Delta Барроуза и проблема авторства «Тихого Дона»

Борис Орехов

НИУ Высшая школа экономики

nevmenandr@gmail.com

14 февраля 2020

Тупиковая ситуация вокруг авторства

- Научная дискуссия вокруг авторства «Тихого Дона» зашла в тупик.
- Сторонники каждой из версий любой фактический аргумент трактуют в свою пользу.
- В таких случаях обычно могут помочь количественные методы определения авторства.
- Однако в связи с этой проблемой они также дискредитированы.
- Выхода из этого тупика скорее всего нет.
- Но если бы он всё же был, то он был бы там, где можно было преодолеть кризис доверия к количественным методам.

Кризис доверия

- Методика, которая будет использоваться для доказательства авторства «Тихого Дона» не должна быть создана специально для решения этой проблемы. Иначе создатели уже будут подозреваться в ангажированности.
- Методика должна показывать хорошие результаты для русского языка и на широком доказательном материале вообще.
- Такая методика есть, это Delta.

Delta

Входной параметр — распределения самых частотных слов:

$$\Delta = \sum_{i=1}^{n} \frac{|z(x_i) - z(y_i)|}{n} \tag{1}$$

Burrows J. F. Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship // Literary and Linguistic Computing 2002. 17(3): 267–287.

Барроуз первоначально показал эту методику на «Потерянном рае».

Чуть точнее

Берем не частотность слов, а z-score:

$$z = \frac{x - \mu}{\sigma} \tag{2}$$

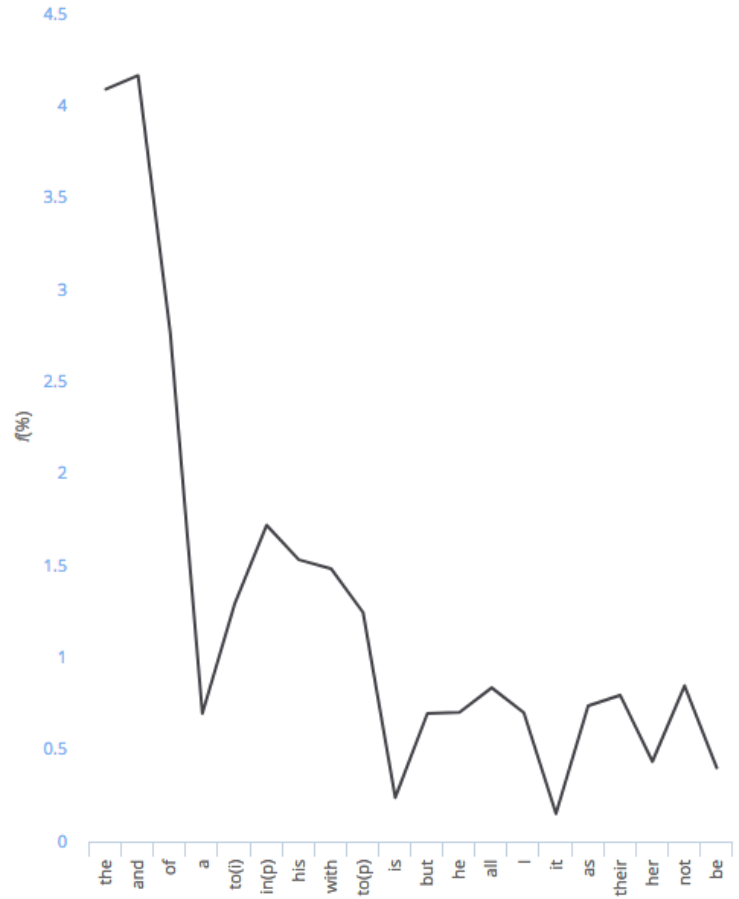
- x — частотность слова в тексте.
- μ — общая частотность слова по корпусу.
- σ — стандартное отклонение частотности слова по корпусу

Описание методики

- До начала расчетов необходимо задать количество наиболее частотных словоформ для всего корпуса. Например, 100.
  В большинстве случаев этого хватает для успешной атрибуции.
- Дальнейшие расчеты проводятся только для этих слов.
- Далее для каждого из этих слов в каждом из текстов корпуса вычисляется z-score
- Это отношение разницы взятой в процентах от общего числа слов в тексте частотности слова в данном тексте и общей частотности слова по всему корпусу к стандартному отклонению частотности слова по корпусу.
- Среднее арифметическое взятых по модулю разниц между z-score у двух сравниваемых текстов — это и есть искомое значение Delta.

Самые частотные слова «Потерянного рая»

PARADISE LOST
John Milton
William Congreve
Matthew Prior
Abraham Cowley
Nahum Tate
John Denham
Andrew Marvell
John Oldham
John Dryden
Thomas D'Urfey
Elkanah Settle
Thomas Shadwell
Jonathan Swift
Samuel Butler
Anne Wharton
Edmund Waller
Charles Cotton
Aphra Behn
Robert Gould
Charles Sedley
Charles Sackville
Alexander Radcliffe
Alexander Brome
John Wilmot
Katherine Phillips



NB.: Нет связи с темой текста!

Сравним суммы

PARADISE LOST

John Milton

William Congreve

Matthew Prior

Abraham Cowley

Nahum Tate

John Denham

Andrew Marvell

John Oldham

John Dryden

Thomas D'Urfey

Elkanah Settle

Thomas Shadwell

Jonathan Swift

Samuel Butler

Anne Wharton

Edmund Waller

Charles Cotton

Aphra Behn

Robert Gould

Charles Sedley

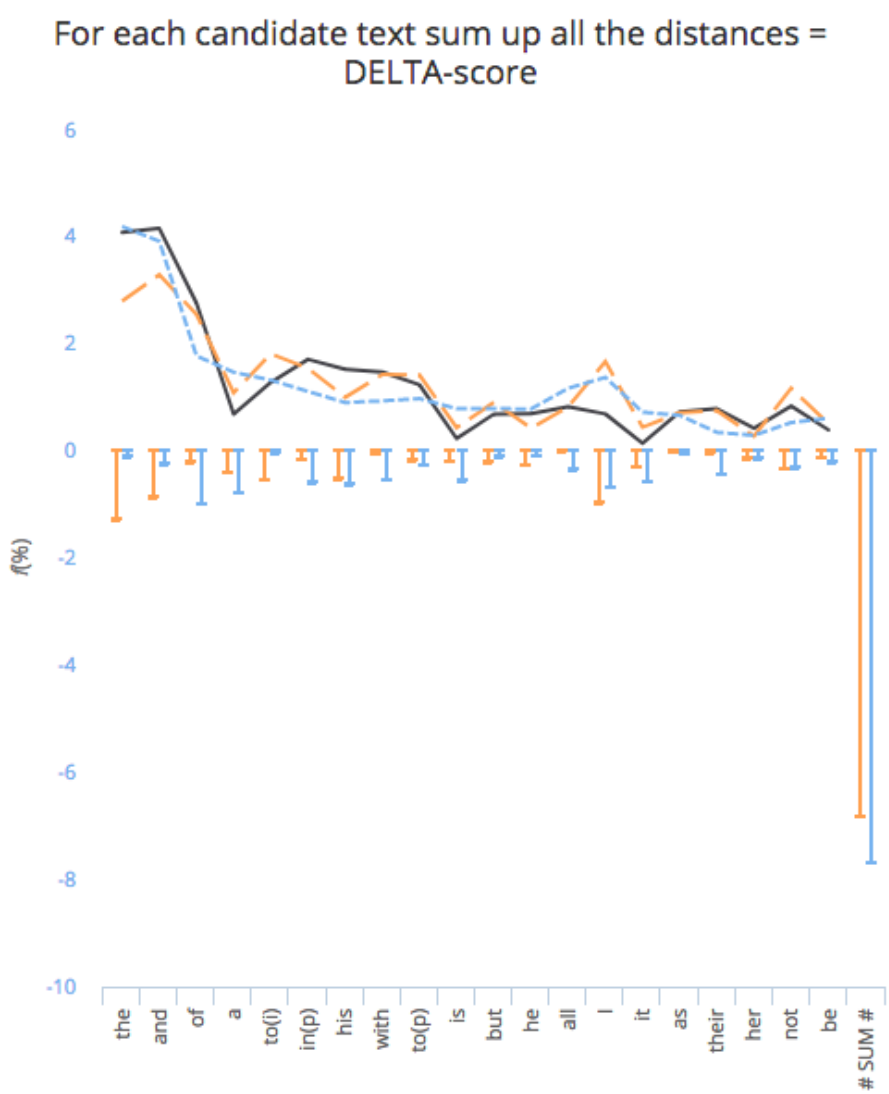
Charles Sackville

Alexander Radcliffe

Alexander Brome

John Wilmot

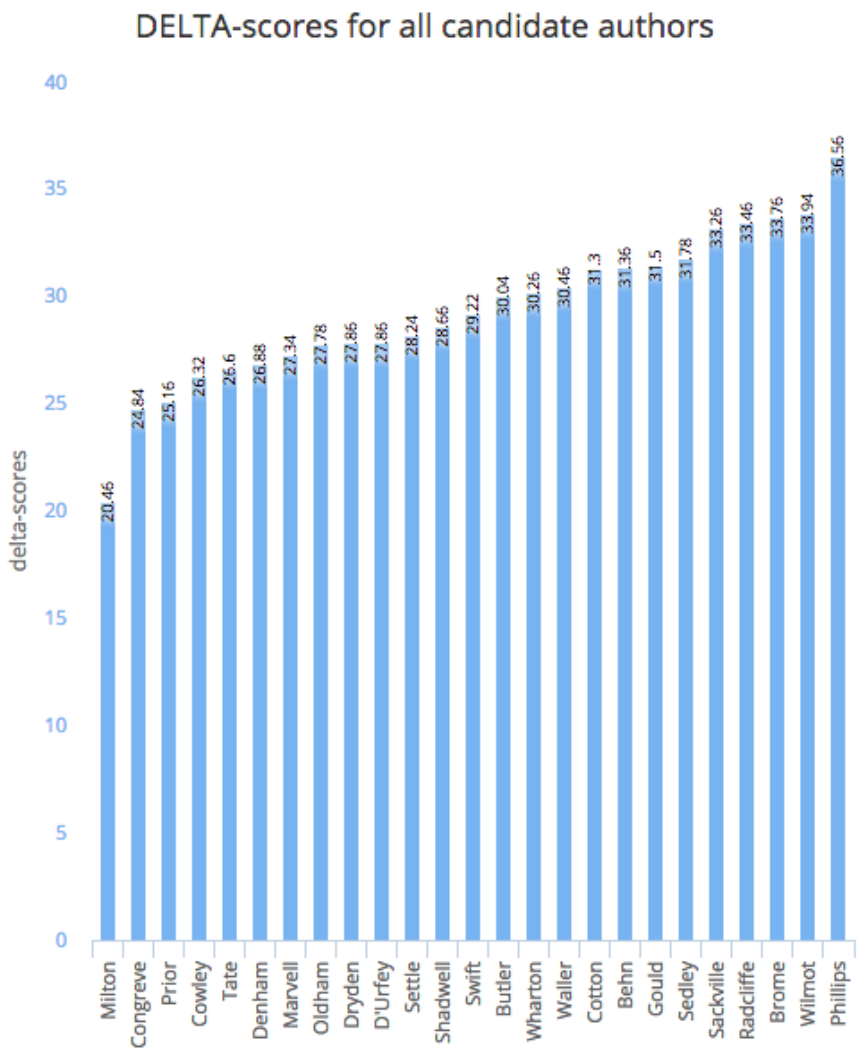
Katherine Phillips



Посмотрим на всех авторов

PARADISE LOST

John Milton
William Congreve
Matthew Prior
Abraham Cowley
Nahum Tate
John Denham
Andrew Marvell
John Oldham
John Dryden
Thomas D'Urfey
Elkanah Settle
Thomas Shadwell
Jonathan Swift
Samuel Butler
Anne Wharton
Edmund Waller
Charles Cotton
Aphra Behn
Robert Gould
Charles Sedley
Charles Sackville
Alexander Radcliffe
Alexander Brome
John Wilmot
Katherine Phillips



Кризис воспроизводимости

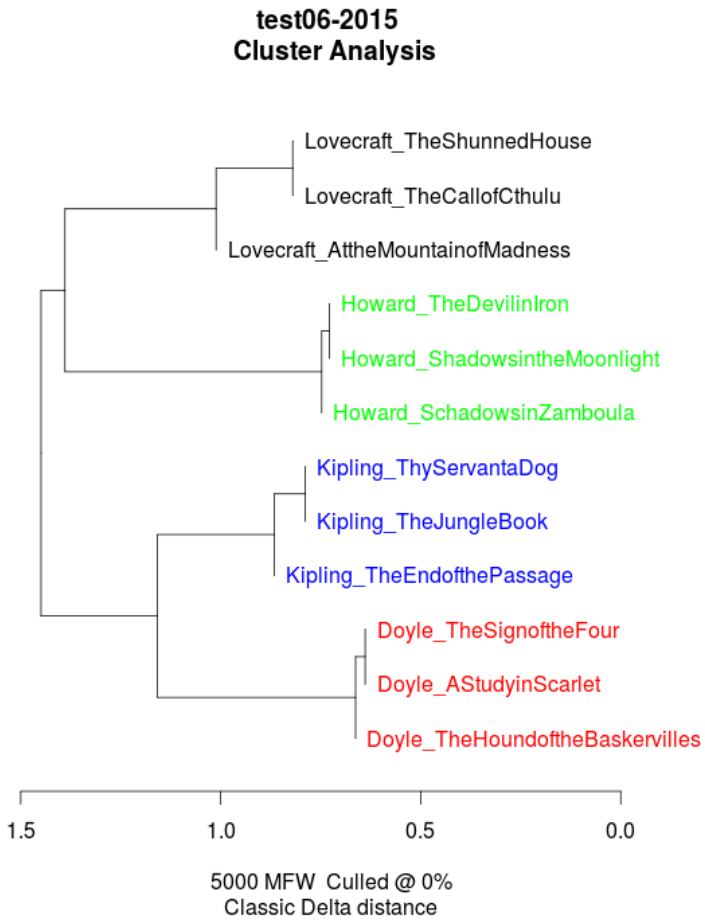
Большинство результатов, на которых часто основываются последующие исследования, попросту не воспроизводятся.

stylo

В последние годы появился пакет stylo, который позволяет быстро посчитать Delta для пользовательских текстов.
Eder M., Rybicki J., Kestemont M. Stylometry with R: a package for computational text analysis // R Journal. 2016. 8(1): 107-121.

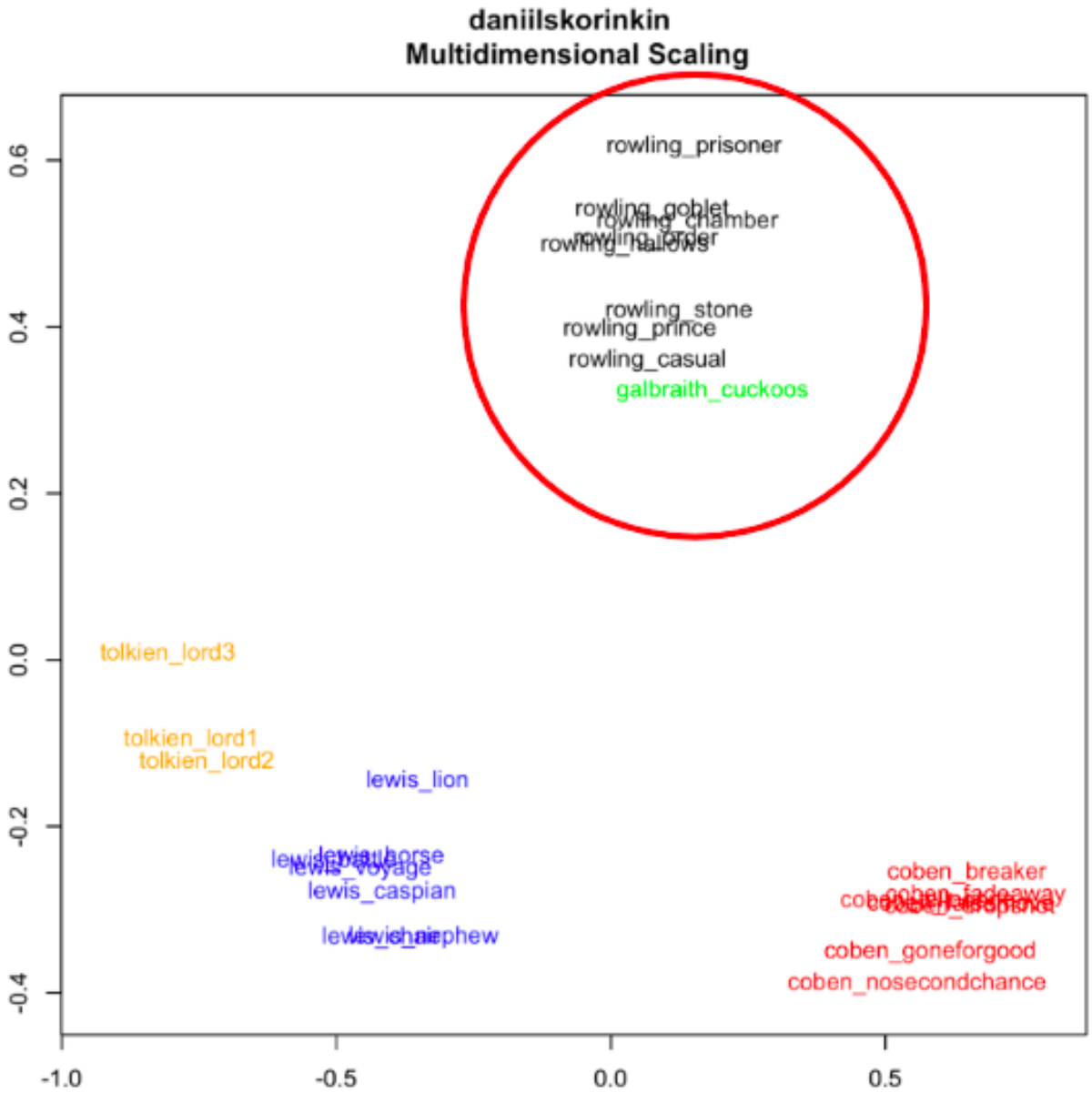
- Бесплатно.
- Есть графический интерфейс.
- Каждый в любой момент может проверить, как работает Delta на собственных текстах.

Как пользоваться результатами Delta?



NB.: Цветовая маркировка.

История успеха



Важно

Delta появилась как универсальный метод поиска «отпечатка пальца» автора в тексте, а не чтобы решить конкретную задачу атрибуции.

Есть ограничения:

- Слишком короткие тексты: до 10000 или до 5000 слов.
- Жанрово инородные тексты.

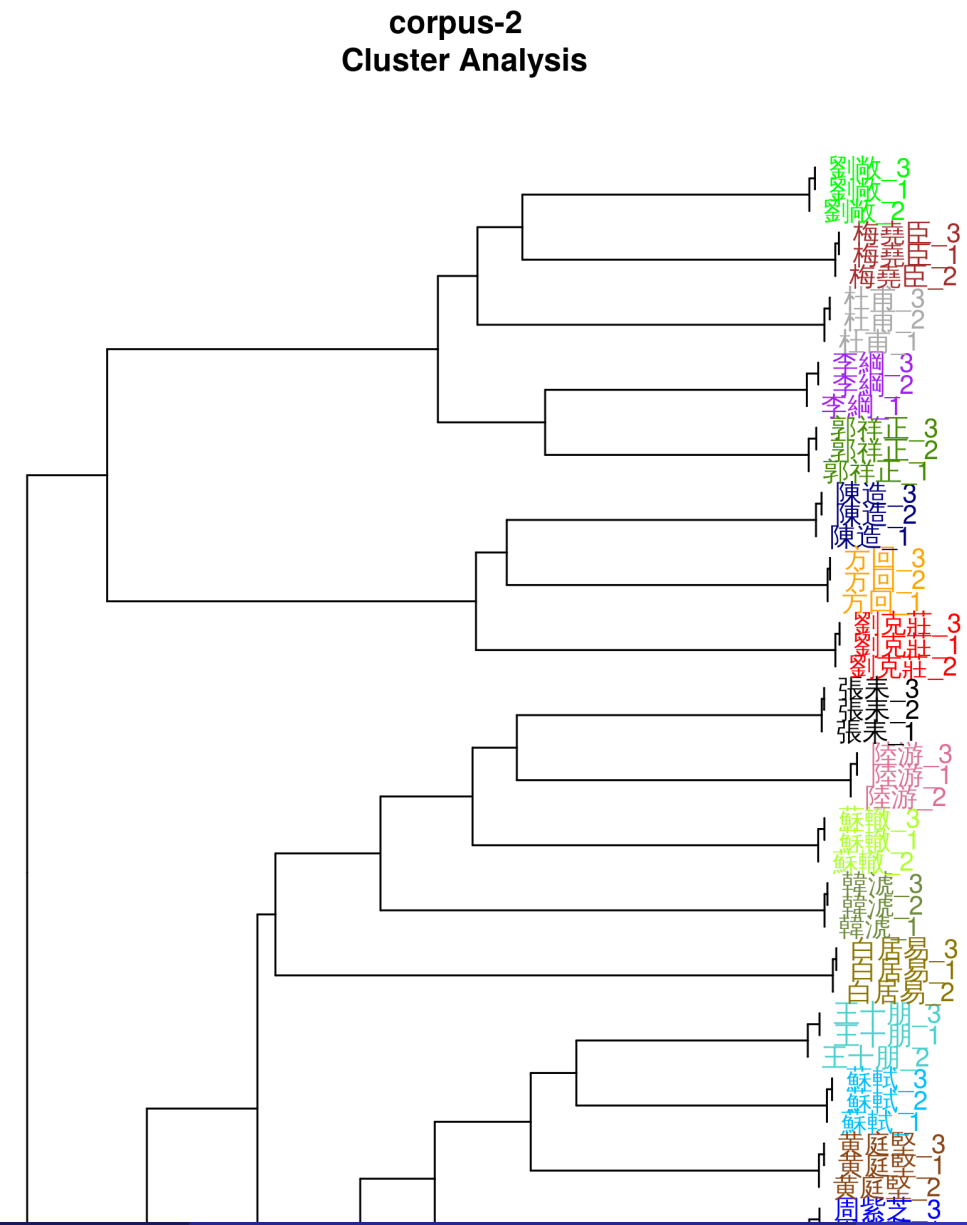
Для каких языков Delta работает?

- Арабский: AbdulRazzaq A.A., Mustafa T. K. Burrows-Delta Method Fitness for Arabic Text Authorship Stylometric Detection // International Journal of Computer Science and Mobile Computing, Vol.3 Issue.6, June- 2014, pg. 69 – 78;
- древнеанглийский: García A. M., Martín J. C. Function Words in Authorship Attribution Studies // Literary and Linguistic Computing. 2007. Vol. 22. № 1. P. 49 – 66;
- польский: Eder M., Rybicki J. PCA, Delta, JGAAP and Polish Poetry of the 16th and the 17th Centuries: Who Wrote the Dirty Stuff?' // Digital Humanities 2009: Conference Abstracts. : MD College Park, 2009. P. 242 – 244;
- немецкий: Jannidis F., Lauer G. Burrows's Delta and Its Use in German Literary History // Distant Readings. Topologies of German Culture in the Long Nineteenth Century Studies in German Literature Linguistics and Culture. / под ред. M. Erlin, L. Tatlock. Rochester: Camden House, 2014. P. 29 – 54.

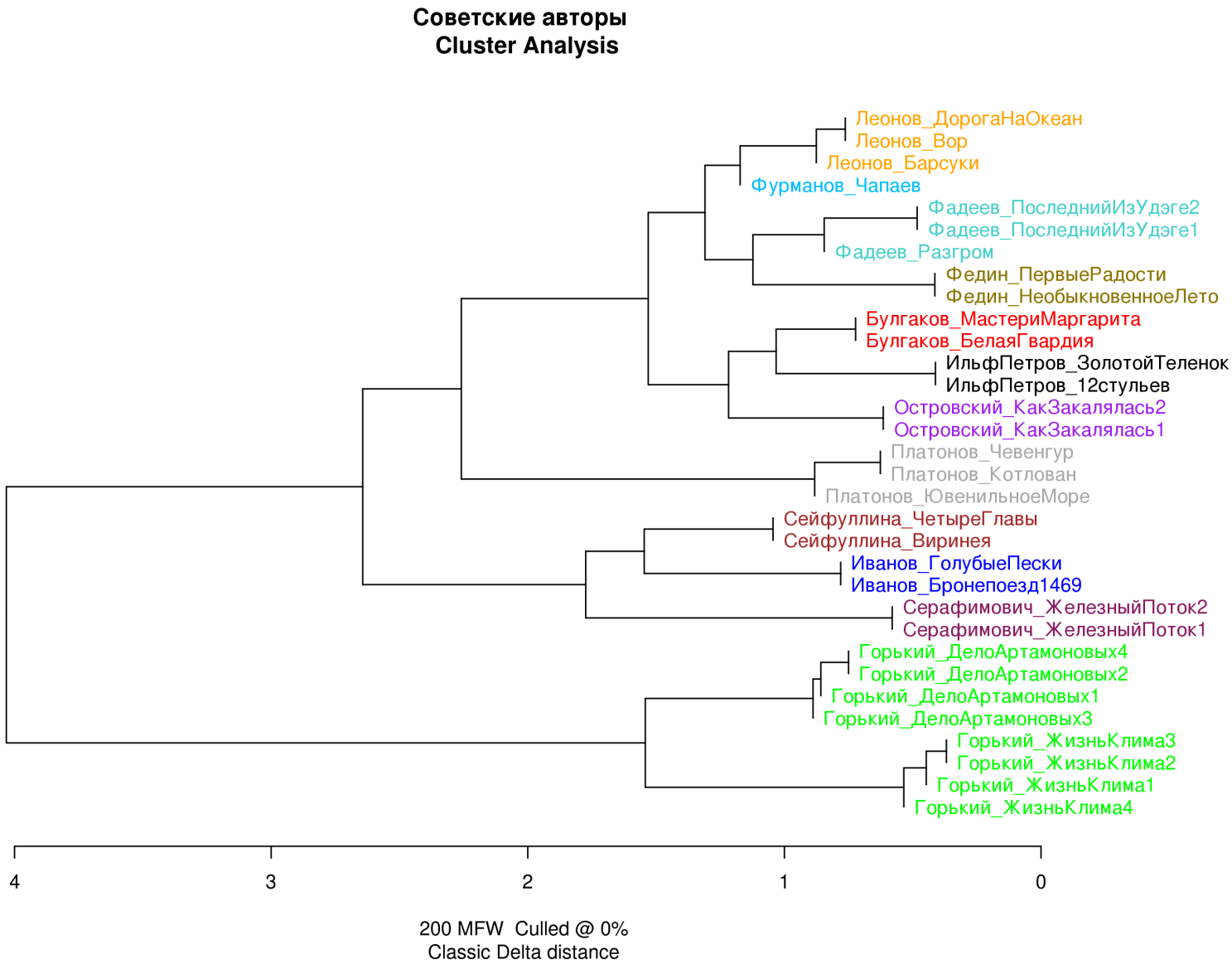
Легкость применения = широта использования

- *Alexander A. G. Gladwin Matthew J. Lavin Daniel M.* Look Stylometry and collaborative authorship: Eddy, Lovecraft, and 'The Loved Dead' // Digital Scholarship in the Humanities, Volume 32, Issue 1, 1 April 2017, Pages 123–140, https://doi.org/10.1093/llc/fqv026;
- *José Calvo Tello* What does Delta see inside the Author?: Evaluating Stylometric Clusters with Literary Metadata 153–161; Hartmut Ilsemann Stylometry approaching Parnassus // Digital Scholarship in the Humanities, Volume 33, Issue 3, 1 September 2018, Pages 548–556, https://doi.org/10.1093/llc/fqx058;
- *Michael P. Oakes* Computer stylometry of C. S. Lewis's The Dark Tower and related texts // Digital Scholarship in the Humanities, Volume 33, Issue 3, 1 September 2018, Pages 637–650, https://doi.org/10.1093/llc/fqx043;
- *Jacques Savoy* Is Starnone really the author behind Ferrante? // Digital Scholarship in the Humanities, Volume 33, Issue 4, 1 December 2018, Pages 902–918, https://doi.org/10.1093/llc/fqy016.

Для китайского



Работает ли на русских текстах?



«Тихий Дон»

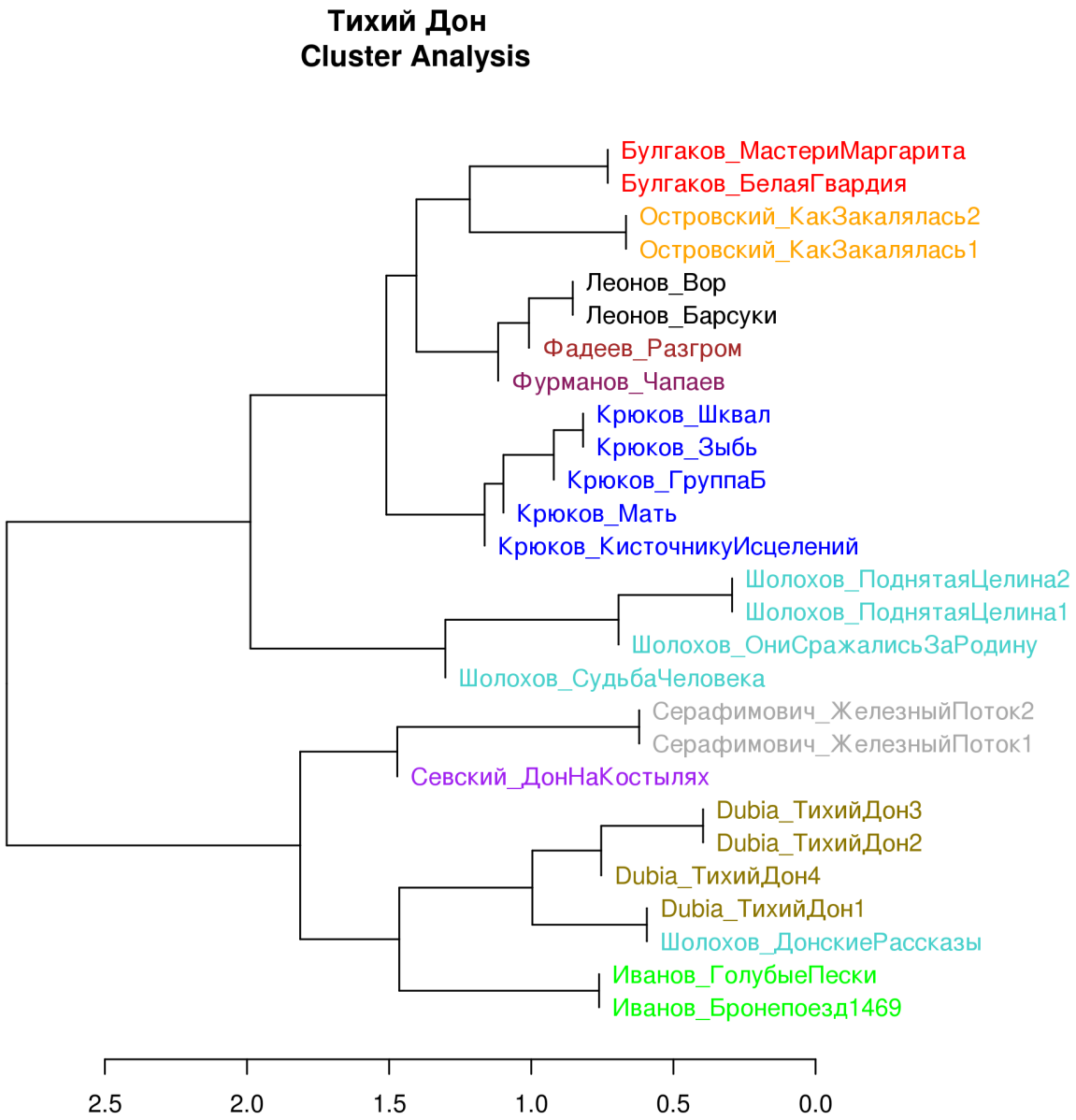
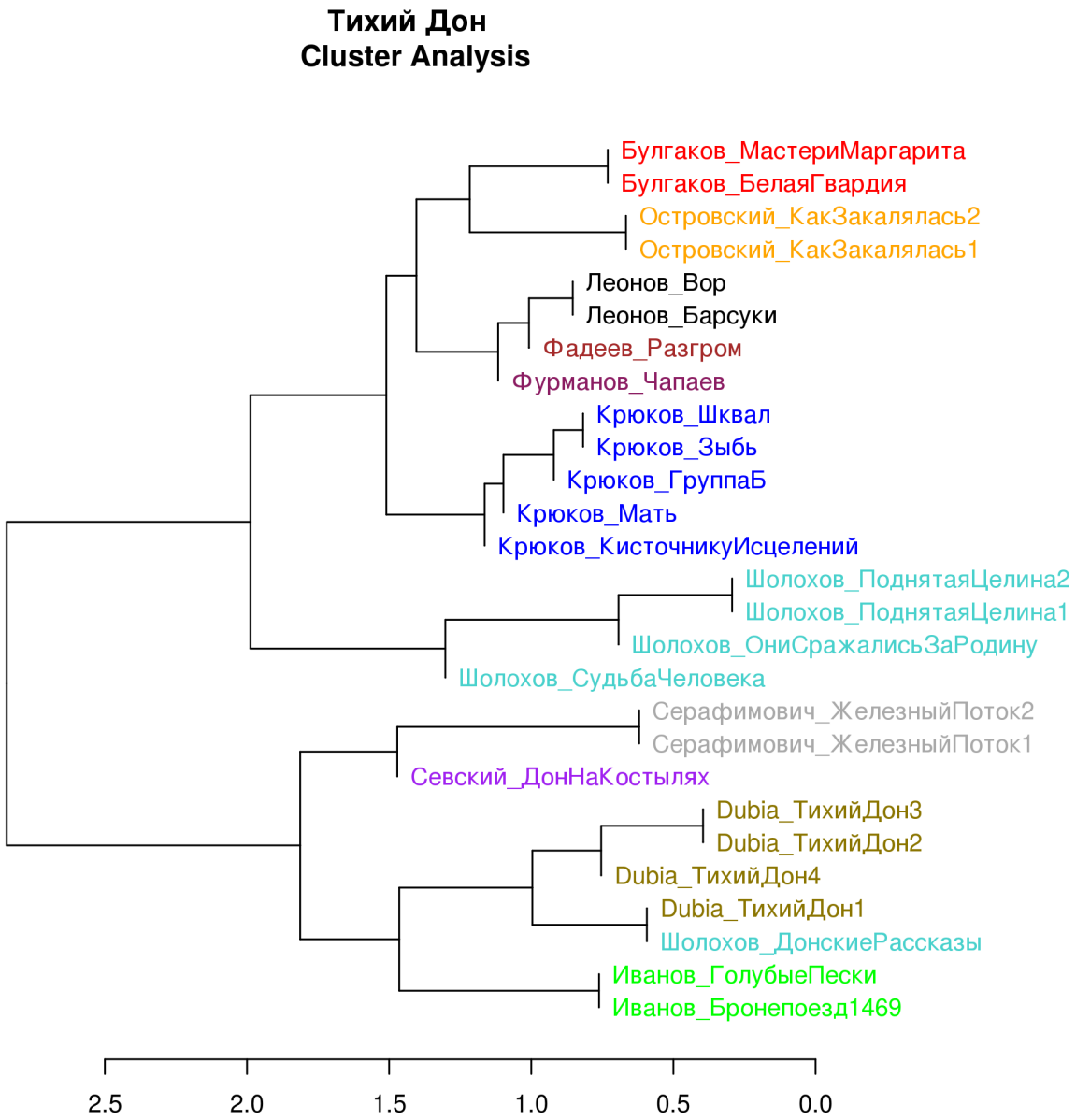

Тихий Дон
Cluster Analysis
Булгаков_МастериМаргарита
Булгаков_БелаяГвардия
Островский_КакЗакалялась2
Островский_КакЗакалялась1
Леонов_Вор
Леонов_Барсуки
Фадеев_Разгром
Фурманов_Чапаев
Крюков_Шквал
Крюков_Зыбь
Крюков_ГруппаБ
Крюков_Мать
Крюков_КисточникуИсцелений
Шолохов_ПоднятаяЦелина2
Шолохов_ПоднятаяЦелина1
Шолохов_ОниСражалисьЗаРодину
Шолохов_СудьбаЧеловека
Серафимович_ЖелезныйПоток2
Серафимович_ЖелезныйПоток1
Севский_ДонНаКостылях
Dubia_ТихийДон3
Dubia_ТихийДон2
Dubia_ТихийДон4
Dubia_ТихийДон1
Шолохов_ДонскиеРассказы
Иванов_ГолубыеПески
Иванов_Бронепоезд1469
2.5
2.0
1.5
1.0
0.5
0.0

Таблица с данными

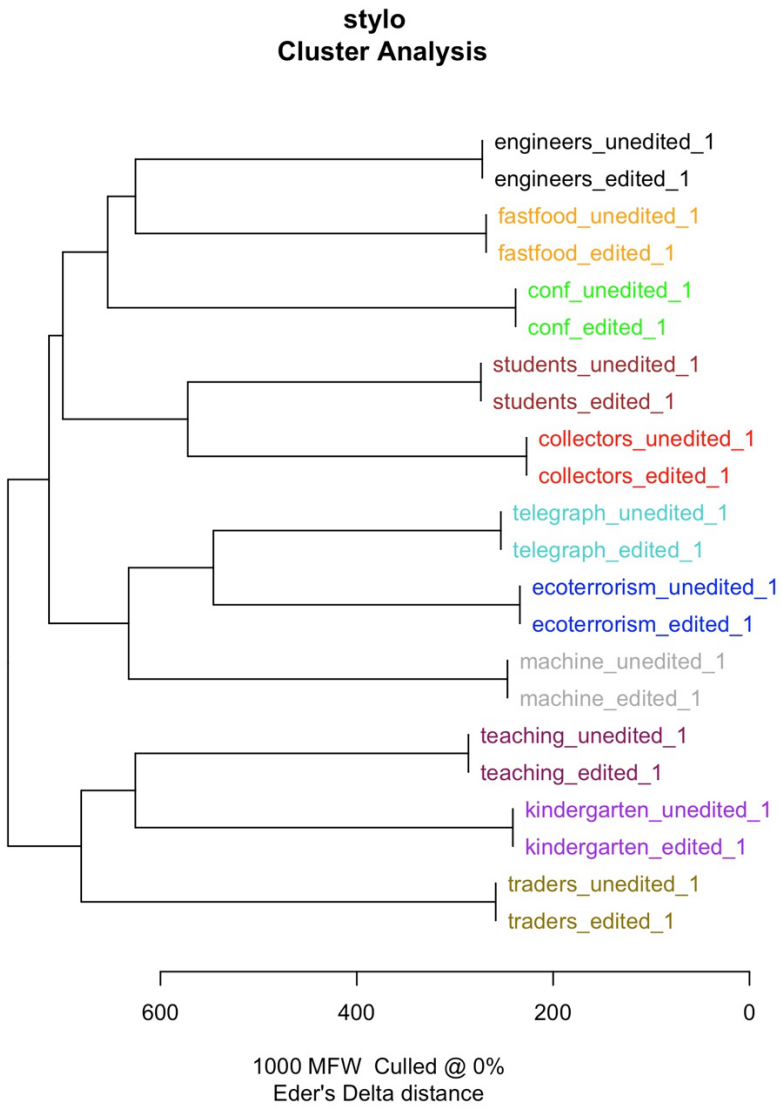
Таблица:

Текст	Булгаков_БГ	Булгаков_МиМ	Иванов_Бр1469
Булгаков_БГв	0	0.7092	1.0459
Булгаков_МиМ	0.7092	0	1.29103
Иванов_Бр1469	1.0459	1.2910	0
Иванов_ГолПес	0.8375	1.0805	0.74697

Текстологически исправный «Тихий Дон»



Отредактированный текст



Итоги

- Delta достаточно надёжный инструмент (есть и исключения).
- Благодаря stylo очень доступный.
- Создан не под конкретную задачу, а как универсальный.
- Отлично работает для русского языка.
- Свидетельствует, что все тома «Тихого Дона» написаны одним человеком.
- Это не Федор Крюков.
- Видно, что «Тихий Дон» написан тем же человеком, что и «Донские рассказы».