Stylowe kalkulacje

(zajawka o atrybucji autorstwa)

Atrybucja Autorstwa

- **ustalenie autora** dzieła anonimowego lub o niepewnym autorstwie na podstawie właściwości samego tekstu
- w oparciu o wyniki **analizy** treści, **cech językowo-stylistycznych tekstu** oraz wskaźników historyczno-bibliograficznych, edytorskich i drukarskich
- zakłada istnienie **stylistycznego odcisku palca** - nieuświadomionych nawyków stylistycznych, do badania których wykorzystuje **stylometrię**

STYLOMETRIA

- bazuje na **badaniach ilościowych** (wspomaganych komputerowo, chociaż przez długi czas szło to ręcznie)
- pomija cechy stylistyczne w ujęciu tradycyjnym - metafory, epitety, stylizację, ironię...
- **analiza porównawcza** - konieczny i bardzo istotny kontekst

- rozkład części mowy
- rozkład form czasownikowych
- połączenia międzywyrazowe
- zbitki literowe
- bogactwo słownictwa
- typowe konstrukcje zdaniowe
- relacje składniowe
- frekwencja wyrazów

DELTA BURROWSA (NIEPEŁNA)

T1: “Cupcake ipsum dolor sit amet.”

T2: “Cupcake ipsum dolor sit amet.”

T3: “Muffin ipsum dolor sit amet.”

T4: “Muffin cupcake cupcake sit amet.”

t_1 = { cupcake: 1, ipsum: 1, dolor: 1, sit: 1, amet: 1 }

t_2 = { cupcake: 1, ipsum: 1, dolor: 1, sit: 1, amet: 1 }

t_3 = { muffin: 1, ipsum: 1, dolor: 1, sit: 1, amet: 1 }

t_4 = { muffin: 1, cupcake: 2, sit: 1, amet: 1 }

```ruby
def diff_between(text_1, text_2)
  diff = text_1.merge(text_2) do |_token, freq_1, freq_2|
    (freq_1-freq_2).abs
  end
  diff.values.sum
end
```

t_1: t_2 {cupcake: 0, ipsum: 0, dolor: 0, sit: 0, amet: 0}

t_1:t_3 {cupcake: 1, ipsum: 0, dolor: 0, sit: 0, amet: 0, muffin: 1}

t_1:t_4 {cupcake: 1, ipsum: 1, dolor: 1, amet: 0, muffin: 1}

diff_by_token_1_2 = 0	“Cupcake ipsum dolor sit amet.”
diff_by_token_1_3 = 2	“Muffin ipsum dolor sit amet.”
diff_by_token_1_4 = 4	“Muffin cupcake cupcake sit amet.”

POKA PRZYKŁAD!

KORPUS

102 powieści angielskie - trzy powieści Jane Austen z Project Gutenberg + 99 powieści (też z Gutenberg Project, ale pobranych z repo: https://github.com/computationalstylistics/100_english_novels)

łącznie: **14 597 120** wyrazów

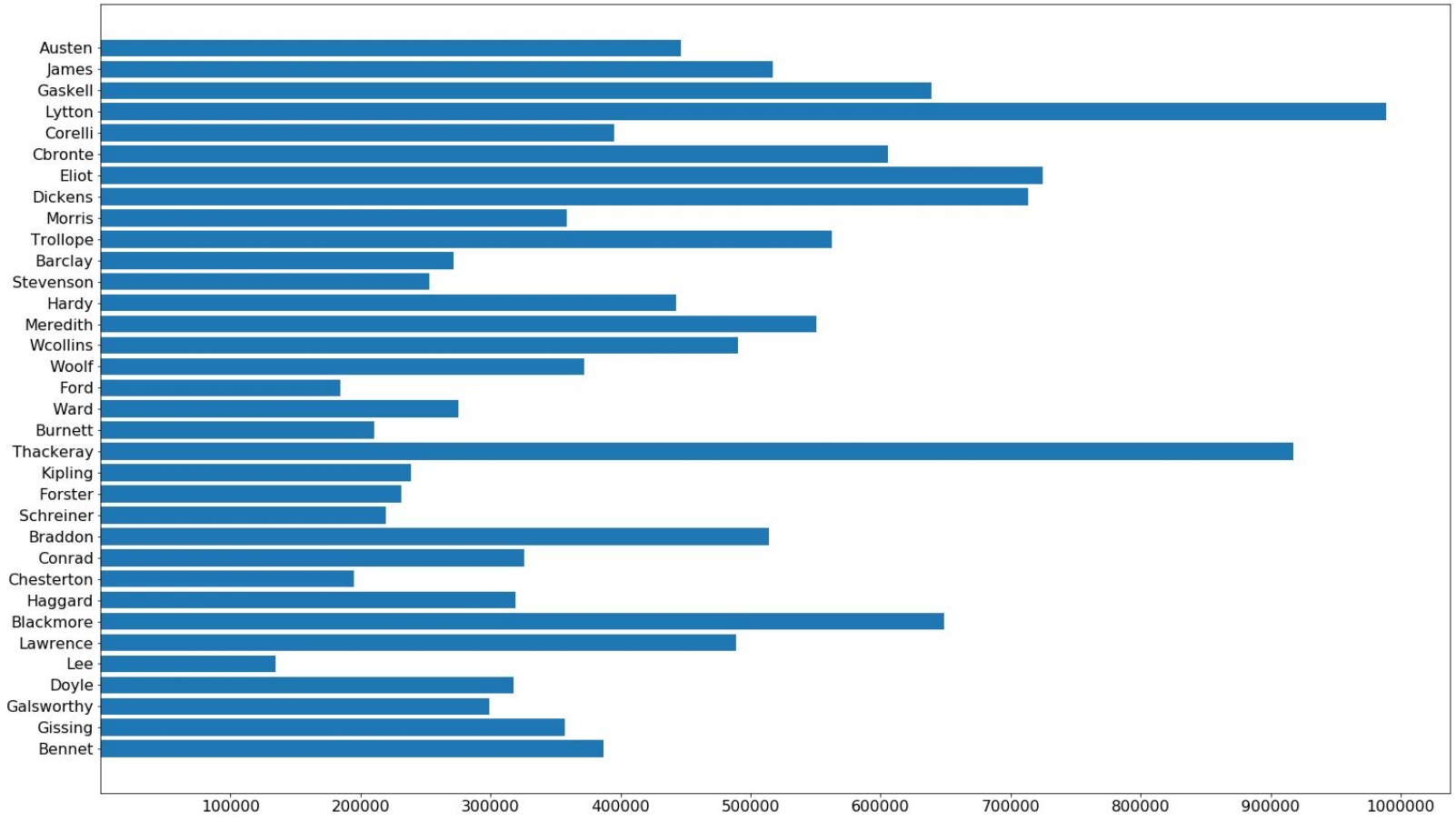
the, and, to, of, a, i, in, he, was, that, it, her, you, his, she, had, with, as, for, not, s, at, but, be, on, is, him, my, have, me, said, all, so, which, by, this, from, they, no, if, would, were, there, what, one, t, when, been, an, or, we, who, could, do, out, up, are, very, your, will, them, mr, now, more, man, then, little, like, their, into, about, some, did, than, know, can, see, well, should, any, time, good, never, only, come, upon, has, how, before, down, must, old, much, own, say, over, think, go, am, such, might, other, after, made, again, don, thought, himself, our, came, great, mrs, day, two, too, here, us, way, back, lady, face, young, eyes, went, sir, life, where, shall, away, long, hand, may, looked, first, house, miss, even, its, yet, nothing, still, make, room, last, though, father, look, men, just, these

GRUPY

34 autorów, po 3 teksty każdego

długość od **134 367** do **988 525** wyrazów w grupie

w każdej grupie autorskiej zliczamy wyrazy, tworząc **frekwencyjny słownik odniesienia** dla danego autora

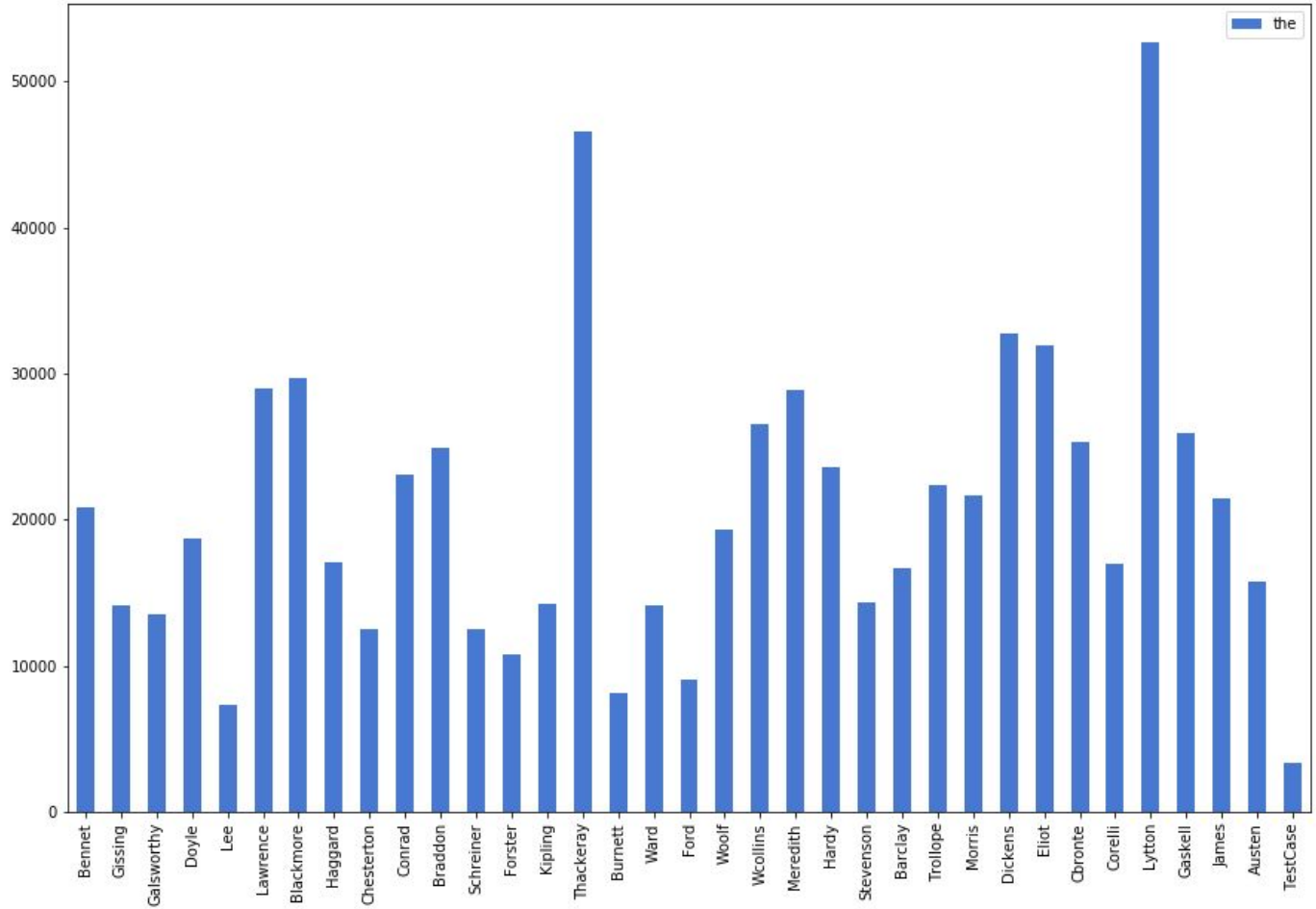

Austen
James
Gaskell
Lytton
Corelli
Cbronte
Eliot
Dickens
Morris
Trollope
Barclay
Stevenson
Hardy
Meredith
Wcollins
Woolf
Ford
Ward
Burnett
Thackeray
Kipling
Forster
Schreiner
Braddon
Conrad
Chesterton
Haggard
Blackmore
Lawrence
Lee
Doyle
Galsworthy
Gissing
Bennet
100000
200000
300000
400000
500000
600000
700000
800000
900000
1000000

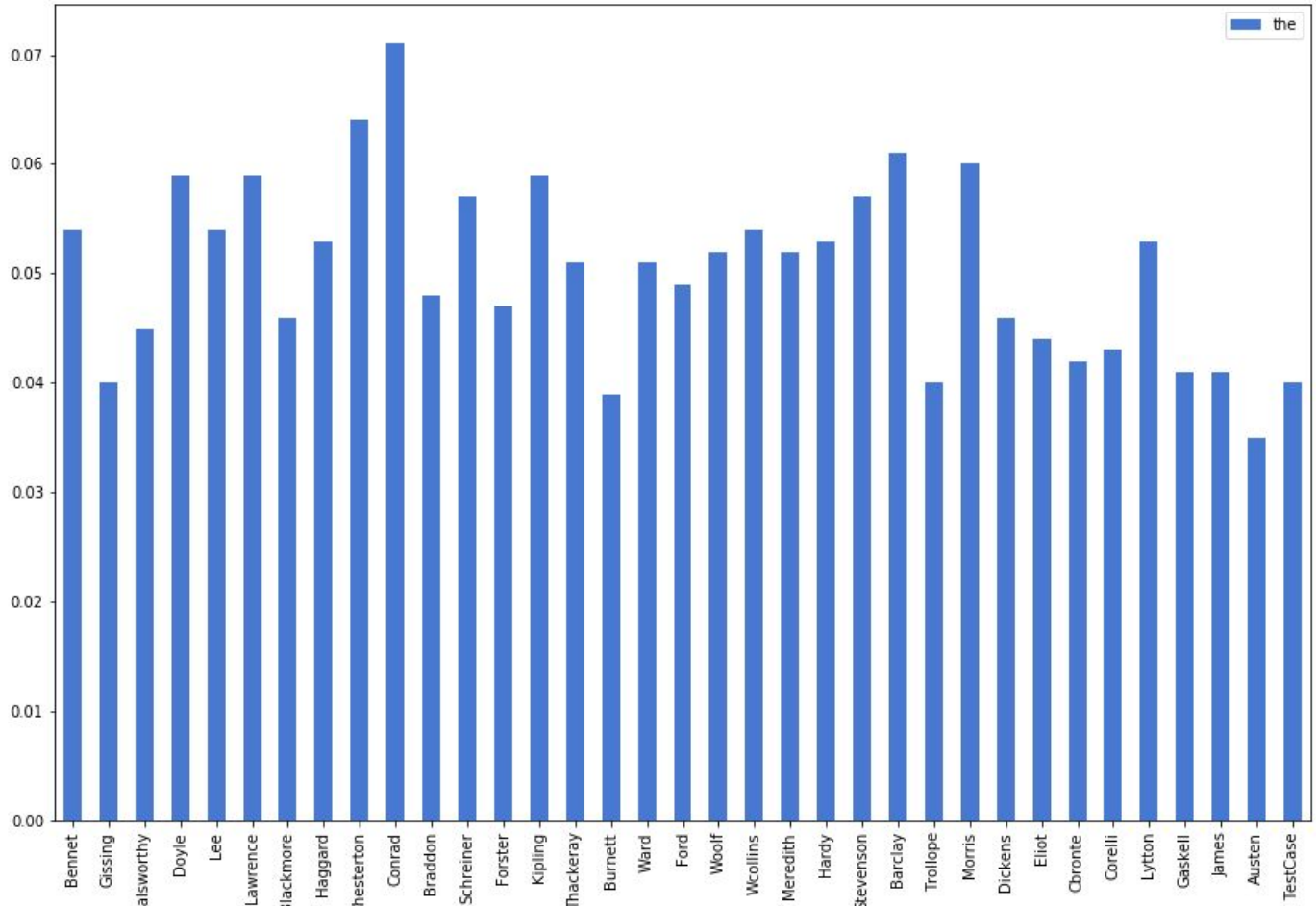
TEKST ZE STRYCHU

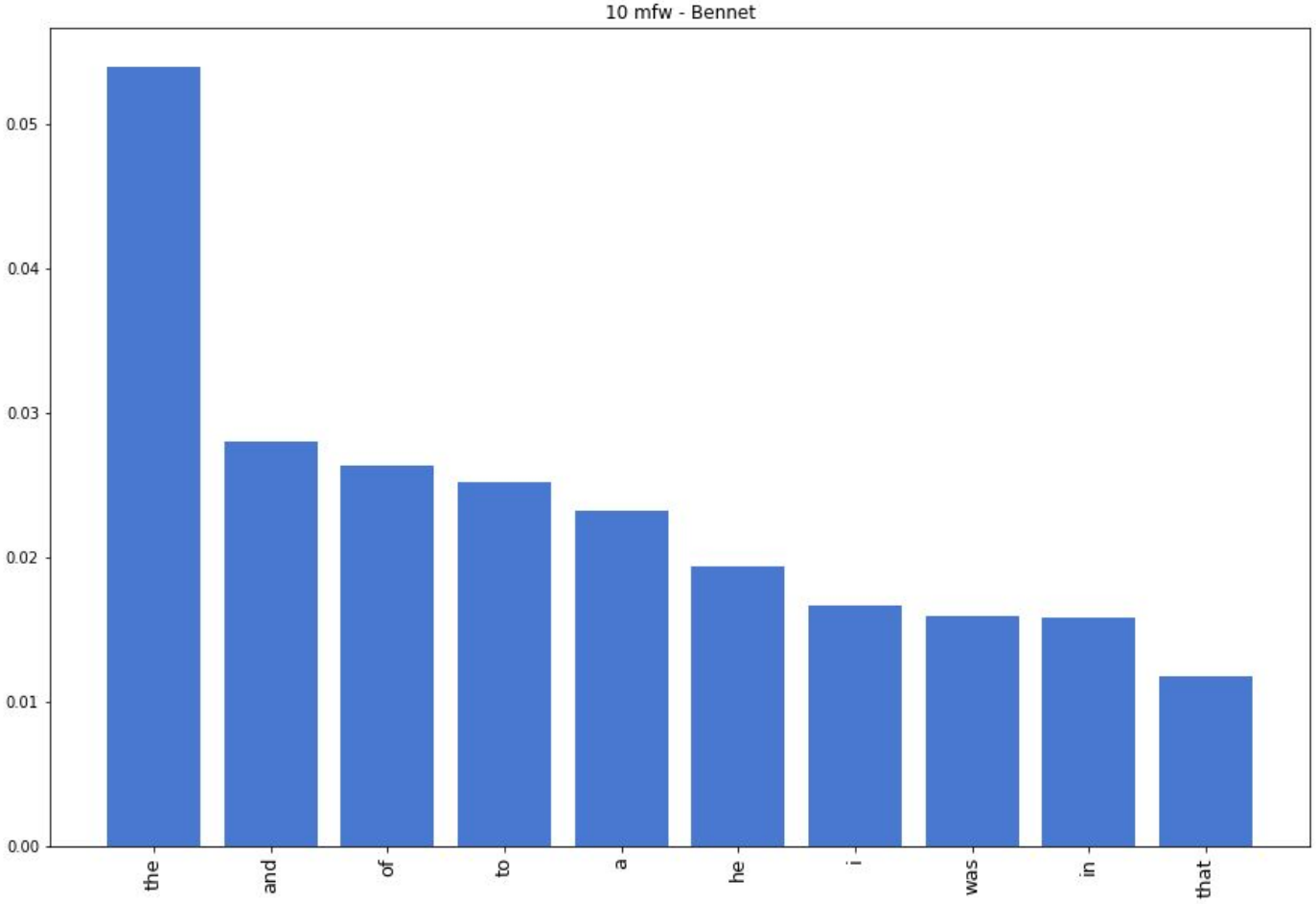
będzie go udawać powieść **Jane Austen - “Persuasion”**

84092 wyrazy, które również zliczamy

PORÓWNAJMY SOBIE


the
50000
40000
30000
20000
10000
0
Bennet
Gissing
Galsworthy
Doyle
Lee
Lawrence
Blackmore
Haggard
Chesterton
Conrad
Braddon
Schreiner
Forster
Kipling
Thackeray
Burnett
Ward
Ford
Woolf
Wcollins
Meredith
Hardy
Stevenson
Barclay
Trollope
Morris
Dickens
Eliot
Cbronte
Corelli
Lytton
Gaskell
James
Austen
TestCase


the
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
Bennet
Gissing
Galsworthy
Doyle
Lee
Lawrence
Blackmore
Haggard
Chesterton
Conrad
Braddon
Schreiner
Forster
Kipling
Thackeray
Burnett
Ward
Ford
Woolf
Wcollins
Meredith
Hardy
Stevenson
Barclay
Trollope
Morris
Dickens
Eliot
Cbronte
Corelli
Lytton
Gaskell
James
Austen
TestCase


10 mfw - Bennet
0.05
0.04
0.03
0.02
0.01
0.00
the
and
of
to
a
he
i
was
in
that

Z-STANDARYZACJA

$$\textbf{Z\_SCORE} = \frac{\text{rzeczywista wartość zmiennej - średnia}}{\text{odchylenie standardowe}}$$

Dla każdego wyrazu W w słownikach frekwencyjnych:

$$\textbf{Z\_SCORE} = \frac{\text{częstość względna W - średnia częstość względna W w korpusie}}{\text{odchylenie standardowe częstości względnej W w korpusie}}$$

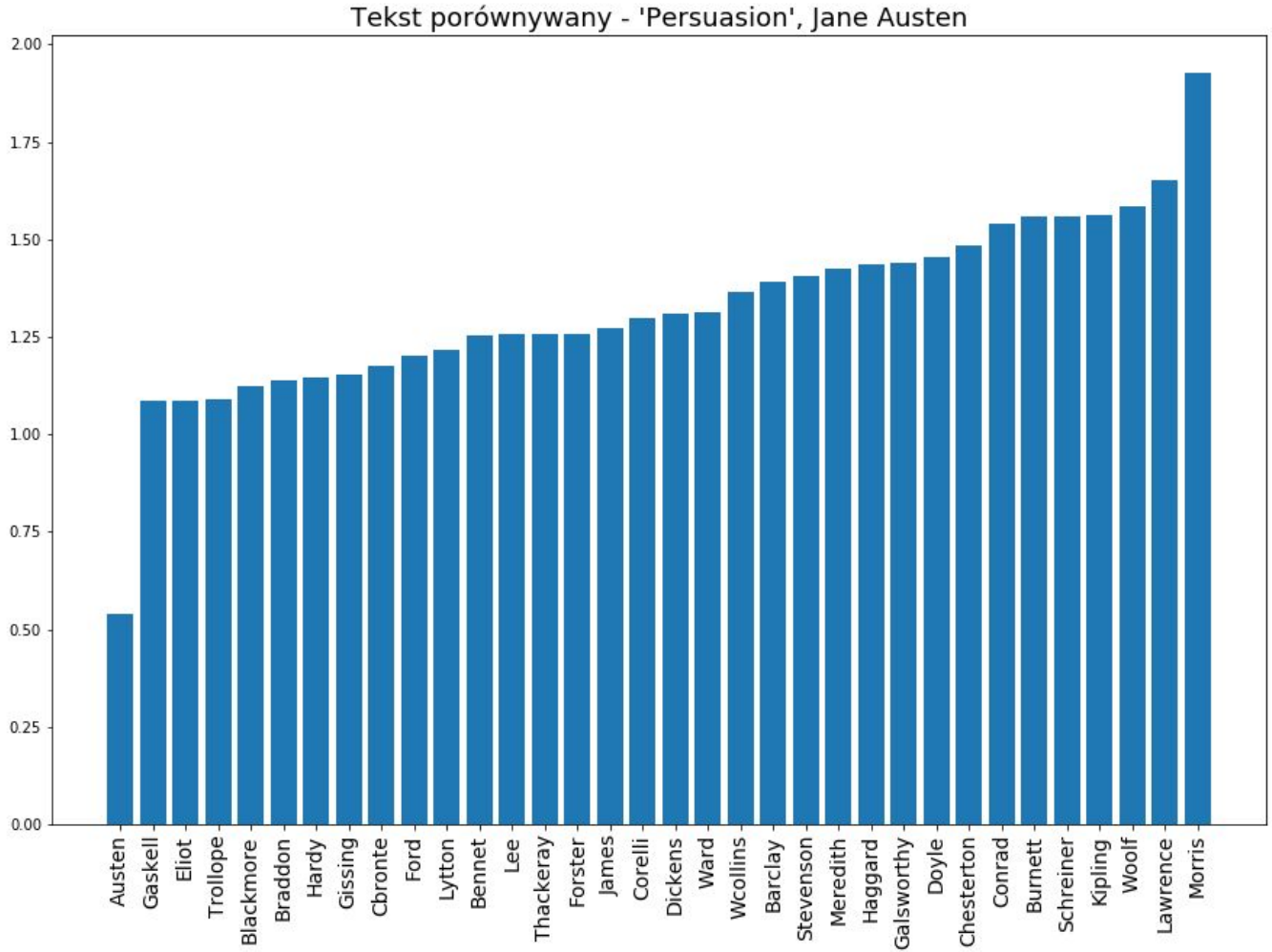
```
most_frequent_terms.each do |term|
  mean = 0
  groups.each do |group|
   mean += group.relative_frequency_for(term)
  end
  mean /= groups.count
end
```

```
most_frequent_terms.each do |t|
  std = 0
  groups.each do |g|
    diff = g.relative_frequency_for(t) - corpus.mean_for(t)
    std += diff**2
  end
  std = Math.sqrt(std / (groups.count - 1))
end
```

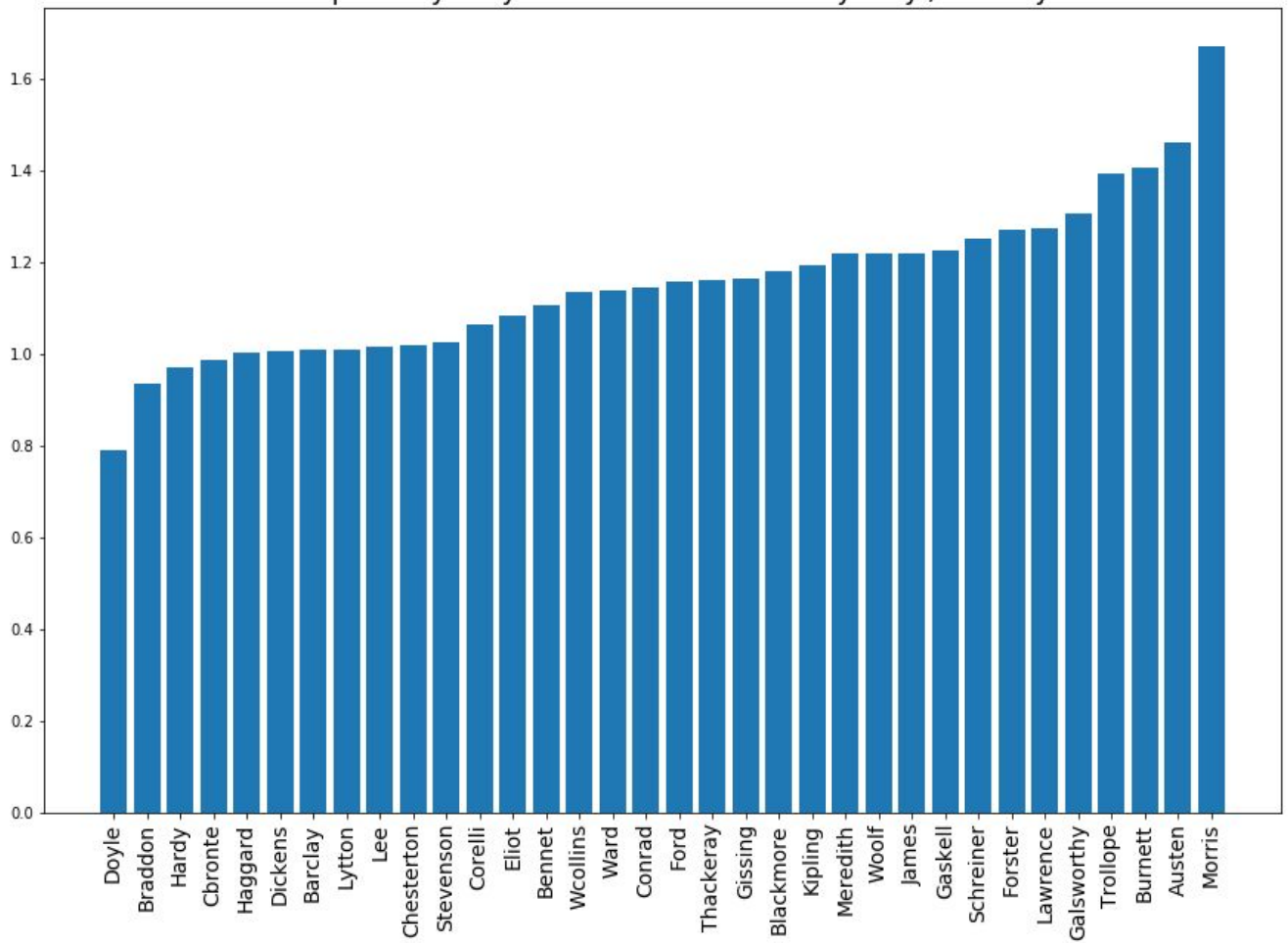
```
def diff_between(z_scores1, z_scores2)
  diff = z_scores1.merge(z_scores2) do |_token, z1, z2|
    (z1 - z2).abs
  end
  diff.values.sum
end
```

```
deltas = groups.map do |group|
   diff_between(test.z_scores, group.z_scores)
end
```

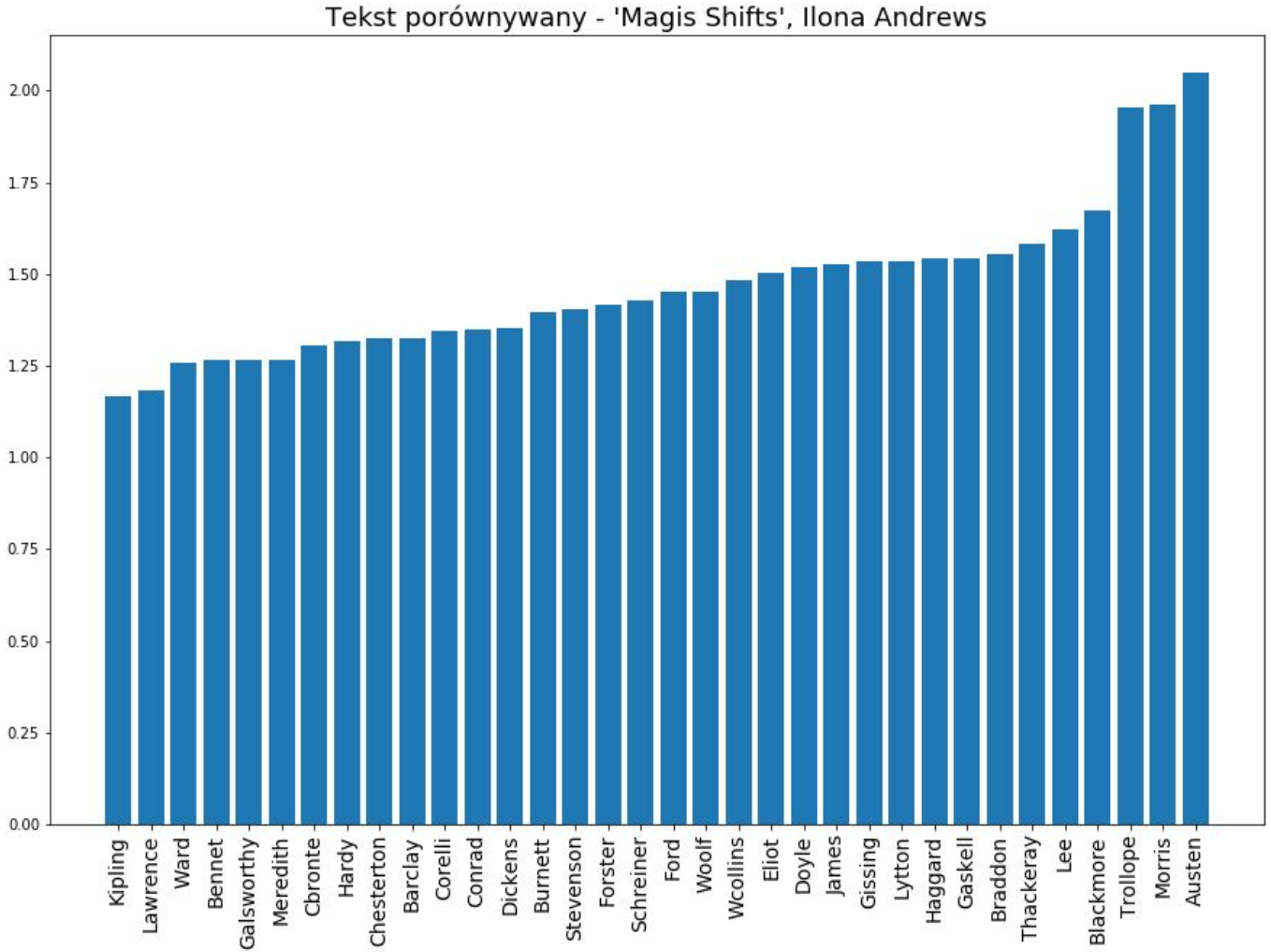
WYNIK?


Tekst porównywany - 'Persuasion', Jane Austen
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Austen
Gaskell
Eliot
Trollope
Blackmore
Braddon
Hardy
Gissing
Cbronte
Ford
Lytton
Bennet
Lee
Thackeray
Forster
James
Corelli
Dickens
Ward
Wcollins
Barclay
Stevenson
Meredith
Haggard
Galsworthy
Doyle
Chesterton
Conrad
Burnett
Schreiner
Kipling
Woolf
Lawrence
Morris

TO MOŻE INNY TEKST?


Tekst porównywany - 'Tales of Terror and Mystery', A.C.Doyle
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Doyle
Braddon
Hardy
Cbronte
Haggard
Dickens
Barclay
Lytton
Lee
Chesterton
Stevenson
Corelli
Eliot
Bennet
Wcollins
Ward
Conrad
Ford
Thackeray
Gissing
Blackmore
Kipling
Meredith
Woolf
James
Gaskell
Schreiner
Forster
Lawrence
Galsworthy
Trollope
Burnett
Austen
Morris

A JEŚLI AUTOR Z ZUPEŁNIE INNEJ BECZKI?


Tekst porównywany - 'Magis Shifts', Ilona Andrews
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Kipling
Lawrence
Ward
Bennet
Galsworthy
Meredith
Cbronte
Hardy
Chesterton
Barclay
Corelli
Conrad
Dickens
Burnett
Stevenson
Forster
Schreiner
Ford
Woolf
Wcollins
Eliot
Doyle
James
Gissing
Lytton
Haggard
Gaskell
Braddon
Thackeray
Lee
Blackmore
Trollope
Morris
Austen

TO SKĄD MAM WIEDZIEĆ?

Arthur Conan Doyle:

- "The Sign of the Four",
- "Tales of Terror and Mystery",
- "The White Company",
- "Memoirs of Sherlock Holmes"

Jane Austen:

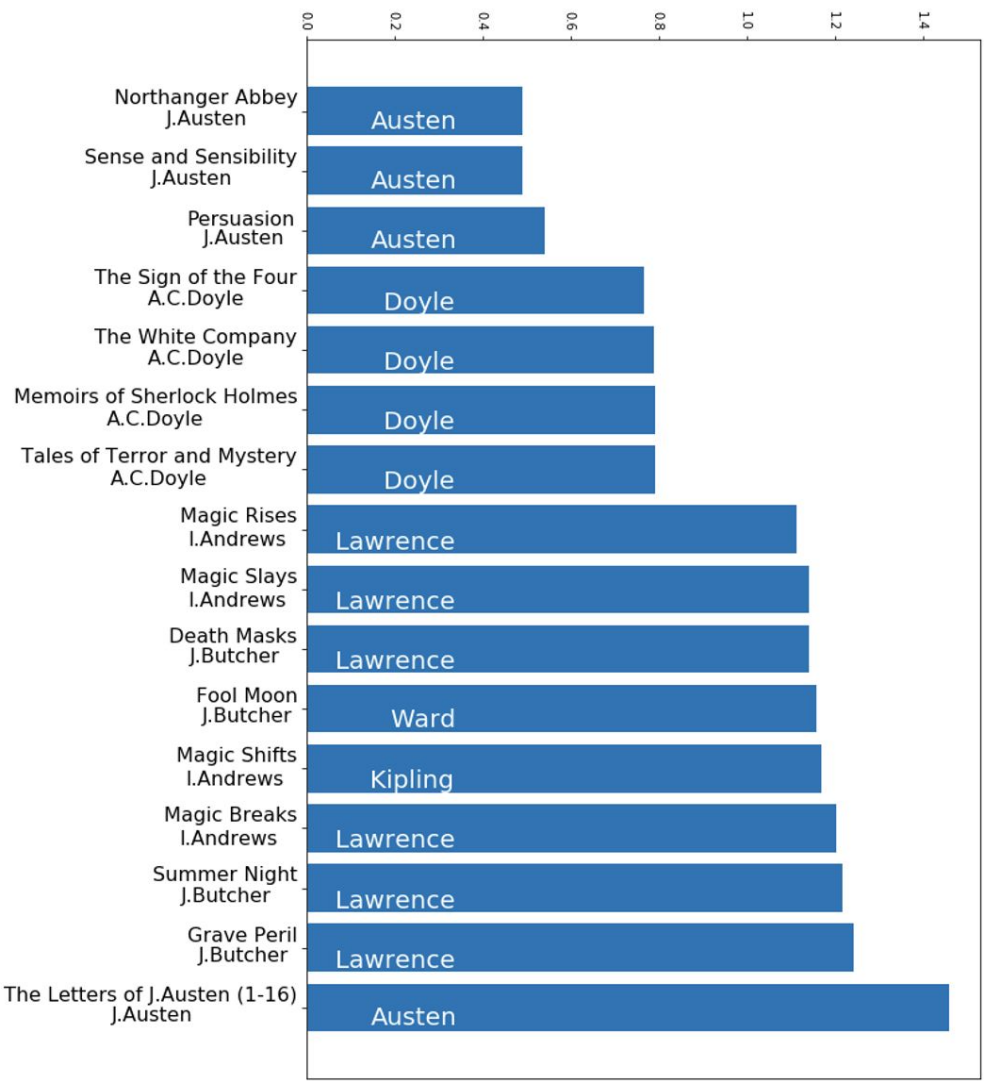
- "The Letters of Jane Austen (1-16)",
- "Northanger Abbey",
- "Persuasion",
- "Sense and Sensibility"

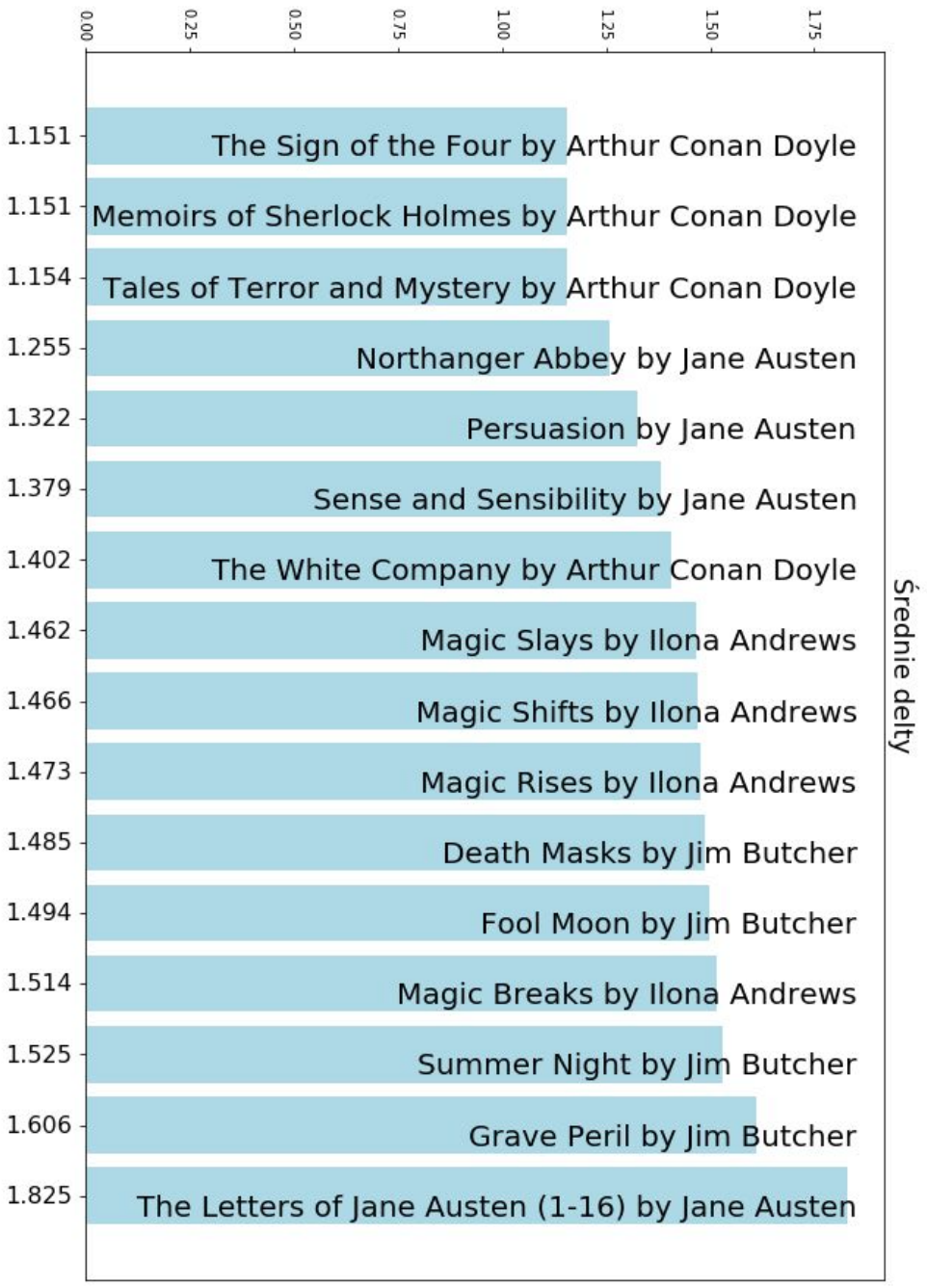
Jim Butcher:

- "Fool Moon",
- "Grave Peril",
- "Summer Night",
- "Death Masks"

Ilona Andrews:

- "Magic Slays",
- "Magic Rises",
- "Magic Breaks",
- "Magic Shifts"


0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
Northanger Abbey
J.Austen
Austen
Sense and Sensibility
J.Austen
Austen
Persuasion
J.Austen
Austen
The Sign of the Four
A.C.Doyle
Doyle
The White Company
A.C.Doyle
Doyle
Memoirs of Sherlock Holmes
A.C.Doyle
Doyle
Tales of Terror and Mystery
A.C.Doyle
Doyle
Magic Rises
I.Andrews
Lawrence
Magic Slays
I.Andrews
Lawrence
Death Masks
J.Butcher
Lawrence
Fool Moon
J.Butcher
Ward
Magic Shifts
I.Andrews
Kipling
Magic Breaks
I.Andrews
Lawrence
Summer Night
J.Butcher
Lawrence
Grave Peril
J.Butcher
Lawrence
The Letters of J.Austen (1-16)
J.Austen
Austen


0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
1.151 The Sign of the Four by Arthur Conan Doyle
1.151 Memoirs of Sherlock Holmes by Arthur Conan Doyle
1.154 Tales of Terror and Mystery by Arthur Conan Doyle
1.255 Northanger Abbey by Jane Austen
1.322 Persuasion by Jane Austen
1.379 Sense and Sensibility by Jane Austen
1.402 The White Company by Arthur Conan Doyle
1.462 Magic Slays by Ilona Andrews
1.466 Magic Shifts by Ilona Andrews
1.473 Magic Rises by Ilona Andrews
1.485 Death Masks by Jim Butcher
1.494 Fool Moon by Jim Butcher
1.514 Magic Breaks by Ilona Andrews
1.525 Summer Night by Jim Butcher
1.606 Grave Peril by Jim Butcher
1.825 The Letters of Jane Austen (1-16) by Jane Austen
Średnie delty

O CO CHODZI Z LISTAMI?

PODSUMOWANIE

- **poszlakowe** potwierdzenie autorstwa
- mocno osadzona w **kontekście** - ważny jest odpowiedni dobór korpusu
- wyliczenia są dość proste, pogubić się można tylko w natłoku tekstów, grup, częstości i wyrazów

formalnie: Burrows oryginalnie nie bazował na grupach autorskich, tj. porównywał tekst testowy z każdym innym tekstem, a dopiero potem wyniki łączył po autorach, jednak dla uproszczenia zastosowałam swoją wariację grupową

statystycznie: delta Burrowsa bazuje na standaryzacji Z (przewidzianej dla rozkładów normalnych) i odległości miejskiej (Manhattan distance), która lepiej działa dla rozkładów Laplace'a - dla standaryzacji Z lepsza byłaby odległość euklidesowa, dla odległości miejskiej lepsza byłaby standaryzacja medianą i rozrzutem

lista najczęstszych słów w korpusie - możemy zmieniać jej warianty: wybrać tylko te słowa, które są obecne we wszystkich tekstach (lub we wszystkich grupach autorskich) lub w jakimś ich procencie, najczęstsze słowa w danym języku w ogóle (w oparciu o statystyki językoznawców)

inne cele: ponieważ metoda ta mierzy różnice między tekstami, możemy zamiast grup autorskich sprawdzać znak emocjonalny tekstu, płeć autora, czas powstania tekstu i wiele innych rzeczy

ŹRÓDŁA

- *Questions of Attribution: Attribution and Beyond*, J.Burrows
- *Delta: A Measure of Stylistic Difference and Guide to Likely Authorship*, J.Burrows
- *Understanding and Explaining Delta measures for Authorship Attribution*, S.Evert i in.
- *Interpreting Burrows' Delta: Geometric and Probabilistic Foundations*, S.Argamon
- *Authorship Attribution*, P.Joula
- *An Open Stylometric System Based on Mutlilevel Text Analysis*, M.Eder, M.Piasecki, T.Walkowiak
- *Stylometria komputerowa w służbie tłumacza*, J.Rybicki
- *Metody ścisłe w literaturoznawstwie i pułapki pozornego obiektywizmu - przykład stylometrii*, M.Eder