Практическая работа #1

Энтропия языка

26 февраля 2024 г.

1 Введение

Понятие энтропии используется во многих областях науки. Впервые термин был введен в рамках термодинамики, но он используется также в разделах физики, статистики, теории управления и т.д. Понятие информационной энтропии в математике было определено Клодом Шенноном. В 1948 году, исследуя проблему рациональной передачи информации через зашумлённый коммуникационный канал, К. Шеннон предложил революционный вероятностный подход к пониманию коммуникаций и создал первую, истинно математическую, теорию энтропии. Его сенсационные идеи быстро послужили основой разработки теории информации, которая использует понятие вероятности.

2 Задача

В данной работе требуется провести оценку энтропии текста на естественном языке. Нужно вычислить ряд последовательных приближений $F_0, F_1, F_2, \ldots, F_n$ к H, как к пределу, которые учитывают все большее число и более тонкие статистические закономерности языка и показать, что с увеличением числа учитываемых символов H уменьшается.

В качестве представляемого результата работы оформить отчет, в котором указать результаты следующих исследований и прикрепить исходный код программы.

- Построить гистрограммы частотности встречаемости последовательностей из 1, 2, 3 букв и слов текстов на русском и английском языках.
- Проанализировать поведение частотных компонент на гистограммах.
- Произвести расчет последовательных приближений $F_0, F_1, F_2, \ldots, F_n$ к H для 2-3 текстов на русском и английском языках, проанализировать полученные результаты.

Тексты должны содержать достаточный объем статистики для анализа статистических особенностей в языке. Например, в качестве текстов могут быть выбраны электронные книги.

3 Энтропия языка

Энтропия есть статистический параметр, который измеряет в известном смысле среднее количество информации, приходящейся на одну букву языкового текста. Если данный язык перевести на язык двоичных знаков (0 или 1) наиболее эффективным образом, то энтропия H равна среднему числу двоичных знаков (бит), приходящихся на одну букву исходного языка.

Существует метод для оценки этих количеств, учитывающий длительные статистические связи, влияние отдельных фраз друг на друга и т. д. Этот метод основан на изучении возможности предсказания английского текста: насколько точно может быть предсказана очередная буква, когда известны предыдущие N букв текста.

По одному из методов вычисления энтропии задается ряд последовательных приближений значений N-граммной энтропии $F_0, F_1, F_2, \ldots, F_n$ к истинной энтропии H, как к пределу, которые учитывают все большее число и более тонкие статистические закономерности языка. Приближение F_N может быть названо N-граммной энтропией; она измеряет количество информации, или энтропию, с учетом статистических связей не длиннее, чем на N следующих друг за другом букв текста. F_N дается формулой $\mathrm{F}_N = -\sum_{i,j} p(b_i, j) \log_2 p(j|b_i) = -\sum_{i,j} p(b_i, j) \log_2 p(b_i, j) + \sum_i p(b_i) \log_2 p(b_i)$, в которой b_i — блок из $N-1$ буквы [$(N-1)$-грамма]; j — произвольная буква, следующая за b_i; $p(b_i, j)$ — вероятность N-граммы $[b_i, j]$; $p(j|b_i)$ — условная вероятность буквы j следовать за блоком b_i, равная $\frac{p(b_i,j)}{p(b_i)}$.

Соотношение (1) можно интерпретировать как формулу для вычисления средней неопределенности (условной энтропии) последующей буквы j, когда известны предыдущие $N-1$ букв. При возрастании N в величине F_N учитываются все более и более далекие статистические связи и энтропия H является предельным значением F_N при $N \to \infty$,

$$H = \lim_{N\to\infty} F_N. \tag{1}$$

N-граммная энтропия для малых значений может быть подсчитана из обычных частотных таблиц отдельных букв, двухбуквенных (диграмм) и трехбуквенных сочетаний (триграмм). Если промежутком между буквами и знаками препинания пренебречь, то получим 27-буквенный алфавит и F_0 может быть взята (по определению) равной $\log_2 27$ или 4,7 бита на букву. Для вычисления F_1 в тексте подсчитывается число появлений каждого символа из заданного алфавита, и каждое из полученных значений делится на общее число символов в тексте, тем самым получая частоту символов в тексте или вероятность их появления $p(i)$. На основе полученных вероятностей может быть вычисленно F_1:

$$F_1 = -\sum_{i=1}^{26} p(i) \log_2 p(i). \tag{2}$$

Согласно выражению (3) диграммное приближение F_2 может быть вычисленно следующим образом: вычисляются частоты появления в тексте всевозможных пар из символов алфавита, получая значения $p(i,j)$, а также частоты появления отдельных символов $p(i)$, после чего F_2 получается, подставляя значения в (3).

$$F_2 = -\sum_{i,j} p(i,j) \log_2 p(j|i) = -\sum_{i,j} p(i,j) \log_2 p(i,j) + \sum_i p(i) \log_2 p(i). \quad (3)$$

Вычисление $F_3, F_4 \ldots$ выполнятеся подобным образом.