Relative N-Gram Signatures: Document Visualization at the Level of Character N-Grams

Magdalena Jankowska, Evangelos Milios, Vlado Kešelj
Faculty of Computer Science, Dalhousie University

June 2013

DALHOUSIE UNIVERSITY
Inspiring Minds

Relative N-Gram Signatures

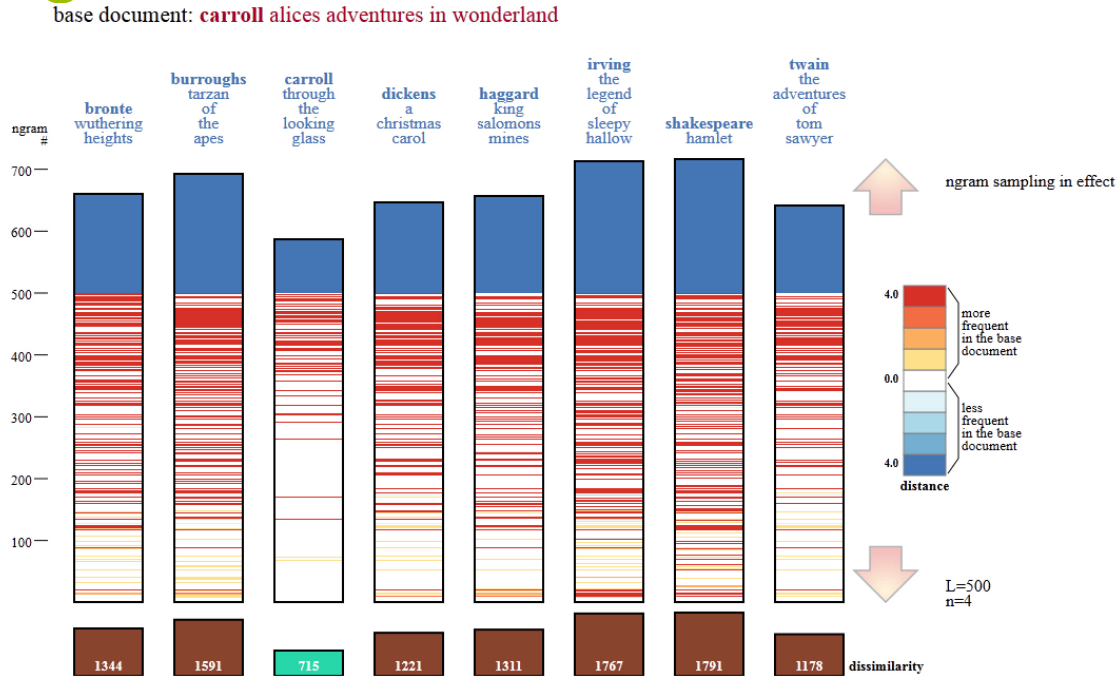
Interactive classification of a document

Who wrote this book?

Analysis of characteristics of a document

What are the characteristics of the author's style?

Language independent method



Character N-Grams

Strings of n consecutive characters from a given text

Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do:

Alice's Adventures in the Wonderland
by Lewis Carroll

Character N-Grams

Strings of n consecutive characters from a given text

Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do:

Alice's Adventures in the Wonderland
by Lewis Carroll

n=4
4-grams

ALIC

Character N-Grams

Strings of n consecutive characters from a given text

Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do:

Alice's Adventures in the Wonderland
by Lewis Carroll

n=4
4-grams

ALIC
LICE

Character N-Grams

Strings of n consecutive characters from a given text

Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do:

Alice's Adventures in the Wonderland
by Lewis Carroll

n=4
4-grams

ALIC
LICE
ICE_

Character N-Grams

Strings of n consecutive characters from a given text

Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do:

Alice's Adventures in the Wonderland
by Lewis Carroll

n=4
4-grams

ALIC
LICE
ICE_
CE_W

Character N-Grams

Strings of n consecutive characters from a given text

Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do:

Alice's Adventures in the Wonderland
by Lewis Carroll

n=4
4-grams

ALIC
LICE
ICE_
CE_W

n-grams in our system:
- uppercase
- each sequence of non-word characters replaced by an underscore

Common N-Gram (CNG) Classifier

assigns a document to a class from a given set of classes

works of Carrol

works of Twain

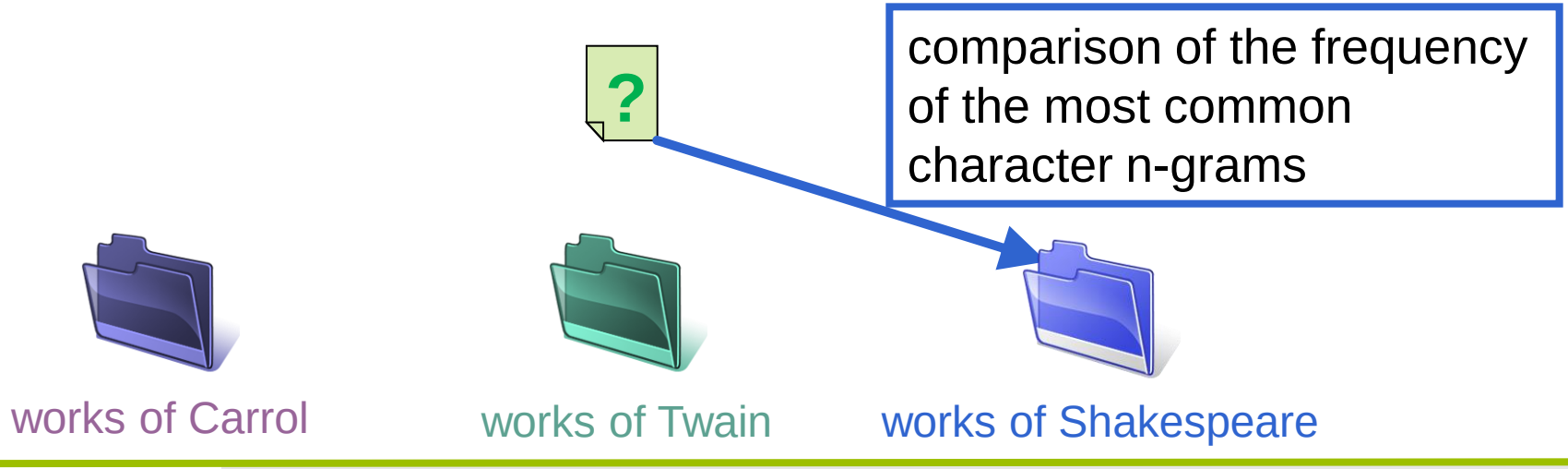
works of Shakespeare

Proposed by

Vlado Kešelj, Fuchun Peng, Nick Cercone, and Calvin Thomas.
N-gram-based author profiles for authorship attribution.
In Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING'03, 2003.

Common N-Gram (CNG) Classifier

assigns a document to a class from a given set of classes

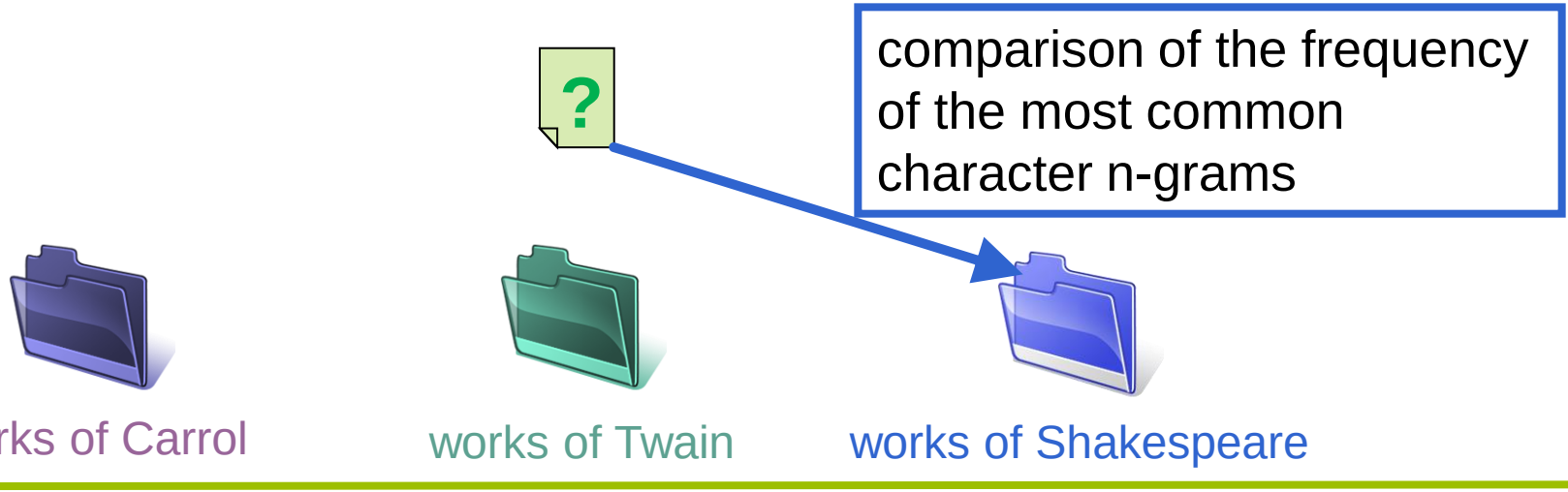


Proposed by

Vlado Kešelj, Fuchun Peng, Nick Cercone, and Calvin Thomas.
N-gram-based author profiles for authorship attribution.
In Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING'03, 2003.

Common N-Gram (CNG) Classifier

assigns a document to a class from a given set of classes



Applications: Authorship attribution
Malicious code detection
Gene classification
Web page genre classification…

CNG Classifier - Dissimilarity

Profile

a sequence of L most common n-grams of a given length n

n=4, L=6

document 1:

Alice's Adventures in the Wonderland

by Lewis Carroll

_ T O _	
_ A N D	
I N G _	
A N D _	
T H E _	
_ T H E	f_1

document 2:

Tarzan of the Apes

by Edgar Rice Burroughs

I N G _
_ A N D
_ O F _
A N D _
T H E _
_ T H E

n-gram | **normalized frequency**

CNG Classifier - Dissimilarity

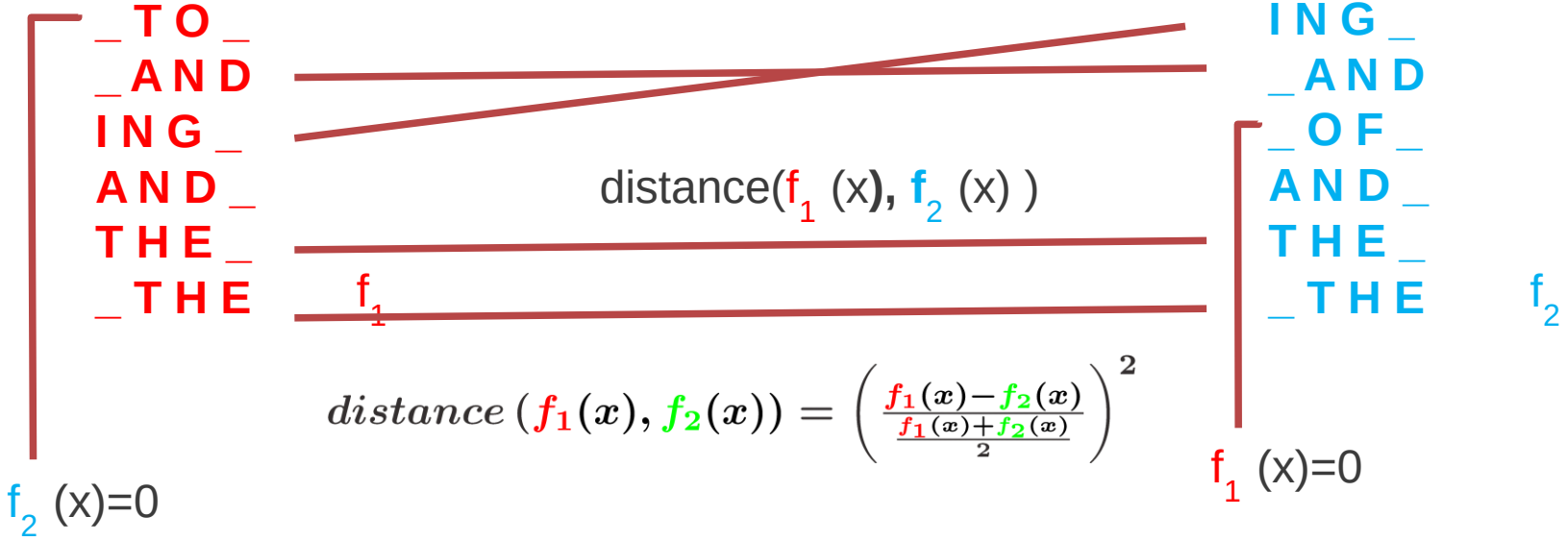
Profile

a sequence of L most common n-grams of a given length n

n=4, L=6

document 1:
Alice's Adventures in the Wonderland
by Lewis Carroll

document 2:
Tarzan of the Apes
by Edgar Rice Burroughs



$$distance\,(f_1(x), f_2(x)) = \left(\frac{f_1(x)-f_2(x)}{\frac{f_1(x)+f_2(x)}{2}}\right)^2$$

CNG Classifier - Dissimilarity

Profile

a sequence of L most common n-grams of a given length n

n=4, L=6

document 1:
Alice's Adventures in the Wonderland
by Lewis Carroll

_ T O _
_ A N D
I N G _
A N D _
T H E _
_ T H E

f_1

distance(f_1 (x), f_2 (x))

document 2:
Tarzan of the Apes
by Edgar Rice Burroughs

I N G _
_ A N D
_ O F _
A N D _
T H E _
_ T H E

f_2

CNG dissimilarity between two documents
sum of the distances with respect to all n-grams in the union of the profiles

Motivation

- text visualization on the language-independent level of character n-grams
  - similarity of documents
  - characteristics of documents

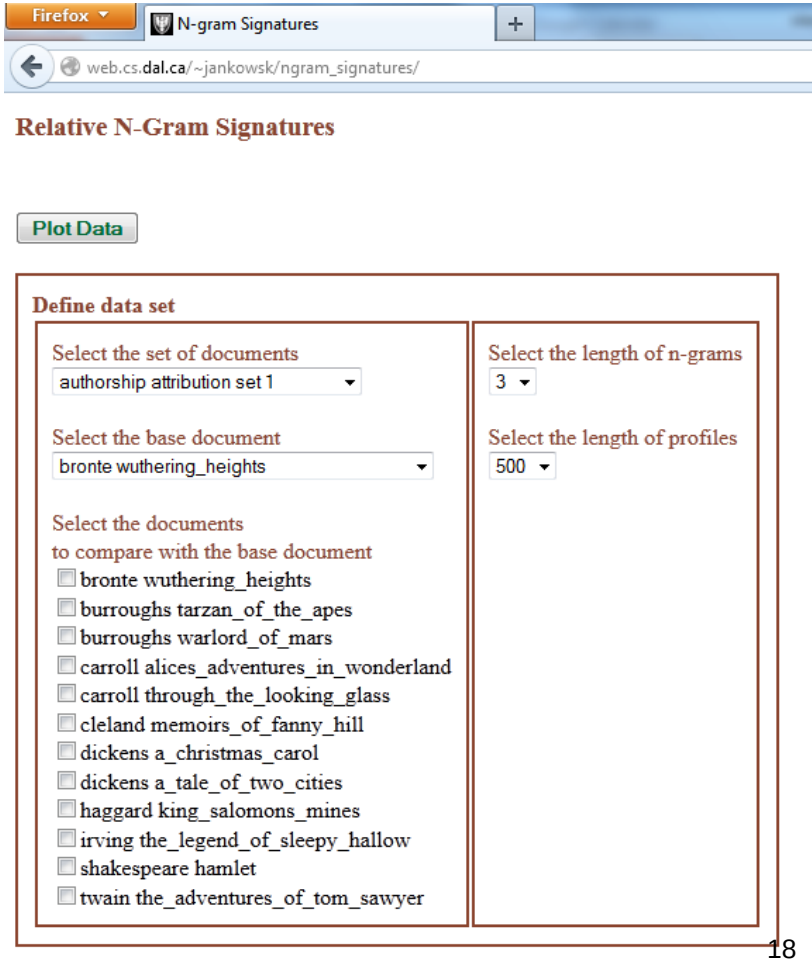
- visualization of the CNG classifier
  - “reasons” for the classification result
  - possibility of influencing the classification

RNG-Sig
Web application

Implemented as a **web application**

d3.js JavaScript library for visualization

Available online with pre-loaded data at:
http://cs.dal.ca/~jankowsk/ngram_signatures



Relative N-Gram Signature

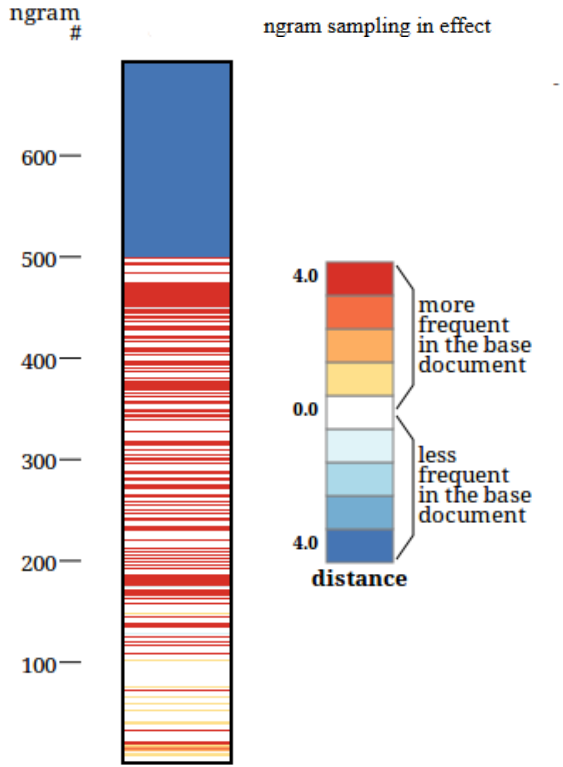
Relative signature of
Tarzan of the Apes
by Burroughs

with respect to
(“on the background of”)

Alice's Adventures in the Wonderland
by Carroll
(base document)

n=4 (4-grams)
L=500 (500 most common n-grams)

Visual representation of the **CNG dissimilarity** between two documents



Relative N-Gram Signature

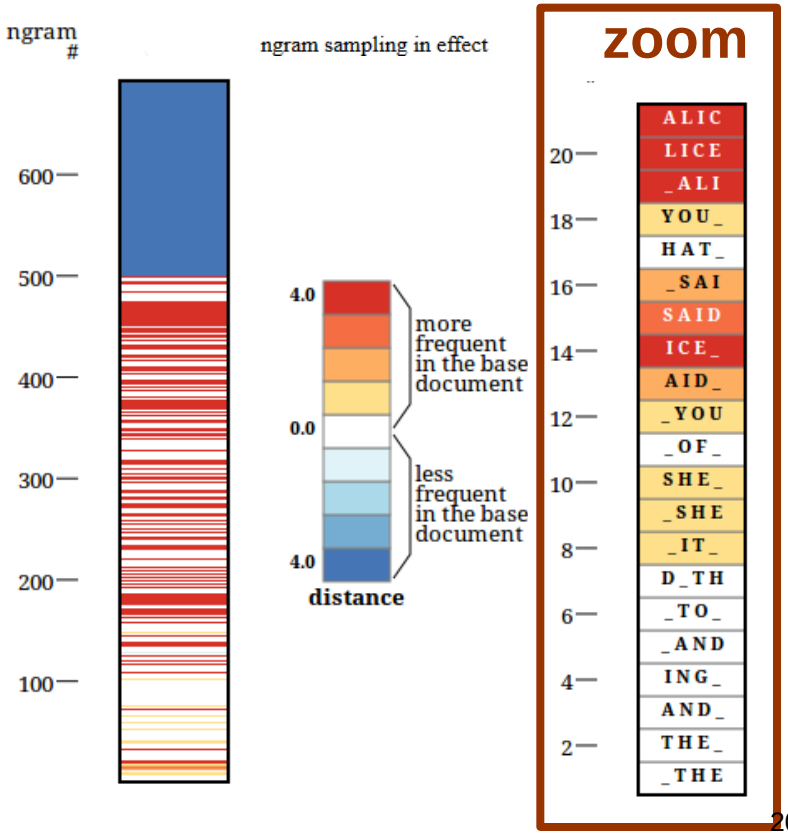
Relative signature of
Tarzan of the Apes
by Burroughs

with respect to
(“on the background of”)

Alice's Adventures in the Wonderland
by Carroll
(base document)

n=4 (4-grams)
L=500 (500 most common n-grams)

Each **strip** represents an **n-gram**



Relative N-Gram Signature

Relative signature of
Tarzan of the Apes
by Burroughs

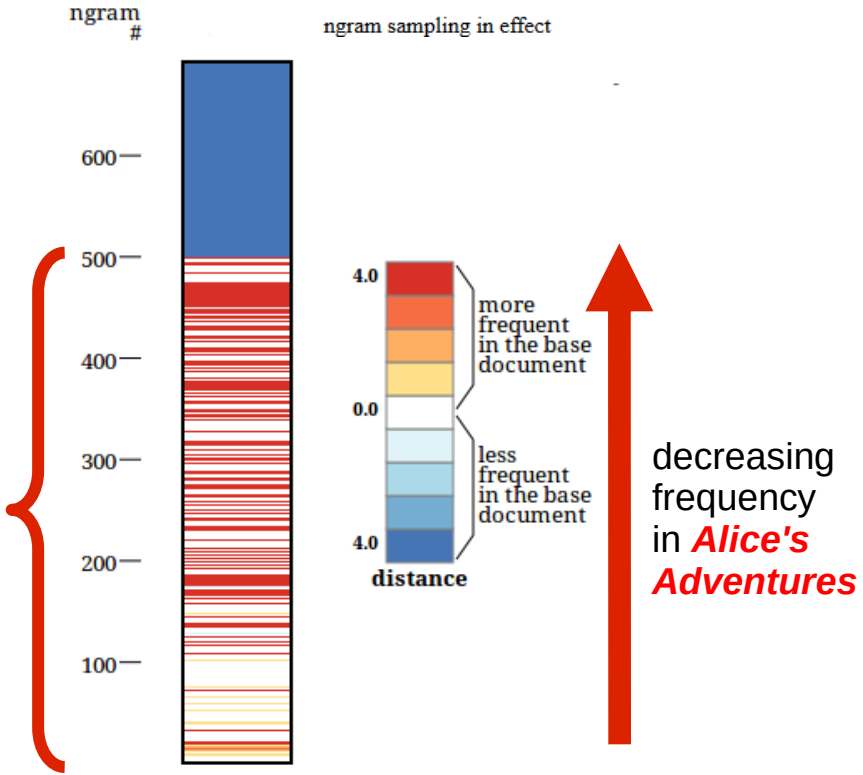
with respect to
(“on the background of”)

Alice's Adventures in the Wonderland
by Carroll
(base document)

n=4 (4-grams)
L=500 (500 most common n-grams)

Each **strip** represents an **n-gram**

500 most common n-grams of ***Alice's Adventures***



Relative N-Gram Signature

Relative signature of
Tarzan of the Apes
by Burroughs

with respect to
(“on the background of”)

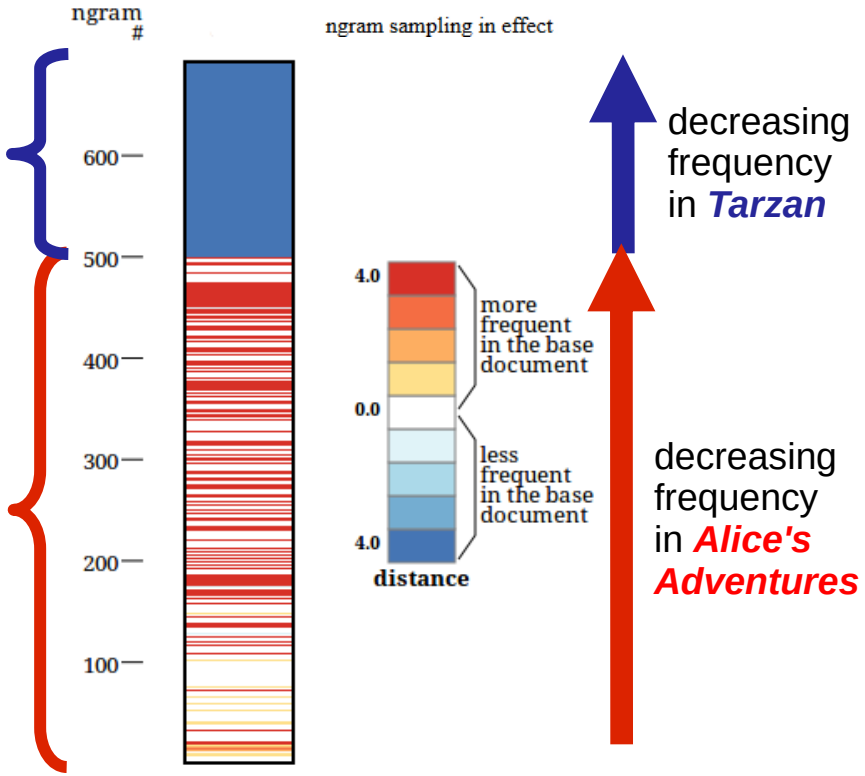
Alice's Adventures in the Wonderland
by Carroll
(base document)

n=4 (4-grams)
L=500 (500 most common n-grams)

these among
500 most common n-grams
of ***Tarzan***
that are not among
500 most common n-grams
of ***Alice's Adventures***

Each **strip** represents an **n-gram**

500 most common n-grams
of ***Alice's Adventures***



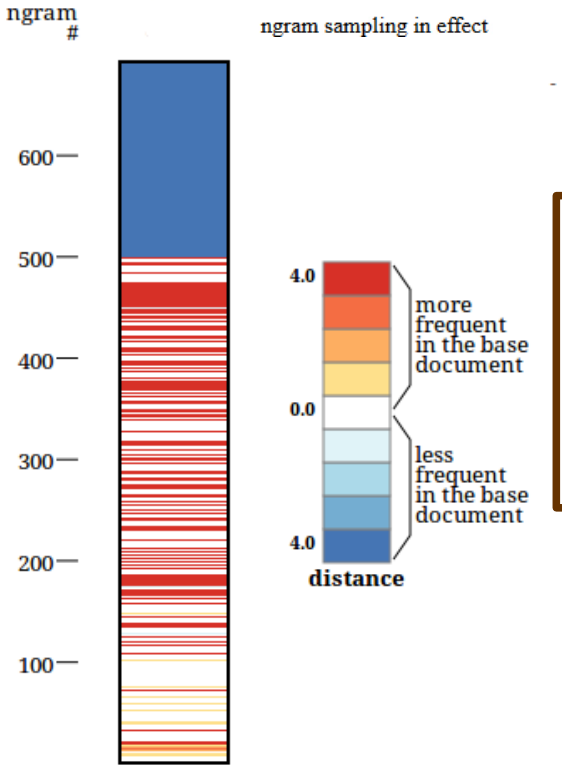
Relative N-Gram Signature

Relative signature of
Tarzan of the Apes
by Burroughs

with respect to
(“on the background of”)

Alice's Adventures in the Wonderland
by Carroll
(base document)

n=4 (4-grams)
L=500 (500 most common n-grams)



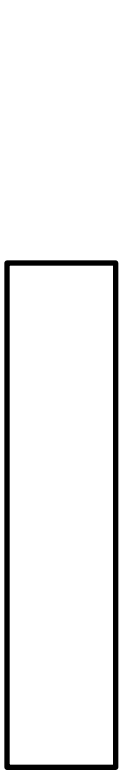
Color:

“distance”
of two documents with respect to this n-gram

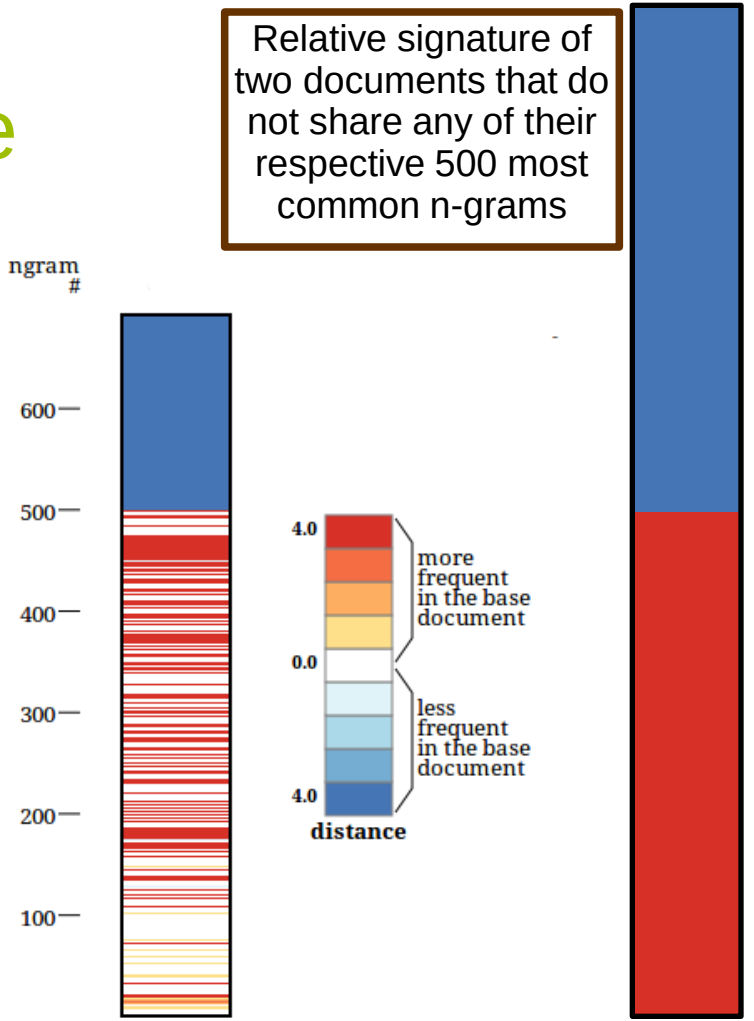
Relative N-Gram Signature

Visualizes **similarity** between documents on the level of character n-grams

Signature of a document with respect to itself



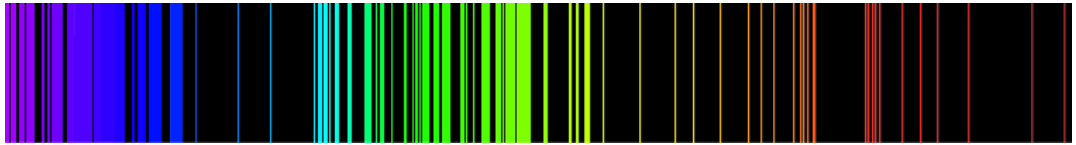
Relative signature of two documents that do not share any of their respective 500 most common n-grams



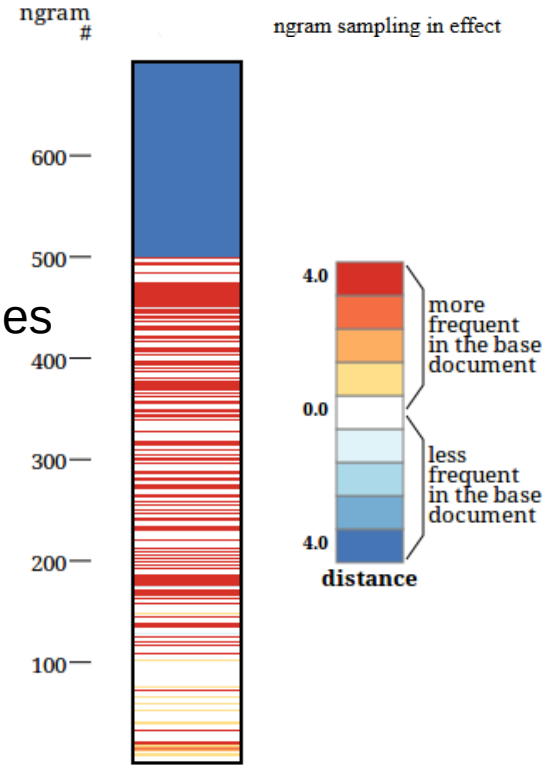
Relative N-Gram Signature
Visual metaphor

Inspiration: emission spectrum

spectrum of frequencies
of electromagnetic emissions by atoms or molecules



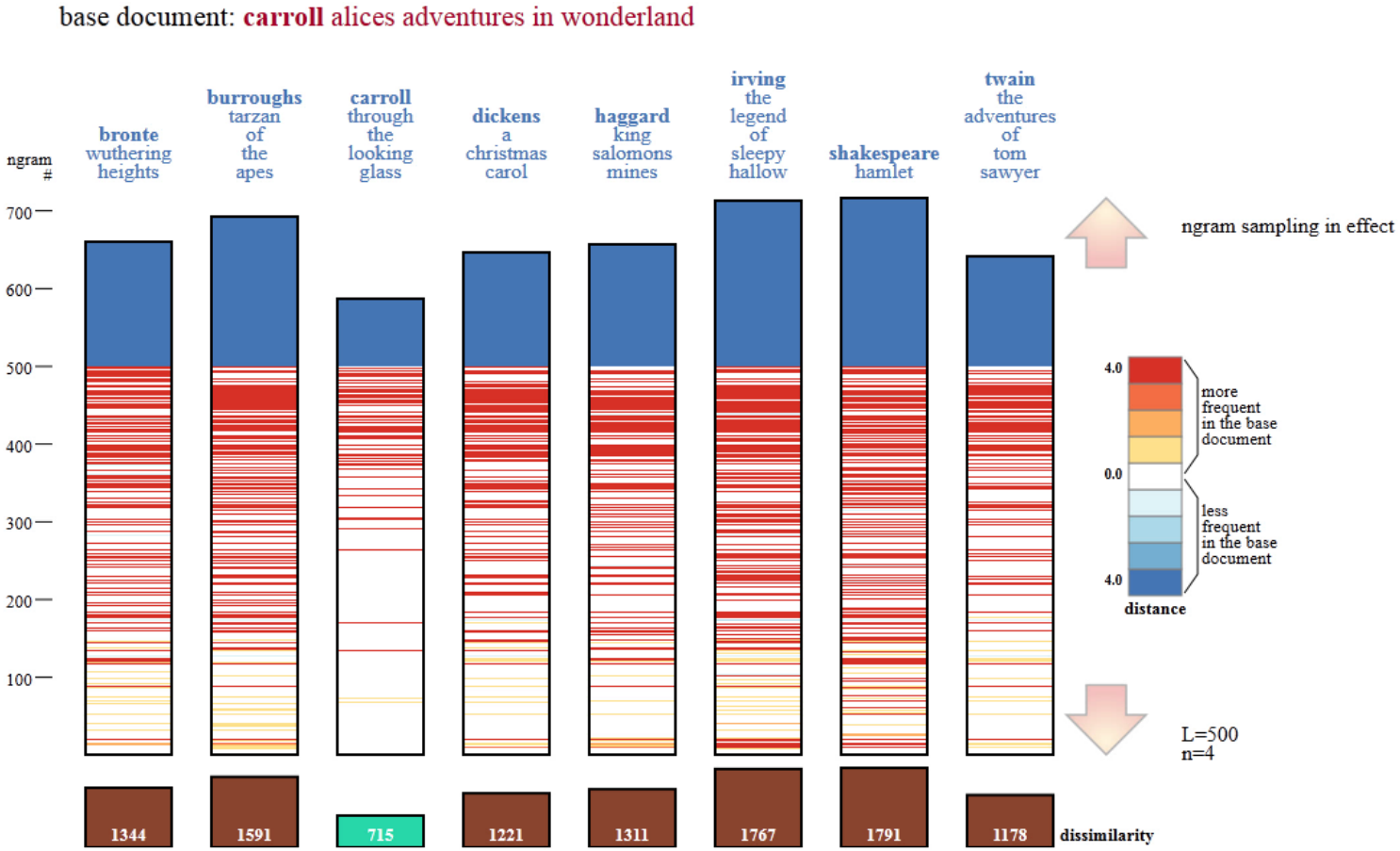
picture from Wikipedia



Sequence of signatures

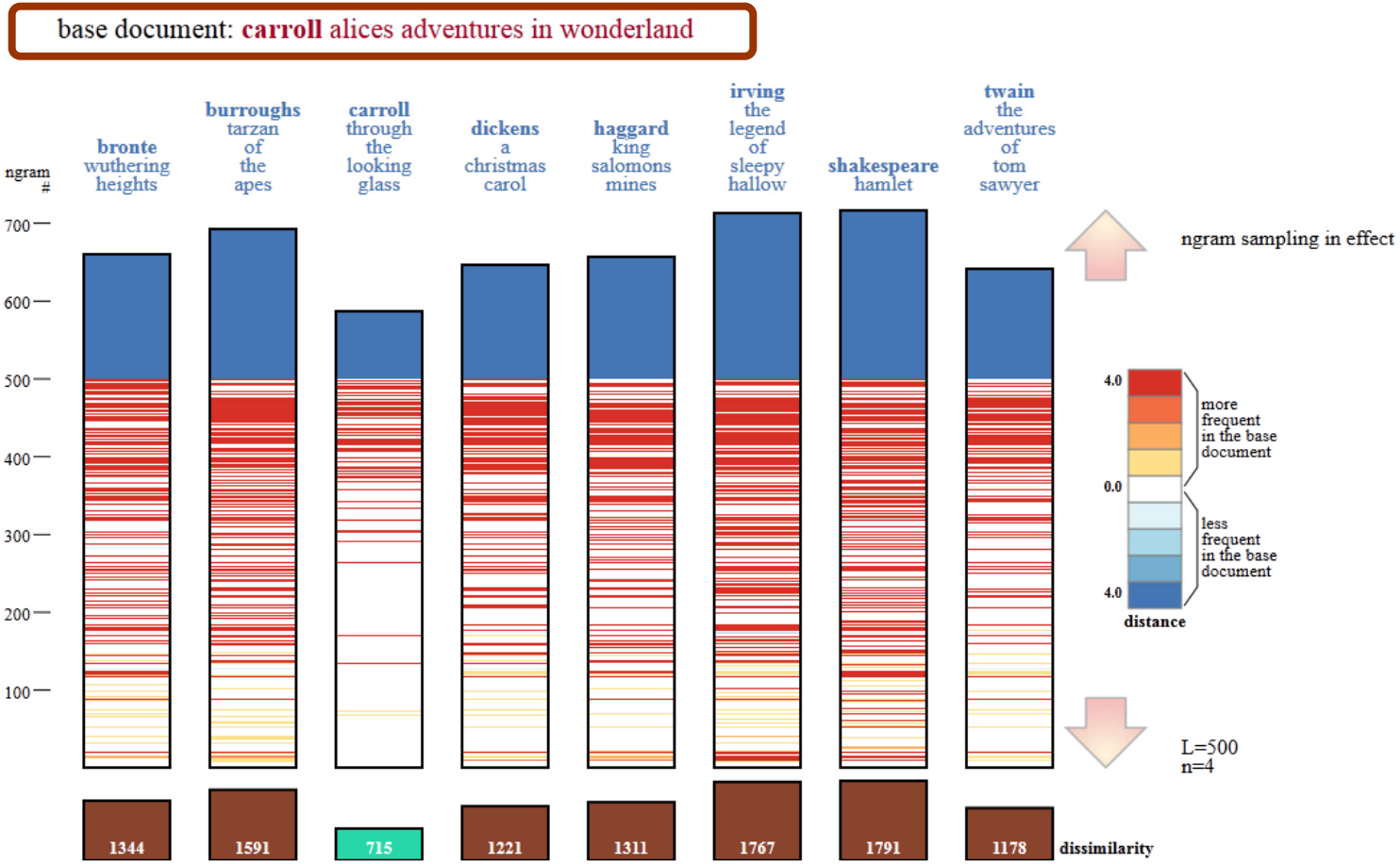
scenario

authorship analysis



Sequence of signatures

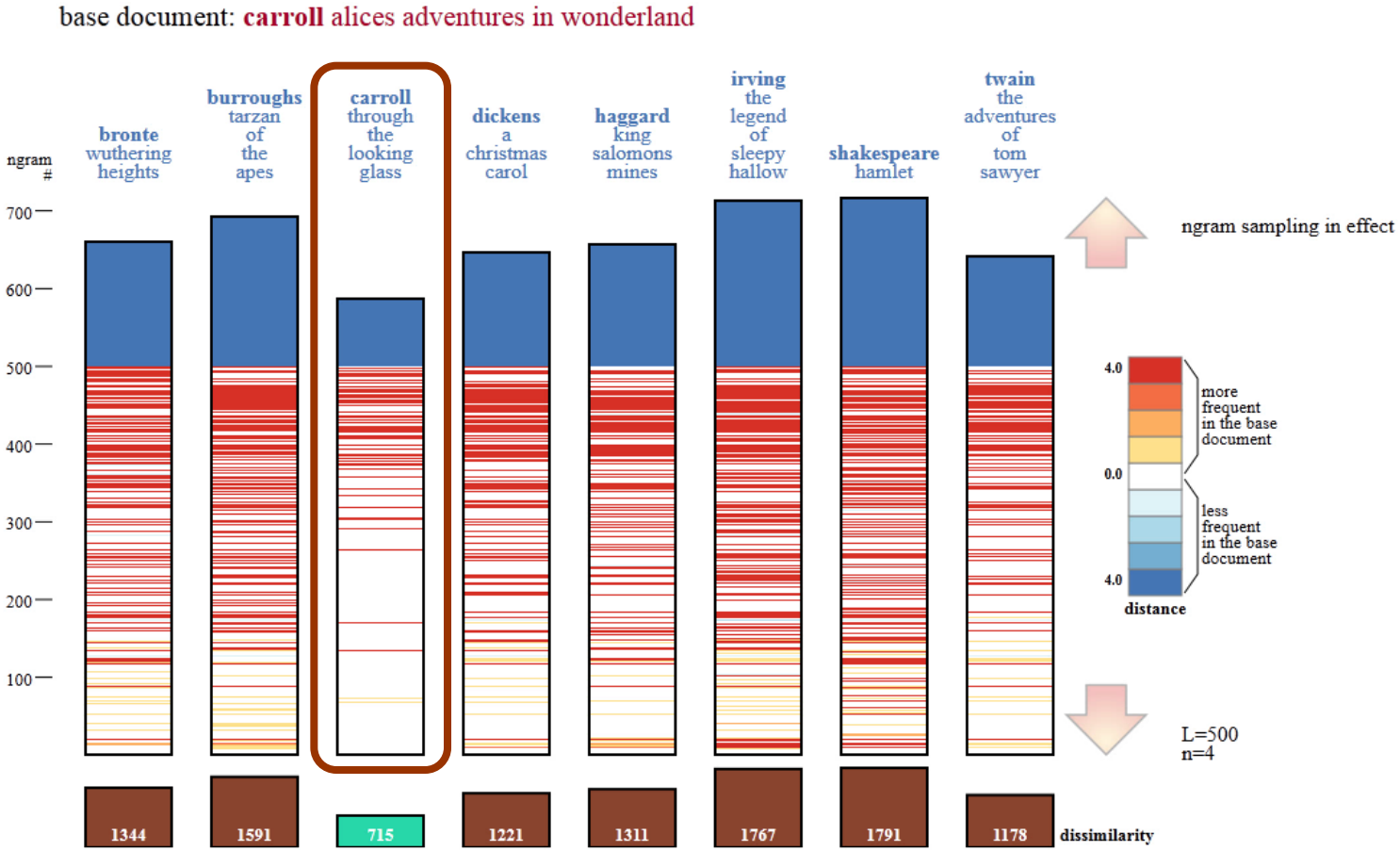
The same **base document**



Sequence of signatures

Signature of **the most similar document**

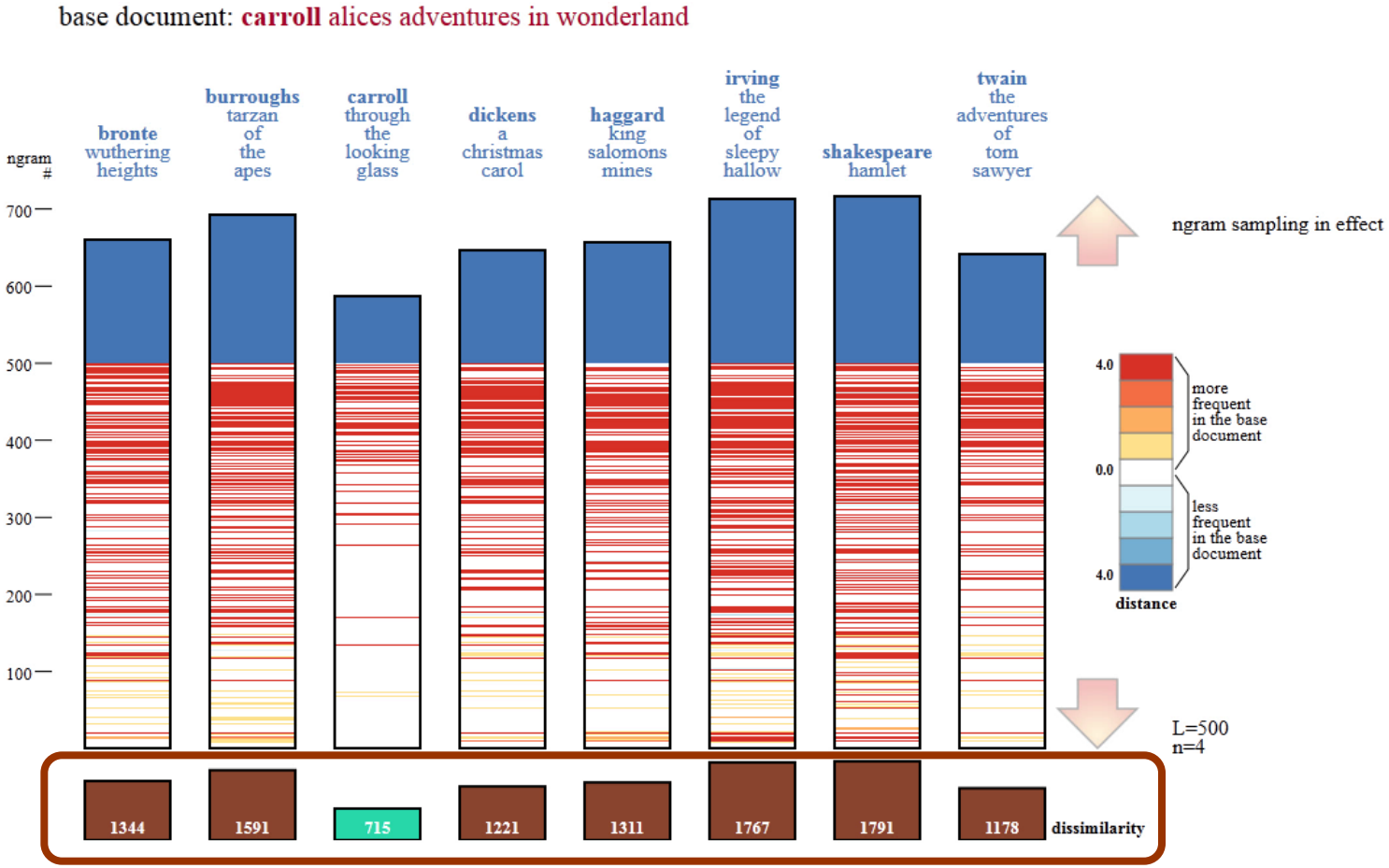
Carrol's "Through the looking glass"



Sequence of signatures

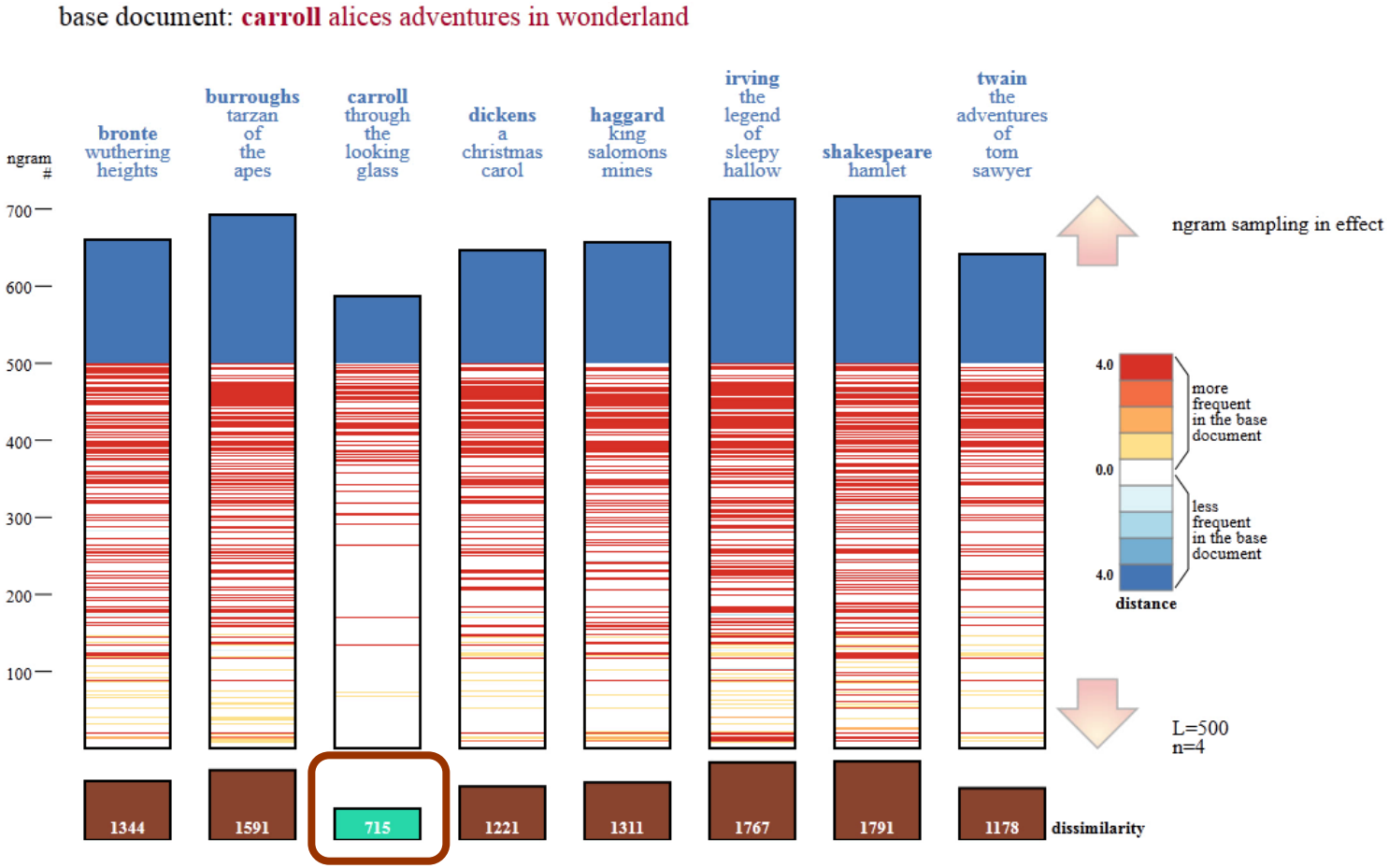
CNG **dissimilarity score**

sum of the distances over all n-grams in a signature



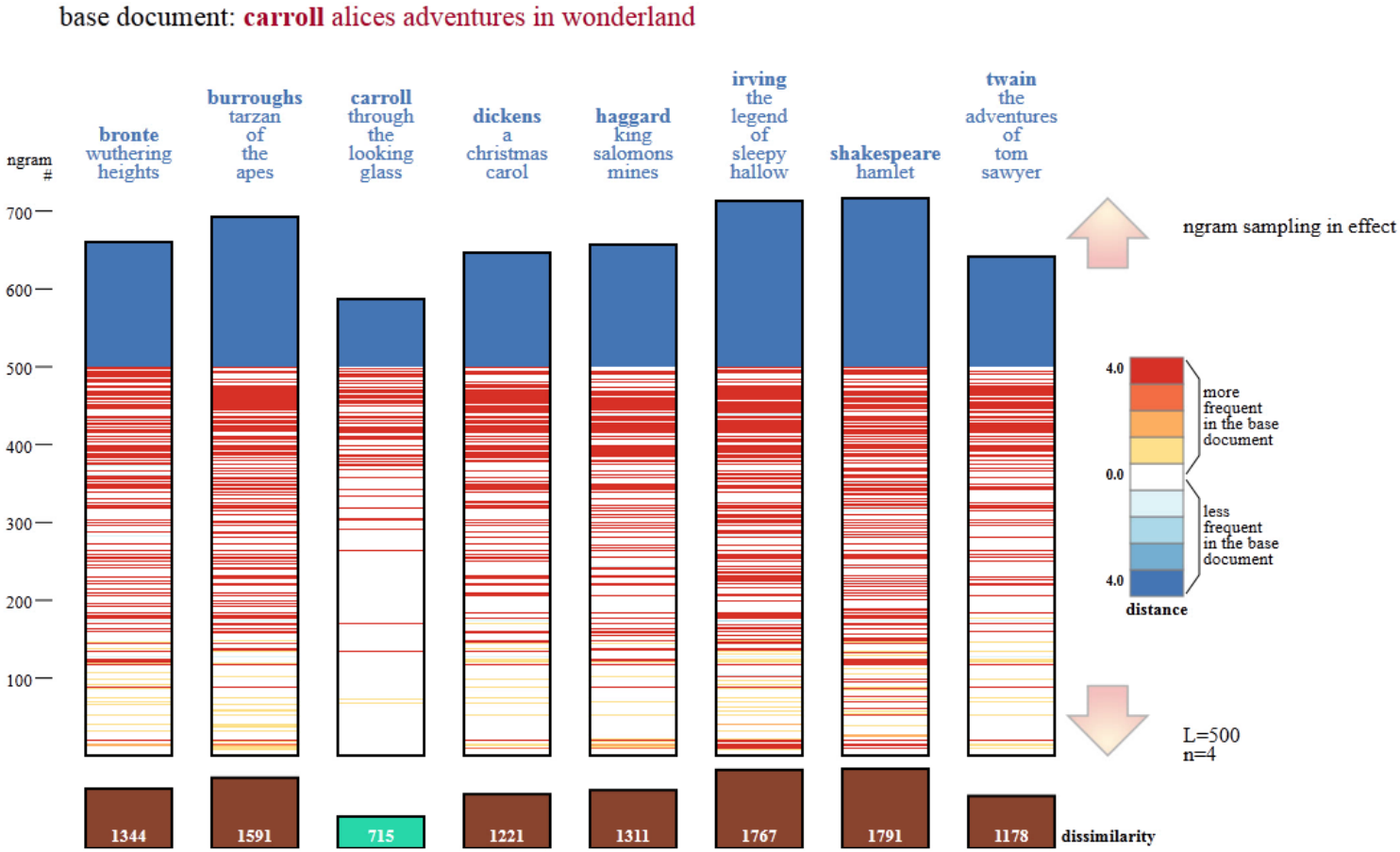
Sequence of signatures

minimum dissimilarity = **classifier result**



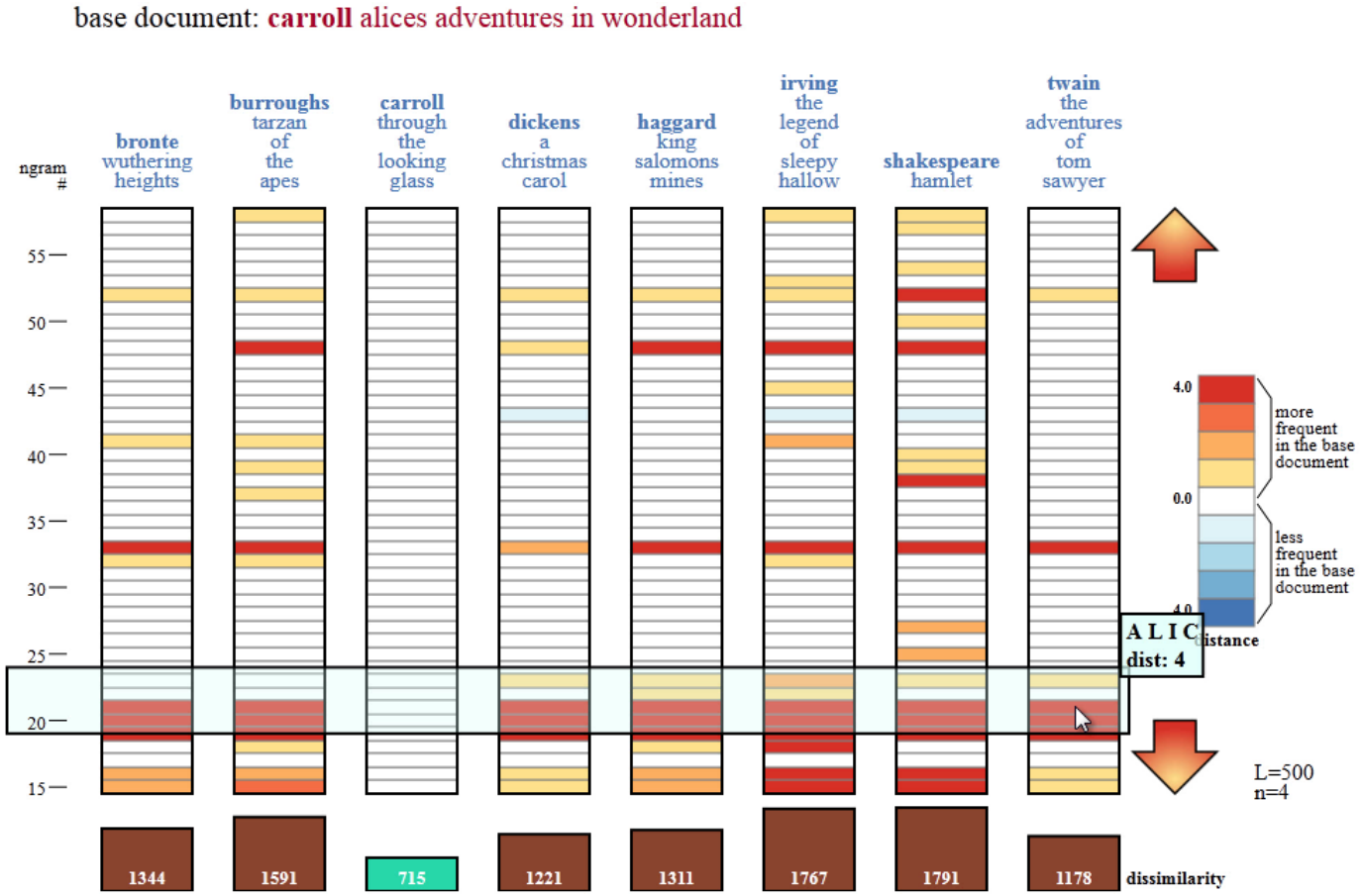
Sequence of signatures

zooming in



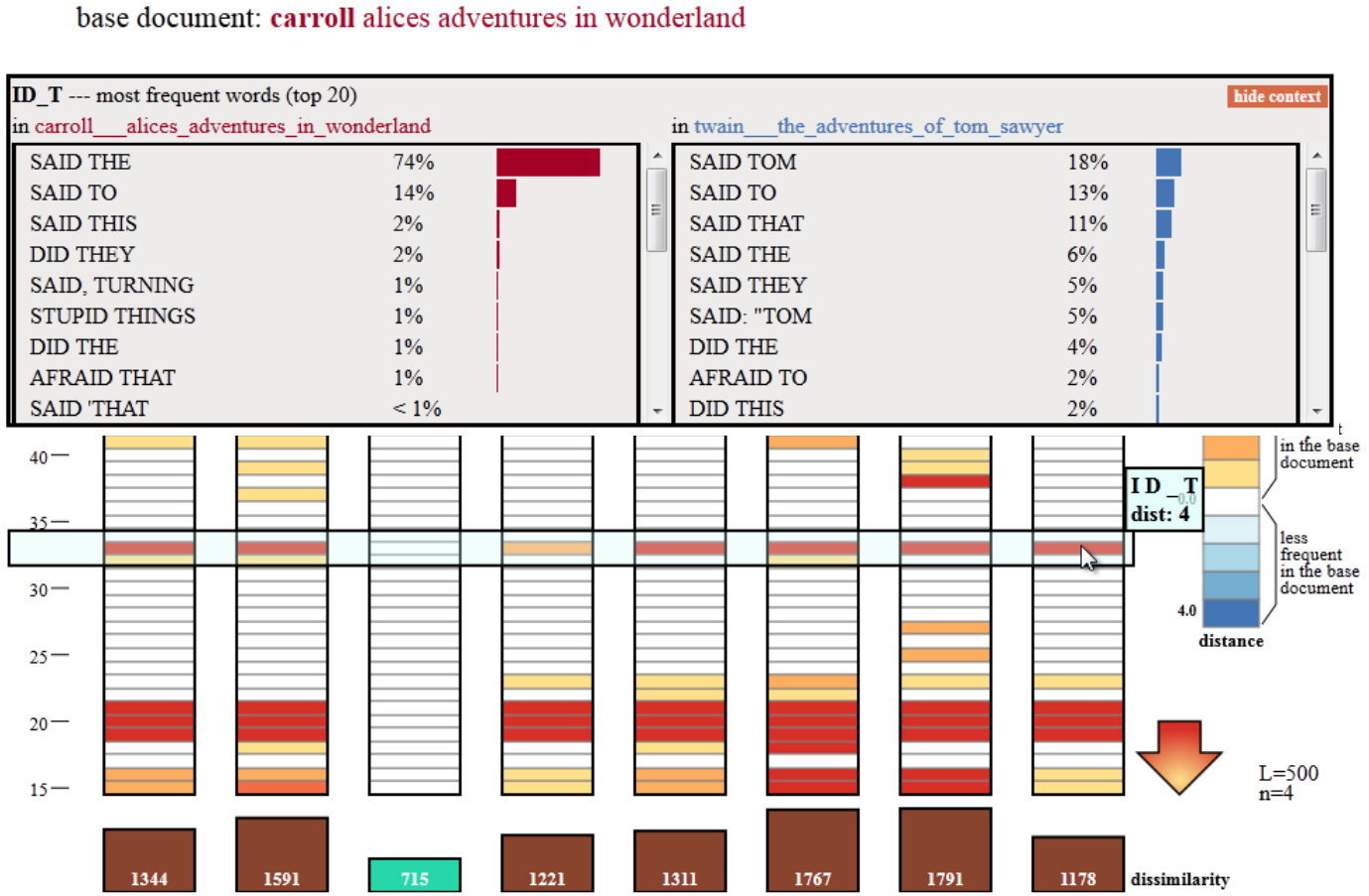
Interactive exploration of signatures

browsing

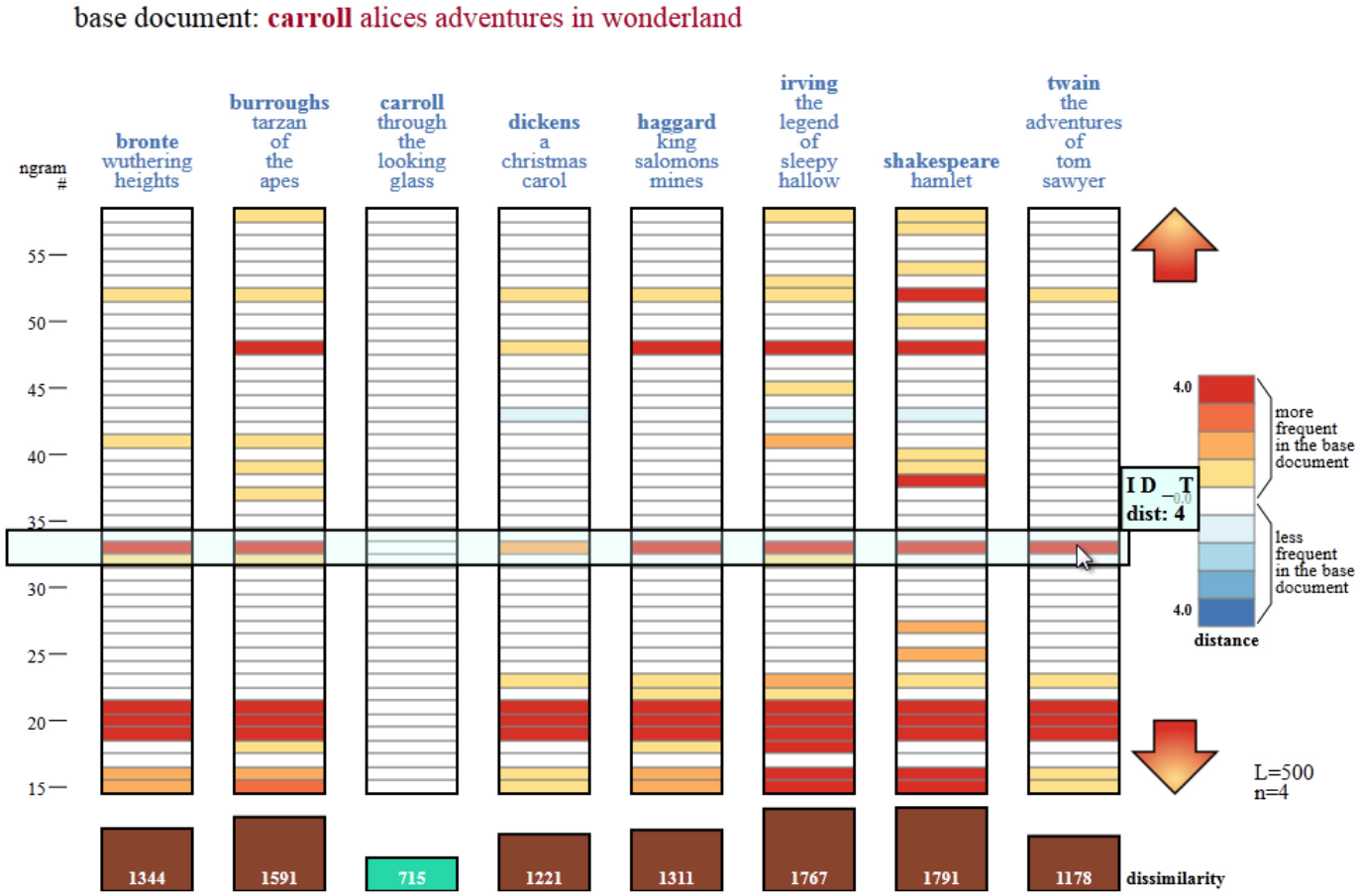


Interactive exploration of signatures

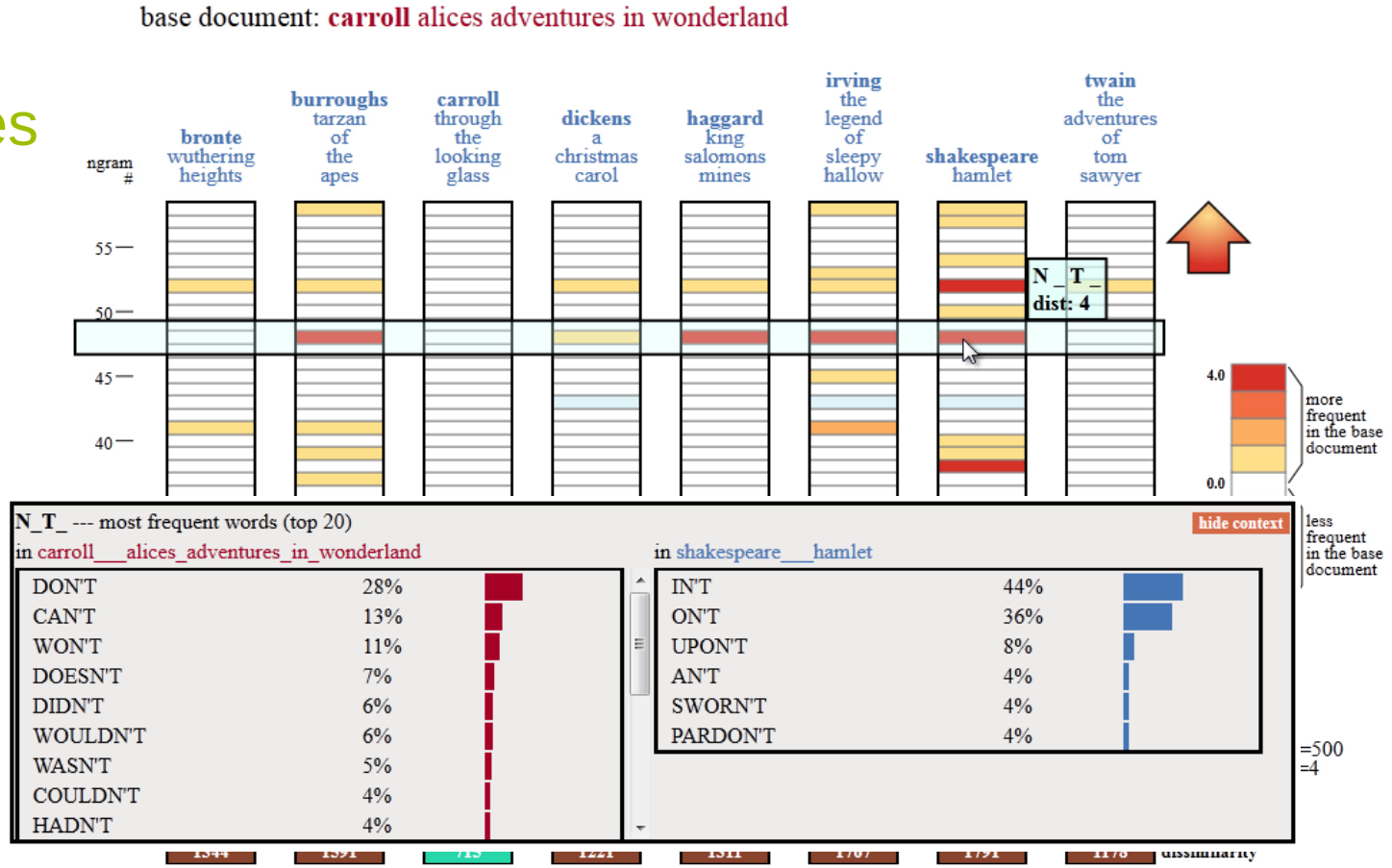
context: most common words



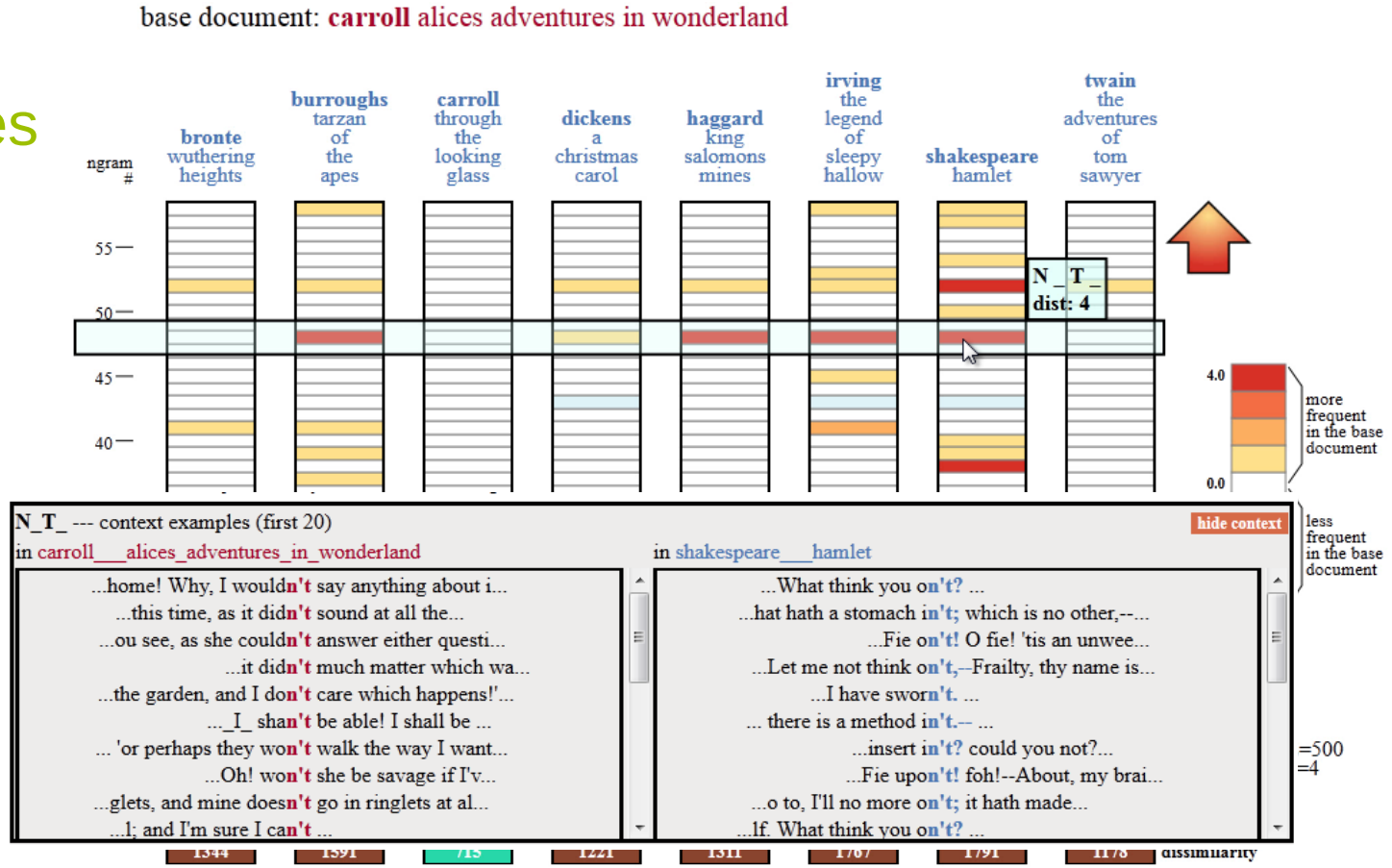
Interactive exploration of signatures



Interactive exploration of signatures



Interactive exploration of signatures

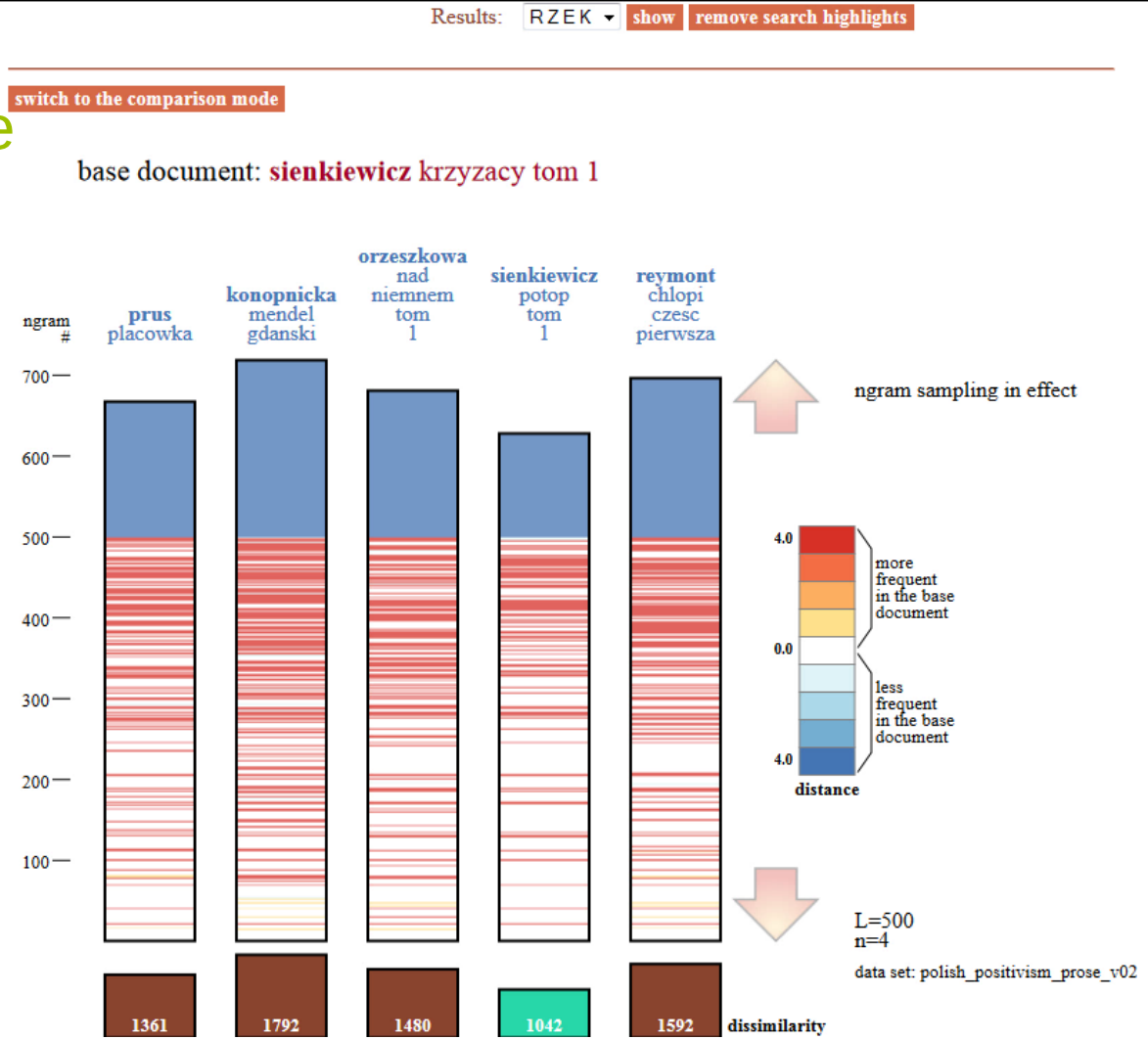


context: concordance style

given n-gram within the text

Language independence

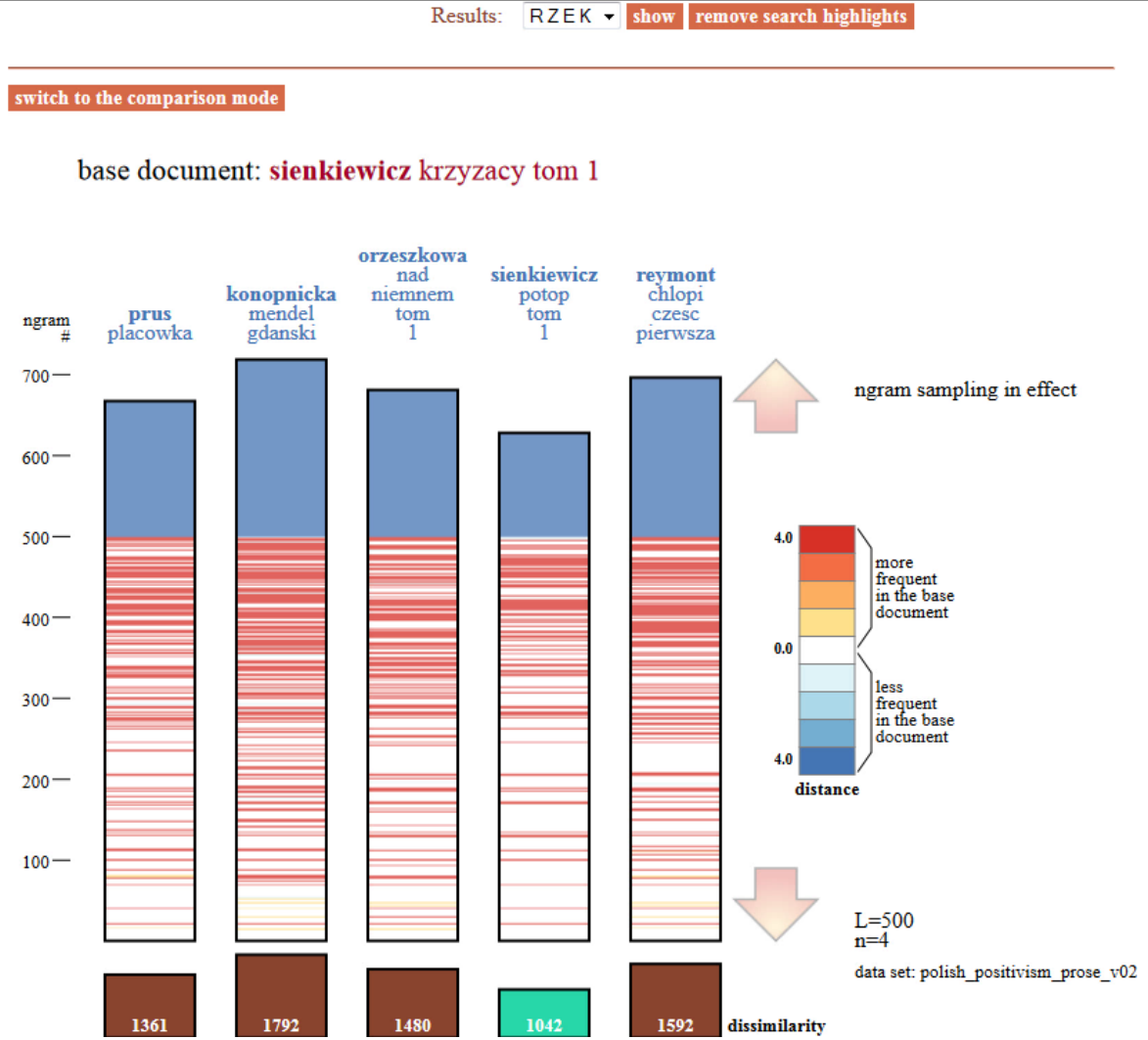
Polish authors



Language independence

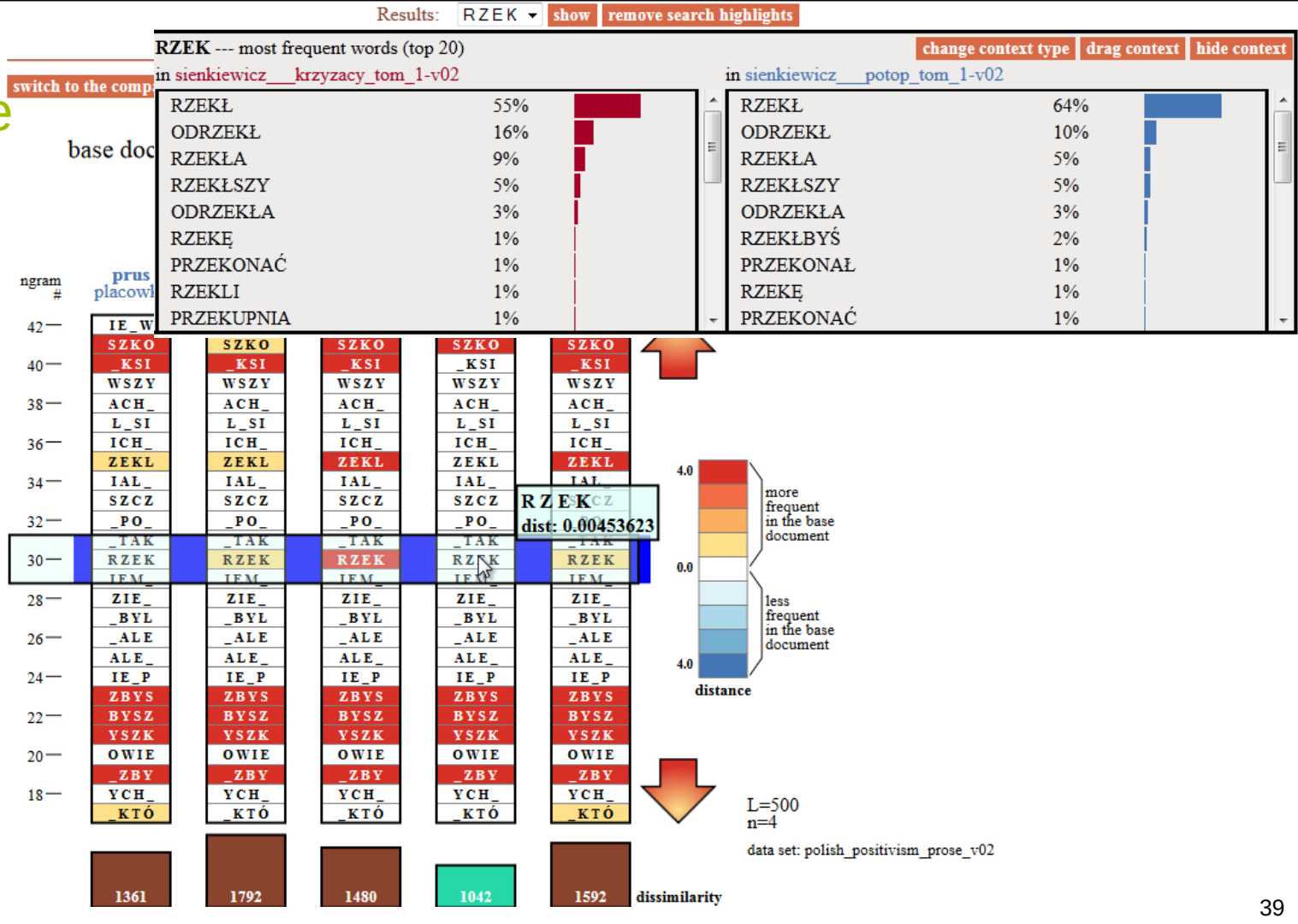
Polish authors

searching for n-grams



Language independence

Polish authors



Motivation for analysis of Mark Twain novels

D. A. Keim and D. Oelke.
Literature Fingerprinting: A New Method for Visual Literary Analysis.
In *Proceedings of the 2007 IEEE Symposium on Visual Analytics Science and Technology,* 2007.

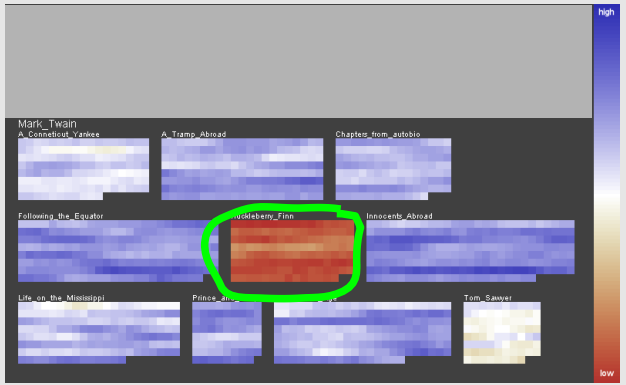
Visual analysis of works of Mark Twain:

Adventures of Huckleberry Finn stands out from the other works of Mark Twain with respect to:

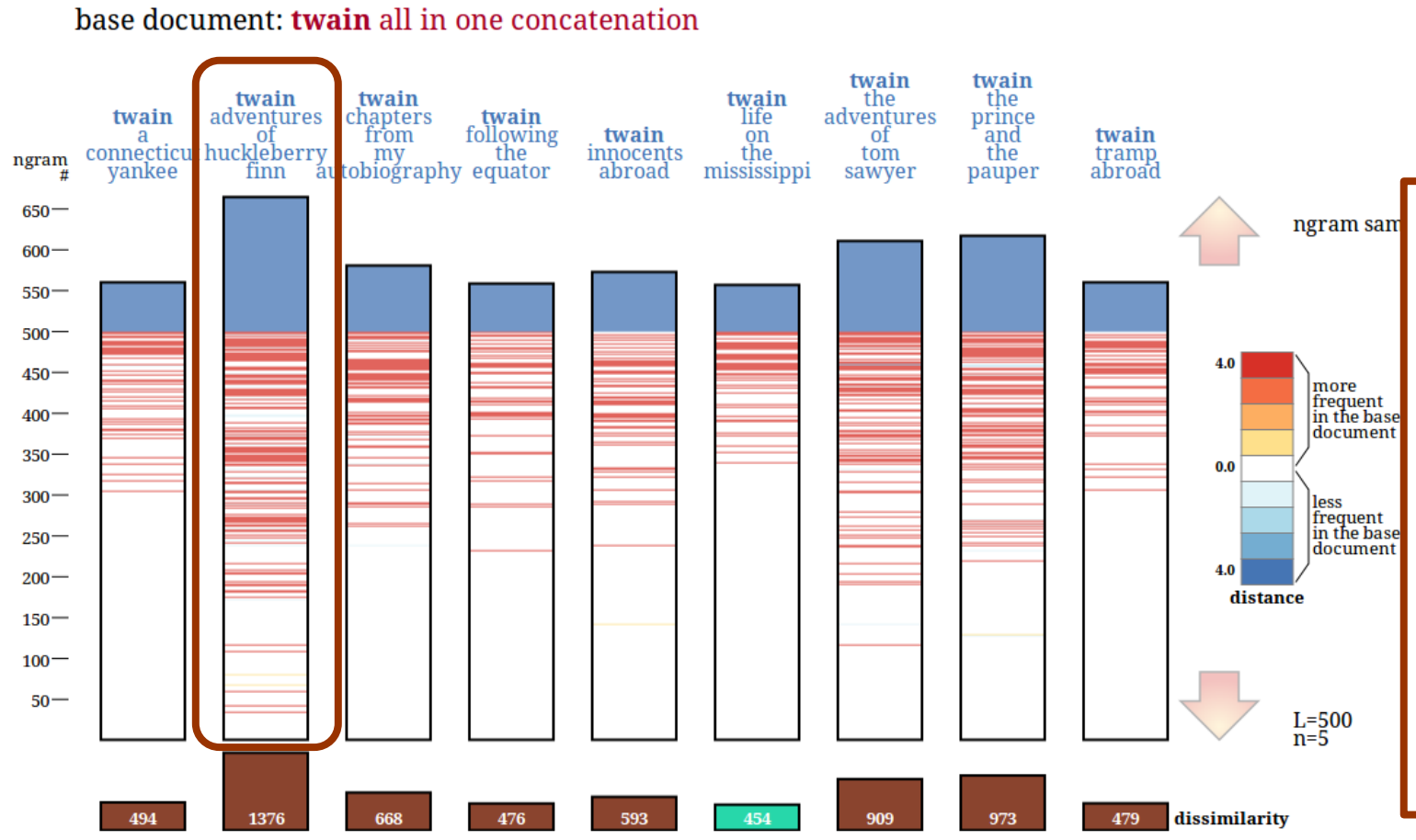
- Function words frequency
- Simpson's index
- Hapax Legomena

Hapax Legomena

Function words (first dimension after PCA)



Example analysis: comparison of novels by Mark Twain

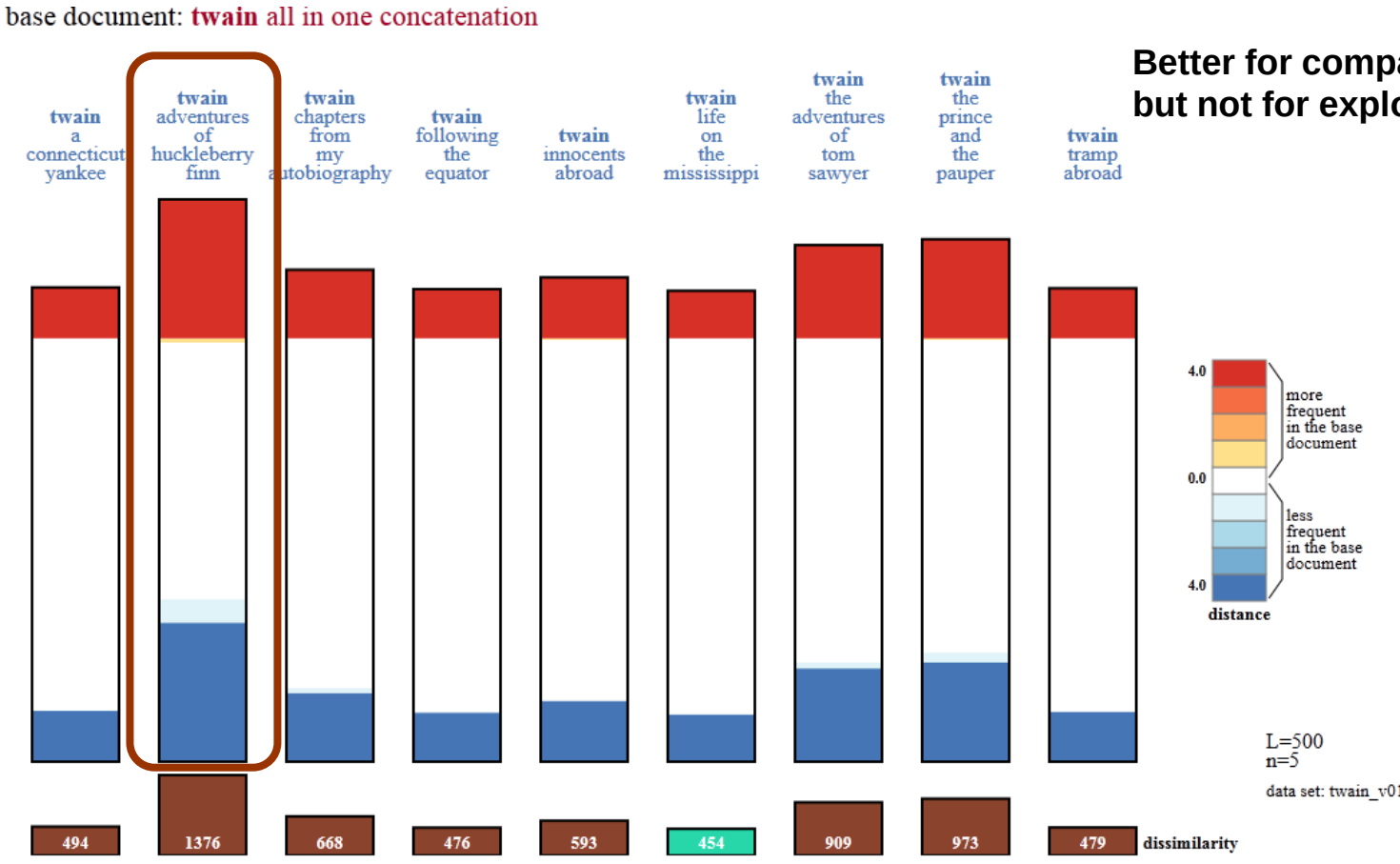


Adventures of Huckleberry Finn

Complementary Comparison View

N-grams ordered separately in each signature, according to their distance

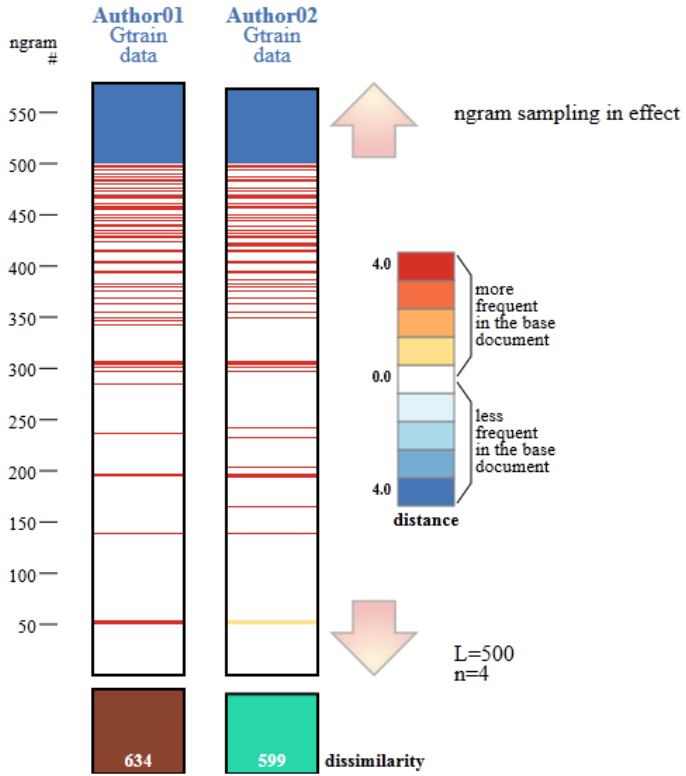
Better for comparison of signatures but not for exploring



Interactively influencing the visualization and the classifier

Ad-hoc Authorship Attribution Competition, 2004

Problem G, sample 02

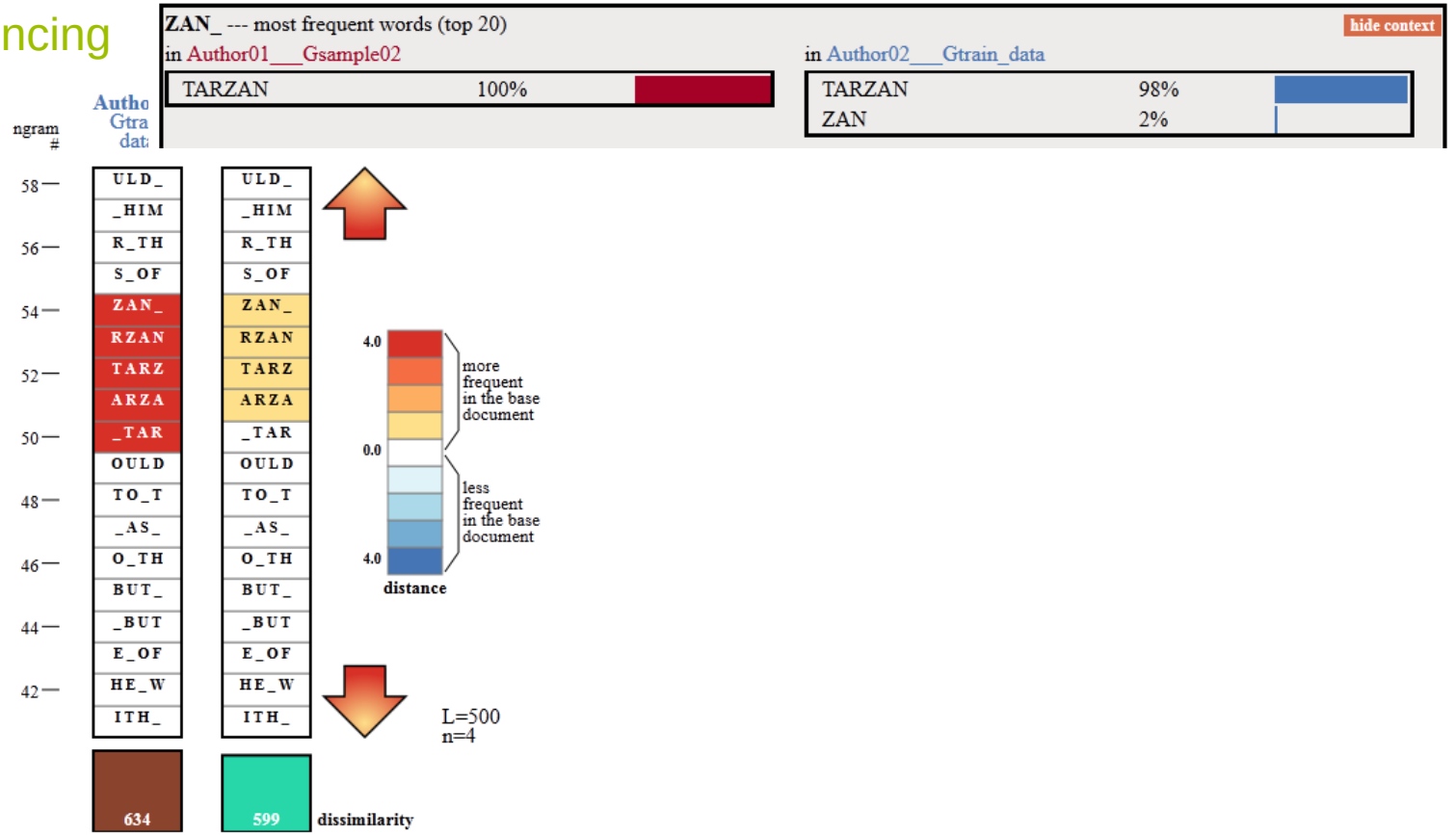


Manual, task-dependent adaptation of the classification process

Interactively influencing the visualization and the classifier

Ad-hoc Authorship Attribution Competition, 2004

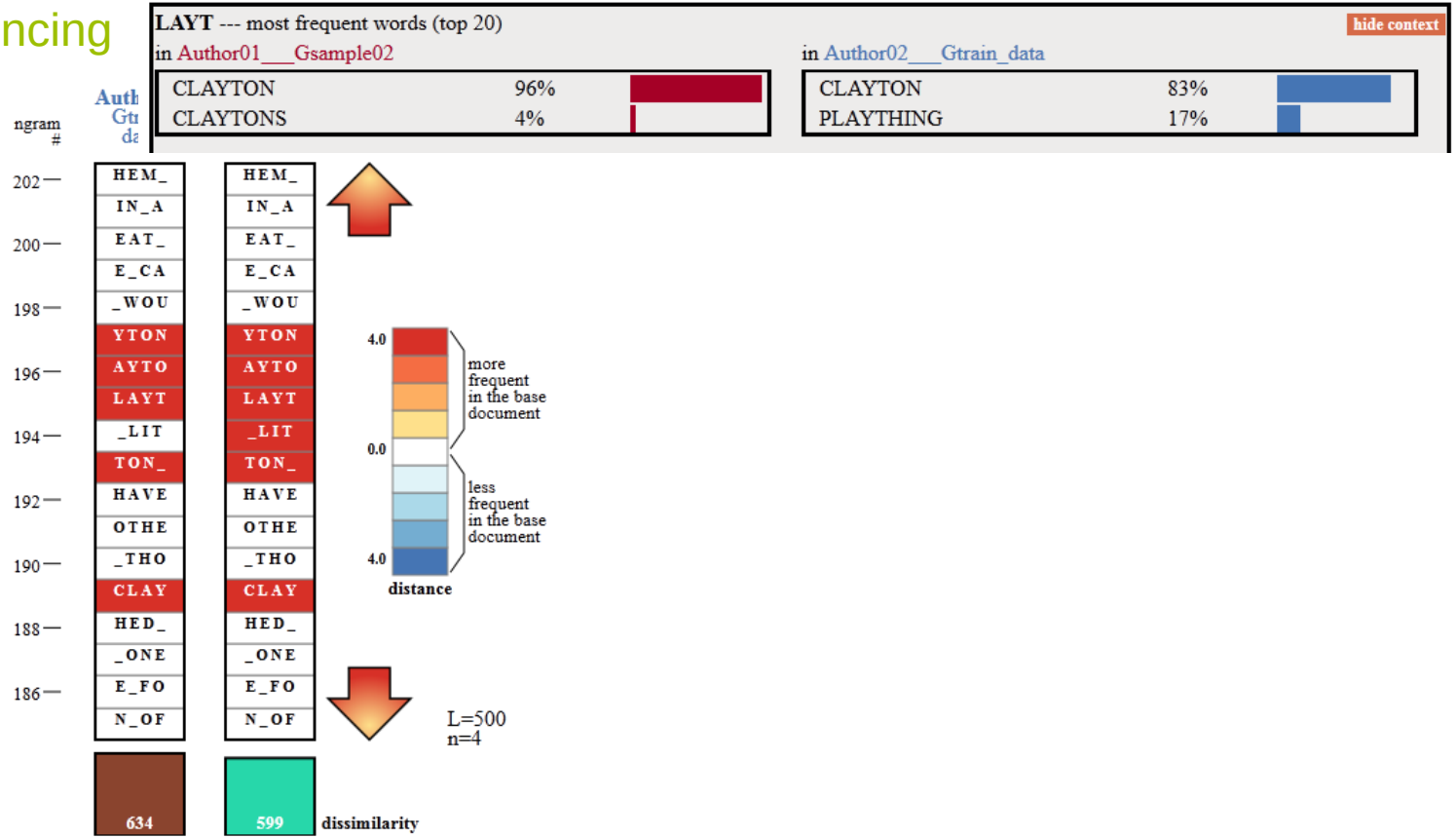
Problem G, sample 02



Interactively influencing the visualization and the classifier

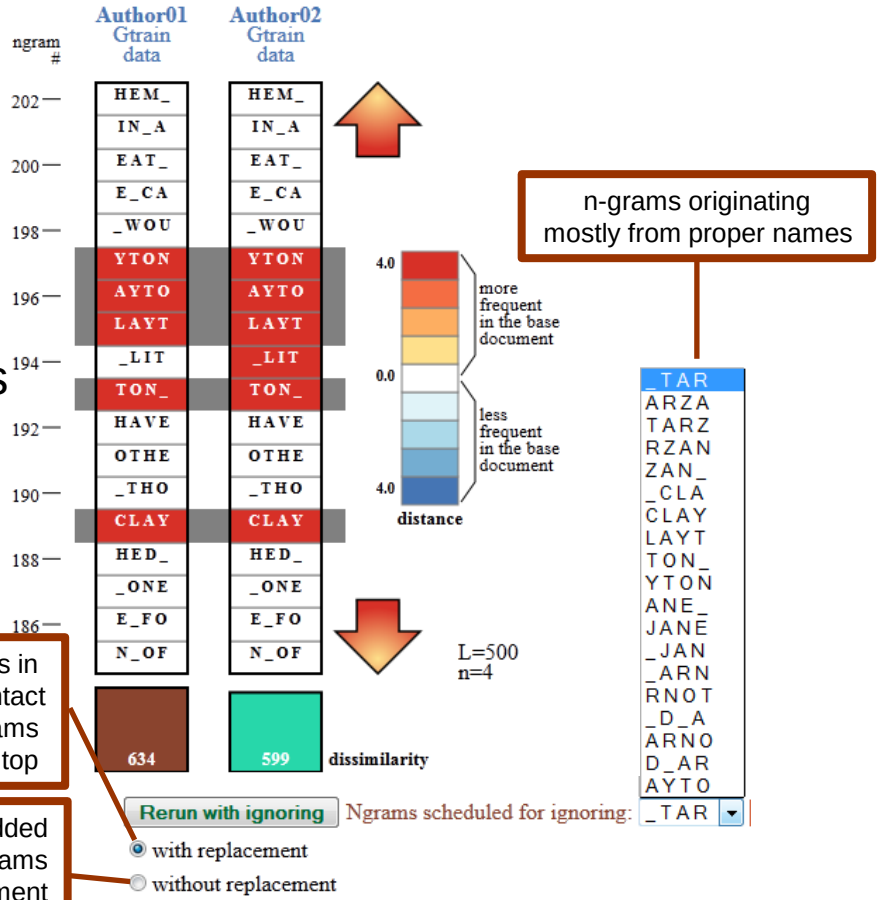
Ad-hoc Authorship Attribution Competition, 2004

Problem G, sample 02



Interactively influencing the visualization and the classifier

ignoring selected n-grams in the base document



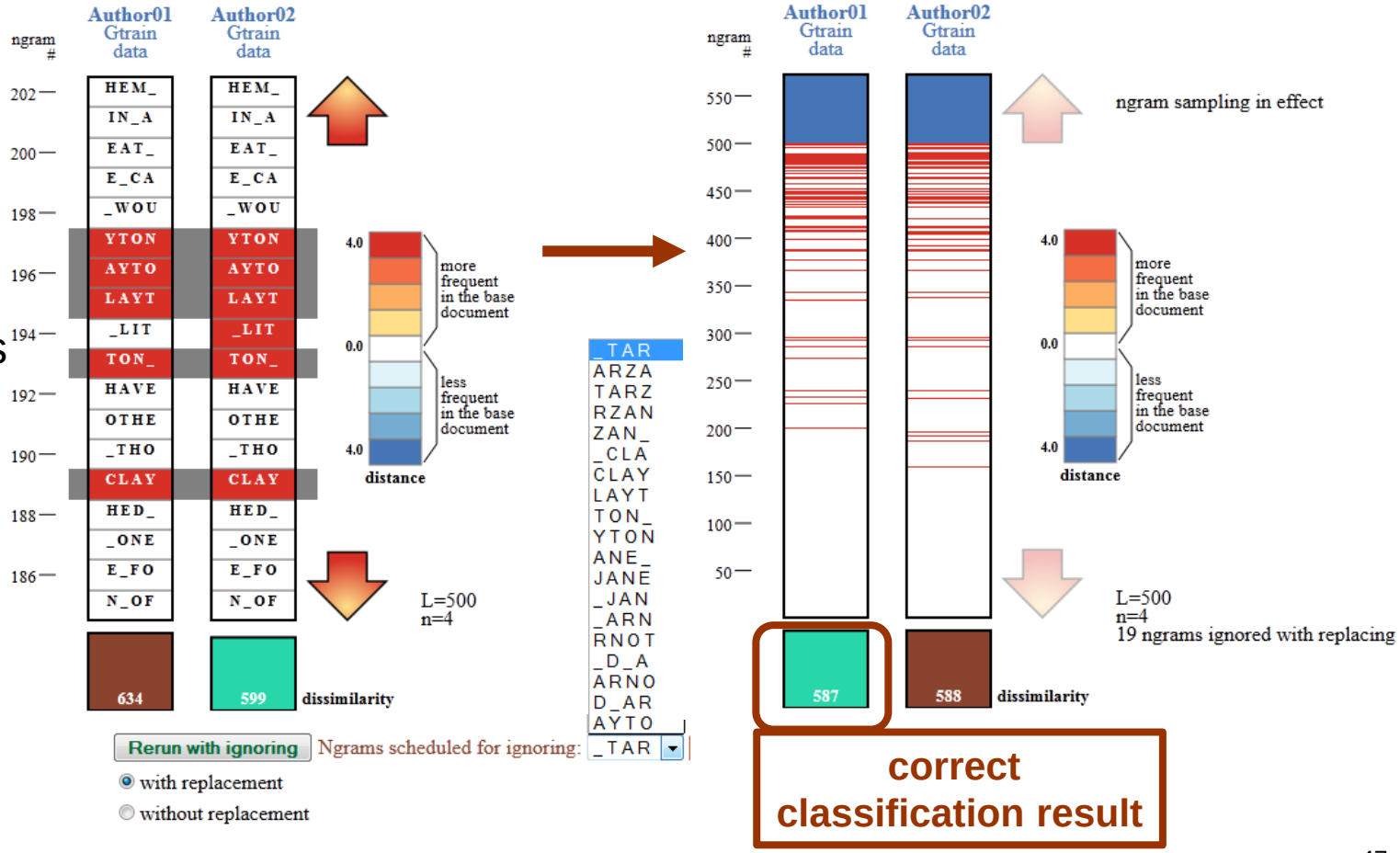
Two options:

the length of the list of n-grams in the base document is kept Intact by adding less frequent n-grams at the top

no new n-grams are added the list of n-grams for the base document becomes shorter

Interactively influencing the visualization and the classifier

ignoring selected n-grams in the base document



Thank you!