Tackling Societal Challenges with Style

Analysis

Martin Potthast
Leipzig University
www.temir.org

joint work with the
Webis Group
www.webis.de

Challenges

Search

Question queries

Axiomatic re-ranking

Argument search

Assessment

Clickbait

Text quality

Fake News and Hyperpartisanship

Offensive language

Detection

Vandalism

Plagiarism

Authorship

Synthesis

Summarization

Paraphrasing

Obfuscation

Web as Corpus

Mnemonic passwords

Challenges

Web as Corpus

Mnemonic passwords

Synthesis

Summarization

Paraphrasing

Obfuscation

Search

Question queries

Axiomatic re-ranking

Argument search

Assessment

Clickbait

Text quality

Fake News and Hyperpartisanship

Offensive language

Detection

Vandalism

Plagiarism

Authorship

Challenges

Search

Question queries
Axiomatic re-ranking
Argument search

Assessment

Clickbait
Text quality
Fake News and Hyperpartisanship
Offensive language

Web as Corpus

Mnemonic passwords

Synthesis

Summarization
Paraphrasing
Obfuscation

Detection

Vandalism
Plagiarism
Authorship

Challenges

Search

Question queries
Axiomatic re-ranking
Argument search

Assessment

Clickbait
Text quality
Fake News and Hyperpartisanship
Offensive language

Web as Corpus

Mnemonic passwords

Synthesis

Summarization
Paraphrasing
Obfuscation

Detection

Vandalism
Plagiarism
Authorship

①

Challenges

Web as Corpus

Mnemonic passwords

Synthesis

Summarization

Paraphrasing

Obfuscation

Search

Question queries

Axiomatic re-ranking

Argument search

Assessment

Clickbait

Text quality

Fake News and Hyperpartisanship

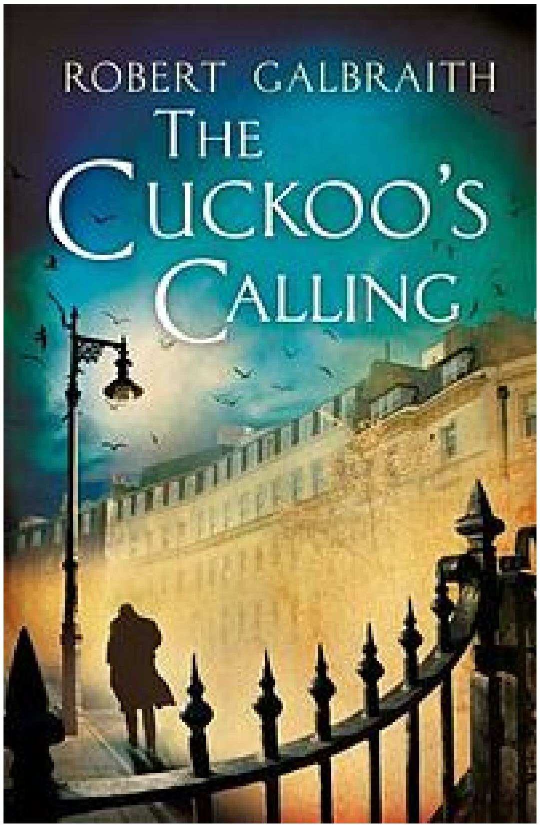
Offensive language

Detection

Vandalism

Plagiarism

Authorship



Fake likes

Fake news

Fake clicks

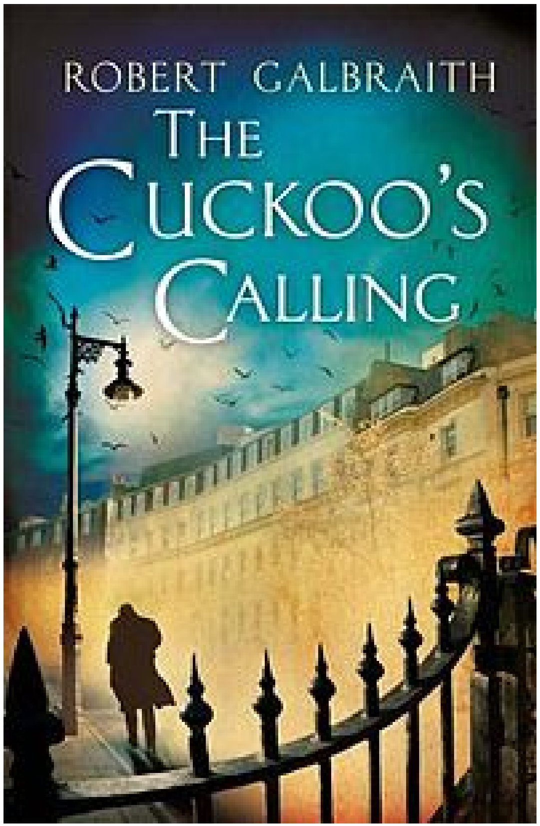
Fake users

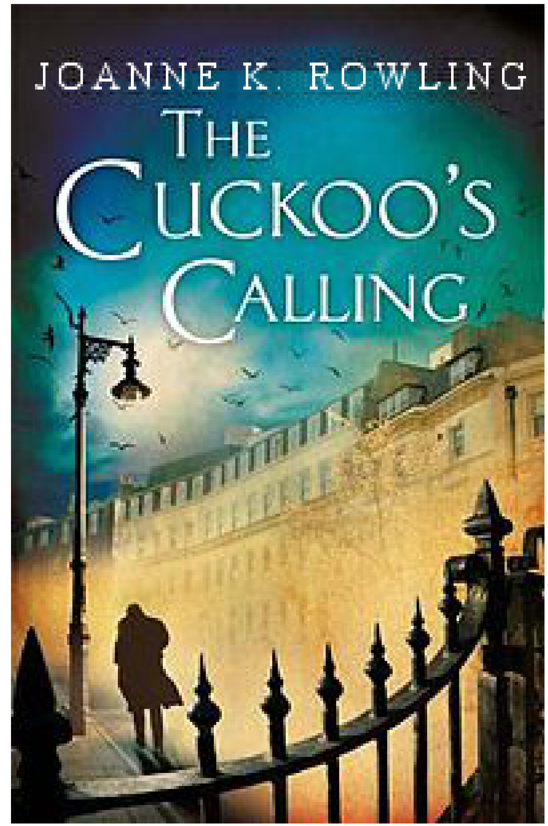
Fake reviews

Fake comments

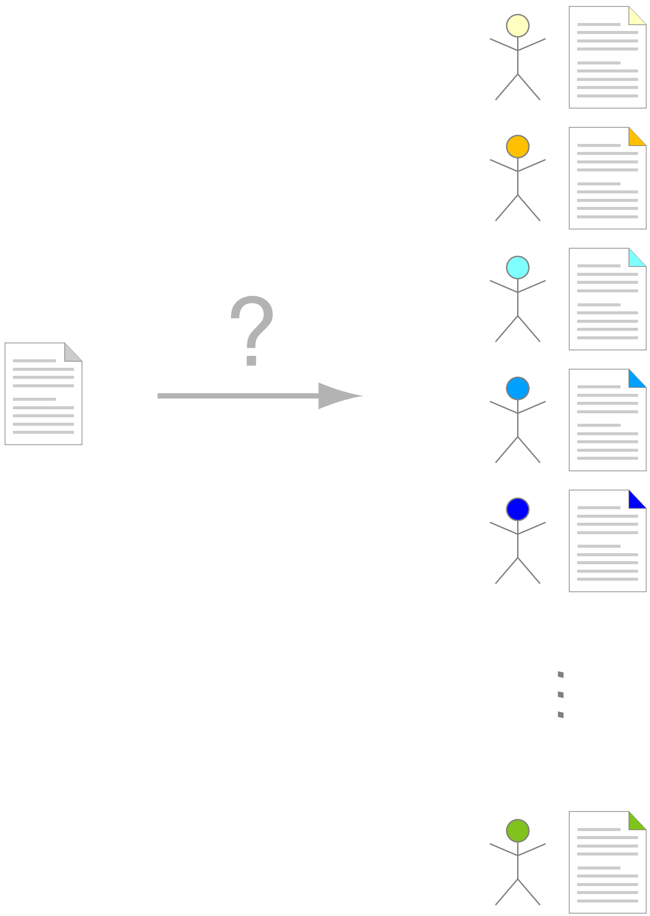
⋮

Fake identities (pseudonyms)


ROBERT GALBRAITH
THE CUCKOO'S CALLING


JOANNE K. ROWLING
THE CUCKOO'S CALLING

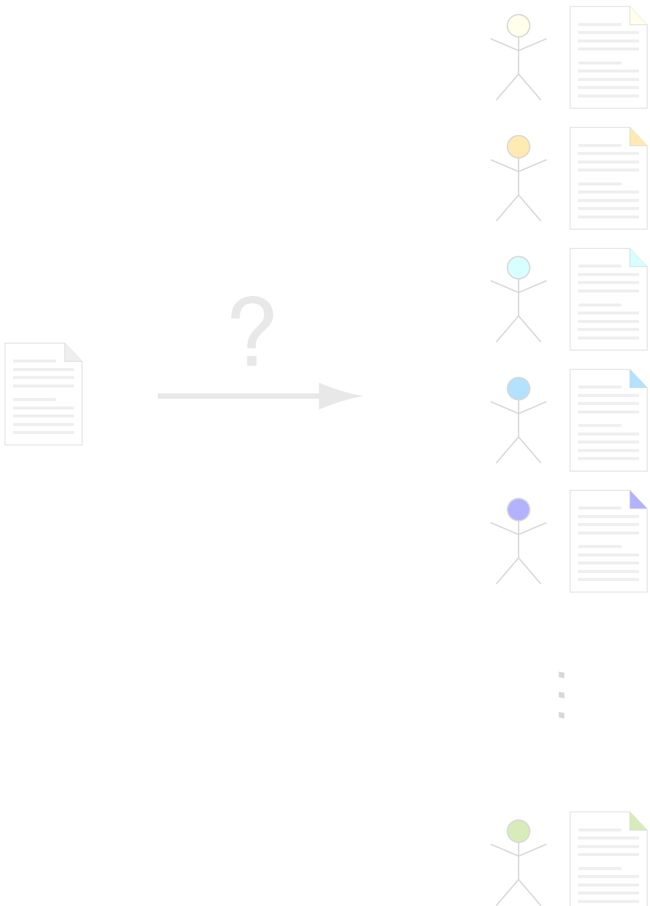
Authorship Attribution

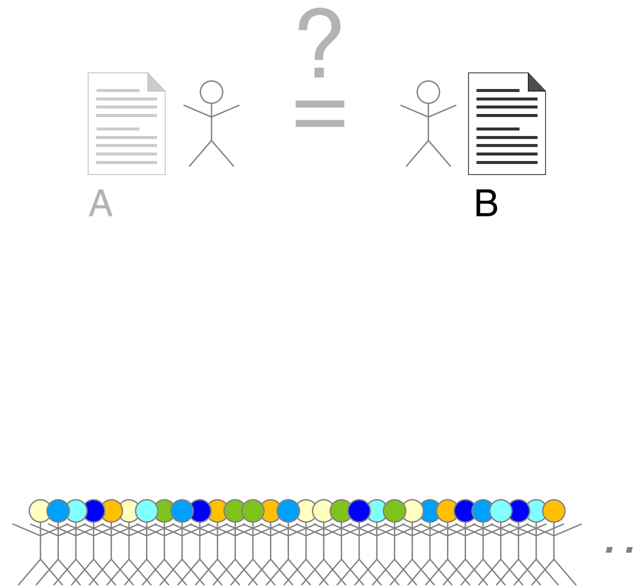


To which author does a text belong?

Authorship Attribution

Authorship Verification

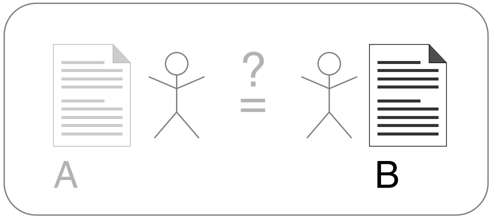




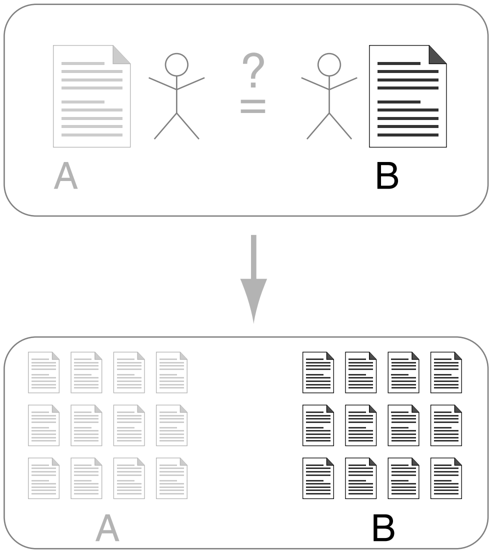
To which author does a text belong?

Originate two texts from the same author?

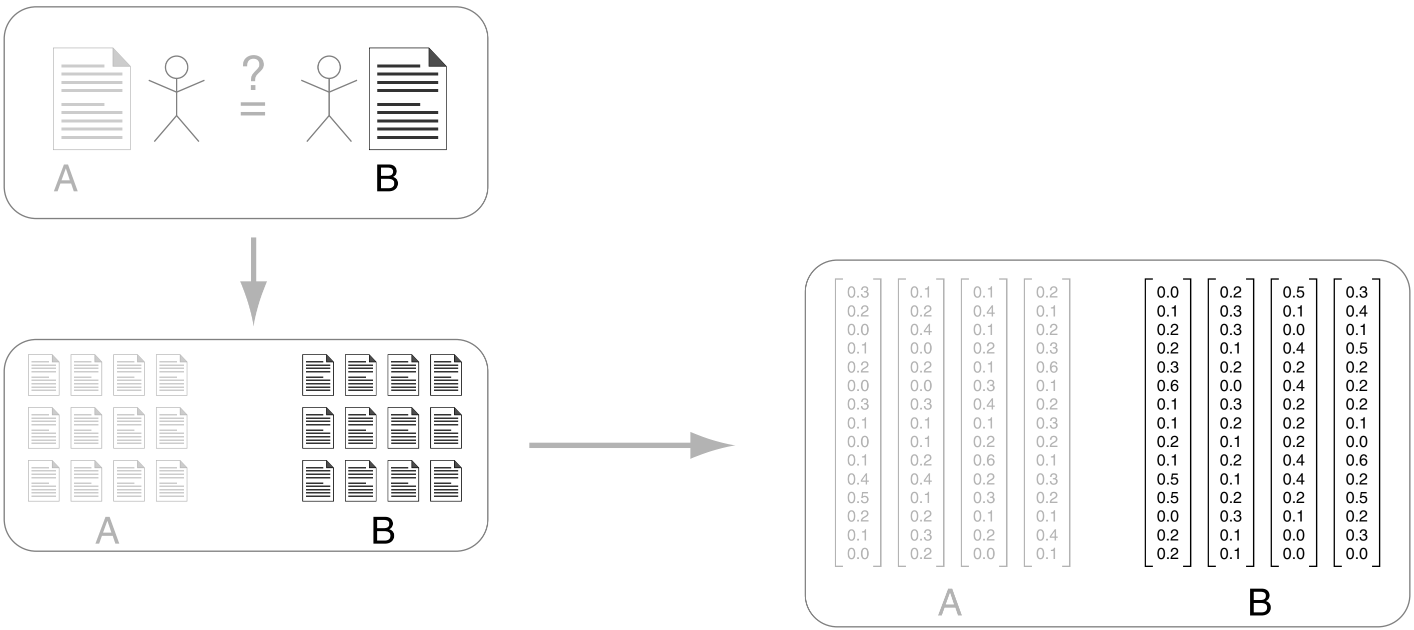
Authorship Verification “Unmasking” [Koppel/Schler 2004]



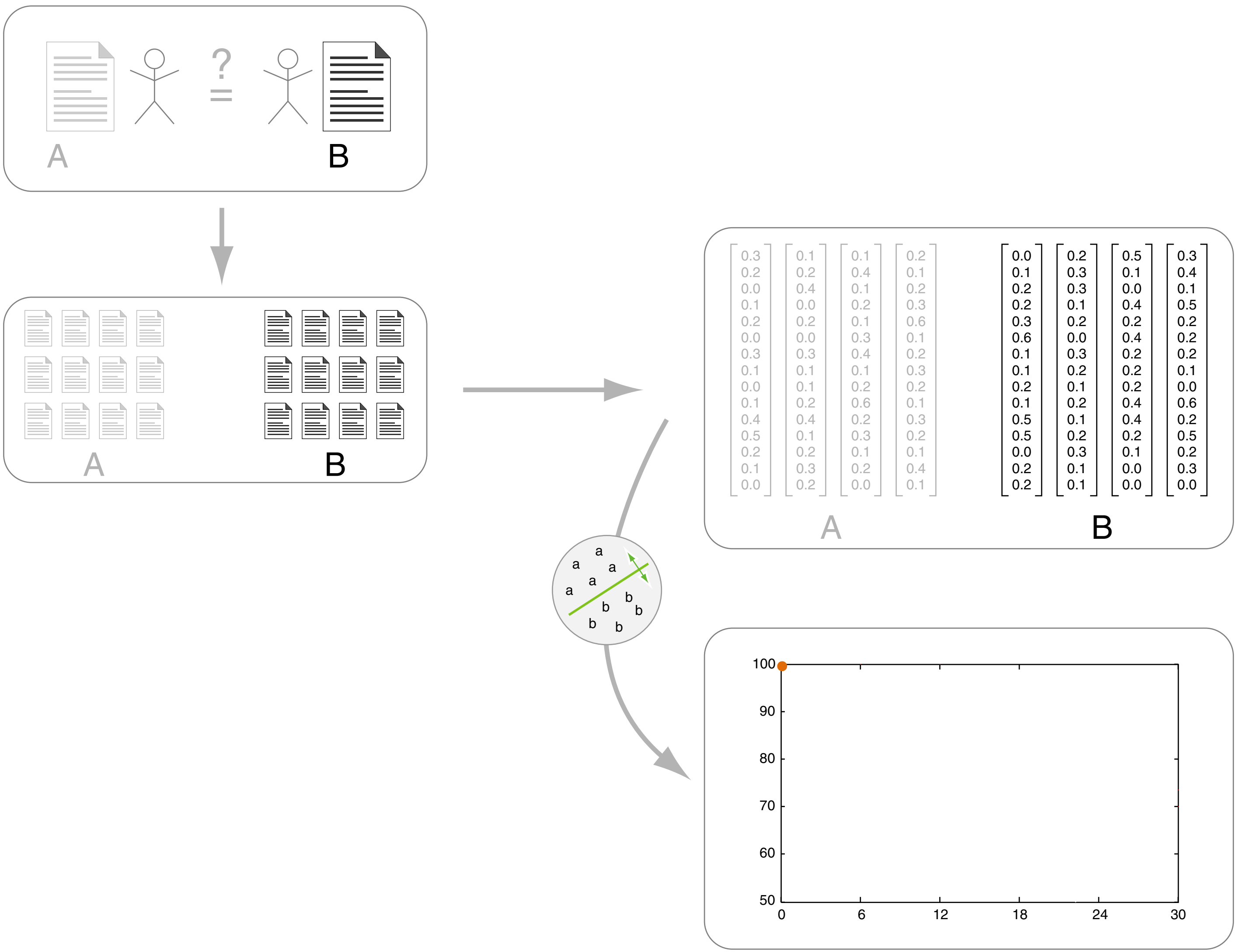
Authorship Verification "Unmasking" [Koppel/Schler 2004]



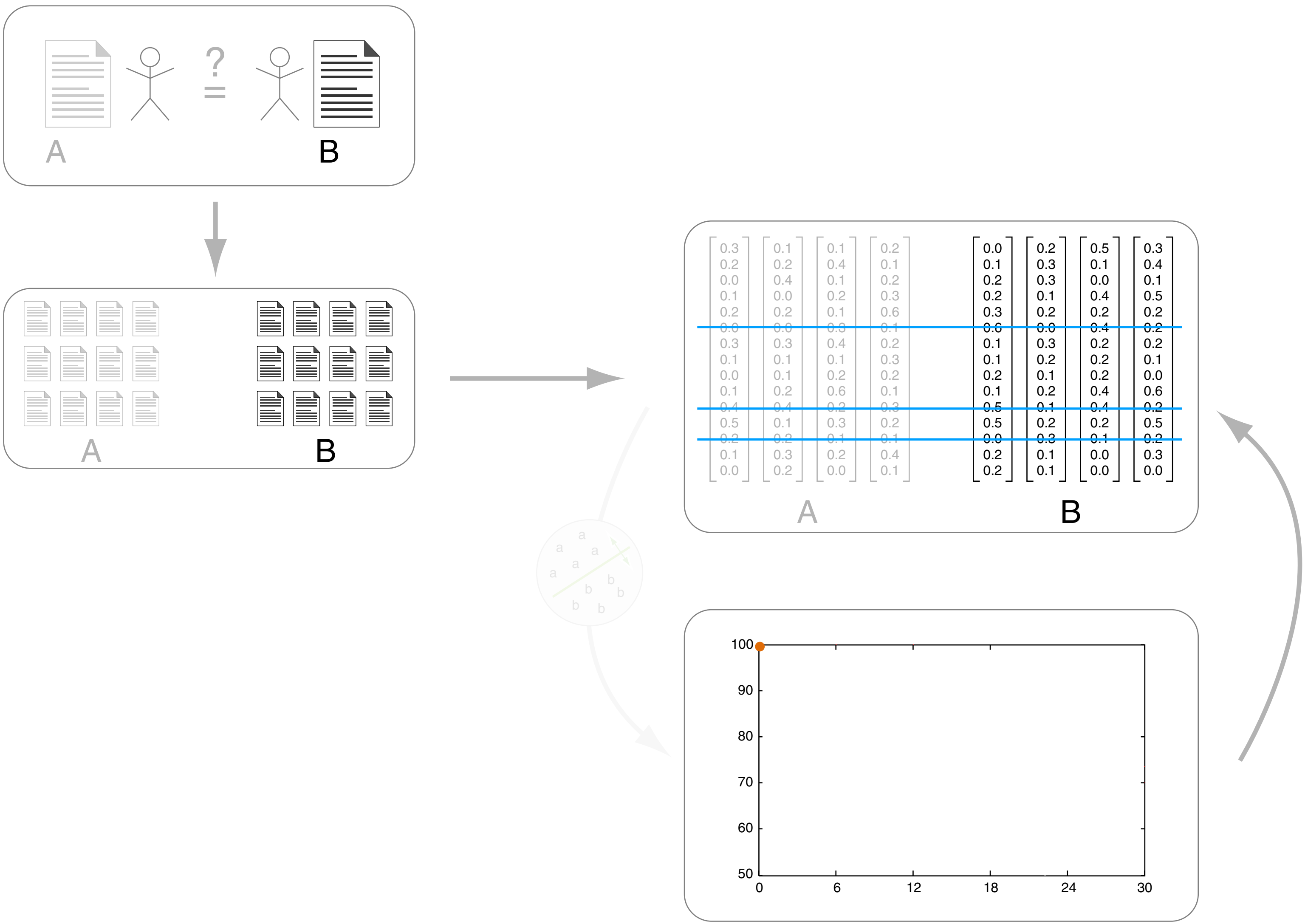
Authorship Verification “Unmasking” [Koppel/Schler 2004]



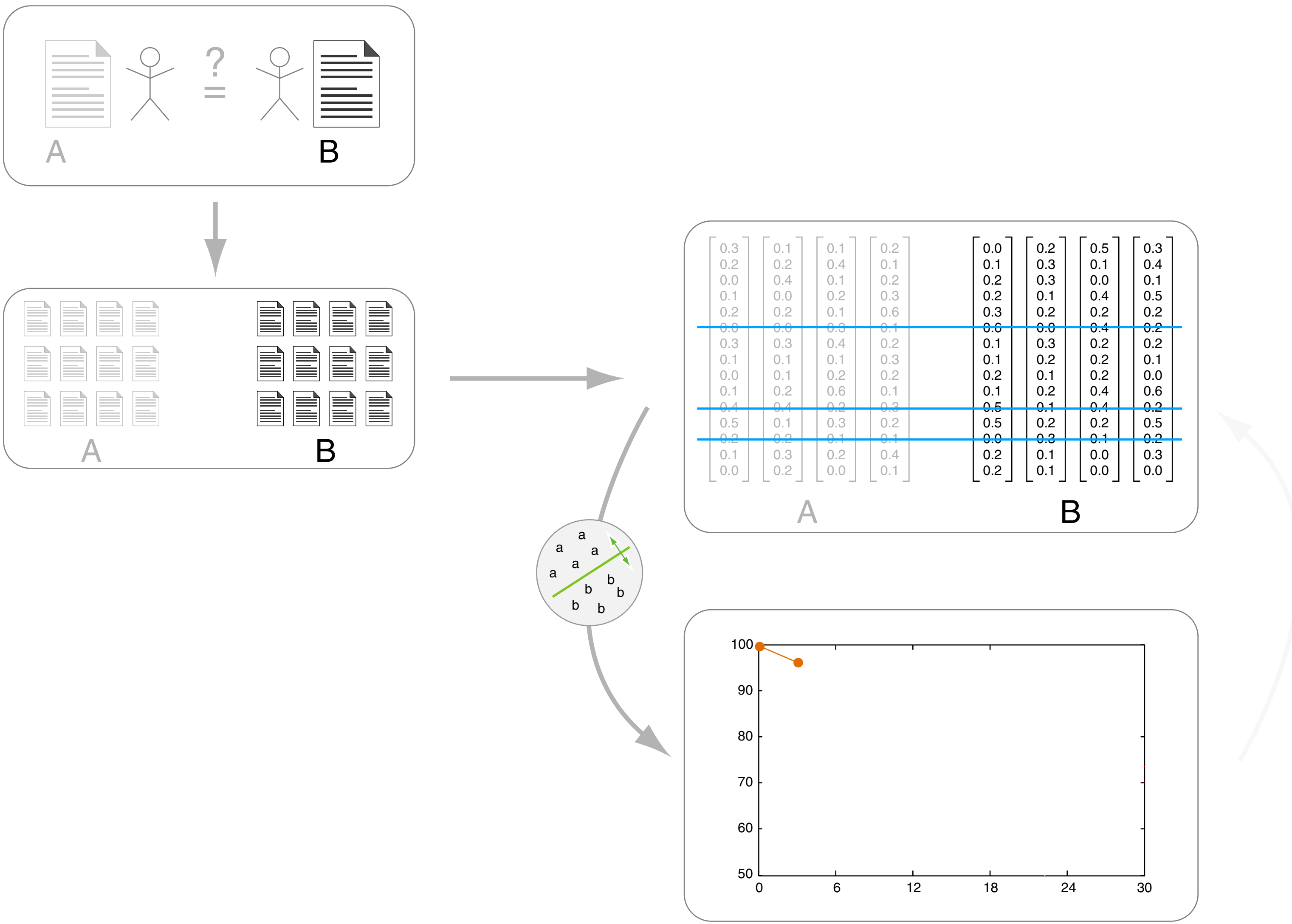
Authorship Verification “Unmasking” [Koppel/Schler 2004]



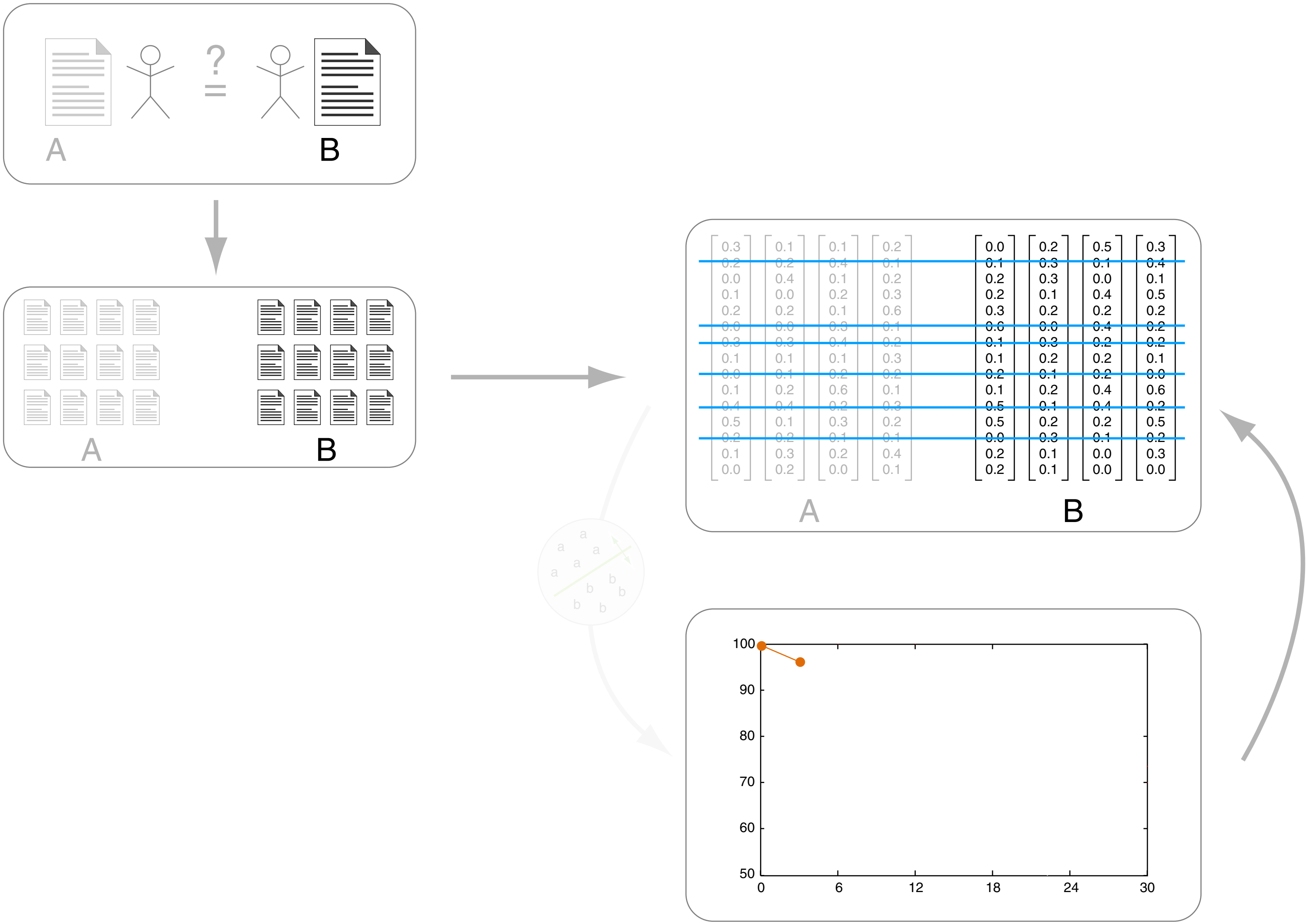
Authorship Verification “Unmasking” [Koppel/Schler 2004]



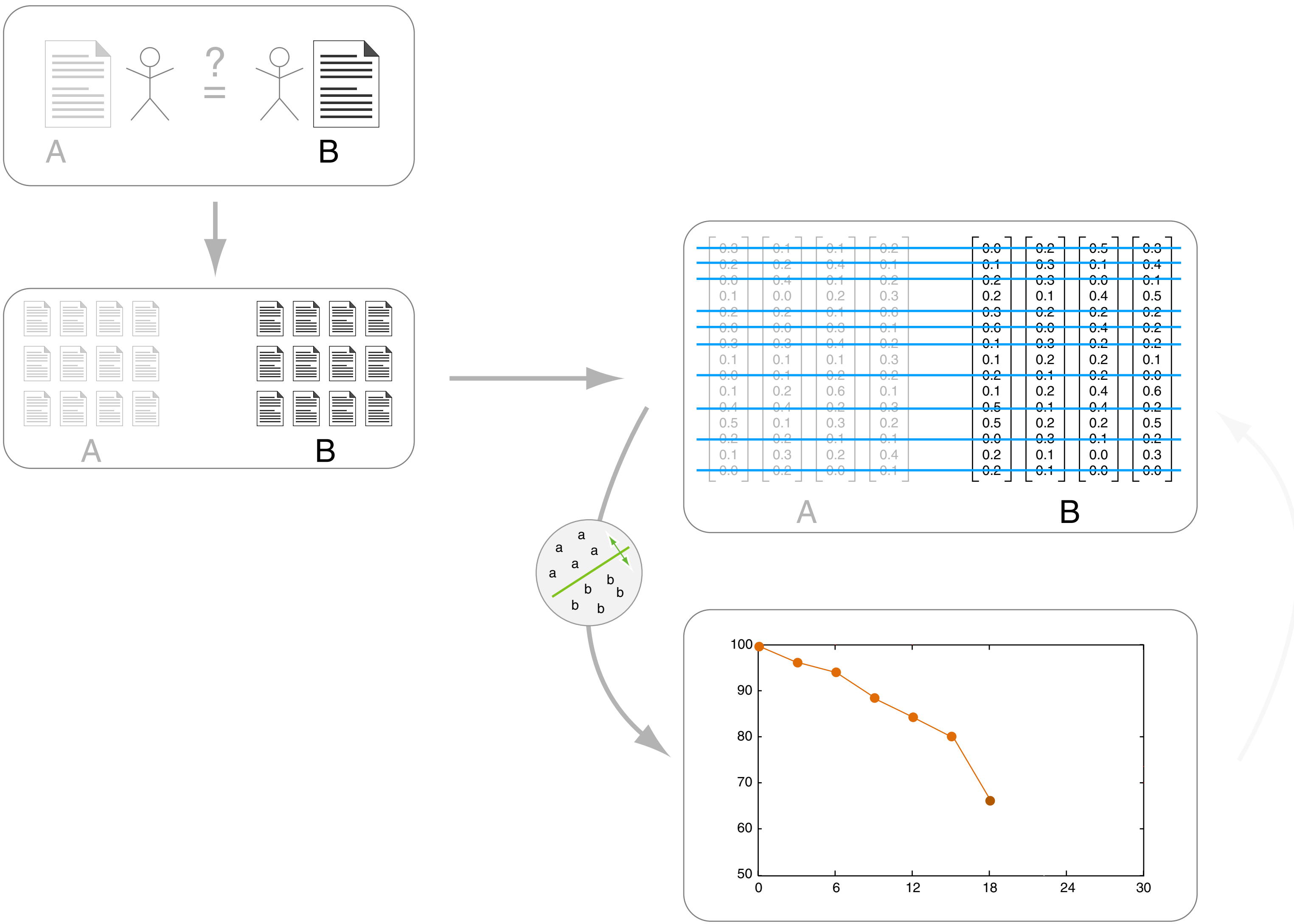
Authorship Verification “Unmasking” [Koppel/Schler 2004]



Authorship Verification “Unmasking” [Koppel/Schler 2004]

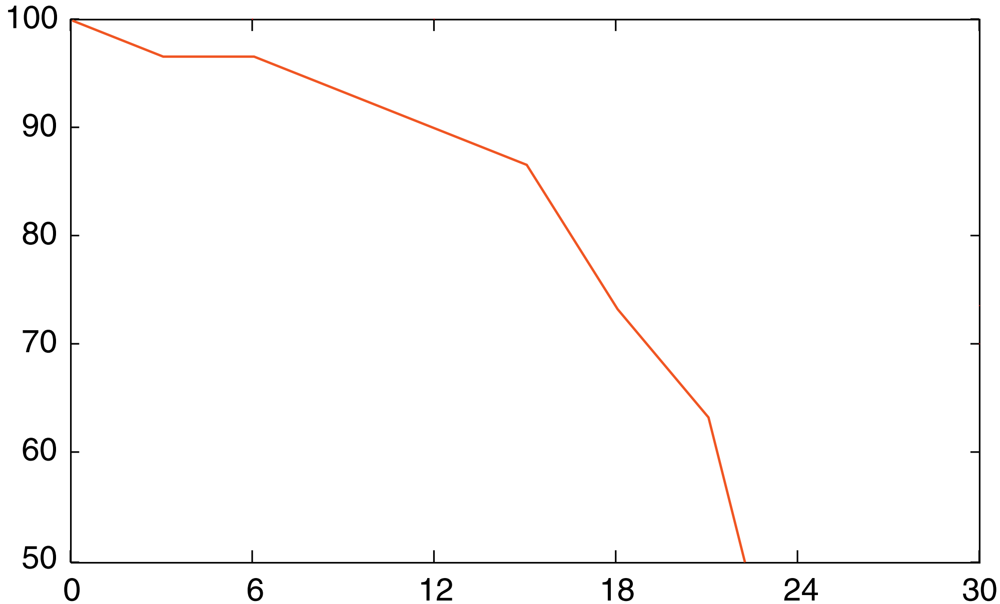


Authorship Verification “Unmasking” [Koppel/Schler 2004]



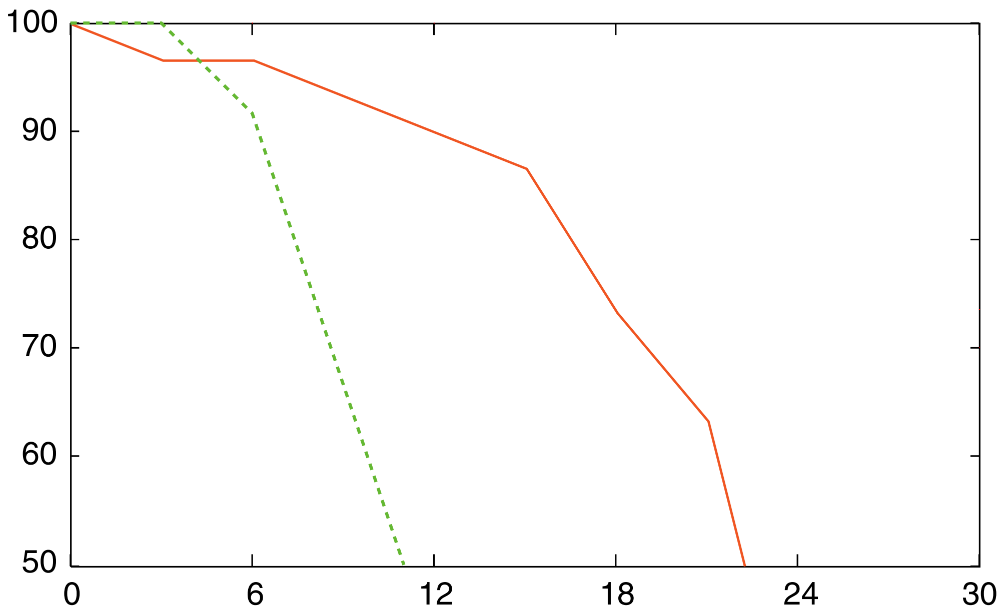
Authorship Verification Unmasking at Work

Typical learning characteristic for . . .



different authors $(A \neq B)$

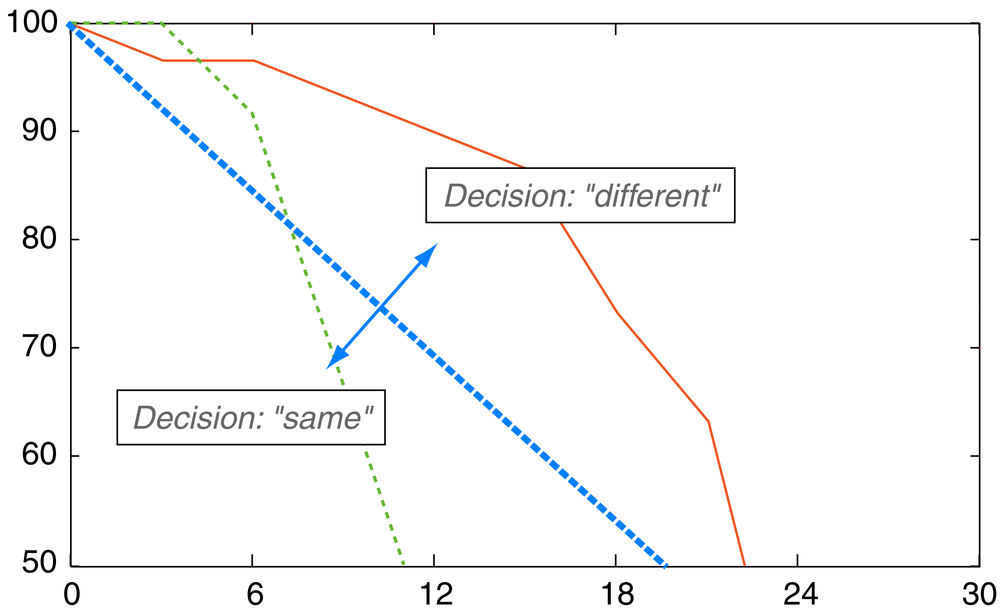
Typical learning characteristic for . . .



different authors $(A \neq B)$

same author $(A = B)$

Typical learning characteristic for . . .



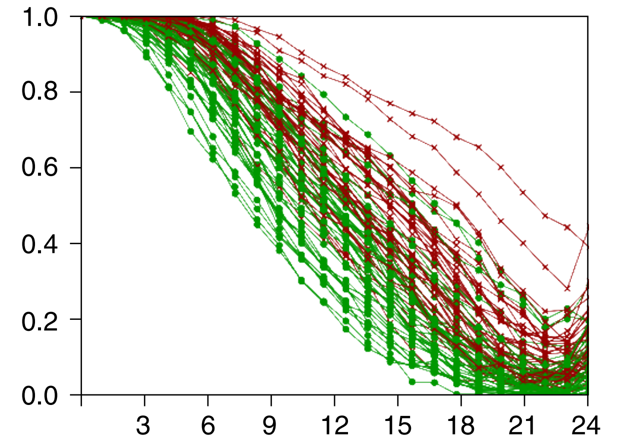
different authors $(A \neq B)$

same author $(A = B)$

The typical learning characteristic can be learned.

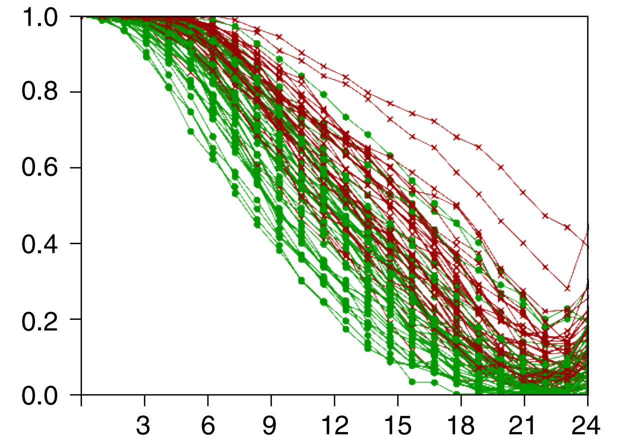
Authorship Verification Recent Results [Bevendorff et al., 2019]

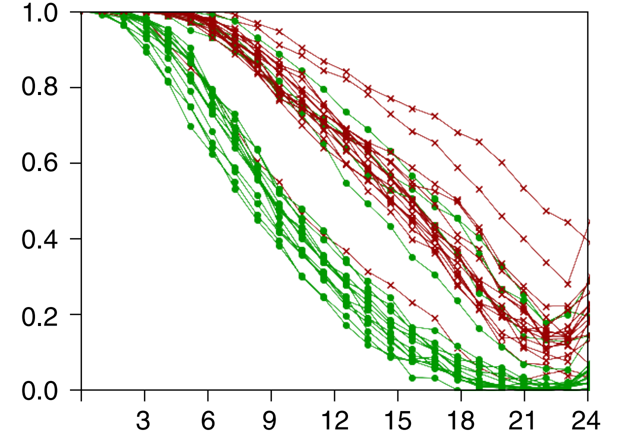
Experiment	**I**	**II**
Performance		
Precision	0.96	1.00
Accuracy	0.63	0.91
Classified	100%	26%
Omitted	0%	74%
Configuration		
Number of cases	180 training / 78 test	
Size of each case	4 000 words	
Number of authors	135	
Number of chunks	25	
Size of each chunk	600 words	
Vocabulary	250 words	
Removed per round	10 words	
Smoothing	no	



Authorship Verification Recent Results [Bevendorff et al., 2019]

Experiment	I	II
Performance		
Precision	0.96	1.00
Accuracy	0.63	0.91
Classified	100%	26%
Omitted	0%	74%
Configuration		
Number of cases	180 training / 78 test	
Size of each case	4 000 words	
Number of authors	135	
Number of chunks	25	
Size of each chunk	600 words	
Vocabulary	250 words	
Removed per round	10 words	
Smoothing	no	





②

Challenges

Web as Corpus

Mnemonic passwords

Synthesis

Summarization

Paraphrasing

Obfuscation

Search

Question queries

Axiomatic re-ranking

Argument search

Assessment

Clickbait

Text quality

Fake News and Hyperpartisanship

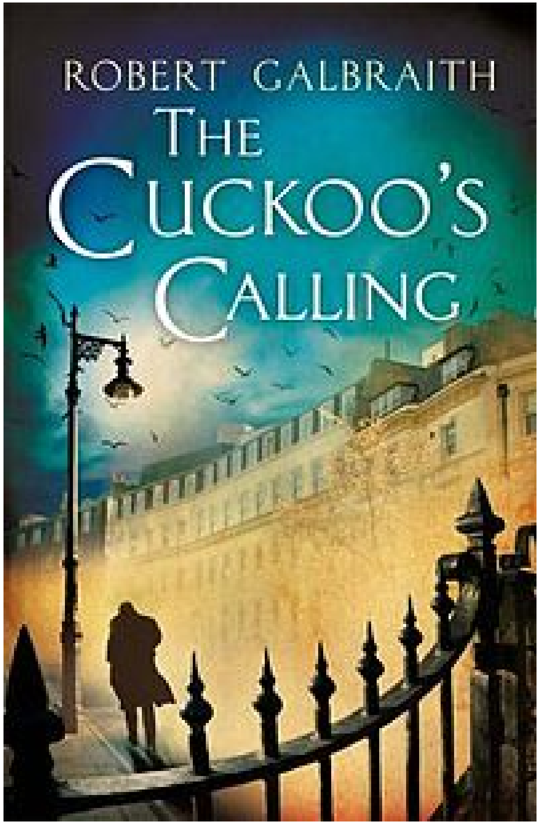
Offensive language

Detection

Vandalism

Plagiarism

Authorship



Fake likes

Fake news

Fake clicks

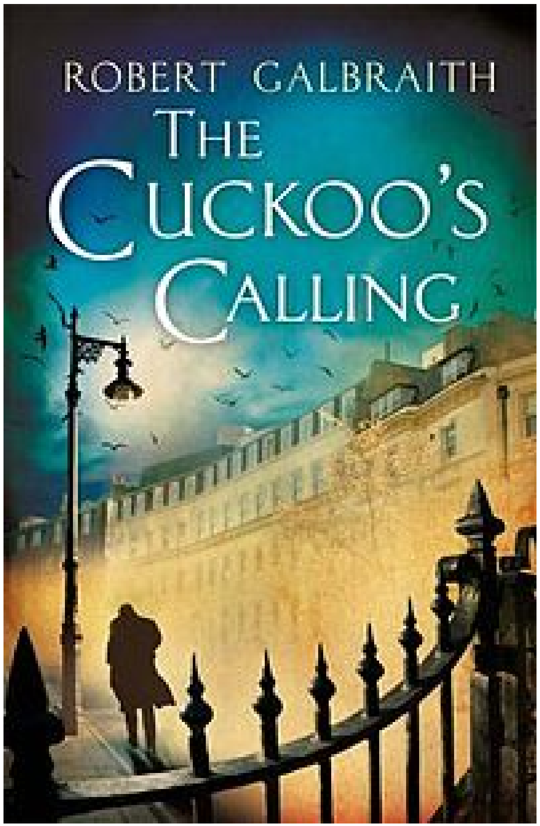
Fake users

Fake reviews

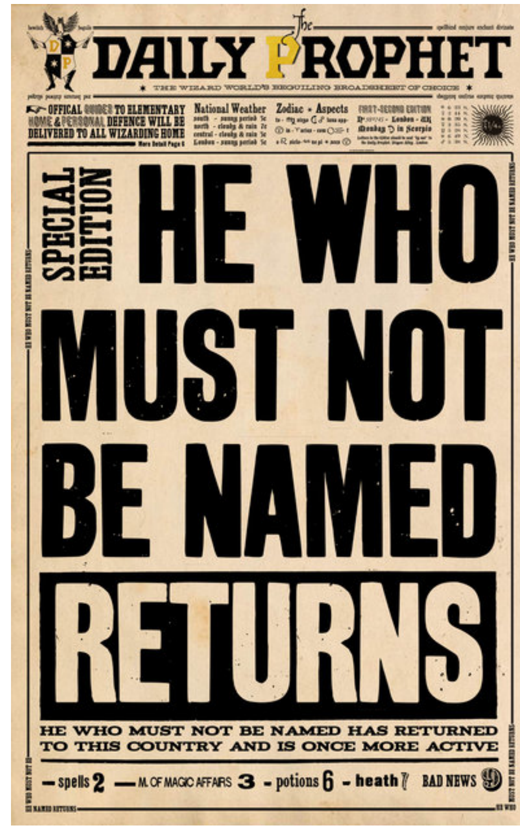
Fake comments

⋮

Fake identities (pseudonyms)


ROBERT GALBRAITH
THE CUCKOO'S CALLING

→


The DAILY PROPHET
THE WIZARD WORLD'S BEGUILING BROADSHEET OF CHOICE
OFFICAL GUIDES TO ELEMENTARY HOME & PERSONAL DEFENCE WILL BE DELIVERED TO ALL WIZARDING HOME
National Weather
Zodiac - Aspects
SPECIAL EDITION
HE WHO MUST NOT BE NAMED RETURNS
HE WHO MUST NOT BE NAMED HAS RETURNED TO THIS COUNTRY AND IS ONCE MORE ACTIVE
spells 2
M. OF MAGIC AFFAIRS 3
potions 6
heath 7
BAD NEWS 9

Countermeasure: Obfuscation [Bevendorff et al., 2019]

A

beautiful␣christmas you know jesus our saviour w
patiently stooping to hunger and pain, so he mig
ones,␣from shame; now if we love him, he bids us
brothers and sisters who need. blessed old nick! i
it, you would remember and certainly do it; this
you empty your pack, pray give a portion to all wh
there's anything left and you can bring a small gi
wasn't that dandy? sure, little mary␣ann has a wo
she has! she␣takes after her own mother. i was jus
that age. and you're just like␣her still, mollie mullig

B

sure, little mary␣ann has a wonderful education, s
s after her own mother. i was just like her when i wa
e just like␣her still, mollie mulligan. sure you're␣th
an alley and the belle of shantytown. whist now! it
lushes. but, hush! i think the show is about to begi
o, samson symbolical! come and see slivers,␣clow
me and see zip, the foremost of freaks! come an
ister sheiks! eager equestriennes,␣each unexcelle
enagerie ever beheld, the giant, the fat girl, the lio
artists from far-off japan, audacious acrobats sh

Idea: Obfuscate by increasing the Kullback-Leibler Divergence (KLD).

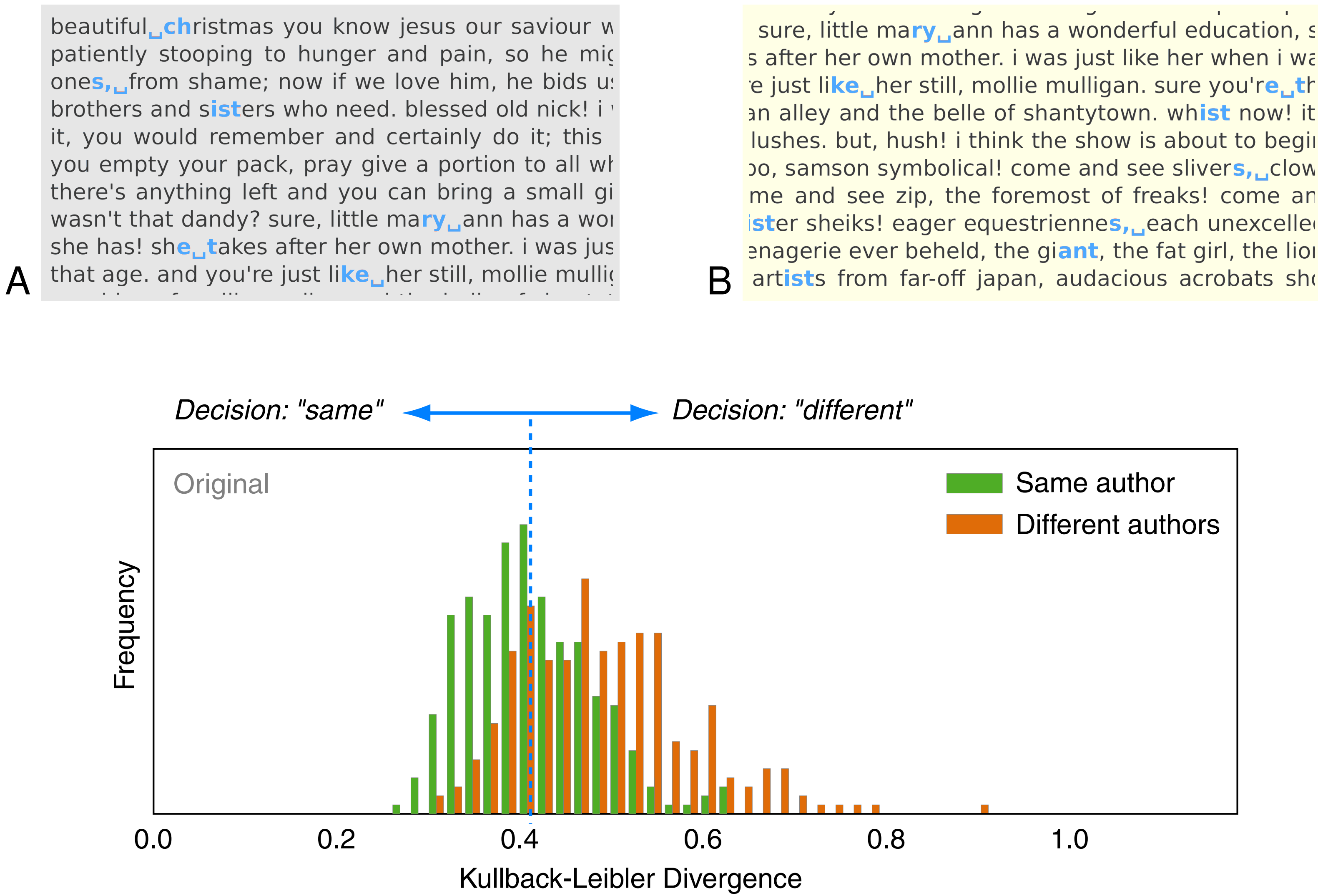
Desired: A “minimally invasive” procedure.

Strategy: Determine ”high-impact” n-grams. $\frac{\partial}{\partial q}\left(p \log_2 \frac{p}{q}\right) \to \max,$

where p and q denote the n-gram-specific occurrence probabilities in texts A and B respectively.

Countermeasure: Obfuscation

[Bevendorff et al., 2019]



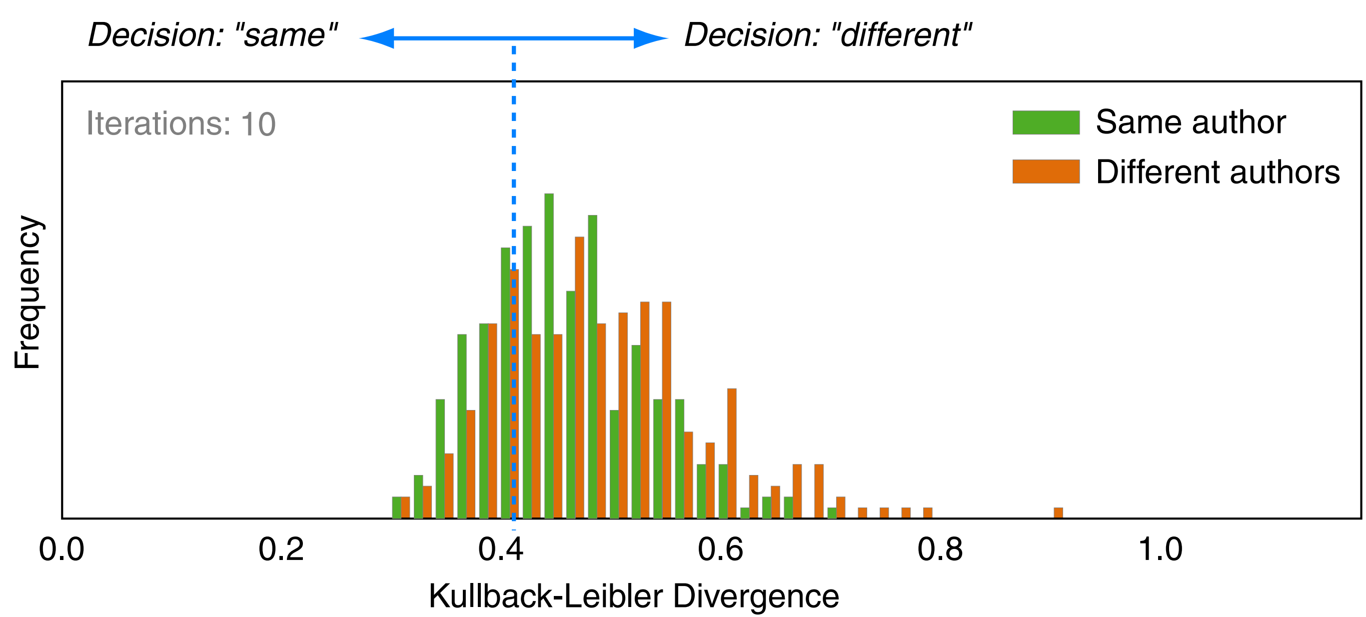
Countermeasure: Obfuscation [Bevendorff et al., 2019]

A

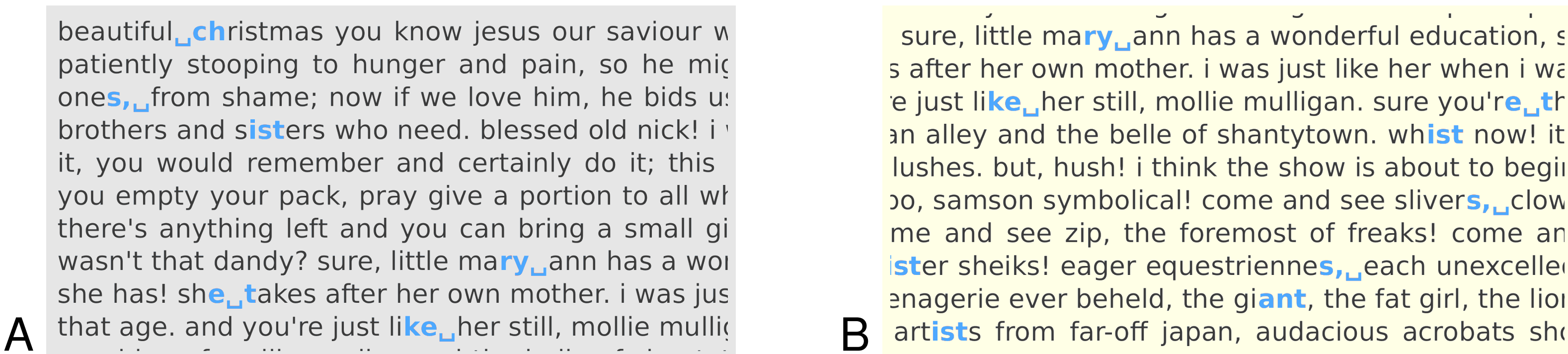
beautiful␣christmas you know jesus our saviour w
patiently stooping to hunger and pain, so he mig
ones,␣from shame; now if we love him, he bids u
brothers and sisters who need. blessed old nick! i
it, you would remember and certainly do it; this
you empty your pack, pray give a portion to all wh
there's anything left and you can bring a small gi
wasn't that dandy? sure, little mary␣ann has a wo
she has! she␣takes after her own mother. i was jus
that age. and you're just like␣her still, mollie mulli

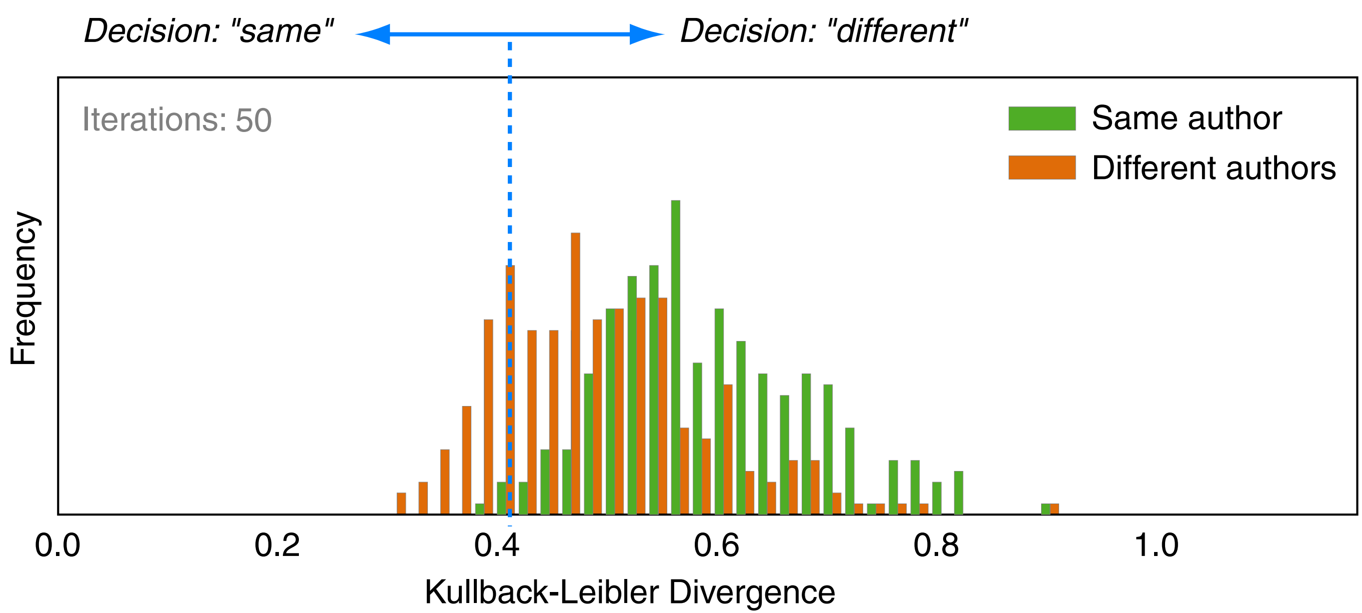
B

sure, little mary␣ann has a wonderful education, s
s after her own mother. i was just like her when i wa
e just like␣her still, mollie mulligan. sure you're␣th
an alley and the belle of shantytown. whist now! it
lushes. but, hush! i think the show is about to begi
o, samson symbolical! come and see slivers,␣clow
me and see zip, the foremost of freaks! come an
ister sheiks! eager equestriennes,␣each unexcelle
enagerie ever beheld, the giant, the fat girl, the lio
artists from far-off japan, audacious acrobats sh



Countermeasure: Obfuscation [Bevendorff et al., 2019]





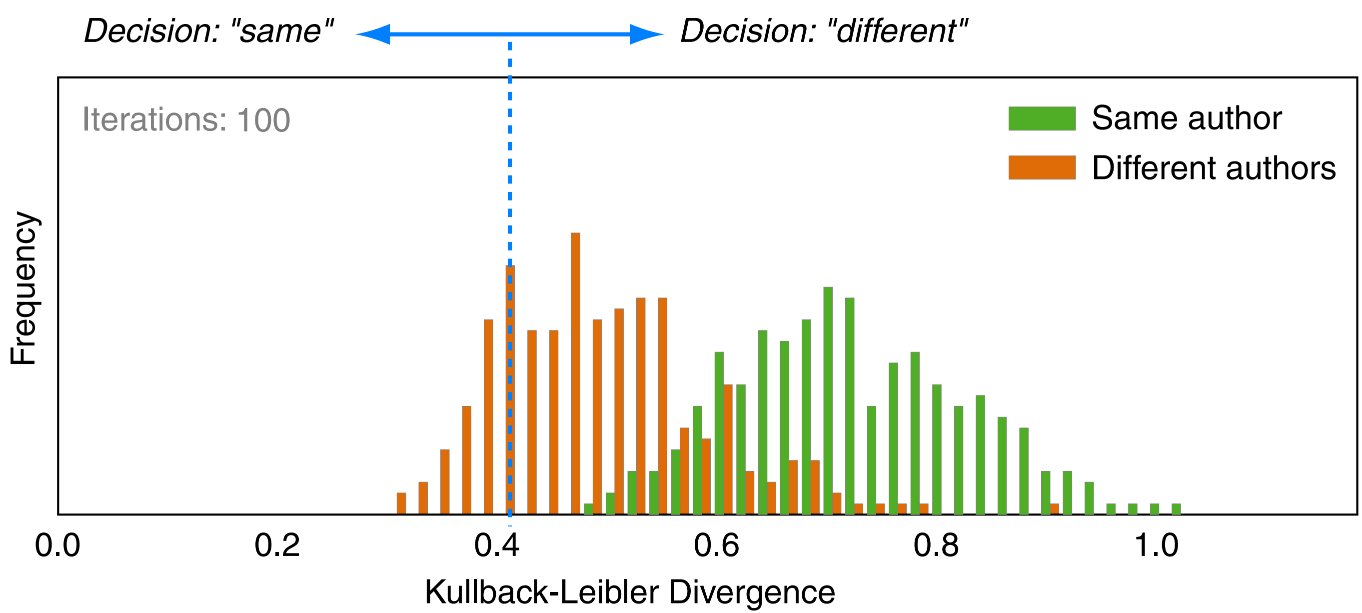
Countermeasure: Obfuscation [Bevendorff et al., 2019]

A

beautiful␣christmas you know jesus our saviour w
patiently stooping to hunger and pain, so he mig
ones,␣from shame; now if we love him, he bids u
brothers and sisters who need. blessed old nick! i
it, you would remember and certainly do it; this
you empty your pack, pray give a portion to all wh
there's anything left and you can bring a small gi
wasn't that dandy? sure, little mary␣ann has a wo
she has! she␣takes after her own mother. i was jus
that age. and you're just like␣her still, mollie mulli

B

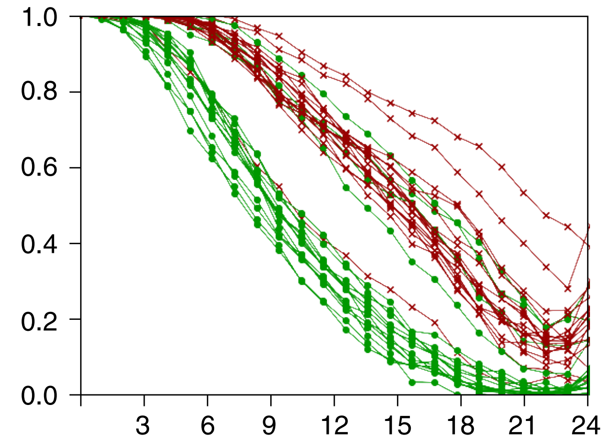
sure, little mary␣ann has a wonderful education, s
s after her own mother. i was just like her when i wa
e just like␣her still, mollie mulligan. sure you're␣th
an alley and the belle of shantytown. whist now! it
lushes. but, hush! i think the show is about to begi
oo, samson symbolical! come and see slivers,␣clow
me and see zip, the foremost of freaks! come an
ister sheiks! eager equestriennes,␣each unexcelle
enagerie ever beheld, the giant, the fat girl, the lio
artists from far-off japan, audacious acrobats sh



Countermeasure: Obfuscation

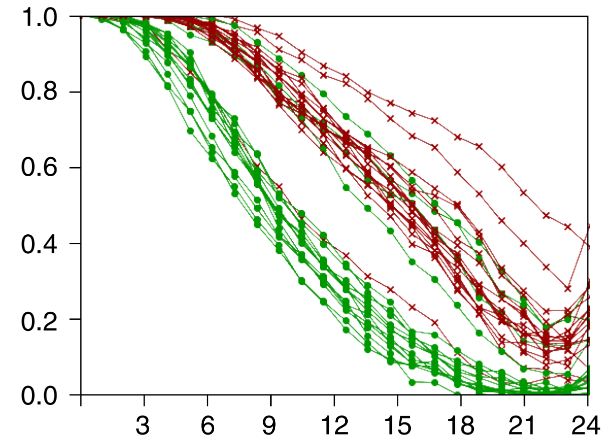
Recent Results

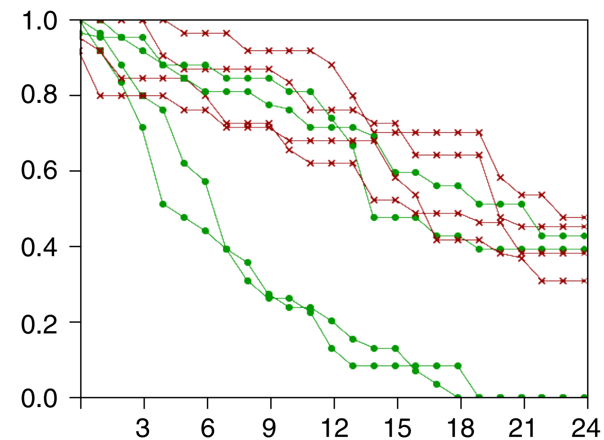
Experiment	I	II	III
Performance			
Precision	1.00	1.00	1.00
Accuracy	0.91	0.75	0.83
n-gram removals	0	80	200
Text coverage	0%	1%	2.6%
Classified	26%	10%	7%
Omitted	74%	90%	93%
Configuration			
Number of cases	180 training / 78 test		
Size of each case	4 000 words		
Number of authors	135		
Number of chunks	25		
Size of each chunk	600 words		
Vocabulary	250 words		
Removed per round	10 words		
Smoothing	no		



Countermeasure: Obfuscation Recent Results

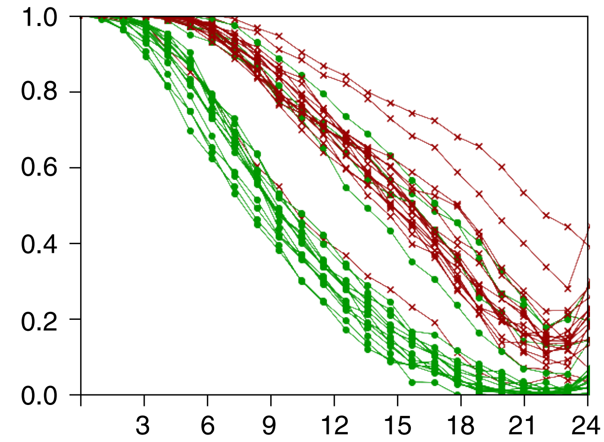
Experiment	**I**	**II**	**III**
Performance			
Precision	1.00	1.00	1.00
Accuracy	0.91	0.75	0.83
n-gram removals	0	80	200
Text coverage	0%	1%	2.6%
Classified	26%	10%	7%
Omitted	74%	90%	93%
Configuration			
Number of cases	180 training / 78 test		
Size of each case	4 000 words		
Number of authors	135		
Number of chunks	25		
Size of each chunk	600 words		
Vocabulary	250 words		
Removed per round	10 words		
Smoothing	no		

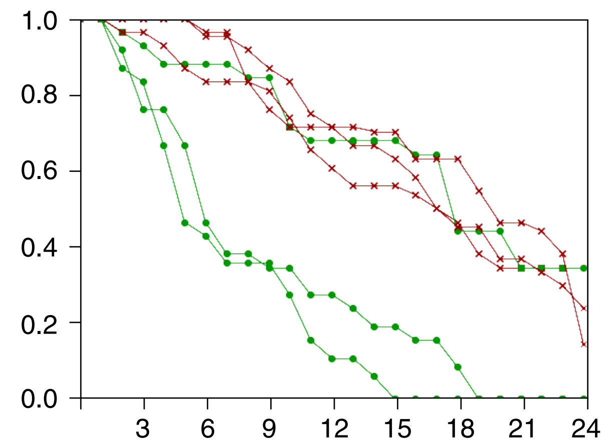




Countermeasure: Obfuscation
Recent Results

Experiment	**I**	**II**	**III**
Performance			
Precision	1.00	1.00	1.00
Accuracy	0.91	0.75	0.83
n-gram removals	0	80	200
Text coverage	0%	1%	2.6%
Classified	26%	10%	7%
Omitted	74%	90%	93%
Configuration			
Number of cases	180 training / 78 test		
Size of each case	4 000 words		
Number of authors	135		
Number of chunks	25		
Size of each chunk	600 words		
Vocabulary	250 words		
Removed per round	10 words		
Smoothing	no		





3

Challenges

Search

Question queries
Axiomatic re-ranking
Argument search

Assessment

Clickbait
Text quality
Fake News and Hyperpartisanship
Offensive language

Web as Corpus

Mnemonic passwords

Synthesis

Summarization
Paraphrasing
Obfuscation

Detection

Vandalism
Plagiarism
Authorship

To the Members of the California State Assembly:

I am returning Assembly Bill 1176 without my signature.

For some time now I have lamented the fact that major issues are overlooked while many unnecessary bills come to me for consideration. Water reform, prison reform, and health care are major issues my Administration has brought to the table, but the Legislature just kicks the can down the alley.

Yet another legislative year has come and gone without the major reforms Californians overwhelmingly deserve. In light of this, and after careful consideration, I believe it is unnecessary to sign this measure at this time.

Sincerely,

Arnold Schwarzenegger

[Veto message for the Shipyard project, Port of San Francisco. Oct. 12th, 2009]

To the Members of the California State Assembly:

I am returning Assembly Bill 1176 without my signature.

For some time now I have lamented the fact that major issues are overlooked while many
*****nnecessary bills come to me for consideration. Water reform, prison reform, and health
*****are are major issues my Administration has brought to the table, but the Legislature just
*****icks the can down the alley.

Yet another legislative year has come and gone without the major reforms Californians
*****verwhelmingly deserve. In light of this, and after careful consideration, I believe it is
*****nnecessary to sign this measure at this time.

Sincerely,

Arnold Schwarzenegger

[Veto message for the Shipyard project, Port of San Francisco. Oct. 12th, 2009]

To the Members of the California State Assembly:

I am returning Assembly Bill 1176 without my signature.

For some time now I have lamented the fact that major issues are overlooked while many
*nnecessary bills come to me for consideration. Water reform, prison reform, and health
*are are major issues my Administration has brought to the table, but the Legislature just
*icks the can down the alley.

Yet another legislative year has come and gone without the major reforms Californians
*verwhelmingly deserve. In light of this, and after careful consideration, I believe it is
*nnecessary to sign this measure at this time.

Sincerely,

Arnold

"*My goodness. What a coincidence* [. . .] "

[Aaron McLear, Schwarzenegger spokesman, Oct. 2009]

Paraphrasing "The Acrostify Benchmark"

An acrostic is a poem or other form of writing in which the first letter, syllable or word of each line, paragraph or other recurring feature in the text spells out a word or a message.

[Wikipedia]

A poem [Kuperavage 2000] :

```
H   He broke my heart
E   Every piece, shattered
A   All I wanted was his love
R   Real, as he promised
T   True, as mine for him
...
```

Paraphrasing “The Acrostify Benchmark”

An acrostic is a poem or other form of writing in which the first letter, syllable or word of each line, paragraph or other recurring feature in the text spells out a word or a message. [Wikipedia]

A poem [Kuperavage 2000] :

```
H   He broke my heart
E   Every piece, shattered
A   All I wanted was his love
R   Real, as he promised
T   True, as mine for him
...
```

Task

Given: (1) A text T and an acrostic x.

(2) Lower and upper bounds on the desired line lengths.

Task: Find a paraphrased version T^* of T in monospaced font that encodes x in some consecutive lines, if possible. Each line of T^* has to meet the length constraints.

Paraphrasing “The Acrostify Benchmark”

An acrostic is a poem or other form of writing in which the first letter, syllable or word of each line, paragraph or other recurring feature in the text spells out a word or a message. [Wikipedia]

A poem [Kuperavage 2000] :

```
H   He broke my heart
E   Every piece, shattered
                              e
```

A complex search problem.

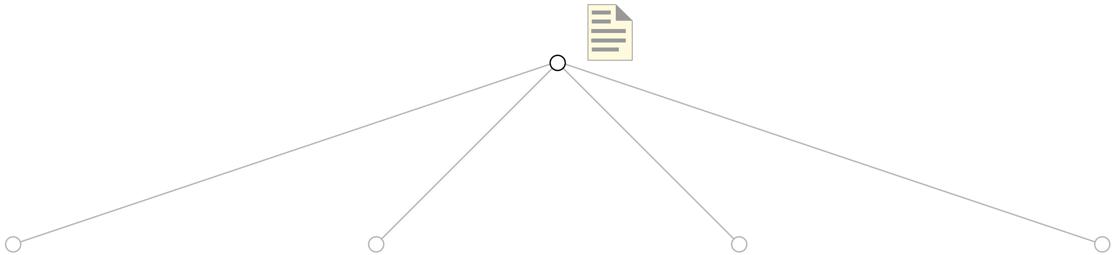
(that may be tackled with AI technologies)

Task

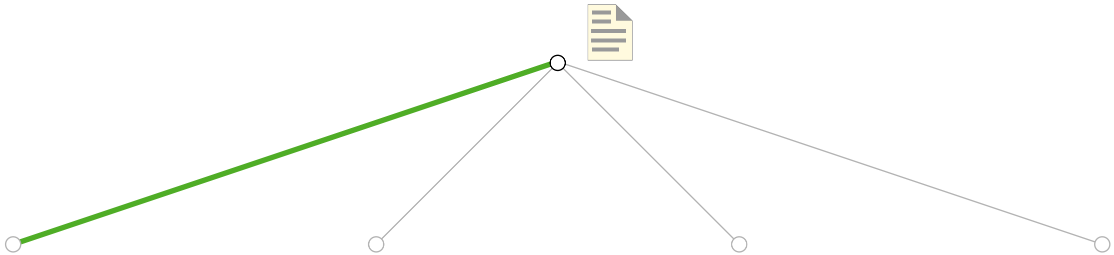
Given: (1) A text T and an acrostic x.

(2) Lower and upper bounds on the desired line lengths.

Task: Find a paraphrased version T^* of T in monospaced font that encodes x in some consecutive lines, if possible. Each line of T^* has to meet the length constraints.



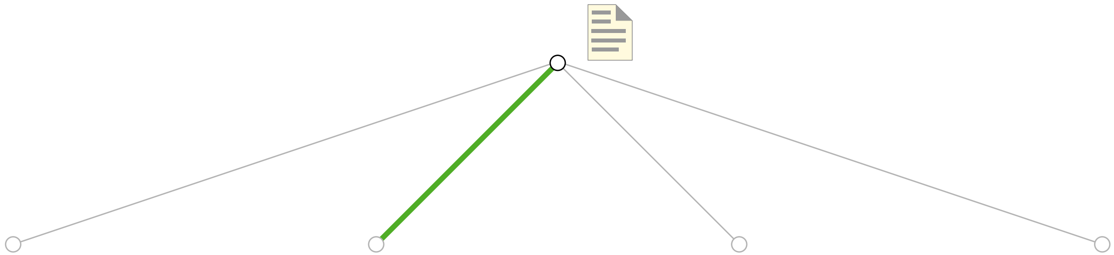
Subtask: Create the character **b**auhaus



Before some time ~~now~~ I have lamented the fact that major issues are overlooked while many bills come to

«Preposition»

Subtask: Create the character **b**auhaus



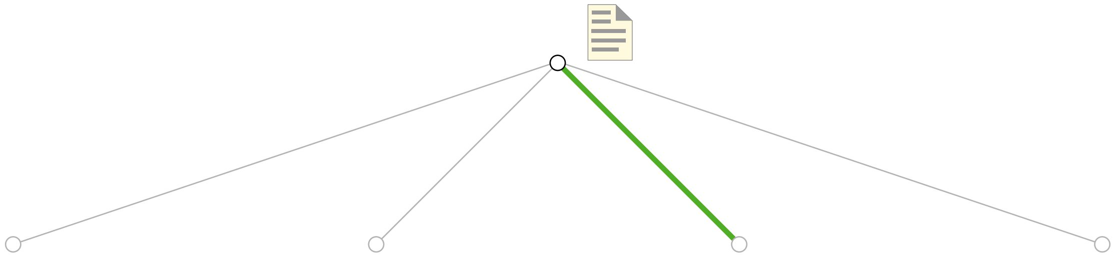
Before some time ~~now~~ I have lamented the fact that major issues are overlooked while many bills come to

«Preposition»

For some time now I have lamented | **but** the fact that major issues are overlooked while many bills

«Add Connective»

Subtask: Create the character **b**auhaus



Before some time ~~now~~ I have lamented the fact that major issues are overlooked while many bills come to

«Preposition»

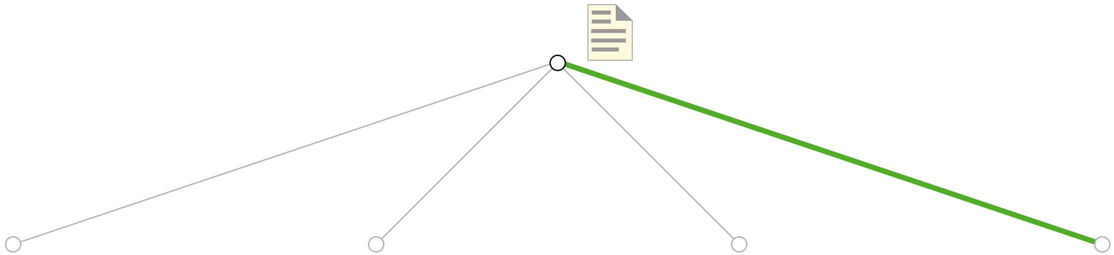
For some time now I have lamented | **but** the fact that major issues are overlooked while many bills

«Add Connective»

Been for some time now I have lamented the fact that major issues are overlooked while many bills come to

«Change Tense»

Subtask: Create the character **b**auhaus



Before some time ~~now~~ I have lamented the fact that major issues are overlooked while many bills come to

«Preposition»

For some time now I have lamented | **but** the fact that major issues are overlooked while many bills

«Add Connective»

Been for some time now I have lamented the fact that major issues are overlooked while many bills come to

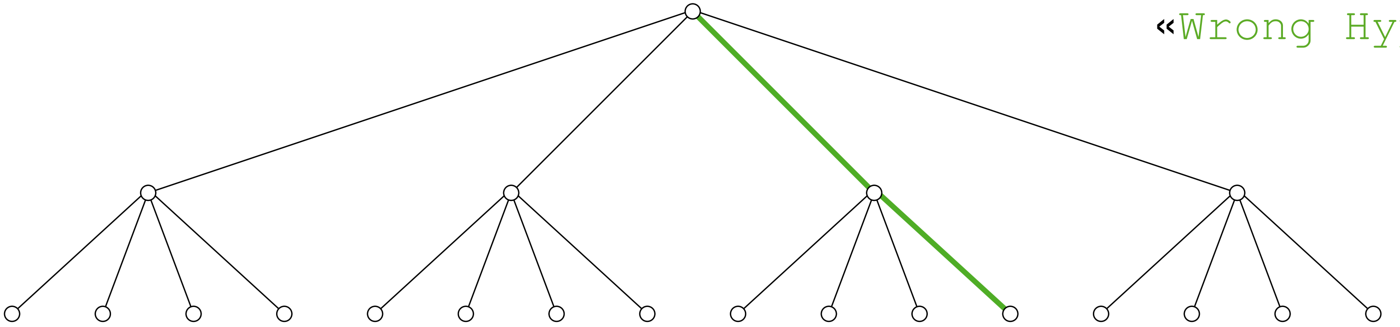
«Change Tense»

For some time now I have lamented the fact that major issues are overlooked while many |
bills come to

«Linebreak»

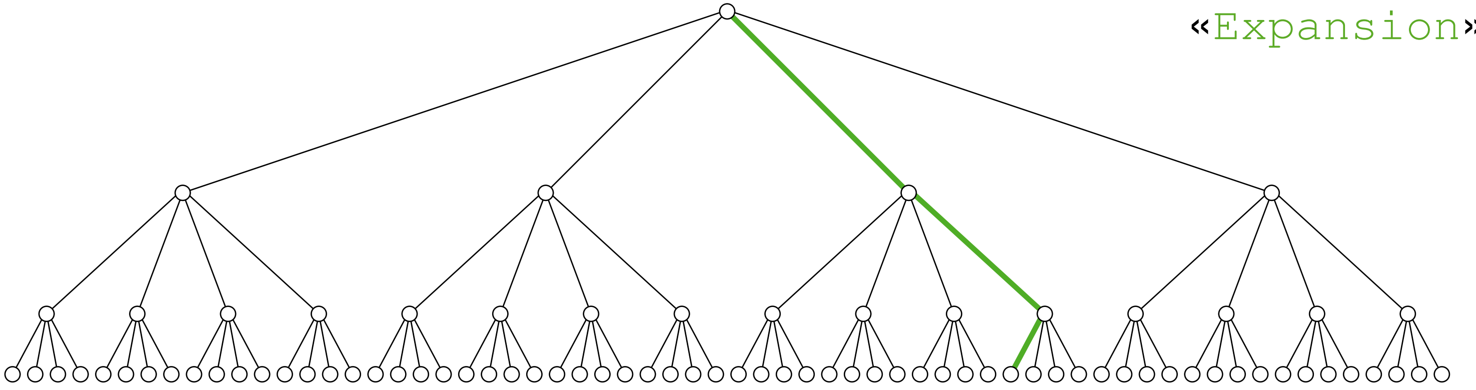
Subtask: Create the character **b**auhaus

«Wrong Hyphen»



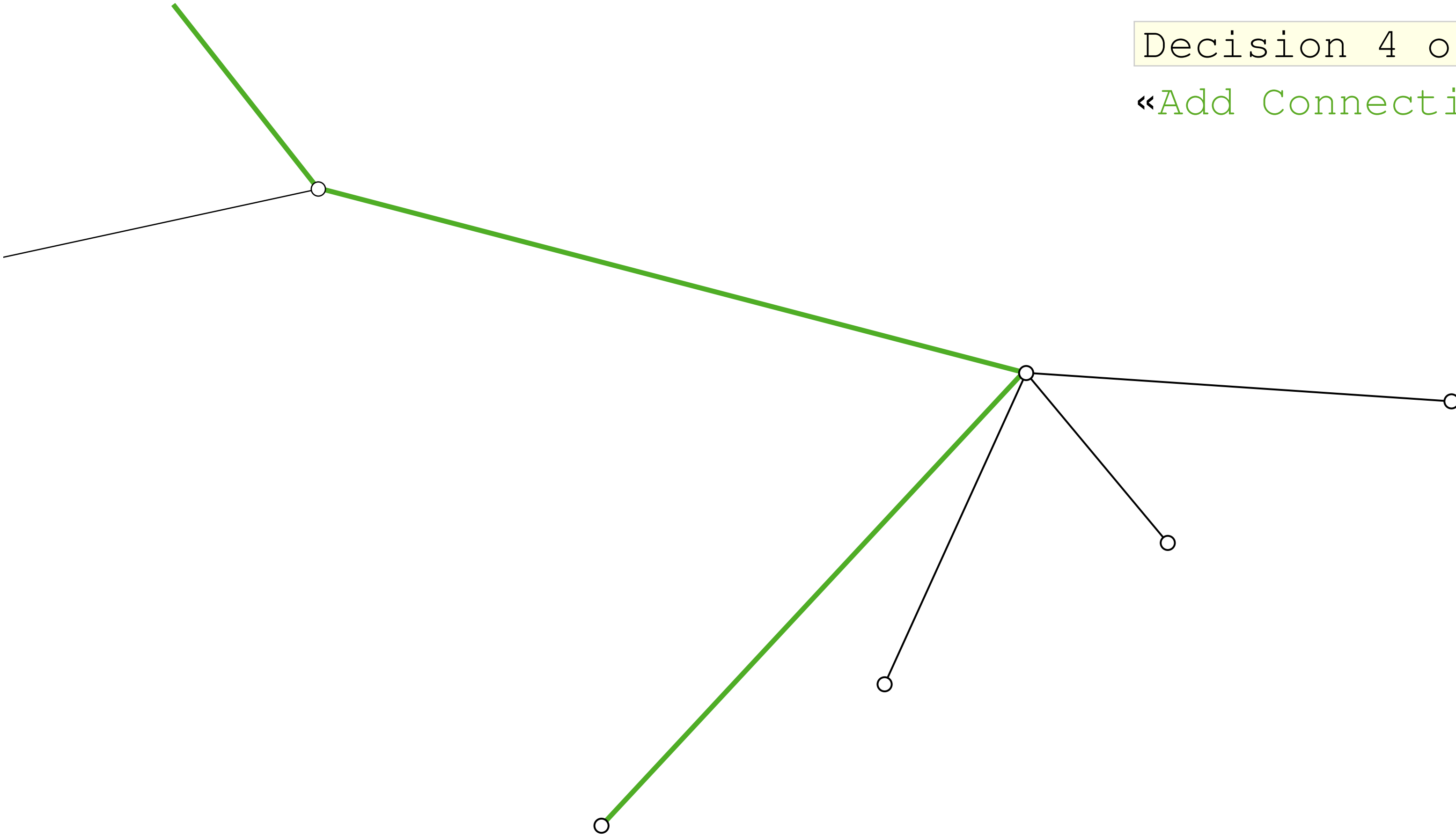
Subtask: Create the character bauhaus

«Expansion»



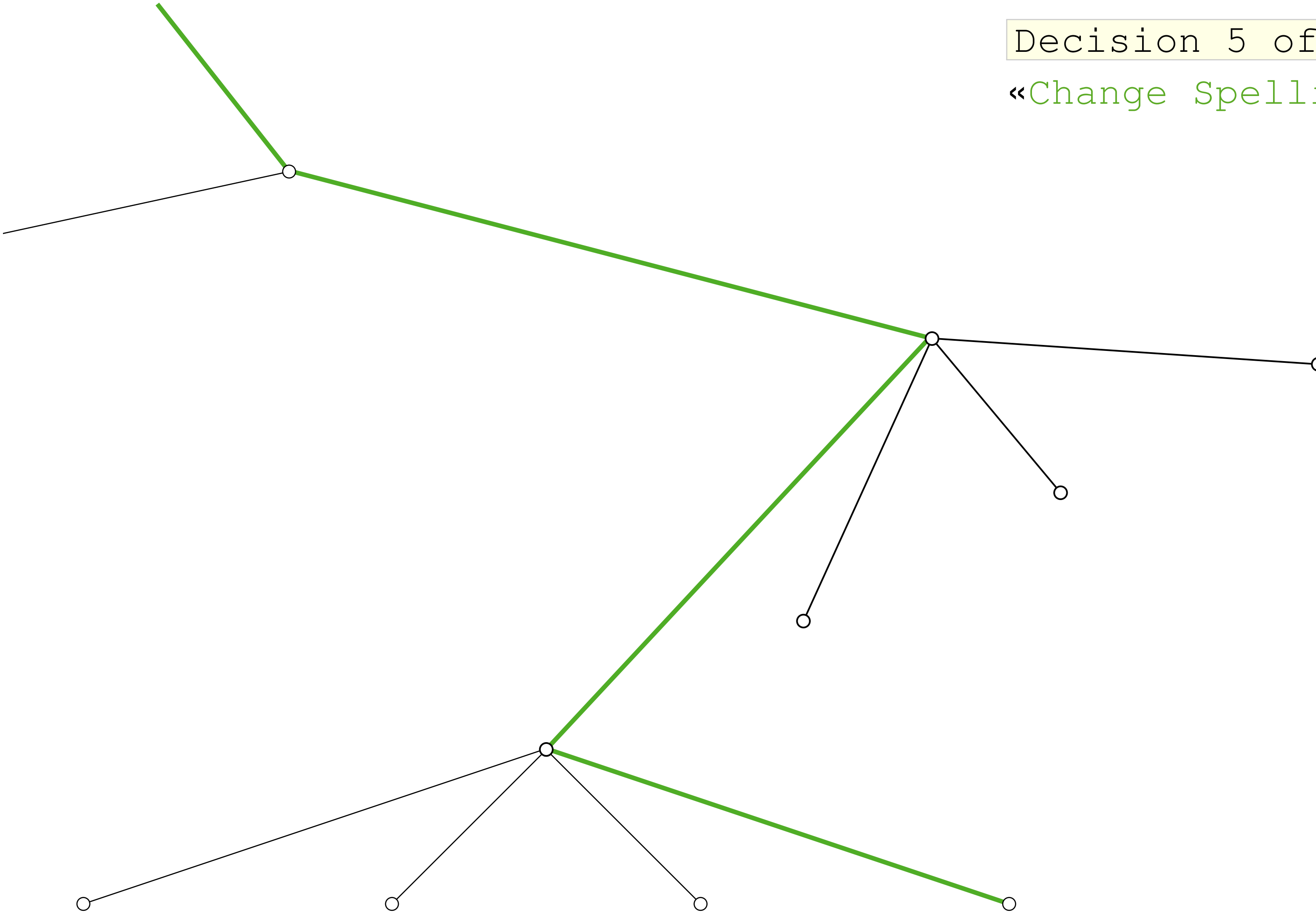
Subtask: Create the character ba**u**haus

«Add Connective»



Subtask: Create the character bau**h**aus

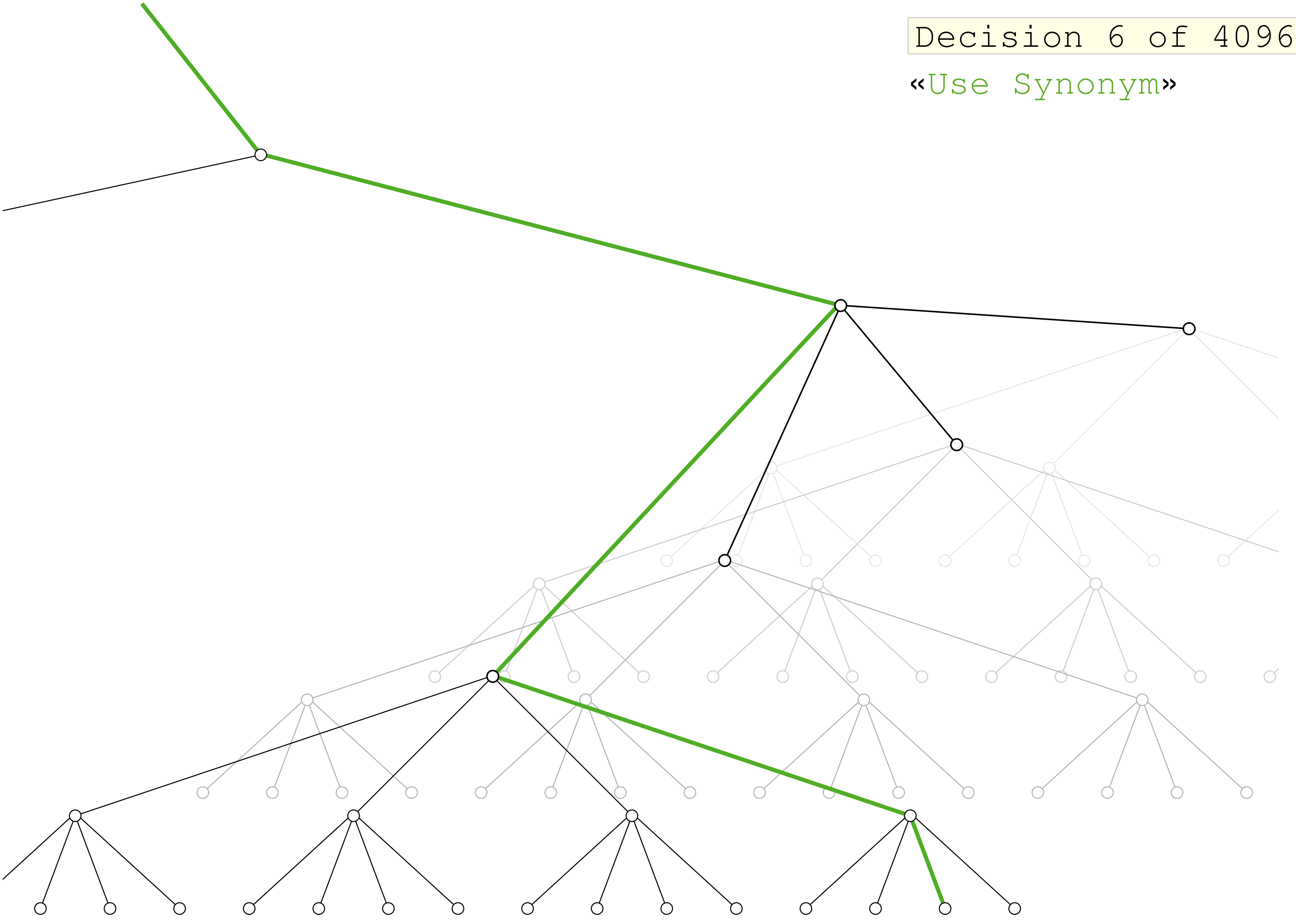
«Change Spelling»



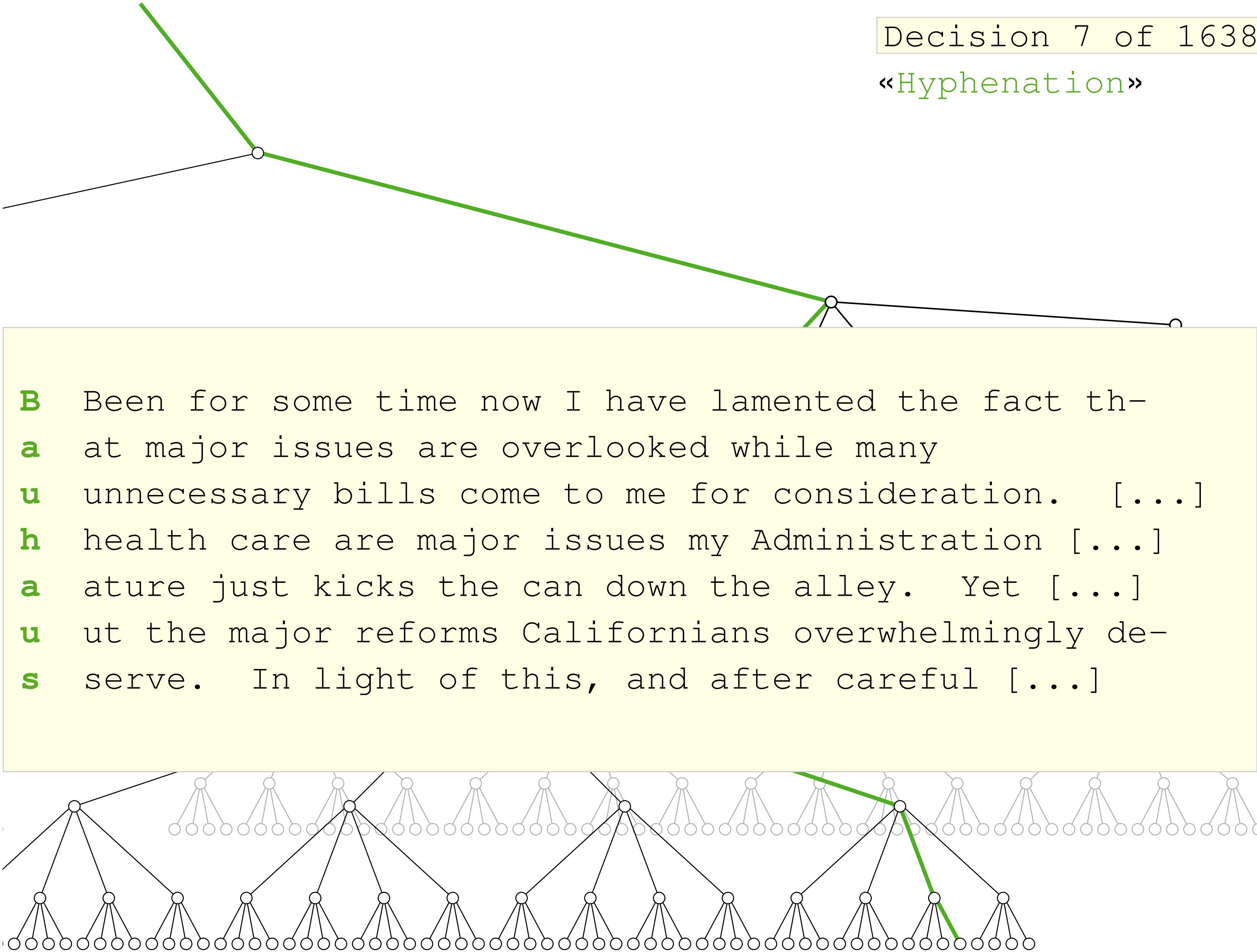
Subtask: Create the character bauh**a**us

Decision 6 of 4096

«Use Synonym»



«Hyphenation»



B Been for some time now I have lamented the fact th-
a at major issues are overlooked while many
u unnecessary bills come to me for consideration. [...]
h health care are major issues my Administration [...]
a ature just kicks the can down the alley. Yet [...]
u ut the major reforms Californians overwhelmingly de-
s serve. In light of this, and after careful [...]

Paraphrasing Searchspace Facts

Consider a text with a length of 100 words (the Schwarzenegger Letter) . . .

- $\approx$ 10 $\cdot$ 3 possibilities to change tense
- $\approx$ 100 possibilities to break a line
- $\approx$ 100 $\cdot$ 3 possibilities to introduce a synonym
- $\approx$ 100 $\cdot$ 3 possibilities to introduce filler words
- $\approx$ 100 $\cdot$ 5 possibilities to hyphenate a word
- $\gg$ 100 possibilities to introduce tautologies
- . . .

Paraphrasing Searchspace Facts

Consider a text with a length of 100 words (the Schwarzenegger Letter) . . .

- $\approx$ 10 · 3 possibilities to change tense
- $\approx$ 100 possibilities to break a line
- $\approx$ 100 · 3 possibilities to introduce a synonym
- $\approx$ 100 · 3 possibilities to introduce filler words
- $\approx$ 100 · 5 possibilities to hyphenate a word
- $\gg$ 100 possibilities to introduce tautologies
- . . .

➜ $>$ 1 000 possible operations to generate a single letter of an acrostic

➜ $O(10^{3n})$ possibilities to synthesize an $n = 7$ letter word like '`Bauhaus`'

Compare the following numbers:

10^{80} atoms in the observable universe

10^{123} game-tree complexity of chess

Paraphrasing Selected Results

Acrostic type	Length (in letters)	Runtime (total in s)	Nodes (total)	Quality-related measures △ WFC	△ ARI	△ SMOG
Common English words						
Adjective						
Noun						
Verb						
Common US first names						
Male						
Female						
Self-referential						
First words						
Average						

Setup details:

- Text genres: Reuters newspaper articles, Enron emails, English Wikipedia articles
- Hardware: standard quad-core PC with 16GB RAM

Paraphrasing Selected Results

Acrostic type	Length (in letters)	Runtime (total in s)	Nodes (total)	Quality-related measures		
				△ WFC	△ ARI	△ SMOG
Common English words						
Adjective	4.4	3.3	287 000			
Noun	4.8	3.4	285 000			
Verb	3.6	2.8	251 000			
Common US first names						
Male	6.0	9.3	852 000			
Female	6.1	7.8	740 000			
Self-referential						
First words	10.3	36.1	3 165 000			
Average	**5.2**	**8.5**	**760 000**			

Setup details:

- Text genres: Reuters newspaper articles, Enron emails, English Wikipedia articles
- Hardware: standard quad-core PC with 16GB RAM

Paraphrasing Selected Results

Acrostic type	Length (in letters)	Runtime (total in s)	Nodes (total)	Quality-related measures		
				△ WFC	△ ARI	△ SMOG
Common English words						
Adjective	4.4	3.3	287 000	-1.0	-1.6	-0.9
Noun	4.8	3.4	285 000	-0.4	-1.0	-0.5
Verb	3.6	2.8	251 000	-1.0	-1.6	-0.9
Common US first names						
Male	6.0	9.3	852 000	-0.7	-1.9	-0.9
Female	6.1	7.8	740 000	-0.6	-1.8	-0.9
Self-referential						
First words	10.3	36.1	3 165 000	-0.3	-0.1	0.2
Average	**5.2**	**8.5**	**760 000**	**-0.8**	**-1.5**	**-0.8**

Setup details:

- Text genres: Reuters newspaper articles, Enron emails, English Wikipedia articles
- Hardware: standard quad-core PC with 16GB RAM

4

Challenges

Search

Question queries
Axiomatic re-ranking
Argument search

Assessment

Clickbait
Text quality
Fake News and Hyperpartisanship
Offensive language

Detection

Vandalism
Plagiarism
Authorship

Synthesis

Summarization
Paraphrasing
Obfuscation

Web as Corpus

Mnemonic passwords

Fake news detection

- Knowledge-based (also called fact checking)
  - Information retrieval
    - Etzioni et al., 2008
    - Magdy and Wanas, 2010
    - Ginsca et al., 2015
  - Semantic web / LOD
    - Wu et al., 2014
    - Ciampaglia et al., 2015
    - Shi and Weninger, 2016
- Context-based
  - Social network analysis
    - Acemoglu et al., 2010
    - Budak et al., 2011
    - Nguyen et al., 2012
    - Kwon et al., 2013
    - Mocanu et al., 2015
    - Tambuscio et al., 2015
- Style-based
  - Deception detection
    - Wei et al., 2013
    - Chen et al., 2015
    - Conroy et al., 2015
    - Rubin et al., 2015
  - Text categorization
    - Badaskar et al., 2008
    - Afroz et al., 2012
    - Rubin et al., 2016


ADDICTING INFO
ALL THE KNOWLEDGE YOU CRAVE
WWW.ADDICTINGINFO.ORG
abc NEWS
we are the 99%
OCCUPY DEMOCRATS
CNN
F
Other98
POLITICO
RIGHT WING News
STAND UP, SPEAK OUT, FIGHT BACK!

Orientation	**Fact-checking results**				
Publisher	true	mix	false	n/a	Σ
Mainstream	806	8	0	12	826
ABC News	90	2	0	3	95
CNN	295	4	0	8	307
Politico	421	2	0	1	424
Left-wing	182	51	15	8	256
Addicting Info	95	25	8	7	135
Occupy Democrats	59	25	7	0	91
The Other 98%	28	1	0	1	30
Right-wing	276	153	72	44	545
Eagle Rising	106	47	25	36	214
Freedom Daily	49	24	22	4	99
Right Wing News	121	82	25	4	232
Σ	1264	212	87	64	1627

Annotations provided by journalists at BuzzFeed

Fake News and Hyperpartisanship Selected Results

Orientation Publisher	**Fact-checking results** true	mix	false	n/a	Σ
Mainstream	806	8	0	12	826
ABC News				3	95
CNN				8	307
Politico				1	424
Left-wing				8	256
Addicting Info				7	135
Occupy Democr				0	91
The Other 98%				1	30
Right-wing	276	153	72	44	545
Eagle Rising	106	47	25	36	214
Freedom Daily	49	24	22	4	99
Right Wing News	121	82	25	4	232
Σ	1264	212	87	64	1627

Fake News Detection

Precision $\approx$ 42%

Recall $\approx$ 41%

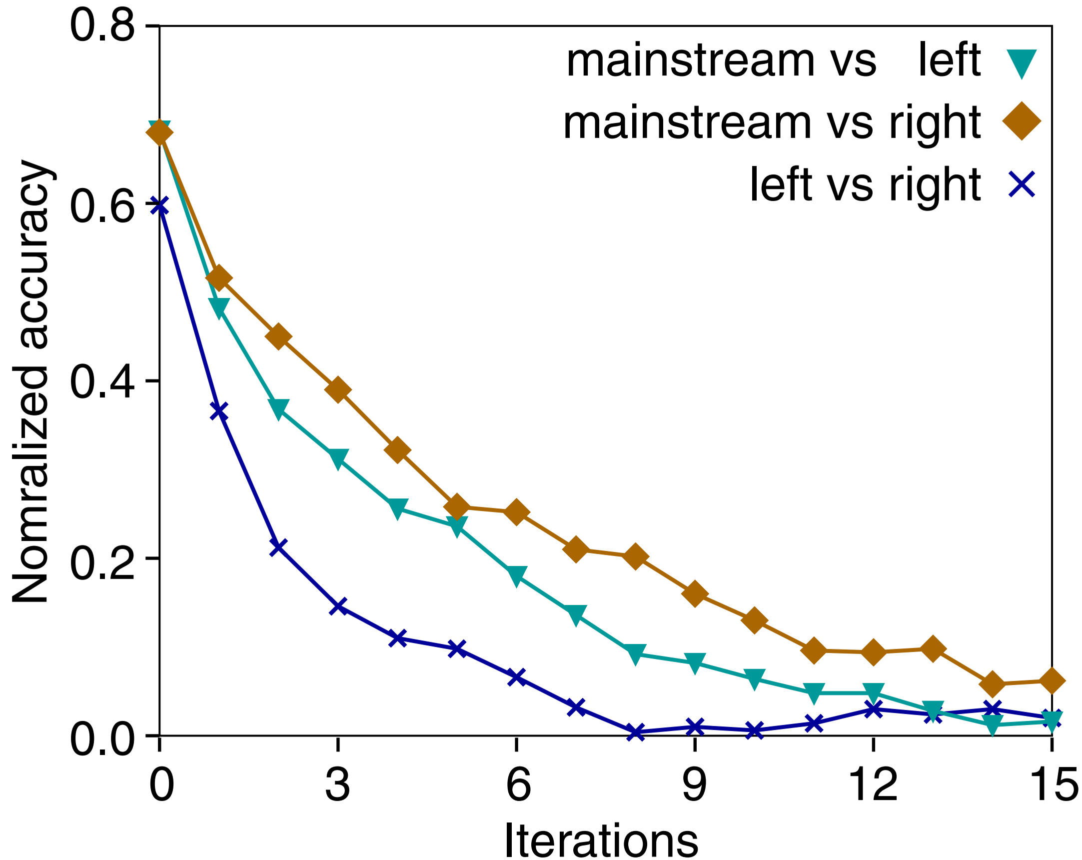
Fake News and Hyperpartisanship Selected Results

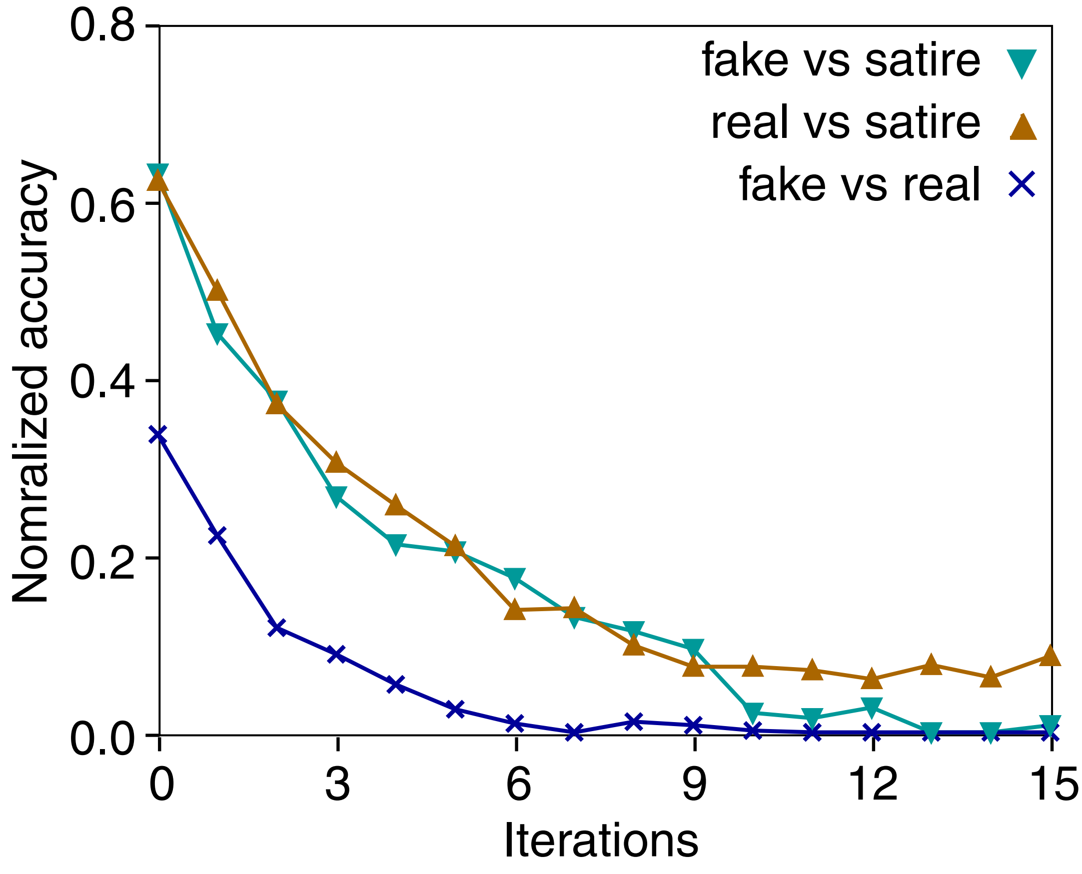
Orientation Publisher	**Fact-checking results** true	mix	false	n/a	Σ
Mainstream	806	8	0	12	826
ABC News					95
CNN					307
Politico					424
Left-wing					256
Addicting In					135
Occupy Der					91
The Other 9					30
Right-wing	276	153	72	44	545
Eagle Rising	106	47	25	36	214
Freedom Daily	49	24	22	4	99
Right Wing News	121	82	25	4	232
Σ	1264	212	87	64	1627

Hyperpartisanship Detection

Precision ≈ 69%

Recall ≈ 89%


0.8
0.6
0.4
0.2
0.0
Nomralized accuracy
0
3
6
9
12
15
Iterations
mainstream vs left
mainstream vs right
left vs right


0.8
0.6
0.4
0.2
0.0
Nomralized accuracy
0
3
6
9
12
15
Iterations
fake vs satire
real vs satire
fake vs real

5

Challenges

Search

Question queries

Axiomatic re-ranking

Argument search

Assessment

Clickbait

Text quality

Fake News and Hyperpartisanship

Offensive language

Detection

Vandalism

Plagiarism

Authorship

Web as Corpus

Mnemonic passwords

Synthesis

Summarization

Paraphrasing

Obfuscation

“You won’t believe how far cats can count.”

Source: BildBlog (http://www.bildblog.de/ressort/fuer-sie-geklickt/); Spoiler: 125 years

Huffington Post Deutschland
23 Std. ·

Ob sie den Job des Geldes wegen macht?

Nach Gehaltserhöhung: So viel verdient Angela Merkel jetzt

Es gibt wohl begehrtere Jobs als den des Bundeskanzlers.

HUFFINGTONPOST.DE

Source: BildBlog (http://www.bildblog.de/ressort/fuer-sie-geklickt/); Spoiler: 226.000 Euro p.a.

BuzzFeed
@BuzzFeed

Follow

Hier ist der absolut genialste Weg, ein schlechtes Pokémon-Tattoo zu retten! bzfd.it/1C3yToz

Source: Twitter @BuzzFeed (http://www.twitter.com/buzzfeed/); Spoiler: No clickbait

Clickbait Algorithmic Assessment in Twitter

Register	Publisher*	Impact (retweets in 2015)	Tweets (in week 24)	Clickbait probability
Print + online	New York Times	$23.8 \cdot 10^6$	875	21%
	The Guardian	$14.0 \cdot 10^6$	744	15%
	Forbes	$11.5 \cdot 10^6$	721	38%
	Daily Mail	$6.9 \cdot 10^6$	516	22%
	Wall Street Journal	$6.5 \cdot 10^6$	747	19%
Online only	Mashable	$20.6 \cdot 10^6$	803	33%
	Huffington Post	$11.6 \cdot 10^6$	770	46%
	Bleacher Report	$10.2 \cdot 10^6$	196	9%
	BuzzFeed	$10.0 \cdot 10^6$	695	42%
	Yahoo!	$8.2 \cdot 10^6$	195	23%
Television	BBC News	$39.6 \cdot 10^6$	694	17%
	ABC News	$17.6 \cdot 10^6$	279	9%
	CNN	$15.0 \cdot 10^6$	345	17%
	Fox News	$10.2 \cdot 10^6$	378	8%
	NBC News	$9.7 \cdot 10^6$	408	14%
			Average:	28%

* Top publishers on Twitter in 2014.

Register	Publisher*	Impact (retweets in 2015)	Tweets (in week 24)	Clickbait probability
Print + online	New York Times	$23.8 \cdot 10^6$	875	21%
	The Guardian	$14.0 \cdot 10^6$	744	15%
	Forbes	$11.5 \cdot 10^6$	721	38%
	Daily Mail	$6.9 \cdot 10^6$	516	22%
	Wa			19%
Online only	Ma			33%
	Hu			46%
	Ble			9%
	Bu			42%
	Yal			23%
Television	BB			17%
	AB			9%
	CNN	$15.0 \cdot 10^6$	345	17%
	Fox News	$10.2 \cdot 10^6$	378	8%
	NBC News	$9.7 \cdot 10^6$	408	14%

Average: 28%

Clickbait detection**

Precision $\approx$ 71%

Recall $\approx$ 73%

* Top publishers on Twitter in 2014.

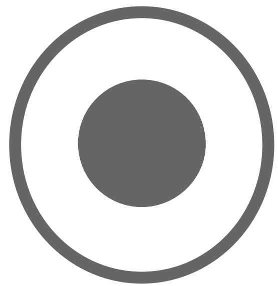
** [Potthast et al., ECIR'16]

Clickbait A Challenge!

http://www.clickbait-challenge.org/
A challenge is organized to encourage researchers in this.

Corpus size	40 000 tweets
Votes per tweet	5
Votes per “check instance”	> 60
Number of AMT workers	3 500

Dedicated acquisition technology and statistics.



Summary

Web as Corpus

Mnemonic passwords

Synthesis

Summarization

Paraphrasing

Obfuscation

Search

Question queries

Axiomatic re-ranking

Argument search

Assessment

Clickbait

Text quality

Fake News and Hyperpartisanship

Offensive language

Detection

Vandalism

Plagiarism

Authorship

Summary

Search

Question queries

Axiomatic re-ranking

Argument search

Assessment

Clickbait

Text quality

Fake News and Hyperpartisanship

Offensive language

Detection

Vandalism

Plagiarism

Authorship

Synthesis

Summarization

Paraphrasing

Obfuscation

Web as Corpus

Mnemonic passwords

Summary

Search

Question queries

Axiomatic re-ranking

Argument search

Assessment

Clickbait

Text quality

Fake News and Hyperpartisanship

Offensive language

Detection

Vandalism

Plagiarism

Authorship

Web as Corpus

Mnemonic passwords

Synthesis

Summarization

Paraphrasing

Obfuscation

Summary

Search

Question queries

Axiomatic re-ranking

Argument search

Assessment

Clickbait

Text quality

Fake News and Hyperpartisanship

Offensive language

Detection

Vandalism

Plagiarism

Authorship

Synthesis

Summarization

Paraphrasing

Obfuscation

Web as Corpus

Mnemonic passwords

Summary

Web as Corpus

Mnemonic passwords

Synthesis

Summarization

Paraphrasing

Obfuscation

Search

Question queries

Axiomatic re-ranking

Argument search

Assessment

Clickbait

Text quality

Fake News and Hyperpartisanship

Offensive language

Detection

Vandalism

Plagiarism

Authorship

Summary

Web as Corpus

Mnemonic passwords

Synthesis

Summarization

Paraphrasing

Obfuscation

Search

Question queries

Axiomatic re-ranking

Argument search

Assessment

Clickbait

Text quality

Fake News and Hyperpartisanship

Offensive language

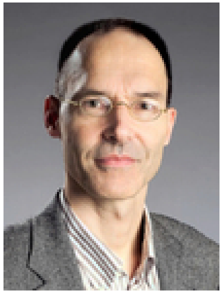
Detection

Vandalism

Plagiarism

Authorship

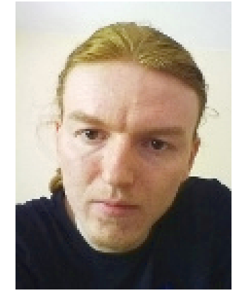
Thank you!



Benno Stein

Matthias Hagen

Henning Wachsmuth



Janek Bevendorff

Michael Völske

Thank you!

Johannes Kiesel

Khalid Al-Khatib

Tim Gollub

Shahbaz Syed

Yamen Ajjour