N-Gram-Based Author Profiles for Authroship Attribution

Florian Friedrich

Motivation

How to identify the author of an anonymous text?

works of Carroll

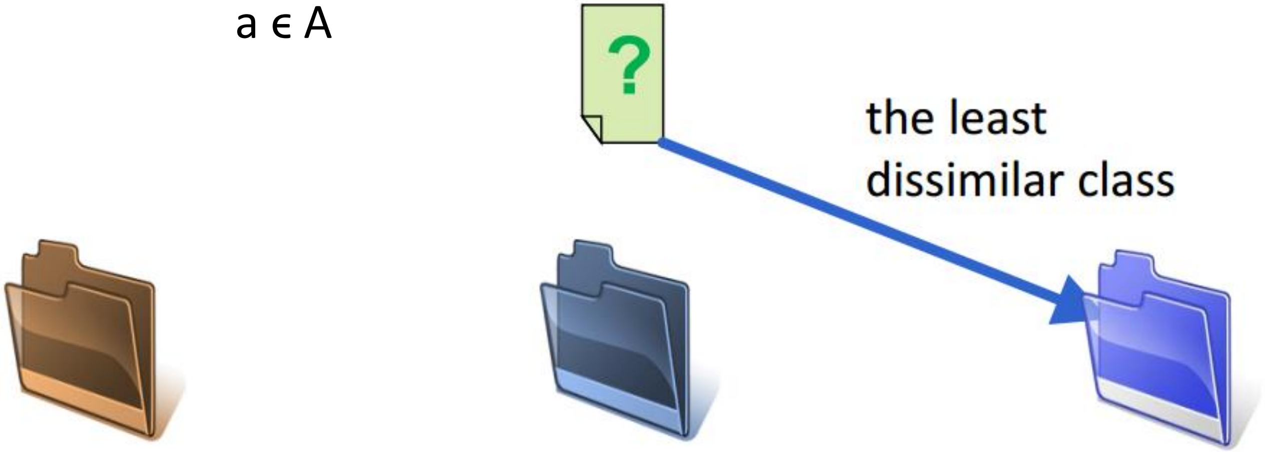
works of Twain

works of Shakespeare

Solution

Measurement of dissimilarity using character-level *n*-gram author profiles.

$$\text{author}(x) = \arg \min_{a \in A} \; d(\, PR(x), PR(x_a) \,)$$



?

the least dissimilar class

works of Carroll

works of Twain

works of Shakespeare

N-Gram

Definition n-gram: contiguous sequence of n items from a given sequence of text

Example: 2-Gram (Bigram)

- Text = {„Welcome to come“}
- n = 2 // length of n-gram

Bigram:	We	el	lc	co	om	me	_t	to	o_	_c
Amount:	1	1	1	2	2	2	1	1	1	1

Character N-Gram Statistics

- authorship attribution
- plagiarism detection
- speech recognition

German text data of
8 million characters:

Trigram	Frequency
ICH	1,15 %
EIN	1,08 %
UND	1,05 %
DER	0,97 %
NDE	0,83 %
SCH	0,65 %
DIE	0,64 %
DEN	0,62 %
END	0,60 %
CHT	0,60 %

Advantages of n-grams

- Language independant
- No word segmentation required (Asian languages)
- No text preprocessing (e.g. no style markers)

Profile Dissimilarity Algorithm

Profile: sequence of **L** most common n-grams of a given length **n**

Profile Dissimilarity Algorithm

Profile: sequence of **L** most common n-grams of a given length **n**

Example for n = 4, L = 6

document 1:

Alice's Adventures in the Wonderland

by Lewis Carroll

profile P_1	
n-gram	normalized frequency f_1
_the	0.0127
the_	0.0098
and_	0.0052
_and	0.0049
ing_	0.0047
_to_	0.0044

document 2:

Tarzan of the Apes

by Edgar Rice Burroughs

profile P_2	
n-gram	normalized frequency f_2
_the	0.0148
the_	0.0115
and_	0.0053
_of_	0.0052
_and	0.0052
ing_	0.0040

Profile Dissimilarity Algorithm

Profile: sequence of **L** most common n-grams of a given length **n**

Example for n = 4, L = 6

document 1:

Alice's Adventures in the Wonderland

by Lewis Carroll

profile P_1	
n-gram	normalized frequency f_1
_the	0.0127
the_	0.0098
and_	0.0052
_and	0.0049
ing_	0.0047
_to_	0.0044

dissimilarity between these documents

$$D = \sum_{x \in P_1 \cup P_2} \left(\frac{f_1(x) - f_2(x)}{\left(\frac{f_1(x) + f_2(x)}{2} \right)} \right)^2$$

where
$f_i(x) = 0$
if x does not appear in P_i

document 2:

Tarzan of the Apes

by Edgar Rice Burroughs

profile P_2	
n-gram	normalized frequency f_2
_the	0.0148
the_	0.0115
and_	0.0053
_of_	0.0052
_and	0.0052
ing_	0.0040

Experiment (Phyton)

- used dataset: PAN 12
- 3 authors with 2 texts each (ca 5.000 words, 25.000 characters)

- Profile: L = 100

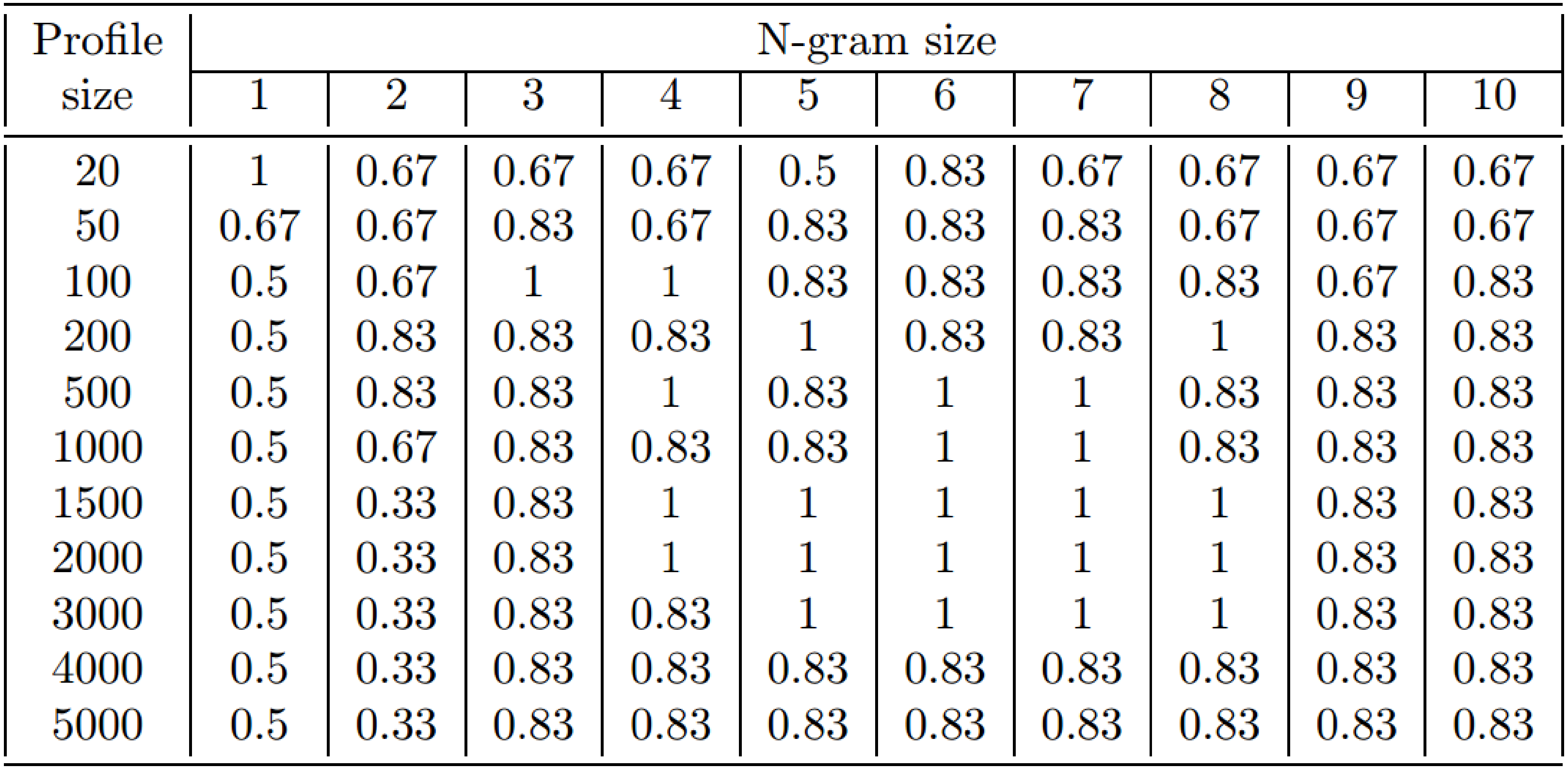
N-Gram	2	3	4	5	6	7	8	9	10
Correct attribution	2/3	2/3	2/3	2/3	1/3	1/3	1/3	1/3	2/3

- Profile: L = 10

N-Gram	2	3	4	5	6	7	8	9	10
Correct attribution	1/3	1/3	2/3	2/3	1/3	0/3	0/3	0/3	0/3

Results (Kešelj, et al., 2003)

Accuracy in English:



Profile size	N-gram size									
	1	2	3	4	5	6	7	8	9	10
20	1	0.67	0.67	0.67	0.5	0.83	0.67	0.67	0.67	0.67
50	0.67	0.67	0.83	0.67	0.83	0.83	0.83	0.67	0.67	0.67
100	0.5	0.67	1	1	0.83	0.83	0.83	0.83	0.67	0.83
200	0.5	0.83	0.83	0.83	1	0.83	0.83	1	0.83	0.83
500	0.5	0.83	0.83	1	0.83	1	1	0.83	0.83	0.83
1000	0.5	0.67	0.83	0.83	0.83	1	1	0.83	0.83	0.83
1500	0.5	0.33	0.83	1	1	1	1	1	0.83	0.83
2000	0.5	0.33	0.83	1	1	1	1	1	0.83	0.83
3000	0.5	0.33	0.83	0.83	1	1	1	1	0.83	0.83
4000	0.5	0.33	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83
5000	0.5	0.33	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83

Reproducibility

- difficult to get data
- large size of dissimilarity

References

- Vlado Kešelj, Fuchun Peng, Nick Cercone, and Calvin Thomas. *N-gram-based author profiles for authorship attribution*. Pacific Association for Computational Linguistics, 2003.

- https://de.wikipedia.org/wiki/N-Gramm
- https://www.uni-weimar.de/medien/webis/events/pan-13/pan13-talks/pan13-authorship-verification/jankowska13-slides.pdf