

Семинар №11

# Дельта Бёрроуза в атрибуции текста

НИУ ВШЭ – Санкт-Петербург

19.06.23

# Что такое дельта Бёрроуза

Дельта Бёрроуза — это мера стилистической близости между текстами.

Дельту используют во многих исследованиях, большая часть которых посвящена установлению авторства различных произведений.

Специалист по компьютерной лингвистике Джон Бёрроуз предложил дельту в 2002 году в своей статье «Дельта: мера стилистической разницы и проводник к определению авторства».

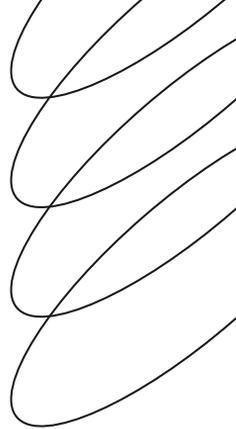
# Как считать дельту Бёрроуза

$m$  - кол-во текстов (обозначим их  $D_1, D_2, \dots, D_m$ );  
 $n$  - кол-во словоформ (обычно 200-500);  
 $w(i)$  - частота  $i$ -ой словоформы в  $D_j$ ;  
 $\mu_i$  — средняя частота  $w_i$  по корпусу;  
 $\sigma_i$  — среднеквадратичное отклонение.



$$\mu_i = \frac{1}{m} \sum_{k=1}^m f_i(D_k) \quad \sigma_i = \sqrt{\frac{1}{m} \sum_{k=1}^m (f_i(D_k) - \mu_i)^2}$$

$$z(f_i(D_j)) = \frac{f_i(D_j) - \mu_i}{\sigma_i}$$



# Как считать дельту Бёрроуза

Рассмотрим два текста, например, D1 и D2 и посчитаем дельту Бёрроуза:

$$\begin{aligned}\Delta(D_1, D_2) &= \frac{1}{n} \sum_{i=1}^n |z(f_i(D_1)) - z(f_i(D_2))| = \\ &= \frac{1}{n} \sum_{i=1}^n \left| \frac{f_i(D_1) - \mu_i}{\sigma_i} - \frac{f_i(D_2) - \mu_i}{\sigma_i} \right| = \frac{1}{n} \sum_{i=1}^n \frac{|f_i(D_1) - f_i(D_2)|}{\sigma_i}\end{aligned}$$

# Как считать дельту Бёрроуза

Чем меньше  $\Delta(D1, D2)$ , тем ближе по стилю тексты  $D1$  и  $D2$ . А посчитав дельты для всех пар текстов, можно их классифицировать.

Обычно для этого используют один из алгоритмов кластеризации, который объединяет в группы близкие друг к другу тексты. Например, тексты могут быть сгруппированы по авторам, жанрам или периоду написания.

# Суть метрики

	a	b	c	d	e	f	g	h	
8	6	5	4	3	2	3	4	5	8
7	5	4	3	2	1	2	3	4	7
6	4	3	2	1		1	2	3	6
5	5	4	3	2	1	2	3	4	5
4	6	5	4	3	2	3	4	5	4
3	7	6	5	4	3	4	5	6	3
2	8	7	6	5	4	5	6	7	2
1	9	8	7	6	5	6	7	8	1
	a	b	c	d	e	f	g	h	

# Вклад Аргамона (2008)

Результаты вероятностного ранжирования элементов на основе Манхэттенского расстояния будут соответствовать распределению Лапласа (Argamon, 2008).



Стандартизация относительных частот через z-score в классической формуле некорректна и требует переосмысления.

Argamon, Shlomo. (2007). Interpreting Burrows's Delta: Geometric and Probabilistic Foundations. *Literary and linguistic computing*. 23. 131-147.

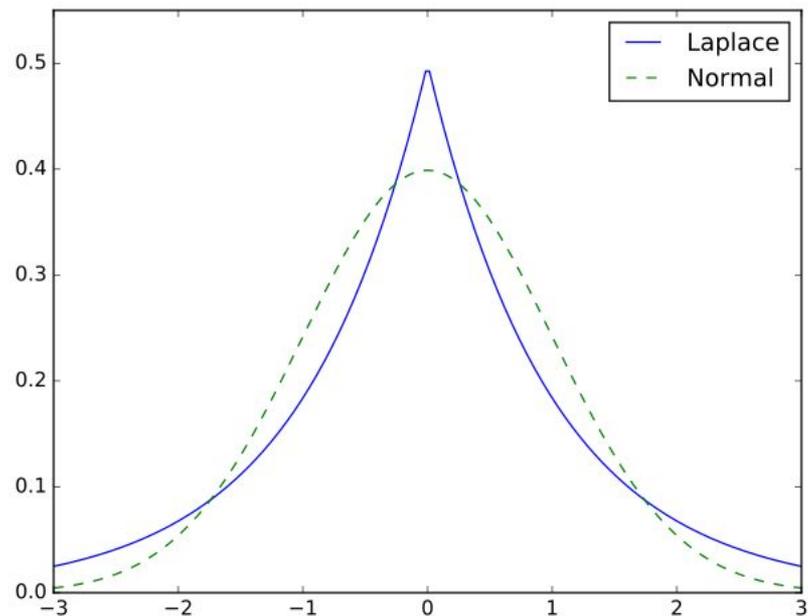


Рис. 1 – Сравнение распределения Лапласа с распределением Гаусса

# «Линейная» дельта

- Вместо z-score используем параметры распределения Лапласа – медиану ( $a_i$ ) и среднее абсолютное отклонение от медианы ( $b_i$ )

$$\longrightarrow \Delta_{L\perp}^{(n)}(D, D') = \sum_{i=1}^n \frac{1}{b_i} |f_i(D) - f_i(D')|$$

# «Квадратичная» дельта

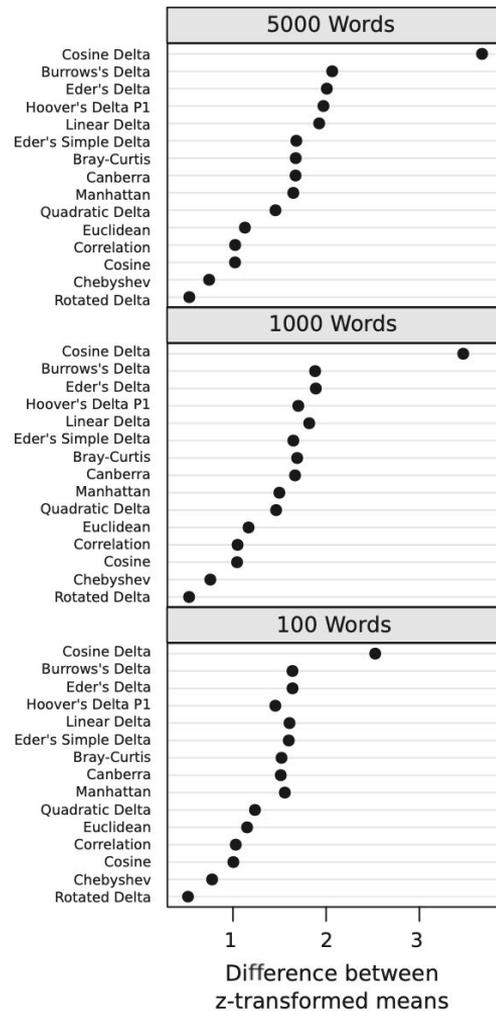
- Вероятностное ранжирование элементов по Евклидову расстоянию = ранжированию по наивысшей вероятности в многомерном Гауссовом распределении. Отсюда возможность использовать z-score.

$$\longrightarrow \Delta_{Q\perp}^{(n)}(D, D') = \sum_i \frac{1}{\sigma_i^2} (f_i(D) - f_i(D'))^2$$

# Performance

- Несмотря на теоретические аргументы, квадратичная дельта менее эффективна в сравнении с дельтой Бёрроуза 😞
- Эффективность линейной дельты близка к той, что характерна для дельты Бёрроуза.
- Лучшей же модификацией (по эффективности) является косинусная дельта.

Evert, Stefan & Proisl, & Jannidis, Fotis & Reger, Isabella & Pielström, Steffen & Schöch, Christof & Vitt, Thorsten. (2017). Understanding and explaining Delta measures for authorship attribution. Digital Scholarship in the Humanities. 32. 4-16.



# Косинусная дельта

Вектор z-оценок вектора слов, связанного с произвольным документом  $D$ , обозначается как:

$$z(D) = \left( \frac{f_1(D) - \mu_1}{\sigma_1}, \frac{f_2(D) - \mu_2}{\sigma_2}, \dots, \frac{f_d(D) - \mu_d}{\sigma_d} \right)$$

В то время как косинусная дельта (дельта Олдриджа и Смита):

$$\Delta_z(D, D_{\text{Author}}) = \frac{\langle z(D), z(D_{\text{Author}}) \rangle}{\|z(D)\| \cdot \|z(D_{\text{Author}})\|}$$

# References

- Argamon, S. (2007). Interpreting Burrows's Delta: Geometric and Probabilistic Foundations. *Literary and Linguistic Computing*, 23, 131-147. <https://doi.org/10.1093/llc/fqn003>
- Evert, S., Proisl, Jannidis, F., Reger, I., Pielström, S., Schöch, C., & Vitt, T. (2017). Understanding and explaining Delta measures for authorship attribution. *Digital Scholarship in the Humanities*, 32, 4-16. <https://doi.org/10.1093/llc/fqx023>
- Smith, P., & Aldridge, W. (2011). Improving Authorship Attribution: Optimizing Burrows' Delta Method\*. *Journal of Quantitative Linguistics*, 18, 63-88. <https://doi.org/10.1080/09296174.2011.533591>
- <https://sysblok.ru/glossary/a-avtor-kto-rassudit-delta-bjorrouza/>
- <https://iq.hse.ru/news/367813734.html>