Семинар №7

Атрибуция: определение авторства текста

НИУ ВШЭ – Санкт-Петербург

05.04.23

Авторство как полемика

Шекспировский и Гомеровский вопросы:

Действительно ли корпус произведений, предписываемый Уильяму Шекспиру, принадлежит ему с точки зрения авторства?

Являются ли эпические поэмы Гомера продуктом индивидуального труда или же это порождение коллективного народного творчества?

Why bother?

– Общее развитие этики и науки
– Внесение исторической ясности
– Улучшение результатов литературоведческого анализа

Формализация

Атрибуция – совокупность процедур и методов, нацеленных на установление авторства (текста, музыки и т.д.)

Классификация:

– Экспертная атрибуция
– Формальная атрибуция

Сферы применения:

– Криминология
– Лингвистика
– История

Экспертная атрибуция

Автороведческая экспертиза:

– Базируется на предположении о том, что каждый автор обладает уникальными особенностями использования языка

– Эксперт использует пятиуровневый подход к анализу текста, также принимая во внимание личностные характеристики автора

Трудности:

– Дорого
– Общая субъективность
– Трудоемко (очень и очень трудоемко)

Формальная атрибуция

Стилометрия – метод определения авторства текста, основанный на распределении частотностей слов.

Исторический контекст:

– Эмпирические основы заложены в начале 20 века Морозовым Н.А. [1]

– В 2002 году случился поворотный момент, когда Джон Бёрроуз вывел первый универсальный инструмент атрибуции текста [2]

✨ Метод Дельты ✨

Запись семинара на YouTube – https://www.youtube.com/watch?v=KjqHNpFAGxo

Метод дельты Бёрроуза

1. Формируем выборку из 200-500 самых частотных слов в документах,
2. Используем преобразованные частоты в формуле:

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

$z(x_i)$ – стандартизированная частота определенного слова в документе A

$z(y_i)$ – стандартизированная частота того же слова, но в документе B

Пример

1. Каждый документ это вектор частотностей:

	Документ 1	Документ 2	Документ 3	...
the	4.52	3.54	4.41	
and	2.26	2.63	2.34	
to	2.20	2.62	2.33	
of	2.17	2.10	2.18	
...				

NB: значения стандартизованы

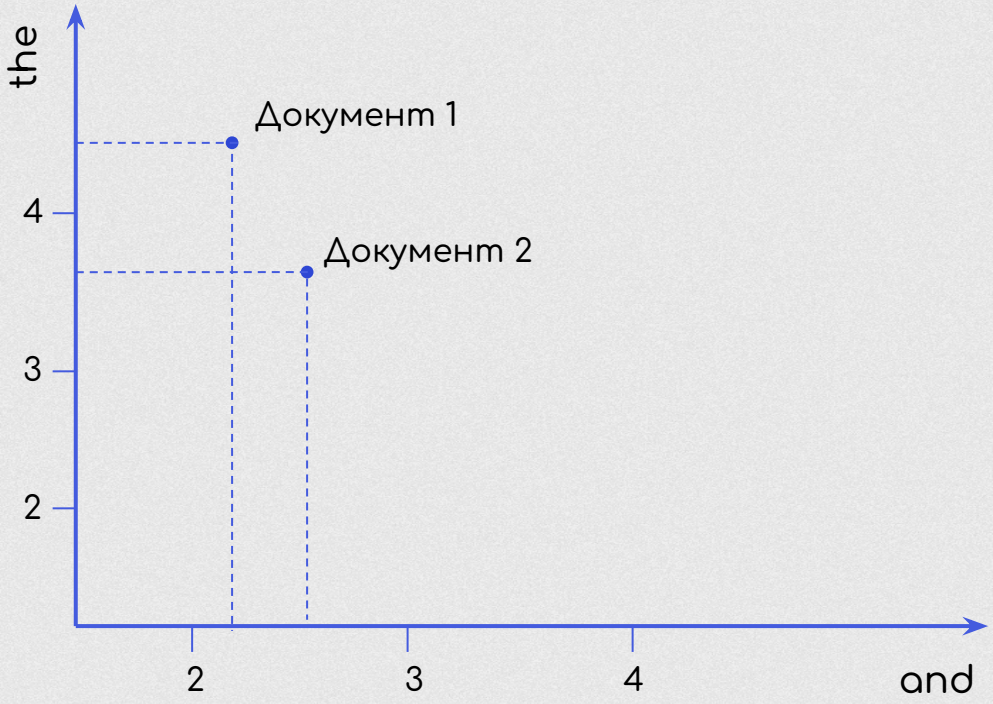
Пример

2. Представим выделенные значения в декартовом пространстве

	Документ 1	Документ 2	Документ 3	...
the	4.52	3.54	4.41	
and	2.26	2.63	2.34	
to	2.20	2.62	2.33	
of	2.17	2.10	2.18	
...				

NB: значения стандартизованы

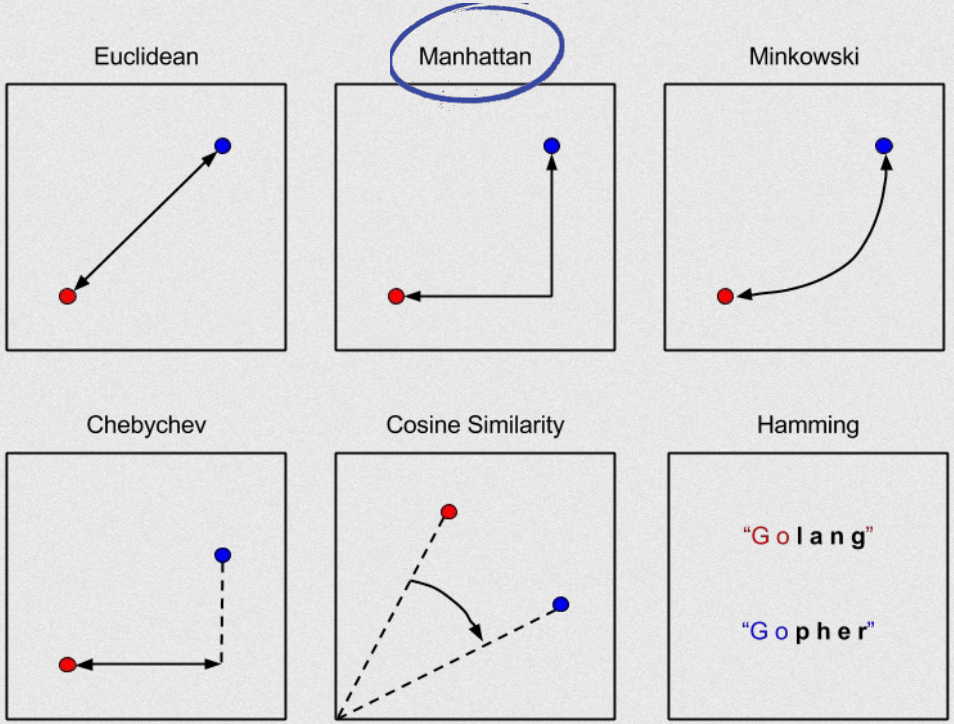
Пример



3. Рассчитаем степень близости между двумя документами. Как?

Ответ: с помощью расстояния

Где-то я уже это видел...

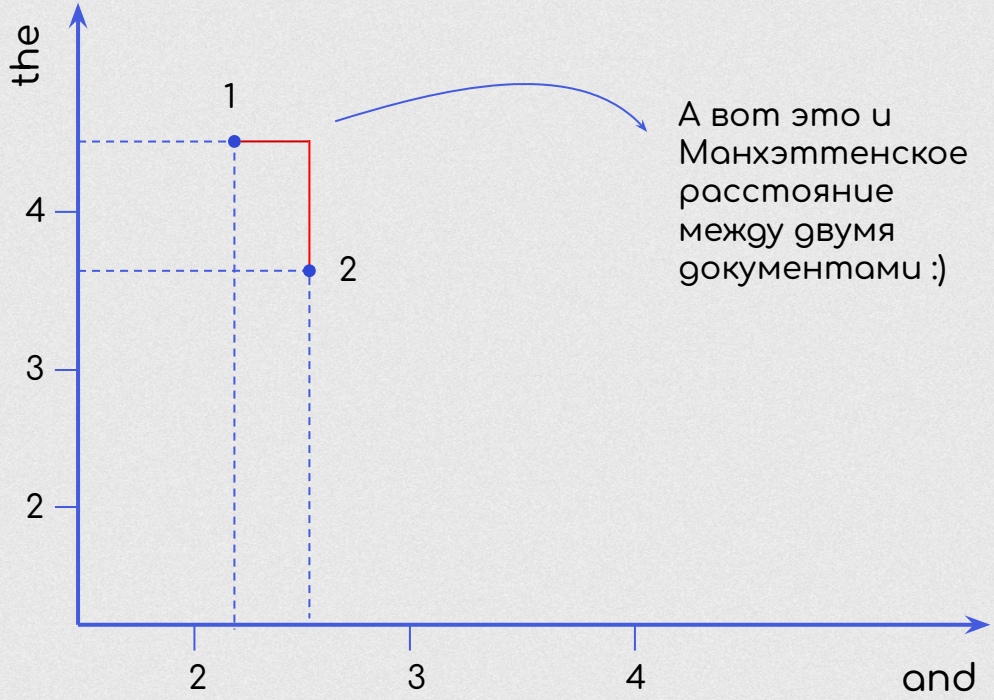


– Дашь списать домашку?
– Да, только не точь-в-точь, чтобы не спалили
– Ок:

$$distance = \sum_{1}^{n} |p_i - q_i|$$

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

Пример


the
1
4
2
3
2
2
3
4
and
А вот это и Манхэттенское расстояние между двумя документами :)

Интерпретация и ограничения

Rule of thumb:

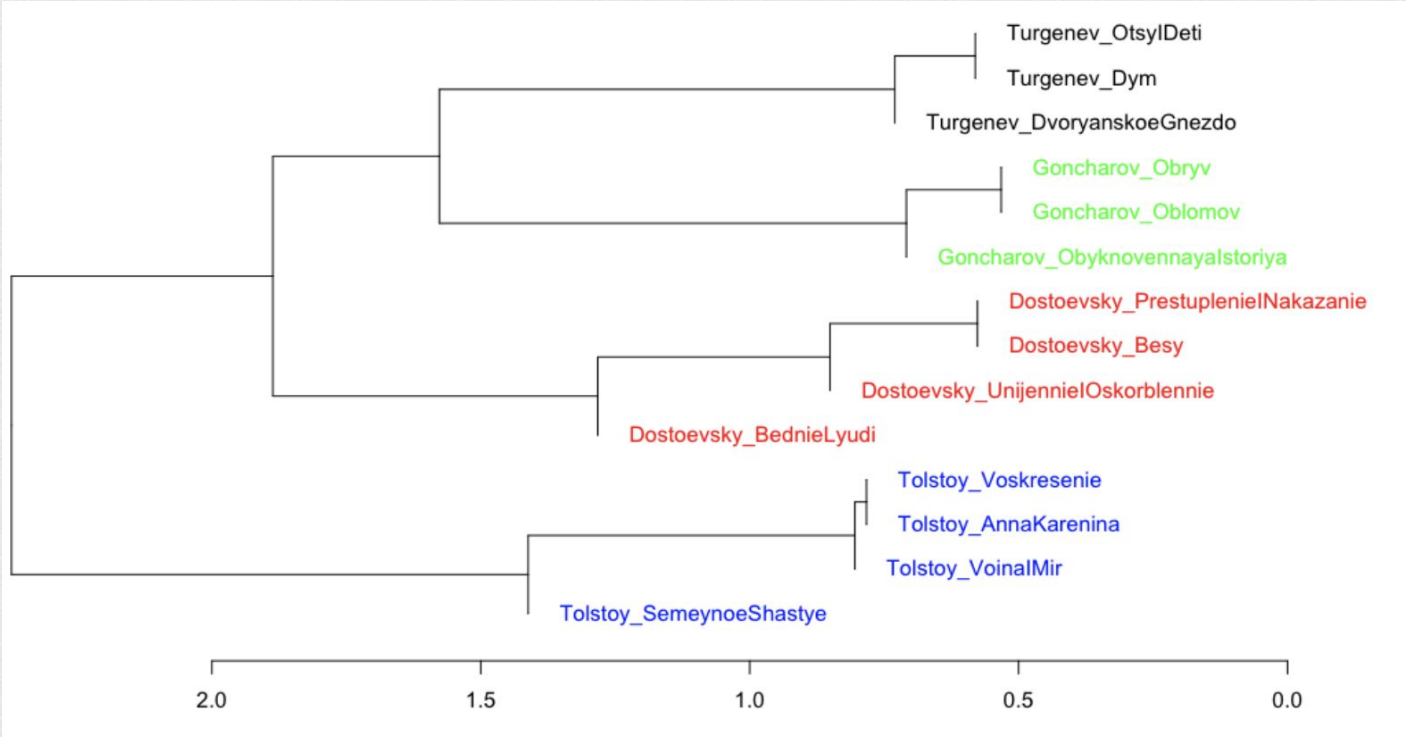
– Чем меньше значение дельты между двумя текстами, тем больше шансов того, что они были написаны одним и тем же человеком.

Где считать дельту?

– Пакет *Stylo* в R
– Пакет *faststylometry* в Python

Ограничения:

– Метод хорошо работает только на **больших текстах** (5-10 тысяч слов)
– Метод хорошо работает только с произведениями **одного жанра**
– Плохое выделение текстов, написанных в соавторстве



Source: Системный блокъ, 2021 [3]

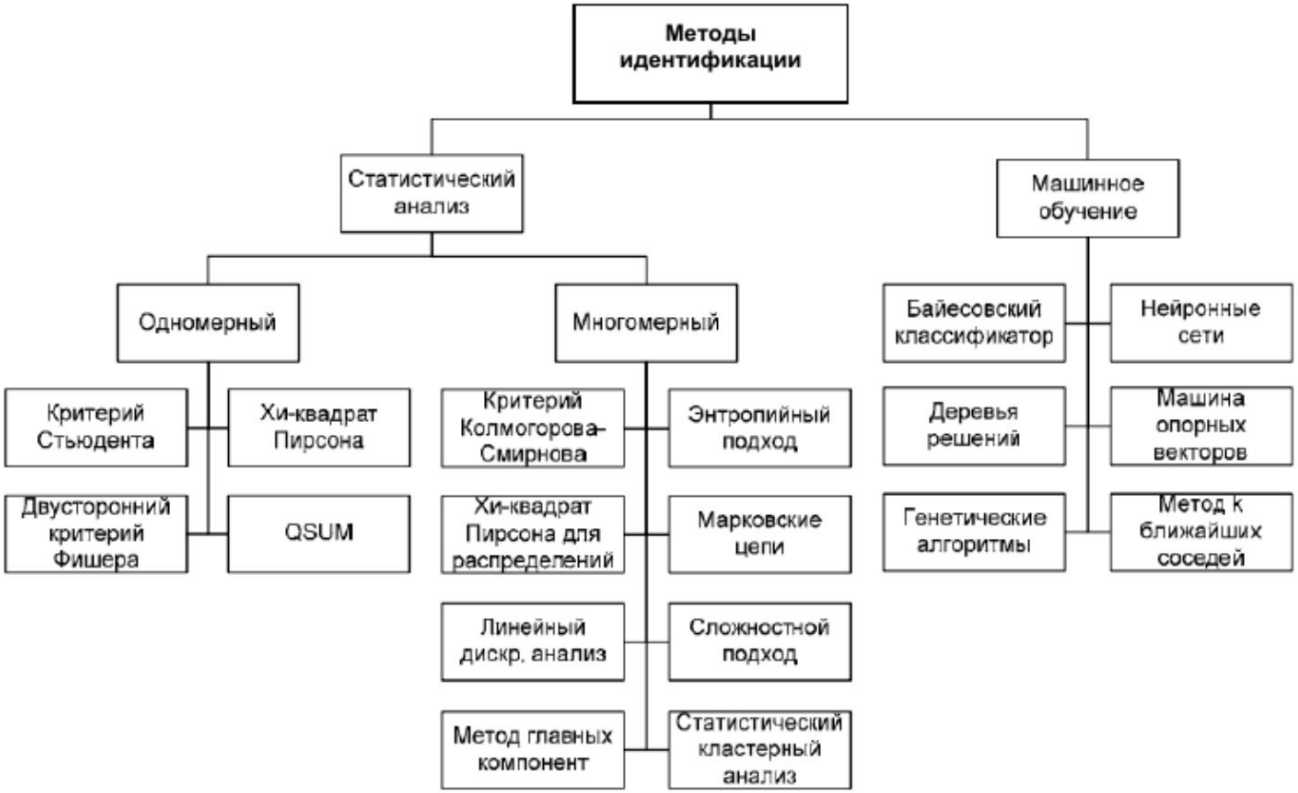
Авторский инвариант

Используют ли люди что-то кроме относительной частоты самого частотного слова? Да!

– Среднее число слогов в тексте
– Процент содержания одной (одного) частицы / союза / предлога
– Процент содержания служебных частей речи
– Среднее число употреблений существительных / прилагательных и т.д.

Как правило, авторский инвариант используется с **классическими статистическими тестами** и **методами машинного обучения**.

Как обычно, предполагается, что есть **эталонный текст**, с которым производится сравнения для установления авторства.



Source: Батура, 2012 [4]

References

[1] Морозов, Н. А. (n.d.). *Лингвистические спектры, как средство для отличения плагиатов от истинных произведений того или другого известного автора и для определения их эпохи*. https://www.textology.ru/library/book.aspx?bookId=1&textId=3#S1

[2] Burrows, J. (2002). 'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing*, *17*(3), 267–287. https://doi.org/10.1093/llc/17.3.267

[3] Скоринкин, А. З., Даниил. (2021, February 26). Что такое стилометрия и как измерить стилистические особенности текста. *Системный Блокъ*. https://sysblok.ru/knowhow/stilometrija-kak-v-raznoe-vremja-ljudi-iskali-avtorov-tekstov/

[4] Батура, Т. В. (2012). Формальные методы определения авторства текстов. *Вестник Новосибирского Государственного Университета. Серия: Информационные Технологии*, *10*(4).