

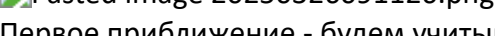
# Следующая пара - лабы, скипнешь - въебу

Идеи А.Н. Колмогорова:

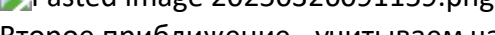
- Энтропия - мера сложности объекта
- Сложность объекта по Колмогорову - длина алгорита, реализованного машиной Тьюринга, описывающей объект

## Энтропия языка

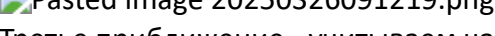
Нулвеое приближение - берём 32 буквы русского алфавита и пробел, поместим их в "ящик" и будем составлять их них текста так - перемешаем быквы, достанем одну, запишем, положим обратно и перемашаем ещё раз. Итог - что-то вроде этого:



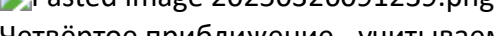
Первое приближение - будем учитывать частоты каждой из букв



Второе приближение - учитываем частоты диаграмм



Третье приближение - учитываем частоты триграмм



Четвёртое приближение - учитываем частоты тетраграмм



## Роль вероятностных параметров слов для измерения содержащейся в тексте информации

Первое приближение - учтены частоты появления слов



Второе приближение - учитываются частоты сочетаний двух соседних слов



## Свойства энтропии языка

Энтропией можно обозначить меру сложности объекта, в том числе языка

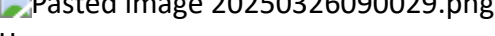
**Первое свойство энтропии языка** - Если энтропия языка равна  $H$ , то существует примерно  $2^{Hk}$  текстов длиной  $k$ , принадлежащих данному языку

**Второе свойство энтропии языка** - если энтропия языка равна  $H$ , то при оптимальном способе кодирования каждый текст языка удлинится в среднем в  $H$  раз

**Третье свойство** - (Пусть каждому тексту языка соответствует вероятность - вероятность события, что из всех мыслимых текстов заданной длины появится именно этот) Если энтропия языка равна  $H$ , то для подавляющего большинства текстов длины  $k$  такая вероятность равна  $2^{-Hk}$

**Энтропия русского языка** - по Колмогорову 1.33

Энтропия английского языка по Шеннону - 0.6-1.33



Чем меньше корпус допустимых текстов, тем меньше величина энтропии языка

## Тексты языка:

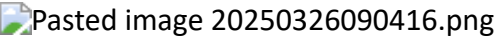
- Актуальные - реально существующие к данному моменту времени тексты на данным языке
- Потенциальные - все возможные тексты

Мы рассматриваем все тексты как потенциальные

## Остаточная энтропия

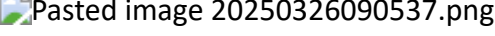
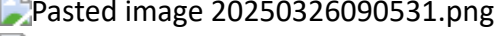
Пусть каждое предложение иностранного языка можно перевести на русский п способами (среднее количество переводов). Текст из 100 предложений можно перевести  $n^{100}$  способами

Остаточная энтропия по Колмогорову



Рассмотрим два языка - полный русский (энтропия A) и ограниченный русский с энтропией B,

Тогда примерные количество текстов длины k N1 & N2 будут равны



Если для заданного текста имеется N переводов длины k, то допустимых переводов этого текста должно быть в  $2^{ak}$  меньше

Чтобы допустимые переводы существовали, должно выполняться неравенство  $N \geq 2^{ak}$  или  $h \geq a$ , где h - остаточная энтропия

## Колмогоровская сложность

**Условная сложность** - сложность текста, вычисленная при условии, что указанные в тексте данные уже известны и могут быть использованы при составлении описаний

Условная сложность < Абсолютная сложность

**Удельная сложность** - сложность целого текста, поделённая на длину текста (сложность, в среднем приходящаяся на один знак)

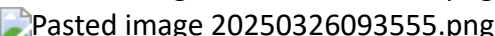
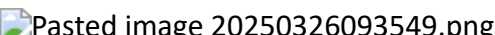
Удельная сложность < Энтропия языка | Для длинных текстов

## Три подхода к определению понятия "количество информации"

- Комбинаторный подход по Колмогорову:  
Энтропия переменного  $x$ :  $H(x) = \log_2 N$   
Указывая на определённое значение, энтропия снимается сообщением информации:  $I = \log_2 N$   
Интересен при кодировании информации
- Вероятностный подход по Колмогорову  
Колмогоров отмечает - придание переменным характера случайных переменных, обладающих совместным распределением вероятностей, позволяет получить значительно более богатую систему понятий и соотношений  
  
(x и y - это сообщения)  
При вероятностном подходе можно образовать матожидания  $MH_w(y/x)$  и  $MI_w(x:x)$   
Величина  $I_w(x,y) = MI_w(x:y) = MI_w(y:x)$  характеризует “тесноту связи” между x и y  
симметричным образом Колмогоров отмечает один парадокс:  
Величина  $I(x:y)$  при комбинаторном подходе всегда неотрицательна (что естественно), величина же  $I_w(x:y)$  может быть и отрицательной  
Подлинной мерой “количества информации” теперь становится усредненная величина  $I_w(x,y)$ 
  - Алгоритмический подход по Колмогорову  
  
Относительная сложность объекта y при заданном x - минимальная длина l(p) программы p получения y из x  
Метод программирования  $\phi(p,x) = y$ , где функцию  $\phi(p,x)$  считаем частично рекурсивной  
Для любой такой функции полагаем  $K_\phi(y/x) = \min I(p)$   
Если не существует p, удовлетворяющее  $\phi(p,x) = y$ , то  $K_\phi(y/x) = \infty$

## Основная теорема Колмогорова

Существует такая частично рекурсивная функции A(p,x), что для любой другой частично рекурсивной функции  $\phi(p,x)$  выполняется:  $K_A(y|x) \leq K_\phi(y|x) + C_\phi$ , где  $C_\phi$  не зависит от x и y



Если конечное множество M из очень большого числа элементов N допускает определение при помощи программы длины, пренебрежимо малой по сравнению с log2N, то почти все элементы множества M имеют сложность K(x), близкую к log2N. Элементы x ∈ M этой сложности и рассматриваются как случайные элементы множества M

## Теорема о симметрии взаимной информации

Соотношение H(x:y)=H(y:x) сохраняется при замене H на K и знака равенства на знак "приблизительно равно"  
 $K(x:y) = K(y:x) + O(\log \max(|\nu_1|, ..., |\nu_n|))$ , где  $\nu_1, ..., \nu_n$  суть компоненты кортежей x и y.

## Теорема об относительном описании

Для любых двух слов a и b существует программа p, которая преобразует a в b, при этом имеет минимальную возможную длину (то есть ее длина равна K(b|a)) и при этом p имеет очень малую сложность относительно b. Другими словами, можно вычислить некоторое «хэш-значение» b длины K(b|a), которого достаточно для восстановления b при заданном слове a

## Практическое использование колмогоровской сложности

Математически близость текстов S и T характеризуется неотрицательным числом - расстоянием

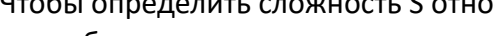
Расстояние d должно удовлетворять трем условиям:

- Расстояние неотрицательно d(S, T) ≥ 0, если d(S, T) = 0 → S = T
- Расстояние не меняется от перестановки текстов d(S, T) = d(T, S)
- Для любых трех текстов S, N, T выполняется неравенство треугольника d(S, T) ≤ d(S, N) + d(N, T)

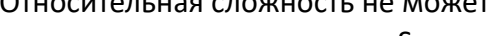
Сжатый файл - набор инструкций для разжимающей программы, который позволяет без потерь восстановить исходный текст. Минимальное количество информации, необходимое для восстановления текста - колмогоровская сложность

Колмогоровская сложность текста T - K(T)

Чтобы определить сложность S относительно T, нужно “подклеить” S к концу T и посмотреть, насколько хорошо эта добавка сжимается

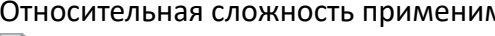


Относительная сложность не может служить метрикой, поскольку нарушаются условия 2 и 3: Условие 2 часто нарушается, если взять текст S маленькой длины и текст T большой:



Если для тех же текстов взять среднее арифметическое или геометрическое, то получится нечто симметричное, но неудовлетворяющее условию 3.

Относительная сложность применима в классификации текстов по автору. Метрика расстояния тогда имеет вид:



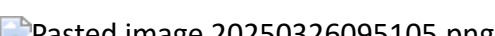
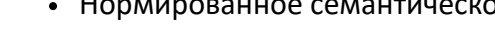
## Метрика схожести

Колмогоровская сложность невычислима для конечных объектов, но её можно заменить длиной сжатых через gzip/GenCompress объектов

Эта идея получила развитие в статье "The Similarity Metric"

Идеи из статьи выше нашли развитие в поисковой машине Google. Пример:

- Общее количество проиндексированных страниц - 8'058'044'651
- horse - 46'700'000 ссылок
- rider - 12'200'000 ссылок
- horse rider - 2'630'000 ссылок
- Нормированное семантическое расстояние NGD(horse, rider) = 0.443



NCD - будем искать на лабах