

D-Scribe - Automatic Authorship Identification and Clustering

Measurement Science and Standards in Forensic Handwriting Analysis

D-Scribe – Automatic Authorship Identification

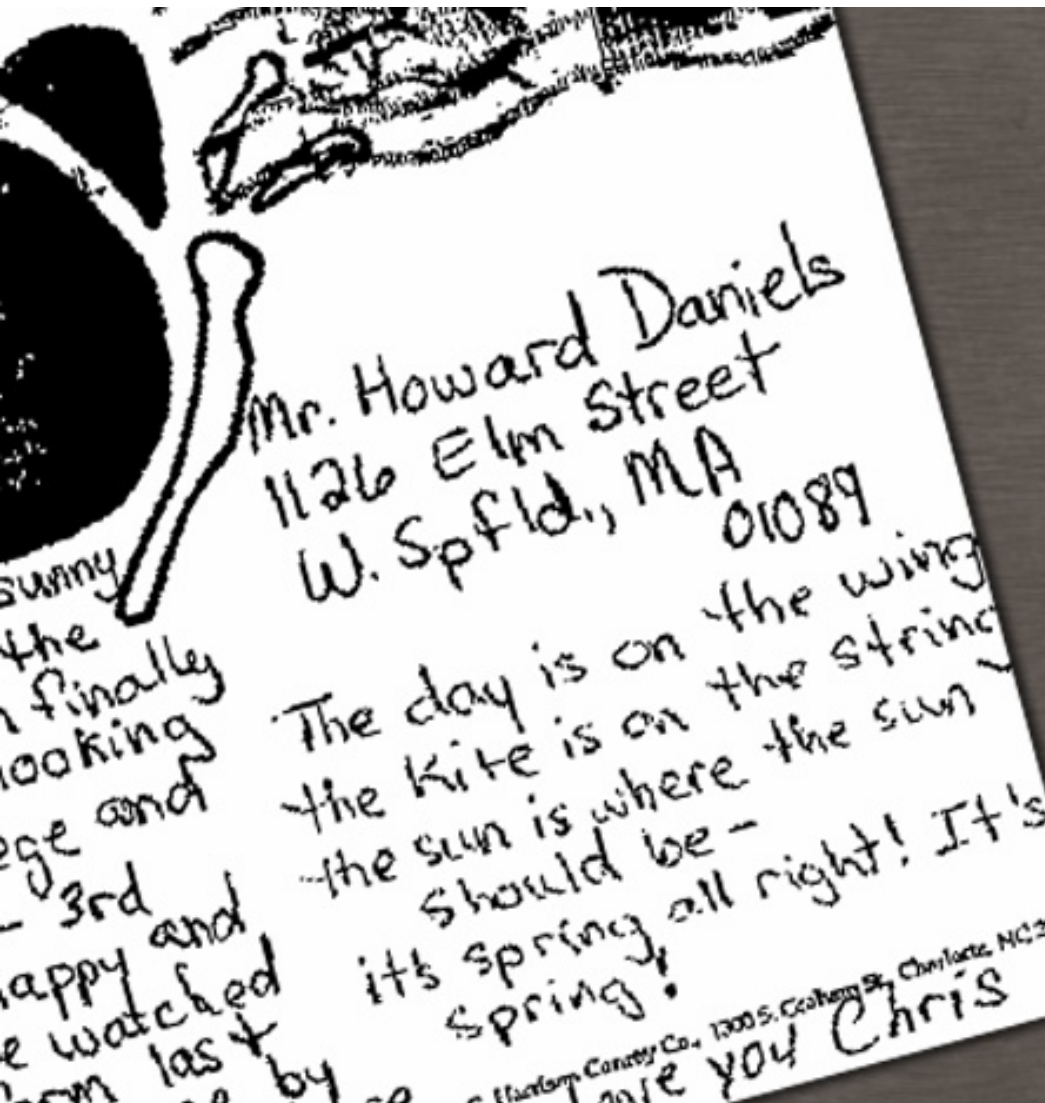
Table of content

The logo for D-Scribe, featuring a stylized black 'D' with a white pen nib inside, followed by the word 'Scribe' in a black, italicized serif font.A close-up photograph of a hand holding a black fountain pen, writing the word 'ME' in blue ink on a white background. The letters are written on a set of three horizontal blue lines. The word 'ME' is written in a cursive, handwritten style.

- Postal roots
- Global OCR
- Use cases
- Architecture
- Structured feature extractor
- Textual feature extractor
- Allograph feature extractor
- Feature analysis
- Cluster analysis
- Potential applications

D-Scribe – Automatic Authorship Identification

Postal roots

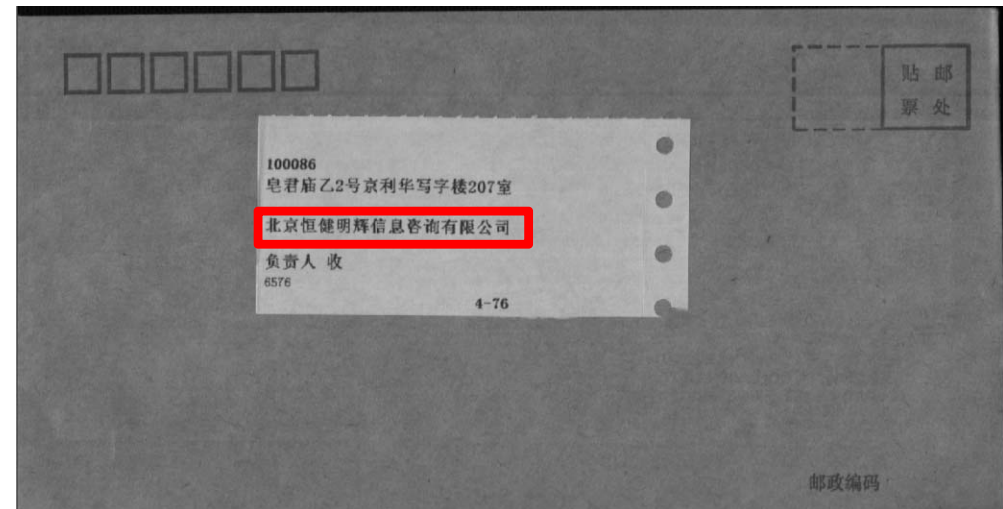
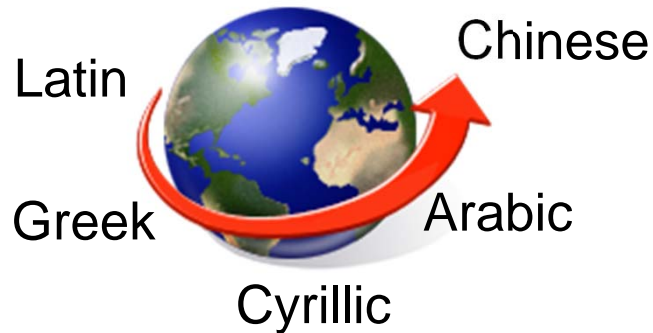


Postal recognition tasks represent a difficult proving ground

- Siemens reads everything on a letter in <1 second
- 150 Billion mail pieces per year

D-Scribe – Automatic Authorship Identification

Global OCR



Region of Interest

北京市朝阳区望京广顺南大街嘉润花园19号写字楼B座2层

Recognition results of chars

北京市朝阳区望京广顺南大街嘉润花园19号写字楼B座2层

Interpretation results

(City, district, road, block, number), (Building, house, floor)

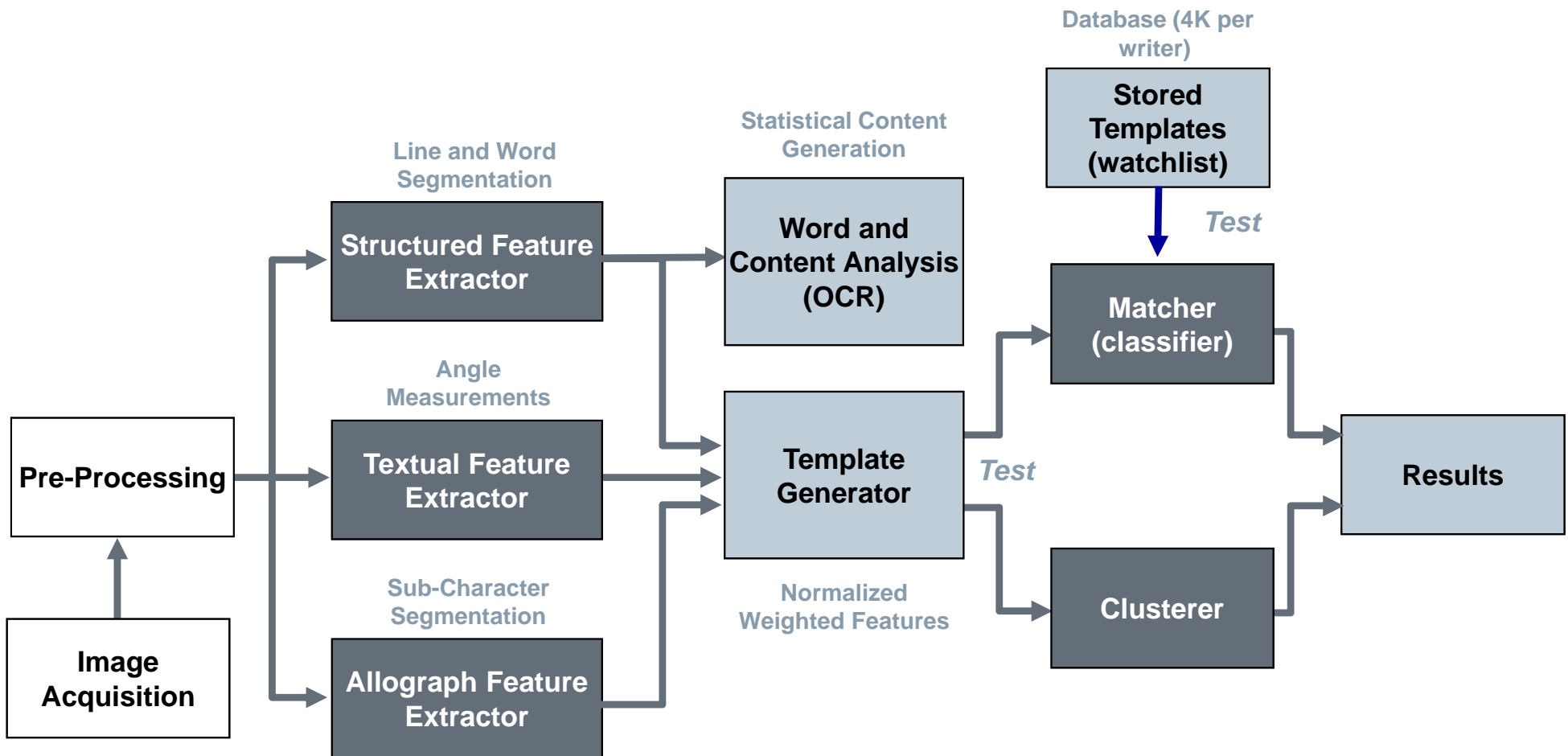
D-Scribe – Automatic Authorship Identification

Use cases

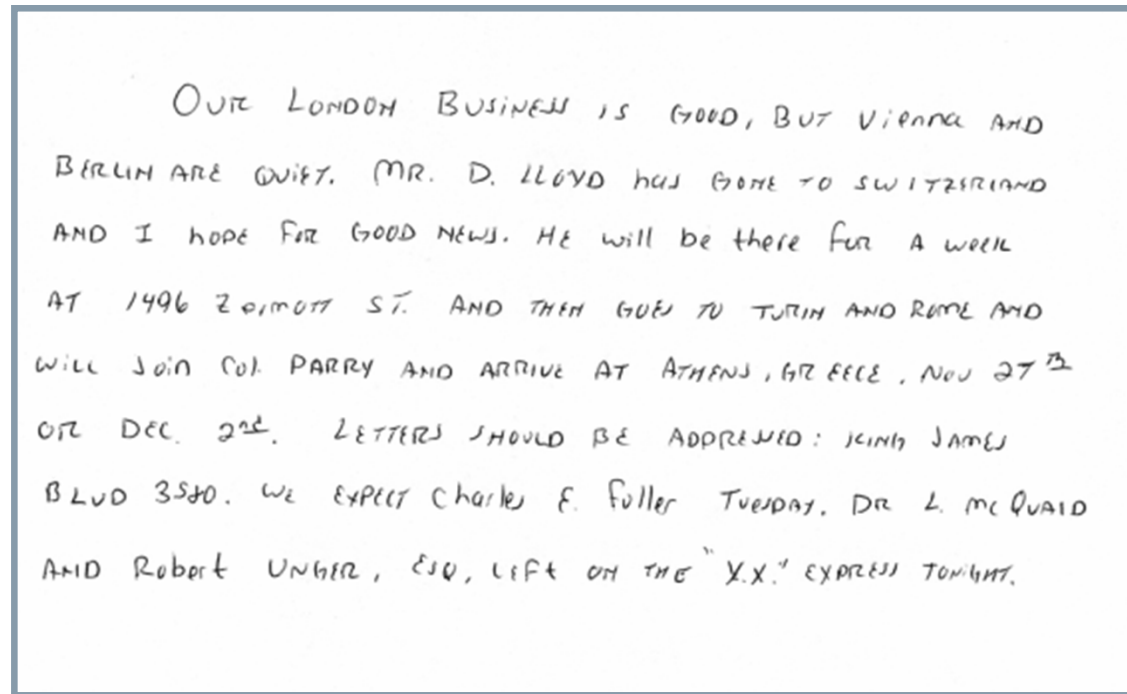
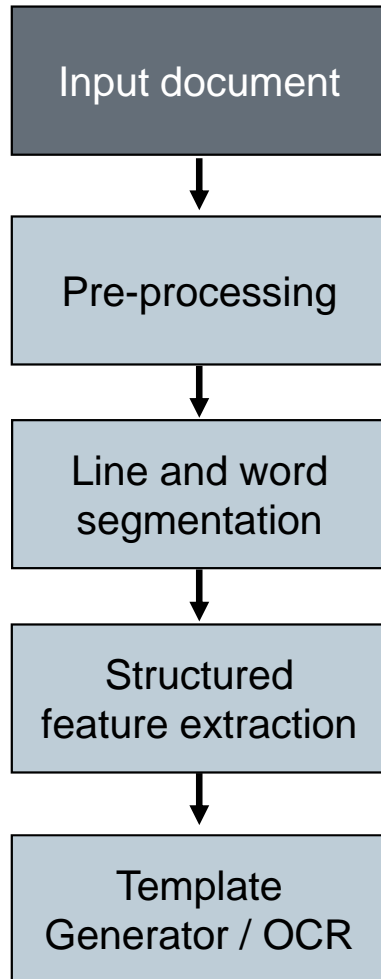
Questions to be answered by biometric handwriting detection

- Which handwriting from a group is similar to a given handwriting sample?
- How similar are two handwriting samples?
- How many authors wrote a set of documents, and how can these be organized by author?

D-Scribe – Automatic Authorship Identification Architecture

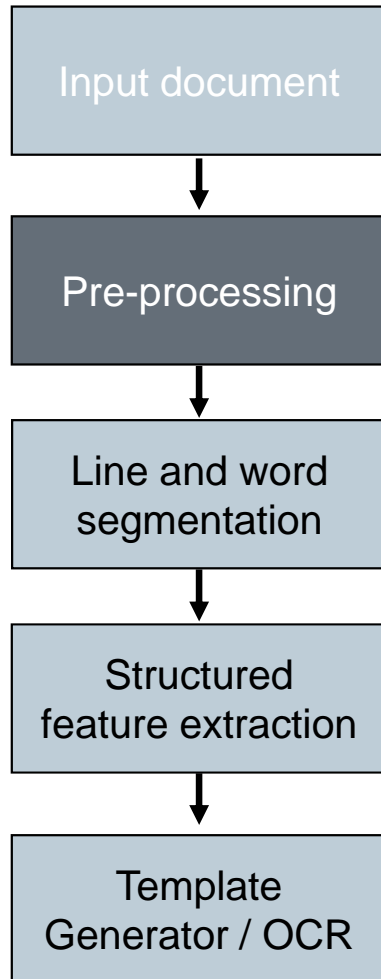


D-Scribe – Automatic Authorship Identification Structured Feature Extractor (1/4)



Input document (200 DPI gray)

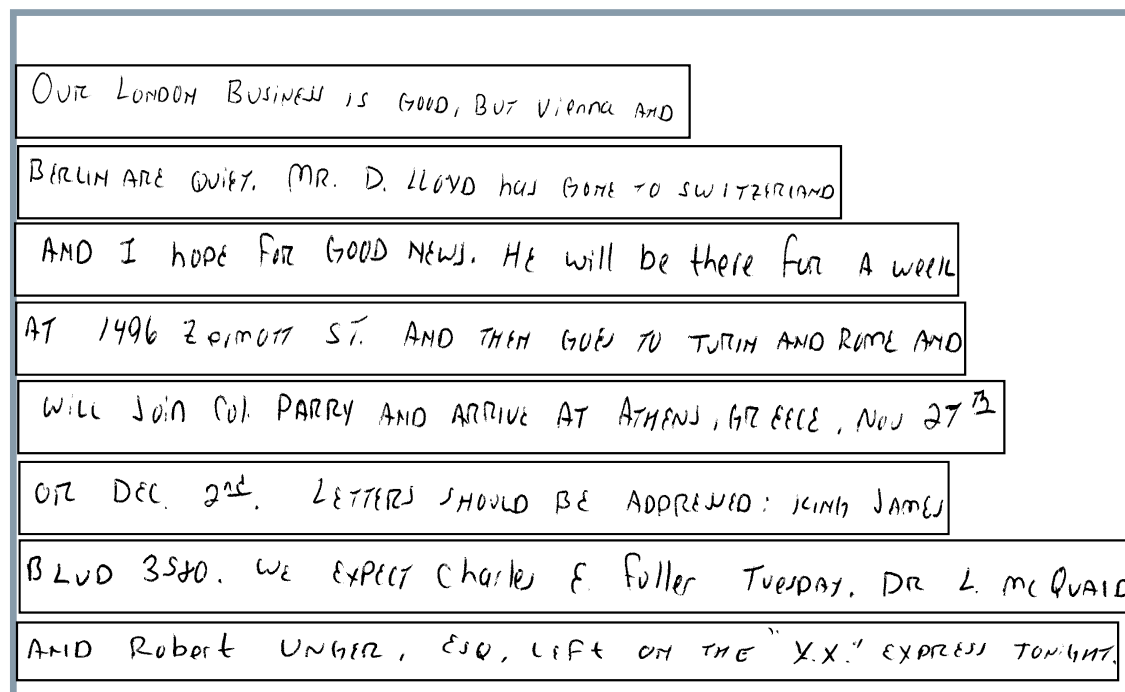
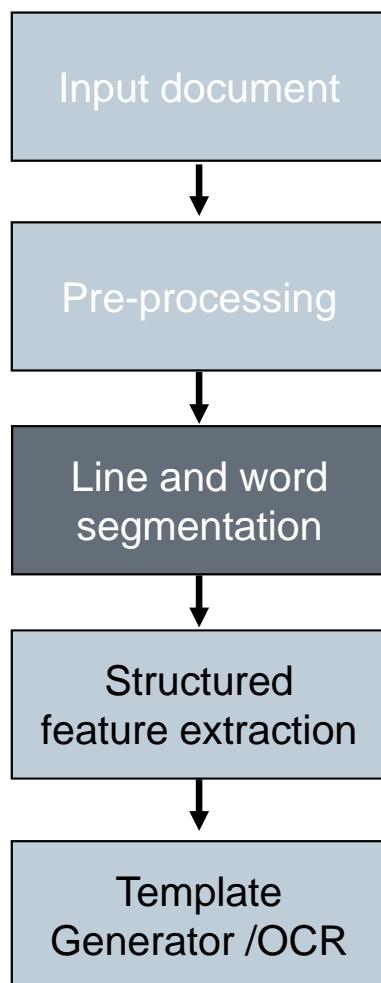
D-Scribe – Automatic Authorship Identification Structured Feature Extractor (2/4)



OUT LONDON BUSINESS IS GOOD, BUT VIENNA AND
BERLIN ARE QUIET. MR. D. LLOYD HAS GONE TO SWITZERLAND
AND I HOPE FOR GOOD NEWS. HE WILL BE THERE FOR A WEEK
AT 1496 ZEMOTT ST. AND THEN GOES TO TURIN AND ROME AND
WILL JOIN COL. PARRY AND ARRIVE AT ATHENS, GREECE, NOV 27TH
OR DEC. 2ND. LETTERS SHOULD BE ADDRESSED: JOHN JAMES
BLVD 3540. WE EXPECT CHARLES E. FULLER TUESDAY. DR. L. McQUAID
AND ROBERT UNGER, ESQ, LEFT ON THE "X.X." EXPRESS TONIGHT.

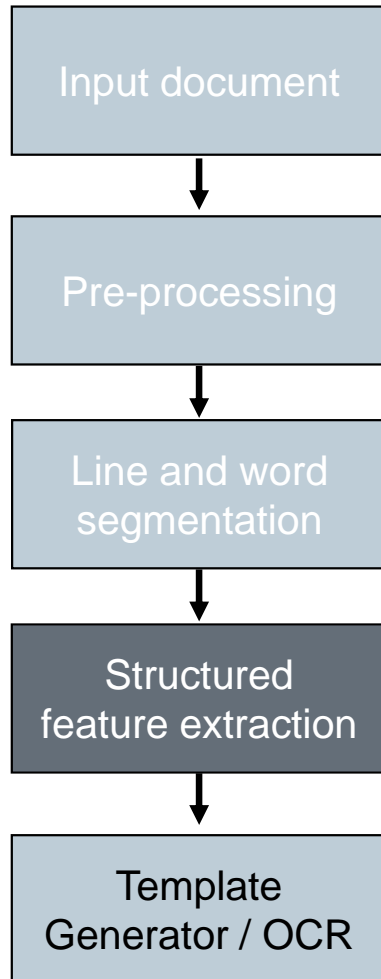
Binarized Document/Line Removal/Skew Correction

D-Scribe – Automatic Authorship Identification Structured Feature Extractor (3/4)



Segmented lines and words

D-Scribe – Automatic Authorship Identification Structured Feature Extractor (4/4)



```
W002-1 ./statistic/docfeature/n_lines 8
W002-1 ./statistic/docfeature/avg_height_of_line 69.000000
W002-1 ./statistic/docfeature/avg_length_of_line 2091.125000
W002-1 ./statistic/docfeature/avg_blackness_of_line 0.079064
W002-1 ./statistic/docfeature/avg_runlength_x_of_line 5.692302
W002-1 ./statistic/docfeature/avg_runlength_y_of_line 6.725820
W002-1 ./statistic/docfeature/avg_slope_of_line 0.012695
W002-1 ./statistic/docfeature/avg_segdist_of_line 28.944444
W002-1 ./statistic/docfeature/avg_n_words_per_line 3.125000
```

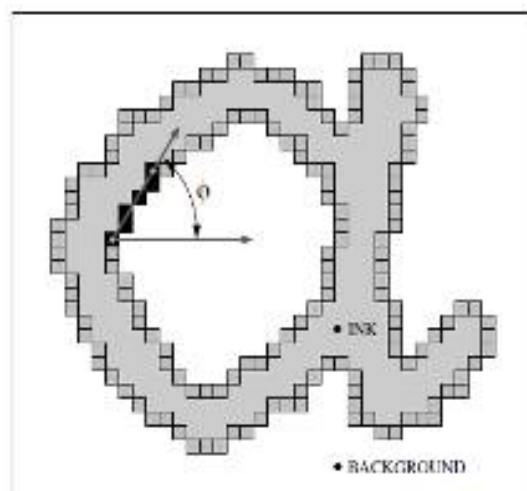
Line and word features

D-Scribe – Automatic Authorship Identification Textual Feature Extractor (1/3)

Pre processed image

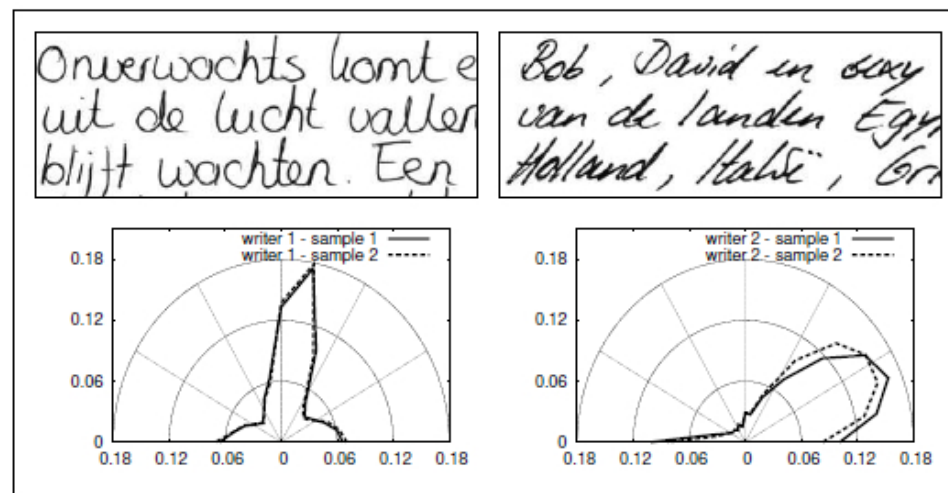


Contour direction comparison



Contour Direction PDF (Bulacu, 2007)

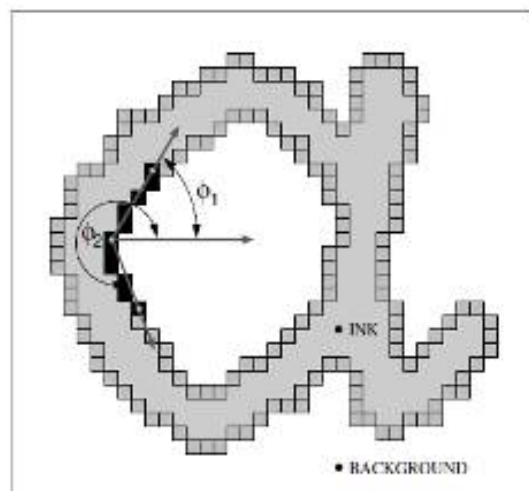
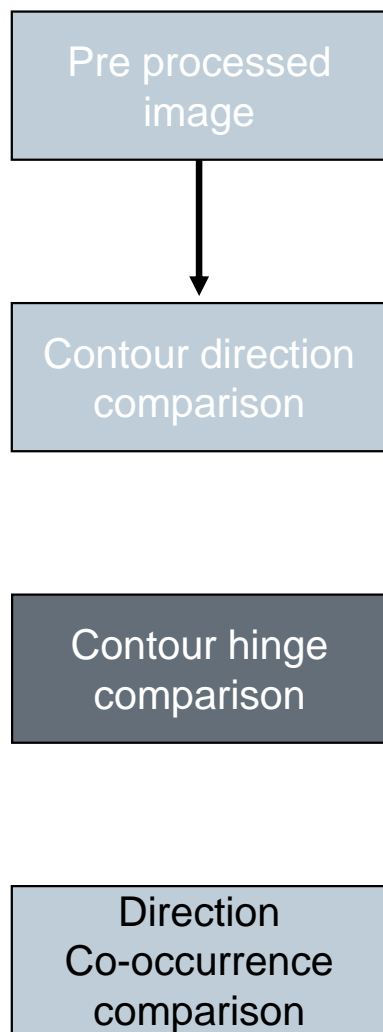
Contour hinge comparison



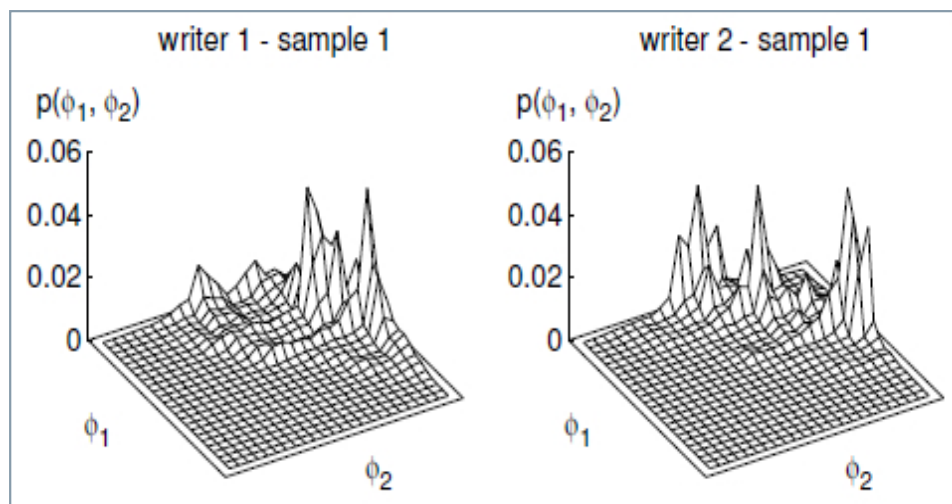
CD Histogram Comparison (Bulacu, 2007)

Direction Co-occurrence comparison

D-Scribe – Automatic Authorship Identification Textual Feature Extractor (2/3)

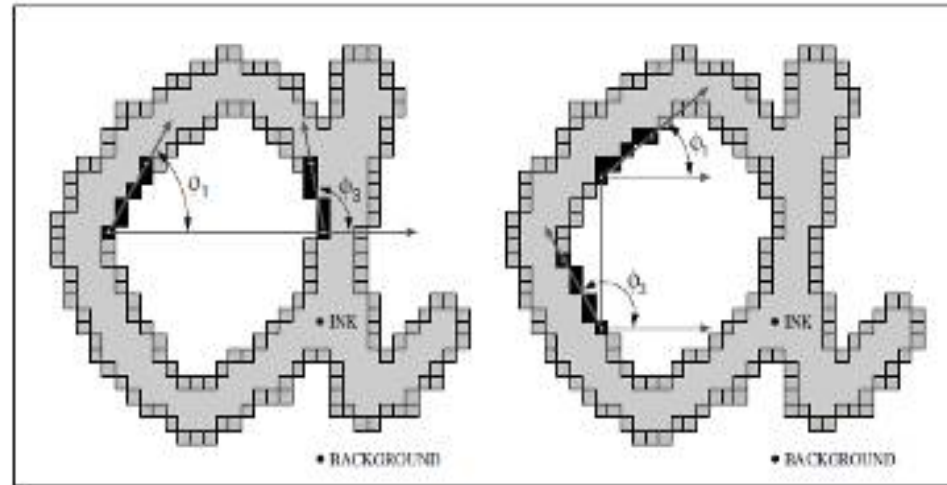
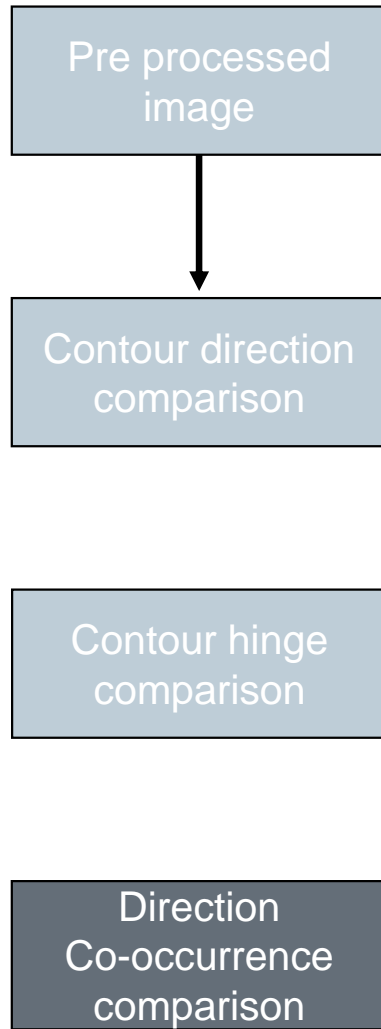


Contour Hinge PDF
(Bulacu, 2007)

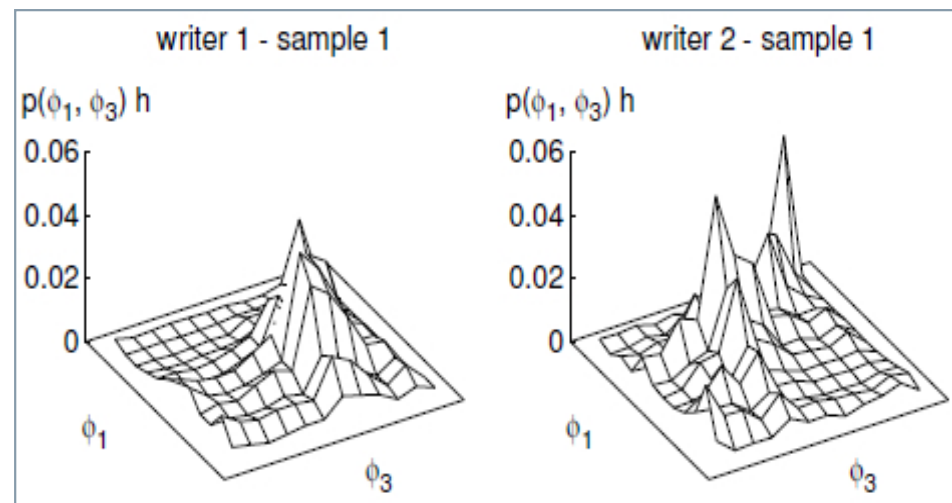


CH Histogram Comparison
(Bulacu, 2007)

D-Scribe – Automatic Authorship Identification Textual Feature Extractor (3/3)



Direction Co-Occurrence PDF (Bulacu, 2007)



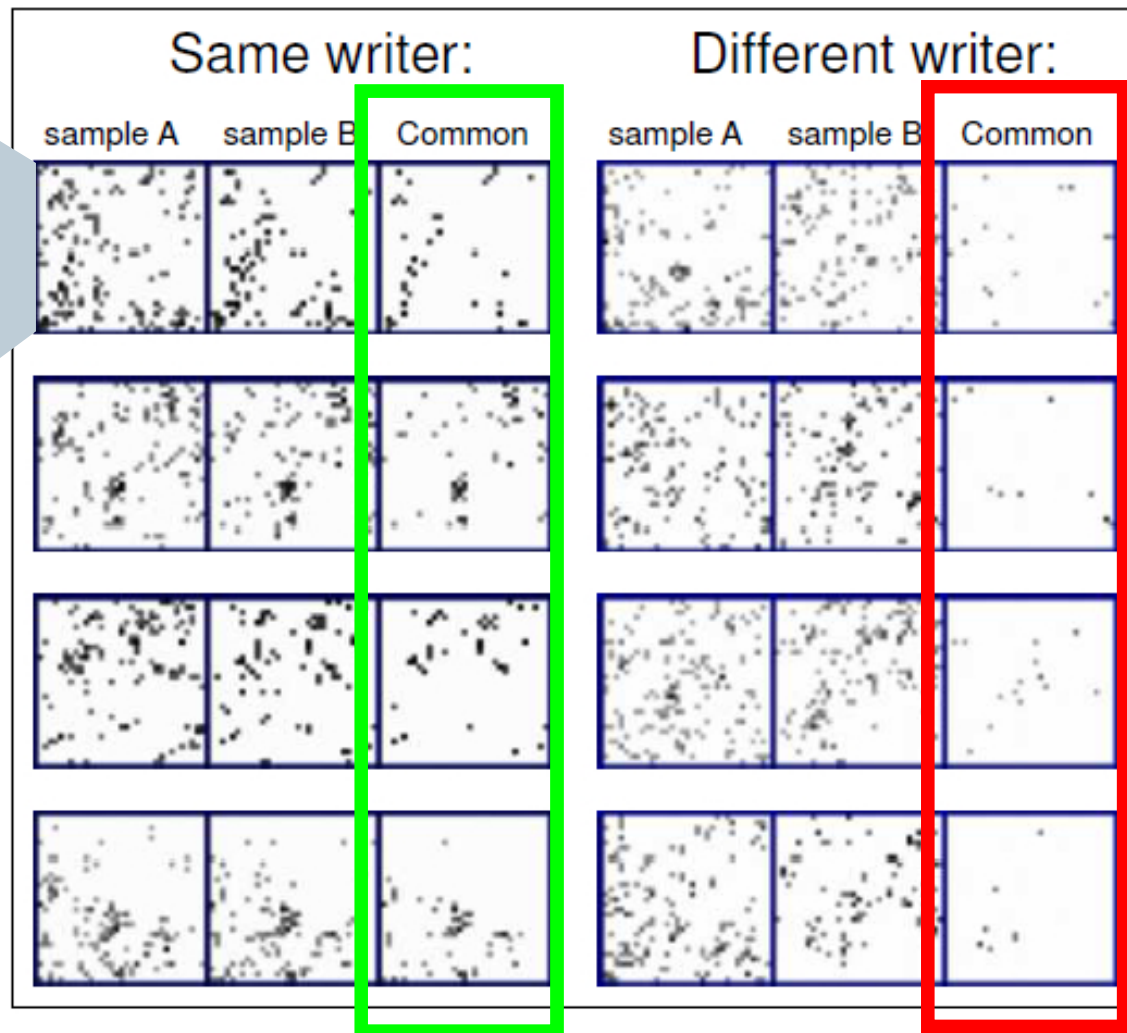
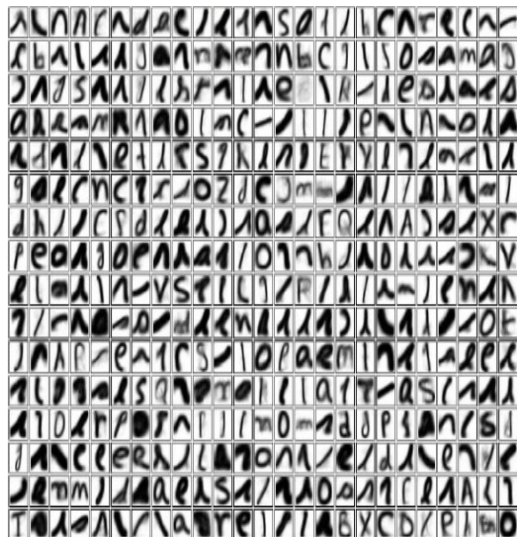
Direction COO Histogram Comparison (Bulacu, 2007)

D-Scribe – Automatic Authorship Identification Allograph Feature Extractor (1/2)

*„Mr. Powell finds it easier to take
childrens and sick people than to
industrie,“ Mr. Brown commented 'ci
full inquiry into the cost of drugs an*



D-Scribe – Automatic Authorship Identification Allograph Feature Extractor (2/2)



(Bulacu, 2007)

D-Scribe – Automatic Authorship Identification

Key features

- Uses advanced and proven image preprocessing algorithm basis
 - Image improvements
 - Text Line extraction
 - Hand/Machine decision
 - Advanced underline and noise removal

- Simple decisions and confidences
- Portable and efficient software
 - Small footprint
 - Low runtime
 - Enables large database analysis
 - Android port available

D-Scribe – Feature analysis

Used test sets

Data Set	Language	Documents	Writers	Remarks
ENG	English	2535	507	5 documents per writer, 4 predefined texts + an arbitrary text
ARAB_1	Arabic	5000	45	50-55 documents per writer, large variety of documents (different background, different pens, artificial documents)
ARAB_2	Arabic	1000 (including unknown writer docs)	200	2-3 documents per writer

D-Scribe – Feature analysis

Two feature configuration sets

■ 4 Features:

- grapheme_snn_split (500)
- grapheme_snn_points (150)
- hinge_improved_rotated_fragment (1536)
- hinge_improved_broi_textline (1536)

(feature vector size)

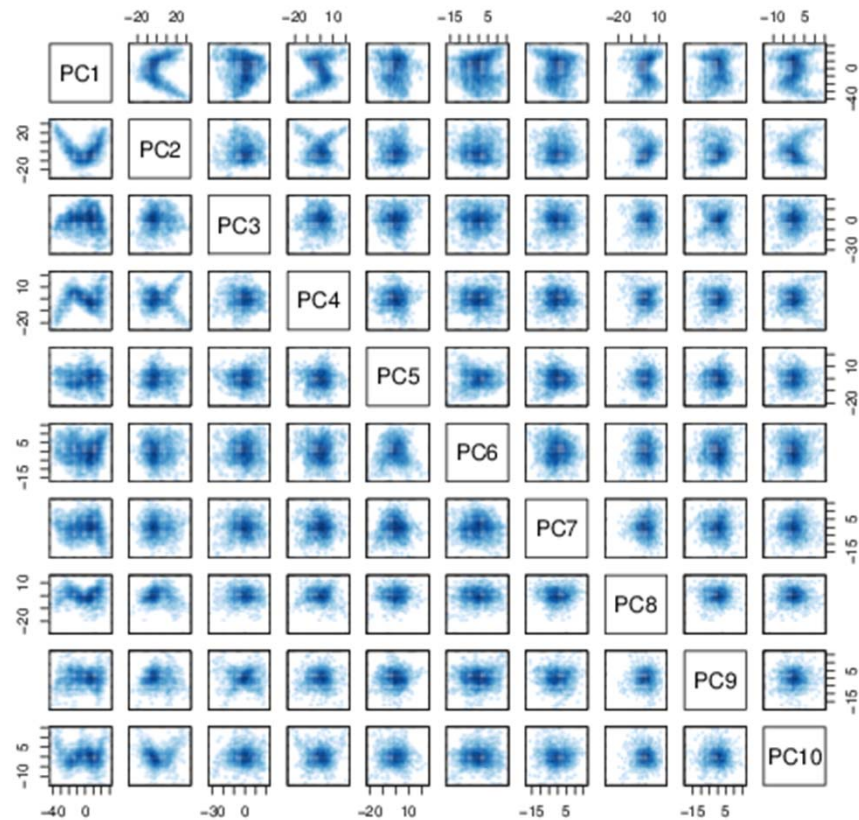
■ 6 Features:

- grapheme_snn_split (500)
- grapheme_snn_points (150)
- hinge_improved_rotated_fragment (1536)
- hinge_improved_broi_textline (1536)
- simple_writing_direction (12)
- hinge_contour_approximation (144)

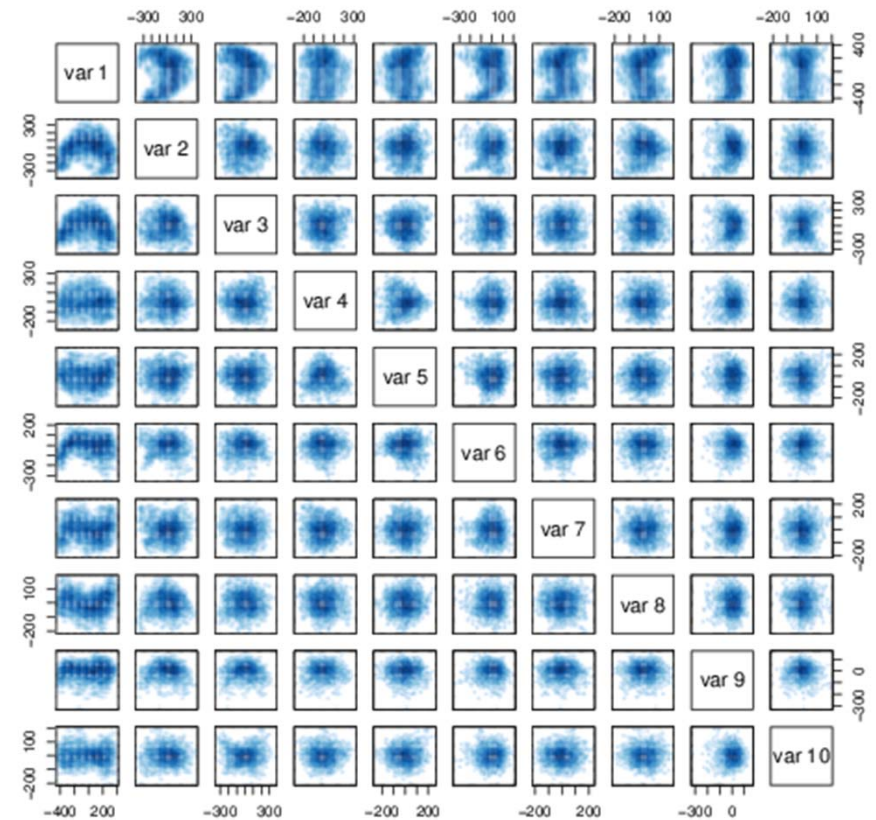
D-Scribe – Feature analysis

Different projections of the ENG data set

PCA



MDS



D-Scribe – Feature analysis

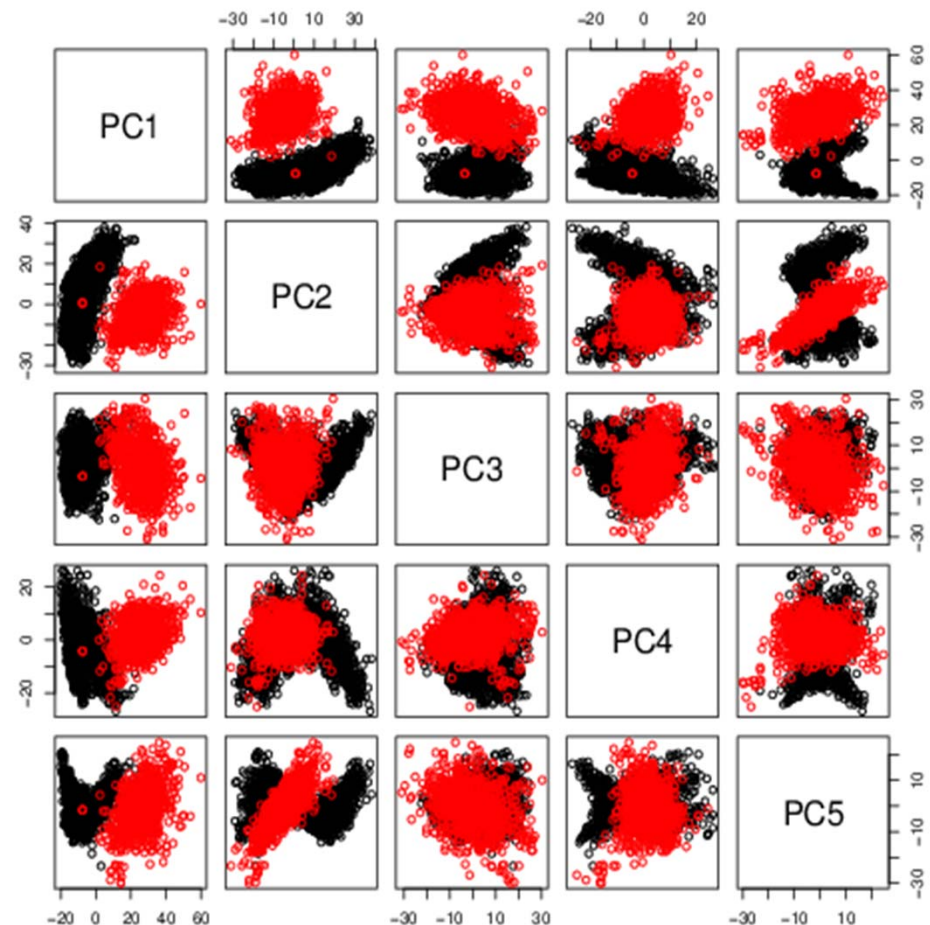
PCA Projections of the ENG and ARAB_2 Data Sets

Side effect:

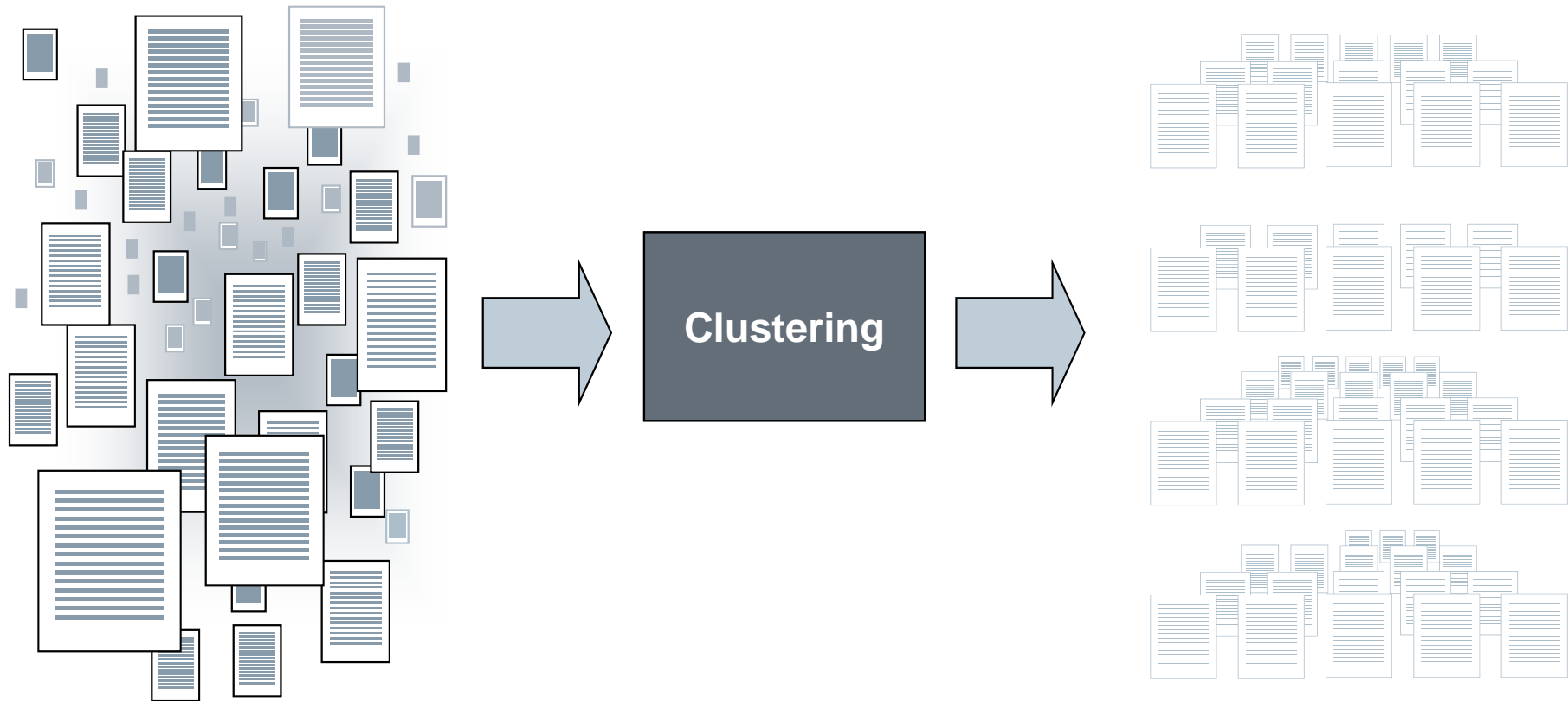
Features can be used to separate by writing system (e.g. Arabic language from Latin)

Black: Latin, Red: Arabic

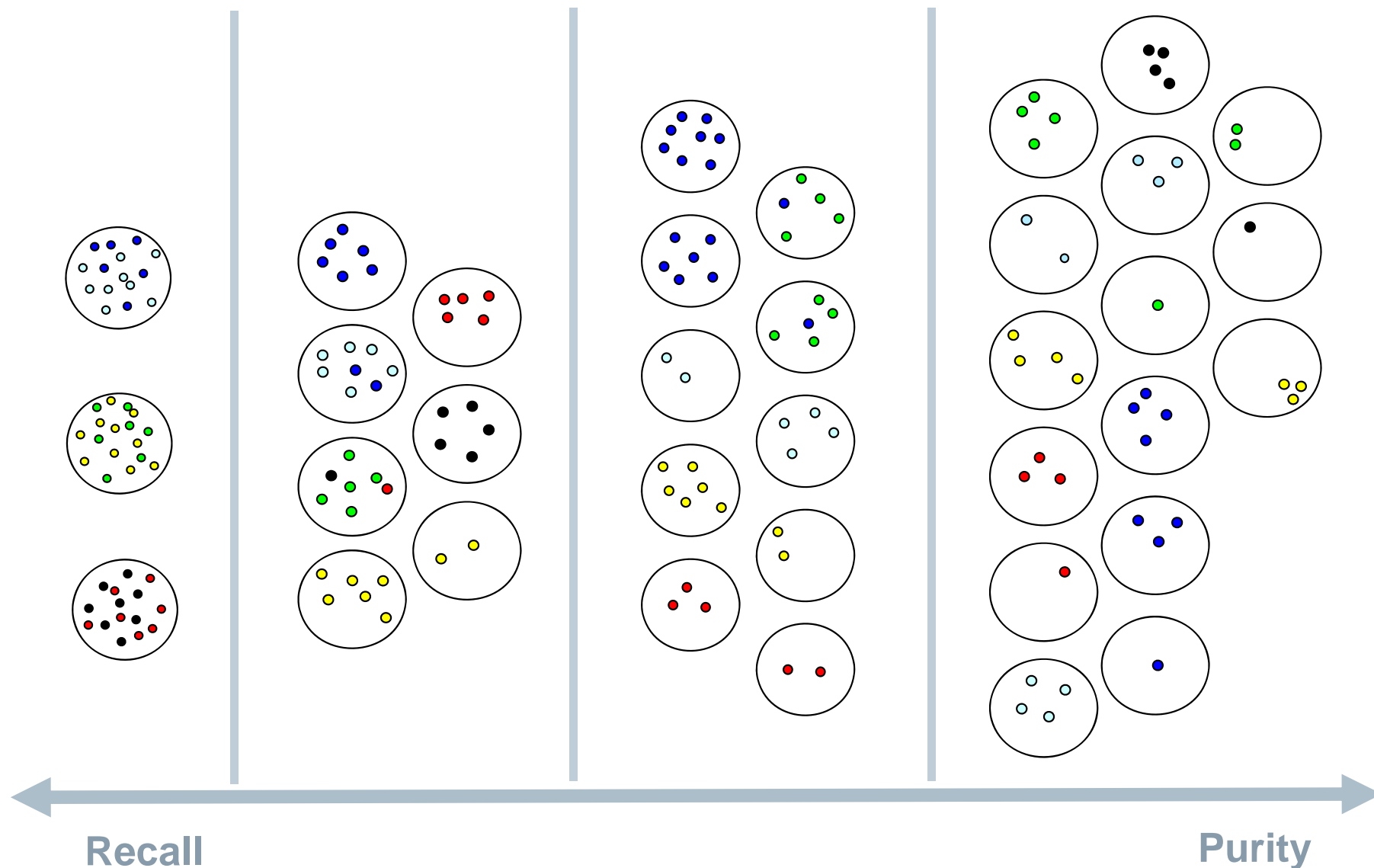
Works also with other data sets



Clustering



How effective was the clustering?



Final Cluster Quality Score

Purity is not as relevant as Recall, because the error putting different writers into same cluster is not as important as putting the same writer in different clusters

Set of elements: $S = \{o_1, \dots, o_n\}$ Ground Truth: $X = \{X_1, \dots, X_r\}$
 Cluster assignment: $Y = \{Y_1, \dots, Y_s\}$ Purity: $\text{purity} = \frac{1}{n} \sum_{y_i \in Y} \max_{x_j \in X} |y_i \cap x_j|$

Recall: $R = \left(\frac{|\{\text{pairs in same set in X and in Y}\}|}{|\{\text{pairs in same set in X}\}|} \right)$

Harmonic Mean of Purity and Recall

$$\text{score}_\beta = \left(1 + \beta^2\right) \frac{\text{purity} \cdot R}{\left(\beta^2 \cdot \text{purity}\right) + R} \qquad \text{score}_1 = 2 \frac{\text{purity} \cdot R}{\text{purity} + R}$$

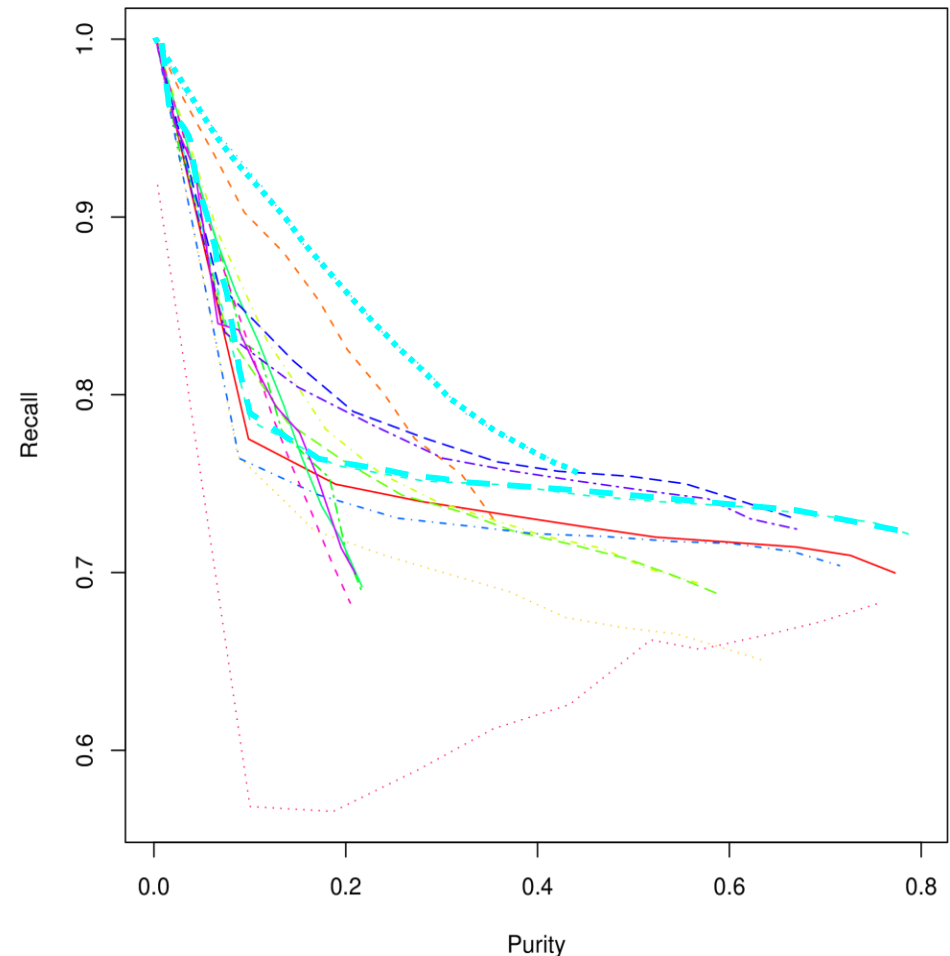
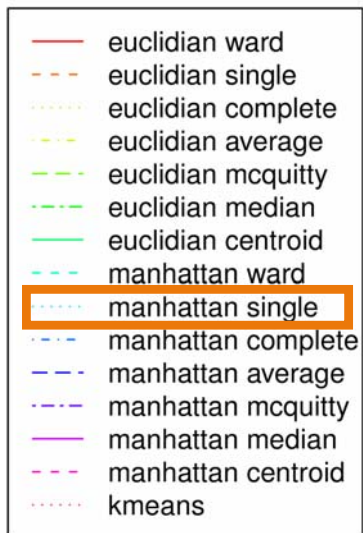
Recall is β times important than purity

Finding Optimal Cluster Method on the ENG Data Set with 6 feature sets

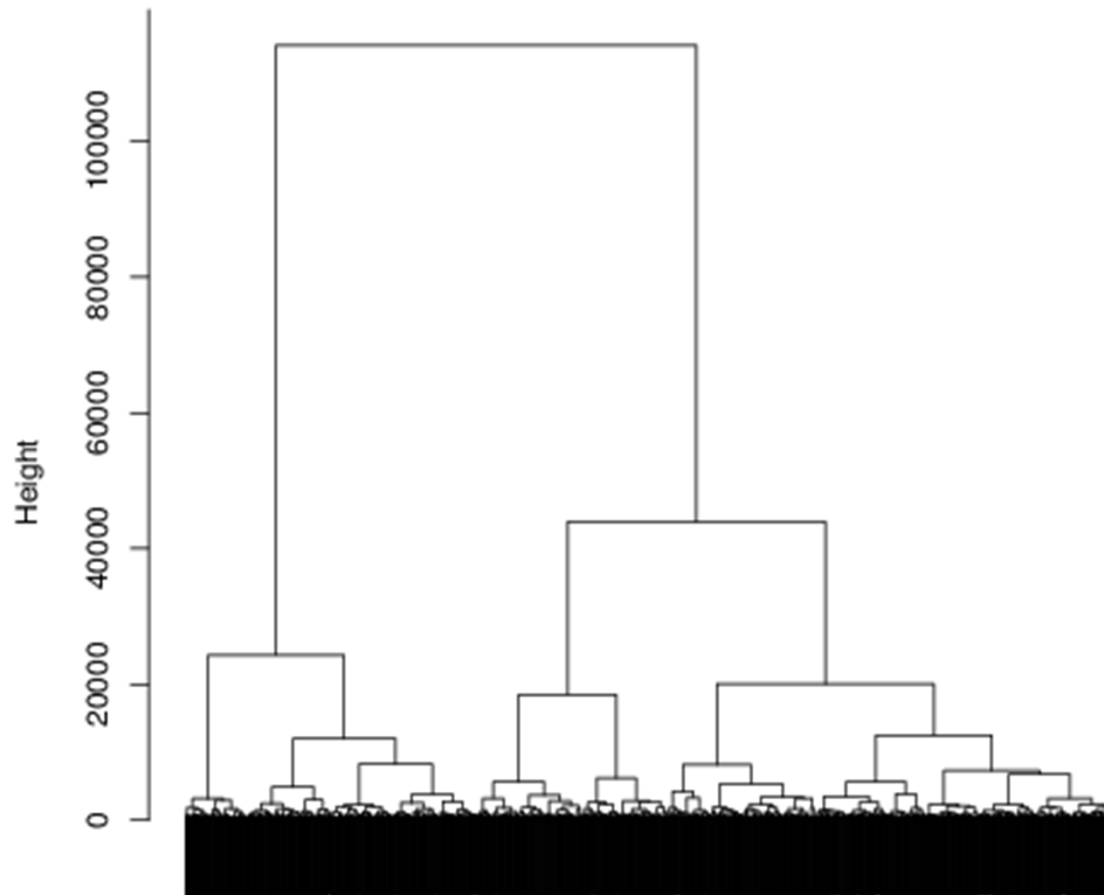
The clustering parameters are a trade off between recall and purity, requiring high recall automatically means low purity and vice versa

The manhattan distance performs in all cases better than the euclidian distance

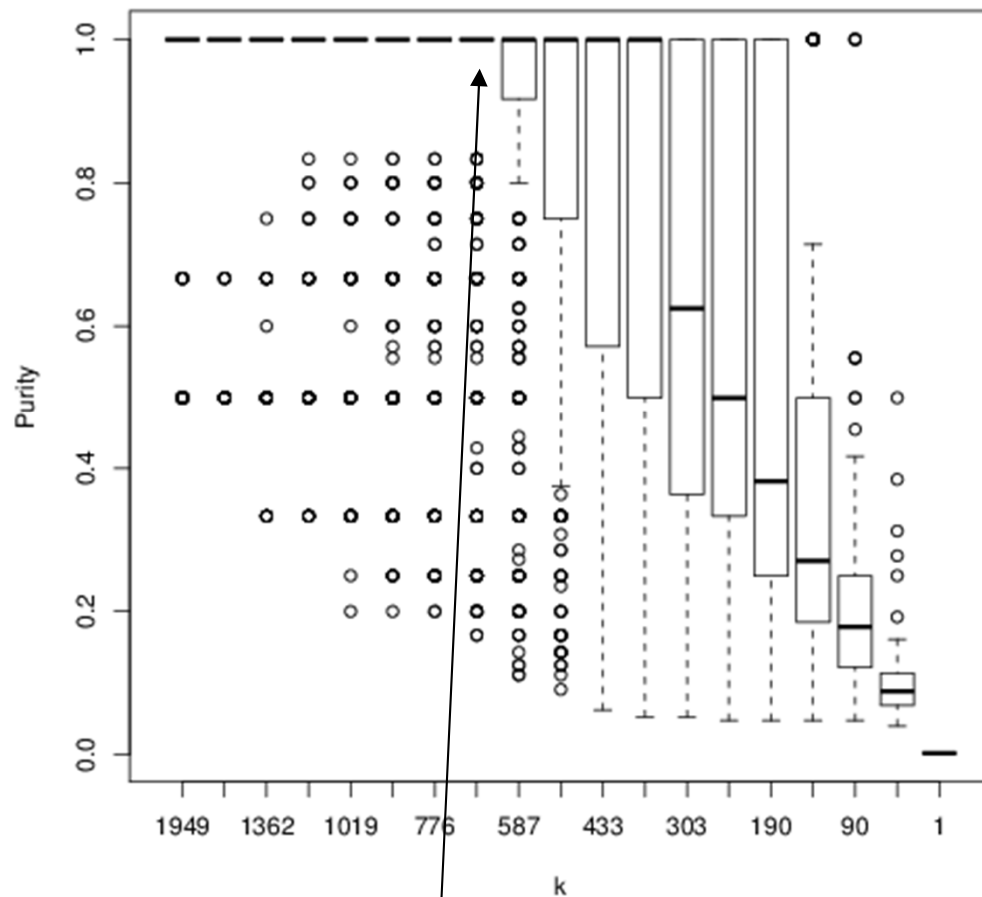
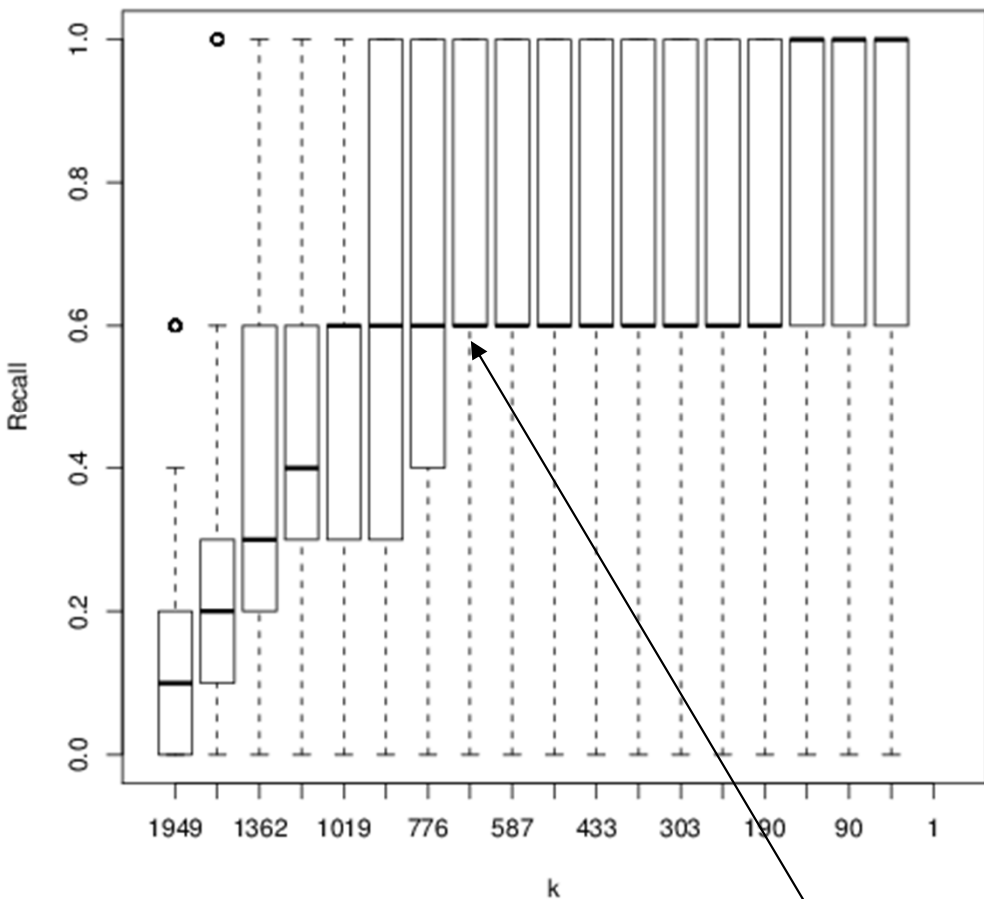
The different cluster methods have different behaviors, either single linkage or Ward's method are considered best



Dendrogram of Clustering with Ward's Method



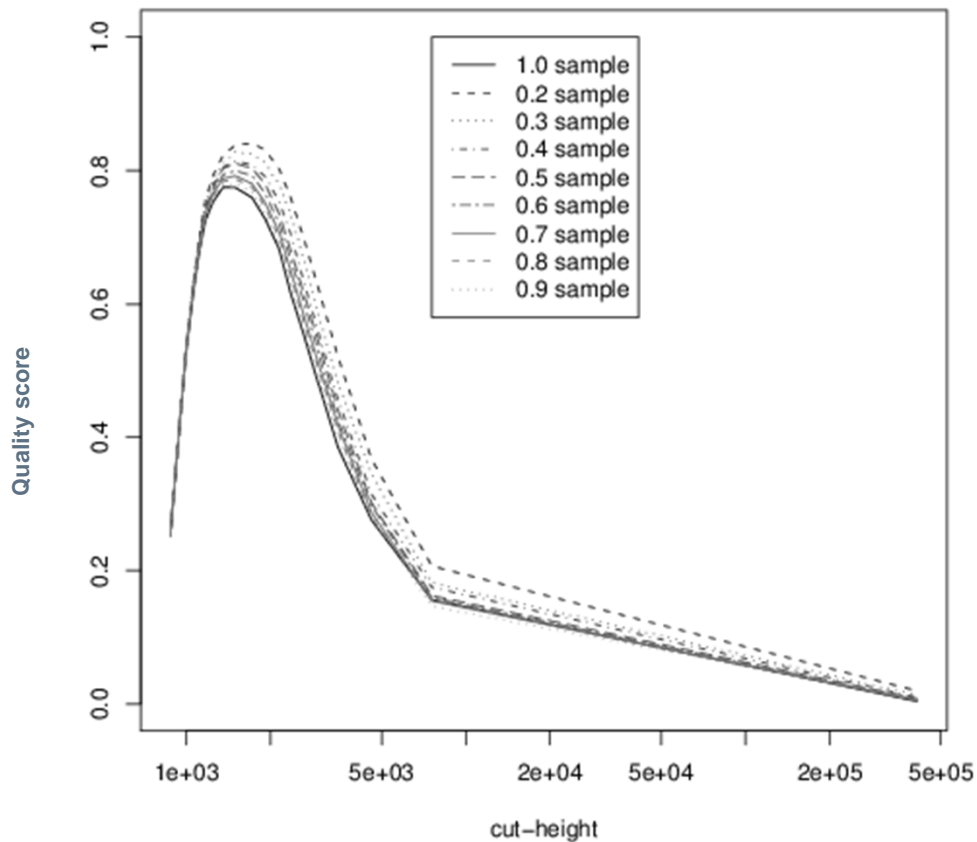
Recall and Purity



Optimal according to quality score

Optimal Cut Point with different samples of data

The height in the dendrogram (intra cluster distance) is the same for different samples of the data set



D-Scribe – Automatic Authorship Identification

Potential Applications

- **OCR**
 - Keyword screening
 - Machine translation (subsystem)
 - OCR combination framework (voting)

- **Writer ID as a biometric**

- **Clustering**

D-Scribe – Automatic Authorship Identification

Potential Applications

- **OCR**
- **Writer ID as a biometric**
 - Screening an questioned document against a known document
 - Document to document
 - Screening documents against a watch list
 - Document to documents
- **Clustering**

D-Scribe – Automatic Authorship Identification

Potential Applications

- **OCR**
- **Writer ID as a biometric**
- **Clustering**
 - Clustering handwritten and machine printed
 - Clustering by writing system / language
 - Clustering by author

D-Scribe – Automatic Authorship Identification

Potential Applications

- **OCR**
- **Writer ID as a biometric**
- **Clustering**
 - Basic Triage:
 - *Eliminating non-relevant documents*
 - Separating Machine Printed from Handwritten
 - Separating documents according to writing system
 - Separating documents by author
 - *Focus on potentially relevant documents*

D-Scribe – Automatic Authorship Identification

Contact information



Mike Carpenter
Product Manager
Logistics and Airport Solutions

2700 Esters Blvd., Suite 200B
DFW Airport, TX 75261

Phone: +1 (972) 947-7491
Mobile: +1 (817) 307-2228

E-mail:
michael.carpenter@siemens.com

siemens.com/answers