M. Kestemont, K. Luyckx & W. Daelemans

Intrinsic Plagiarism Detection Using Character Trigram Distance Scores

Under a novel document representation

PAN 2011 @ CLEF

PLAGIARISM DETECTION

- External detection:
  - reference corpus = ALL source documents
  - **‘Closed’ world**
- **Realistic?**
  - Growing potential reference collection (cf. web)
    - Computationally complex!
  - Not all sources digitally/publicly available
  - E.g. *student hiring ghost writer for sections in master thesis: what if ghost writer himself did not plagiarize?*
- Practically **relevant**

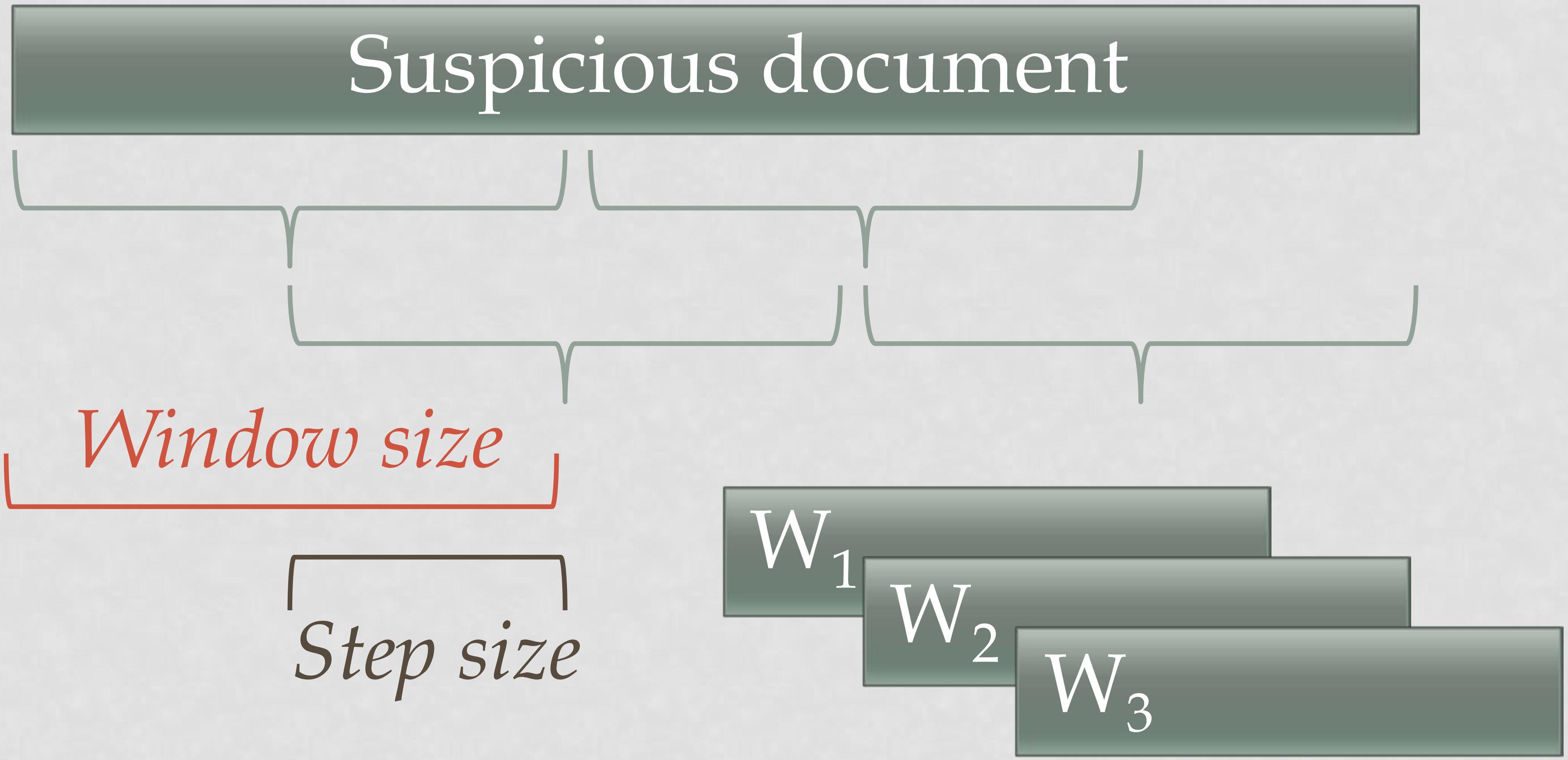
APPROACH?

- **Limited resources**
- Only document itself…
- Seminal work: standard methodology

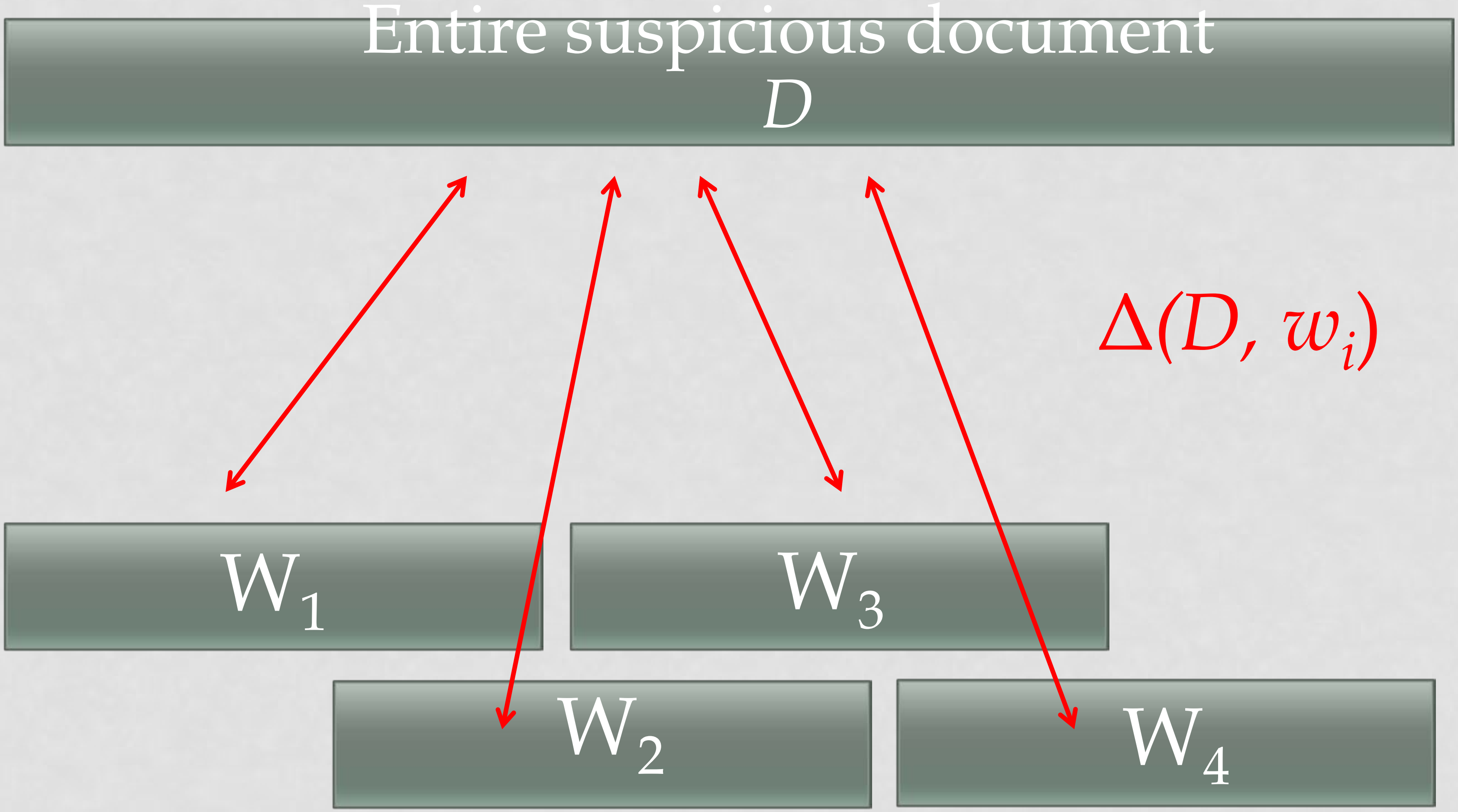
"The underlying approach to intrinsic plagiarism detection has not changed: a suspicious document *d* is chunked, and […] **each chunk is compared with the whole of *d***. Then, chunks whose writing style differs significantly from the average writing style of the document are identified using outlier detection." (PAN overview 2010)

- (Negative undertone?)

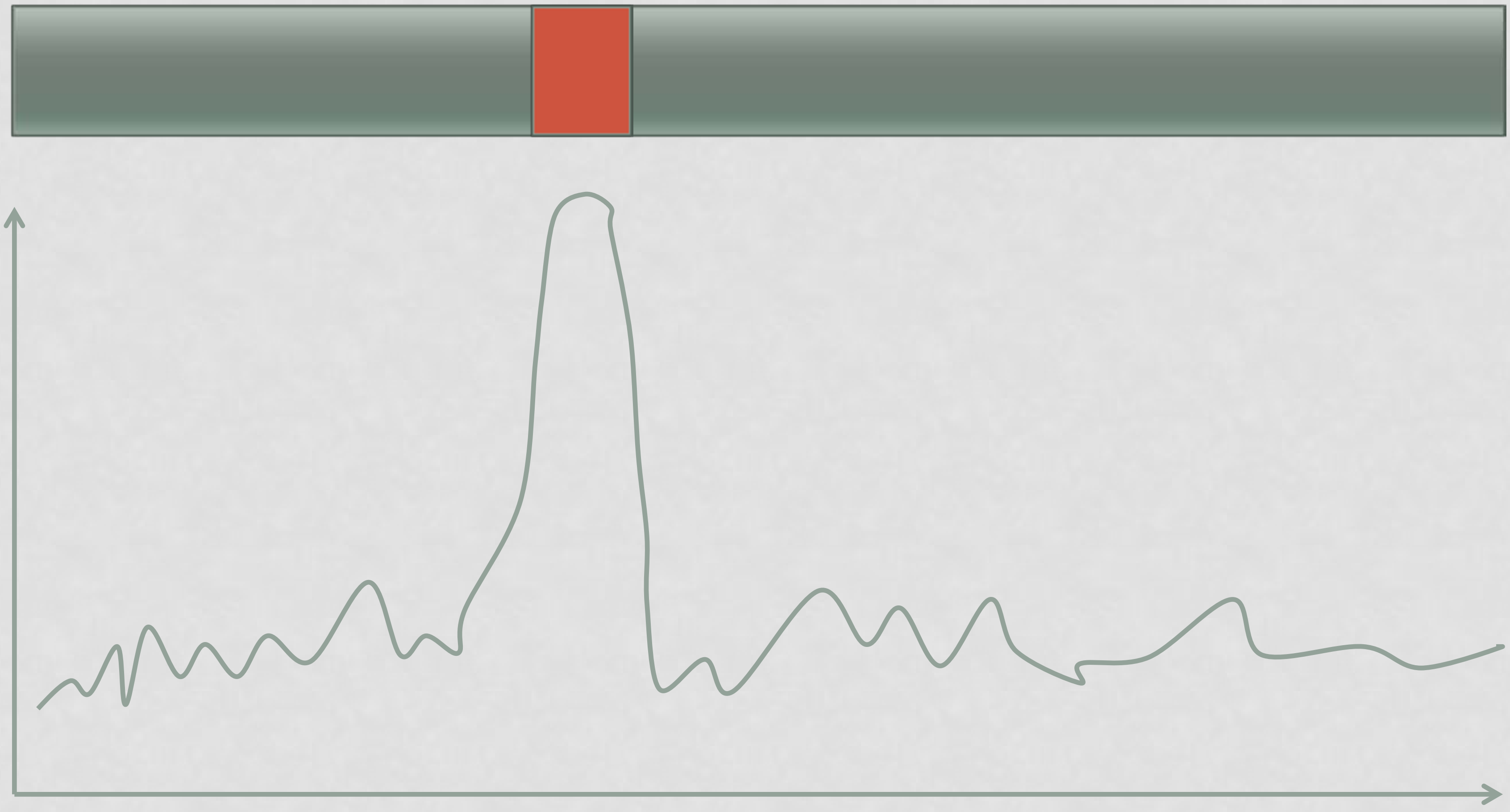
Segments, chunks, windows, …



D vs. $w_1, w_2, w_3, \ldots, w_n$



BEST-CASE SCENARIO

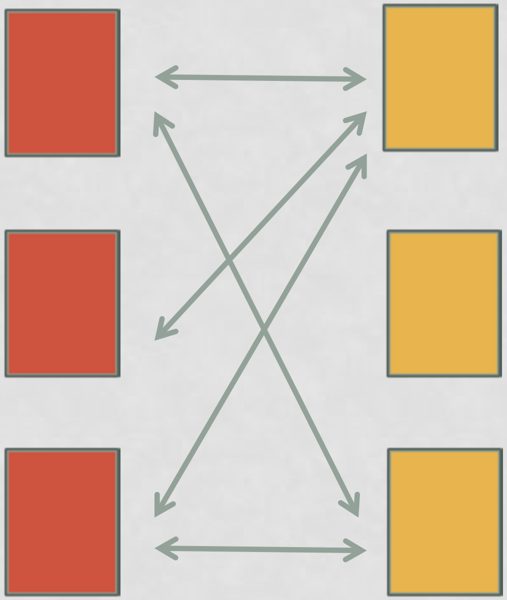


IMPLICIT ASSUMPTIONS?

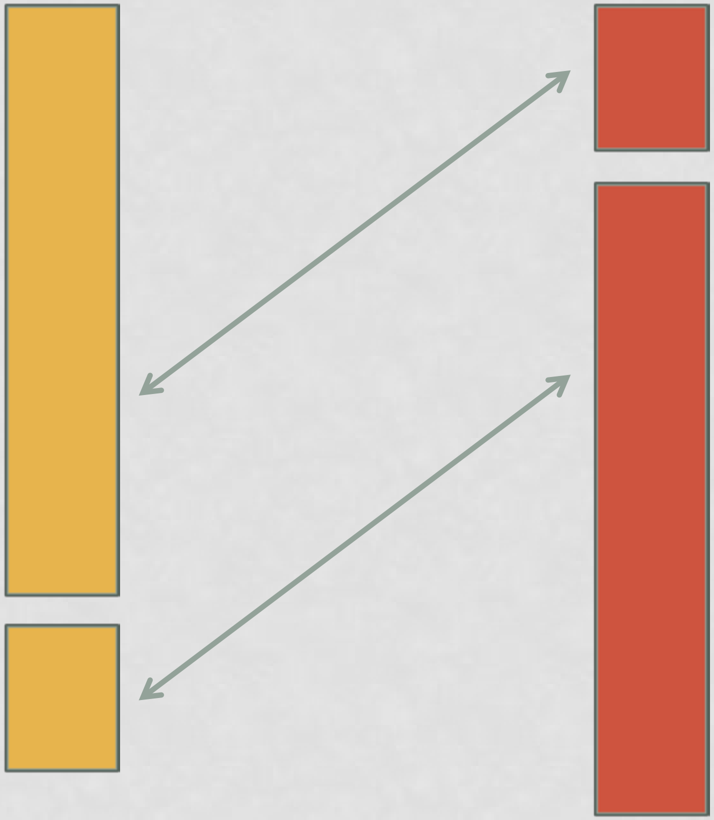
1 – "It's okay to compare a chunk to the document as a whole."

2 – "The whole document is a reliable point of stylistic reference."

COMMON PRACTICE?



Equal size



Different size

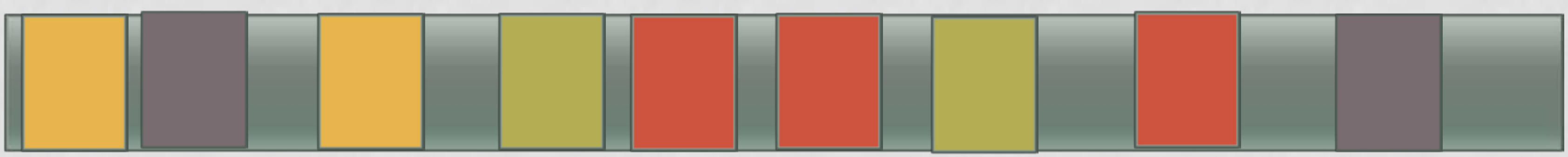
IMPLICIT ASSUMPTIONS?

~~1 – "It's okay to compare a chunk to the document as a whole."~~

2 – "The whole document is a reliable point of stylistic reference."

WORST-CASE SCENARIOS

Original text will be marked as plagiarized?



Which one is the original author?

QUESTIONABLE ASSUMPTIONS

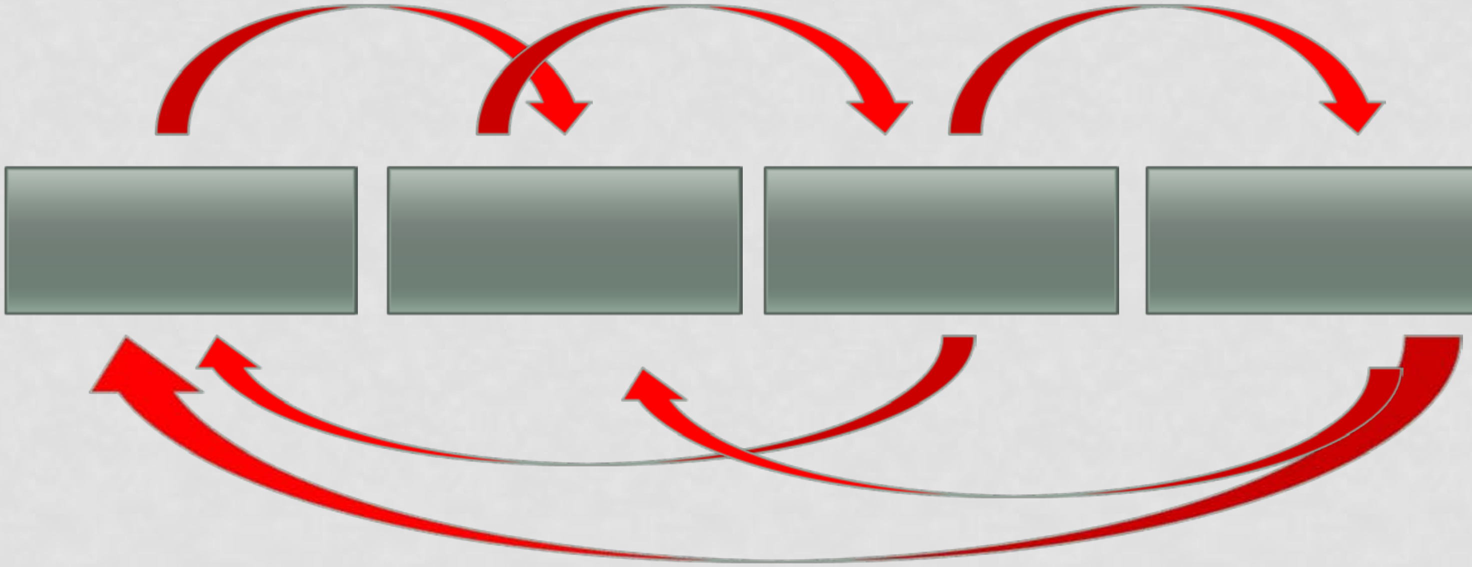
~~1 – "It's ok to compare a chunk to the document as a whole"~~

~~2 – "Whole document is reliable point of stylistic reference"~~

But is there an alternative?

WINDOW VS. WINDOW

- Instead of *Document* vs. *Window*…
- *Window* versus *Window*
  - No assumption of reliability of *D* as a whole
  - Comparing blocks of equal size

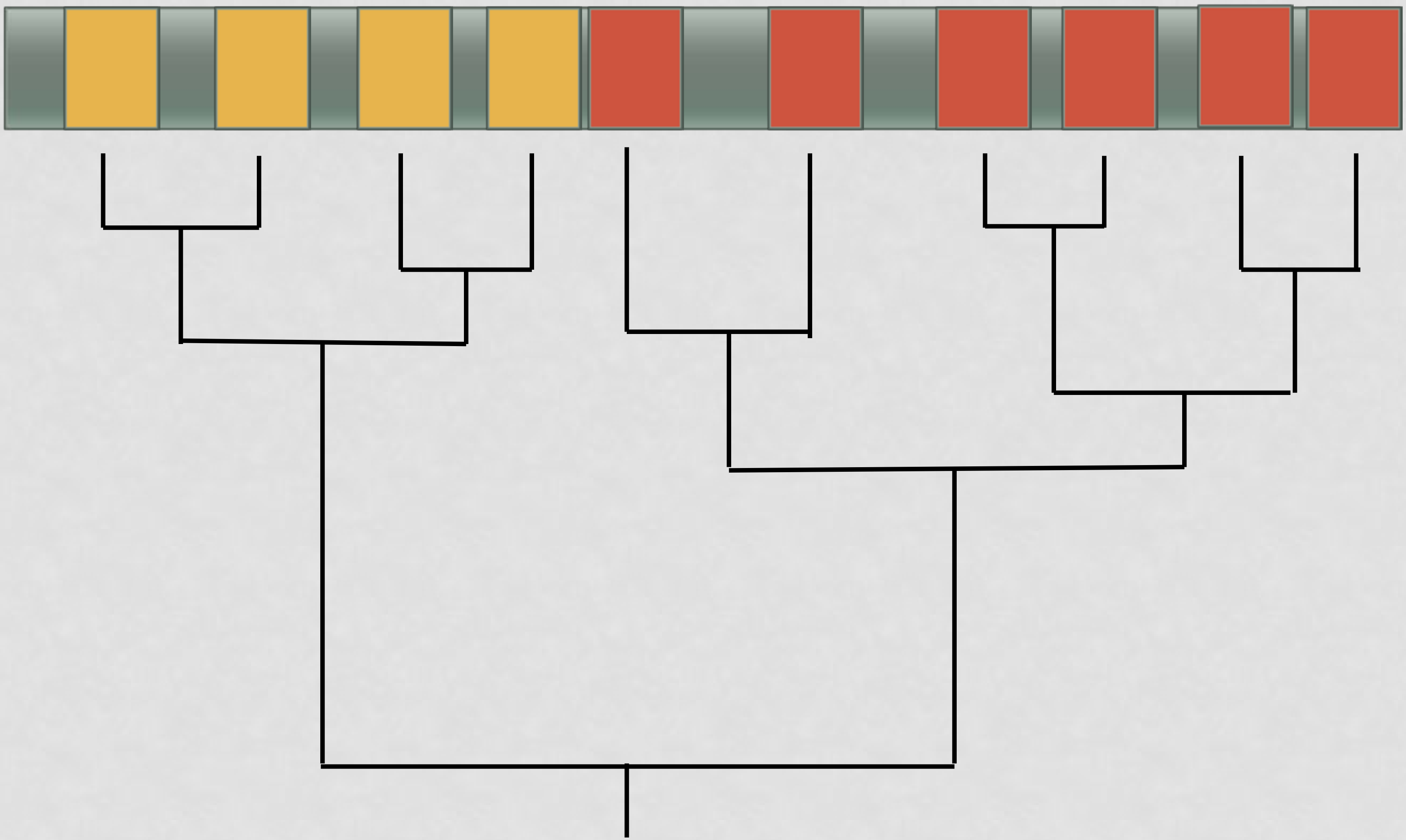


SYMMETRICAL DISTANCE MATRIX

	w_1	w_2	...	w_{n-1}	w_n
w_1	0	$\Delta(w_1,w_2)$	...	$\Delta(w_1,w_{n-1})$	$\Delta(w_1,w_n)$
w_2	$\Delta(w_1,w_2)$	0	...	$\Delta(w_2,w_{n-1})$	$\Delta(w_2,w_n)$
...	...	...	...	...	...
w_{n-1}	$\Delta(w_{n-1},w_1)$	$\Delta(w_{n-1},w_2)$	...	0	$\Delta(w_{n-1},w_n)$
w_n	$\Delta(w_n,w_1)$	$\Delta(w_n,w_2)$	...	$\Delta(w_n,w_{n-1})$	0

Cf. Distance tables for clustering

CLUSTERING OF PLAGIARISMS OF SAME SOURCE



DISTANCE MEASURE

- Stamatatos's normalized distance
- Distance between two 'text profiles'
- Profile = bag-of-character-trigrams

$$\sum_{g \in P(w_x)} \frac{\left(\frac{2(f_{w_x}(g) - f_{w_y}(g))}{f_{w_x}(g) + f_{w_y}(g)}\right)^2}{4|P(w_x)|}$$

SYMMETRIC ADAPTATION

- Originally: **all** trigrams from **1** document
- Asymmetrical: distance(A,B) != distance(B,A)
- Adaptation: restrict to *n*=1000 most frequent character trigrams from entire corpus
- Stylometric inspiration
- Computationally simple: symmetry!

	w_1	w_2	...	w_{n-1}	w_n
w_1	0	$\Delta(w_1,w_2)$	...	$\Delta(w_1,w_{n-1})$	$\Delta(w_1,w_n)$
w_2		0	...	$\Delta(w_2,w_{n-1})$	$\Delta(w_2,w_n)$
...			...	...	...
w_{n-1}				0	$\Delta(w_{n-1},w_n)$
w_n					0

OUTLIERS?

- Distance table (cf. clustering)
- Multivariate, higher-dimensional
- Mvoutlier (*R*, Filzmoser et al.)
- Principal Components Analysis
- Reduces dimensionality before detection

CHUNKING?

ws	ss	plagdet	recall	precision	granularity
20,000	20,000	19.48	20.02	19.01	1.00
20,000	15,000	20.59	21.84	19.88	1.01
20,000	10,000	23.80	27.79	21.00	1.01
20,000	5,000	25.84	39.55	19.52	1.02
20,000	1,000	26.36	44.99	18.91	1.01
15,000	15,000	20.04	20.29	20.71	1.01
15,000	11,250	22.41	23.09	22.41	1.02
15,000	7,500	25.97	29.69	23.44	1.01
15,000	3750	26.79	40.17	20.63	1.02
15,000	750	**27.21**	45.09	19.89	1.02
10,000	10,000	21.33	20.35	23.34	1.03
10,000	7,500	24.14	24.05	25.95	1.05
10,000	5,000	27.26	29.98	25.89	1.03
10,000	2,500	**27.53**	40.00	22.03	1.04
5,000	5,000	21.77	20.38	28.09	1.12
5,000	3,750	24.03	24.18	29.79	1.16
5,000	2,500	**27.52**	30.42	28.50	1.10
5,000	1,250	**27.49**	37.56	24.55	1.11

The smaller the windows, the better (but more expensive)

OUTBOUND PARAMETER

outbound	ws	ss	plagdet	recall	precision	granularity
.20	20,000	20,000	19.92	21.17	18.84	1.00
.20	20,000	5,000	25.87	41.84	19.06	1.02
.30	20,000	5,000	25.66	36.60	20.09	1.01
.30	15,000	3,750	26.82	37.24	21.48	1.02
.35	15,000	3,750	25.68	30.01	22.91	1.02
.30	10,000	2,500	**27.61**	36.93	23.13	1.04
.20	10,000	2,500	27.29	42.25	21.17	1.04

- Controlled ratio of outliers detected
- Higher outbound pushed precision
- Lower outbound pushed recall (even more)

RESULTS

Training corpus (PAN 2010)

- Plagdet: 28.60
- Recall: 36.57
- Precision: 26.70
- Granularity: 1.11

Test corpus (PAN 2011-INTR)

- Plagdet: 16.79 (2nd place)
- Recall: **42.79** (!)
- Precision: 10.75 (?)
- Granularity: 1.03

<u>Comparison</u>

- *ws* = 5000, *ss* = 2500, *n* = 2500, *outbound* = .20
- Disappointing precision – dramatic drop
- Method does invariably great in recall
- Shorter documents in test?

REFERENCES

- **Filzmoser, P. ,Maronna, R. ,Werner, M.** (2008). Outlier identification in high dimensions. *Computational Statistics and Data Analysis* 52(3).
- **Potthast, M., Barrón Cedeño, A., Eiselt, A. ,Stein, B., Rosso, P.** (2010). Overview of the 2nd International Competition on Plagiarism Detection. *Notebook Papers of CLEF 2010 LABs and Workshops.*
- **Stamatatos, E.** (2009). A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology* 60(3).
- **Stamatatos, E.** (2009). Intrinsic Plagiarism Detection Using Character Ngram Profiles. *Proceedings of the 3rd International Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse* (2009)
- **Stein, B., Lipka, N., Prettenhoffer, P.** (2011). Intrinsic Plagiarism Analysis. *Natural Language Engineering* 45(1).
- **Luyckx, K., Daelemans,** W. (2011). The effect of author set size and data size in authorship attribution. *Literary and Linguistic Computing* 26(1).