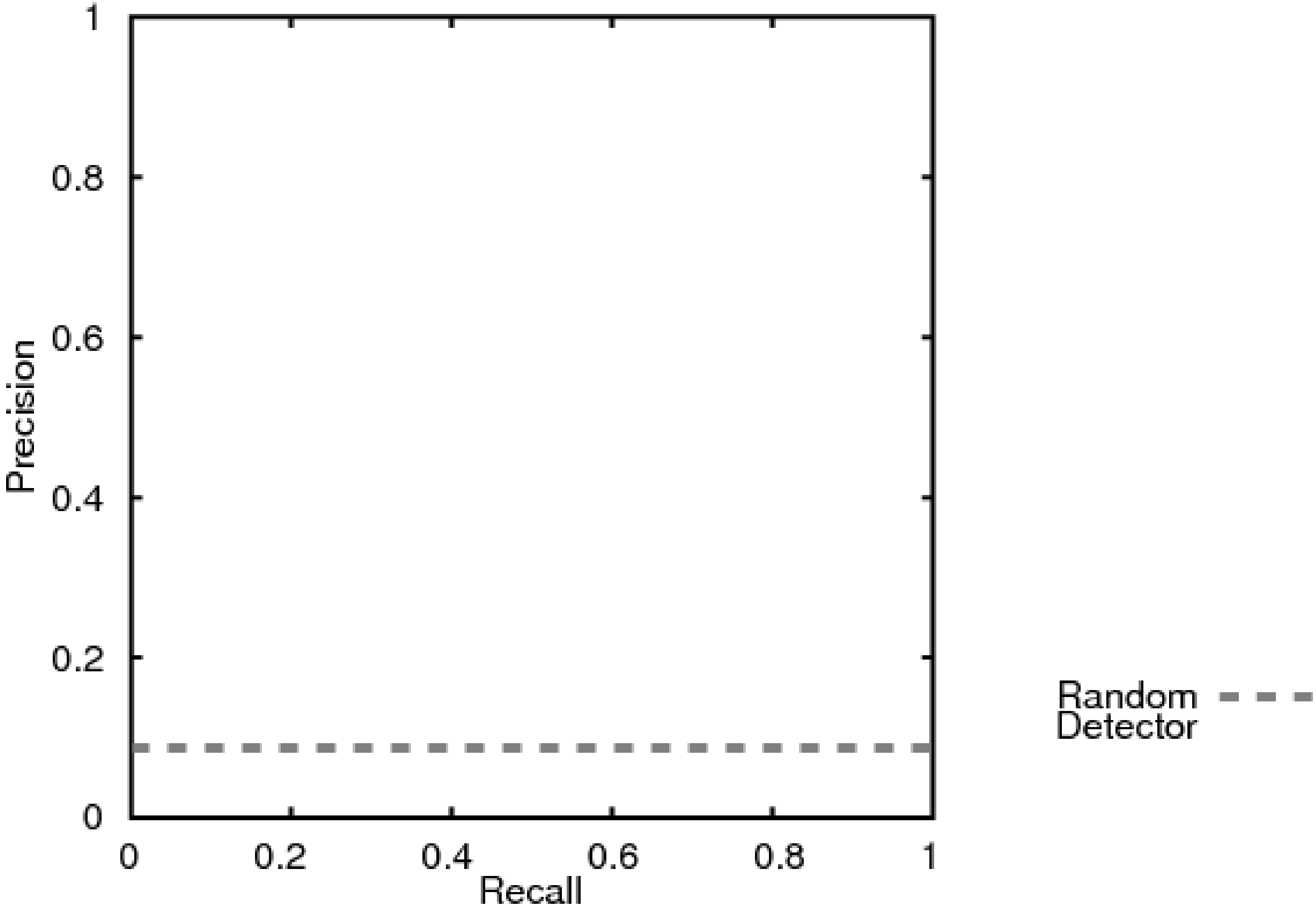
- 32 452 edits (sampled from a week's worth of Wikipedia edit logs)
- 28 468 different edited articles (edit frequency resembles article importance)
- 2391 edits are vandalism (a 7% ratio is in concordance with the literature)

The edits in PAN-WVC-10 have been reviewed by 753 human annotators, recruited at Amazon's Mechanical Turk:

- Each edit was reviewed by at least 3 different annotators.
- If the annotators did not agree, the edit was reviewed again by 3 other.
- If still less than 2/3 of the annotators agreed, 3 more annotators were asked.
- After 8 iterations only 70 edits remained in a tie, which proofed to be tough choices.

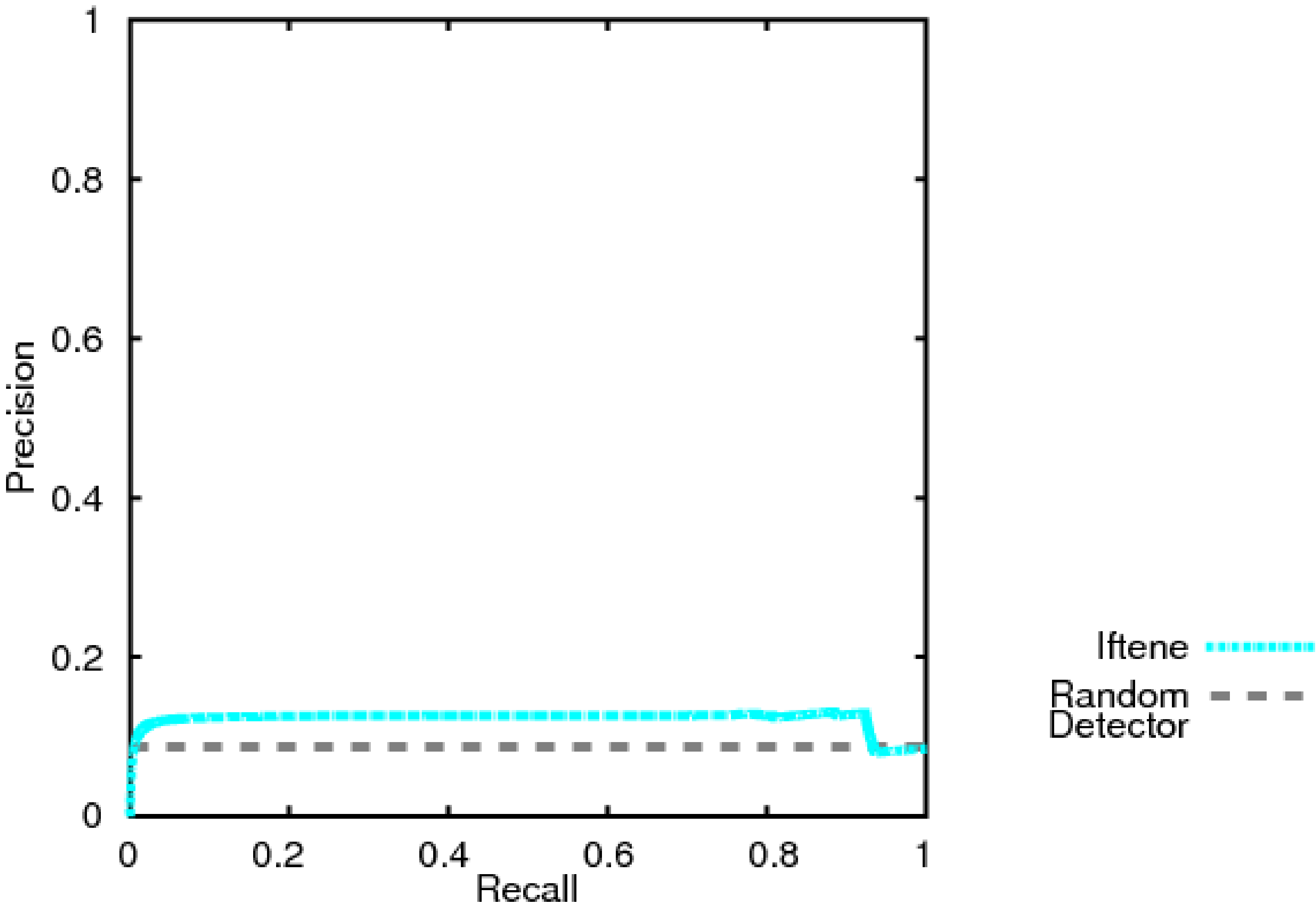
The PAN Competition Continued

Vandalism Detection Results



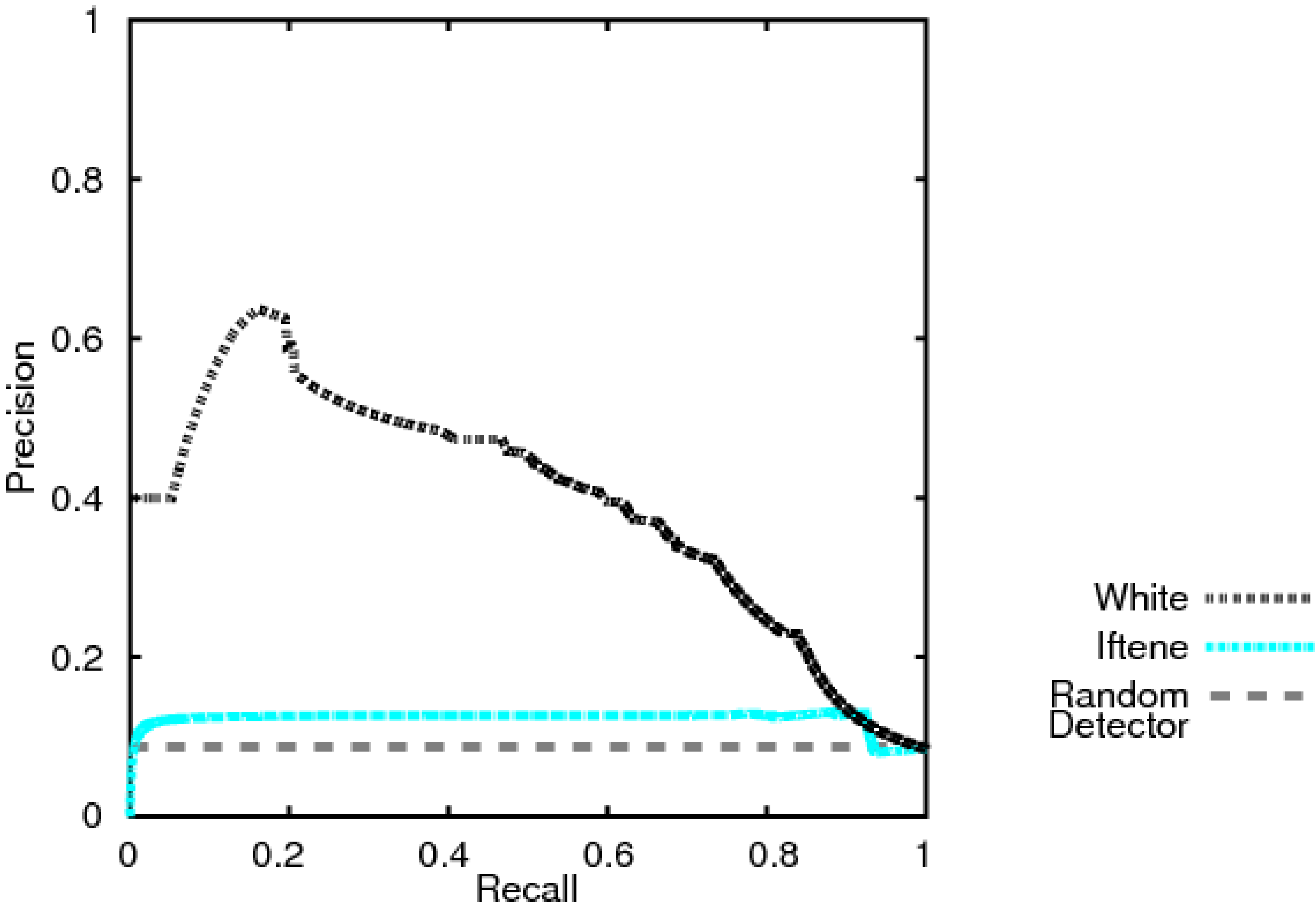
The PAN Competition Continued

Vandalism Detection Results



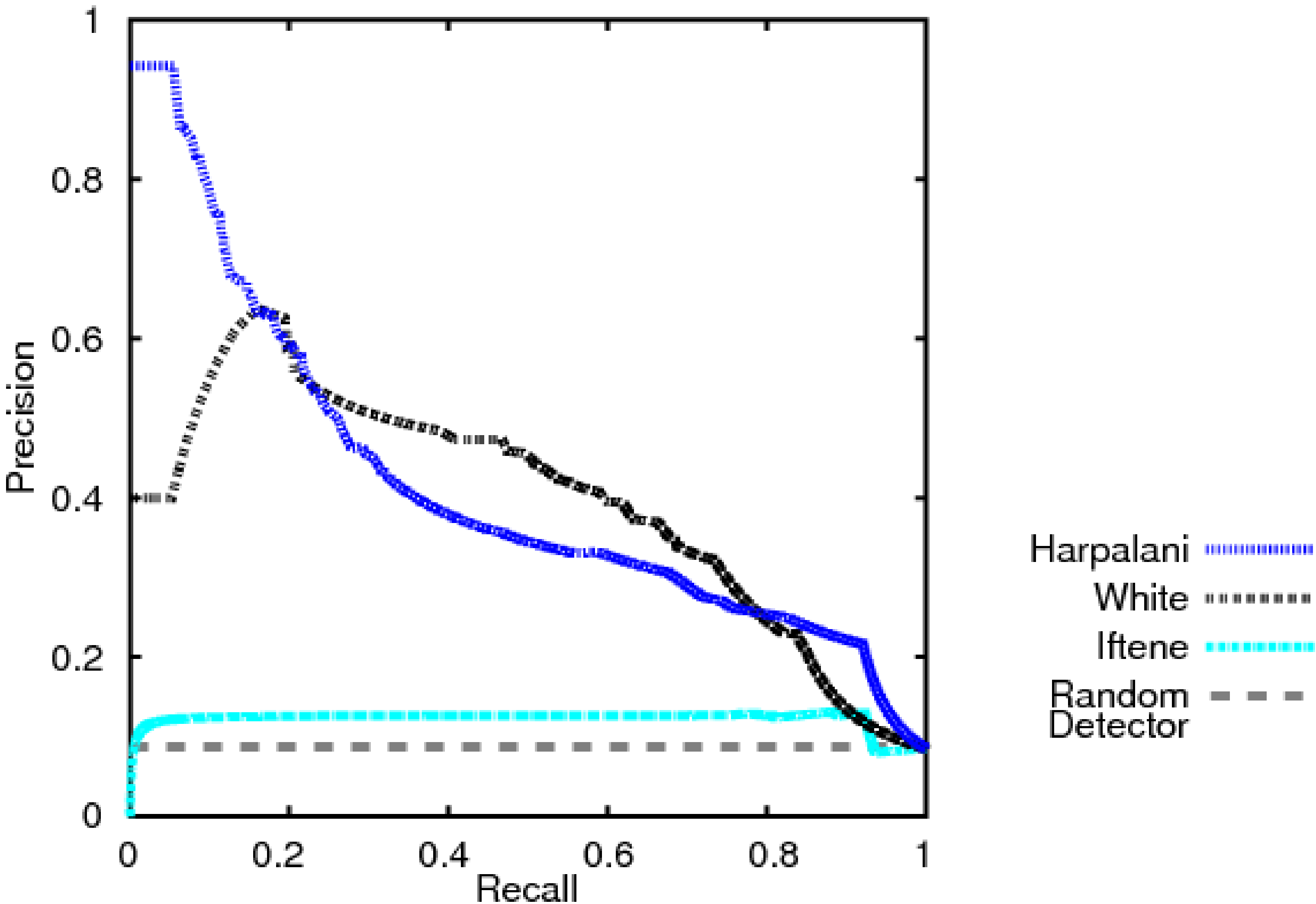
The PAN Competition Continued

Vandalism Detection Results



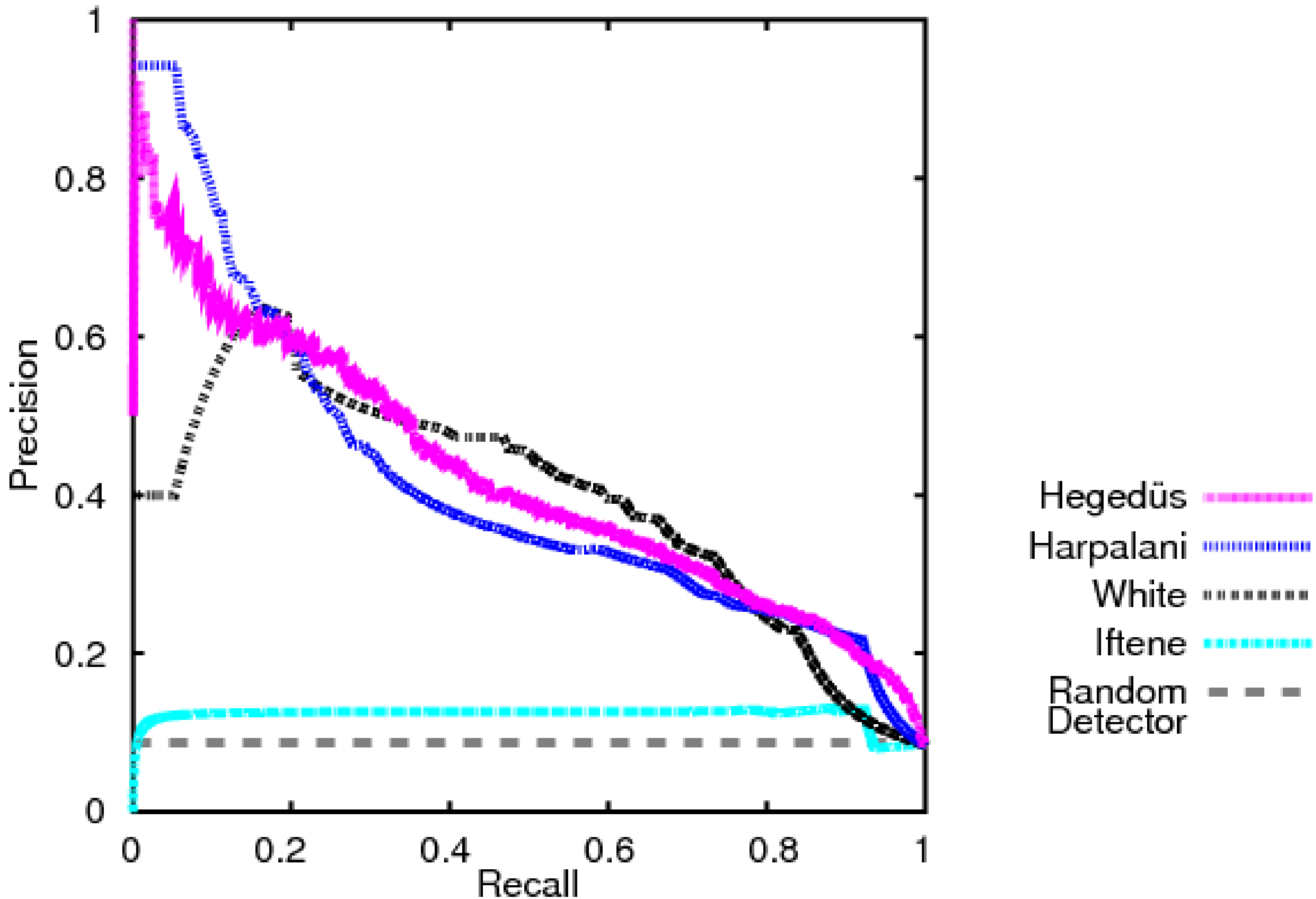
The PAN Competition Continued

Vandalism Detection Results



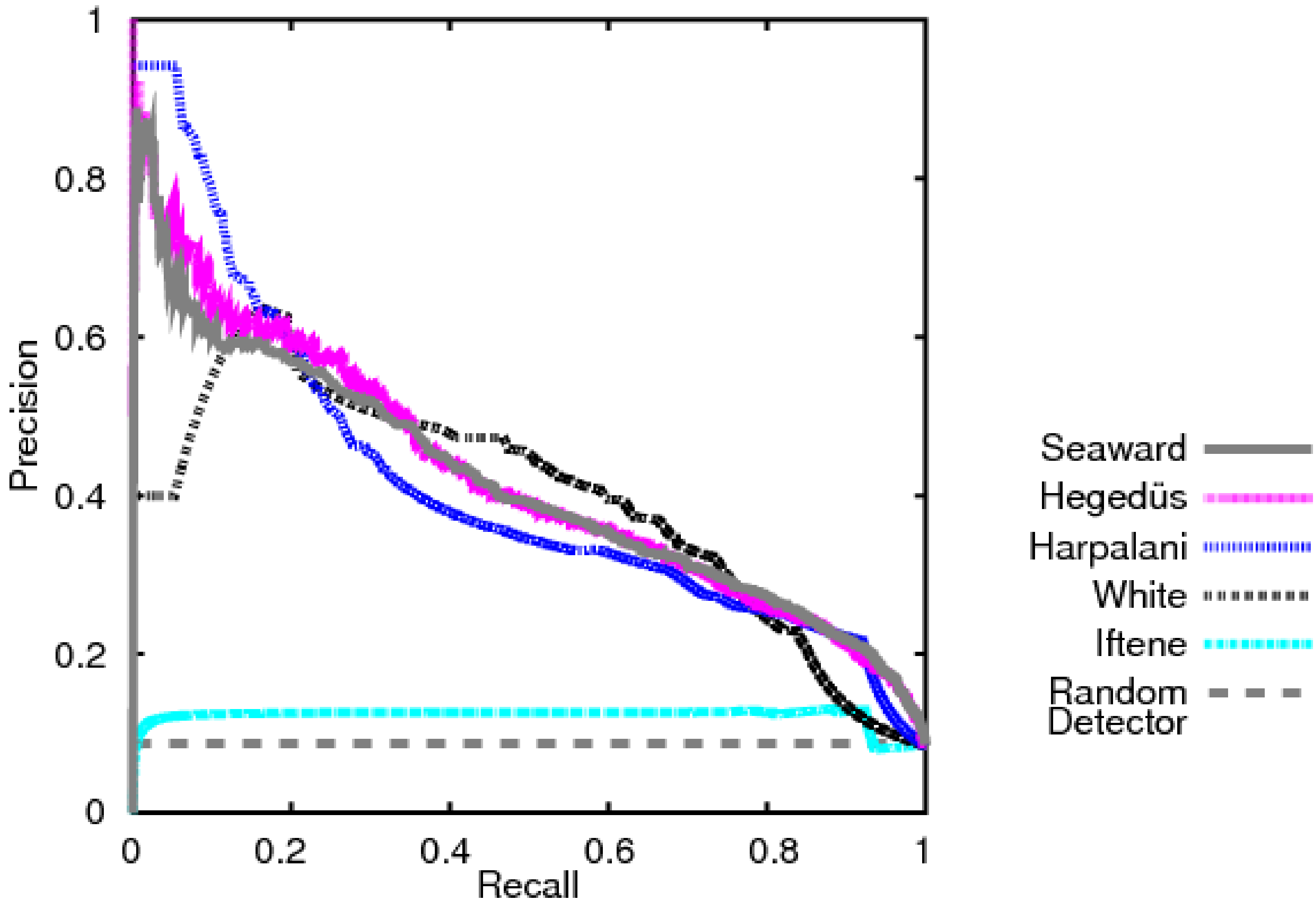
The PAN Competition Continued

Vandalism Detection Results



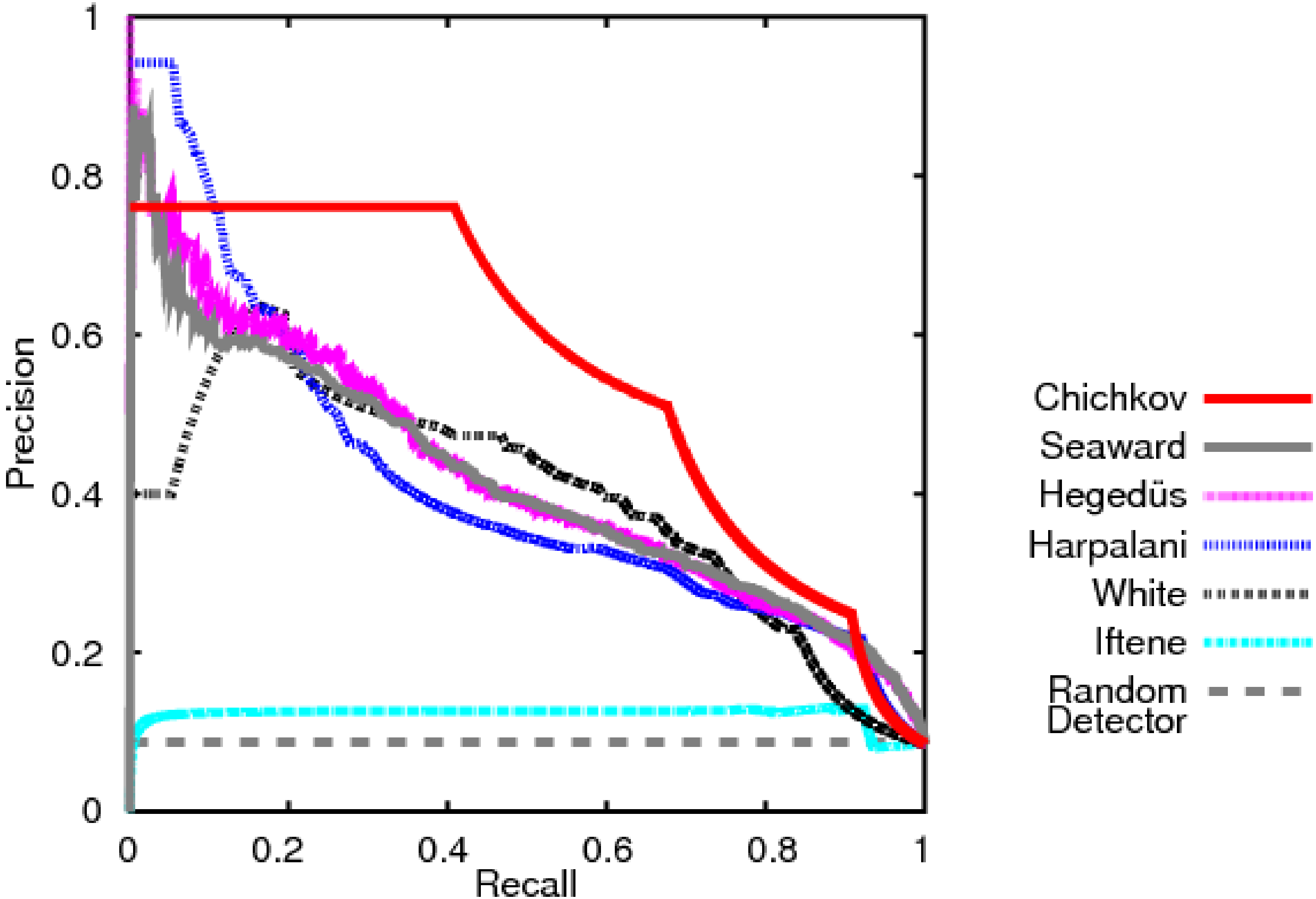
The PAN Competition Continued

Vandalism Detection Results



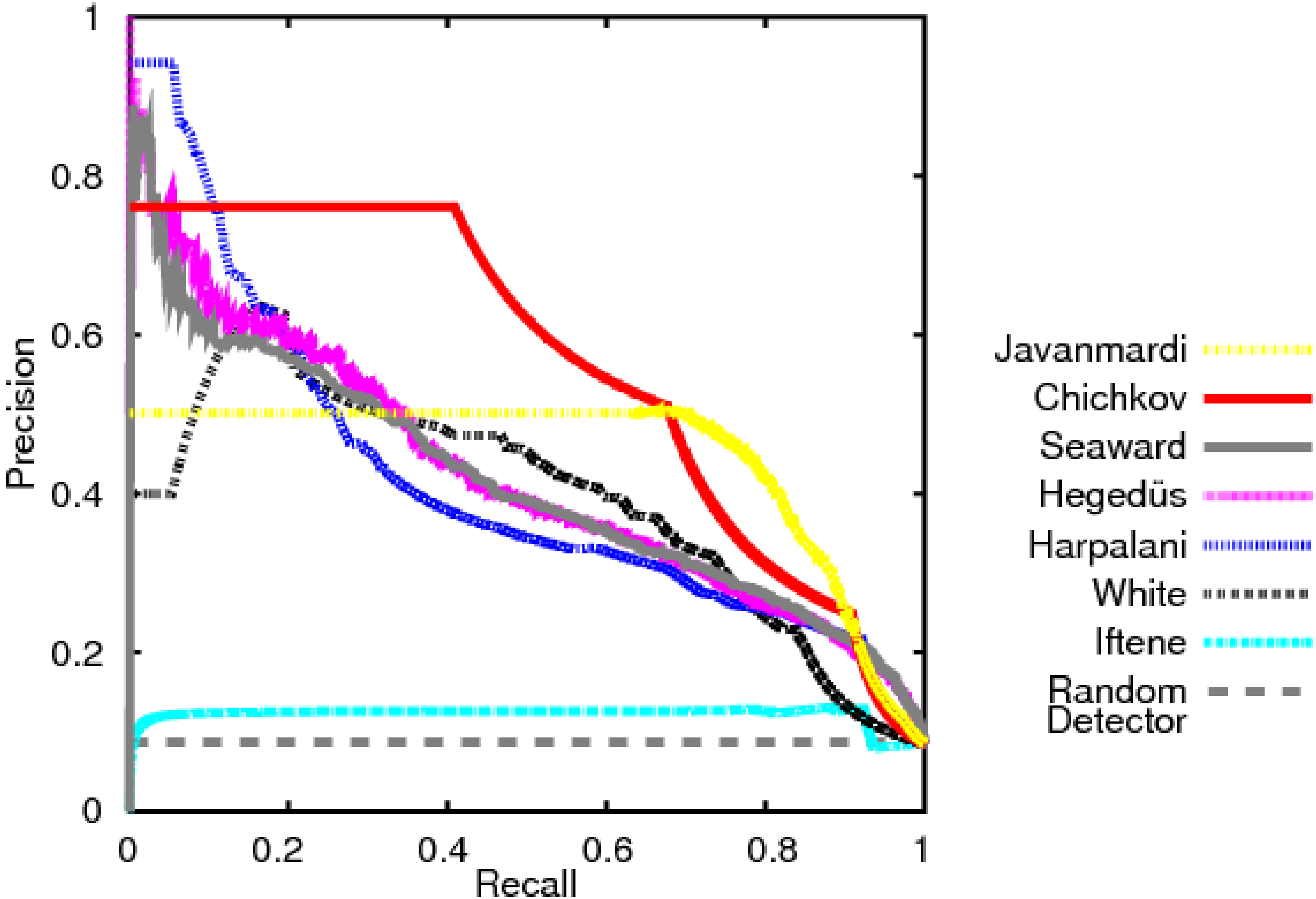
The PAN Competition Continued

Vandalism Detection Results



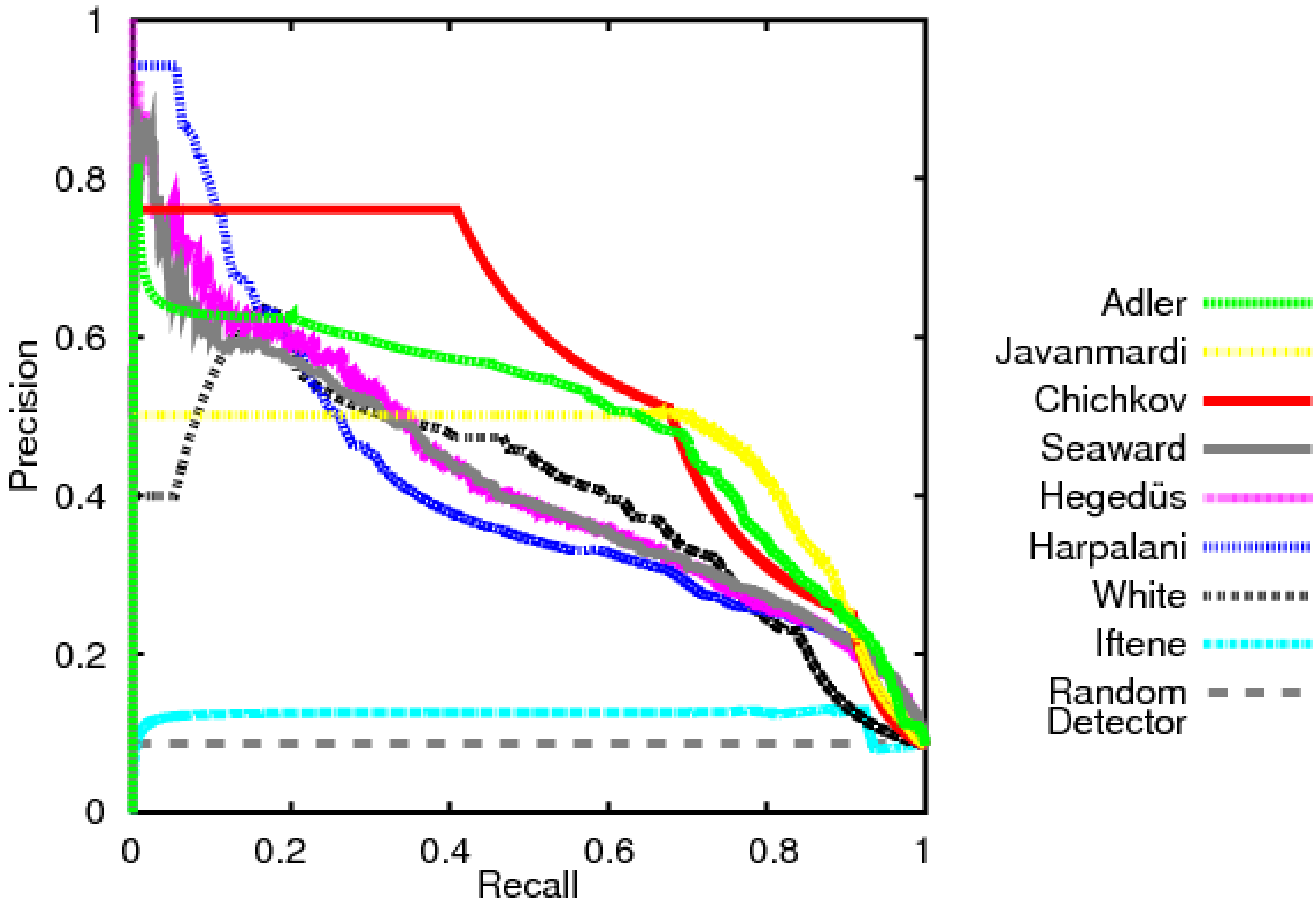
The PAN Competition Continued

Vandalism Detection Results



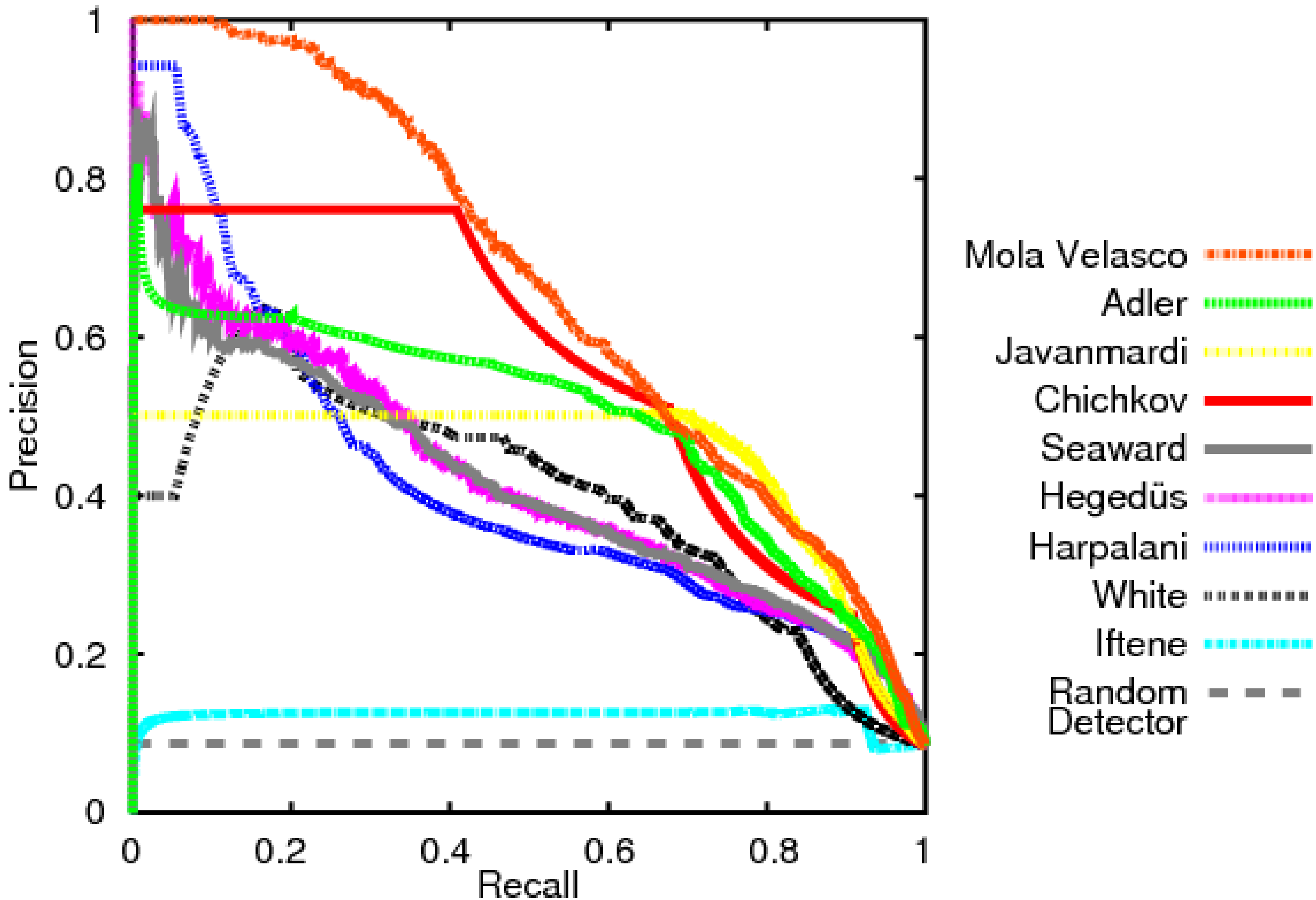
The PAN Competition Continued

Vandalism Detection Results



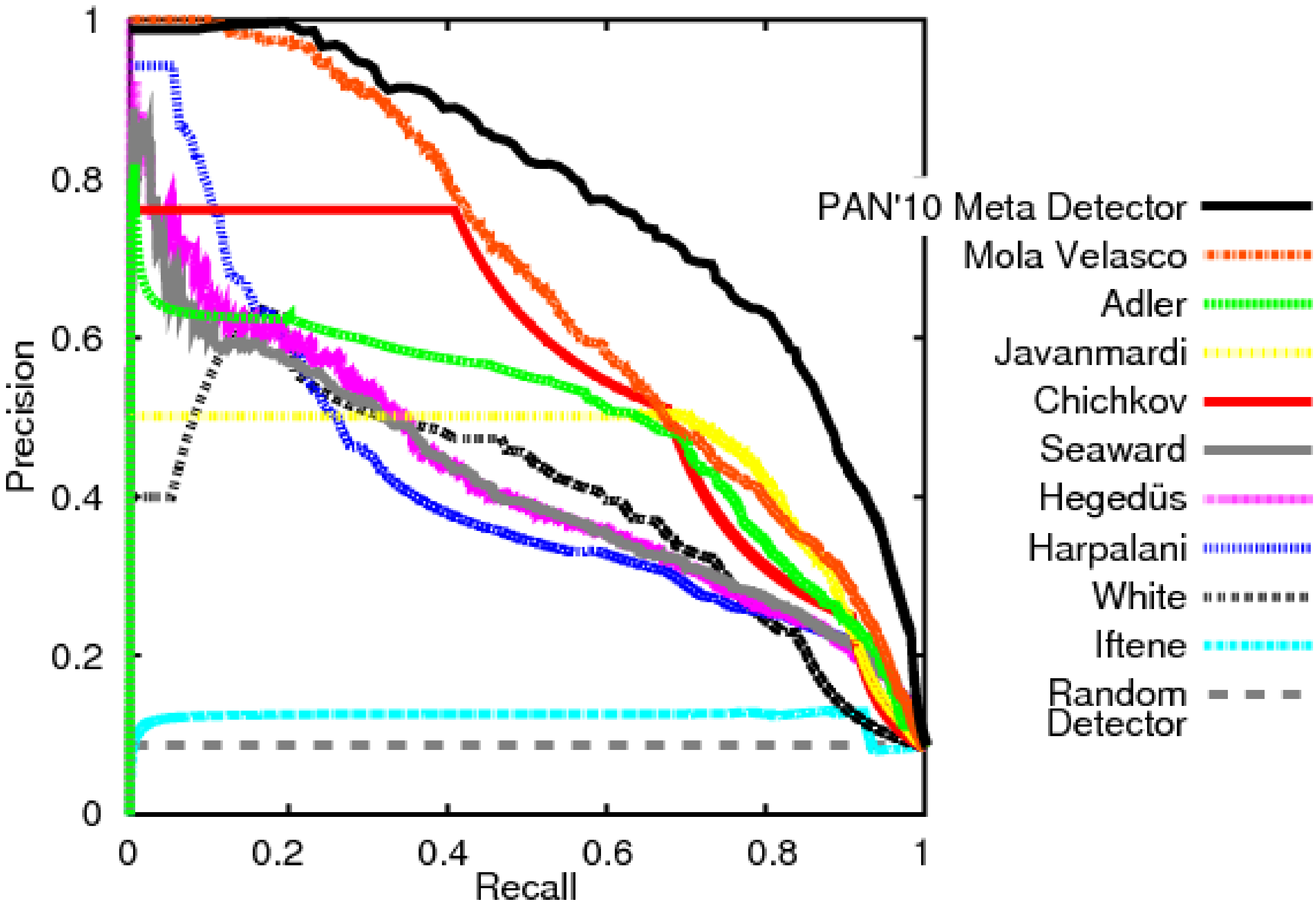
The PAN Competition Continued

Vandalism Detection Results



The PAN Competition Continued

Vandalism Detection Results



Almost the End

Almost the End

What We have Seen

- Machine-executable plagiarism detection (external)
  1. keyword extraction
  2. heuristic search
  3. document selection
  4. detailed analysis
  5. citation analysis
- Machine-executable plagiarism detection (intrinsic)
  1. impurity assessment
  2. chunking strategy
  3. style model construction
  4. outlier identification
  5. authorship verification
- Machine-executable vandalism detection
- Selected PAN competition results
- Various details

Almost the End

Some Take-away Messages ;–)

- The frontiers of external plagiarism detection
  - document access
  - processing time
  - understanding of human search behavior

Almost the End

Some Take-away Messages ;–)

- The frontiers of external plagiarism detection
  - document access
  - processing time
  - understanding of human search behavior
- The frontiers of intrinsic plagiarism detection
  - text length (at least 500 words)

Almost the End

Some Take-away Messages ;–)

- The frontiers of external plagiarism detection
  - document access
  - processing time
  - understanding of human search behavior
- The frontiers of intrinsic plagiarism detection
  - text length (at least 500 words)
- The frontiers of vandalism detection
  - deep text understanding (semantics and pragmatics)

Almost the End

Some Take-away Messages ;–)

- The frontiers of external plagiarism detection
  - document access
  - processing time
  - understanding of human search behavior
- The frontiers of intrinsic plagiarism detection
  - text length (at least 500 words)
- The frontiers of vandalism detection
  - deep text understanding (semantics and pragmatics)
- Support by the crowd will increase
  - human cheaper than machine—sometimes (currently and medium term)
  - human intelligence tasks at AMT

Almost the End

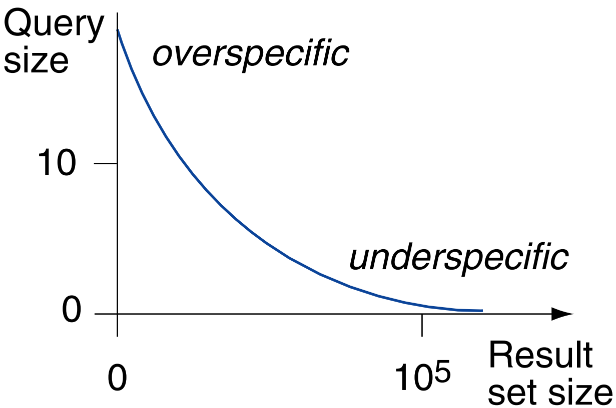
Some Take-away Messages ;–)

- The frontiers of external plagiarism detection
  - document access
  - processing time
  - understanding of human search behavior
- The frontiers of intrinsic plagiarism detection
  - text length (at least 500 words)
- The frontiers of vandalism detection
  - deep text understanding (semantics and pragmatics)
- Support by the crowd will increase
  - human cheaper than machine—sometimes (currently and medium term)
  - human intelligence tasks at AMT
- Untapped potential
  - integration of NLP into IR and IE (becomes popular)
  - integration of AI into IR and IE (in its infancy)

Thank you!

The User over Ranking Hypothesis [Stein/Hagen 2010]

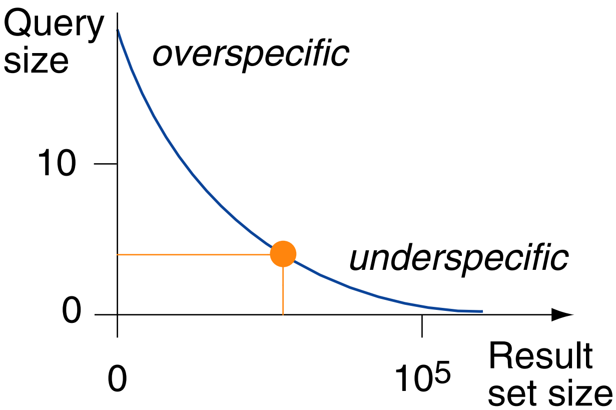
The User over Ranking Hypothesis



Query Specificity

The User over Ranking Hypothesis

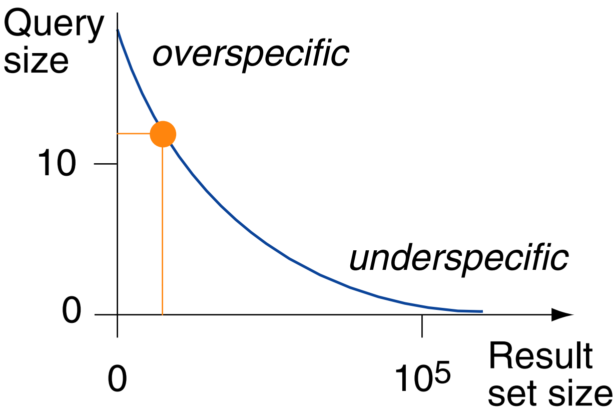
- User / keyword extractor has enough information to overspecify a search.



Query Specificity

The User over Ranking Hypothesis

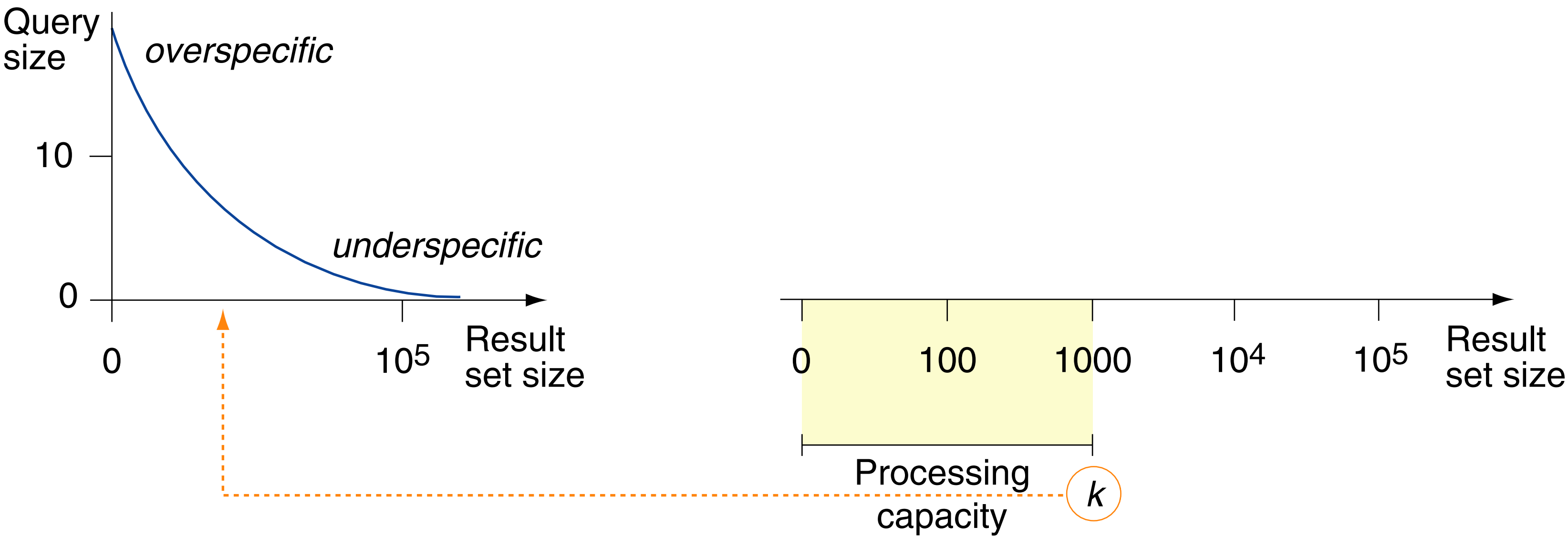
- User / keyword extractor has enough information to overspecify a search.



Query Specificity

The User over Ranking Hypothesis

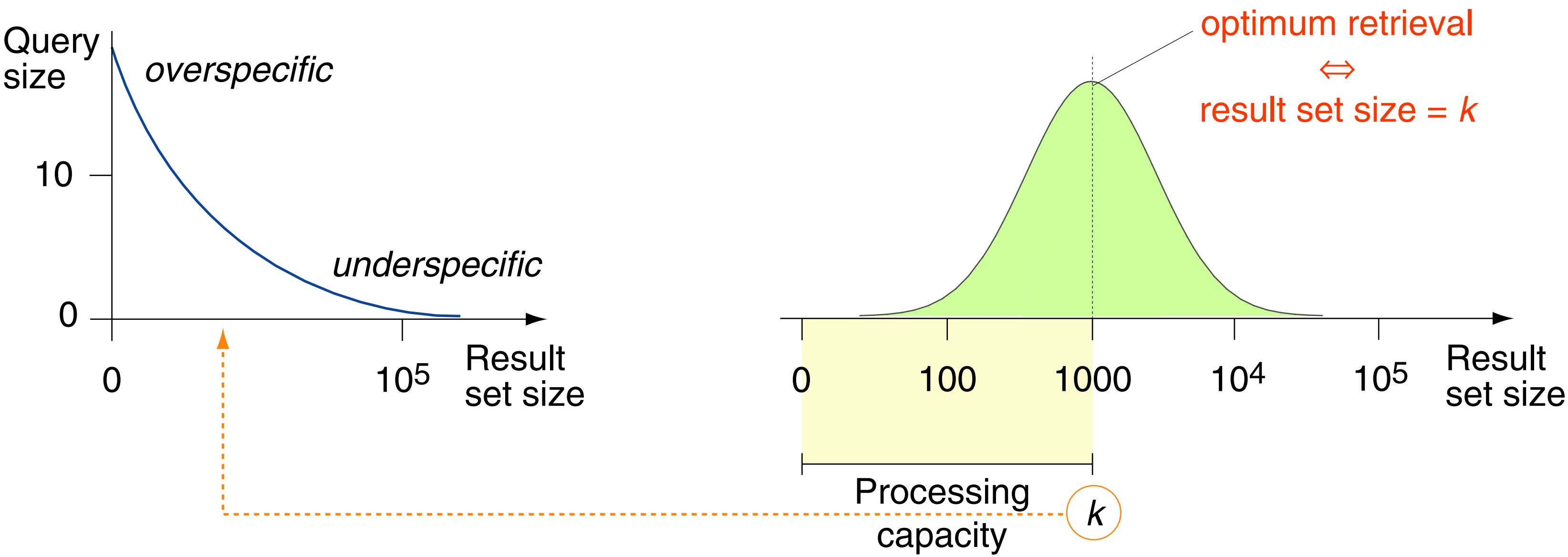
- User / keyword extractor has enough information to overspecify a search.
- Machine can spent a certain amount of time to analyze results.



Query Specificity

The User over Ranking Hypothesis

- User / keyword extractor has enough information to overspecify a search.
- Machine can spent a certain amount of time to analyze results.
- Rely on user / keyword extractor rather than on ranking algorithms: exploit processing capacity, considering “as many keywords as possible”.



Query Specificity — **Probability for Retrieval Success**

Obfuscation Technology

Rationale: emulate a plagiarist's text modification efforts.

Our task:

Given a section s_x, create a section s_q that has a high content similarity to s_x under some retrieval model but a different wording.

Obfuscation Technology

Rationale: emulate a plagiarist's text modification efforts.

Our task:

Given a section s_x, create a section s_q that has a high content similarity to s_x under some retrieval model but a different wording.

Obfuscation strategies:

1. random text operations
2. semantic word variation
3. POS-preserving word shuffling

Perfect obfuscation:

$s_x =$ "The quick brown fox jumps over the lazy dog."

- $s_q^* =$ "Over the dog, which is lazy, quickly jumps the fox which is brown."
- $s_q^* =$ "Dogs are lazy which is why brown foxes quickly jump over them."
- $s_q^* =$ "A fast bay-colored vulpine hops over an idle canine."

Obfuscation Technology: Random Text Operations

s_q is created from s_x by shuffling, removing, inserting, or replacing words or short phrases at random.

$s_x =$ “The quick brown fox jumps over the lazy dog.”

Examples:

- $s_q =$ “over The. the quick lazy dog context jumps brown fox”
- $s_q =$ “over jumps quick brown fox The lazy. the”
- $s_q =$ “brown jumps the. quick dog The lazy fox over”

Obfuscation Technology: Semantic Word Variation

s_q is created from s_x by replacing each word by one of its synonyms, antonyms, hyponyms, or hypernyms, chosen at random.

$s_x =$ “The quick brown fox jumps over the lazy dog.”

Examples:

- $s_q =$ “The quick brown dodger leaps over the lazy canine.”
- $s_q =$ “The quick brown canine jumps over the lazy canine.”
- $s_q =$ “The quick brown vixen leaps over the lazy puppy.”

Obfuscation Technology: POS-preserving Word Shuffling

Given the part of speech sequence of s_x, s_q is created by shuffling words at random while retaining the original POS sequence.

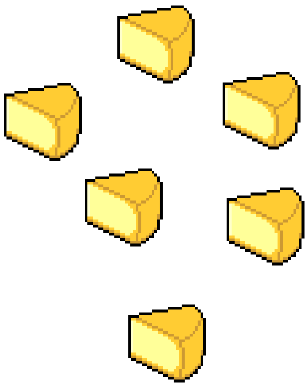
$s_x =$ "The quick brown fox jumps over the lazy dog."

POS $=$ "DT JJ JJ NN VBZ IN DT JJ NN ."

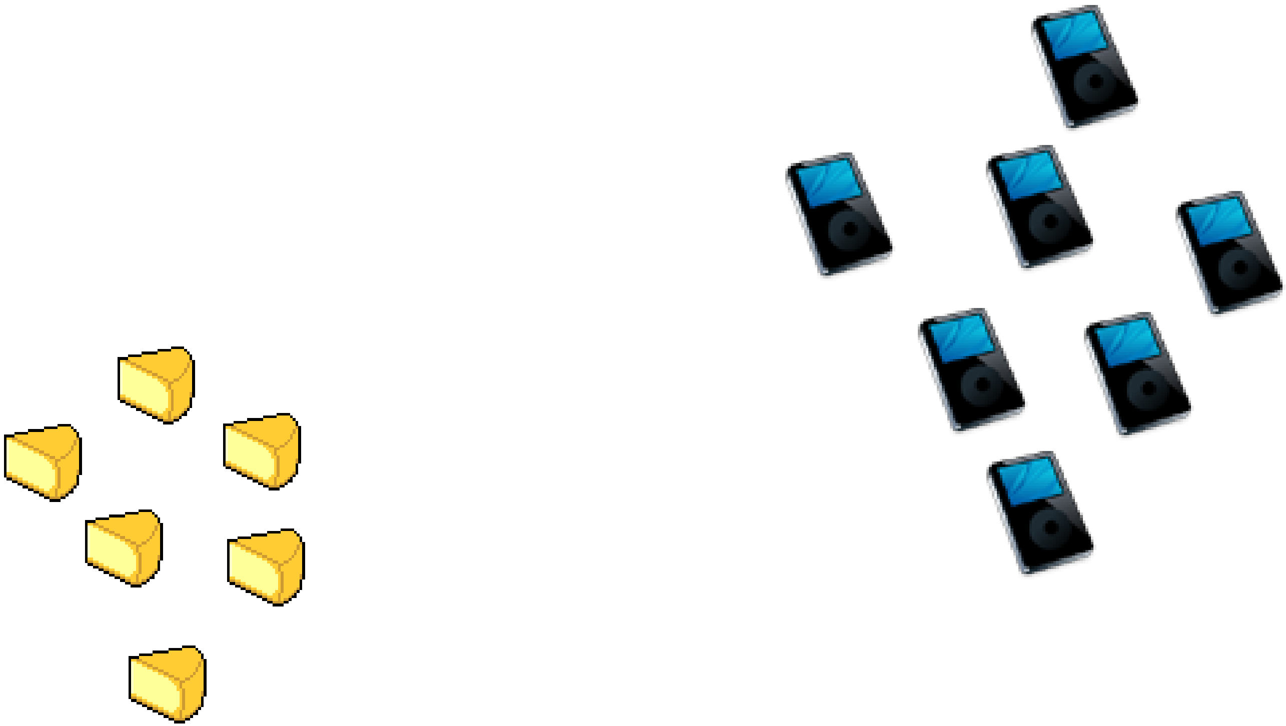
Examples:

- $s_q =$ "The brown lazy fox jumps over the quick dog."
- $s_q =$ "The lazy quick dog jumps over the brown fox."
- $s_q =$ "The brown lazy dog jumps over the quick fox."

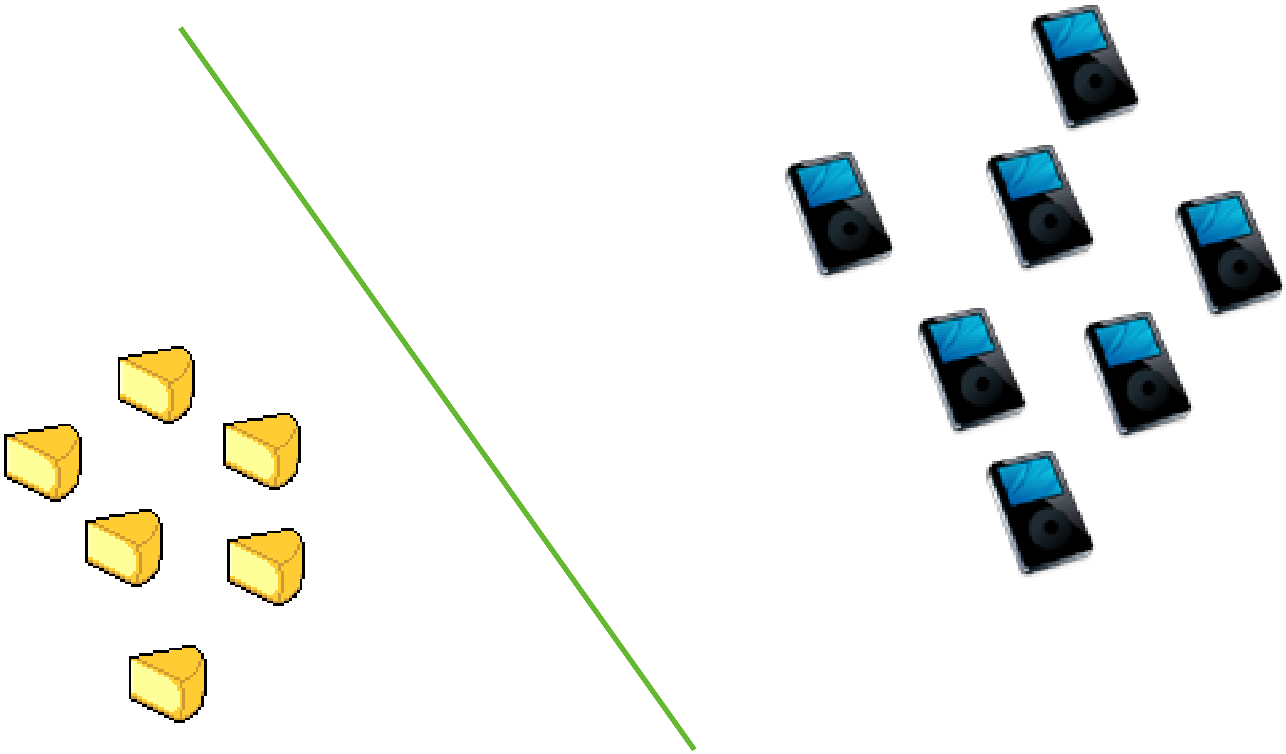
One Class Classification



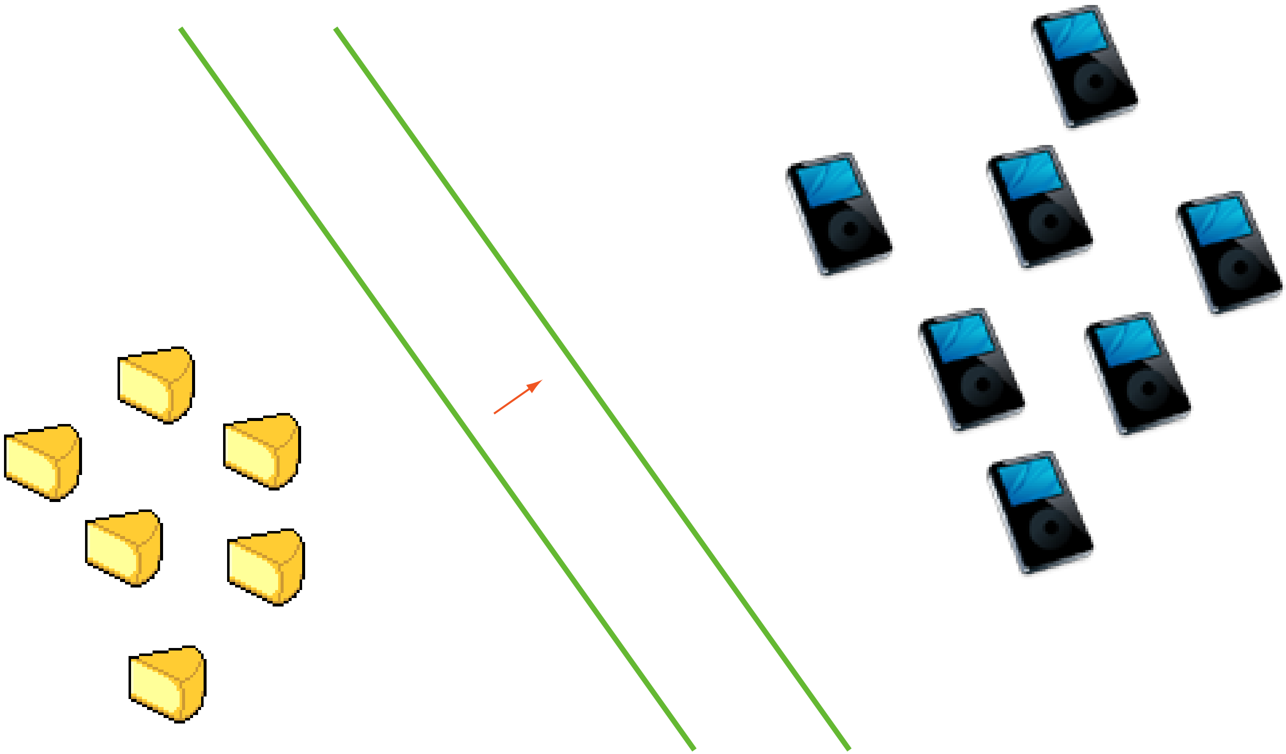
One Class Classification



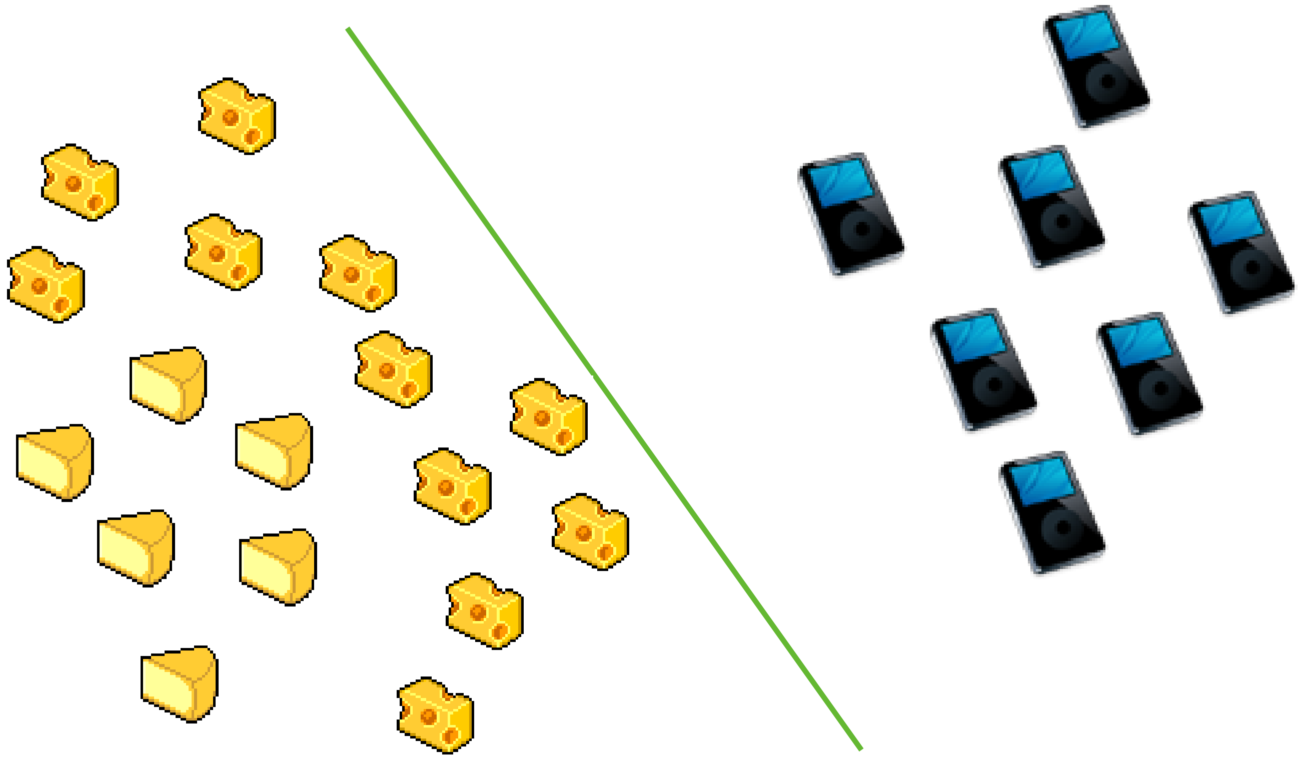
One Class Classification



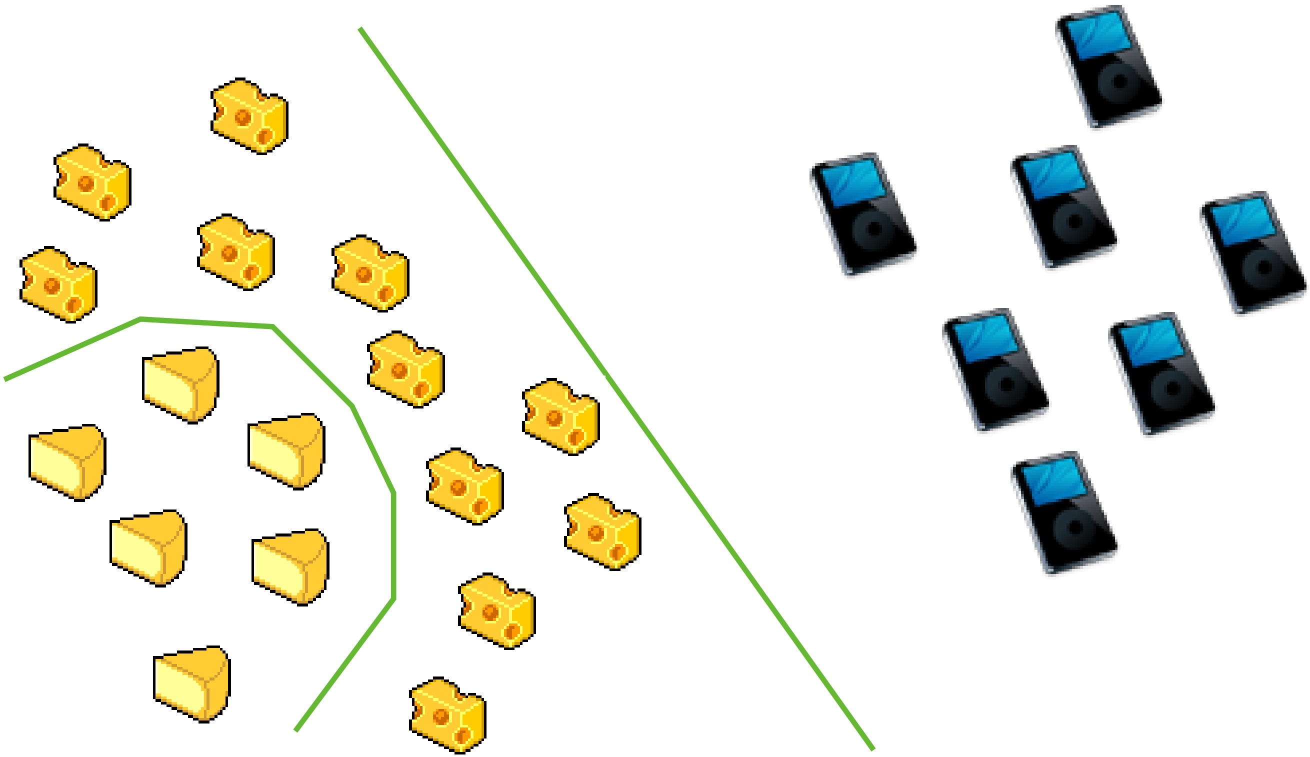
One Class Classification



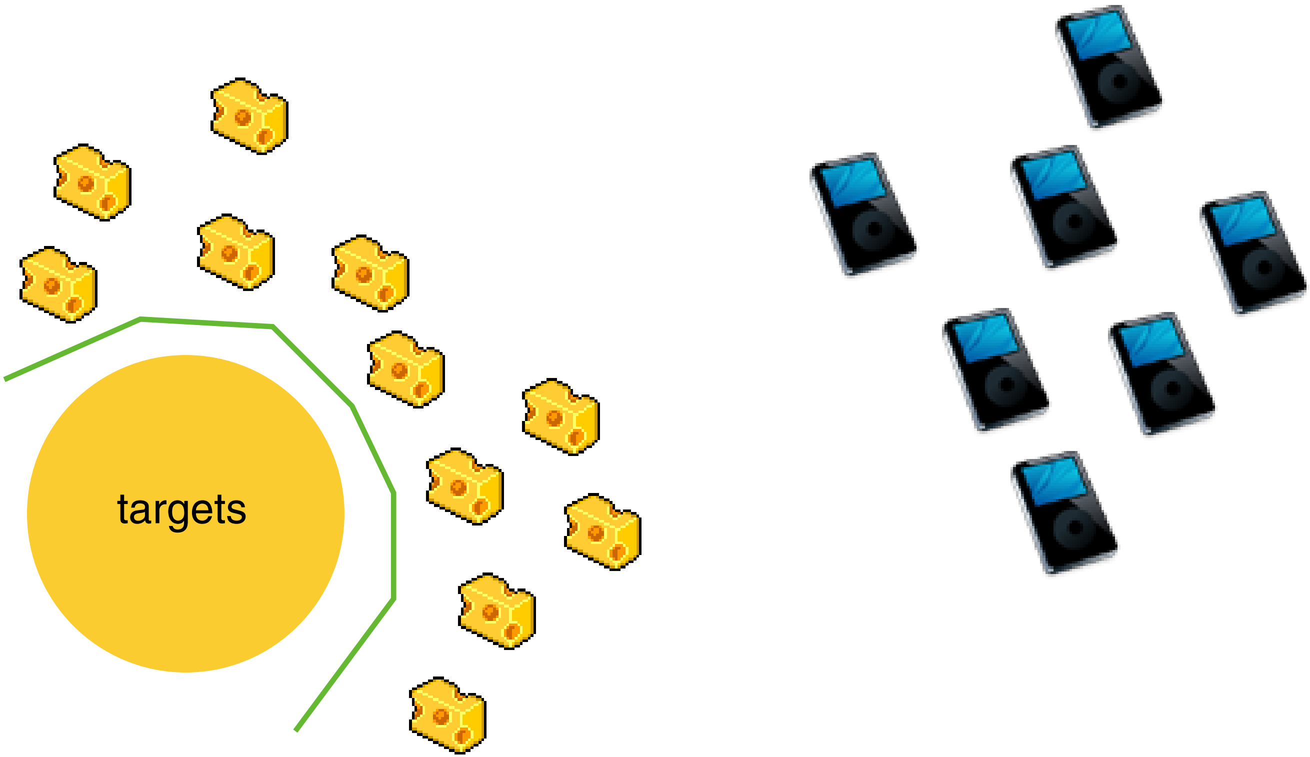
One Class Classification



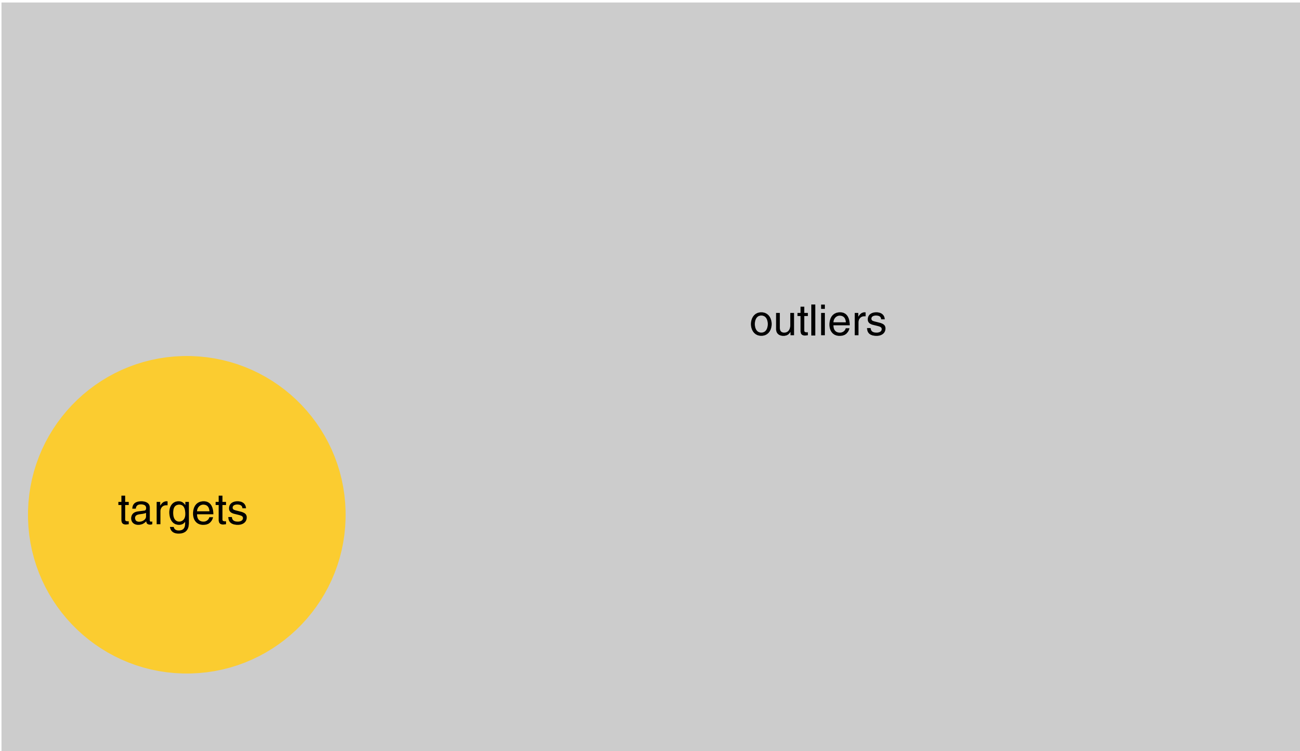
One Class Classification



One Class Classification



One Class Classification



One Class Classification