

Математические модели в лингвистике

10. Измерение расстояний между корпусами текстов

Мати Пентус, Александр Пиперски,
Алексей Сорокин

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

- *У попа была собака, он её любил. Она съела кусок мяса, он её убил.*
- *Я вас любил: любовь еще, быть может, в душе моей угасла не совсем; но пусть она вас больше не тревожит; я не хочу печалить вас ничем.*
- *Я больше люблю собак, чем кошек, но у меня нет собаки, потому что мне не хватает денег, чтобы кормить её МЯСОМ.*

Расстояние между корпусами

- Задача: найти меру $d(A, B)$, которая позволит оценить, насколько похожи между собой тексты / корпуса текстов A и B , и сравнивать такие оценки между собой.
- Пример:
 - На что больше похожи английские любовные романы: на научную фантастику или на приключенческие романы?
 - На что больше похожи стихи Мандельштама: на стихи Ахматовой, Блока или Гумилёва?

Сходные задачи

- Информационный поиск
- Детекция плагиата

Информационный поиск: отличия

- Оценивается сходство текстов существенно различного размера (длинных документов с коротким поисковым запросом).
- Существенно только ранжирование документов, наиболее сходных с запросом.

Детекция плагиата: отличия

- Ищутся точные текстуальные совпадения.
- Подразумевается злая воля противоположной стороны.

Частотные списки

- Проще подступиться к задаче сравнения корпусов между собой, если сравнивать не корпуса, а их частотные списки.

она	3	вы	4	я	3
он	2	не	3	нет	2
быть	1	я	2	собака	2
кусок	1	больше	1	больше	1
любить	1	быть	1	деньги	1
мясо	1	душа	1	кормить	1
поп	1	еще	1	кошка	1
собака	1	любить	1	любить	1
съесть	1	любовь	1	мясо	1
у	1	мой	1	но	1
убить	1	мочь	1	она	1
ВСЕГО	14	ничто	1	потому	1
		но	1	у	1
		она	1	хватать	1
		печалить	1	чем	1
		пусть	1	что	1
		совсем	1	чтобы	1
		тревожить	1	ВСЕГО	21
		угаснуть	1		
		хотеть	1		
		ВСЕГО	26		

Нормированные частотные списки

- Полученные списки плохи тем, что числа в них зависят от размера текста
- Нормируем их — например, приводя к процентам (в корпусной лингвистике обычно приводят к единицам на миллион, ipm)

она	3	21	вы	4	15	я	3	14
он	2	14	не	3	12	нет	2	10
быть	1	7	я	2	8	собака	2	10
кусок	1	7	больше	1	4	больше	1	5
любить	1	7	быть	1	4	деньги	1	5
мясо	1	7	душа	1	4	кормить	1	5
поп	1	7	еще	1	4	кошка	1	5
собака	1	7	любить	1	4	любить	1	5
съесть	1	7	любовь	1	4	мясо	1	5
у	1	7	мочь	1	4	но	1	5
убить	1	7	мой	1	4	она	1	5
ВСЕГО	14	98	ничто	1	4	потому	1	5
			но	1	4	у	1	5
			она	1	4	хватать	1	5
			печалить	1	4	чем	1	5
			пусть	1	4	что	1	5
			совсем	1	4	чтобы	1	5
			тревожить	1	4	ВСЕГО	21	104
			угаснуть	1	4			
			хотеть	1	4			
			ВСЕГО	26	103			

Для простоты пренебрежём погрешностью округления

	С1	С2	С3		С1	С2	С3
больше	0	4	5	она	21	4	5
быть	7	4	0	печалить	0	4	0
вы	0	15	0	поп	7	0	0
деньги	0	0	5	потому	0	0	5
душа	0	4	0	пусть	0	4	0
еще	0	4	0	собака	7	0	10
кормить	0	0	5	совсем	0	4	0
кошка	0	0	5	съесть	7	0	0
кусок	7	0	0	тревожить	0	4	0
любить	7	4	5	у	7	0	5
любовь	0	4	0	убить	7	0	0
мой	0	4	0	угаснуть	0	4	0
мочь	0	4	0	хватать	0	0	5
мясо	7	0	5	хотеть	0	4	0
не	0	12	0	чем	0	0	5
нет	0	0	10	что	0	0	5
ничто	0	4	0	чтобы	0	0	5
но	0	4	5	я	0	8	14
он	14	0	0				

Частотные списки как векторы

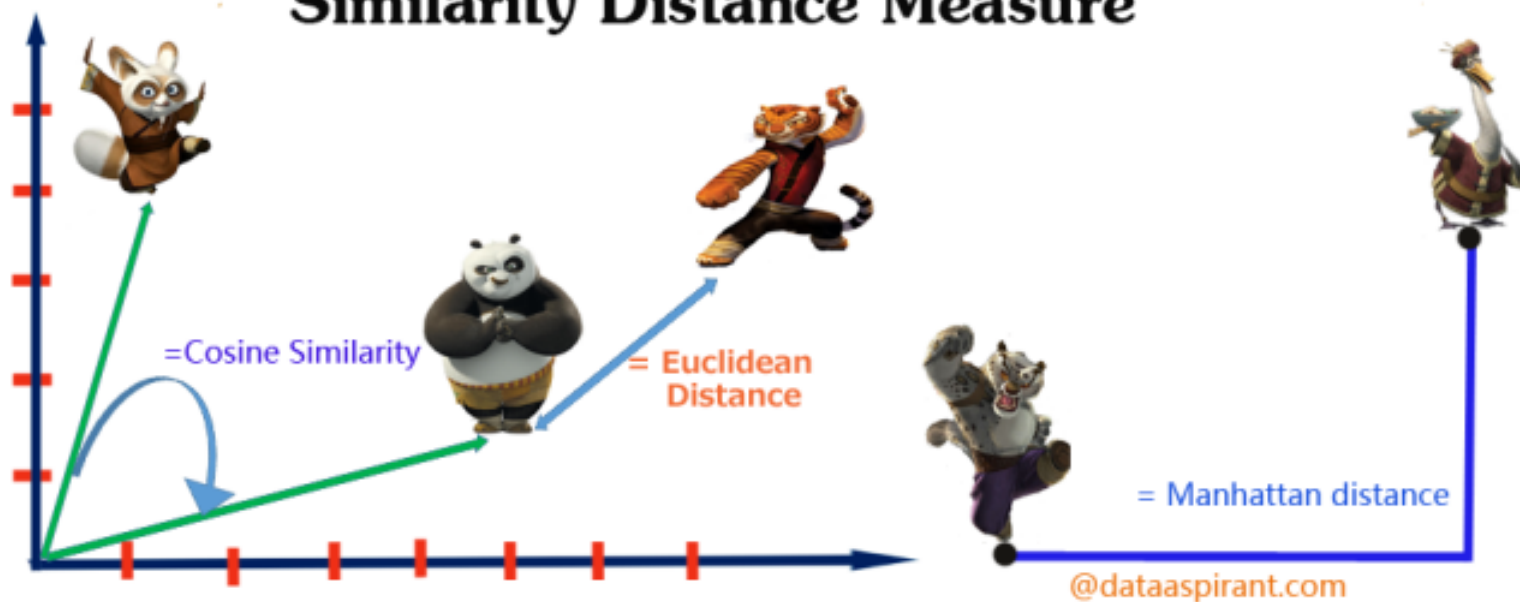
- Каждому корпусу теперь соответствует упорядоченная последовательность чисел — **вектор (кортеж)**
- C2: (4; 4; 15; 0; 4; 4; 0; ...; 0; 8)
- Как можно измерять расстояния между такими векторами?

Геометрические меры

1. Евклидово расстояние
2. Манхэттенское расстояние
(1–2. Расстояние Минковского)
3. Косинусное расстояние

Геометрические меры

Similarity Distance Measure



Евклид, Манхэттен, Минковский

- Евклидово расстояние:

$$D(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2}$$

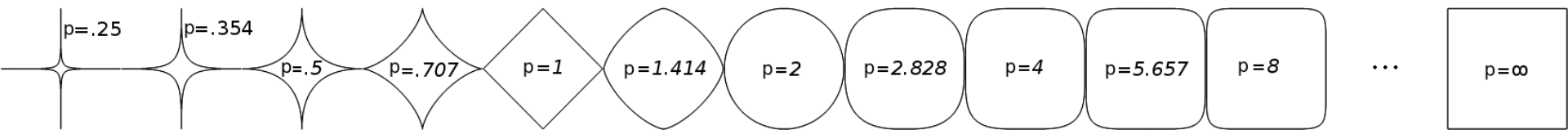
- Манхэттенское расстояние:

$$D(x, y) = |x_1 - y_1| + |x_2 - y_2| + \dots + |x_n - y_n|$$

- Расстояние Минковского:

$$D(x, y) = \sqrt[p]{|x_1 - y_1|^p + |x_2 - y_2|^p + \dots + |x_n - y_n|^p}$$

Единичная окружность при различных p Минковского



Косинусное расстояние

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|_2 \|\mathbf{B}\|_2} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$$

distance = 1 - similarity

Статистические меры

4. χ^2

5. Коэффициент Спирмена

Мера 4: χ^2

- Строим частотные списки корпусов A и B .
- Выбираем топ- N слов из объединённого частотного списка для двух корпусов.
- Представляем данные в виде таблицы сопряжённости $(N+1) \times 2$ (добавив строку для всех остальных слов) и вычисляем χ^2 обычным образом.

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

χ^2 : сравнение двух корпусов по Ляшевская & Шаров (2009)

	Худ50-60	Публ50-60	Σ
<i>и</i>	0.038	0.037	0.075
<i>в</i>	0.024	0.035	0.059
<i>не</i>	0.021	0.016	0.037
<i>на</i>	0.017	0.015	0.032
<i>я</i>	0.017	0.014	0.031
<i>быть</i>	0.012	0.014	0.026
<i>он</i>	0.020	0.012	0.032
<i>с</i>	0.011	0.011	0.022
<i>что</i>	0.008	0.008	0.016
<i>а</i>	0.010	0.006	0.016
ПРОЧЕЕ	0.822	0.832	1.654
Σ	1.000	1.000	2.000

χ^2 : сравнение двух корпусов по Ляшевская & Шаров (2009)

	Худ50-60	Публ50-60	
<i>и</i>	7×10^{-6}	7×10^{-6}	
<i>в</i>	0.0010	0.0010	
<i>не</i>	0.0003	0.0003	
<i>на</i>	6×10^{-5}	6×10^{-5}	
<i>я</i>	0.0001	0.0001	
<i>быть</i>	7×10^{-5}	7×10^{-5}	
<i>он</i>	0.0010	0.0010	
<i>с</i>	0	0	
<i>что</i>	0	0	
<i>а</i>	0.0005	0.0005	
ПРОЧЕЕ	3×10^{-5}	3×10^{-5}	
			0.0064

Мера 4: χ^2

- NB: в таблицу внесены относительные частоты, а не абсолютные, потому что нас интересует только размер эффекта, а не статистическая значимость (не нужно, чтобы пропорциональное увеличение выборок меняло расстояние между корпусами).
- Диапазон: от 0 (абсолютно одинаковые частоты) до 2 (абсолютно разные частоты; *кстати, какие?*).

χ^2 : расстояния между 4 корпусами по Ляшевская & Шаров (2009)

	Худ50–60	Худ70–80	Публ50–60	Публ70–80
Худ50–60	0	0.0003	0.0064	0.0039
Худ70–80	0.0003	0	0.0065	0.0035
Публ50–60	0.0064	0.0065	0	0.0012
Публ70–80	0.0039	0.0035	0.0012	0

Мера 5: Спирмен

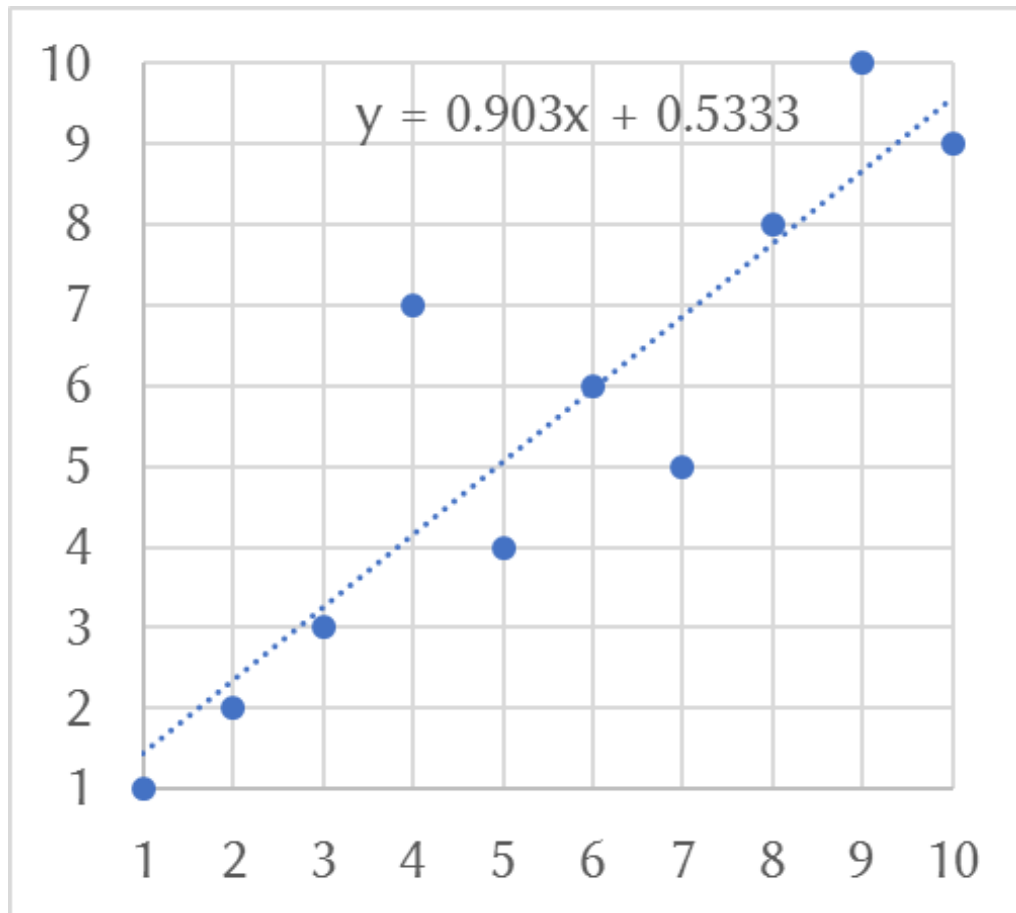
- Строим частотные списки корпусов A и B .
- Выбираем топ- N слов из объединённого частотного списка для двух корпусов.
- Определяем ранги слов в частотном списке, включающем в себя эти N слов, для A и в таком же списке для B и вычисляем коэффициент корреляции Спирмена r_s обычным образом исходя из разностей между рангами для каждого слова D_i :

$$r_s = 1 - \frac{6 \sum D_i^2}{N(N^2 - 1)}$$

Спирмен: сравнение двух корпусов по Ляшевская & Шаров (2009)

	Худ50-60		Публ50-60	
<i>и</i>	0.038	1	0.037	1
<i>в</i>	0.024	2	0.035	2
<i>не</i>	0.021	3	0.016	3
<i>на</i>	0.0172	5	0.015	4
<i>я</i>	0.0170	6	0.0139	6
<i>быть</i>	0.012	7	0.0145	5
<i>он</i>	0.020	4	0.012	7
<i>с</i>	0.011	8	0.011	8
<i>что</i>	0.008	10	0.008	9
<i>а</i>	0.010	9	0.006	10

Спирмен: сравнение двух корпусов по Ляшевская & Шаров (2009)



Мера 5: Спирмен

- Мера слишком чувствительна к частотам внизу списка (если переставить местами 1-е и 2-е слово, эффект тот же, как если переставить 100-е и 101-е слово)
- Диапазон: от -1 (противоположный порядок) до 1 (идентичный порядок)
- Мера расстояния: $1 - r_s$

Спирмен: расстояния между 4 корпусами по Ляшевская & Шаров (2009)

	Худ50–60	Худ70–80	Публ50–60	Публ70–80
Худ50–60	0	0.012	0.097	0.097
Худ70–80	0.012	0	0.121	0.085
Публ50–60	0.097	0.121	0	0.036
Публ70–80	0.097	0.085	0.036	0

Мера б: сходство по ключевым словам (Kilgarriff 2009; SketchEngine)

- Берём по топ-5000 частотных слов из каждого из корпусов
- Для каждого из слов в объединении этих множеств вычисляем, насколько оно характерно для одного или другого корпуса.
- $k(w) = \max \left(\frac{f_A(w)+n}{f_B(w)+n}, \frac{f_B(w)+n}{f_A(w)+n} \right)$
- Берём среднее значение k по N наиболее характерным словам.

Reference corpus: **Early English Books Online (EEBO)**[Switch focus and reference \(sub\)corpus](#)Page [Next >](#)

word	<i>British Academic Spoken English Corpus (BASE)</i>		<i>Early English Books Online (EEBO)</i>		Score
	frequency	frequency/mill ?	frequency	frequency/mill	
i	15,450	12337.7	145,195	147.1	50.3
okay	3,552	2836.5	0	0.0	29.4
n't	6,681	5335.2	89,115	90.3	28.6
just	4,066	3246.9	84,060	85.1	18.1
going	3,965	3166.3	94,406	95.6	16.7
actually	2,407	1922.1	26,344	26.7	16.0
got	3,037	2425.2	87,386	88.5	13.4
yeah	1,414	1129.2	0	0.0	12.3
sort	2,560	2044.3	74,752	75.7	12.2
get	2,783	2222.4	121,150	122.7	10.4
really	1,806	1442.2	50,276	50.9	10.2
something	1,687	1347.2	41,626	42.2	10.2

КС: сравнение 4 английских корпусов в SketchEngine

($n = 100 / 1000000 = 0.0001$)

	(B A S E)	(B A W E)	B r o w n	(E E B O)
British Academic Spoken English Corpus (BASE)	1.00	<u>3.28</u>	<u>3.12</u>	<u>4.20</u>
British Academic Written English Corpus (BAWE)	<u>3.28</u>	1.00	<u>2.21</u>	<u>3.45</u>
Brown	<u>3.12</u>	<u>2.21</u>	1.00	<u>2.80</u>
Early English Books Online (EEBO)	<u>4.20</u>	<u>3.45</u>	<u>2.80</u>	1.00

Мера б: сходство по ключевым словам (Kilgarriff 2009; SketchEngine)

- Диапазон: от 1 (абсолютно идентичные частотные списки) и более.

Пожелания к мерам расстояния между корпусами (1)

- Мера должна являться метрикой (в математическом смысле)
 - Тождество: $d(A, A) = 0$
Корпус находится от на расстоянии 0 от самого себя.
 - Симметрия: $d(A, B) = d(B, A)$
Расстояние от A до B равно расстоянию от B до A .
 - Неравенство треугольника:
 $d(A, C) \leq d(A, B) + d(B, C)$

Пожелания к мерам расстояния между корпусами (2)

- **Ограниченность:**
мера должна иметь фиксированный диапазон значений (тогда расстояние легко преобразуется в близость и наоборот).
- **Устойчивость к окружению:**
 $d(A, B)$ не должно меняться в зависимости от того, какие корпуса есть в общем наборе для исследования.
- **Соответствие реальности**

Оценка мер расстояния: Known-Similarity Corpora (KSC)

- Kilgarriff 2001
- Берутся два исходных корпуса A и B , которые априорно считаются существенно различными
- Эти корпуса нарезаются на куски одинакового размера, из которых составляются корпуса известной степени сходства (Known-Similarity Corpora)

Оценка мер расстояния: Known-Similarity Corpora



C_0	B_1	B_2	B_3	B_4	B_5	B_6	B_7	B_8	B_9	B_{10}
C_1	A_1	B_{11}	B_{12}	B_{13}	B_{14}	B_{15}	B_{16}	B_{17}	B_{18}	B_{19}
C_2	A_2	A_3	B_{20}	B_{21}	B_{22}	B_{23}	B_{24}	B_{25}	B_{26}	B_{27}
C_3	A_4	A_5	A_6	B_{28}	B_{29}	B_{30}	B_{31}	B_{32}	B_{33}	B_{34}
C_4	A_7	A_8	A_9	A_{10}	B_{35}	B_{36}	B_{37}	B_{38}	B_{39}	B_{40}
C_5	A_{11}	A_{12}	A_{13}	A_{14}	A_{15}	B_{41}	B_{42}	B_{43}	B_{44}	B_{45}
C_6	A_{16}	A_{17}	A_{18}	A_{19}	A_{20}	A_{21}	B_{46}	B_{47}	B_{48}	B_{49}
C_7	A_{22}	A_{23}	A_{24}	A_{25}	A_{26}	A_{27}	A_{28}	B_{50}	B_{51}	B_{52}
C_8	A_{29}	A_{30}	A_{31}	A_{32}	A_{33}	A_{34}	A_{35}	A_{36}	B_{53}	B_{54}
C_9	A_{37}	A_{38}	A_{39}	A_{40}	A_{41}	A_{42}	A_{43}	A_{44}	A_{45}	B_{55}
C_{10}	A_{46}	A_{47}	A_{48}	A_{49}	A_{50}	A_{51}	A_{52}	A_{53}	A_{54}	A_{55}

Оценка мер расстояния: Known-Similarity Corpora

- На наборе из 11 KSC:
660 сравнений $d(C_a, C_b) ? d(C_c, C_d)$, где $c \leq a < b \leq d$, при этом $a \neq c$ или $b \neq d$
- Мера расстояния хорошо отражает реальность, если даёт правильный знак в большой доле из этих 660 случаев.
- Доля правильных знаков и есть оценка меры.

Kilgarriff 2001: сравнение мер

Table 7. Comparison of four measures

	spear	χ^2	closed	type 1	type 2
KSC-set					
acc_gua	93.33	91.33	82.22	81.11	80.44
art_gua	95.60	93.03	84.00	83.77	84.00
bmj_gua	95.57	97.27	88.77	89.11	88.77
env_gua	99.65	99.31	87.07	84.35	86.73

Table 8. Spearman/ χ^2 comparison on all KSCs

	spear	χ^2	tie	total
Highest score	5	13	3	21

- Победитель — χ^2 .

Сравнение мер

- Воспроизводим эксперимент Килгарриффа, но с бóльшим количеством мер и на большем количестве корпусов.
- Два этапа:
 - подбор параметров
 - собственно тестирование мер

Материал исследования

- 20 подкорпусов Британского национального корпуса (BNC) объёмом $\geq 550\ 000$ токенов
- 10 случайно выбранных подкорпусов используются на этапе подбора параметров, 10 подкорпусов — на этапе тестирования мер.

ID	Source	Token count	BNC file ID's
acc	Accountancy	598,379	CBT-CBY
art	The Art Newspaper	696,904	CKT-CKY, EBS-EBX
bel	The Belfast Telegraph	847,445	HJ3-HJ4, K29-K35
bio	The Dictionary of National Biography: Missing persons	773,750	GSX-GTH
eco	The Economist	929,886	ABD-ABK, CR7-CRC
gua	The Guardian	984,553	A7S-AAX
gut	Gut: Journal of Gastroenterology and Hepatology	808,792	HU2-HU4, HWS-HWT
han	Hansard Extracts	1,234,630	HHV-HHX
ind	The Independent	1,131,300	A1D-A5X
kee	Keesings Contemporary Archives	2,875,490	HKP-HLT
law	The Weekly Law Reports	721,892	FBS-FE3
liv	Liverpool Daily Post and Echo	921,162	K36-K4M, K97
mir	The Daily Mirror	833,598	CH1-CH3, CH5-CH7
new	Central television news scripts	1,332,831	K1B-K28
nor	Northern Echo	1,272,118	K4N-K55
nsc	New Scientist	980,897	ANX, B71-B7N
sco	The Scotsman	1,394,751	K56-K5M
tel	Daily Telegraph	1,319,936	AH9-AL6
tod	Today	1,419,457	CBC-CBG, CEK-CEP
uni	Unigram X	609,537	CMW-CN0, CS8-CTV

Подбор параметров

- Для Евклидова, манхэттенского и косинусного расстояния, χ^2 , Спирмена — количество слов, участвующих в сравнении.
- Для ключевых слов — количество отбираемых ключевых слов и константа, добавляемая к числителю и знаменателю.

Подбор параметров: χ^2

<i>N</i>	Медиана	Ср. арифм.		<i>N</i>	Медиана	Ср. арифм.
10	97,27	94,65		316	99,39	97,12
15	95,98	94,63		464	99,09	97,09
22	97,50	95,13		681	99,02	96,98
32	98,26	95,37		1000	98,94	96,74
46	99,24	96,01		1468	98,94	96,73
68	99,32	96,38		2154	98,71	96,74
100	99,47	96,68		3162	98,64	96,72
147	99,47	96,99		Все	97,73	96,01
215	99,47	97,19				

Подбор параметров: результаты

	Параметры	Медиана	Ср. арифм.
Евклид	$N = 1000$	99,17	96,36
Манхэттен	$N = 215$	99,47	96,95
Косинус	$N = \text{все слова}$	98,86	96,54
χ^2	$N = 215$	99,47	97,19
Спирмен	$N = 215$	99,09	96,05
Ключевые	$N = 68, n = 0.01$	99,62	97,12

Тестирование

...	acc	bio	gua	gut	law	liv	mir	sco	tod	uni
acc		EMCχK	K	χ	χK	K	M	χ	MC	χK
bio			M	EMCK	EMC	Mχ	EMCχSK	EMCK	EMCK	EMCχSK
gua				EC	M	K	E	S	EC	K
gut					M	M	EMC	M	M	E
law					M	M	M	M	M	M
liv							C	E	S	C
mir								C	E	χK
sco									E	M
tod										M
uni										

E 15
M 24
C 14
χ 9
S 4
K 13

Тестирование

(20 прогонов × 45 пар корпусов)

- Манхэттен 429 побед 48%
- Евклид 415 побед 46%
- Косинус 386 побед 43%
- χ^2 293 победы 33%
- Ключевые 260 побед 29%
- Спирмен 127 побед 14%

Что лучше сравнивать?

- Во всех экспериментах по оценке мер сравнивались словоформы
- Может быть, надо сравнивать что-то другое?
 - не униграммы, а биграммы?
 - не слова, а символы?
 - ...

Что лучше сравнивать: эксперимент

- Материал: 20 подкорпусов BNC
- Мера: χ^2 на топ-400 элементах
- Уровни анализа:
 - символы
 - словоформы
 - леммы
 - части речи (глагол, существительное, etc.)
 - части речи CLAWS5 (более дробные пометы)
- От 1-грамм до 3-грамм

Частотные списки разных уровней (BNC: British Medical Journal)

word1		char2		tag3	
,	0.046	e_	0.023	S S S	0.026
<i>the</i>	0.041	s_	0.020	S Prep S	0.025
.	0.028	_t	0.019	S S Pun	0.021
<i>to</i>	0.027	_a	0.015	Adj S S	0.020
<i>and</i>	0.025	th	0.014	Prep Art S	0.019
<i>of</i>	0.022	in	0.014	S Pun S	0.018
<i>a</i>	0.015	t_	0.013	Art Adj S	0.017
<i>in</i>	0.014	er	0.012	Adj S Prep	0.017
...

Худшие результаты

		char1	char2	char3	word1	word2	word3	lemma1	lemma2	lemma3	tag1	tag2	tag3	claws1	claws2	claws3
184	mir tod	93.64	92.42	69.39	84.55	68.33	75.91	86.36	85.61	51.97	66.67	56.97	60.76	66.97	58.33	69.39
185	gua sco	62.58	74.85	74.09	76.82	86.21	56.82	68.18	71.82	79.09	56.82	66.97	81.52	53.94	76.21	68.94
186	gua tel	78.03	74.7	77.12	64.09	71.82	81.97	71.67	75	82.27	61.36	47.58	46.82	64.24	69.85	72.73
187	ind sco	71.06	89.09	67.42	62.88	76.82	73.48	69.7	71.06	56.97	58.03	57.12	79.24	52.27	81.06	59.85
188	sco tel	79.55	75.91	75.45	77.58	66.67	75.61	66.36	79.09	74.09	60.61	51.97	66.21	53.33	61.52	50.3
189	ind tel	70.45	69.24	73.79	80.3	75.15	59.55	57.73	77.42	84.7	63.79	66.67	56.97	54.7	58.03	64.85
190	gua ind	85.61	67.58	45	53.48	64.85	65.91	55.3	56.06	57.12	36.67	40.3	67.12	54.24	53.48	70.91

Лучшие результаты

		char1	char2	char3	word1	word2	word3	lemma1	lemma2	lemma3	tag1	tag2	tag3	claws1	claws2	claws3
1	bio new	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
2	new uni	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
3	bio uni	100	100	100	100	100	99.85	100	100	100	100	100	100	100	100	100
4	bel bio	100	100	100	100	100	100	100	100	100	100	100	99.7	100	100	100
5	bio mir	100	100	100	99.55	100	100	100	100	99.85	100	100	100	100	99.85	100
6	bio eco	100	100	100	100	100	99.85	99.39	100	99.85	100	100	100	100	100	100
7	bio han	100	100	100	99.85	100	99.09	100	100	99.85	100	100	100	100	100	100

Сравнение уровней анализа

Unit of analysis	Ngrams	Median	Mean	Standard deviation	Highest score (% cases)
Characters	2	99.39	97.16	5.31	43%
Characters	3	96.62	96.62	7.01	39%
Lemmata	1	99.09	96.03	7.12	16%
Characters	1	98.94	96.43	5.93	36%
Words	1	98.94	96.10	6.71	22%
CLAWS5	2	98.86	94.61	9.02	21%
PoS	3	98.79	95.10	8.44	22%
Lemmata	2	98.33	95.79	6.59	12%

Литература

- Kilgarriff, Adam. 2001. Comparing corpora. *International Journal of Corpus Linguistics* 6(1). 97–133. doi:10.1075/ijcl.6.1.05kil.
- Kilgarriff, Adam. 2009. Simple maths for keywords. *Proceedings of Corpus Linguistics Conference CL2009*. University of Liverpool, UK. <https://www.sketchengine.co.uk/wp-content/uploads/2015/04/2009-Simple-maths-for-keywords.pdf>
- Oakes, Michael P. 2009. Corpus linguistics and language variation. In Paul Baker (ed.), *Contemporary corpus linguistics*, 159–183. London: Continuum.
- Ляшевская, Ольга Николаевна & Сергей Александрович Шаров. 2009. *Частотный словарь современного русского языка*. Москва: Азбуковник.
- Шайкевич, Анатолий Янович. 2015. Меры лексического сходства частотных словарей. *Труды международной конференции «Корпусная лингвистика–2015»*, 434–442. Санкт-Петербург: СПбГУ, Филологический факультет.