Факультет гуманитарных наук

Школа философии и культурологии

Москва
2023

Меры расстояния для определения авторства древнегреческих текстов

Ольга Алиева, к.ф.н., доцент, руководитель проектной группы «Цифровая античность»
oalieva@hse.ru

Постановка вопроса

- проблема самодостаточности количественных методов
- не все стилистические особенности могут быть описаны количественно
- количественные методы пригодны только для отрицательного решения вопроса о подлинности (Ф. Бласс, П. Фридлендер, А. Бёк);
- вмешательства редакторов и переписчиков
- авторские стилизации
- подвижность представлений об авторстве
- недостаточное число испытаний на древнегреческом корпусе

quivis praesumitur genuinus liber, donec demonstretur contrarium

Меры расстояния и сходства

1	$D_{Manhattan}$	$\sum_{i=1}^{n} \lvert P_i - Q_i \rvert$	5	$D_{Canberra}$	$\sum_{i=1}^{n} \frac{\lvert P_i - Q_i \rvert}{P_i + Q_i}$
2	$D_{Euclidean}$	$\sqrt{\sum_{i=1}^{n} \lvert P_i - Q_i \rvert^2}$	6	D_{Clark}	$\sqrt{\sum_{i=1}^{n} \left(\frac{\lvert P_i - Q_i \rvert}{P_i + Q_i} \right)^2}$
3	S_{Cosine}	$\frac{\sum_{i=1}^{n} P_i Q_i}{\sqrt{\sum_{i=1}^{n} P_i^2} \sqrt{\sum_{i=1}^{n} Q_i^2}}$	7	$D_{Jeffreys}$	$\sum_{i=1}^{n} (P_i - Q_i) ln \frac{P_i}{Q_i}$
4	$D_{Tanimoto}$	$\frac{\sum_{i=1}^{n} \left(max(P_i, Q_i) - min(P_i, Q_i) \right)}{\sum_{i=1}^{n} max(P_i, Q_i)}$	8	$D_{Labbé}$	$\frac{\sum_{i=1}^{n} \lvert P_i - Q_i \rvert}{2N_P}$

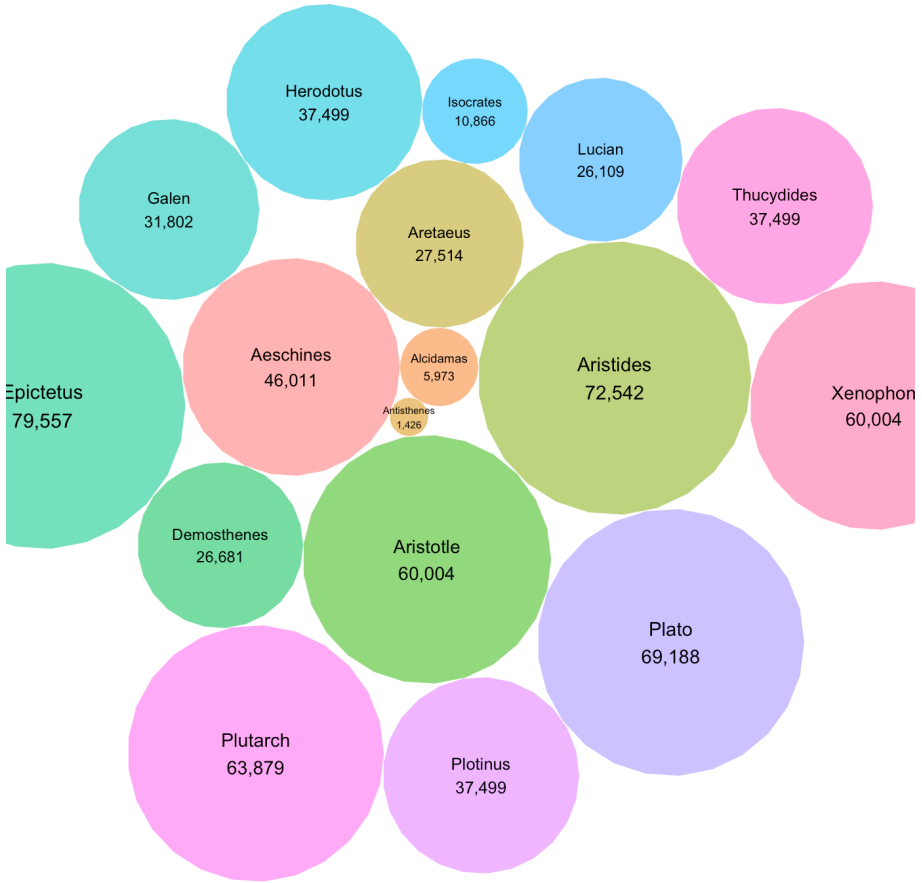
Задачи эксперимента и программные средства

1) выяснить, какие меры расстояния дают наибольшую точность на отрывках разной длины с использование разного числа переменных;
2) установить, обнаруживаются ли различия при использовании стандартизированных и нестандартизированных значений частотности (для тех методов, которые это допускают);
3) сравнить точность атрибуции при использовании словоформ или трехсложных энграм;
4) сделать выводы о том, на каких текстах классификатор чаще ошибается

Все вычисления выполнены в R, при этом токенизация (деления на слова или энграмы) выполнена с использованием библиотеки Stylo; для составления матриц расстояния или сходства привлекалась библиотека Philentropy, для других вычислений и визуализаций использовались Tidyverse и Tidymodels.

Корпус

- в корпус вошли сочинения древнегреческих прозаиков (историков, врачей, философов и ораторов) из библиотеки Perseus
- объем корпуса —694 тыс. слов
- корпус является несбалансированным в двух отношениях: разные авторы представлены разным количеством текстов, а сами эти тексты неравномерны по объему
- всего 57 текстов (включая выборки) и 17 авторов
- для анализа тексты были разделены на токены: словоформы и трехбуквенные энграмы (с сохранением диакритики)



Оценивание

500

500 слов — длина шага для испытания эффективности метода на отрывках 1000-7000 токенов (всего 13 шагов)

100

100 слов — длина шага для испытания эффективности метода с количеством предикторов (mfw) от 100 до 1000 (всего 10 шагов)

10

итераций проведено для каждой длины отрывка и mfw, т.о. для оценки каждого метода выполнено 57 * 10 * 13 * 10 = 74 100 классификаций

Средние показатели точности для всех методов

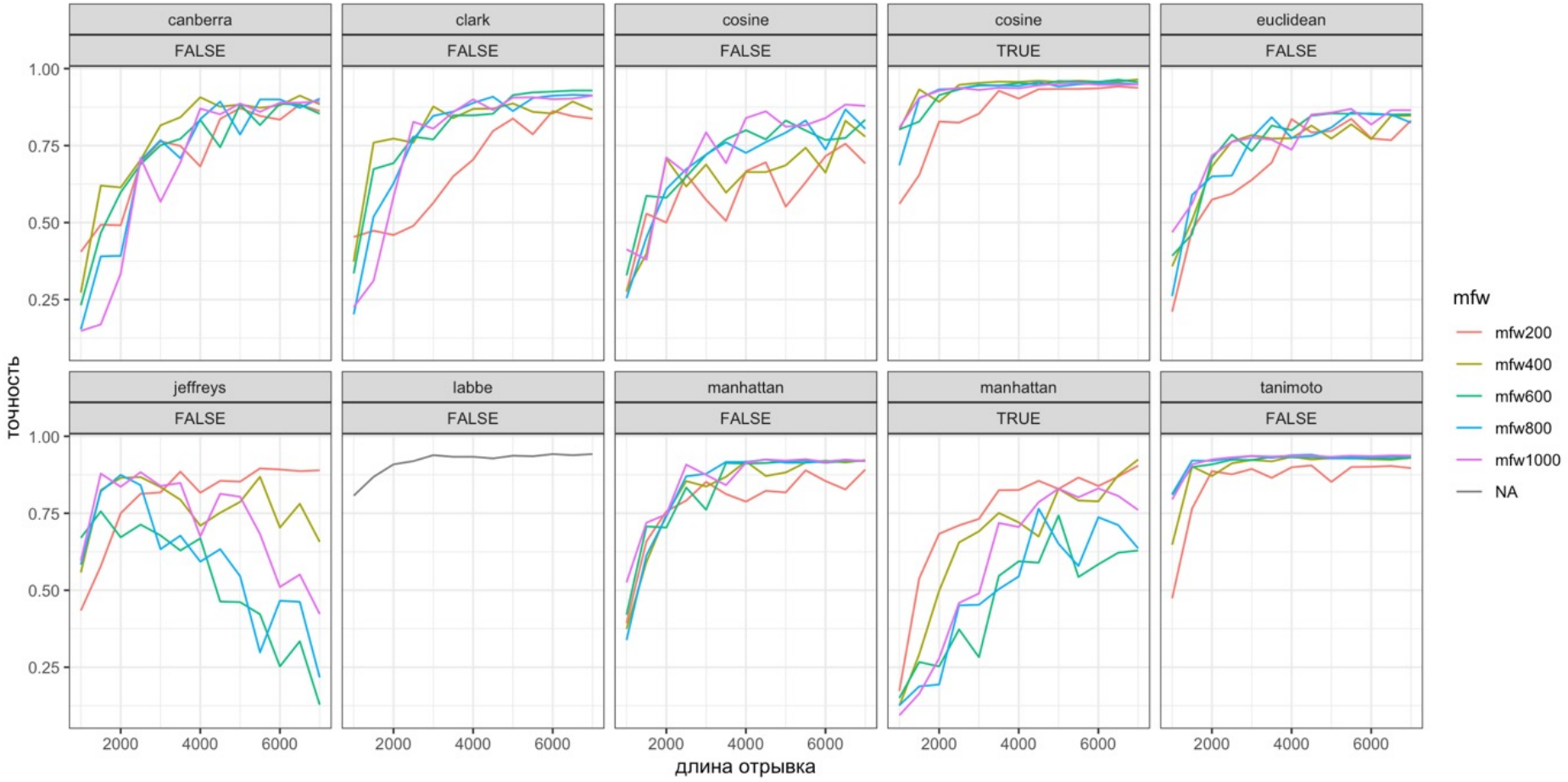
На всех отрывках и mfw наилучшие результаты показали:

- расстояние Лаббе (абсолютные значения частотности)
- косинусное сходство (стандартизированные значения)
- расстояние Танимото (относительная частотность без стандартизации)

NB: методы, давшие менее 0.6 процентов точности, исключены из дальнейшего рассмотрения.

метод	scale	точность
labbe	FALSE	0.918
cosine	TRUE	0.915
tanimoto	FALSE	0.901
manhattan	FALSE	0.821
clark	FALSE	0.765
euclidean	FALSE	0.732
canberra	FALSE	0.721
jeffreys	FALSE	0.68
cosine	FALSE	0.675
manhattan	TRUE	0.595
euclidean	TRUE	0.258
clark	TRUE	0.115
canberra	TRUE	0.053
tanimoto	TRUE	0.042

Результаты классификации



Наилучшие классификации (среднее значение для 10 итераций)

method	scale	mfw	size	.estimate
cosine	TRUE	500	6500	0.968
cosine	TRUE	300	7000	0.965
cosine	TRUE	400	7000	0.965
cosine	TRUE	500	7000	0.965
cosine	TRUE	600	5500	0.965
cosine	TRUE	300	4500	0.963
cosine	TRUE	400	6500	0.963
cosine	TRUE	300	5500	0.961
cosine	TRUE	300	6000	0.961
cosine	TRUE	400	3500	0.96
cosine	TRUE	400	4000	0.96
cosine	TRUE	400	5000	0.96
cosine	TRUE	400	5500	0.96
cosine	TRUE	500	5000	0.96
cosine	TRUE	600	5000	0.96
cosine	TRUE	600	6500	0.96
cosine	TRUE	200	6500	0.958

Точность на небольших отрывках

method	scale	mfw	size	.estimate
cosine	TRUE	700	2000	0.935
cosine	TRUE	400	1500	0.933
cosine	TRUE	400	2000	0.933
cosine	TRUE	300	1500	0.932
tanimoto	FALSE	800	1000	0.882
cosine	TRUE	400	1000	0.875

Наиболее «непредсказуемые» авторы: Аристотель, Демосфен и Аристид.

Итоги сравнения

- для сравнения трех векторов, хранящих данные о точности классификации на отрывках разной длины, был использован парный непараметрический тест Уилкоксона
- классификации с использованием mfw 100 и 200, показавшие наихудший результат, удалены
- результат сохранен в виде трех векторов, хранящих сведения о средней точности классификации по итогам десяти итераций на всех mfw
- тест показал отсутствие статистически значимой разницы между LAB и TAN, но на уровне значимости 0.05 отвергнута гипотеза об отсутствии различия между COS_S и LAB, а также TAN и COS

COS_S

size	.estimate
1000	0.774
1500	0.892
2000	0.917
2500	0.938
3000	0.944
3500	0.946
4000	0.948
4500	0.953
5000	0.952
5500	0.955
6000	0.954
6500	0.956
7000	0.955

LAB

size	.estimate
1000	0.807
1500	0.868
2000	0.909
2500	0.919
3000	0.939
3500	0.933
4000	0.933
4500	0.928
5000	0.937
5500	0.935
6000	0.942
6500	0.939
7000	0.942

TAN

size	.estimate
1000	0.766
1500	0.908
2000	0.906
2500	0.924
3000	0.929
3500	0.929
4000	0.935
4500	0.933
5000	0.93
5500	0.931
6000	0.93
6500	0.931
7000	0.933

Применение и вывод

Используя все три метода, мы попробовали классифицировать следующие тексты:

- послание «К Демонику» Пс.-Исократа,
- трактат «О воспитании детей» Пс.-Плутарха,
- а также два диалога из Corpus Platonicum («Пир» и «Феаг»).

Вывод:

- авторитет традиции действителен до тех пор, пока не опровергнут, в то время как подтвердить его статистически не представляется возможным.

Фридрих Бласс (1843–1907)