МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ

Урок №11

Коллокации

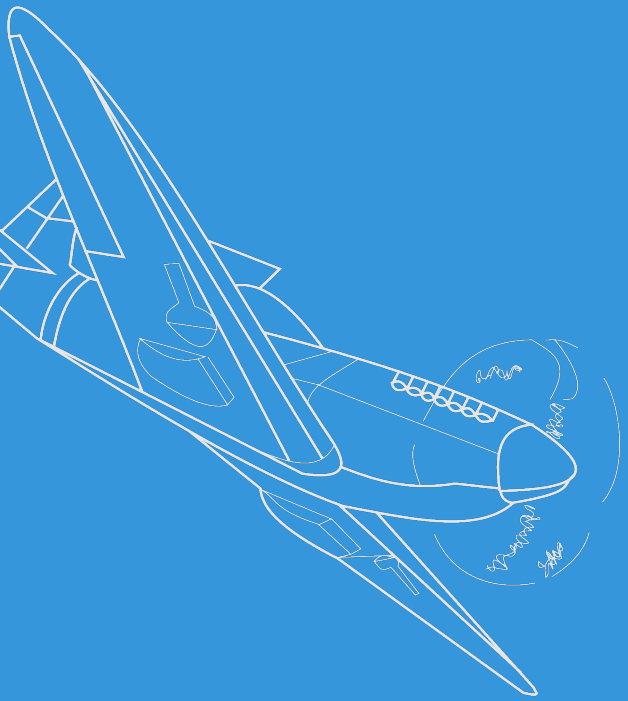
(основано на слайдах Андрея Калинина, Hinrich Schütze, Christina Lioma)

16 декабря 2021 года

Антон Кухтичев

Содержание занятия

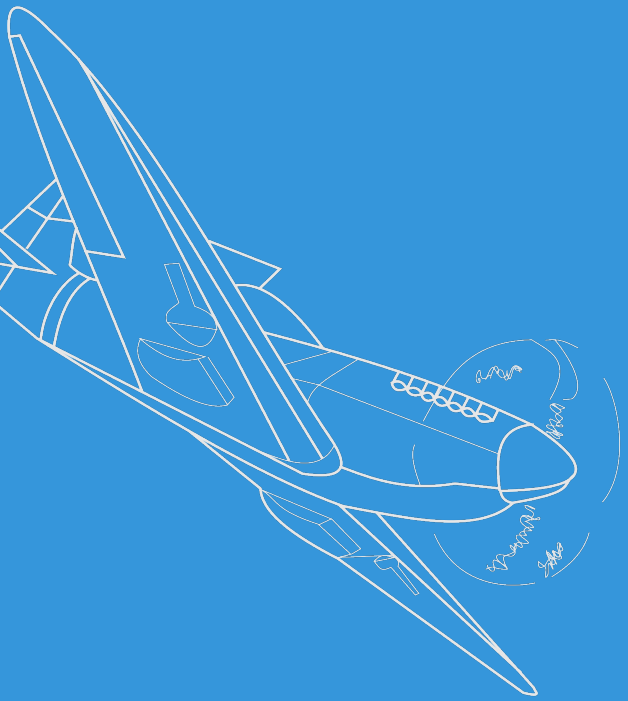
1. Коллокации
2. Частотность
3. Среднее и отклонение
4. Проверка гипотез



Технический долг

Зависимые/независимые смысловые фрагменты

- Независимый смысловой фрагмент содержит подлежащее, глагол и законченную мысль.
- Зависимый смысловой фрагмент содержит подлежащее и глагол, но не содержит законченной мысли.
- Часто зависимый смысловой фрагмент помечается зависимым словом-маркером (after, although, as, as if, because и т.д.).
- Although it is raining, I am going out for a run.
  - I am going out for a run - независимый смысловой фрагмент
  - Although it is raining - зависимый смысловой фрагмент



Коллокации

Литература

Foundations of statistical natural language processing, Christopher Manning, Hinze Schultze.

Коллокация

- Устойчивое словосочетание, целостное синтаксически и семантически.
  - Одно слово при этом сохраняет своё значение.
  - Другое слово обусловлено традицией.
- Например:
  - strong tea, powerful drug.
  - идёт дождь, молоть чушь
- Ограниченная композиционность

Зачем их искать?

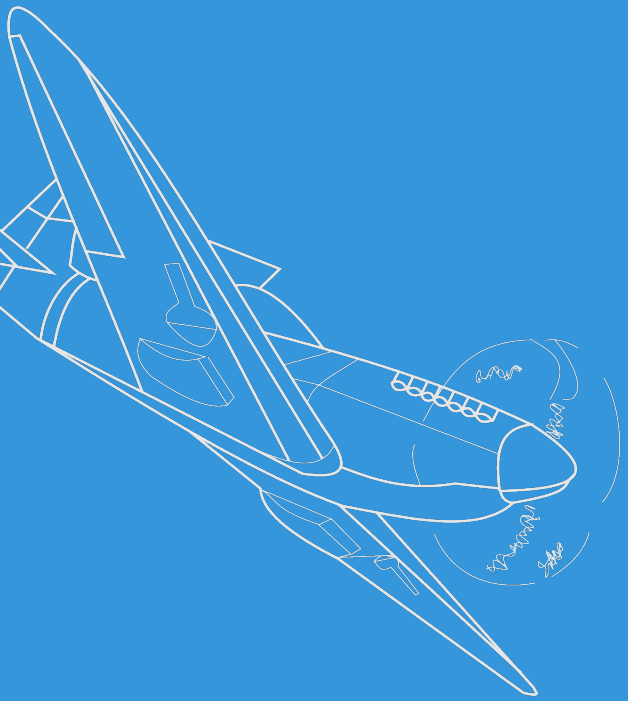
- Генерация текста.
  - Чтобы текст выглядел естественно.
- Лексикография.
  - Вхождения в словари.
- Разметка текста.
- Корпусные исследования.
  - Почему tea – strong, a drug – powerful, но не наоборот?

Признаки коллокации

- Некомпозиционность
  - Смысл коллокации не является композицией смысла её частей.
- Незаменяемость
  - Нельзя заменить зависимое слово на другое подходящее по смыслу или контексту.
- Немодифицируемость
  - Компоненты коллокации не получается свободно модифицировать следуя грамматическим правилам (идиомы).

Корпус

- Архив New York Times за 4 месяца
  - Август – Ноябрь, 1990
- 115 мегабайтов текста
- 14 миллионов слов



Частотность

Частотность биграмм

- Частотная верхушка дает не очень хорошие результаты
- За исключением ни одной коллокации.

$C(w^1\ w^2)$	w^1	w^2
80871	of	the
58841	in	the
26430	to	the
21842	on	the
21839	for	the
18568	and	the
16121	that	the
15630	at	the
15494	to	be
13899	in	a
13689	of	a
13361	by	the
13183	with	the
12622	from	the
11428	New	York
10007	he	said
9775	as	a
9231	is	a
8753	has	been
8573	for	a

Добавим фильтр по частям речи

- A – adjective, прилагательное
- P – preposition, предлог
- N – noun, существительное

Tag Pattern	Example
A N	*linear function*
N N	*regression coefficients*
A A N	*Gaussian random variable*
A N N	*cumulative distribution function*
N A N	*mean squared error*
N N N	*class probability function*
N P N	*degrees of freedom*

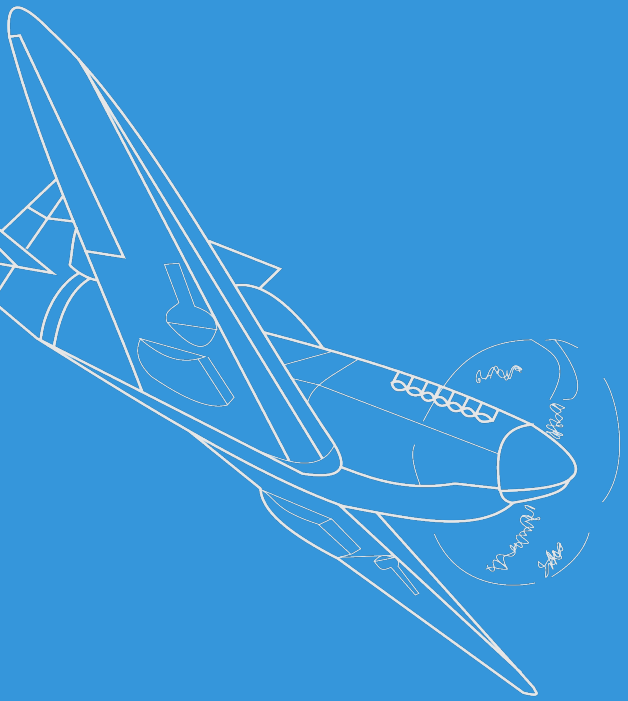
Результат применения фильтра

- Только три исключения:

$C(w^1\ w^2)$	w^1	w^2	Tag Pattern
11487	New	York	A N
7261	United	States	A N
5412	Los	Angeles	N N
3301	last	year	A N
3191	Saudi	Arabia	N N
2699	last	week	A N
2514	vice	president	A N
2378	Persian	Gulf	A N
2161	San	Francisco	N N
2106	President	Bush	N N
2001	Middle	East	A N
1942	Saddam	Hussein	N N
1867	Soviet	Union	A N
1850	White	House	A N
1633	United	Nations	A N
1337	York	City	N N
1328	oil	prices	N N
1210	next	year	A N
1074	chief	executive	A N
1073	real	estate	A N

20 частых употреблений strong или powerful

w	$C(strong, w)$	w	$C(powerful, w)$
support	50	force	13
safety	22	computers	10
sales	21	position	8
opposition	19	men	8
showing	18	computer	8
sense	18	man	7
message	15	symbol	6
defense	14	military	6
gains	13	machines	6
evidence	13	country	6
criticism	13	weapons	5
possibility	11	post	5
feelings	11	people	5
demand	11	nation	5
challenges	11	forces	5
challenge	11	chip	5
case	11	Germany	5
supporter	10	senators	4
signal	9	neighbor	4
man	9	magnet	4



Среднее и отклонение

Коллокация может иметь разрыв

- knock … door:
  - she **knocked** on his **door**
  - they **knocked** at the **door**
  - 100 women **knocked** on Donaldson’s **door**
  - a man **knocked** on the metal front **door**
- Строим биграммы для всех слов внутри окна некоторого размера.

Пример построения биграмм внутри окна

- Stocks crash as rescue plan teeters
- Биграммы для трёхсловного окна:

stocks crash	*stocks as*	*stocks rescue*		
	crash as	*crash rescue*	*crash plan*	
		as rescue	*as plan*	*as teeters*
			rescue plan	*rescue teeters*
				plan teeters

Среднее

- knock … door:
  - she **knocked** on his **door**
  - they **knocked** at the **door**
  - 100 women **knocked** on Donaldson’s **door**
  - a man **knocked** on the metal front **door**
- Среднее расстояние между knock и door:

$$\frac{1}{4}(3+3+5+5)=4.0$$

- door … knock – отрицательное расстояние

Отклонения

- n – количество биграмм с обоими словами
- d_i – i-ое расстояние

$$s^2 = \frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1}$$

- Если расстояние всегда одинаковое, s = 0
- Если расстояние случайное, s велико.
- Для knocked ... door:

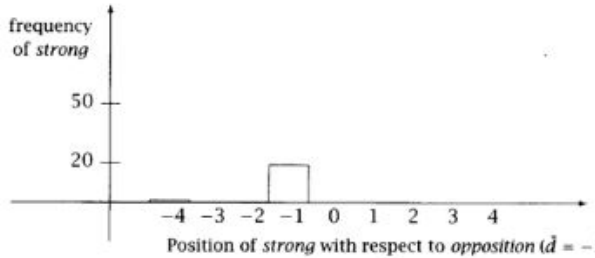
$$s = \sqrt{\frac{1}{3}((3-4.0)^2 + (3-4.0)^2 + (5-4.0)^2 + (5-4.0)^2)} \approx 1.15$$

Гистограммы отклонений

strong / opposition

m = -1.15

s = 0.67

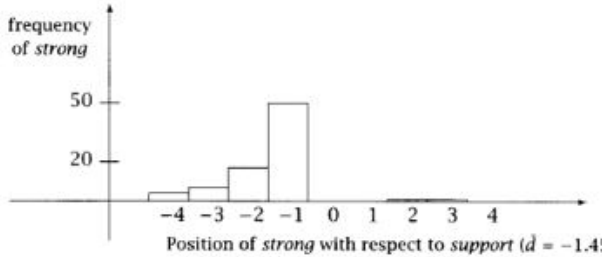


Position of *strong* with respect to *opposition* ($\bar{d} = -1.15, s = 0.67$).

strong / support

m = -1.45

s = 1.07

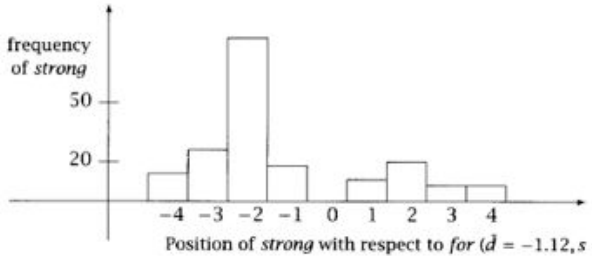


Position of *strong* with respect to *support* ($\bar{d} = -1.45, s = 1.07$).

strong / for

m = -1.12

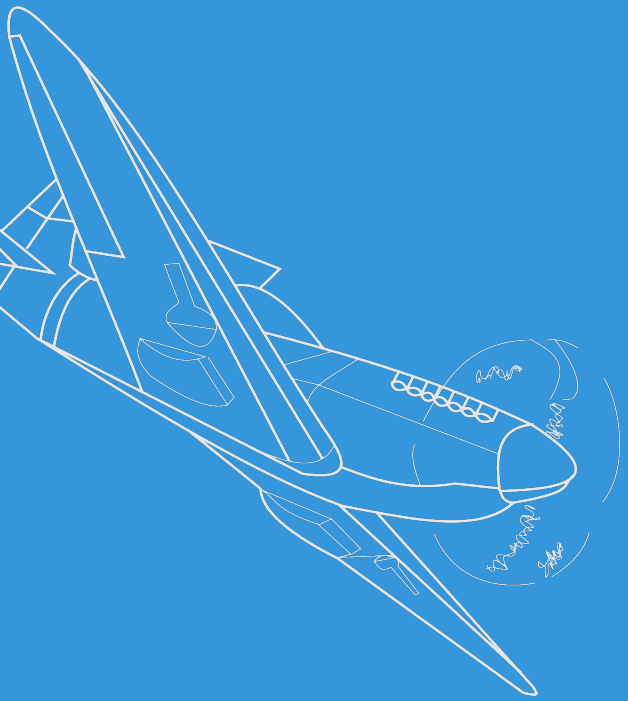
s = 2.15



Position of *strong* with respect to *for* ($\bar{d} = -1.12, s = 2.15$).

Примеры

s	$\bar{d}$	Count	Word 1	Word 2
0.43	0.97	11657	New	York
0.48	1.83	24	previous	games
0.15	2.98	46	minus	points
0.49	3.87	131	hundreds	dollars
4.03	0.44	36	editorial	Atlanta
4.03	0.00	78	ring	New
3.96	0.19	119	point	hundredth
3.96	0.29	106	subscribers	by
1.07	1.45	80	strong	support
1.13	2.57	7	powerful	organizations
1.01	2.00	112	Richard	Nixon
1.05	0.00	10	Garrison	said



Проверка гипотез

Проблема

- Биграмма может быть частой случайно
  - new companies
    - оба слова довольно частотные
    - словосочетание может встретиться и без особенного смысла
- Нужно иметь способ отличить случайное от неслучайного.

Классическая задача из статистики

- Основная гипотеза, H_0 – между словами нет зависимости.
- Вычисляем p с учётом истинности H_0.
- Если p невелико (уровень значимости $p < 0.05$, 0.01, 0.005 или 0.001) – отвергаем H_0.
- В обратном случае – считаем гипотезу вероятной.

Проверка гипотез

- То есть, проверяем не только саму коллокацию, но и наличие достаточного числа подтверждений для неё.
- Если два слова встречаются независимо друг от друга, то

$$P(w^1 w^2) = P(w^1)P(w^2)$$

- Это не совсем корректно, но подходит для демонстрации метода.

Критерий Стьюдента, t-критерий

- Основная гипотеза – выборка взята из нормального распределения с мат. ожиданием μ.
- Сравнивается ожидаемое и наблюдаемое среднее, нормированное на дисперсию:

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}}$$

- Результат проверяется по таблицам.

Пример

- Средний рост людей некоторой национальности – 158 см.
- У нас есть случайная выборка 200 людей,

  $\bar{x}$ =169 и s^2 = 2600.
- Проверяем:

$$t = \frac{169 - 158}{\sqrt{\frac{2600}{200}}} \approx 3.05$$

- Уровень значимости 0,05 – 2.576.
- 3.05 > 2.576.

Поиск коллокаций

- Рассматриваем корпус как последовательность N биграмм.
- Случайная величина: 1 если встретилась рассматриваемая биграмма.

new companies

- Оценим вероятности методом наибольшего правдоподобия:

$$P(new) = \frac{15828}{14307668}$$

$$P(companies) = \frac{4675}{14307668}$$

- Основная гипотеза: new и companies независимы:

$$H_0 : P(new\ companies) = P(new)P(companies) = \frac{15828}{14307668} \times \frac{4675}{14307668} \approx 3.615 \times 10^{-7}$$

new companies

- Биномиальное распределение с p = 3,615×10^{-7}
- μ = 3,615×10^{-7}
- σ^2=p(1−p)≈p=3,615×10^{-7}
- C(new companies) = 8, т.е.

  $\bar{x}$ = 8/14307668≈5,591×10^{-7}
- Тогда

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}} \approx \frac{5.59110^{-7} - 3.61510^{-7}}{\sqrt{\frac{5.59110^{-7}}{14307668}}} \approx 0.999932$$

- t < 2,576 (порог для α=0,05), основную гипотезу отвергнуть нельзя

Примеры применения t-критерия

t	$C(w^1)$	$C(w^2)$	$C(w^1\ w^2)$	w^1	w^2
4.4721	42	20	20	Ayatollah	Ruhollah
4.4721	41	27	20	Bette	Midler
4.4720	30	117	20	Agatha	Christie
4.4720	77	59	20	videocassette	recorder
4.4720	24	320	20	unsalted	butter
2.3714	14907	9017	20	first	made
2.2446	13484	10570	20	over	many
1.3685	14734	13478	20	into	them
1.2176	14093	14776	20	like	people
0.8036	15019	15629	20	time	last

- 824 из 831 биграммы – коллокации по этому критерию.
- Поэтому он полезен как основа ранжирования коллокаций.

Критерий Пирсона, χ^2-критерий

- Критерий Стьюдента подразумевает нормальное распределение (хотя бы аппроксимированное)
- Критерий Пирсона не делает таких предположений.
- Сравниваются наблюдаемые частотности с ожидаемыми в случае независимости.

Критерий Пирсона, χ^2-критерий

- Критерий Стьюдента подразумевает нормальное распределение (хотя бы аппроксимированное)
- Критерий Пирсона не делает таких предположений.
- Сравниваются наблюдаемые частотности с ожидаемыми в случае независимости.

Простейший случай

	w_1 = *new*	$w_1 \neq$ *new*
w_2 = *companies*	8 (*new companies*)	4667 (e.g., *old companies*)
$w_2 \neq$ *companies*	15820 (e.g., *new machines*)	14287181 (e.g., *old machines*)

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$E_{1,1}$= $\frac{8 + 4667}{N} \times \frac{8 + 15820}{N} \times N \approx 5.2$

$$X^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11} + O_{12})(O_{11} + O_{21})(O_{12} + O_{22})(O_{21} + O_{22})}$$

$$\frac{14307668(8 \times 14287181 - 4667 \times 15820)^2}{(8 + 4667)(8 + 15820)(4667 + 14287181)(15820 + 14287181)} \approx 1.55$$

Для α=0,05 порог 3,841, основная гипотеза не может быть отвергнута.

vache – cow?

	cow	¬ *cow*
vache	59	6
¬ *vache*	8	570934

$$\chi^2 = 456400.$$

- Т.е., можно предположить, что vache – хороший вариант перевода cow.

Рекомендуемая литература

Foundations of statistical natural language processing, Christopher Manning, Hinze Schultze

Для саморазвития (опционально)

Чтобы не набирать двумя пальчиками

Спасибо за внимание!

Антон Кухтичев

a.kukhtichev@mail.ru

@toshunster