

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

Relative N-Gram Signatures

Interactive classification of a document

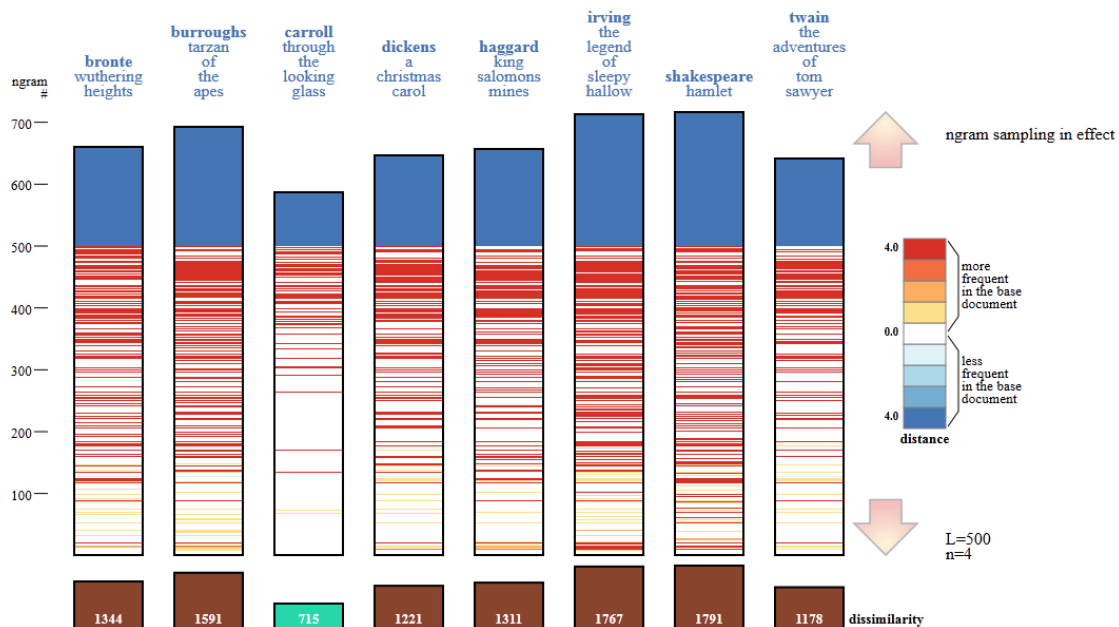
Who wrote this book?

Analysis of characteristics of a document

What are the characteristics of the author's style?

Language independent method

base document: **carroll** *alices adventures in wonderland*



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:

**n=4
4-grams**

ALIC

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
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

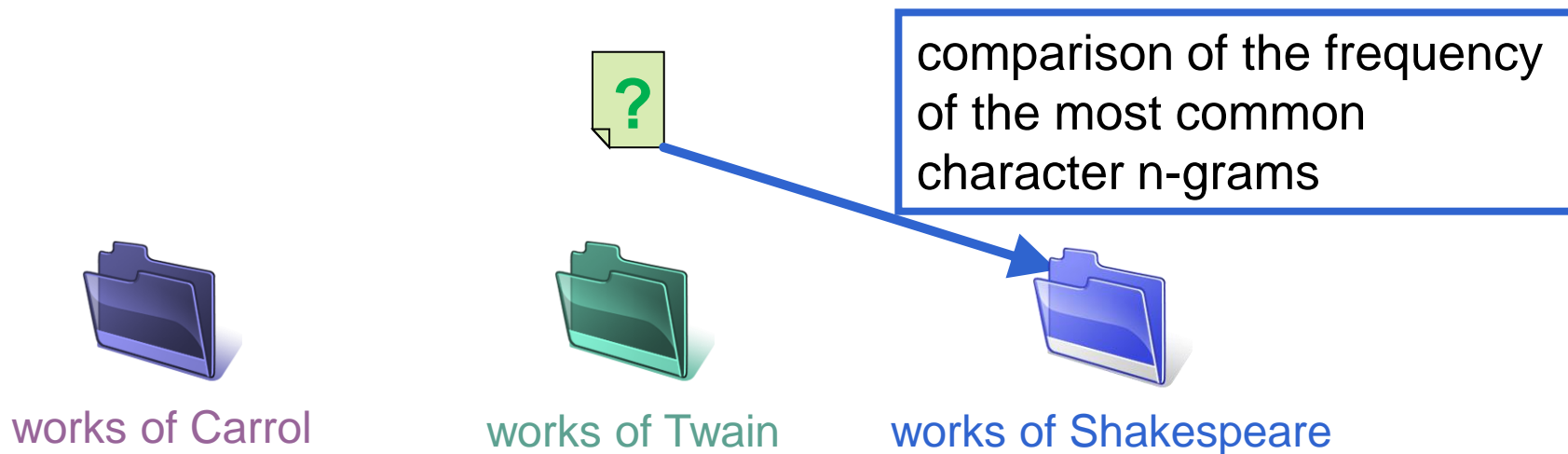
Vlado Kešelj, Fuchun Peng, Nick Cercone, and Calvin Thomas.

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

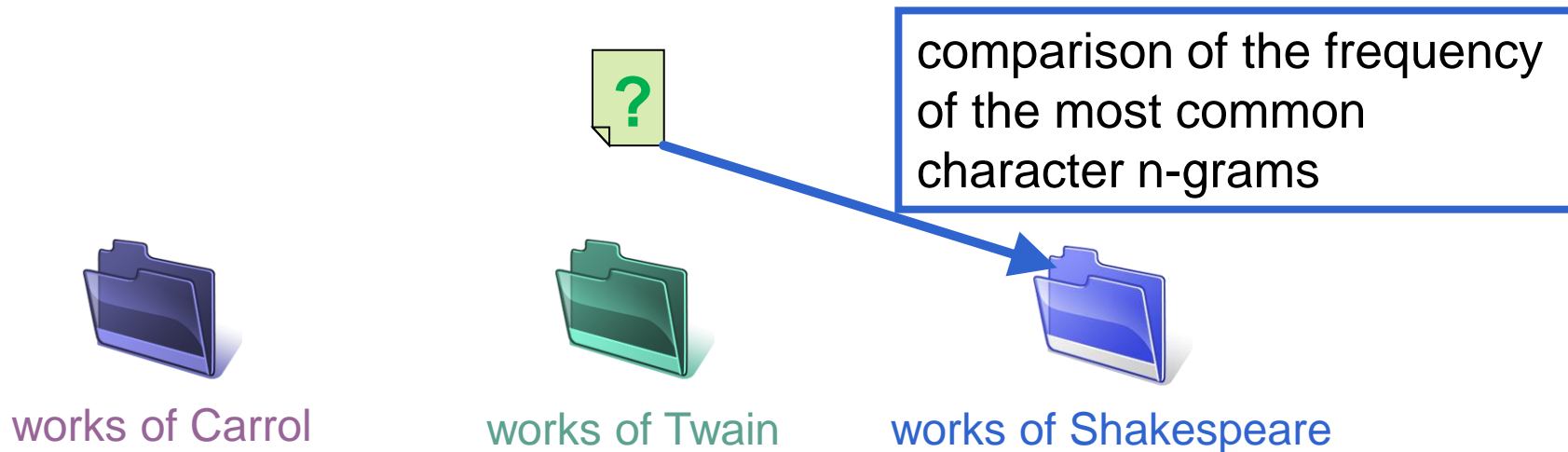
Vlado Kešelj, Fuchun Peng, Nick Cercone, and Calvin Thomas.

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

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

$n=4, L=6$

n-gram

normalized frequency

CNG Classifier - Dissimilarity

Profile

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

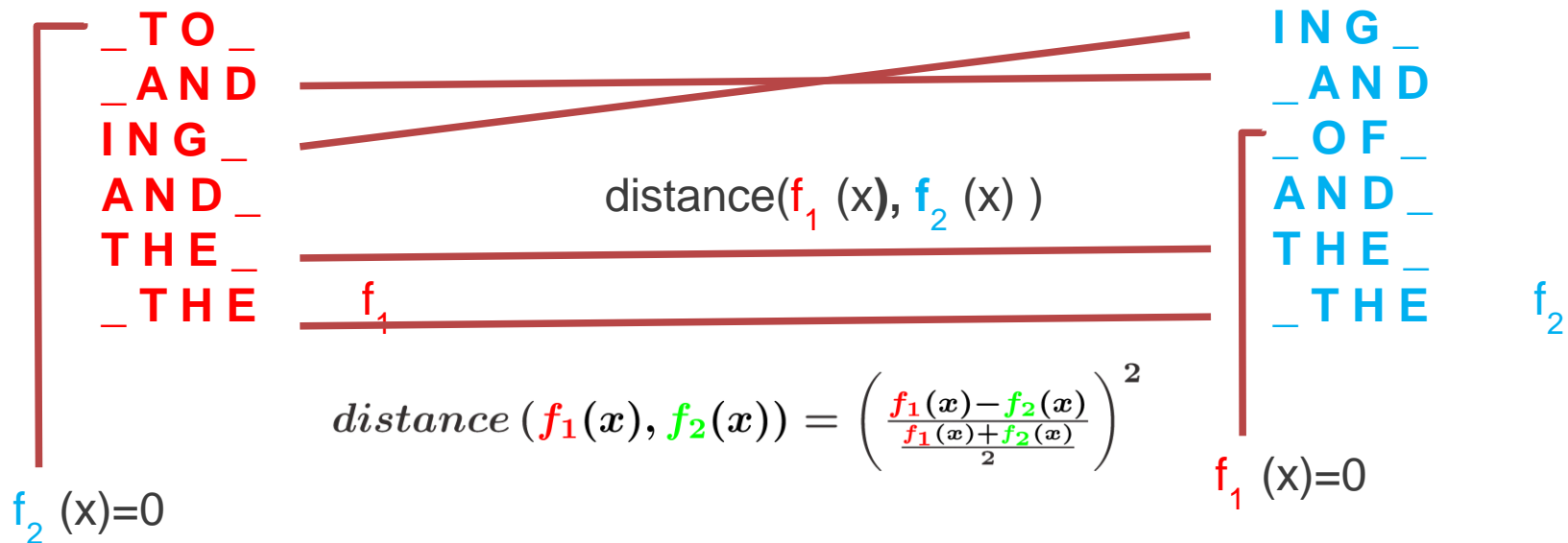
$n=4, L=6$

document 1:

Alice's Adventures in the Wonderland
by Lewis Carroll

document 2:

Tarzan of the Apes
by Edgar Rice Burroughs



CNG Classifier - Dissimilarity

Profile

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

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

$n=4, L=6$

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

f_2

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

Motivation

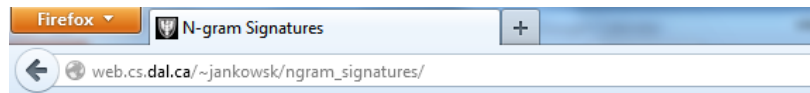
- text visualization on the language-independent level of character n-grams
 - similarity of documents
 - characteristics of documents
- visualization of the CNG classifier
 - “reasons” for the classification result
 - possibility of influencing the classification

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 Signatures

Plot Data

Define data set

Select the set of documents

authorship attribution set 1

Select the base document

bronte wuthering_heights

Select the documents

to compare with the base document

- bronte wuthering_heights
- burroughs tarzan_of_the_apes
- burroughs warlord_of_mars
- carroll alices_adventures_in_wonderland
- carroll through_the_looking_glass
- cleland memoirs_of_fanny_hill
- dickens a_christmas_carol
- dickens a_tale_of_two_cities
- haggard king_salomons_mines
- irving the_legend_of_sleepy_hallow
- shakespeare hamlet
- twain the_adventures_of_tom_sawyer

Select the length of n-grams

3

Select the length of profiles

500

Relative N-Gram Signature

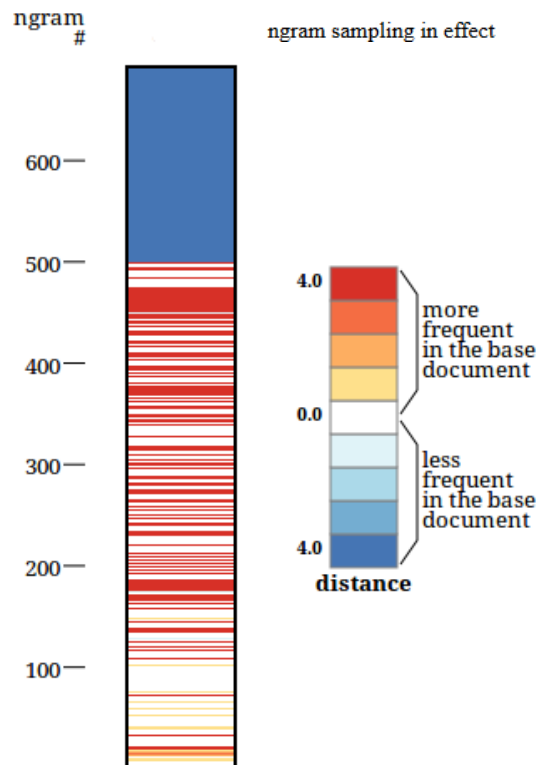
Relative signature of
Tarzan of the Apes
by Burroughs

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

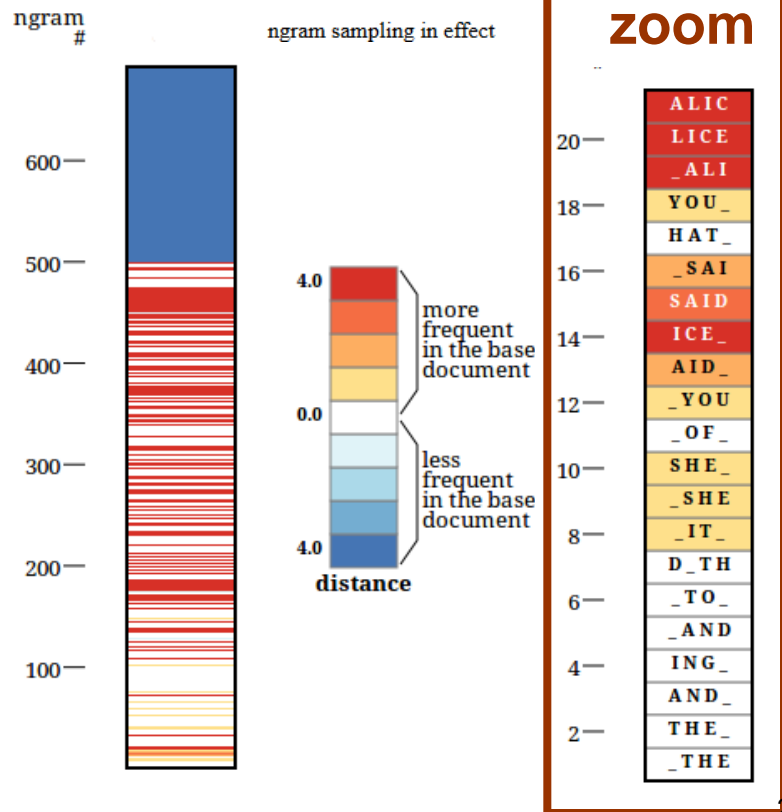
Relative signature of
Tarzan of the Apes
by Burroughs

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



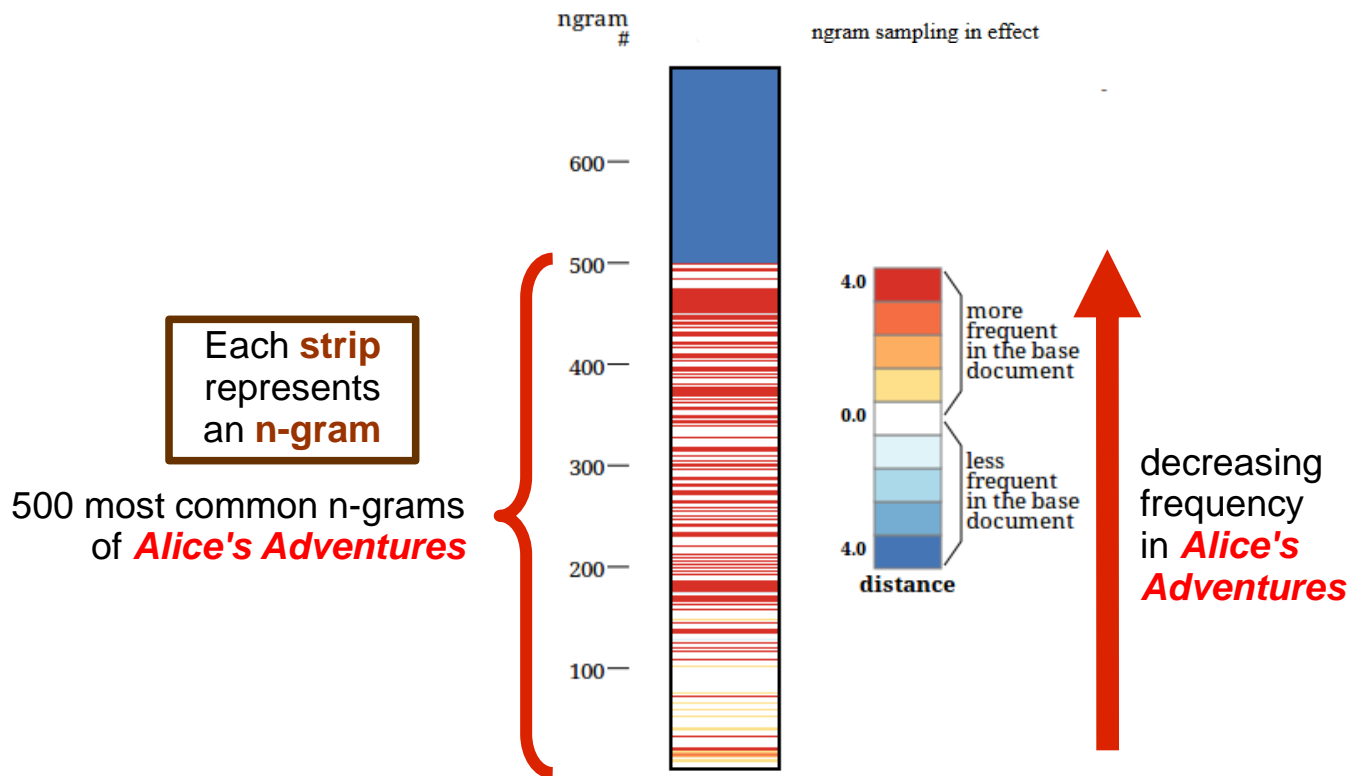
Relative N-Gram Signature

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)



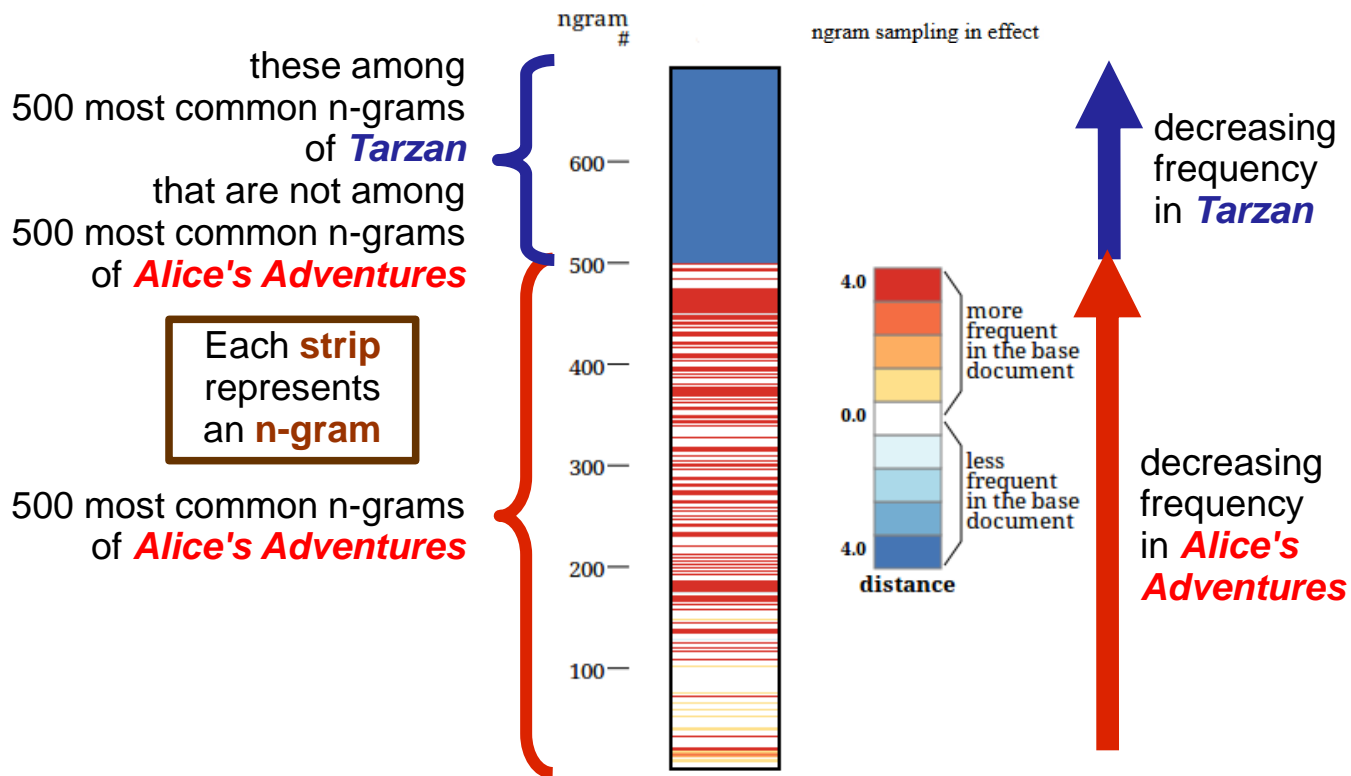
Relative N-Gram Signature

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)



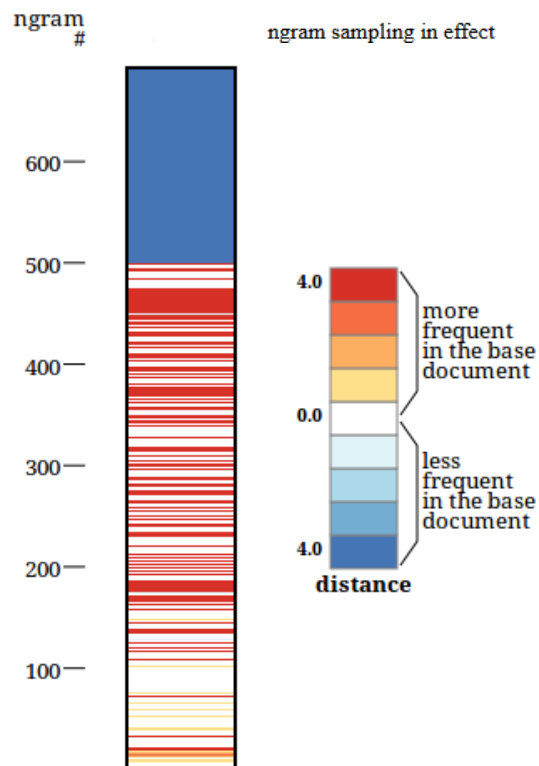
Relative N-Gram Signature

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)

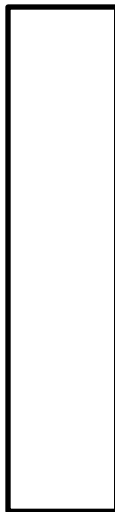


Color:
"distance"
of two
documents
with respect
to this n-gram

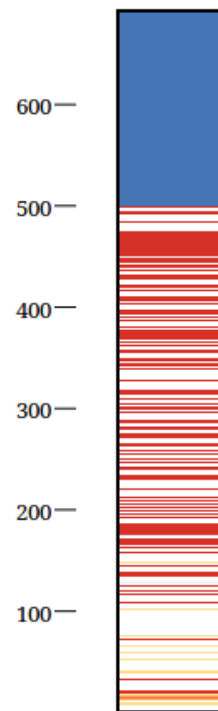
Relative N-Gram Signature

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

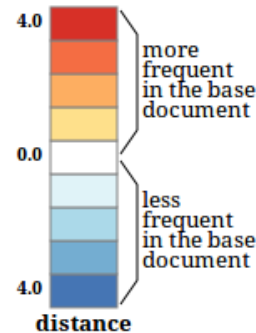
Signature of a document with respect to itself



ngram #



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



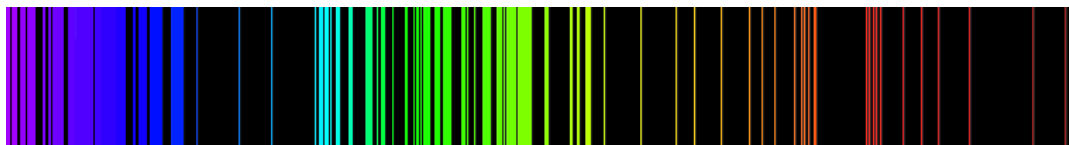
Relative N-Gram Signature

Visual metaphor

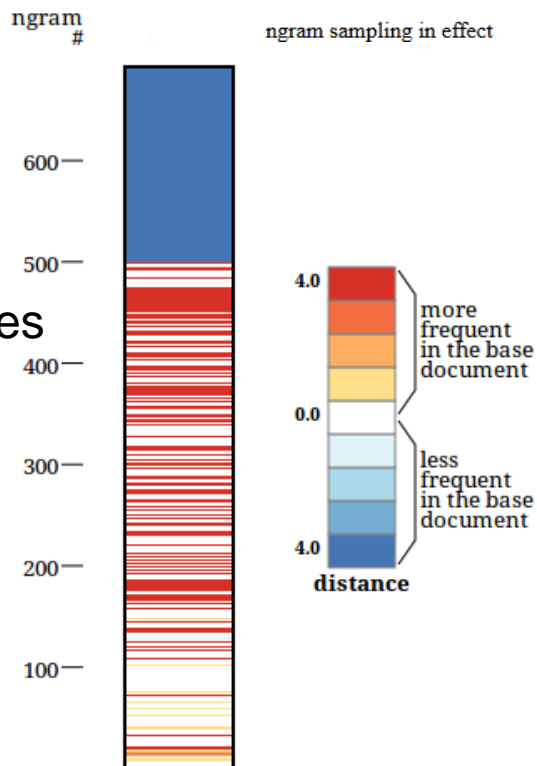
Inspiration: emission spectrum

spectrum of frequencies

of electromagnetic emissions by atoms or molecules



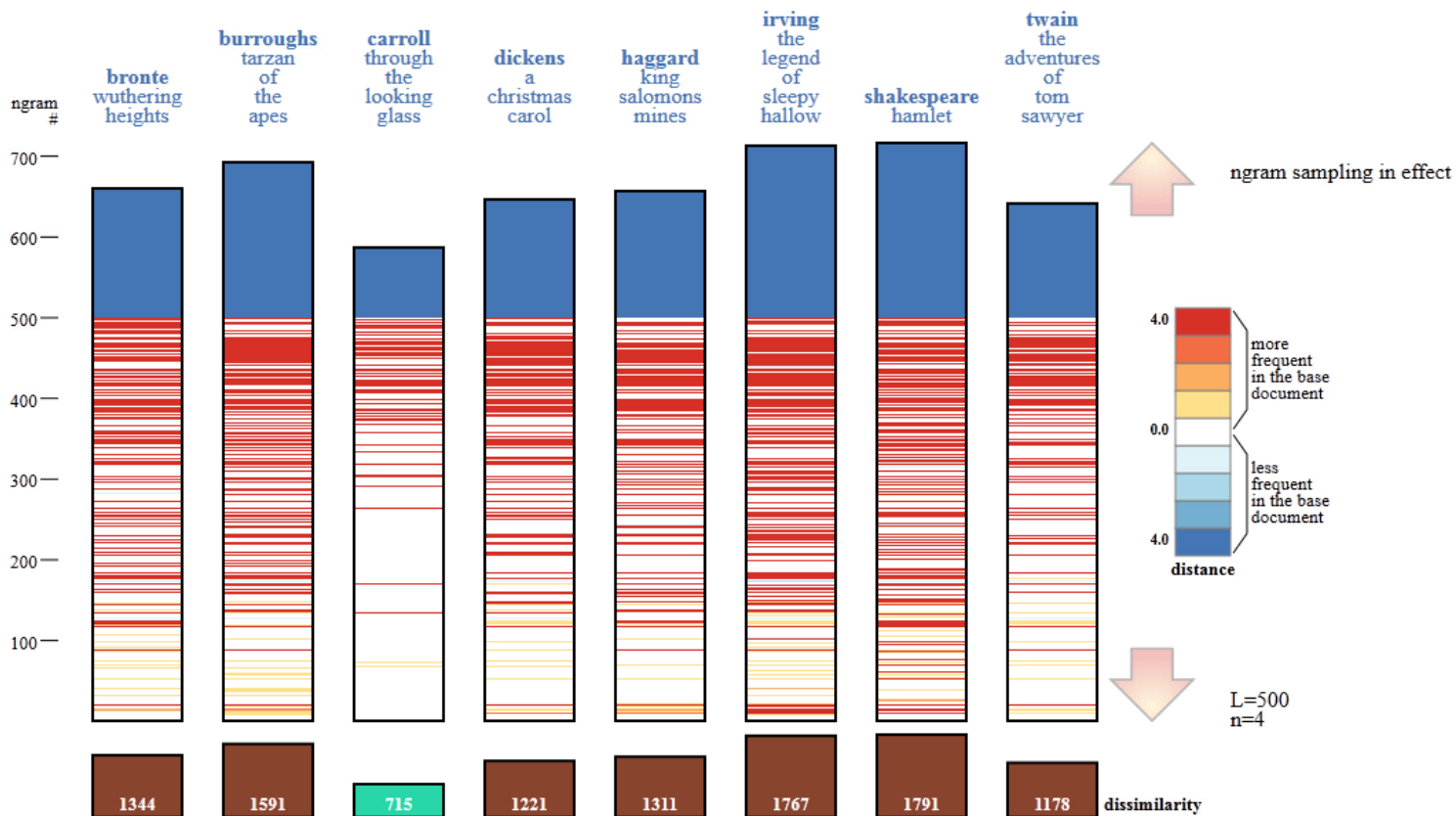
picture from Wikipedia



Sequence of signatures

base document: **carroll** alices adventures in wonderland

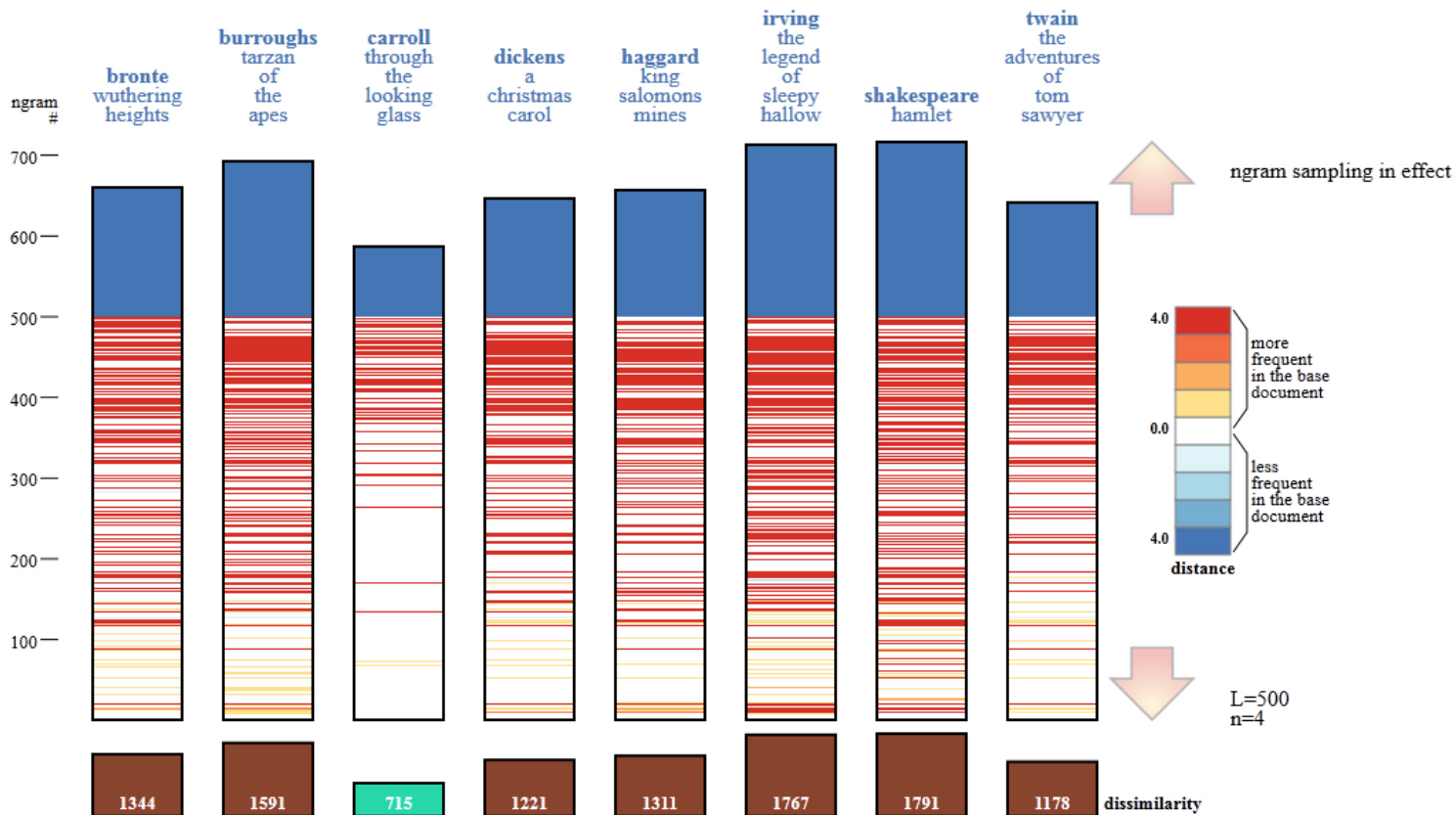
scenario
authorship analysis



Sequence of signatures

base document: **carroll** alices adventures in wonderland

The same
base
document

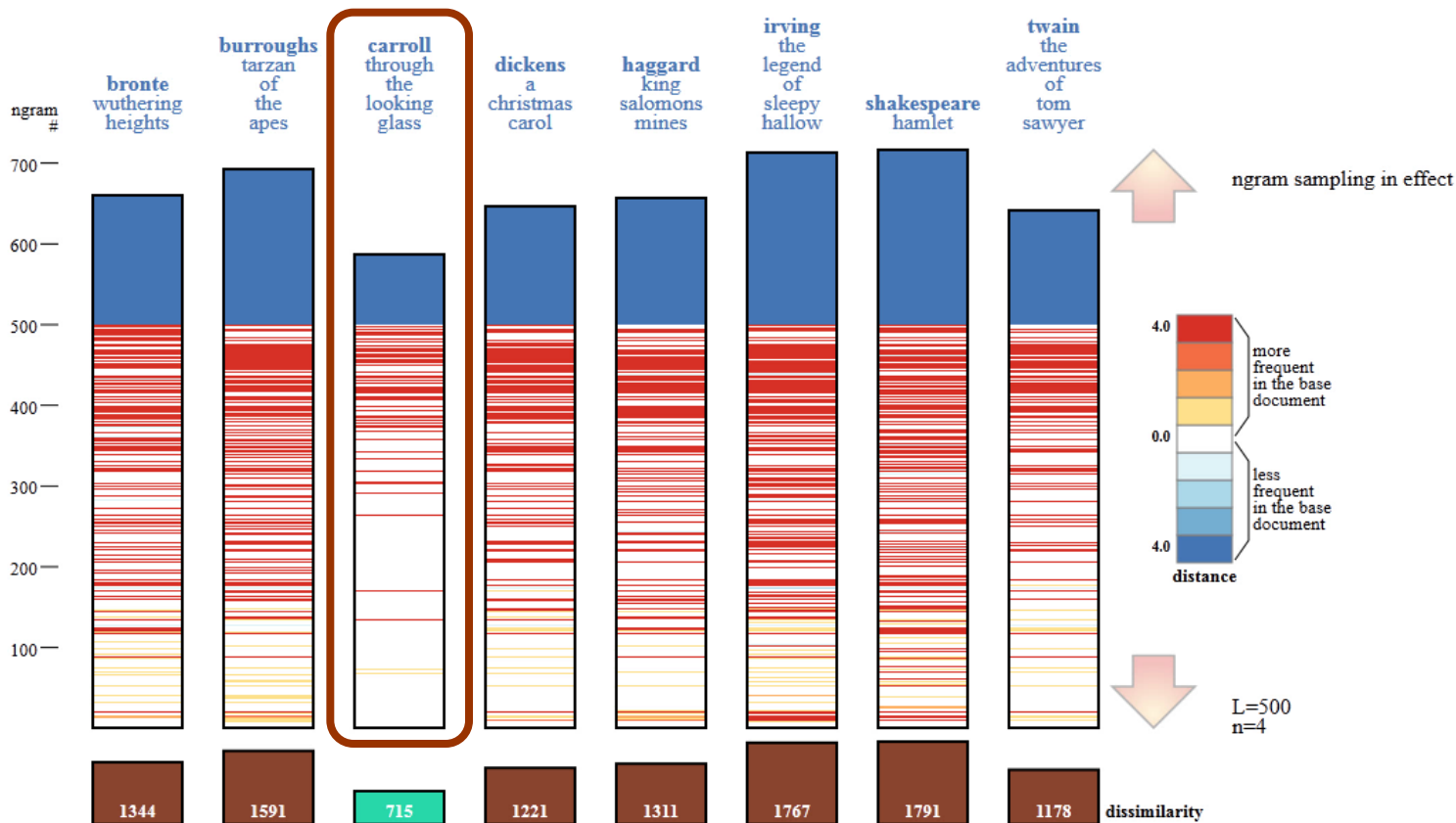


Sequence of signatures

base document: **carroll** alices adventures in wonderland

Signature of the most similar document

Carroll's "Through the looking glass"

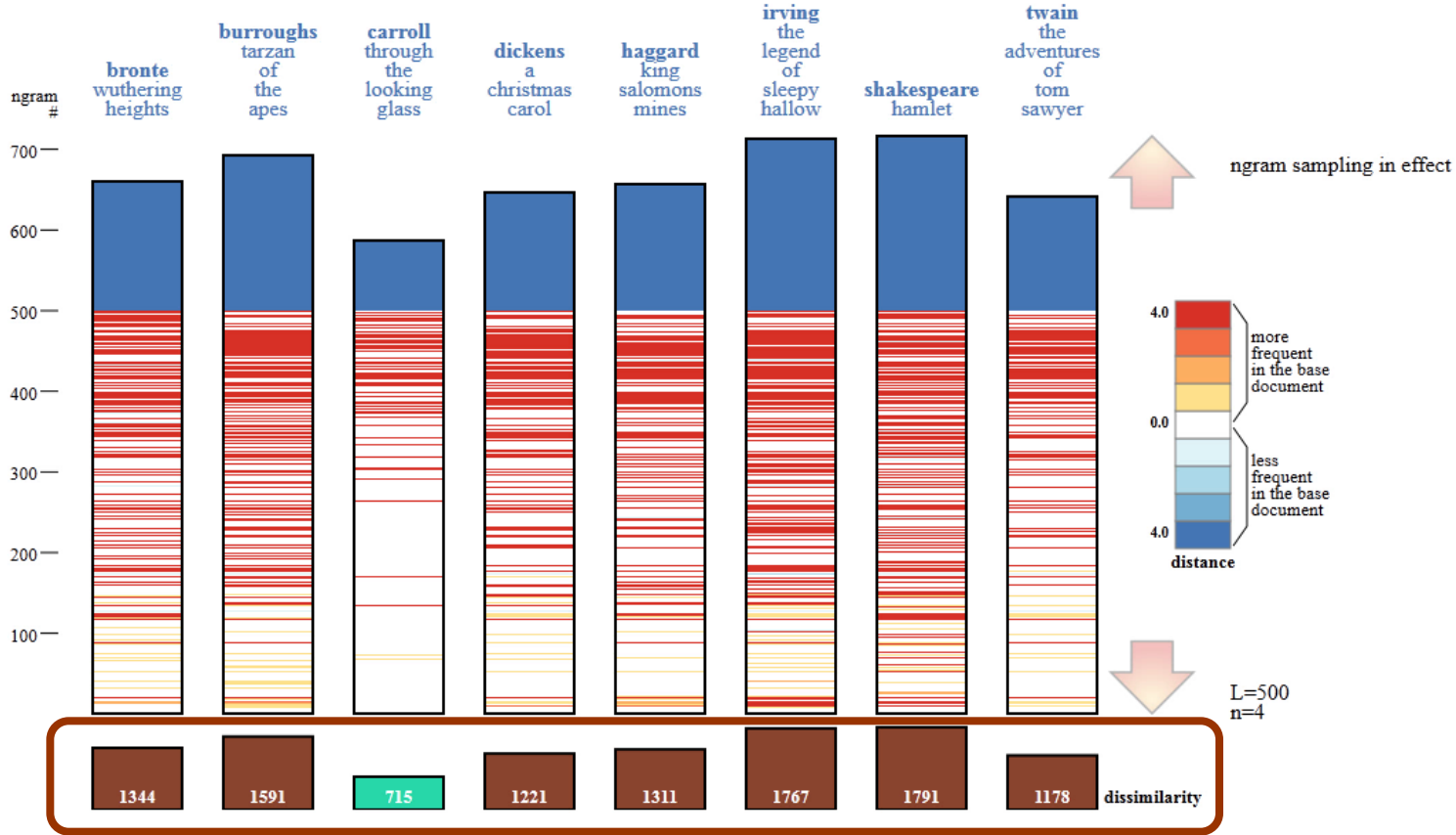


Sequence of signatures

base document: **carroll** alices adventures in wonderland

CNG
dissimilarity
score

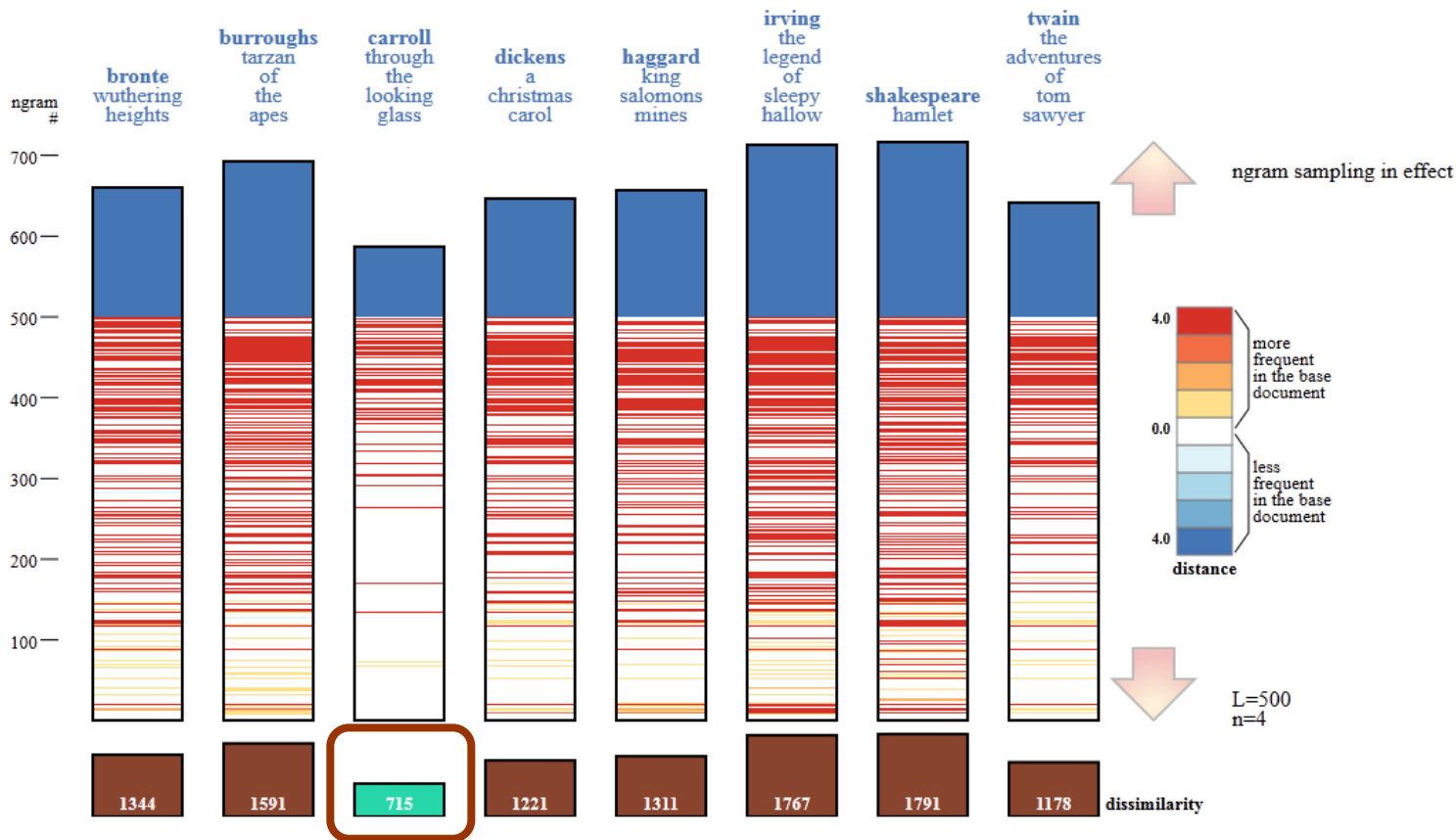
sum of the distances
 over all n-grams in a
 signature



Sequence of signatures

base document: **carroll** alices adventures in wonderland

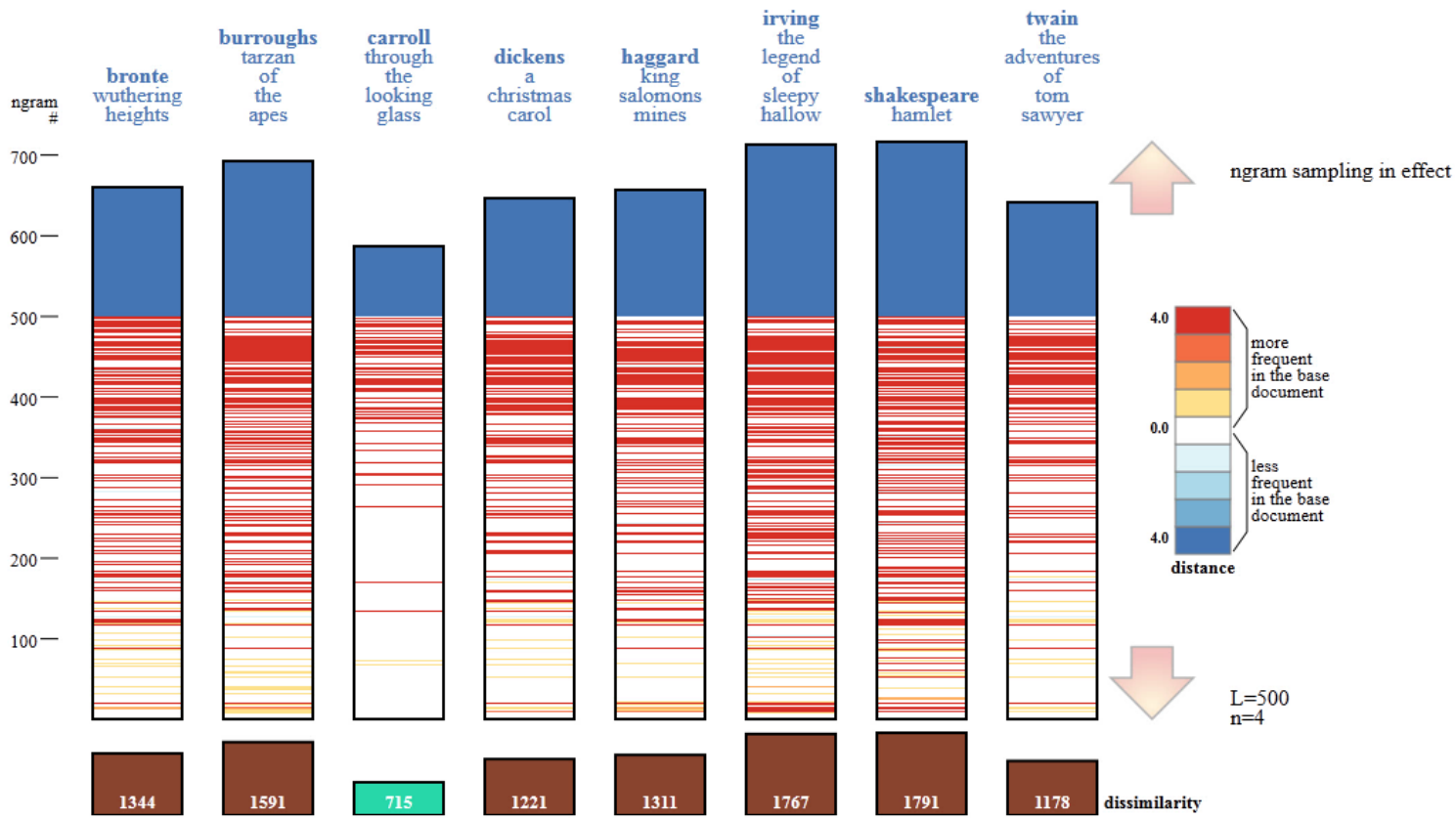
minimum
dissimilarity
=
classifier
result



Sequence of signatures

base document: **carroll** alices adventures in wonderland

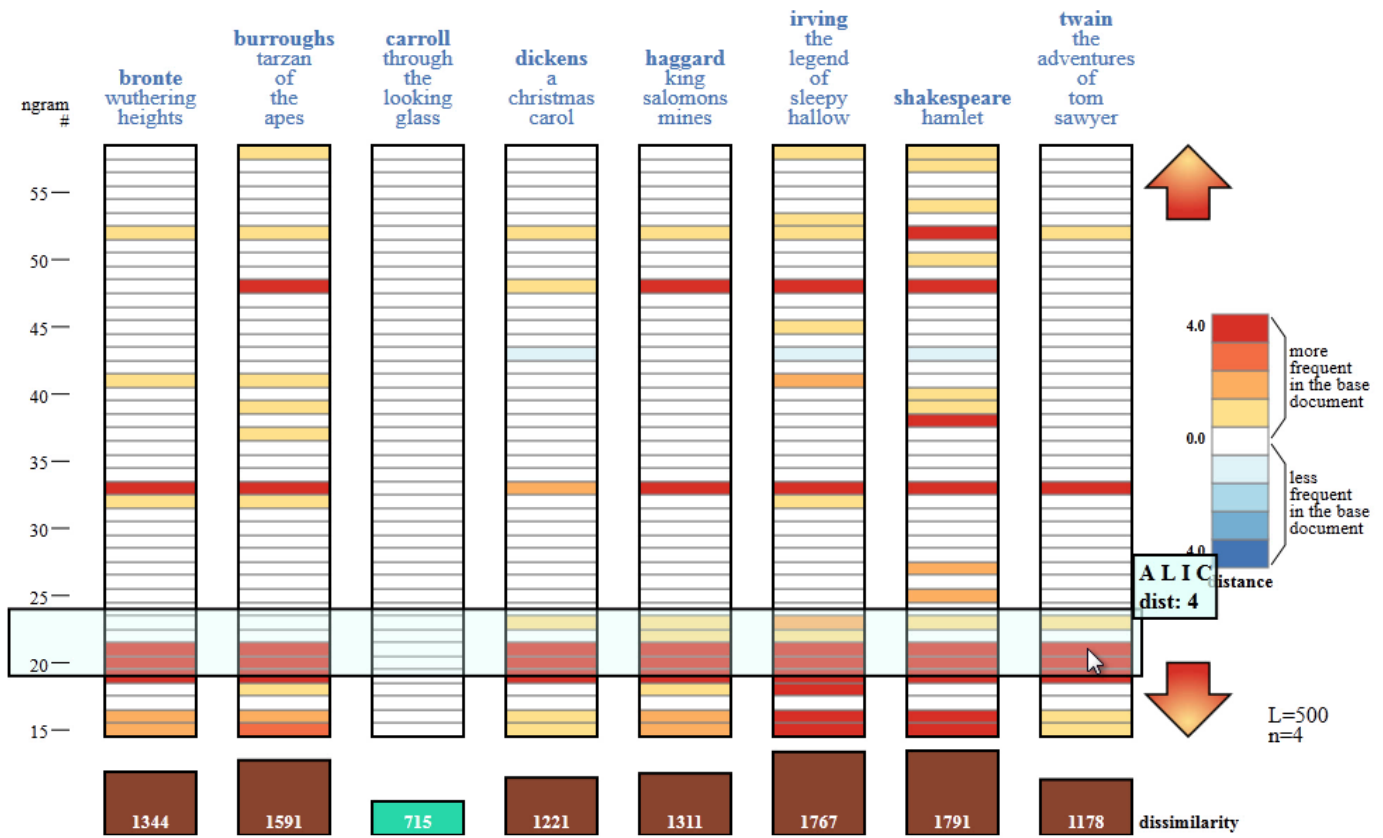
zooming in



Interactive exploration of signatures

base document: **carroll** alices adventures in wonderland

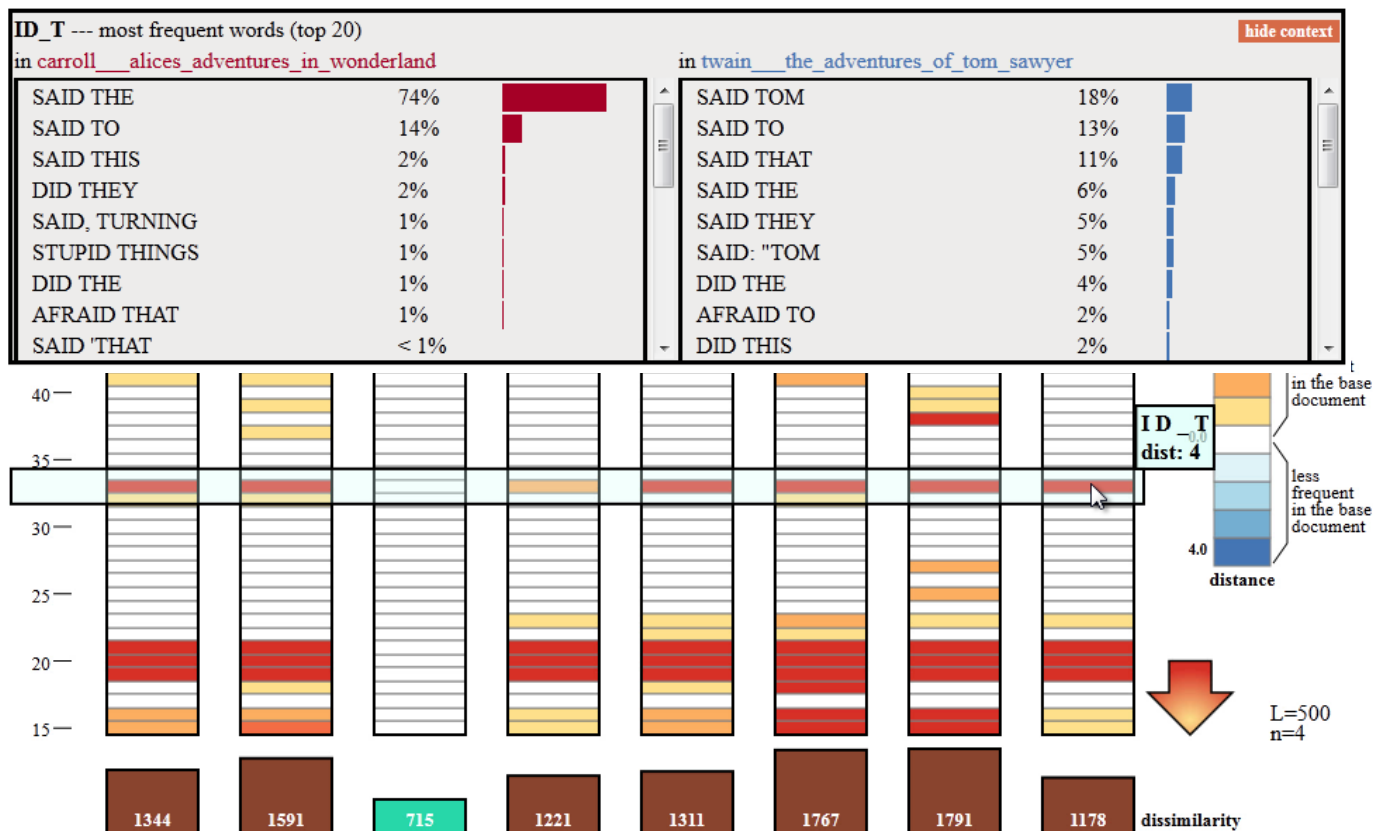
browsing



Interactive exploration of signatures

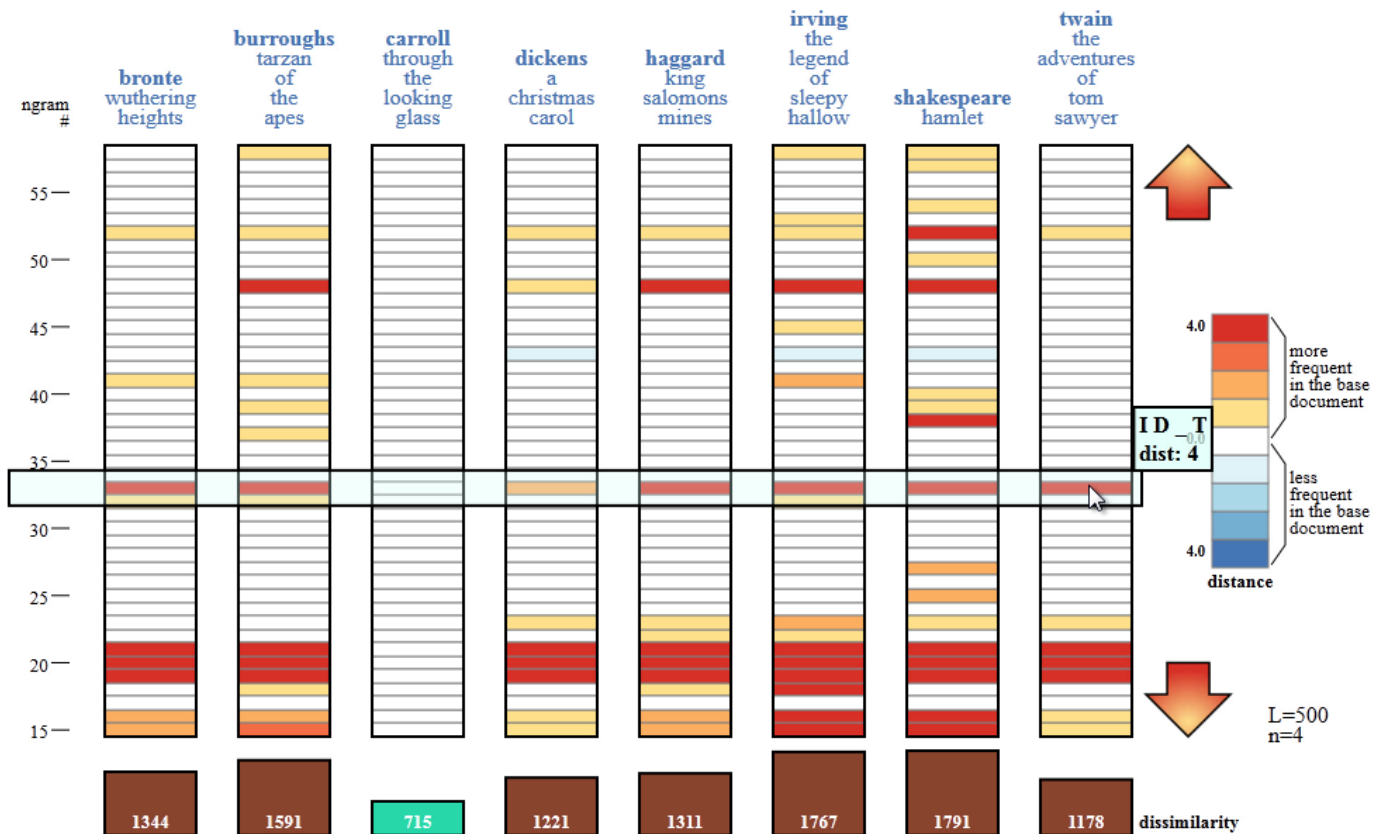
base document: **carroll** alices adventures in wonderland

**context:
most common
words**



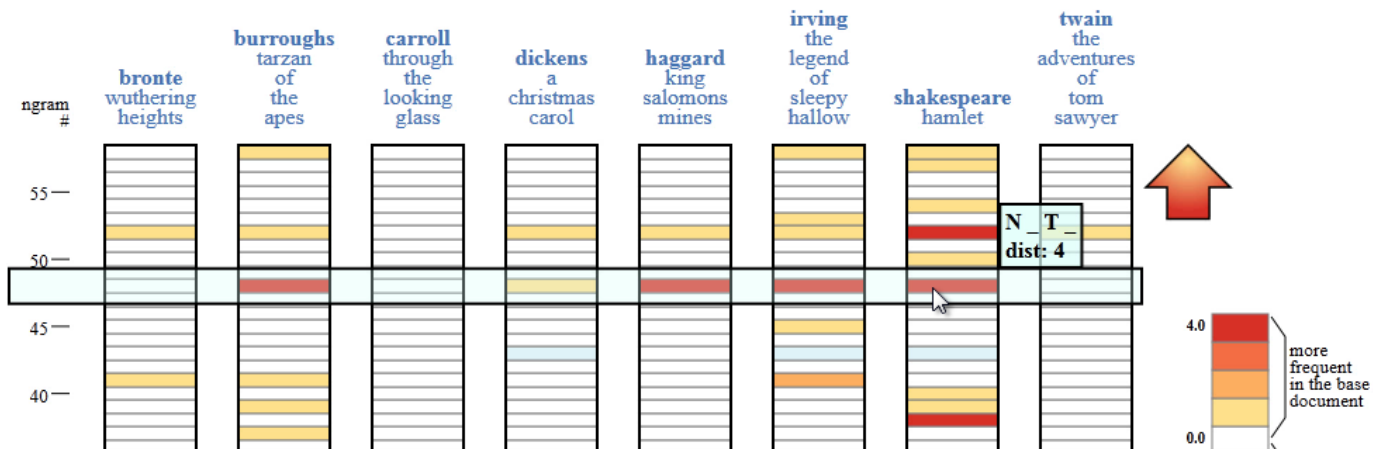
Interactive exploration of signatures

base document: **carroll** alices adventures in wonderland



Interactive exploration of signatures

base document: **carroll** *alices adventures in wonderland*



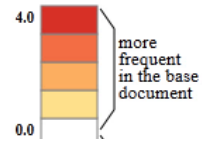
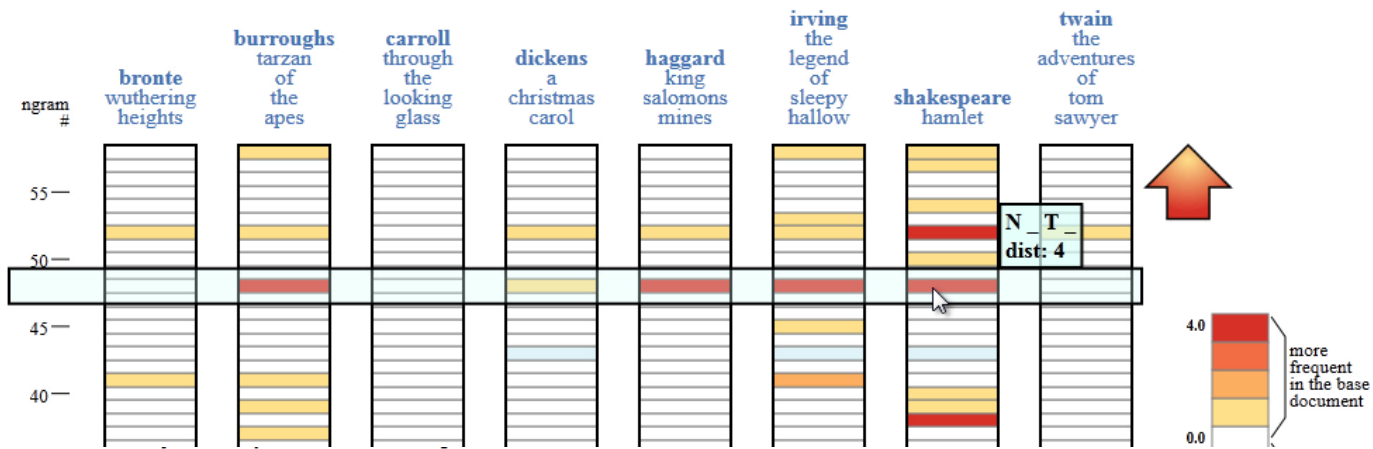
N_T_ --- most frequent words (top 20) hide context

in <u>carroll</u> <u>alices adventures in wonderland</u>		in <u>shakespeare</u> <u>hamlet</u>	
DON'T	28%	IN'T	44%
CAN'T	13%	ON'T	36%
WON'T	11%	UPON'T	8%
DOESN'T	7%	AN'T	4%
DIDN'T	6%	SWORN'T	4%
WOULDN'T	6%	PARDON'T	4%
WASN'T	5%		
COULDN'T	4%		
HADN'T	4%		

=500
=4

Interactive exploration of signatures

base document: **carroll** alices adventures in wonderland



less frequent in the base document

=500
=4

**context:
concordance
style**

**given n-gram
within the text**

N_T_ --- context examples (first 20)

in <u>carroll__alices_adventures_in_wonderland</u>	in <u>shakespeare__hamlet</u>
...home! Why, I wouldn't say anything about i...	...What think you on't? ...
...this time, as it didn't sound at all the...	...hat hath a stomach in't; which is no other,--...
...ou see, as she couldn't answer either questi...	...Fie on't! O fie! 'tis an unwee...
...it didn't much matter which wa...	...Let me not think on't,--Frailty, thy name is...
...the garden, and I don't care which happens!...	...I have sworn't. ...
..._I shan't be able! I shall be there is a method in't.-- ...
... 'or perhaps they won't walk the way I want...	...insert in't? could you not?...
...Oh! won't she be savage if I v...	...Fie upon't! foh!--About, my brai...
...glets, and mine doesn't go in ringlets at al...	...o to, I'll no more on't; it hath made...
...!; and I'm sure I can'tIf. What think you on't? ...

1343 1591 715 1221 1311 1707 1791 1178 **assimilarity**

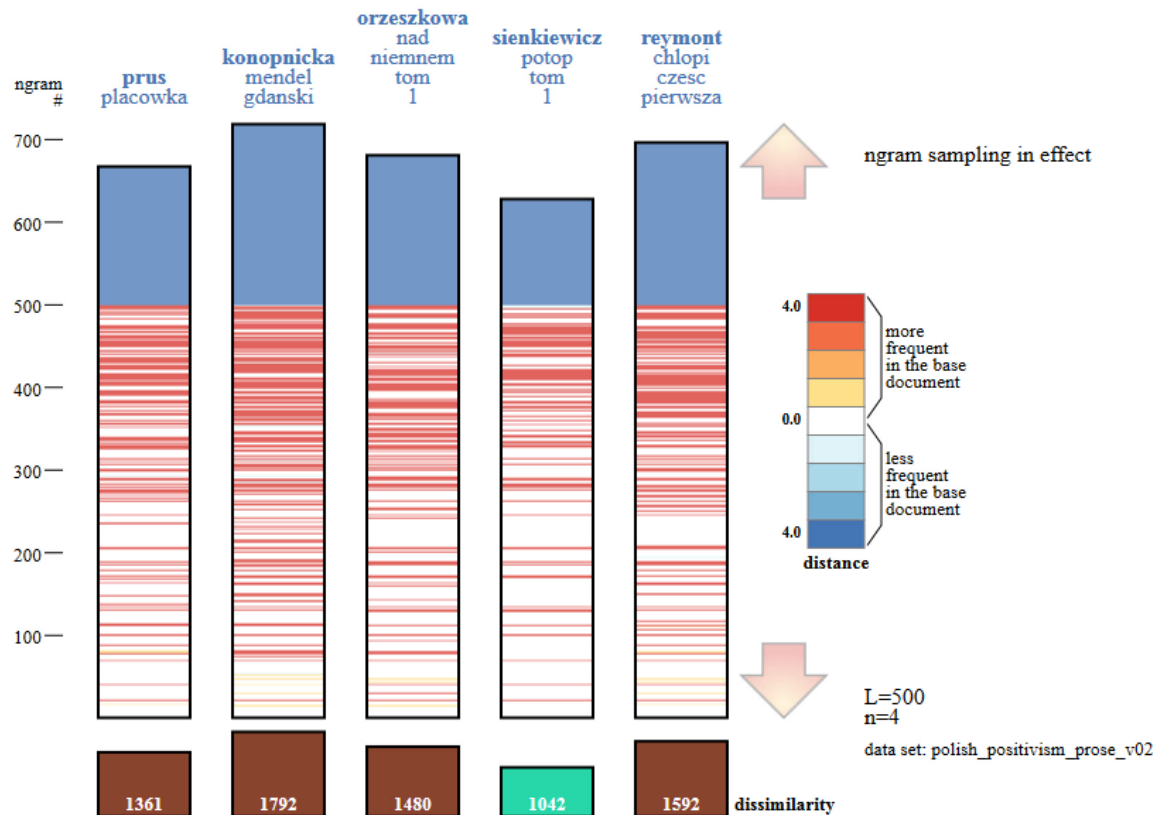
hide context

Language independence

switch to the comparison mode

base document: **sienkiewicz krzyczacy tom 1**

Polish authors



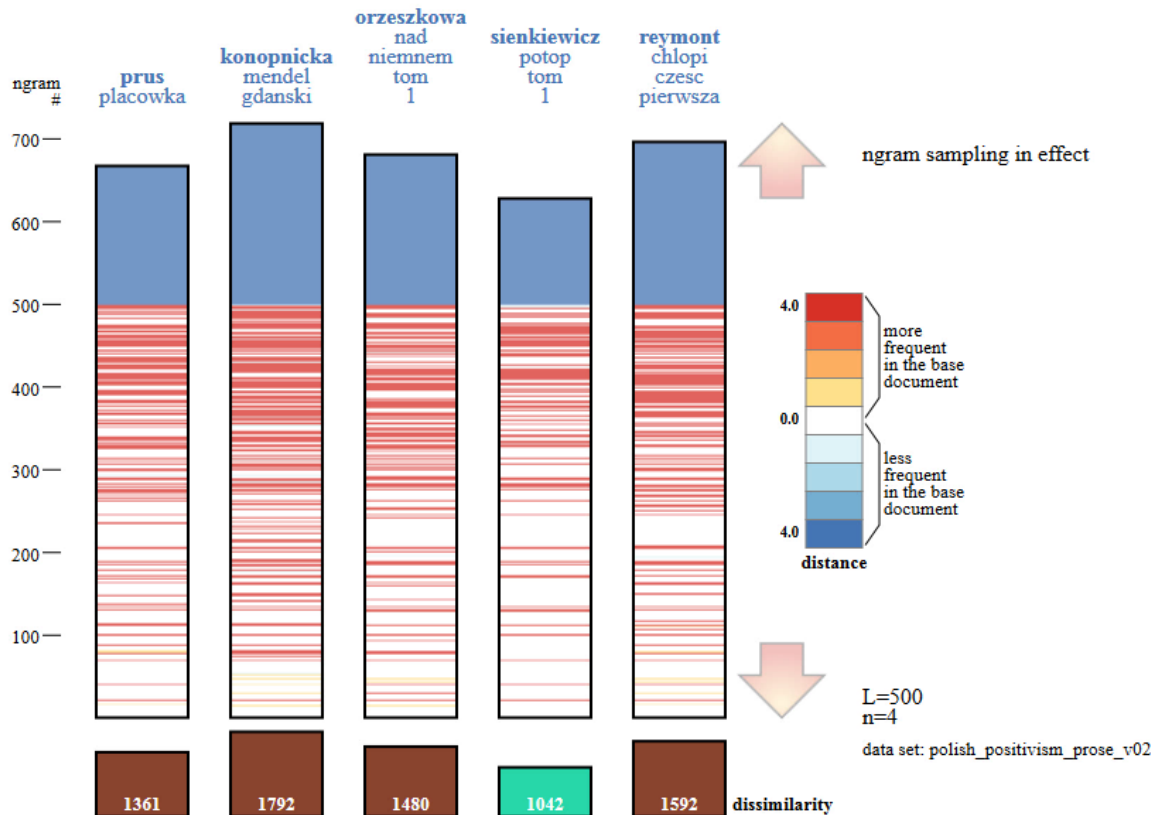
Language independence

Polish authors

searching for n-grams

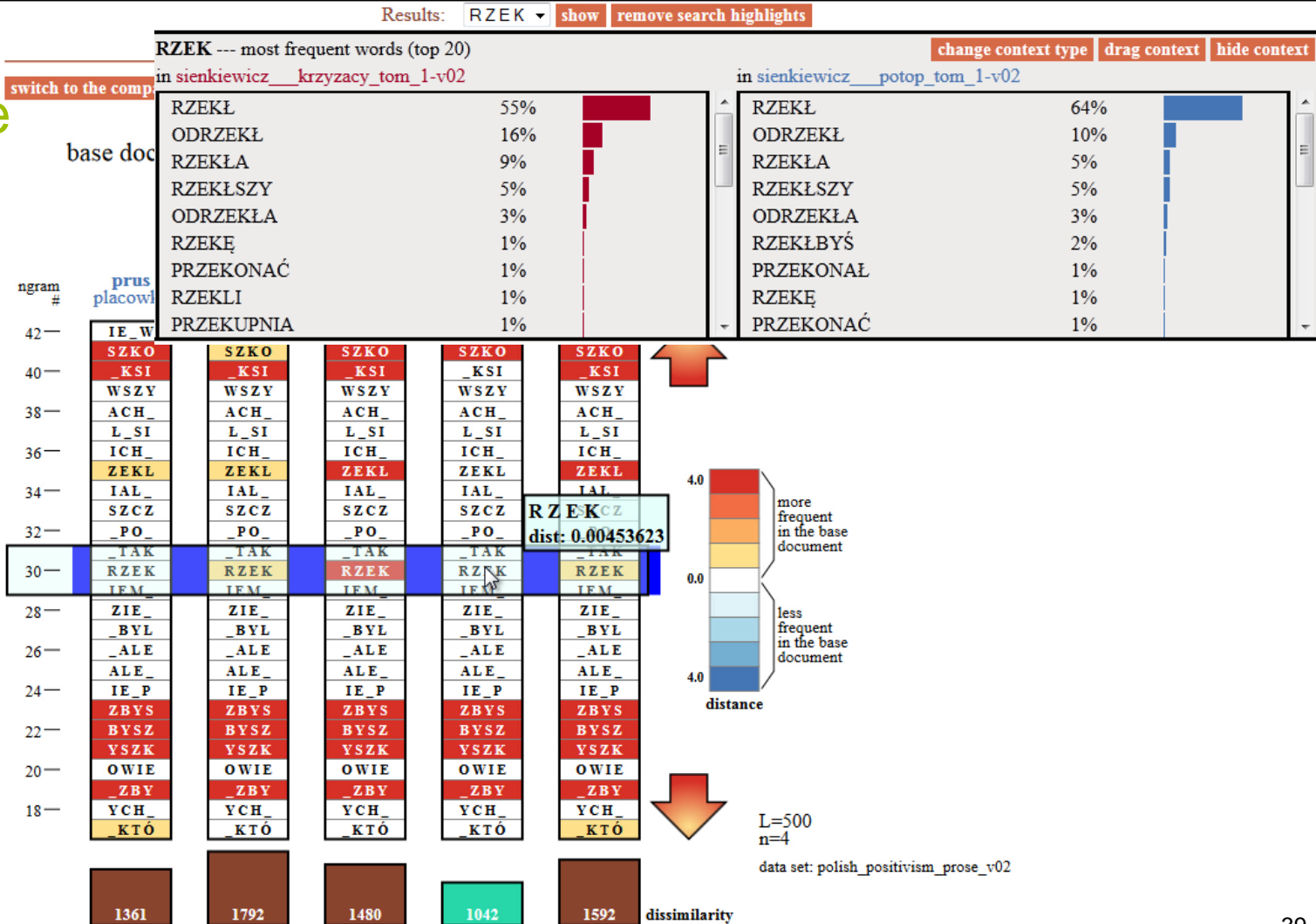
switch to the comparison mode

base document: **sienkiewicz krzyzacy tom 1**



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:

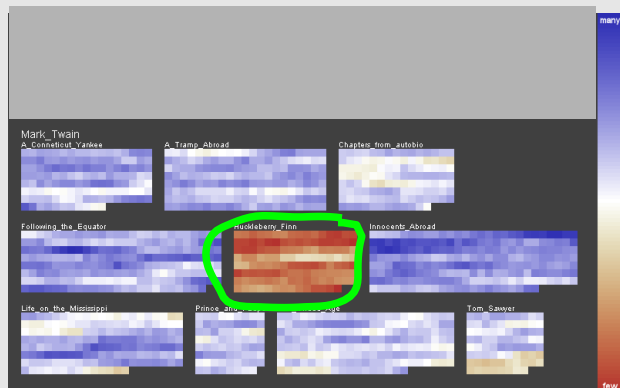
Adventures of Huckleberry Finn stands out from the other works of Mark Twain with respect to:

Function words frequency

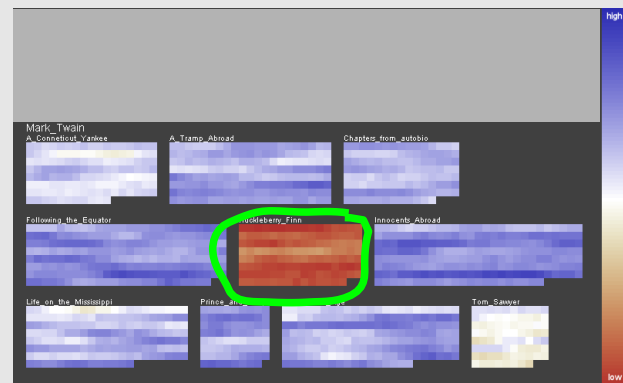
Simpson's index

Hapax Legomena

Hapax Legomena

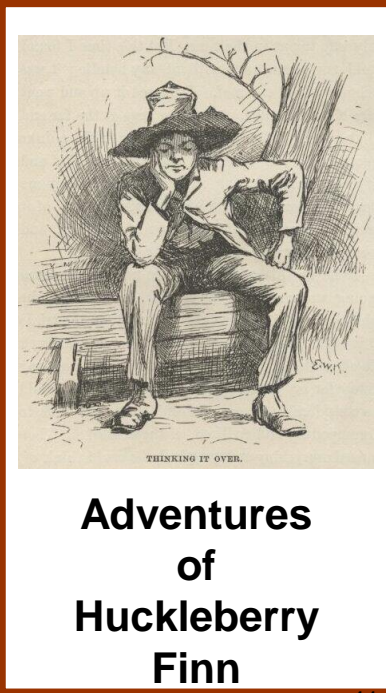
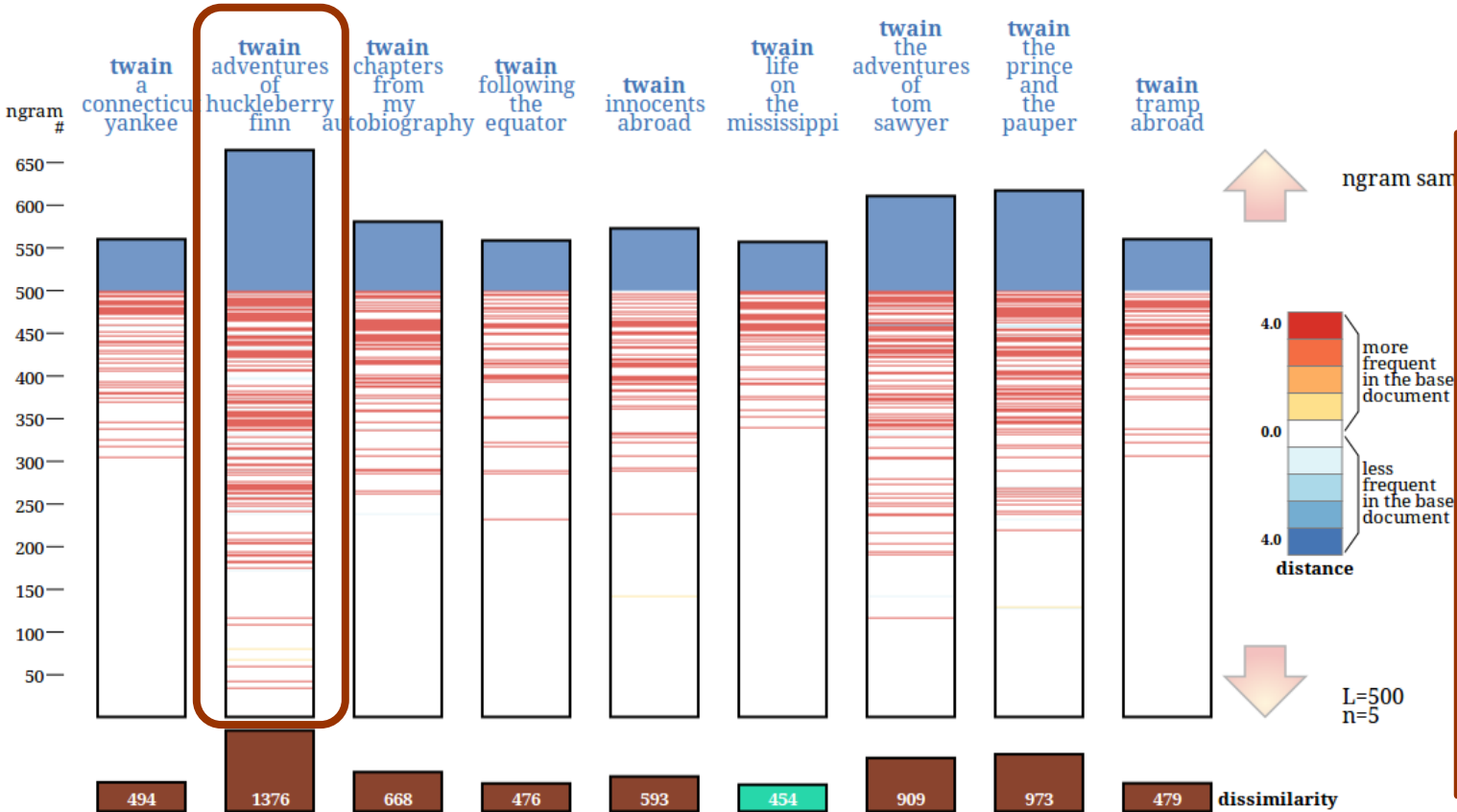


Function words (first dimension after PCA)



Example analysis: comparison of novels by Mark Twain

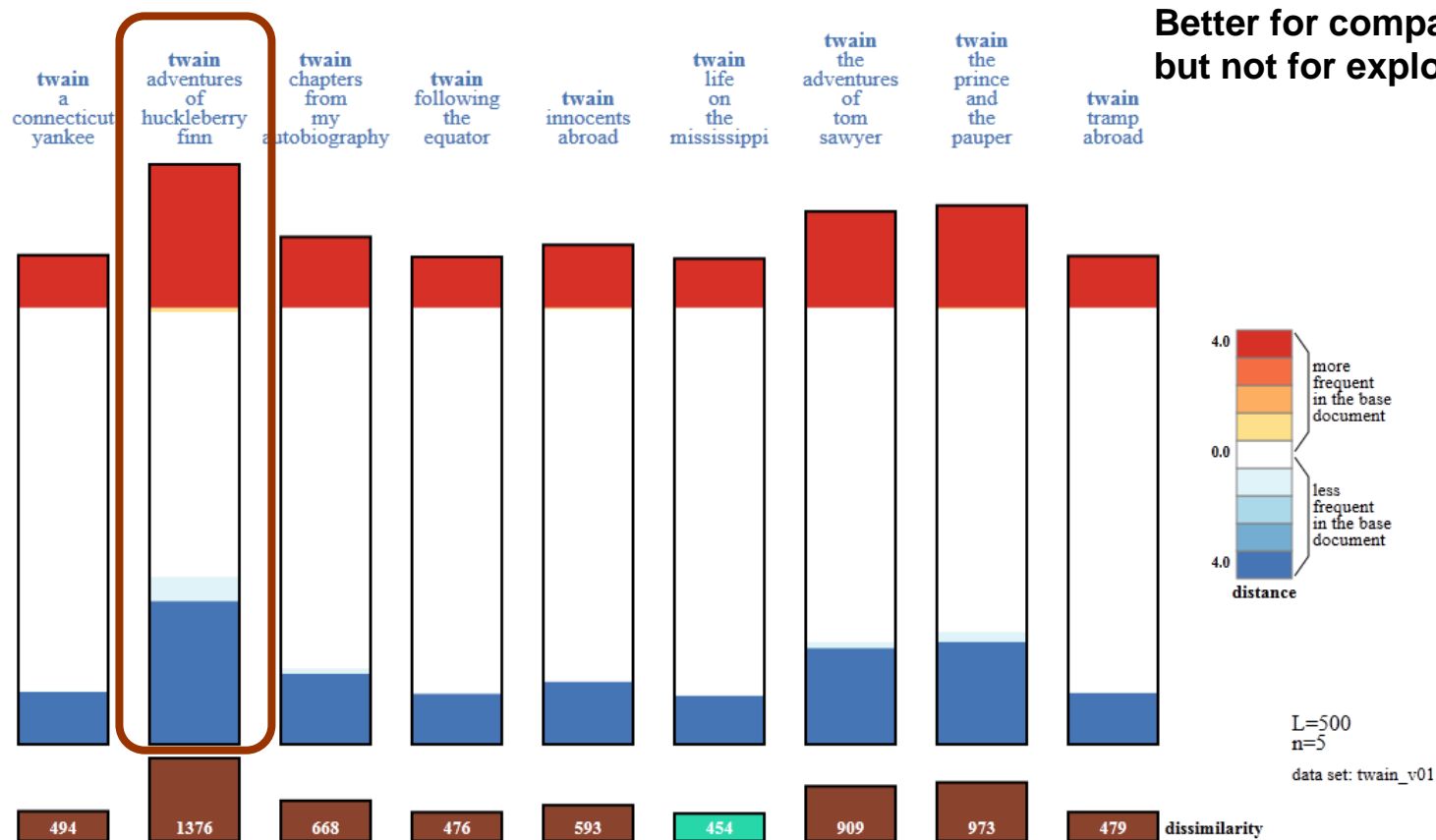
base document: **twain** all in one concatenation



Complementary Comparison View

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

base document: **twain** all in one concatenation

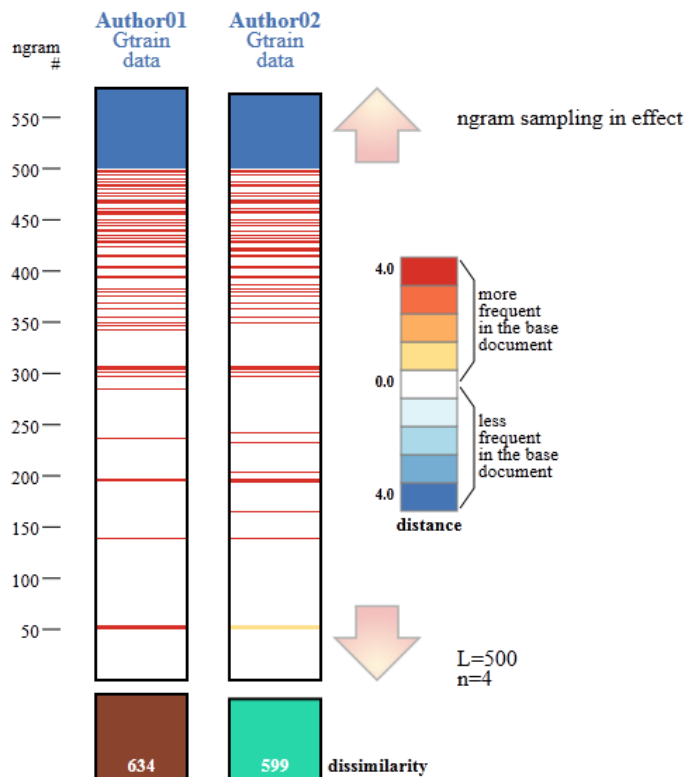


Better for comparison of signatures
but not for exploring

Interactively influencing the visualization and the classifier

Ad-hoc Authorship Attribution Competition, 2004

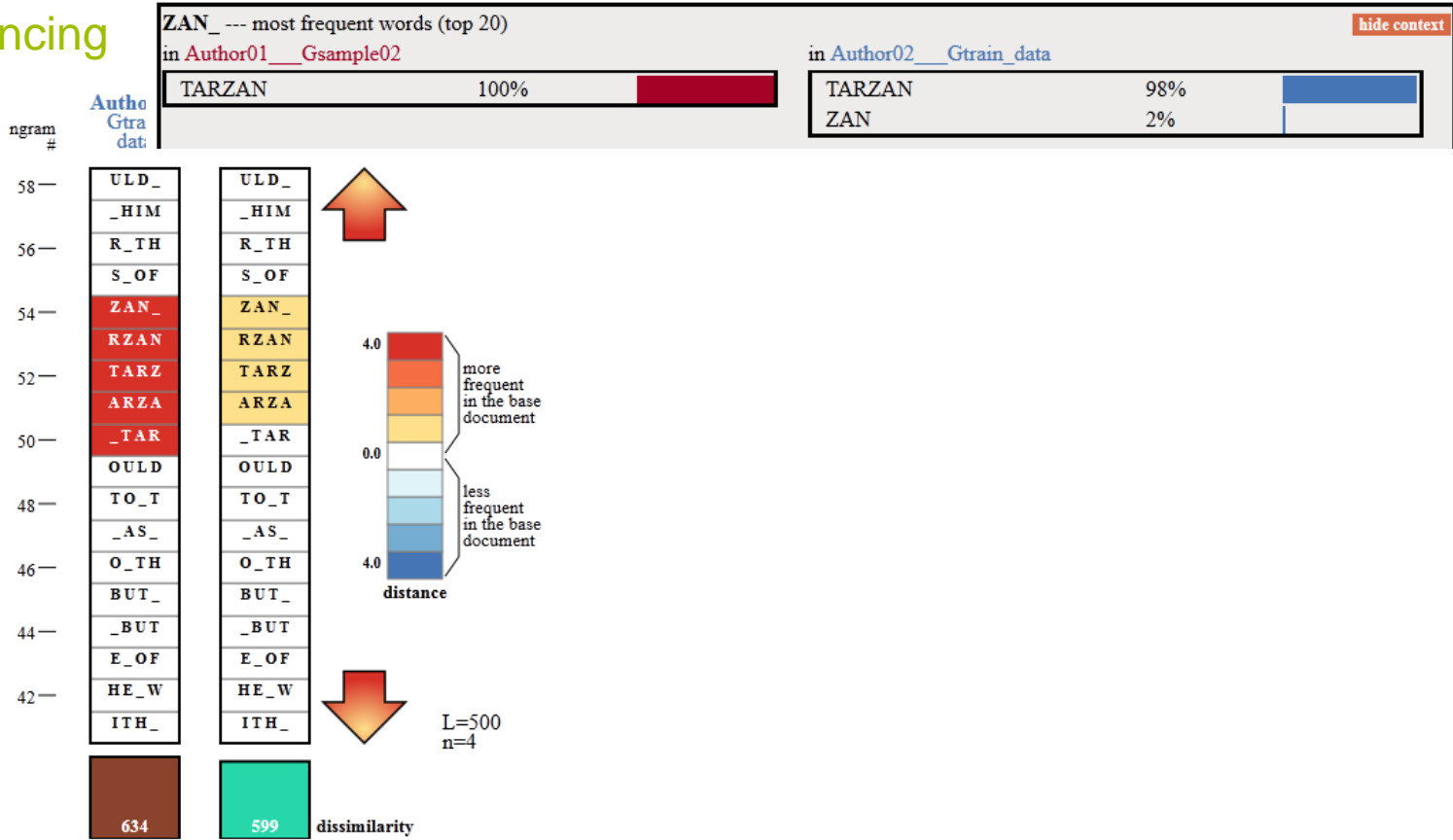
Problem G, sample 02



Manual, task-dependent adaptation of the classification process

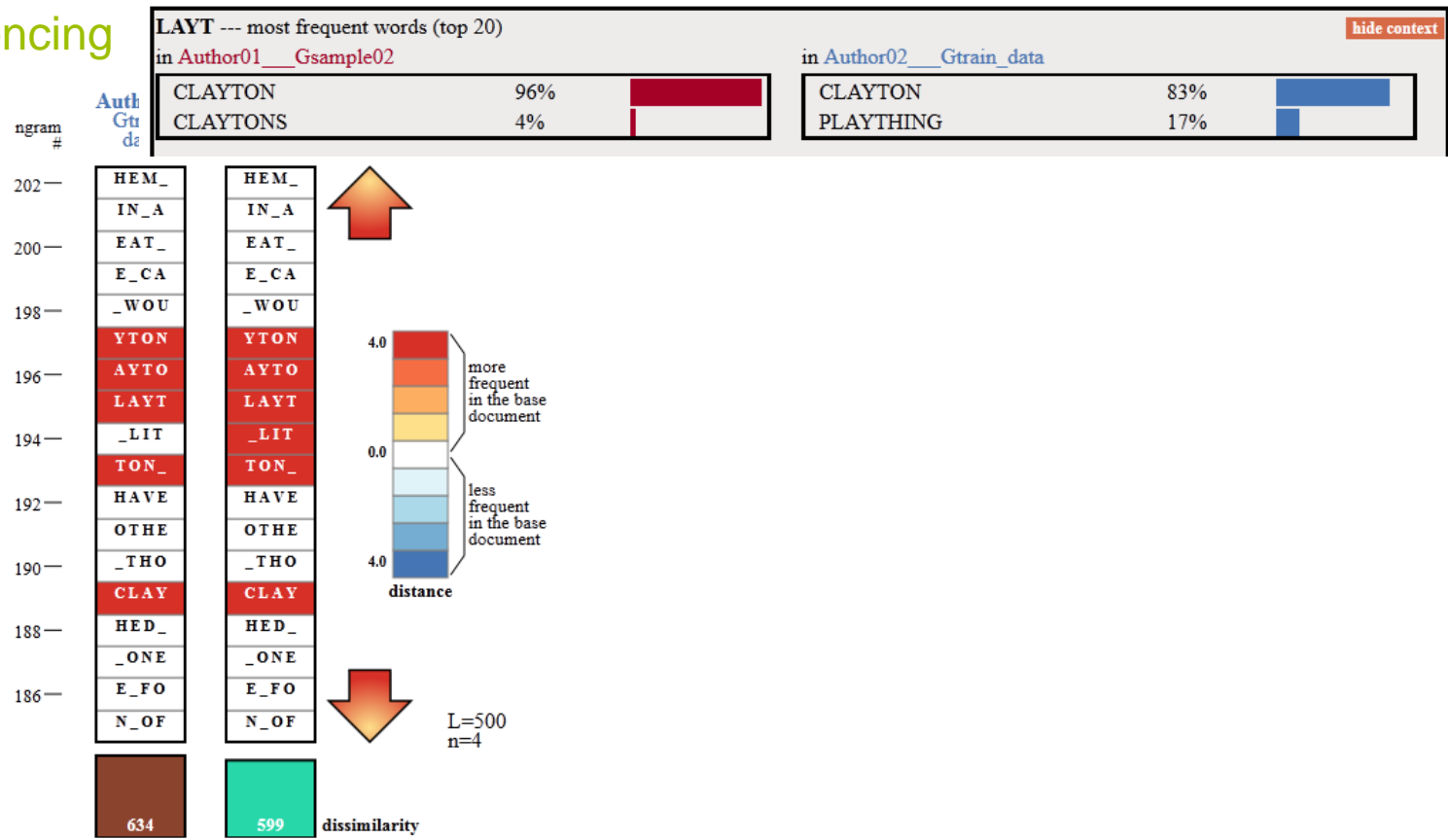
Interactively influencing the visualization and the classifier

Ad-hoc Authorship Attribution Competition, 2004
 Problem G, sample 02



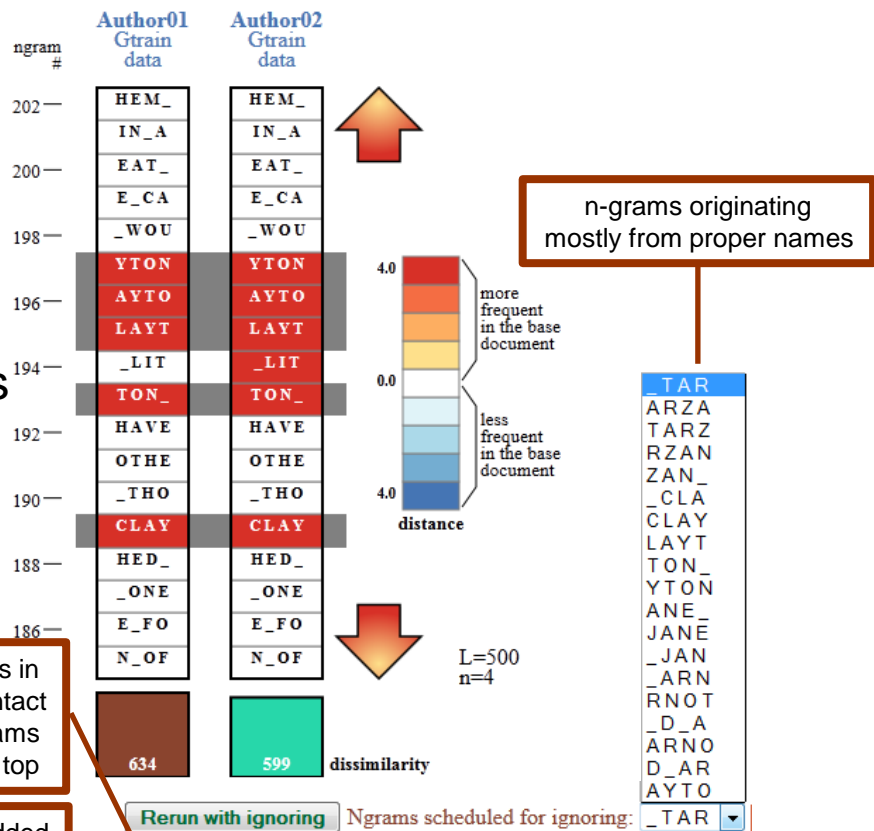
Interactively influencing the visualization and the classifier

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



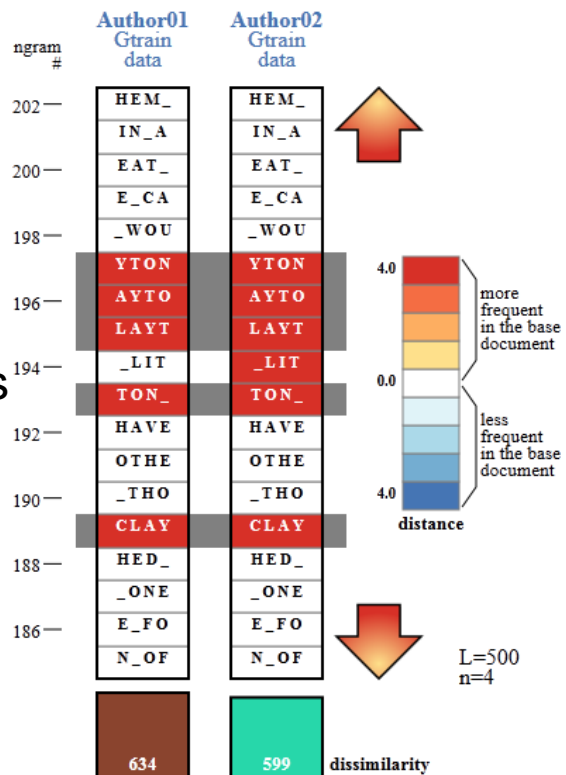
Two options:

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

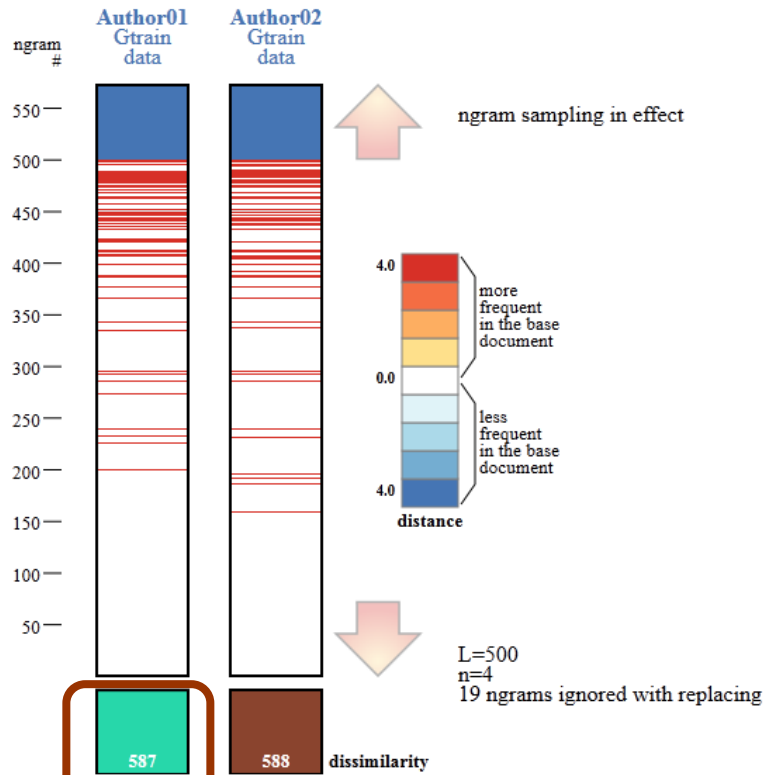
no new n-grams are added the list of n-grams for the base document becomes shorter

Interactively influencing the visualization and the classifier

ignoring selected n-grams in the base document



Ngrams scheduled for ignoring:



correct classification result

Rerun with ignoring Ngrams scheduled for ignoring:

with replacement
 without replacement

Thank you!