Вычисление межтекстового расстояния при количественном определении авторства

Борис Орехов nevmenandr@gmail.com

Эта презентация

Атрибуция

Зачем нужна атрибуция

Атрибуция — определение авторства — имеет много применений как в науке, так и в приземленных задачах, например, в судебной практике.

Сегодня мы будем говорить о текстовой атрибуции.

Иногда от того, кто автор текста, зависит судьба крупной суммы денег, иногда — карьера или свобода человека, а иногда людям просто интересно, кто же написал то или иное произведение.

Атрибуция и художественная литература

С точки зрения науки о литературе, строго говоря, всё равно, является автором «Тихого Дона» Михаил Шолохов или Фёдор Крюков, но общественность этот вопрос волновать не перестанет, наверное, уже никогда.

Иногда определение авторства художественного произведения мыслится как единственная задача цифровой филологии (digital literary studies)

DH и количественная атрибуция

That even after more than a decade of energetic speculation, the phrase and the concept “digital humanities” still frustrates attempts at provisional definition, let alone precision, is a liability and a predicament for anyone who has come to realize that sustained shouting about novelty only deafens. (p. XIII)

Brian Lennon Passwords: Philology, Security, Authentication . Cambridge, Mass.: Harvard University Press, 2018

Атрибуция неколичественная и количественная

Атрибуция может выглядеть по-разному. Надежнее всего такая, которая основана на документах. Если бухгалтерия в своей строгой отчетности зафиксировала выплату денег за определенный текст некоторому лицу, очень высока вероятность, что получатель денег и есть автор этого текста. Вероятность этого выше, чем **при любом другом** способе выяснить, кто автор.

Текст как единственный источник информации

Но у нас не всегда есть надежные документы. И даже чаще их нет. Тогда единственным способом докопаться до истины будет сам текст, и люди верят, что, опираясь только на содержащиеся в нем косвенные свидетельства, можно установить, кто его написал. Хотя вообще-то это не более чем самонадеянная гипотеза, но мифы эпохи модерна слишком сильны.

Мифы атрибуции

Но вокруг атрибуции в принципе много мифов. Например, может показаться, что установить авторство можно по одному-единственному слову. Так, были те, кто считал, что автором известного произведения русской литературы XVIII в. «Отрывок путешествия в*** И*** Т***» является А. Н. Радищев потому что Радищев написал другое, самое известное, произведение русской литературы, содержащее в своем заглавии слово «путешествие». Так себе доказательная база. Чтобы всерьез об этом рассуждать, нужно забыть, что были и другие авторы, писавшие про путешествия. Например, Василий Алексеевич Лёвшин создал роман «Новейшее путешествие» (1784) и Михаил Михайлович Щербатов написал «Путешествие в землю Офирскую господина С..., шведского дворянина» (тоже 1784).

Слово «путешествие» и частотный словарь

Вообще, такое заметное слово, как «путешествие», вряд ли годится как аргумент, оно входит в топ 2700 самых частотных слов русского языка. Так что опираться только на него в рассуждениях наивно. Кто угодно может употребить это слово, язык для всех общий.

Необходимость проверок

Эта область чрезвычайно похожа на доказательную медицину. Тут надо все проверять и всякие «общие соображения» могут оказаться пустышкой.

Каждый второй «знает», как вычислять авторство в тексте, потому что это кажется ему очевидным, или он что-нибудь читал про это много лет назад (например, у родителей академика Фоменко), но никто не знает, насколько эти методы действительно работают. Проверять слишком скучно.

Плохой пример

Дэвид Робинсон из Heap 7 лет назад задался вопросом, кто написал колонку в «New York Times», представившись сотрудником администрации Трампа. Робинсону показалось, что будет удачной идеей найти специфичные для каждого твиттер-аккаунта представителя администрации Трампа слова с помощью метрики TF-IDF и посчитать косинусную близость между статьей и твитами разных близких к Трампу людей. Вышло, что наиболее вероятный кандидат на авторство колонки — сам Трамп. Или кто-то из Госдепа. Про некоторые наборы научных идей говорят, что элементы в них либо тривиальны, либо неверны.

TF-IDF

$$tf(t,d) = \frac{n_t}{\sum_k n_k}$$

где n_t число вхождений слова t в документ, а в знаменателе — общее число слов в данном документе.

$$idf(t,D) = \log \frac{|D|}{|\{\, d_i \in D \mid t \in d_i \,\}|}$$

где

|D| — число документов в коллекции; $|\{\, d_i \in D \mid t \in d_i \,\}|$ — число документов из коллекции D, в которых встречается t (когда $n_t \neq 0$)

Шекспир и военная разведка

But the 1848 revolutions swept many of the European “black chambers” away, and when an imperial United States assembled its first cryptanalytic military intelligence service in 1917, it was recruiting its personnel from two seemingly unlikely places: Riverbank Laboratories, a private research foundation in Illinois whose staff included amateur scholars laboring to decipher Francis Bacon’s ostensibly enciphered authorship of Shakespeare’s plays; and university departments of English, especially at Chicago and Yale, where the Baconians’ Stratfordian opponents in Shakespeare studies served on the faculty.

Brian Lennon Passwords…

Стилометрия

Шире, чем атрибуция

«Стилометрия — это статистический анализ стилистики посредством компьютерного анализа текста».

«Python для хакеров: нетривиальные задачи и проекты» (2023) Л. Вогана

(Глава 2, стр. 57—84)

Но! Стилометрия началась (самое позднее, а есть мнения, что и раньше) в середине XIX века, когда никаких компьютеров еще не было. А само слово «стилометрия» применил к текстам современник Ленина.

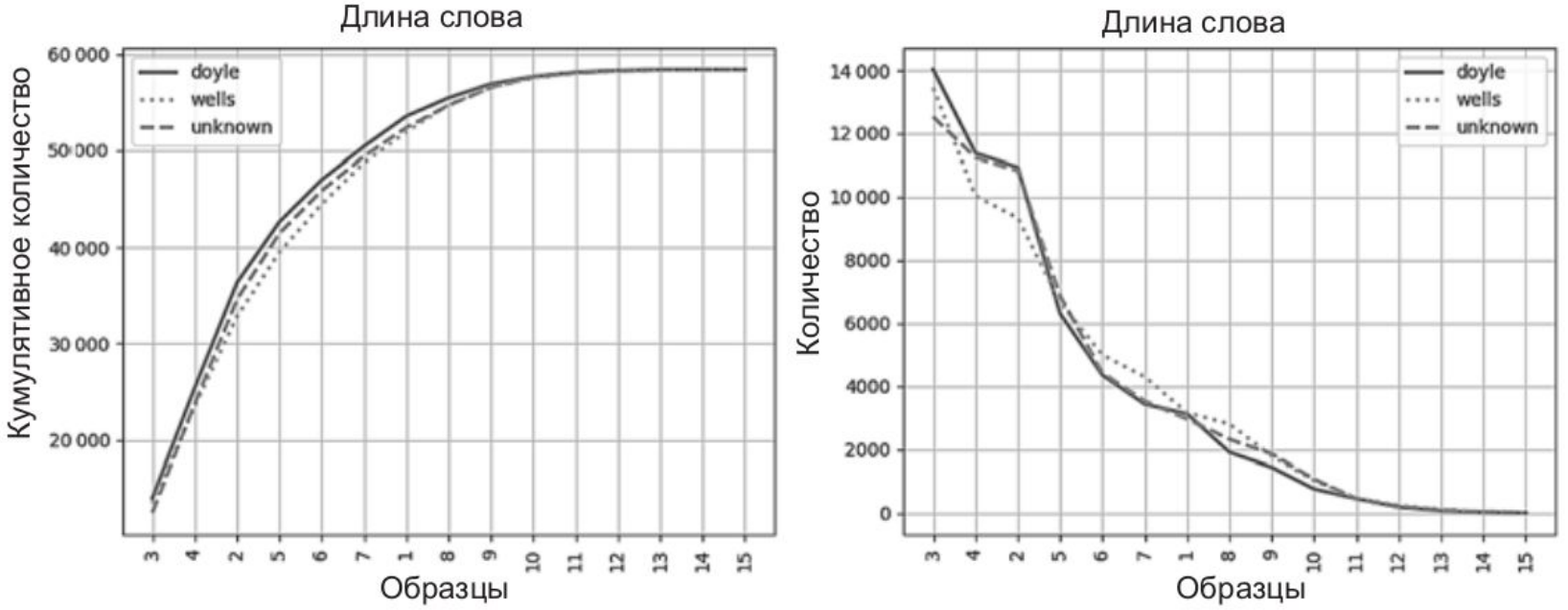
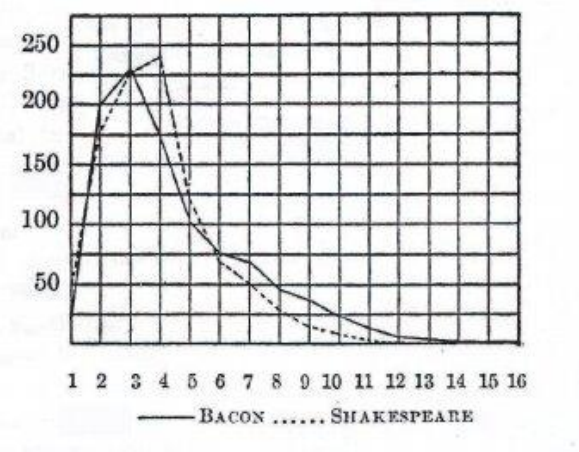


Рис. 2.3. Кумулятивный график NLTK (слева) и распределение частотности по умолчанию (справа)

Графики из книги Л. Вогана, стр. 70

Почему линейный график, а не столбчатый?

Все дело в том, что — осознанно или нет — Воган опирается здесь на работы геофизика Томаса Менденхолла 1880-х годов, который предполагал, что частотность слов в идиостиле автора должна работать как-то так же, как и открытый незадолго до этого спектральный анализ химических элементов.

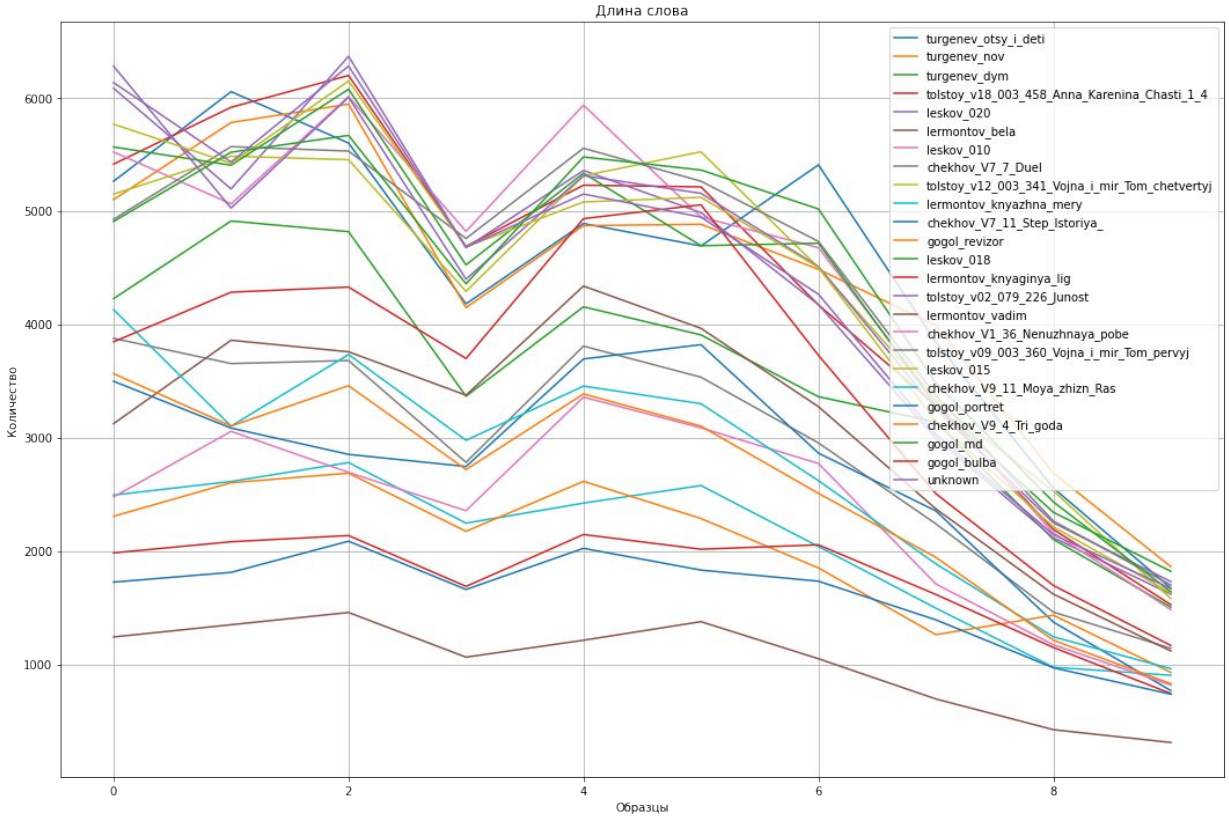


Работает ли такой подход?

Но с 1880-х годов научная мысль все-таки несколько ушла вперед, и все это из 2025-го выглядит достаточно наивно. Да и некритично воспринятый формат графиков показывает, что в питоне автор понимает больше, чем в количественной атрибуции.

Спектральный анализ, геофизики, графики — это все, конечно, интересно, но можно ли доверять этому методу? В книге Вогана есть пример, где он показывает, что «Затерянный мир», вроде как, несколько больше похож на «Собаку Баскервилей», чем на «Войну миров», но это неточно.


Не нужно покупать препараты
без доказанной эффективности
А гематогенку можно?
Гематогенку можно



Если взять побольше текстов и на русском

Филологи-классики и стилометрия

Хотя классическая филология была одной из первых гуманитарных дисциплин, обратившейся к количественным исследованиям, стилометрия никогда не рассматривалась филологами как самодостаточный метод. Суждение об авторстве текста, как писал Фридрих Бласс, должно учитывать данные рукописной традиции, свидетельства современников, соответствие бытовых и исторических реалий времени жизни автора, а также соответствие идей, тем и жанров тому, что известно по подлинным сочинениям автора. <...> Стилистические и языковые особенности — лишь один, не главный и не единственный, инструмент исследователя, а из этих особенностей лишь некоторые могут быть описаны количественно. В этом смысле и современные методы «атрибуции авторства», опирающиеся на статистику и технологии, могут играть только вспомогательную роль и не представляют угрозы традиционным подходам.

Алиева О. В. Меры расстояния для определения авторства древнегреческих текстов

В действительности все не так, как на самом деле

Подходы, основанные на расстоянии

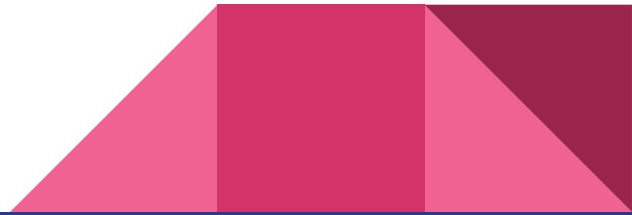
Текст или группа текстов могут быть представлены в виде вектора — упорядоченного множества значений, которые называются координатами или компонентами вектора. Для каждой пары векторов может быть вычислено расстояние или сходство между ними; минимальное расстояние или максимальное сходство будут указывать на возможного автора.

Алиева О. В. Меры расстояния…

Критерии метрики

функция считается метрикой расстояния, только если она удовлетворяет критериям неотрицательности, идентичности и симметричности, а также отвечает дополнительному условию — неравенству треугольника.

Алиева О. В. Меры расстояния…



Ключевая работа

state-of-the-art количественной атрибуции

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

Delta

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

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

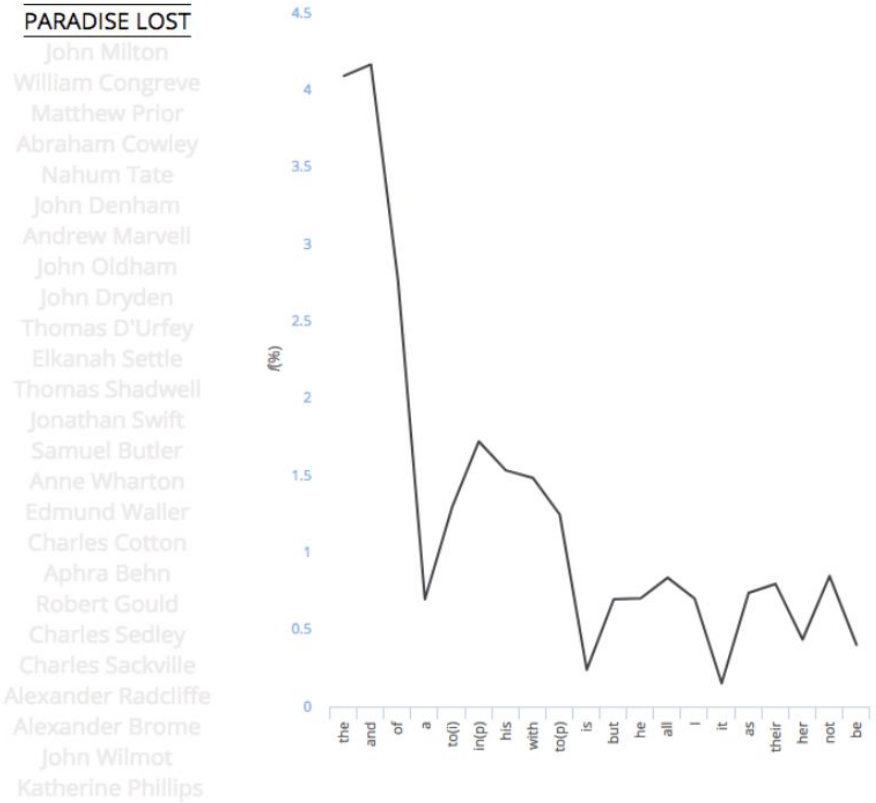
n — число слов для анализа, *x* — частотность, μ — среднее значение, σ — стандартное отклонение

Delta, howto

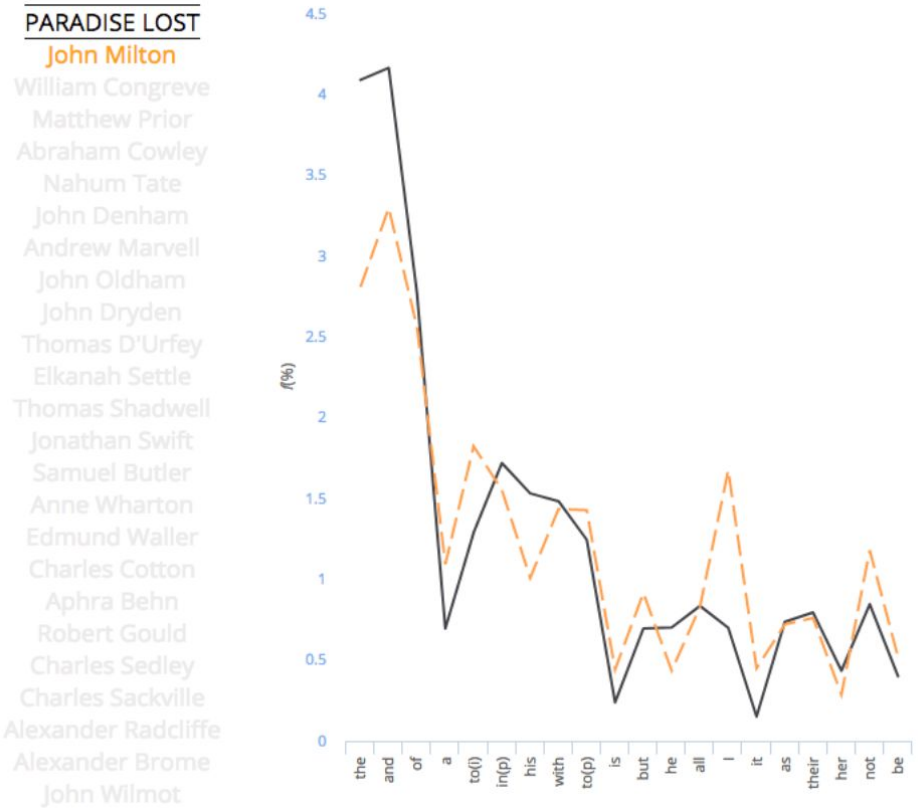
До начала расчетов необходимо задать некоторое количество наиболее частотных словоформ для всего корпуса. Например, 100. В большинстве случаев этого хватает для успешной атрибуции. Дальнейшие расчеты проводятся только для этих слов. Самыми частотными в корпусе являются служебные, а не полнозначные слова. Таким образом, не важна тематика текста. Для каждого из этих слов в каждом из текстов корпуса вычисляется z-score. Среднее арифметическое взятых по модулю разниц между z-score у двух сравниваемых текстов – это и есть искомое значение Delta.

Почему это работает, никто не знает, но это работает, что проверено на множестве разных кейсов и языков — от арабского до древнеанглийского. И даже с китайским все получается, хотя там все не так, как мы привыкли, и со словами, и с графикой.

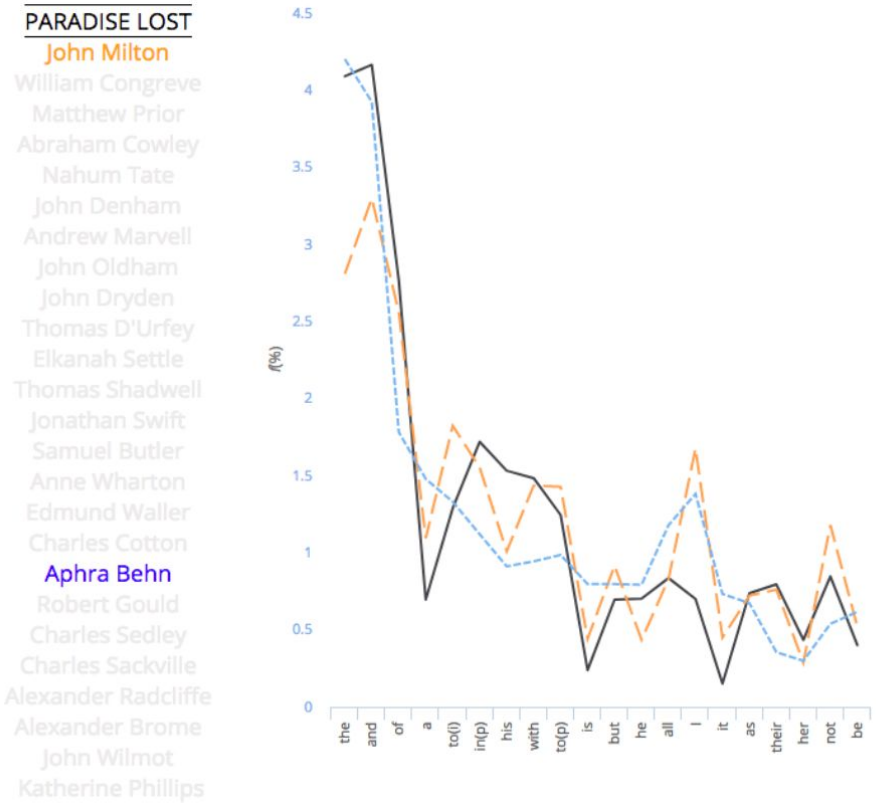
Разбираемся в смысле подсчетов



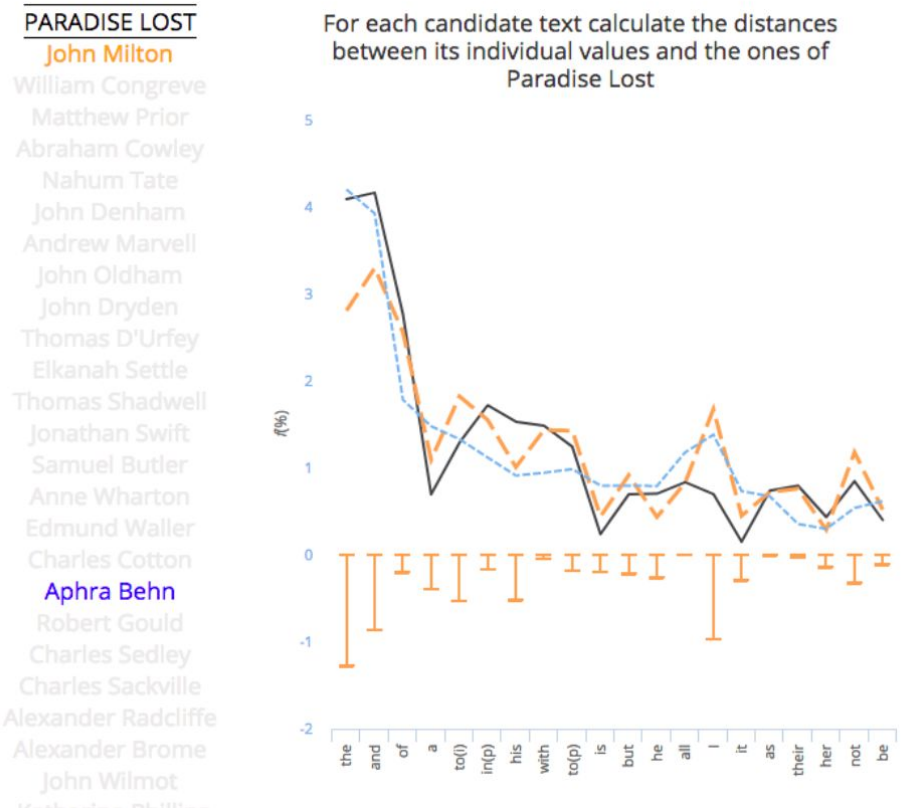
«Потерянный рай» Мильтона, MFW



Мильтон кандидат на авторство



автор-дистрактор



Расстояние между целевым значением и Мильтоном

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

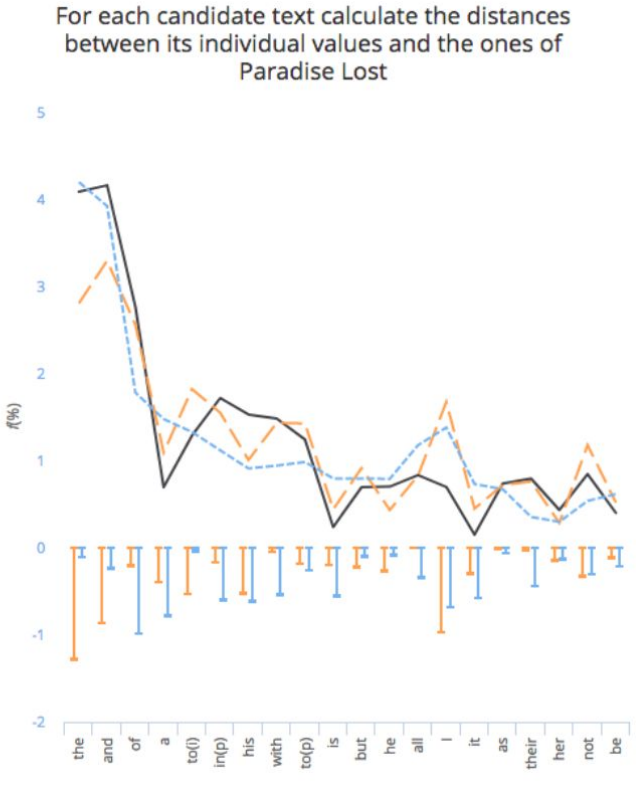
Charles Sedley

Charles Sackville

Alexander Radcliffe

Alexander Brome

John Wilmot



Все расстояния (слова)

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

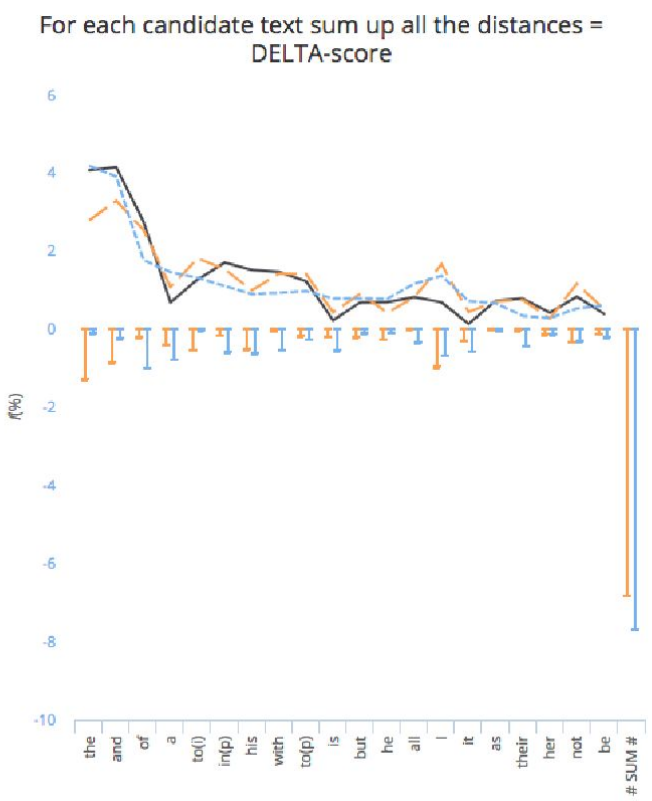
Alexander Brome

John Wilmot

Katherine Phillips

Общее расстояние (тексты)

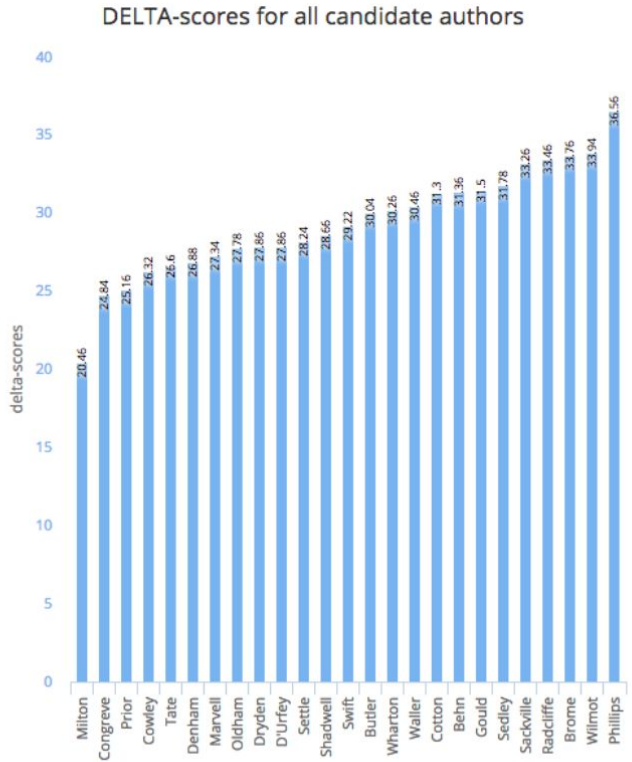
PARADISE LOST
John Milton
William Congreve
Matthew Prior
Abraham Cowley
Nahum Tate
John Denham
Andrew Marvell
John Oldham
John Dryden
Thomas D'Urfey
Elkanah Settle
Thomas Shadwell
Jonathan Swift
Samuel Butler
Anne Wharton
Edmund Waller
Charles Cotton
Aphra Behn
Robert Gould
Charles Sedley
Charles Sackville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



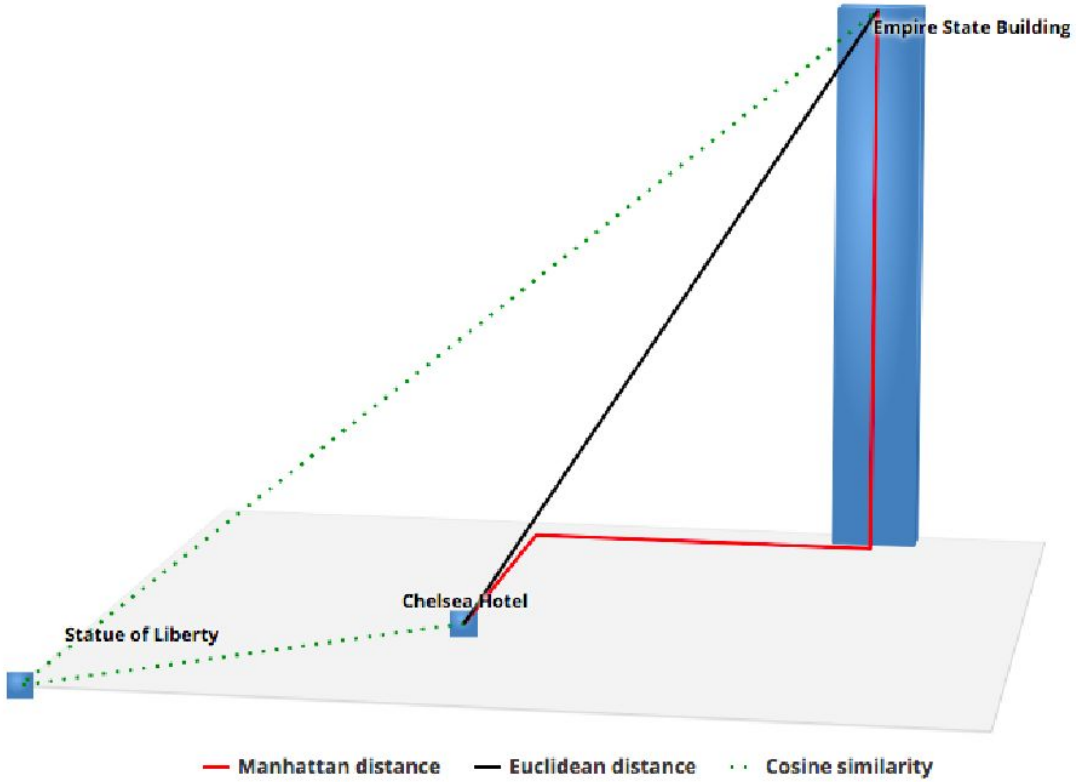
Разница в расстояниях

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



Все расстояния вместе (иллюстрации П. Плехача)


Empire State Building
Chelsea Hotel
Statue of Liberty
Manhattan distance
Euclidean distance
Cosine similarity

Расстояния

Другие расстояния

результаты. Манхэттенское расстояние, оно же расстояние городских кварталов, лежит в основе метода Берроуза [Burrows, 2002]. В качестве альтернативы предлагалось использовать евклидово расстояние [Argamon, 2008], а также косинусное сходство[5] [Smith, Aldridge, 2011. P. 79–80], в том числе с предварительной стандартизацией [Evert, Proisl et al., 2017]. Стандартизация признаков по z-оценке показывает, на сколько стандартных отклонений значения признака больше или меньше среднего арифметического. Высокая точность классификации достигались с использованием сходства Ружечки, оно же *minmax* [Koppel, Winter, 2014; Kestemont et al., 2016]. 1 – *minmax* эквивалентно расстоянию Танимото, и наоборот [Cha, 2007. P. 302]. Канберрское расстояние рекомен-

довалось для использования на арабском корпусе [Ahmed, 2019], а расстояние Кларка, среди прочих, — на английском и французском [Kocher, Savoy, 2019]. Из семейства энтропийных расстояний мы возьмем расхождение Джеффриса [Деза, Деза, 2008. С. 221], которое представляет собой симметричную версию расхождения Кульбака-Лейблера; последнее называют также относительной энтропией [Savoy, 2020. P. 39–42]. Поскольку перекрёстная энтропия H(P, Q) для распределений P и Q определяется как сумма энтропии H(P) и относительной энтропии DKL(P, Q), отдельно перекрёстную энтропию мы в этой работе не рассматриваем[6]. Кроме того, протестировано расстояние Лаббе [Labbé, Labbé, 2006; Labbé, 2007; Cortelazzo et al., 2013]. Формулы приведены в Табл. 1.

1	DManhattan	$\sum_{i=1i-i}^{n} PQ$
2	DEuclidean	$\sum_{i=1i-i}^{n2} PQ$
3	SCosine	$\frac{i=1nPiQi}{i=1nPi2i=1nQi2}$
4	DTanimoto	$\frac{i=1nmaxPi,Qi-minPi,Qi}{i=1nmaxPi,Qi}$
5	DCanberra	$\frac{i=1nPi-Qi}{Pi+Qi}$
6	DClark	$\frac{i=1nPi-Qi}{Pi+Qi2}$
7	DJeffreys	$\frac{i=1n(Pi-Qi)lnPi}{Qi}$
8	DLabbé	$\frac{i=1n\lvert Pi-Qi\rvert}{2NP}$

Таблица 1. Меры расстояния и сходства

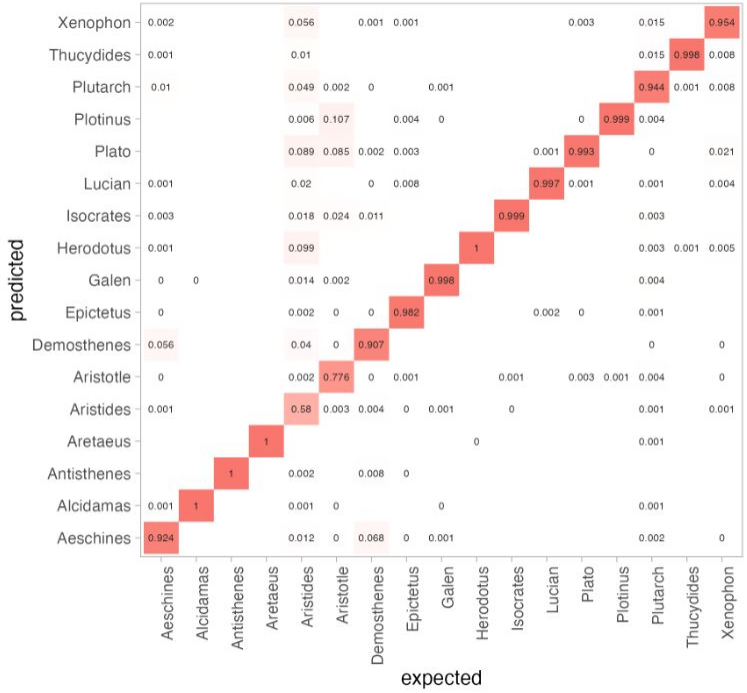


Рис. 8. Матрицы ошибок. COS_S

Как считать?

Не вручную, конечно

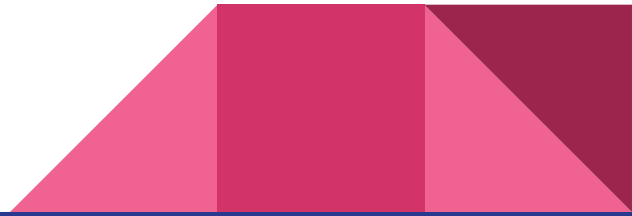
R:

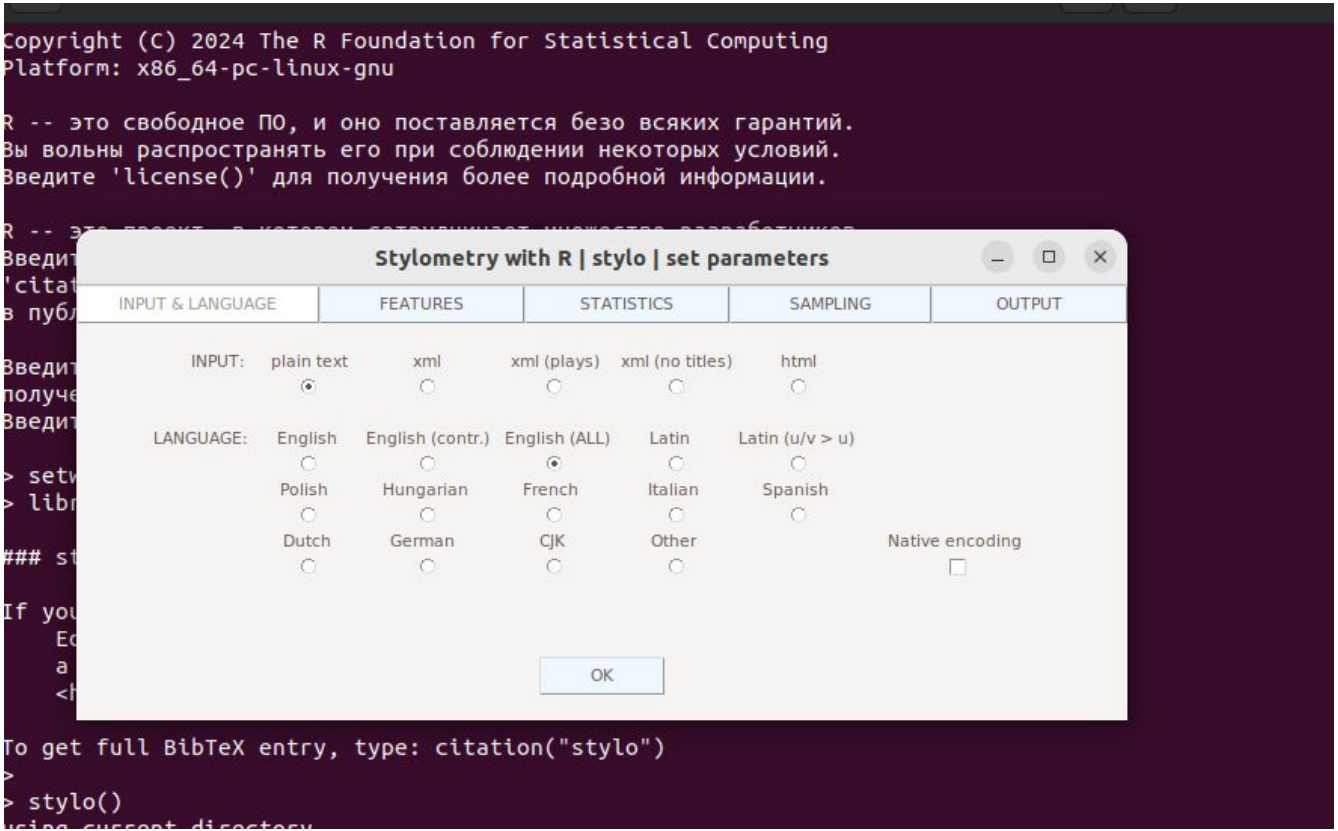
```
install.packages('stylo') # ставим пакет

setwd('R_stylo') # устанавливаем соотв. папку

library(stylo) # подключаем библиотеку

stylo() # вызываем GUI
```

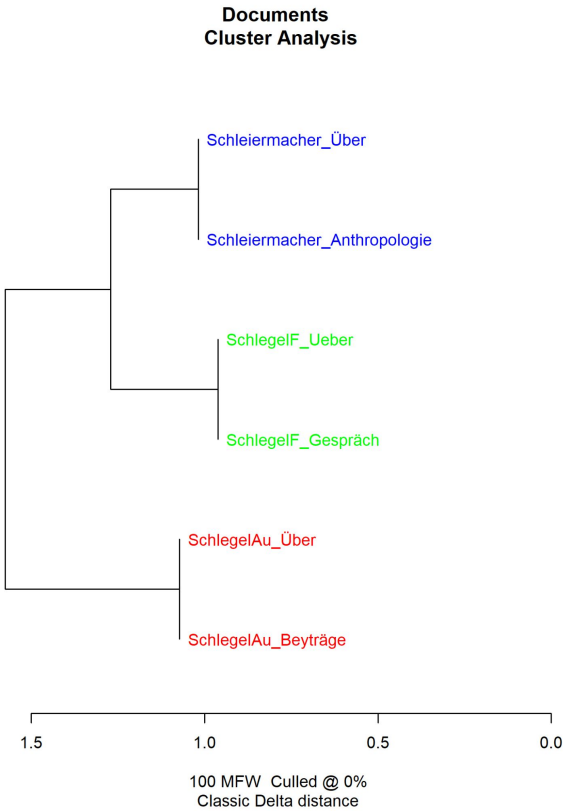



Copyright (C) 2024 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu
R -- это свободное ПО, и оно поставляется безо всяких гарантий.
Вы вольны распространять его при соблюдении некоторых условий.
Введите 'license()' для получения более подробной информации.
Stylometry with R | stylo | set parameters
INPUT & LANGUAGE
FEATURES
STATISTICS
SAMPLING
OUTPUT
INPUT:
plain text
xml
xml (plays)
xml (no titles)
html
LANGUAGE:
English
English (contr.)
English (ALL)
Latin
Latin (u/v > u)
Polish
Hungarian
French
Italian
Spanish
Dutch
German
CJK
Other
Native encoding
OK
To get full BibTeX entry, type: citation("stylo")
> stylo()

GUI

Таблицы расстояний

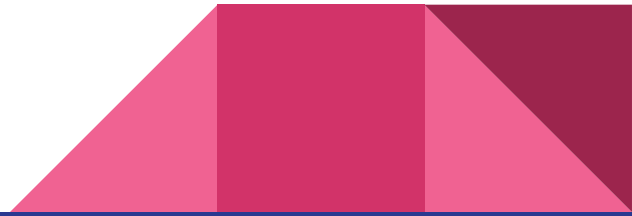
	Булгаков_БелаяГвардия	Булгаков_МастериМаргарита	Иванов_Бронепоезд1469	Иванов_Г
Булгаков_БелаяГвардия	0	0.730950155372916	1.05660639749339	0.8698296
Булгаков_МастериМаргарита	0.730950155372916	0	1.3174371694465	1.1199071
Иванов_Бронепоезд1469	1.05660639749339	1.3174371694465	0	0.7615374
Иванов_ГолубыеПески	0.869829639790795	1.11990710428626	0.761537458827709	
Крюков_ГруппаБ	0.820738135825795	1.01968953959767	1.19738454977875	1.0057418
Крюков_Зыбь	0.973386567603556	1.04412818308415	1.21756759098551	1.0931547
Крюков_КисточникуИсцелений	1.07175010219892	1.21815831877945	1.27472584617636	1.1863708
Крюков_Мать	1.0794748785247	1.18827291371373	1.32158306043481	1.2171530
Крюков_Шквал	0.886868198366372	0.879857137541794	1.22338514122809	1.0956309
Леонов_Барсуки	0.833198143923984	0.949506879831114	1.10983721251759	0.9504977
Леонов_Вор	1.02090155667755	1.04338481244864	1.44858868359607	1.2069326
Островский_КакЗакалялась1	0.885770005513563	0.988091445370399	1.11178251527831	0.9291667
Островский_КакЗакалялась2	0.946615925092193	1.01015168463526	1.27568075417001	1.0020556
Севский_ДонНаКостылях	1.14273385903323	1.33579695568848	1.39339484168868	1.2222270
Серафимович_ЖелезныйПоток1	1.07111641461049	1.39581360021947	1.18060886872941	1.1162498
Серафимович_ЖелезныйПоток2	1.06320835206475	1.40484567726945	1.19767411437329	1.1049972
Фадеев_Разгром	1.04680113990552	0.997192823593271	1.31426456495089	1.2495911
Фурманов_Чапаев	1.12192612286954	1.13760697620434	1.3422335595438	1.2125847
Шолохов_ДонскиеРассказы	0.946194822103831	1.29105568259361	1.0574695449379	0.8776724
Шолохов_ОниСражалисьЗаРодину	0.9001478267005	0.972142581818599	1.18347893215187	1.0399216
Шолохов_ПоднятаяЦелина1	0.888338162388753	0.95133229035956	1.19601648484496	0.9807304



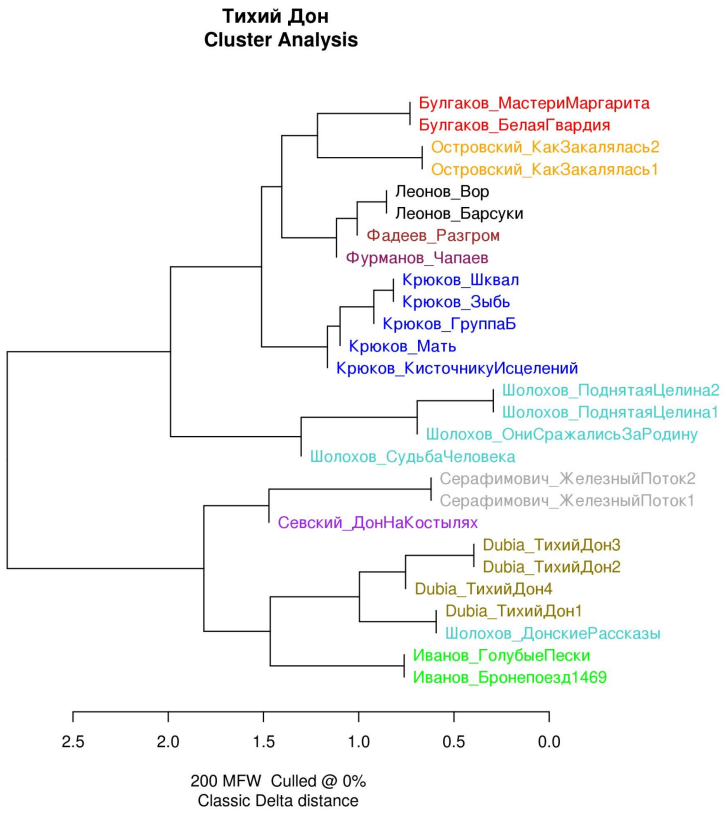
Дендрограмма-визуализация

Ограничения метода

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



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
200 MFW Culled @ 0%
Classic Delta distance

Не Крюков

Литература

Алиева О. В. Меры расстояния для определения авторства древнегреческих текстов // Цифровые гуманитарные исследования. 2024. № 1. С. 8–33.

Орехов Б. В. Как на самом деле определять автора с помощью компьютера? // Хабр. 2024. URL: https://habr.com/ru/articles/834912/

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

Lennon B. Passwords: Philology, Security, Authentication . Cambridge, Mass.: Harvard University Press, 2018.

Дополнительная литература

Великанова Н. П., Орехов Б. В. Цифровая текстология: атрибуция текста на примере романа М. А. Шолохова «Тихий Дон» // Мир Шолохова. Научно-просветительский общенациональный журнал. 2019. № 1. С. 70–82.

Eder M., Rybicki J., Kestemont M. Stylometry with R: a package for computational text analysis // R Journal. 2016. 8(1): 107-121.

Данные и код

Delta:

https://github.com/nevmenandr/delta_illustr/

«Тихий Дон»:

Орехов, Борис, 2020, "Стилеметрические данные «Тихого Дона» и современной ему прозы", https://doi.org/10.31860/openlit-2020.05-R001, Репозиторий открытых данных по русской литературе и фольклору, V1.