STYLE CHANGE DETECTION 2019

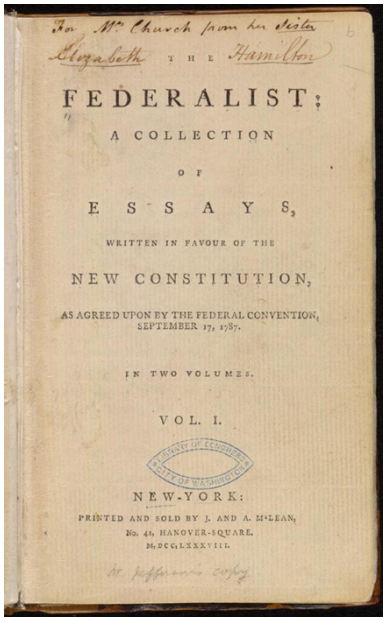
Can we uncover how many authors a document has?

Sukanya Nath
Computer Science Department
Université de Neuchâtel

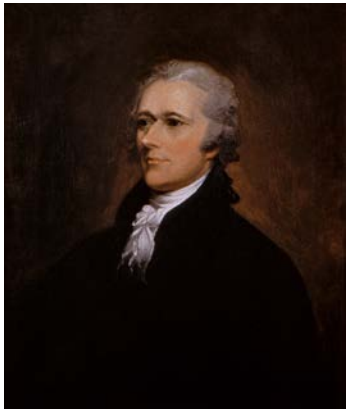
CONTENTS

WHAT IS STYLE CHANGE DETECTION (SCD)?

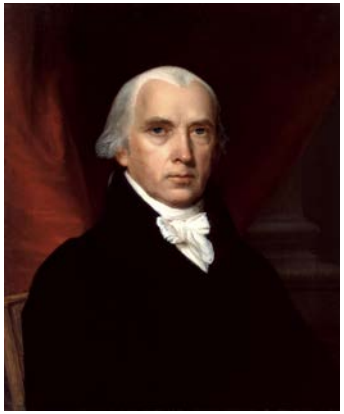
Detect number of authors by determining unique writing styles of each author



Jay



Hamilton



Madison

?

USE CASES

- Establishing the authenticity of a document, fraud detection/ prevention
- Plagiarism detection
- Determine authorship of anonymous posts at forums
- Fake news detection
- Aid automatic speech transcription using transcripts

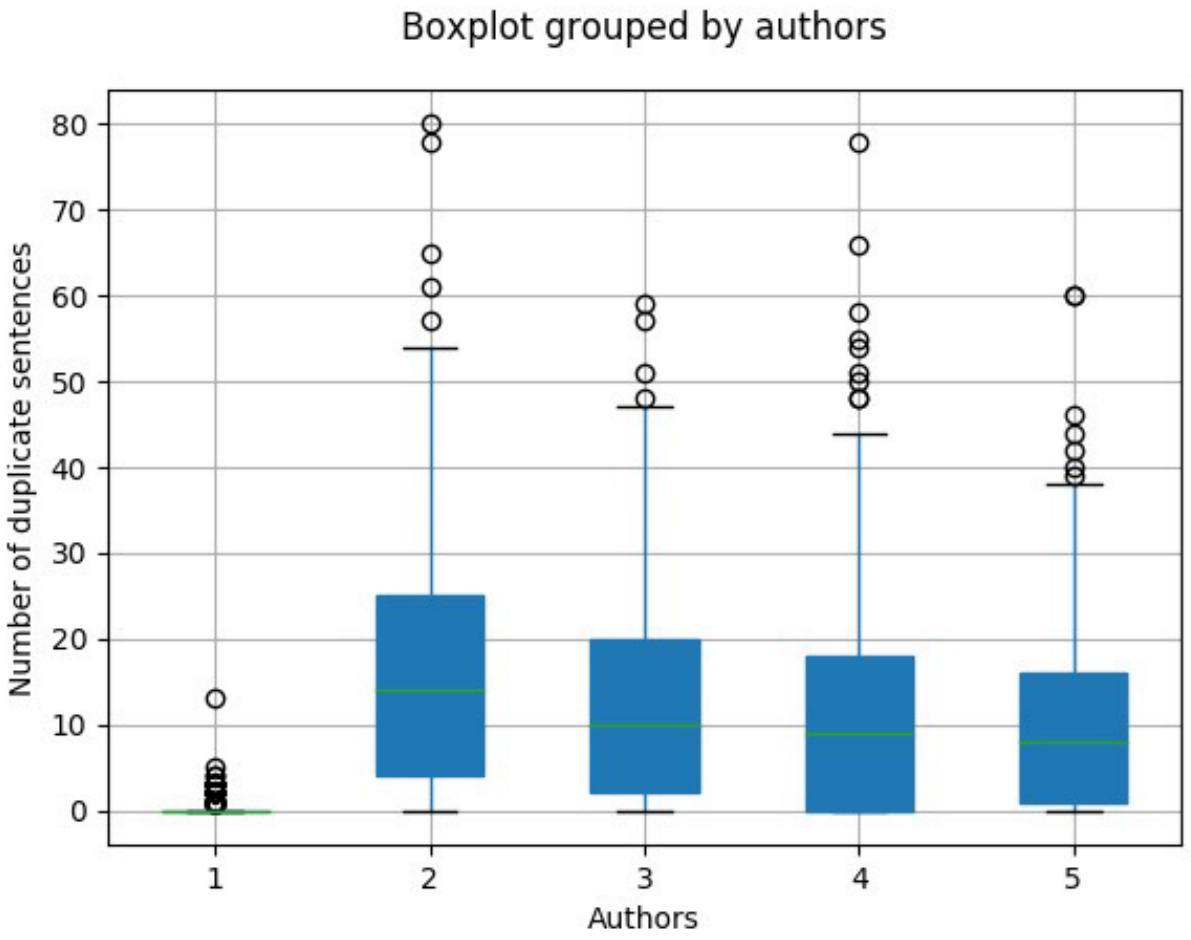
CHALLENGES

- No prior text sample for each author
- Some authors have contributed little
- Hard to determine style change boundaries

DATASET DESCRIPTION

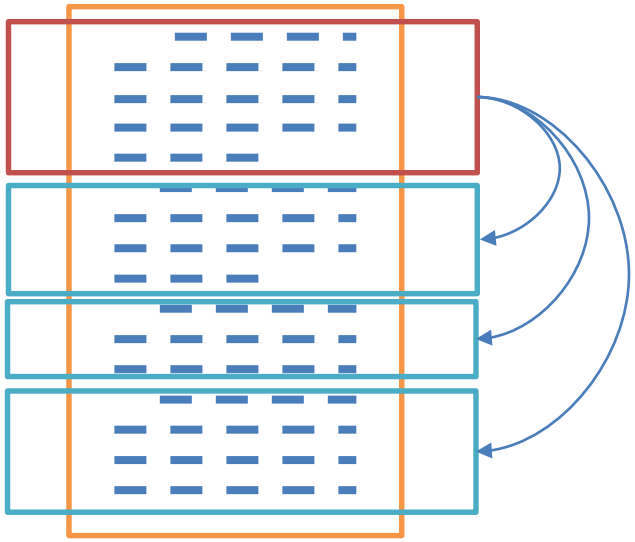
- StackOverFlow Forum threads
- 2546 (training) / 1272 (validation) documents
- 1-5 authors
- Mean number of tokens: 1570

DUPLICATE SENTENCES



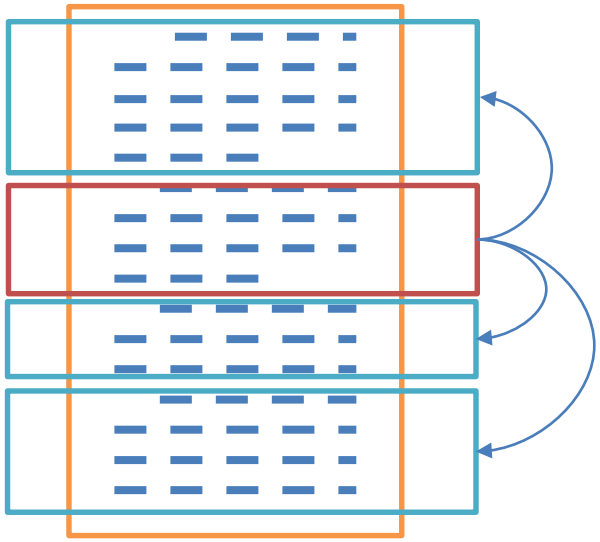
STRATEGY

– Divide the text into paragraph sized windows
– Compare windows among each other using Most Frequent Words (MFW) and measure distance (e.g. Matusita)



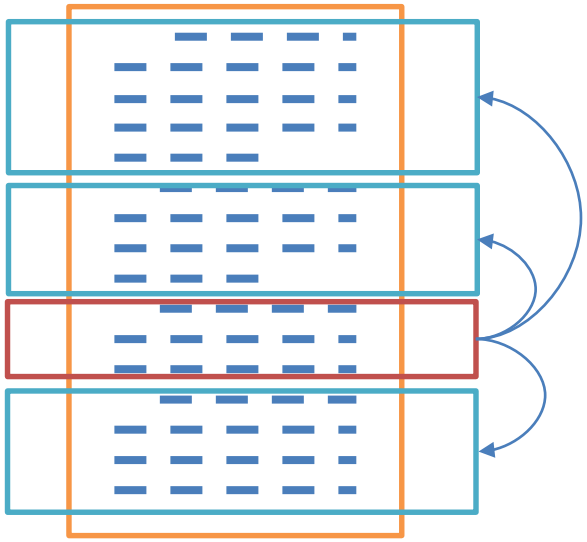
STRATEGY

- Divide the text into paragraph sized windows
- Compare windows among each other using Most Frequent Words (MFW) and measure distance (e.g. Matusita)



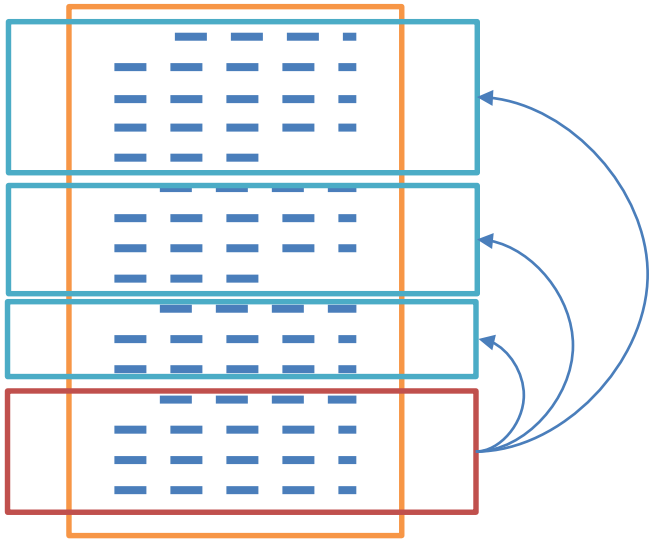
STRATEGY

- Divide the text into paragraph sized windows
- Compare windows among each other using Most Frequent Words (MFW) and measure distance (e.g. Matusita)



STRATEGY

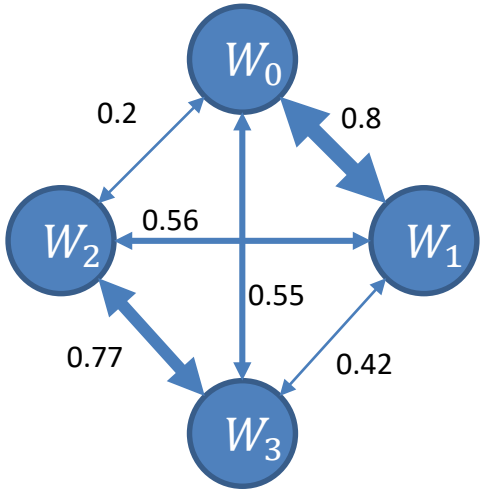
- Divide the text into paragraph sized windows
- Compare windows among each other using Most Frequent Words (MFW) and measure distance (e.g. Matusita)



STRATEGY

	W0	W1	W2	W3
W0	0	0.8	0.2	0.55
W1	0.8	0	0.56	0.42
W2	0.2	0.56	0	0.77
W3	0.55	0.42	0.77	0

Distance Matrix

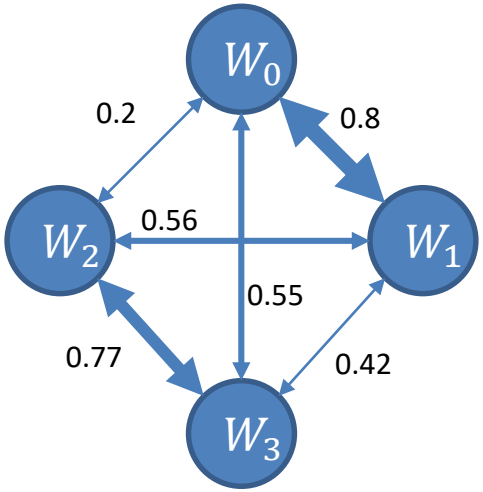


Graph representing distances between windows

STRATEGY

	W0	W1	W2	W3
W0	0	0.8	0.2	0.55
W1		0	0.56	0.42
W2			0	0.77
W3				0

Distance Matrix

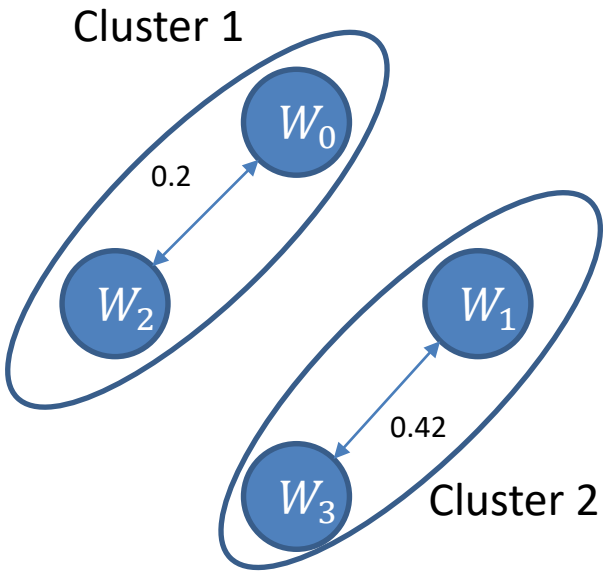


Graph representing distances between windows

GOAL

	W0	W1	W2	W3
W0	0	0.8	0.2	0.55
W1		0	0.56	0.42
W2			0	0.77
W3				0

Distance Matrix



Clusters representing authors

THRESHOLD BASED CLUSTERING ALGORITHM

Select the smallest distance from the distance matrix iteratively and use the corresponding windows to either

A	B	0.2

1. Create a new cluster

2. Add a node to an existing cluster
(subject to Add Node Threshold)

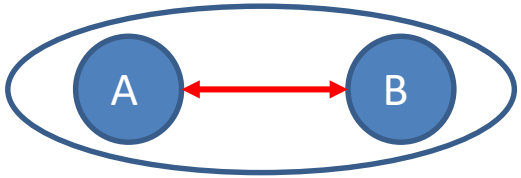
3. Merge two clusters
(subject to Merge Clusters Threshold)

THRESHOLD BASED CLUSTERING ALGORITHM

Select the smallest distance from the distance matrix iteratively and use the corresponding windows to either

A	B	0.2

1. Create a new cluster



2. Add a node to an existing cluster
(subject to Add Node Threshold)

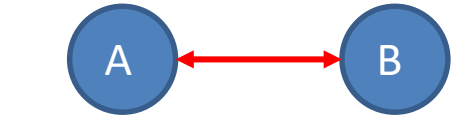
3. Merge two clusters
(subject to Merge Clusters Threshold)

THRESHOLD BASED CLUSTERING ALGORITHM

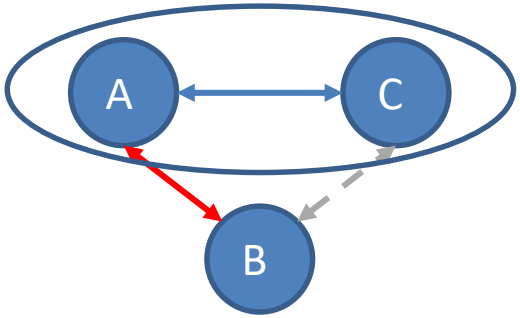
Select the smallest distance from the distance matrix iteratively and use the corresponding windows to either

A	B	0.2

1. Create a new cluster



2. Add a node to an existing cluster
(subject to Add Node Threshold)



3. Merge two clusters
(subject to Merge Clusters Threshold)

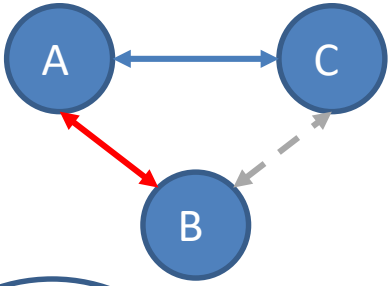
THRESHOLD BASED CLUSTERING ALGORITHM

Select the smallest distance from the distance matrix iteratively and use the corresponding windows to either

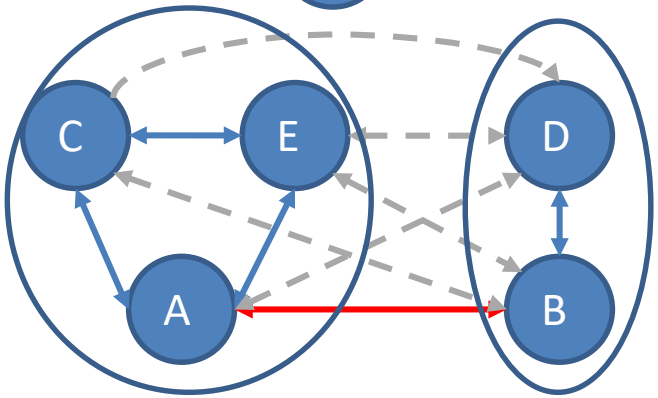
A	B	0.2

1. Create a new cluster

2. Add a node to an existing cluster
(subject to Add Node Threshold)

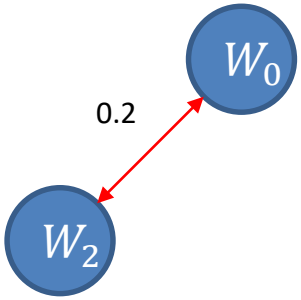


3. Merge two clusters
(subject to Merge Clusters Threshold)



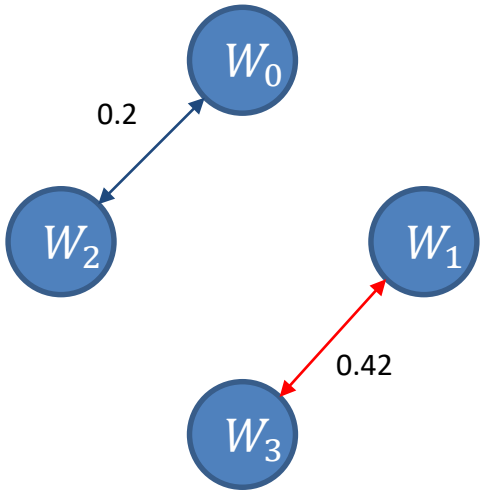
THRESHOLD BASED CLUSTERING ALGORITHM (EXAMPLE)

	W0	W1	W2	W3
W0	0	0.8	**0.2**	0.55
W1		0	0.56	0.42
W2			0	0.77
W3				0



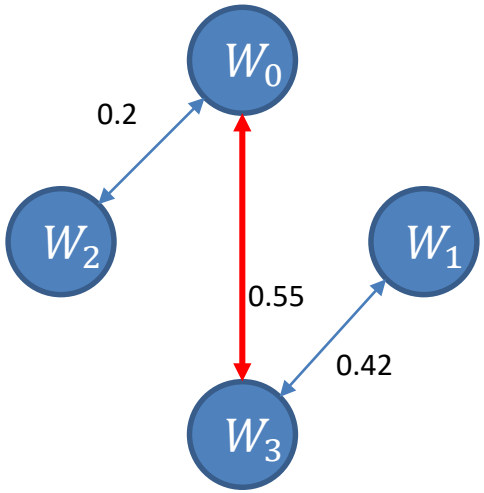
THRESHOLD BASED CLUSTERING ALGORITHM (EXAMPLE)

	W0	W1	W2	W3
W0	0	0.8	0.2	0.55
W1		0	0.56	**0.42**
W2			0	0.77
W3				0



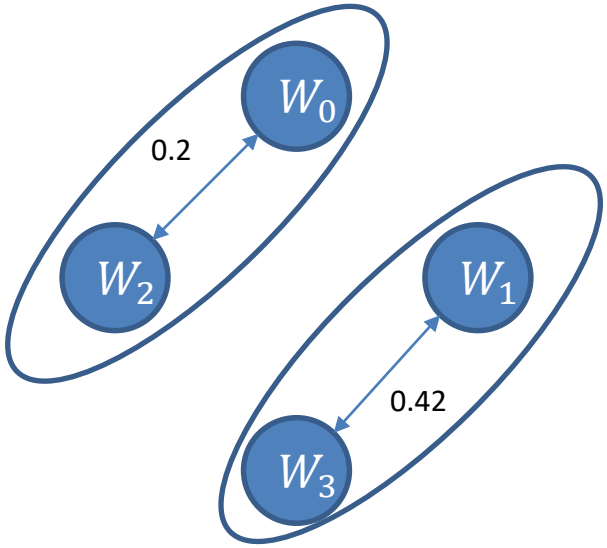
THRESHOLD BASED CLUSTERING ALGORITHM (EXAMPLE)

	W0	W1	W2	W3
W0	0	0.8	0.2	**0.55**
W1		0	0.56	0.42
W2			0	0.77
W3				0



THRESHOLD BASED CLUSTERING ALGORITHM (EXAMPLE)

	W0	W1	W2	W3
W0	0	0.8	0.2	0.55
W1		0	0.56	0.42
W2			0	0.77
W3				0



WINDOW MERGE CLUSTERING ALGORITHM

1

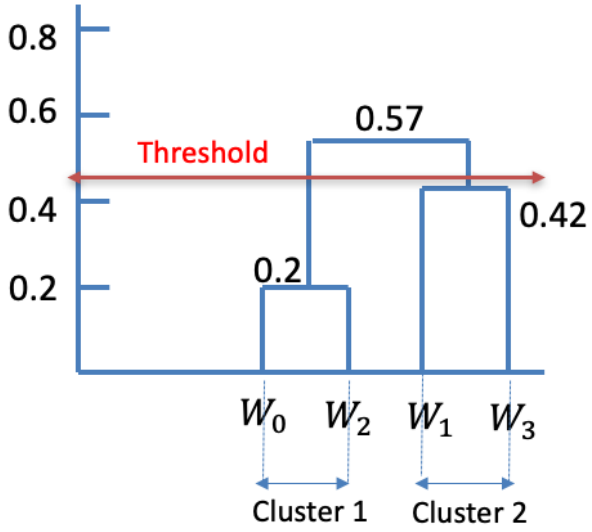
	W0	W1	W2	W3
W0	0	0.8	0.2	0.55
W1		0	0.56	0.42
W2			0	0.77
W3				0

2

	W0W2	W1	W3
W0 W2	0	0.72	0.65
W1		0	0.42
W3			0

3

	W0W2	W1W3
W0 W2	0	0.57
W1 W3		0

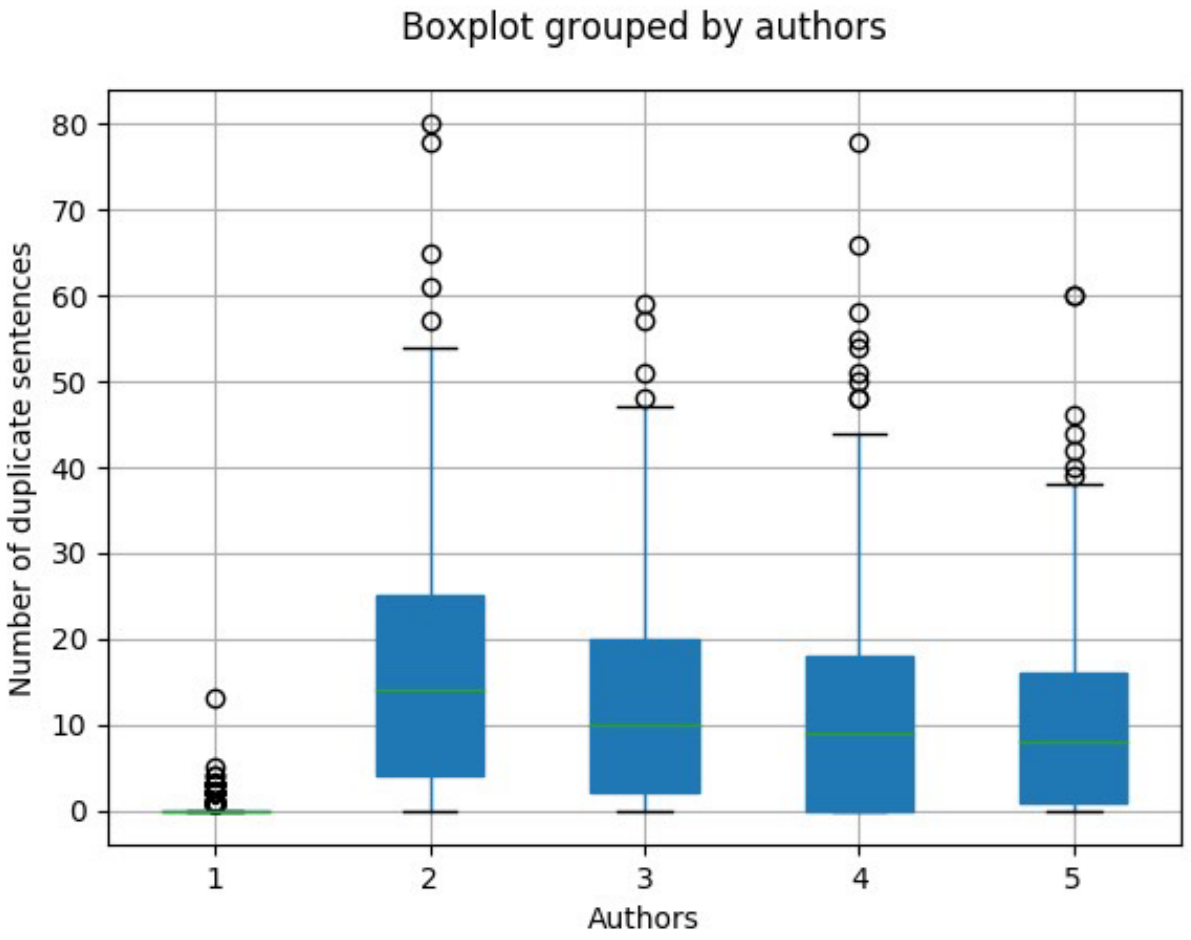


EVALUATION

Initial Evaluation Results

	Training set			Validation set		
Algorithm	**Acc.**	**OCI**	**Rank**	**Acc.**	**OCI**	**Rank**
TBC	**0.66**	**0.83**	**0.42**	**0.65**	**0.82**	**0.42**
WMC	0.62	0.91	0.35	0.63	0.88	0.37
Combined Min.	0.65	0.92	0.36	0.66	0.9	0.38

DUPLICATE SENTENCES



EVALUATION

Final Evaluation Results (using duplicates filter)

	Training			Validation			Official Test		
Algorithm	**Acc.**	**OCI**	**Rank**	**Acc.**	**OCI**	**Rank**	**Rank**	**Acc.**	**OCI**
TBC	**0.83**	**0.87**	**0.48**	**0.83**	**0.85**	**0.49**	**0.85**	**0.87**	**0.49**
WMC	0.72	0.93	0.4	0.74	0.9	0.42	-	-	-
Combined Min.	0.70	0.93	0.39	0.72	0.91	0.41	-	-	-

CONCLUSION AND FUTURE WORK

- Demonstrated that it is possible to detect style changes in a text when no prior dataset is available
- Threshold Based Clustering algorithm performed the best out of the three models
- Splitting of Windows may be improved
- Measure cluster quality for improvement.

THANK YOU FOR YOUR ATTENTION!

Sukanya Nath
Adresse
CH-2000 Neuchâtel
sukanya.nath@unine.ch
www.unine.ch