Воронежский государственный университет,
Воронежский государственный университет
инженерных технологий

Предельный размер словаря писателя и фрактальная размерность его метакниги

А. А. Кретов, М.В. Половинкина,
И. П. Половинкин, М. В. Ломец

Фрактальная размерность текста метакниги и способ ее оценки

- Согласно закону Хипса, количество различных, уникальных слов лемм (N), как функция от общего количества слов в метакниге (M) имеет степенной порядок $\Theta\left(M^{\alpha}\right)$, где $\alpha \in (0,1)$
- Предлагается рассматривать закон Хипса не как асимптотическую оценку, а как точную формулу

$$\alpha = \alpha(M) = \ln N / \ln M.$$

Понятие фрактальной размерности

- Введем в пространстве R^d совокупность конгруэнтных множеств, имеющих топологическую размерность d. Пусть фрактальный объект находится в пространст R^d . Зафиксируем достаточно малый радиус *l*>0. Покроем целиком фрактальный объект шарами радиуса *l*. Предположим, что для этого потребовалось как минимум шаров *N=N(l).* Число $\alpha_0 = -\lim_{l\to 0}(\ln N / \ln l) = \lim_{l\to 0}(\ln N / \ln(1/l))$ называется фрактальной размерностью.

Понятие фрактальной размерности

- Мы не можем устремлять к нулю размер атомарного множества, которым естественно считать слово (словоупотребление). Придется его изменить: $l = 1/M$.
- По определению положим

$$\alpha_0 = -\lim_{l \to 0}(\ln N / \ln l) = \lim_{M \to +\infty}(\ln N / \ln M) = \lim_{M \to +\infty} \alpha(M),$$

- α_0 – назовем фрактальной размерностью текста
- Объем текста M может принимать большие значения, $M \to +\infty$

- Авторы утверждают и иллюстрируют примерами текстов трех разных авторов, что α убывает с возрастанием M.
- Предел убывающей на промежутке функции в правом конце равен точной нижней грани функции на этом промежутке. Поэтому значение при максимальном значении в заданном диапазоне и следует считать наилучшим приближением верхней оценки фрактальной размерности.
- Нижняя оценка фрактальной размерности метакниги: на основе эмпирических данных произведем аппроксимацию функции. Пользуясь полученной зависимостью определим величину метакниги, при превышении которой приращение величины словаря будет пренебрежимо мало.

Коэффициент лексического разнообразия (КЛР)

- Это количественная характеристика текста, отражающая степень богатства словаря при построении текста заданной длины.
- Вычисляется как отношение числа отдельных лексических единиц словаря (лемм, англ. types) к количеству их употреблений в тексте (словоформ, «текстовых слов», англ. tokens).
- Для такого способа вычисления принято обозначение TTR.
- Можно считать предельным размером словаря такое значение размера, при котором КЛР становится пренебрежимо малым.

Оценка фрактальной размерности метакниги Л.Н. Толстого

- Рассмотрено 20 произведений Льва Толстого разного объема и размера, охватывающие отрезок времени в 52 года.
- Совершено 19 шагов, на каждом из которых метакнига наращивалась посредством конкатенации текста произведения, вычислялся ее текущий размер, а также осуществлялись лемматизация, соответствующее наращивание словаря и вычисление его текущего размера.
- Верхняя оценка фрактальной размерности метакниги Л.Н. Толстого - 0,7252.

Фрагмент таблицы «Динамика КЛР в нарастающем корпусе текстов Л.Н. Толстого»

№	Год	Текст	Δ_N	Δ_M	N	M	Y_{TTR}
1	1852	Детство	4253	30326	4253	30326	0,1402
2	1854	Отрочество	1452	23020	5705	53346	0,1069
3	1855	Севастопольские рассказы	2117	36041	7822	89287	0,0875
4	1856	Два гусара	844	17219	8666	106606	0,0813
5	1856	Утро помещика	751	15669	9417	122275	0,0770
6	1857	Юность	1208	49939	10625	172214	0,0617
7	1858	Альберт	147	7927	10772	180141	0,0598
8	1862	Поликушка	725	16879	11497	197020	0,0584
9	1863	Казаки	1391	46002	1288	243022	0,0530

№	Год	Текст	Δ_N	Δ_M	N	M	Y_{TTR}
10	1869	Война и мир	7212	459672	20100	702694	0,0286
11	1877	Анна Каренина	2702	270110	22802	972804	0,0234
12	1884	Записки сумасшедшего	32	3729	22834	976533	0,0234
13	1886	Смерть Ивана Ильича	190	17716	23024	994249	0,0232
14	1889	Крейцерова соната	270	25434	23294	1019683	0,0228
15	1890	Дьявол	395	14246	23689	1033929	0,0229
16	1891	Мать	48	3597	23737	1037526	0,0229
17	1895	Хозяин и работник	268	14270	24005	1051796	0,0228
18	1898	Отец Сергий	153	13706	24158	1065502	0,0227
19	1899	Воскресение	1591	137305	25749	1202807	0,0214
20	1904	Хаджи-Мурат	481	36376	26230	1239183	0,0212

Динамика КЛР в нарастающем корпусе текстов Л.Н. Толстого

- Выберем в качестве линии тренда логарифмическую зависимость

  Коэффициент правдоподобия в таком случае высок $R^2 \approx 0{,}9611$

  Получаем уравнение:

$$Y_{TTR} = 0{,}4081 - 0{,}028 \ln \mathrm{M}$$

- Функция достигает нулевого значения в точке $M_0 \approx 2129565$
- Размер метакниги Л.Н. Толстого составляет примерно 2 129 600 слов.

Динамика КЛР в нарастающем корпусе текстов Л.Н. Толстого

- Предельный объем словаря найдем из тех же соображений при том же выборе базисных функций. Мы получим уравнение:

$$Y_{TTR} = 0{,}6301 - 0{,}0604 \ln M$$

- Эта функция достигает нулевого значения в точке $N_0 \approx 33932$
- Оценка предельного размера словаря Л.Н. Толстого составляет примерно 33 900 слов.

Закон Ципфа

- Для описания зависимости размера словаря от размера текста воспользуемся законом Ципфа: $N = AM^{\gamma}$
- Воспользуемся данными и аппроксимируем степенную функцию в законе Ципфа. Этот подход дает нам $N = 38{,}069M^{0{,}4653}$
- Подставим в эту формулу значение $M_0 \approx 2129565$ и получим $N_0 \approx 33506$
- Относительная погрешность составляет

$$\frac{33932 - 33506}{33506} \times 100\% \approx 1{,}27\%$$

- Предельный размер словаря Л.Н. Толстого составляет 33500 - 34000 слов.
- Размер текста, при котором достигается предельный размер словаря Л.Н. Толстого, составляет 2 129 500 - 2 130 000 слов.
- Вычислим нижнюю оценку фрактальной размерности метакниги Л.Н. Толстого:

$$\alpha_0 \approx \frac{\ln 34000}{\ln 2130000} \approx 0,71605$$

- Фрактальная размерность метакниги Л.Н. Толстого может быть заключена в промежуток [0,7160; 0,7252].

Оценка фрактальной размерности метакниги Ф.М. Достоевского

- Мы рассмотрели 17 произведений Ф.М. Достоевского
- Верхняя оценка фрактальной размерности метакниги Ф.М. Достоевского равняется 0,7190.
- Для нижней оценки понадобилась фиксация всех промежуточных пар значений после каждой конкатенации из таблицы.

Таблица «Динамика КЛР в нарастающем корпусе текстов Ф.М. Достоевского»

№	Год	Текст	Δ_N	Δ_M	N	M	Y_{TTR}
1	1846	Бедные люди	4798	42162	4798	42162	0,1138
2	1846	Двойник	2606	49342	7404	91504	0,0809
3	1847	Хозяйка	1260	23931	8664	115435	0,0751
4	1848	Белые ночи	447	17053	9111	132488	0,0688
5	1849	Неточка Незванова	1264	55290	10375	187778	0,0553
6	1859	Дядюшкин сон	1453	41538	11828	229316	0,0516
7	1859	Село Степанчиково	1693	66470	13521	295786	0,0457
8	1860	Записки из мертвого дома	2733	98476	16256	394262	0,0412

№	Год	Текст	Δ_N	Δ_M	N	M	Y_{TTR}
9	1861	Униженные и оскорбленные	1134	119692	17388	513954	0,0338
10	1862	Скверный анекдот	350	16414	17738	530368	0,0334
11	1864	Записки из подполья	611	35452	18349	531022	0,0346
12	1866	Игрок	537	44630	18886	575652	0,0328
13	1866	Преступление и наказание	2984	172635	21870	748287	0,0292
14	1869	Идиот	1657	209206	23527	957493	0,0246
15	1872	Бесы	1893	197329	25420	1154822	0,0220
16	1875	Подросток	1195	189590	26615	1344412	0,0198
17	1880	Братья Карамазовы	2822	297078	29437	1641490	0,0179

Динамика КЛР в нарастающем корпусе текстов Ф.М. Достоевского

- Выберем в качестве линии тренда логарифмическую зависимость и получим уравнение:

$$Y_{TTR} = 0{,}3603 - 0{,}025 \ln M$$

- Функция достигает нулевого значения в точке $M_0 \approx 1815733$
- Размер метакниги Ф.М. Достоевского составляет 1 815 700 слов.

Динамика КЛР в нарастающем корпусе текстов Ф.М. Достоевского

- Предельный объем словаря найдем из тех же соображений при том же выборе базисных функций. Мы получим уравнение:

$$Y_{TTR} = 0{,}5283 - 0{,}05 \ln \mathrm{M}$$

- Функция достигает нулевого значения в точке $N_0 \approx 38793$
- Оценка предельного размера словаря Ф.М. Достоевского составляет примерно 38 800 слов.

Закон Ципфа

- Воспользуемся первичными данными и аппроксимируем степенную функцию в законе Ципфа: $N = 27{,}484M^{0{,}4908}$
- Подставим в эту формулу значение $M_0 \approx 1815733$ и получим $N_0 \approx 32435$
- Относительная погрешность составляет

$$\frac{38793 - 32435}{32435} \times 100\% \approx 19{,}60\%$$

- Так как относительная погрешность весьма велика, нижняя оценка фрактальной размерности будет не очень точна. Придется выбрать вариант, который дает меньшее значение α_0 :

$$\alpha_0 \approx \frac{\ln 32435}{\ln 1815733} \approx 0,7207$$

- Фрактальная размерность метакниги Ф.М. Достоевского не превосходит 0,7191.

Спасибо за внимание!