Контрастивный анализ

Квантитативный анализ текста

Кирилл Александрович Маслинский

21.02.2022 / 03

ЕУ СПб

Ключевые слова

План

Метод контрастного корпуса

Задача — извлечение лексики, характерной для данного корпуса

- Контрастный корпус (reference corpus) — отражает словоупотребление в языке вообще или в более широкой предметной области
- Составить частотные списки слов для изучаемого и контрастного корпуса
- Отсортировать слова по расхождению частотности с ожидаемой на основании контрастного корпуса
- Ключевые слова изучаемого корпуса — наверху списка

Ключевые слова корпуса

Simple maths (by Adam Kilgarriff)

«это слово встречается в этом корпусе вдвое чаще, чем в том»

- Самый простой подход
  - Нормализовать частотности
    - употреблений на тысячу или употреблений на миллион (IPM)
  - Вычислить отношение нормализованных частотностей
  - Отсортировать список слов по значению отношения

Для примера:

- Два корпуса по миллиону токенов
- Нормализовать частотности не нужно

fc focus corpus — изучаемый корпус

rc reference corpus — контрастный корпус

Проблема 1: нельзя делить на 0

слово	fc	rc	отношение
редкость	10	0	?
помешивать	100	0	?
вкуснотища	1000	0	?

Стандартное решение: прибавить 1:

слово	fc	rc	отношение
редкость	11	1	11
помешивать	101	1	101
вкуснотища	1001	1	1001

Проблема 2: из-за редких слов слишком много больших отношений

Частотность тоже важна. Решение: прибавить n.

- $n = 1$

слово	fc	rc	fc+n	rc+n	отношение
изредка	10	0	11	1	11,00
временами	200	100	201	101	1,99
часто	12000	10000	12001	10001	1,20

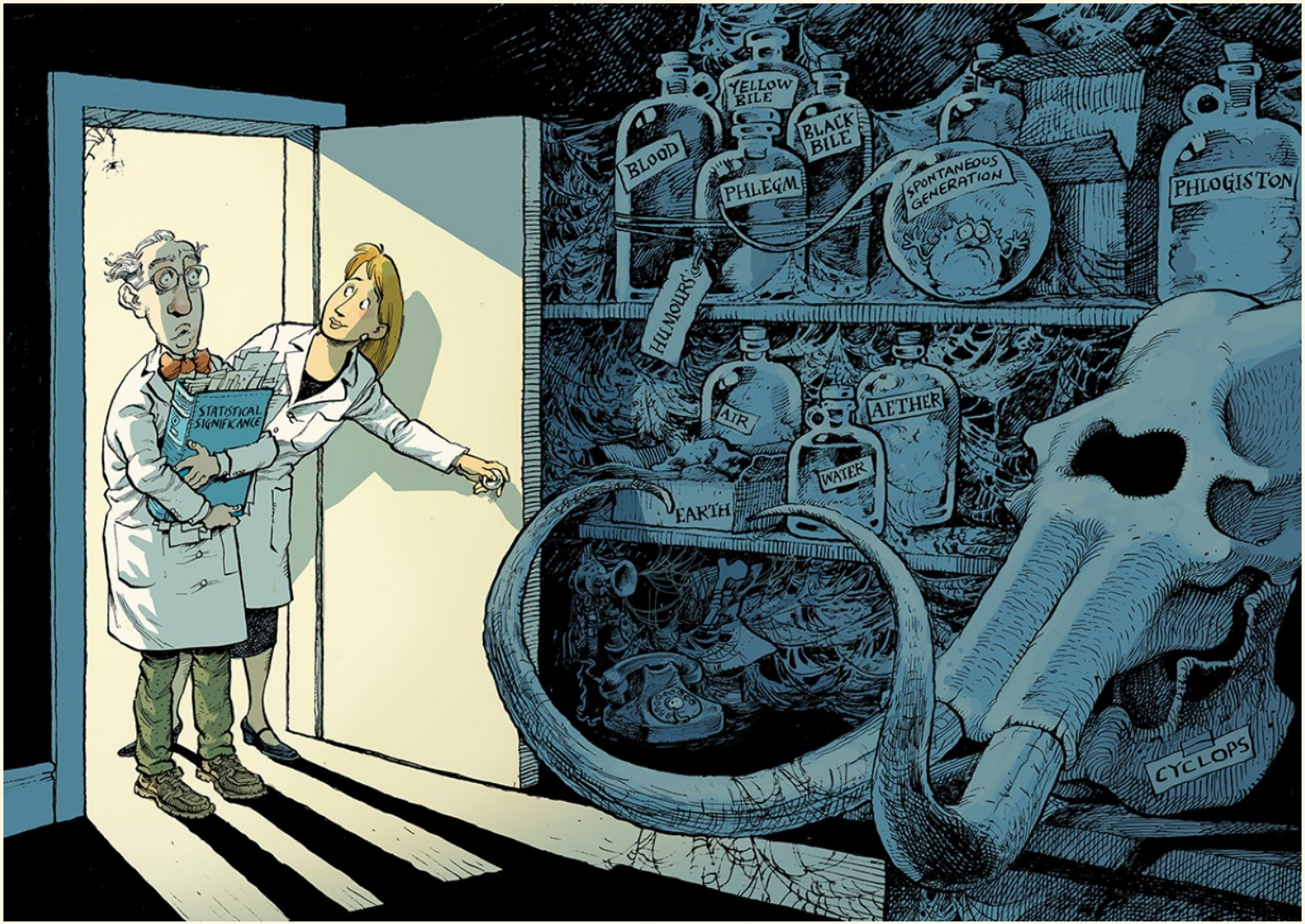
- $n = 100$

слово	fc	rc	fc+n	rc+n	отношение
временами	200	100	300	200	1,50
изредка	10	0	110	100	1,10
часто	12000	10000	12100	10100	1,20

План

- Ключевые слова
- Использование контрастного корпуса
- Отношение правдоподобия
- Взаимная информация
- Байесовский подход

Уходящая эпоха статистической значимости



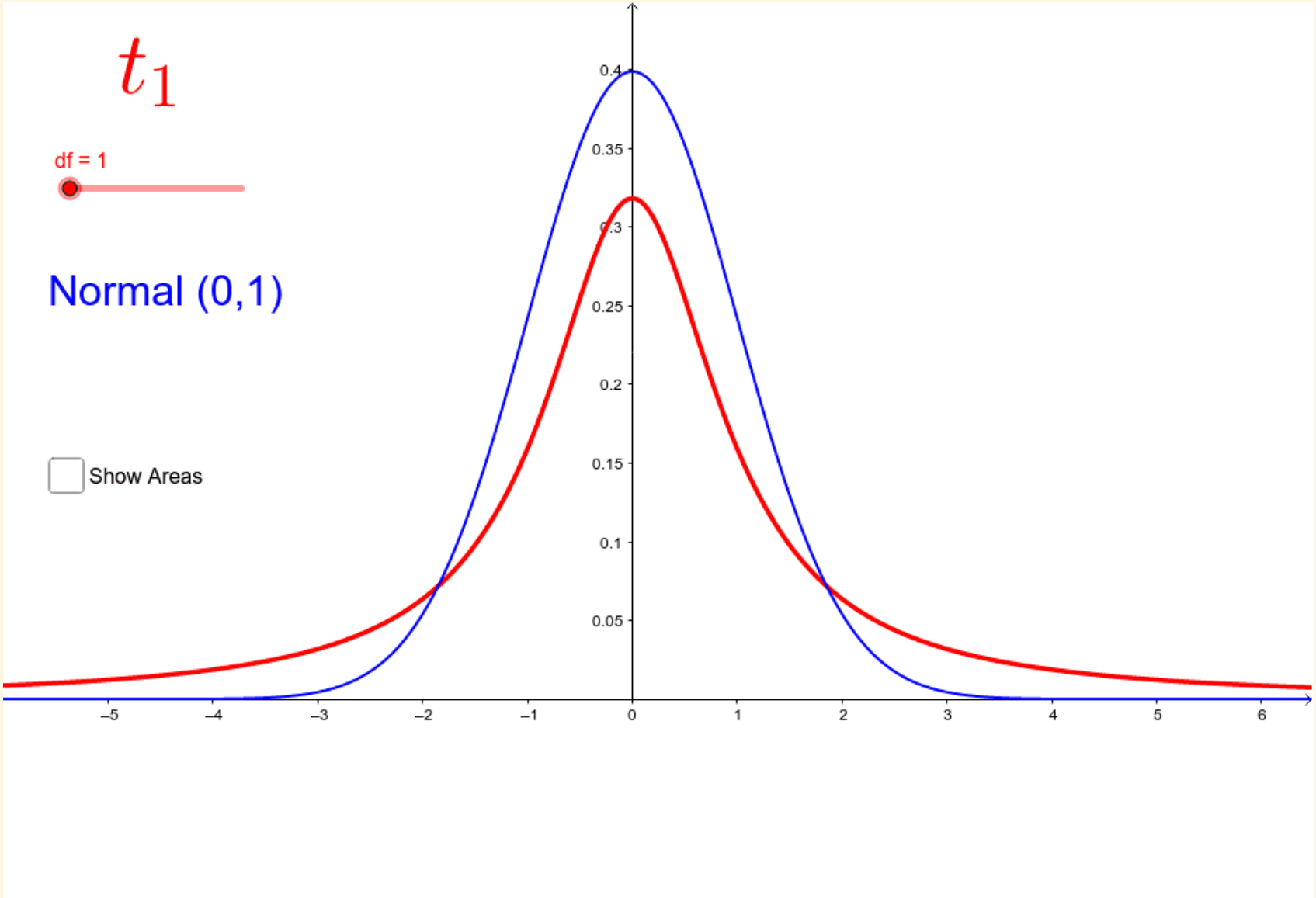
Amrhein et al. “Scientists rise up against statistical significance” (2019)

Нормальность и распределение слов

В парадигме стандартных статистических тестов при сравнении частотностей слов возникали проблемы:

- Предположение о нормальности неверно в случае частотного распределения слов
- В языке слишком много редких событий
- Неприменимость тестов, основанных на предположении о нормальности (напр., хи-кавдрат), как минимум к редким событиям (частотность < 5)

Нормальное распределение излишне удивлено



Отношение правдоподобия: мотивировка

Способ включить частотности слов в парадигму статистических тестов:

Ted Dunning “Accurate Methods for the Statistics of Surprise and Coincidence” (1994)

- Отношение правдоподобия менее зависит от предположения о нормальности распределения данных
- Поэтому не так резко завышает значимость редких событий и может применяться для оценки различий не только самых частотных слов

Отношение правдоподобия: формула

	Корпус 1	Корпус 2	Всего
Частотность слова	a	b	a+b
Частотность остальных слов	c-a	d-b	c+d-a-b
Всего	c	d	c+d

Ожидаемые частотности:

E1 $\frac{c}{c+d}(a+b)$
E2 $\frac{d}{c+d}(a+b)$

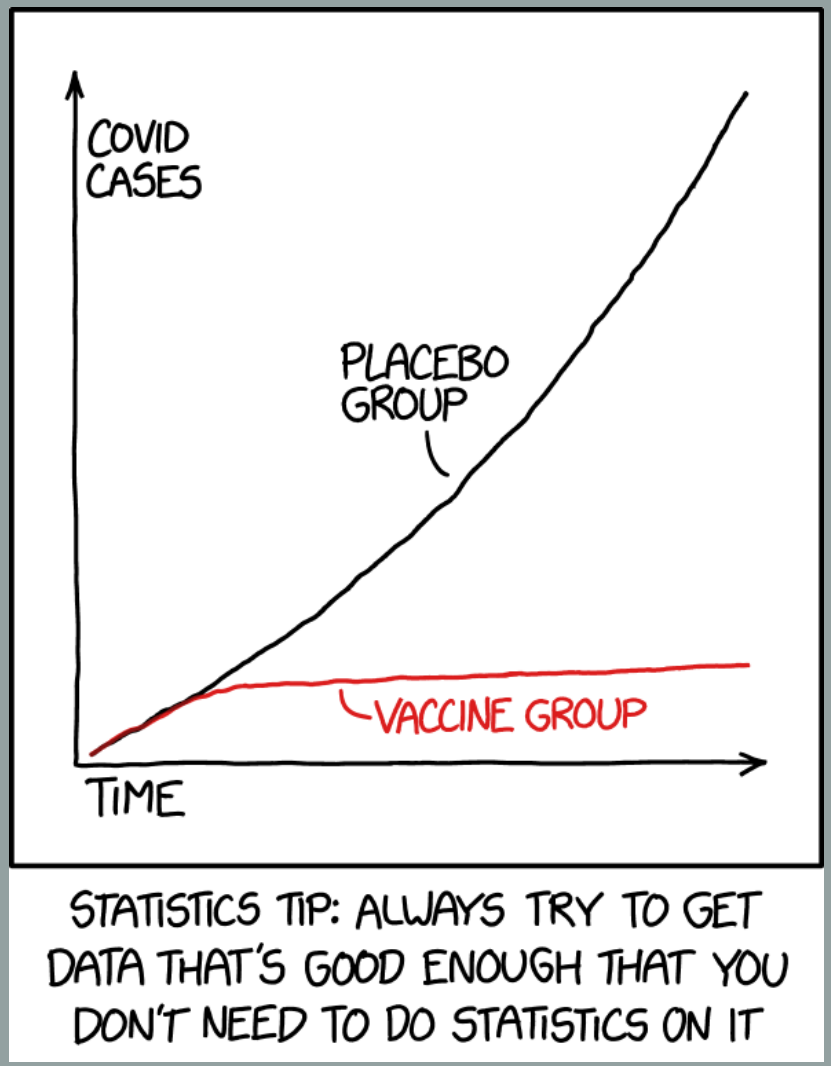
$$LL = G^2 = 2(a \log(a/E1) + b \log(b/E2)) \quad (1)$$

Проблема Log-Likelihood

- Более чувствителен к частотным событиям (словам), чем к менее частотным [занижает степень различия по менее частотным словам]

Эксперимент на царях и цветочках

Практический вывод



План

- Ключевые слова
  - Использование контрастного корпуса
  - Отношение правдоподобия
  - Взаимная информация
  - Байесовский подход

Условная вероятность

$$P(Event|Condition)$$

Условная вероятность

$$P(B|A) = \frac{P(B \wedge A)}{P(A)} \tag{2}$$

$$P(\text{интеллект}|\text{искусственный}) = \frac{P(\text{искусственный интеллект})}{P(\text{искусственный})} =$$

$$= \frac{\frac{4}{46804371}}{\frac{121}{46804371}} = \frac{4}{121} = 0.033$$

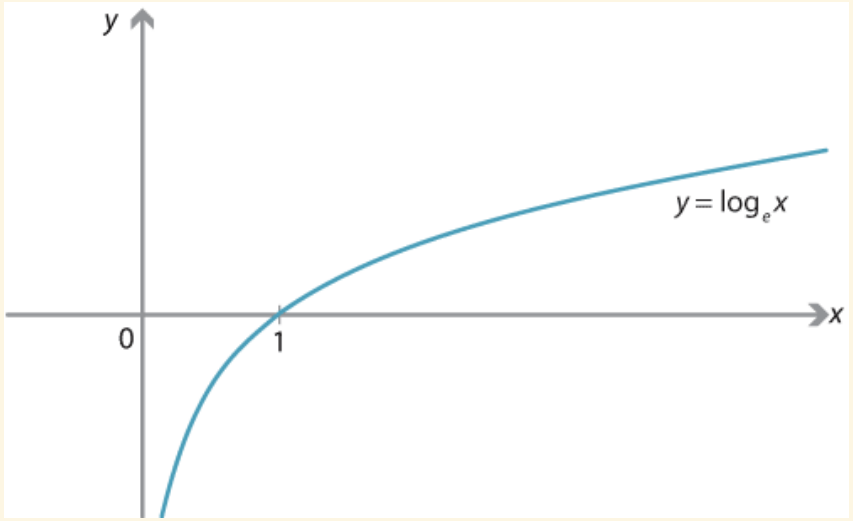
Pointwise mutual information

PMI

$$pmi(x;y) = log\frac{p(x,y)}{p(x)p(y)} =$$

$$= log\frac{p(x|y)}{p(x)} =$$

$$= log\frac{p(y|x)}{p(y)}$$



Positive PMI

$$ppmi(x;y) = max(pmi(x;y), 0)$$

Проблема PMI

- Очень чувствителен к редким, но высоко информативным относительно друг друга событиям (слова встречаются всегда вместе) [завышает степень различия по редким словам]

План

- Ключевые слова
  - Использование контрастного корпуса
  - Отношение правдоподобия
  - Взаимная информация
  - Байесовский подход

Вероятность независимых событий

Независимые события — наступление одного не изменяет вероятности другого.

$$P(B|A) = P(B) \quad |P(B) > 0 \tag{3}$$

$$P(B \cap A) = P(A) \cdot P(B) \tag{4}$$

Совместная вероятность

Для независимых A и B:

$p(\text{A и B}) = p(A)p(B) \quad | \; p(B|A) = p(B)$

В общем случае:

$p(\text{A и B}) = p(A)p(B|A) \quad | \; p(B|A) \neq p(B)$

Совместная вероятность

Для независимых A и B:

$p(\text{A и B}) = p(A)p(B) \quad | \; p(B|A) = p(B)$

В общем случае:

$p(\text{A и B}) = p(A)p(B|A) \quad | \; p(B|A) \neq p(B)$

Правило Байеса

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{5}$$

$$posterior = \frac{likelihood \cdot prior}{evidence} \tag{6}$$

Правило Байеса

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{5}$$

$$posterior = \frac{likelihood \cdot prior}{evidence} \tag{6}$$

1. Log odds ratio:

$$O_1 = \frac{f_{(w,c1)}}{N_{c1} - f_{(w,c1)}}$$

$$O_2 = \frac{f_{(w,c2)}}{N_{c2} - f_{(w,c2)}}$$

$$LO = log \frac{O_1}{O_2}$$

2. Weighted by uninformative Dirichlet prior:

$$\delta = \frac{\frac{f_{(w,c1)} + \alpha_{(w,c1)}}{N_{c1} + \alpha_{c1} - f_{(w,c1)} - \alpha_{(w,c1)}}}{\frac{f_{(w,c2)} + \alpha_{(w,c2)}}{N_{c2} + \alpha_{c2} - f_{(w,c2)} - \alpha(w,c2)}}$$