FASTDOCODE: Finding Approximated Segments of N-Grams for Document Copy Detection

Lab Report for PAN at CLEF 2010

Gabriel Oberreuter
goberreu@ing.uchile.cl
Universidad de Chile - September 2010

Group members : Gastón L'Huillier, Sebastián Ríos and Juan D. Velásquez

DOCODE

- Project *DOCODE:* implement a plagiarism detection software.
- Sponsored by FONDEF & Sistemas Complejos de Ingeniería, Universidad de Chile.

  - Aimed at tackling the increasingly plagiarism problem on our country.
  - Motivate research on this field.

DOCODE

Desirable Features:

- Designed for large amount of documents.
- Capable of finding obfuscated plagiarism.
- Results in reasonable time.
- Good precision.

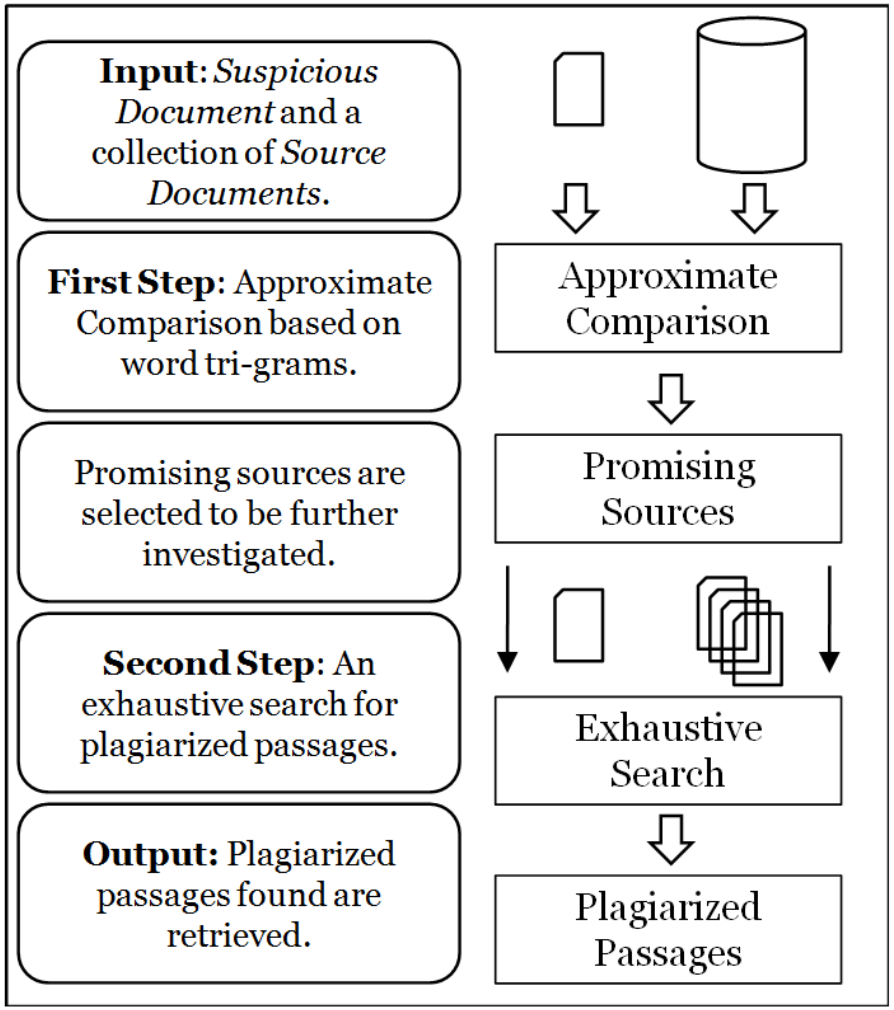
PAN: Uncovering Plagiarism, Authorship, and Social Software Misuse

- Workshop & Competition on Plagiarism Uncovering.
- Includes lot's of documents.
- Verbatim and obfuscated copy.
- Benchmark against other approaches.

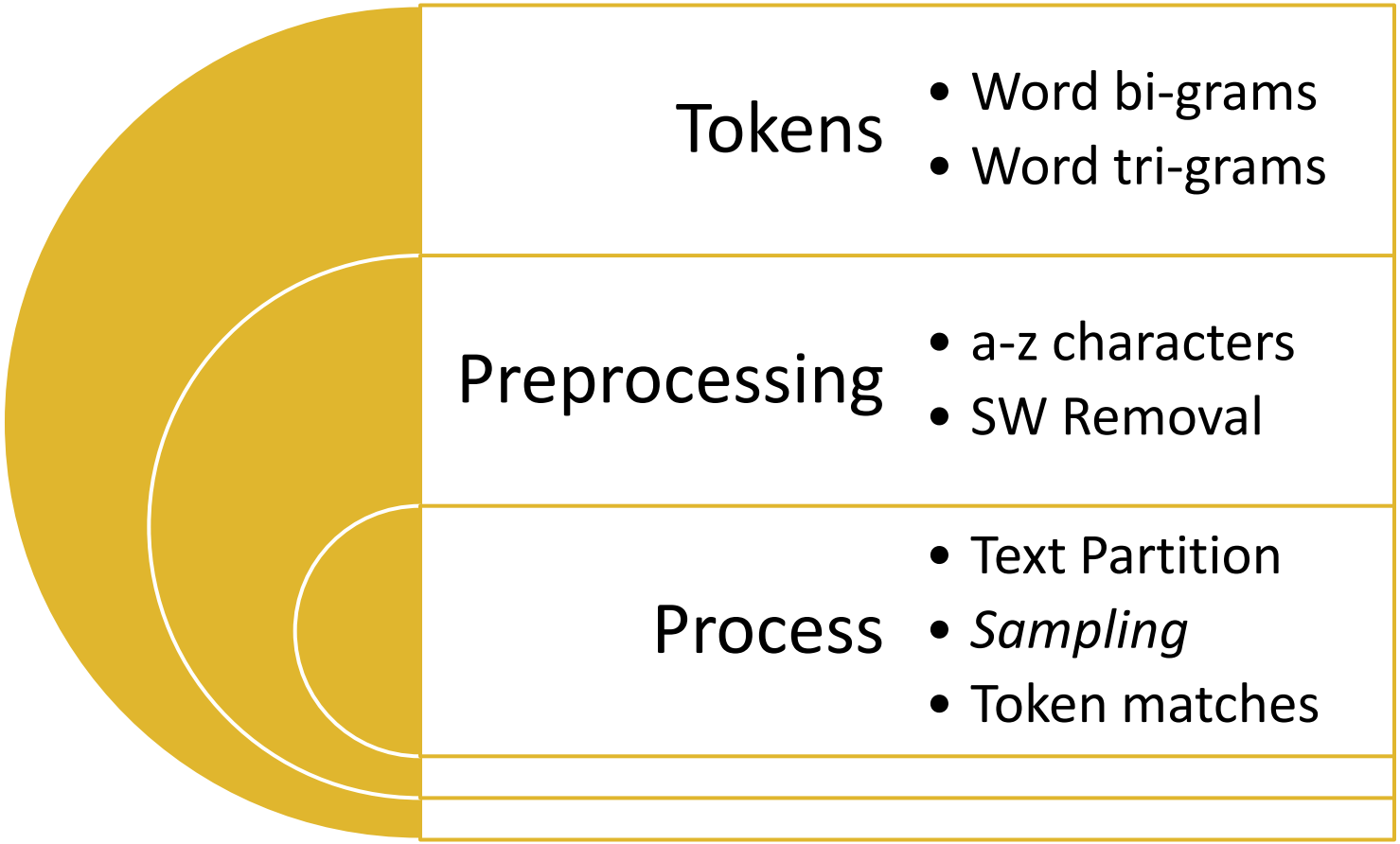
<u>Perfect opportunity and motivation for our purpose.</u>

Our Approach

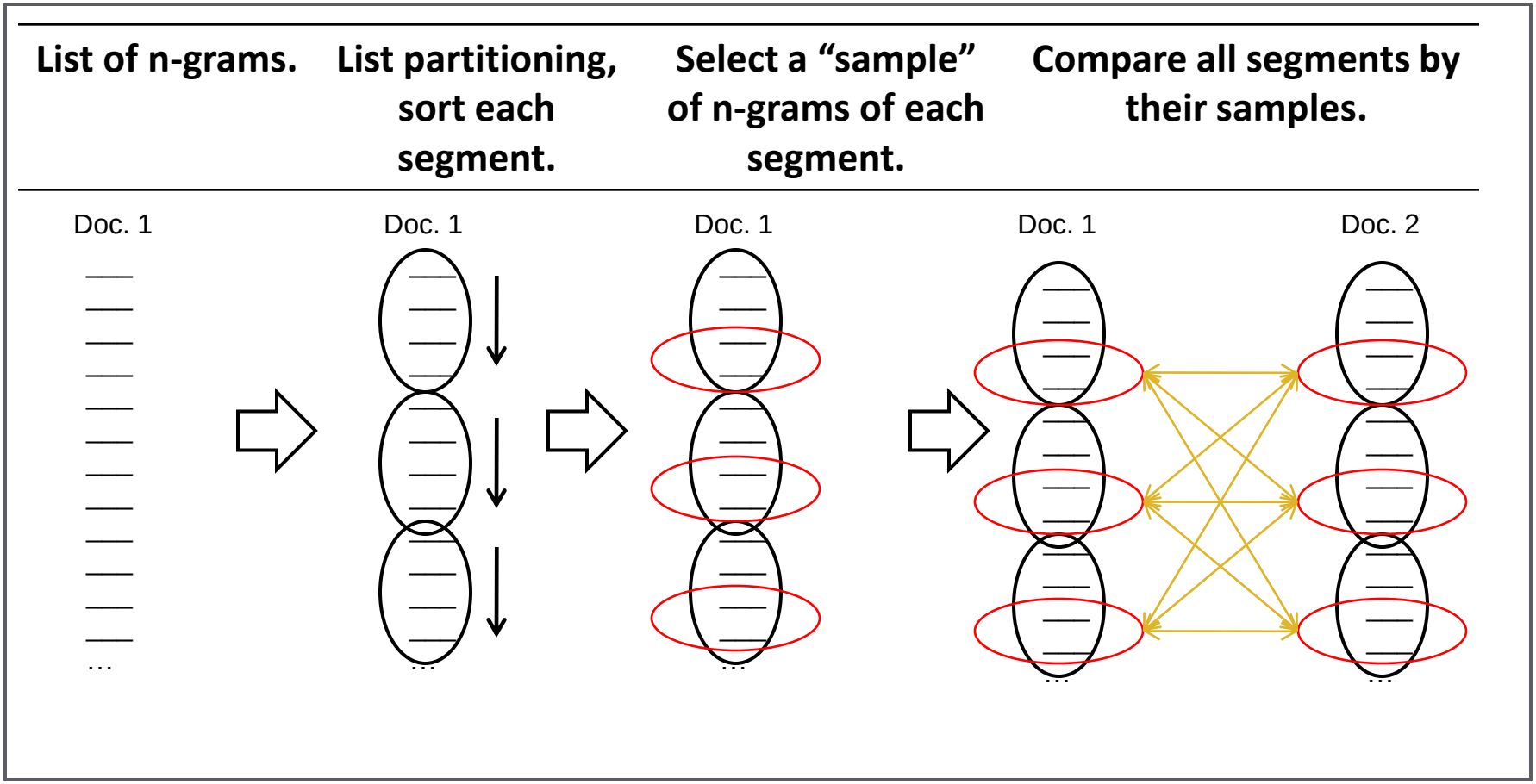
- Focused on external plagiarism detection.
- No cross-lingual consideration.
- Based on word bi-grams and word tri-grams.
- Selecting *samples* for each document in order to reduce compute time.



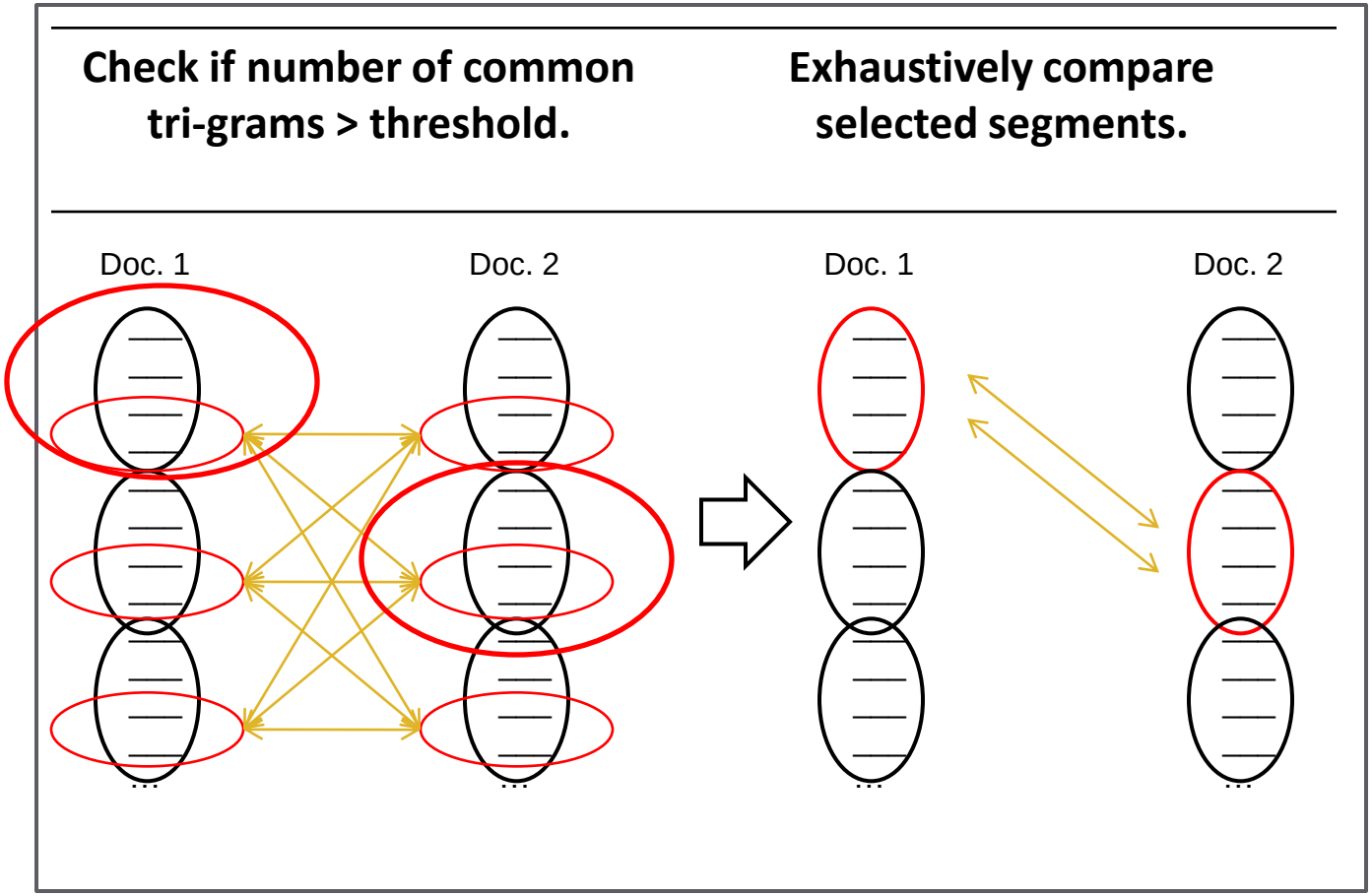
1 → Reducing the Search Space



1 → Reducing the Search Space



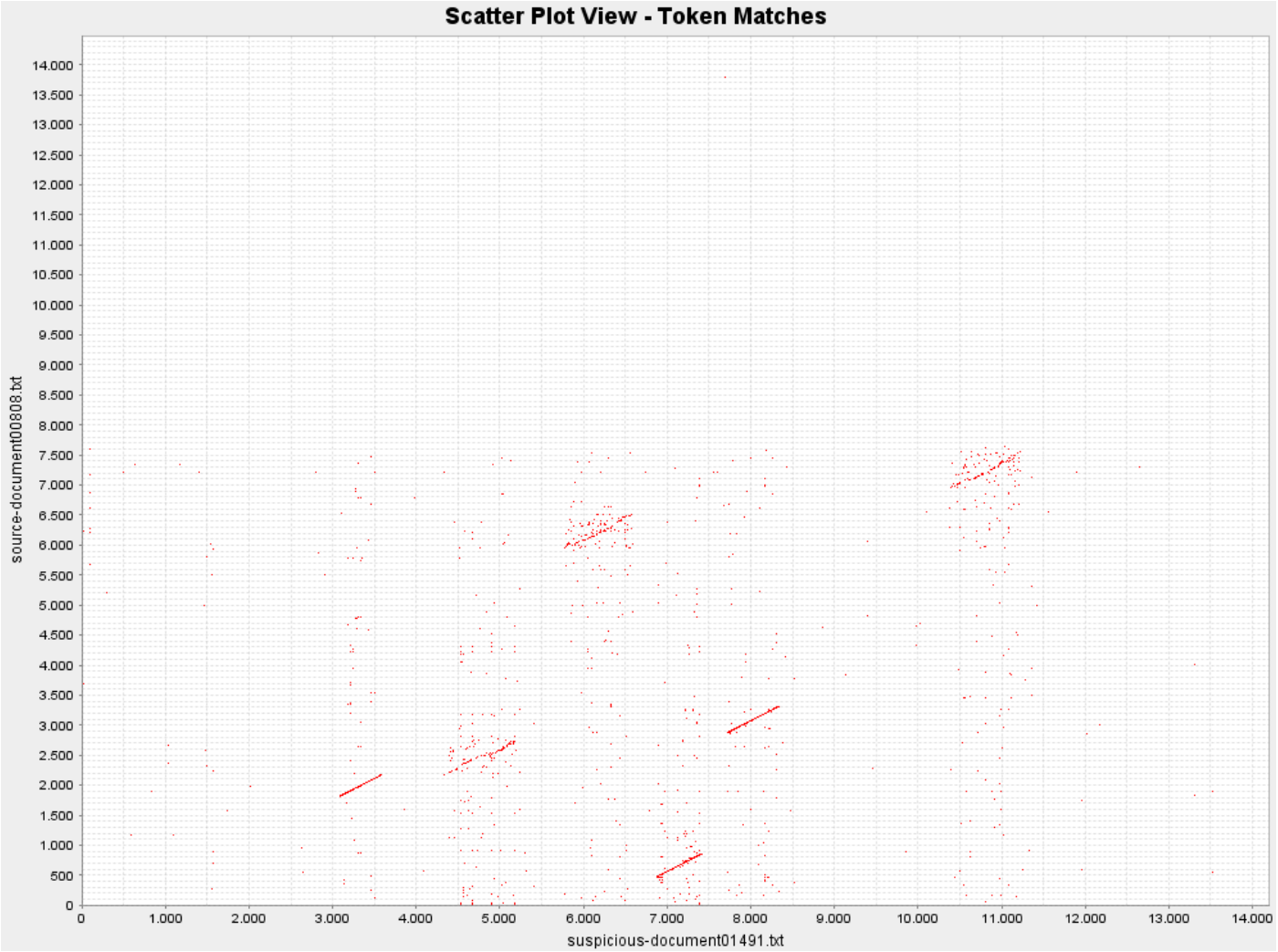
1 → Reducing the Search Space



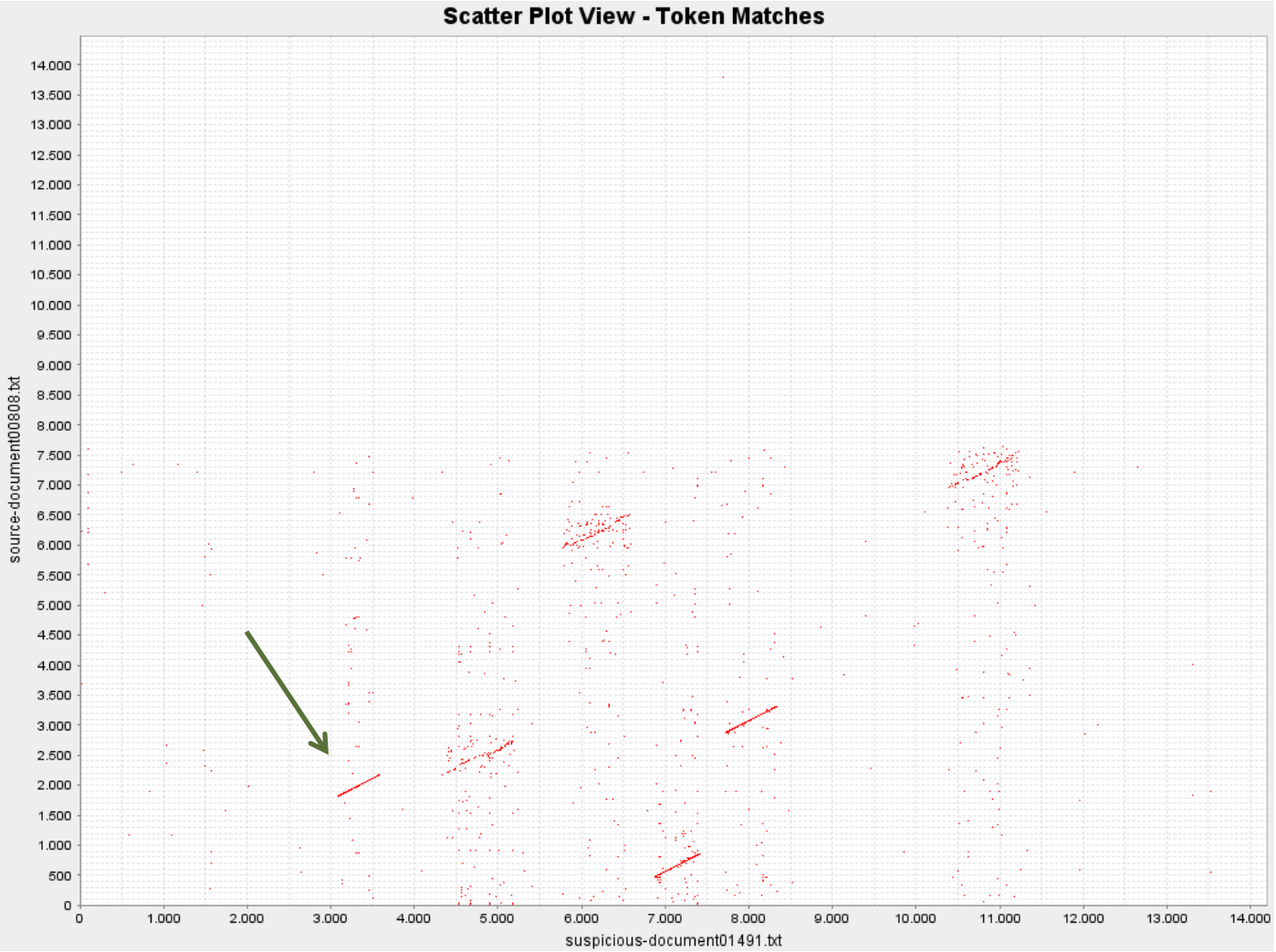
2 → Finding Plagiarized Passages

- Based (again) on word tri-grams and word bi-grams.
- Preprocessing: a-z characters considered only.
- For each word the corresponding offset is indexed.
- Based on three considerations:
  1) *Find a plagiarism "lead" or hint.*
  2) *Consider the possible size of the plagiarism case.*
  3) *Consider different obfuscation levels.*

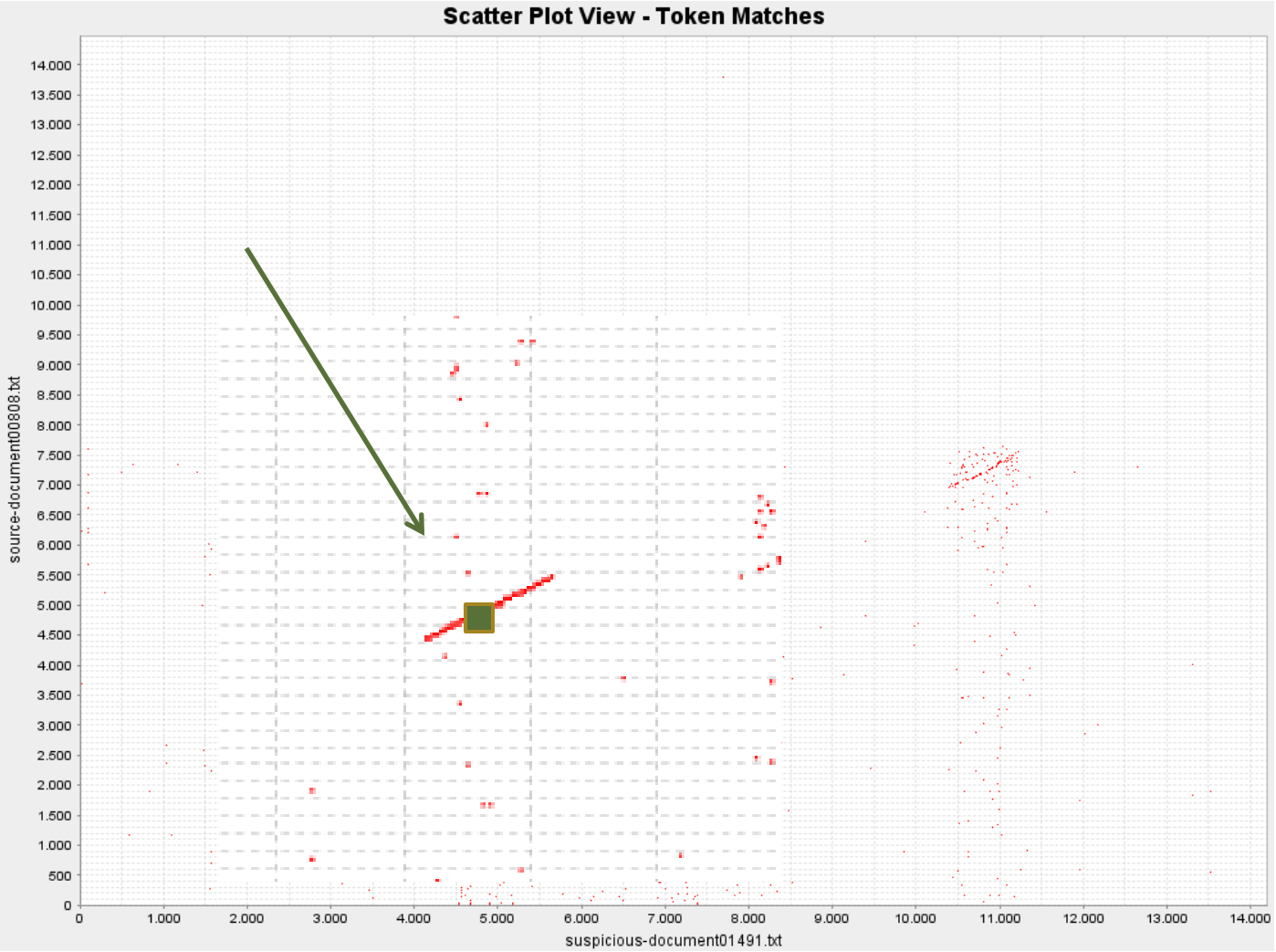
2 → Finding Plagiarized Passages



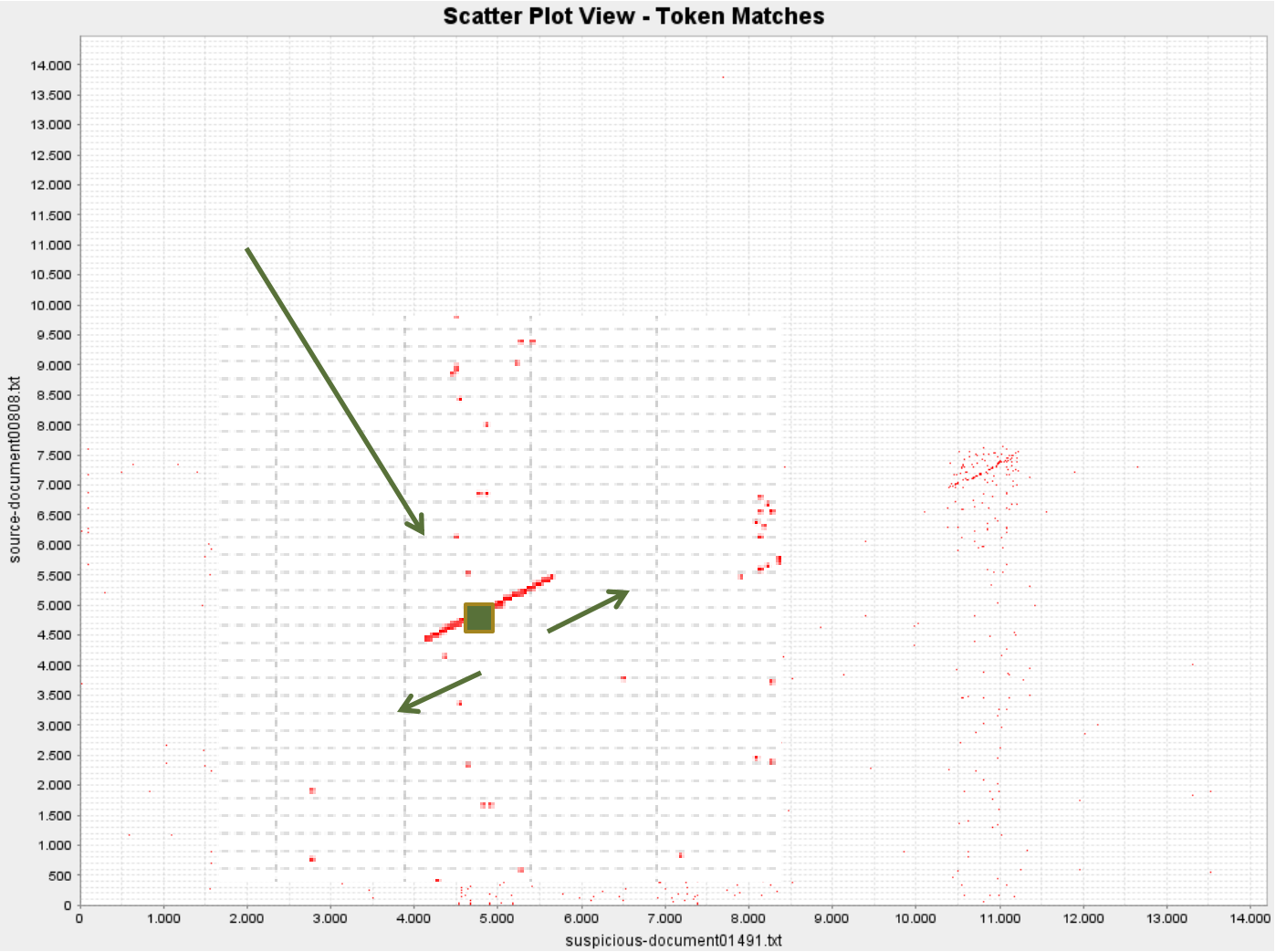
2 → Finding Plagiarized Passages



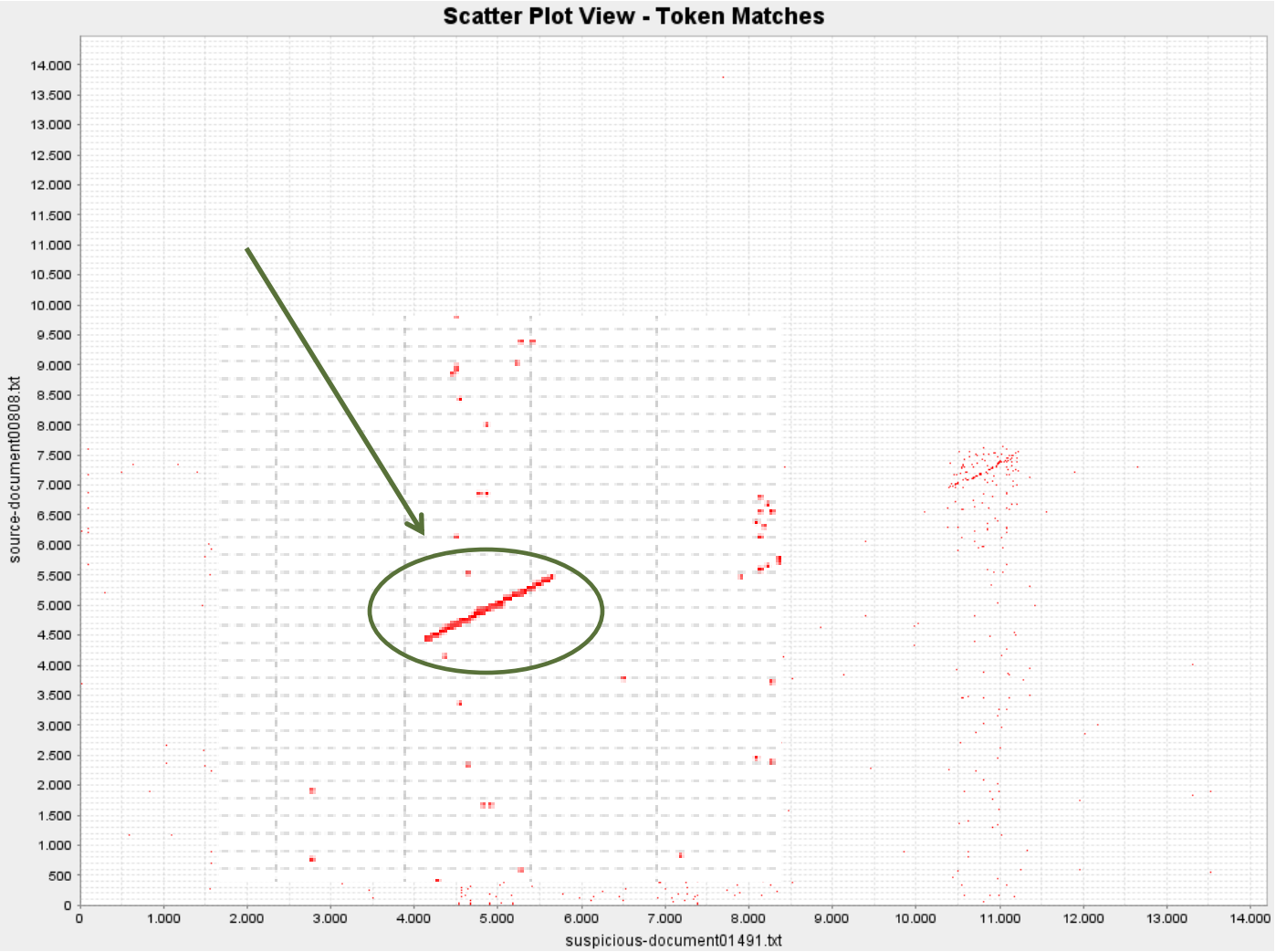
2 → Finding Plagiarized Passages



2 → Finding Plagiarized Passages



2 → Finding Plagiarized Passages



Experiment for 1 Step

Experiment presented for **first algorithm:** Reducing the Search Space.

The idea: prove the benefits of selecting samples over using *all* word n-grams for classifying a pair of documents.

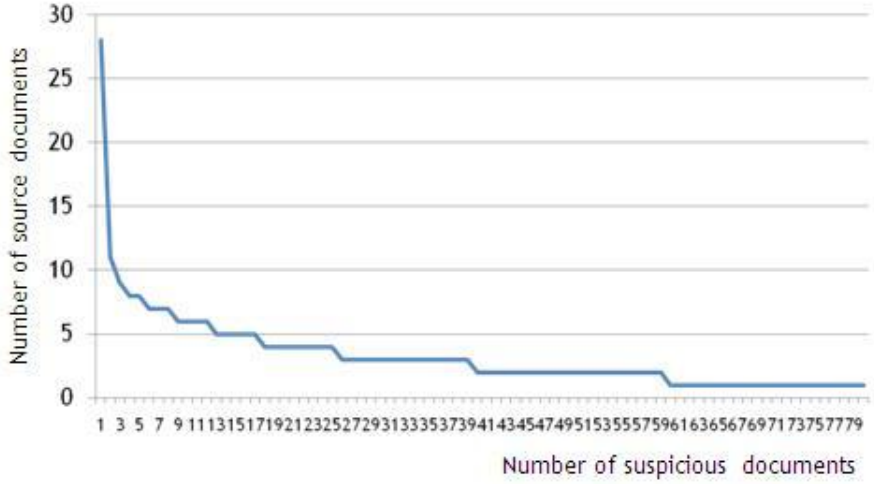
Two variants for **benchmark:**

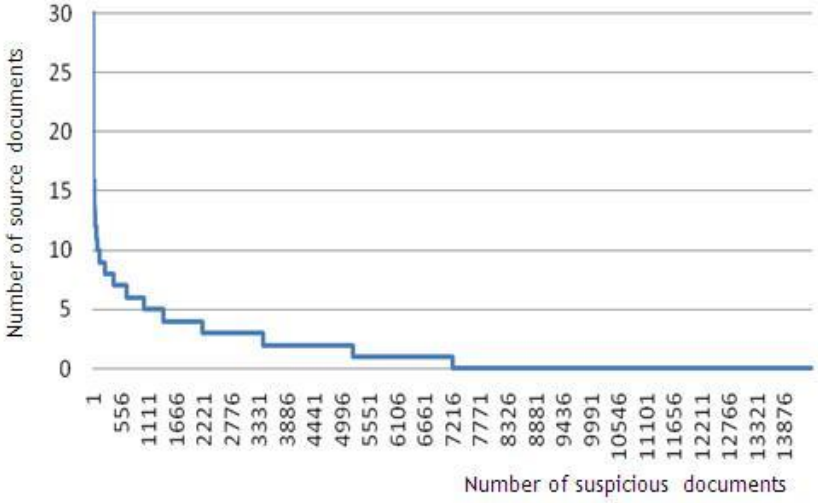
1) Exhaustive

2) Samples selected based on *TF*

Experiment for 1 Step

- Experimental dataset: 80 suspicious documents and 500 source documents from PAN'09 training corpus.
- Selected in order to preserve # of sources per suspicious document.





Results

Copy Detector	Accuracy	Precision	Recall	F-measure	runtime (s)
Exhaustive 0	**0,998**	**0,895**	0,914	**0,904**	20568
Exhaustive 1	0,990	0,616	**0,958**	0,750	21103
Exhaustive 2	0,961	0,882	0,916	0,899	29655
SimTF 0	0,874	0,824	0,821	0,823	6959
SimTF 1	0,923	0,766	0,800	0,783	7451
SimTF 2	0,874	0,836	0,818	0,827	6615
SimAR 0	0,887	0,865	0,856	0,861	5393
SimAR 1	0,899	0,859	0,852	0,855	5596
SimAR 2	0,849	0,828	0,868	0,847	**5231**

Results

Copy Detector	Accuracy	Precision	Recall	F-measure	runtime (s)
Exhaustive 0	**0,998**	**0,895**	0,914	**0,904**	20568
Exhaustive 1	0,990	0,616	**0,958**	0,750	21103
Exhaustive 2	0,961	0,882	0,916	0,899	29655
SimTF 0	0,874	0,824	0,821	0,823	6959
SimTF 1	0,923	0,766	0,800	0,783	7451
SimTF 2	0,874	0,836	0,818	0,827	6615
SimAR 0	0,887	0,865	0,856	0,861	5393
SimAR 1	0,899	0,859	0,852	0,855	5596
SimAR 2	0,849	0,828	0,868	0,847	**5231**

Results

Copy Detector	Accuracy	Precision	Recall	F-measure	runtime (s)
Exhaustive 0	**0,998**	**0,895**	0,914	**0,904**	20568
Exhaustive 1	0,990	0,616	**0,958**	0,750	21103
Exhaustive 2	0,961	0,882	0,916	0,899	29655
SimTF 0	0,874	0,824	0,821	0,823	6959
SimTF 1	0,923	0,766	0,800	0,783	7451
SimTF 2	0,874	0,836	0,818	0,827	6615
SimAR 0	0,887	0,865	0,856	0,861	5393
SimAR 1	0,899	0,859	0,852	0,855	5596
SimAR 2	0,849	0,828	0,868	0,847	**5231**

Results

Copy Detector	Accuracy	Precision	Recall	F-measure	runtime (s)
Exhaustive 0	**0,998**	**0,895**	0,914	**0,904**	20568
Exhaustive 1	0,990	0,616	**0,958**	0,750	21103
Exhaustive 2	0,961	0,882	0,916	0,899	29655
SimTF 0	0,874	0,824	0,821	0,823	6959
SimTF 1	0,923	0,766	0,800	0,783	7451
SimTF 2	0,874	0,836	0,818	0,827	6615
SimAR 0	0,887	0,865	0,856	0,861	5393
SimAR 1	0,899	0,859	0,852	0,855	5596
SimAR 2	0,849	0,828	0,868	0,847	**5231**

Results

Copy Detector	Accuracy	Precision	Recall	F-measure	runtime (s)
Exhaustive 0	**0,998**	**0,895**	0,914	**0,904**	20568
Exhaustive 1	0,990	0,616	**0,958**	0,750	21103
Exhaustive 2	0,961	0,882	0,916	0,899	29655
SimTF 0	0,874	0,824	0,821	0,823	6959
SimTF 1	0,923	0,766	0,800	0,783	7451
SimTF 2	0,874	0,836	0,818	0,827	6615
SimAR 0	0,887	0,865	0,856	0,861	5393
SimAR 1	0,899	0,859	0,852	0,855	5596
SimAR 2	0,849	0,828	0,868	0,847	**5231**

~4x speed bump.

General Results @ PAN

Overall Results

Precision:	0.85
Recall:	0.48
F-measure:	0.61
Granularity:	1.01
Overall:	0.61

- Comparison computed on two eight-core Servers, each with 6 GB of RAM.
- Java Implementation.
- Reducing Search Space:

  ~20 Hours.
- Finding Plagiarized Passages:

  ~12 Hours.

Conclusions

- Word bi-grams and word tri-grams are good tokens for plagiarism uncovering.
- Search Space Reduction adapted to plagiarism.
- Sampling to speed up the process: 4x on current corpus, at little cost of recall.
- Some obfuscated plagiarism left undetected.

Future Work

- Grid optimization over the parameters.
- Further investigation of *sampling* strategies for document fingerprinting.
- Synonym consideration for highly obfuscated plagiarism cases.
- Use of char n-grams with this approach for intrinsic plagiarism detection.
- Cross-language consideration.
- Other ways to increase compute time.

FASTDOCODE: Finding Approximated Segments of N-Grams for Document Copy Detection

Lab Report for PAN at CLEF 2010

Gabriel Oberreuter
goberreu@ing.uchile.cl
Universidad de Chile - September 2010

Group members : Gastón L'Huillier, Sebastián Ríos and Juan D. Velásquez