Математические модели в лингвистике

7. Лингвистические корпуса

Мати Пентус, Александр Пиперски,
Алексей Сорокин

МГУ, межфакультетский курс,
осенний семестр 2017–2018 учебного года

Лингвистические корпуса

- Большие собрания текстов для лингвистических исследований и автоматической обработки текстов
- Лингвистический корпус =
  = тексты + разметка (+ поиск)

Виды разметки:

- Метаразметка (информация о тексте, авторе и т. д.)
- П

Text Encoding Initiative (TEI)

- Стандарт представления текстовой информации различных типов
- От стихов до нот и лингвистических корпусов
- http://www.tei-c.org/release/doc/tei-p5-doc/en/html/

```
<text>
<body>
 <lg>
  <lg type="quatrain">
   <l>My Mistres eyes are nothing like the Sunne,</l>
   <l>Currall is farre more red, then her lips red</l>
   <l>If snow be white, why then her brests are dun:</l>
   <l>If haires be wiers, black wiers grown on her head:</l>
  </lg>
  <lg type="quatrain">
   <l>I have seene Roses damaskt, red and white,</l>
   <l>But no such Roses see I in her cheekes,</l>
   <l>And in some perfumes is there more delight,</l>
   <l>Then in the breath that from my Mistres reekes.</l>
  </lg>
```

```
 <lg type="quatrain">
  <l>I love to heare her speake, yet well I know,</l>
  <l>That Musicke hath a farre more pleasing sound:</l>
  <l>I graunt I never saw a goddesse goe,</l>
  <l>My Mistres when shee walkes treads on the ground.</l>
 </lg>
</lg>
<lg type="couplet">
 <l>And yet by heaven I think my love as rare,</l>
 <l>As any she beli'd with false compare.</l>
</lg>
</body>
</text>
```

Лингвистическая разметка

- Токенизация
- Лемматизация
- Морфологическая разметка
- Синтаксическая разметка

Токенизация

- Токенизация — разбиение текста на единицы (≈ слова) для дальнейшей обработки
- Считать ли токенами знаки препинания?
- Как обрабатывать числа, сокращения, неоднословные названия и т. п.?

Лемматизация и морфологическая разметка

- Лемматизация — приведение слова к начальной форме
- Морфологическая разметка — присвоение слову морфологических помет
- *слову* — *слово*, сущ., ср., дат. ед.

Идеал и реальность

Текст	*три*	*мушкетёра*
Лемма	*три*	*мушкетёр*
Пометы	числ., им.	сущ., м., род. ед.
Лемма	*три*	*мушкетёр*
Пометы	числ., им.	сущ., м., род. ед.
	числ., вин.	сущ., м., вин. ед.
Лемма	*тереть*	
Пометы	гл., пов.	

Синтаксическая разметка

- Чаще всего — синтаксис зависимостей
- *Мама мыла белую раму*
- *мыла* $\longrightarrow$ *Мама*, *мыла* $\longrightarrow$ *раму*, *раму* $\longrightarrow$ *белую* (+ типы отношений)
- На практике автоматический синтаксический разбор пока далёк от идеала

Корпуса русского языка

- Национальный корпус русского языка (http://www.ruscorpora.ru)
- ruTenTen в системе SketchEngine (http://the.sketchengine.co.uk)
- ...

Национальный корпус русского языка

- Самый популярный русский корпус
- **Основной подкорпус** — 283 млн словоформ
- Вручную отобранные тексты
- Морфологическая омонимия вручную снята в 2% текстов
- Синтаксической разметки нет

Род слова *кофе*

- Русский орфографический словарь (2007): кофе, *нескл.*, *м.* и (*разг.*) *с.*
- Посмотрим род слова *кофе* в реальных текстах
- Как искать?

Род слова *кофе*

- Выпил двойное вкусное кофе, дав «на чай» его двойную стоимость [Ю. М. Нагибин. Дневник (1982)]
- Весь мир проходил мимо, и мир этот можно было рассматривать, спокойно размешивая в стакане двадцатисентовое кофе с молоком. [Андрей Седых. Далекие, близкие. Воспоминания (1979)]
- Не помню, сколько мы заплатили за удобный ночлег, вкусный ужин и утреннее кофе. [Н. О. Лосский. Воспоминания: жизнь и философский

Род слова *кофе*

- ... я пил утреннее кофе в молочном баре ... [В. В. Набоков. Лолита (1967)]
- Я тоже провел это время в Париже: <...> поддельное, но все же ароматное кофе... [Ю. П. Анненков. Дневник моих встреч (1966)]
- ... настоящее кофе со сливками можно пить только у Либмана [Дон Аминадо. Поезд на третьем пути (1954)]
- ... жареный кролик с зел. бобами, компот (с сахаром) из апельсинов, хорошее кофе — давно так не ел! [И. А. Бунин. Дневники (1940-1953)]

только и знать/делать

- *только и знать/делать что X*
- Что можно изучать про эти конструкции в современном русском языке?
- Рассмотрим их употребление в текстах с 1980-го года и далее

1. Денис Горелов. Москва кирзам верит. «Молодые». Режиссер Николай Москаленко. Год 1971. (2002) // «Известия», 2002.07.14 [омонимия снята] Все примеры (1)

   Крестьяне 70-х любили уесть городских, что те про деревню **только и знают**, **что** спереди у коровы рога, а сзади вымя. [Денис Горелов. Москва кирзам верит. «Молодые». Режиссер Николай Москаленко. Год 1971. (2002) // «Известия», 2002.07.14] [омонимия снята] ←…→

2. Ольга Андреева, Григорий Тарасевич. Отличники с Манежной // «Русский репортер», № 3 (181), 27 января 2011, 2011 [омонимия не снята] Все примеры (1)

   А мама строгая, **только и знает**, **что** за трояки ругать. [Ольга Андреева, Григорий Тарасевич. Отличники с Манежной // «Русский репортер», № 3 (181), 27 января 2011, 2011] [омонимия не снята] ←…→

3. А. А. Ганиева. Вечер превращается в ночь (2010) [омонимия не снята] Все примеры (1)

   Эти суфии **только и знают**, **что** свою чIанду Пророку приписывать, — сказал Мага и, быстро подтянувшись несколько раз, спрыгнул на землю. [А. А. Ганиева. Вечер превращается в ночь (2010)] [омонимия не снята] ←…→

4. Анна Николаева. Человек-снежинка // «Наука и жизнь», 2009 [омонимия не снята] Все примеры (1)

   Сверстники Уилла **только и знали**, **что** играть в снежки или кататься на коньках. [Анна Николаева. Человек-снежинка // «Наука и жизнь», 2009] [омонимия не снята] ←…→

5. Олег Зайончковский. Счастье возможно: роман нашего времени (2008) [омонимия не снята] Все примеры (1)

   — Все дела да дела, — вздыхает опять Дмитрий Павлович, — а на звезды взглянуть и некогда. **Только и знаю**, **что** Большую Медведицу. Его грусть меня подкупает. [Олег Зайончковский. Счастье возможно: роман нашего времени (2008)] [омонимия не снята] ←…→

6. Михаил Гиголашвили. Чертово колесо (2007) [омонимия не снята] Все примеры (1)

   пустой дым — добычу ленивых духов, которые **только и знают**, **что** парить над кострами и воровать фимиам у храмовых голубей. [Михаил Гиголашвили. Чертово колесо (2007)] [омонимия не снята] ←…→

только и знать/делать

- *только и знать что* — 45 раза (а на самом деле 32)
- *только и делать что* — 285 раз
- Конструкция *только и делать что* почти в 10 раз частотнее, чем *только и знать что*

только и знать/делать

	наст.	прош.
только и знать	24 (75%)	8 (25%)
только и делать	142 (51%)	138 (49%)

- $\chi^2(df = 1, N = 312) = 5{,}87; p = 0{,}015$

	инф.	дублир.
только и знать	20 (62%)	12 (38%)
только и делать	0 (0%)	280 (100%)

только и знать/делать: выводы

- *только и делать что* почти в 10 раз частотнее, чем *только и знать что*
- В *только и делать что* смысловой глагол дублирует форму спрягаемого, а в *только и знать что* встречаются инфинитив (чаще) и дублирование (реже)
- Похоже, что *только и знать что* чаще употребляется в настоящем времени

Ударение в глаголах

- *о́тдал ~ отда́л*
- *о́тдало ~ отда́ло ~отдало́*
- Какие типы текстов позволяют изучать ударение в таких формах?
- **Транскрипты устных текстов или поэзия**

ruTenTen

- Самый большой русский корпус
- 14 млрд словоформ
- Автоматически скачанные из интернета тексты
- Статистическое снятие морфологической омономии
- Синтаксической разметки нет

MULTEXT-East

- Система морфологической разметки русских текстов
- Пометы — буква части речи + последовательность грамматических значений в определенном порядке
- *уборщицах* — Ncfply
- *удовольствием* — Ncfsin
- *экономишь* — Vmip2s-a-p

Частотность

- Корпуса — основной инструмент для изучения частотности различных единиц в языке
- Ср. ранее сделанное исследование про *только и знать/делать что*

Частотность

- Задача: сравнить частотность некоторых единиц в разных (под)корпусах
- Проблема: разный объём корпусов

Слово *телефон* — у кого чаще?

Лев Толстой — 9 вхождений в НКРЯ

Анатолий Рыбаков — 6 вхождений в НКРЯ

Как считать частотность?

Автор	Число вхождений	Объём подкорпуса	Частотность (на млн)
Л. Толстой	9	1 906 467	5
А. Рыбаков	6	180 583	33

- **Частотность** = Число вхождений / Объём подкорпуса (в словах) * 1 000 000
- Дробная частотность — неудобно
  $\Longrightarrow$ домножаем на 1 000 000 (ipm, instances per million words)

Частотные словари русского языка

- Гарри Джоссельсон (Детройт, 1953, 1 млн)
- Эви Штейнфельдт (Таллин, 1963, 400 тыс.)
- Лидия Засорина (ред.; Ленинградский университет + Горьковский университет, 1977, 1 млн)
- Ольга Ляшевская, Сергей Шаров (2009, 92 млн): http://dict.ruslang.ru/freq.php

Частотность в Macmillan Dictionary

- *** — 1–2500 места в частотном списке (*the*, *animal*)
- ** — 2501–5000 места в частотном списке (*appropriate*, *tragedy*)
- * — 5001–7500 места в частотном списке (*restriction*, *allegedly*)
- без звёдочек — остальные (*crescent*, *thatch*)

wordandphrase.info

SEE LISTS	FREQ RANGE	1-500	501-3000	> 3000		ACAD	HELP
	166 WORDS	70 %	15 %	15 %		1 %	

Mr. and Mrs. **Dursley**, of number four, Privet Drive, were **proud** to say that they were **perfectly** normal, thank you very much. They were the last people you'd expect to be involved in anything strange or **mysterious**, because they just didn't hold with such **nonsense** .

Mr. **Dursley** was the director of a firm called Grunnings, which made **drills**. He was a big, **beefy** man with hardly any neck, although he did have a very large **mustache**. Mrs. **Dursley** was thin and **blonde** and had nearly twice the **usual** amount of neck, which came in very **useful** as she spent so much of her time **craning** over garden **fences**, **spying** on the neighbors. The Dursleys had a small son called Dudley and in their opinion there was no finer boy **anywhere** .

The Dursleys had everything they wanted, but they also had a secret, and their greatest fear was that somebody would discover it. They didn't think they could bear it if anyone found out about the Potters.

Случай с Оливером (Р. М. Фрумкина)

Заглонитель Ланс Оливер чуть не погиб в результате наплочения турма. Он ехал ласкунно на лошади покровнательно от Мэнсфилда (Австралия) и увидел вахню турмов, в которой было кастожно 15 животных. Столенно, ничего бы и не случилось, если бы собака Оливера не начала порочить на вахню.

Один из турмов — старый, крупный лователь, выбатушенный корочением собаки, бросился за ней. Та отпешила скумановаться за лошадью, на которой сидел Оливер. Тогда турм бросился уже на Оливера. Он схватил подвешенца отмаленными

Случай с Оливером (Р. М. Фрумкина)

Скотовод Ланс Оливер чуть не погиб в результате нападения кенгуру. Он ехал верхом на лошади неподалеку от Мэнсфилда (Австралия) и увидел стадо кенгуру, в котором было примерно 15 животных. Возможно, ничего бы и не случилось, если бы собака Оливера не начала лаять на стадо.

Один из кенгуру — старый крупный самец, раздраженный лаем собаки, бросился за ней. Та попыталась укрыться за лошадью, на которой сидел Оливер. Тогда кенгуру бросился уже на Оливера. Он схватил всадника передними

Тесты словарного запаса

- Английский язык: testyourvocab.com
- Русский язык: myvocab.info
- В основе — оценка знания слов разных групп частотности

Закон Ципфа

- Джордж Кингсли Ципф (Зиф; Geogre Kingsley Zipf, 1902–1950)
- Известен как автор закона о распределении частотностей слов

Закон Ципфа

- Построим частотный словарь и упорядочим его по убыванию
- Как могут быть распределены частоты?
- В частотных словарях реальных текстов частота обратно пропорциональна рангу

Закон Ципфа

- Простая формулировка: $f(r) = k/r$, где
  k — константа,
  r — ранг слова в частотном списке,
  $f(r)$ — частота слова с рангом r

Закон Ципфа

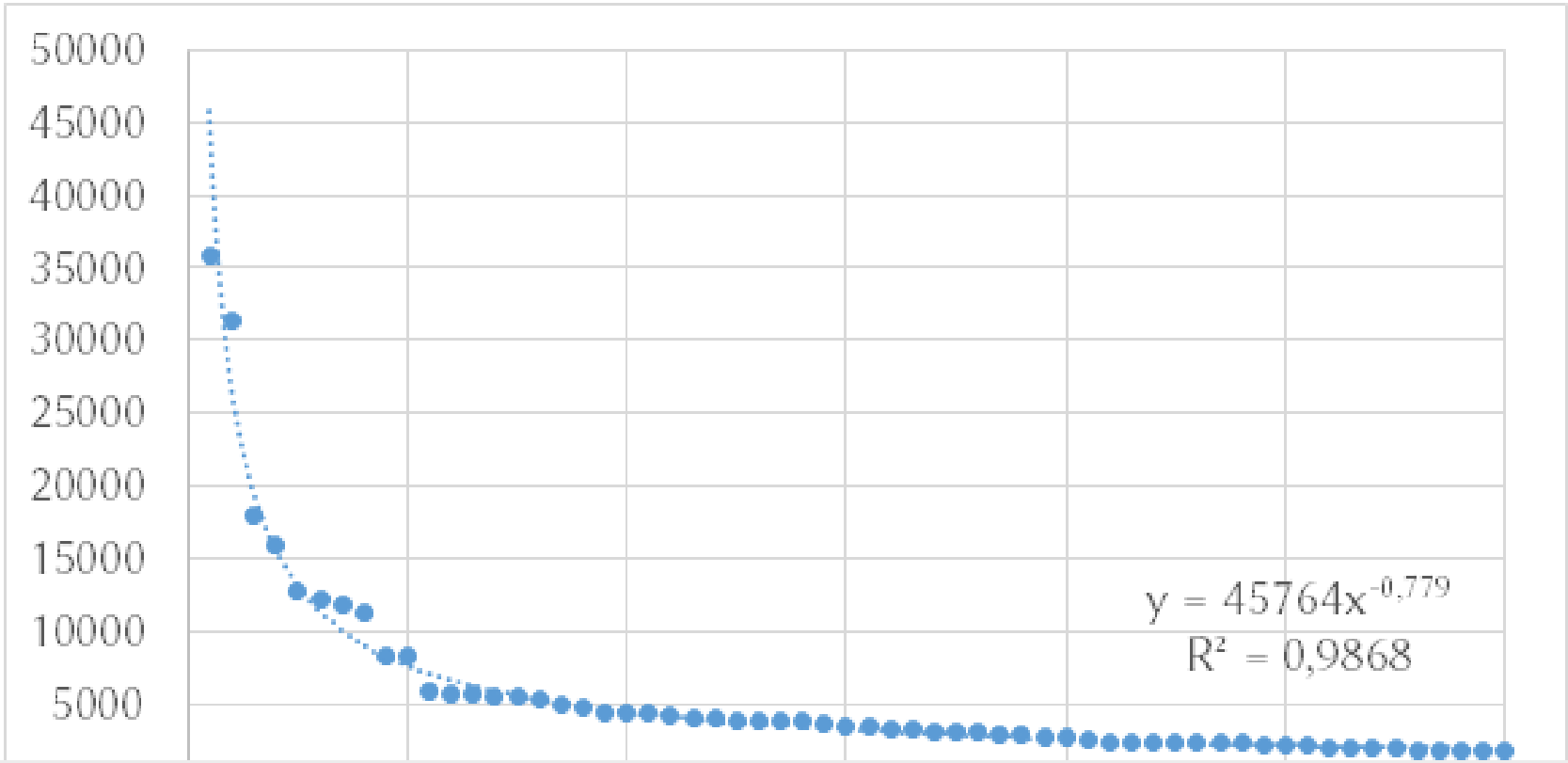
- Более сложная формулировка:
  $f(r) = k/r^a$, где
  a — константа
- Применим ли закон Ципфа для бесконечного количества элементов?

Закон Ципфа

- Простейший инструмент для проверки данных на соответствие закону Ципфа — Excel
- Построить график столбца с упорядоченными частотами и добавить степенную линию тренда

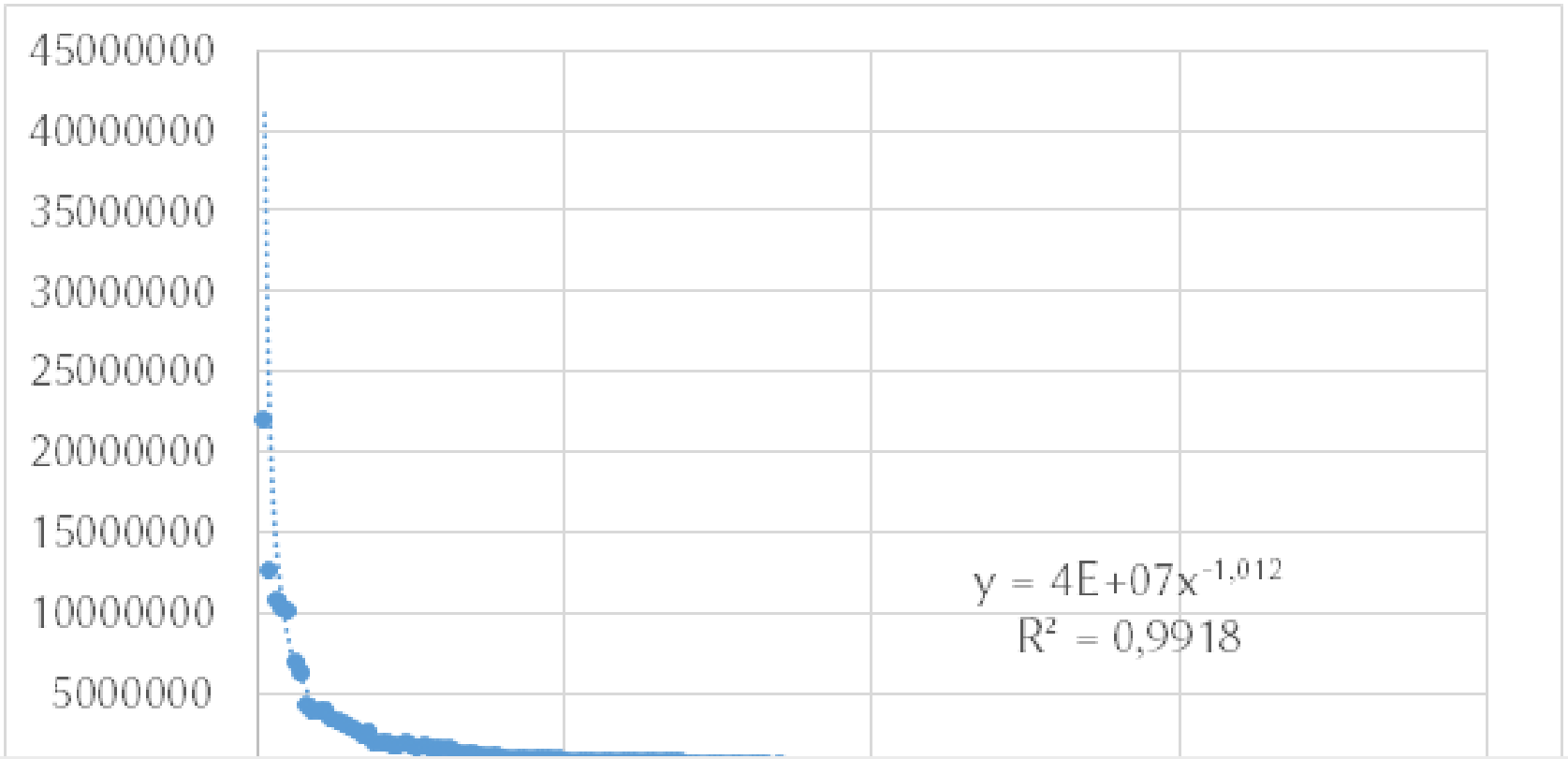
Закон Ципфа

60 наиболее частотных русских слов [Ляшевская, Шаров 2009]:



Закон Ципфа

200 наиболее частотных английских слов (COCA, www.wordfrequency.info):



Закон Ципфа: критика

- Нет лингвистически содержательного объяснения
- Закон Ципфа хорошо работает в середине распределения, но не по краям
- Возможно, нужны уточнения (и они делаются)

n-граммные модели

- Корпуса позволяют оценивать вероятность новых предложений, текстов и т. п.
- Важно для различных компьютерно-лингвистических приложений
- Униграммные, биграммные, триграммные, ... модели

Униграммная модель

- Каждое слово — независимое случайное событие
- Вероятность выбора того или иного слова приравнивается к его частотности в обучающем корпусе (оценка **методом максимального правдоподобия**)
- $P(the) = 0.047205$, $P(if) = 0.002166$, $P(strange) = 0.000033$

(Игрушечный) машинный перевод

Это мой дом

- *Это* $\longrightarrow$ *This is, These are, This*
- *мой* $\longrightarrow$ *my, mine, wash*
- *дом* $\longrightarrow$ *house, home, building*

Это мой дом $\longrightarrow$ *These are mine house; This my building; This wash house; This is my house; This my house; ...*

Униграммная модель

- $P(\textit{These are mine house}) =$
  $P(\textit{these}) \cdot P(\textit{are}) \cdot P(\textit{mine}) \cdot P(\textit{house}) =$
  $0{,}001237 \cdot 0{,}005034 \cdot 0{,}000055 \cdot 0{,}000258 = 9 \cdot 10^{-14}$
- $P(\textit{This is my house}) =$
  $P(\textit{this}) \cdot P(\textit{is}) \cdot P(\textit{my}) \cdot P(\textit{house}) =$
  $0{,}004842 \cdot 0{,}010128 \cdot 0{,}001978 \cdot 0{,}000258 = 3 \cdot 10^{-11}$
- Проблемы?

Биграммная модель

- Каждое слово зависит только от предыдущего слова: марковская цепь 1-го порядка
- Как оценить $P(is|this)$ (т. е. вероятность встретить слово *is* после слова *this*)?

Биграммная модель

- $P(This\ is\ my\ house) = P(< s >) \cdot P(this| < s >) \cdot P(is|this) \cdot P(my|is) \cdot P(house|my) \cdot P(< /s > |house) = 0{,}0036746 \cdot 0{,}026704 \cdot 0{,}112626 \cdot 0{,}003033 \cdot 0{,}003551 \cdot 0{,}111606 = 1 \cdot 10^{-10}$
- $P(This\ my\ house) = P(< s >) \cdot P(this| < s >) \cdot P(my|this) \cdot P(house|my) \cdot P(< /s > |house) = 0{,}0036746 \cdot 0{,}026704 \cdot 0{,}000130 \cdot 0{,}003551 \cdot 0{,}111606 = 5 \cdot 10^{-11}$
- $P(is|this) \cdot P(my|is) = 0{,}112626 \cdot 0{,}003033 = 0{,}000342 > P(my|this) = 0{,}000130$