

Как считать слова?

Дмитрий Пионтковский, Максим Прасолов, Григорий Рыбников

1 Главная задача

Задача 1. В словаре племени Винни-Пухов 100 слов. В фразах их языка возможны любые сочетания этих слов. Существуют два магических заклинания, “Земля стоит на великом крокодиле” и “Каждый вечер крокодил глотает солнце”, которые вызывают ураган, и поэтому вслух можно произносить только такие фразы, в которых эти последовательности слов не встречаются¹. Сколько всего фраз из двадцати слов можно произносить вслух?

Задача 2. У компьютера есть 256 различных команд. Существует одна последовательность из четырёх команд, после которой компьютер ломается. Программисты написали все возможные программы из семи команд. Сколько процентов из них не ломают компьютер?²

Задача 3 (Главная Задача). Алфавит некоторого языка L состоит из N букв. Задано несколько слов v_1, \dots, v_k , которые называются запретными и в языке не употребляются. Слово (то есть ограниченная последовательность букв) называется допустимым, если никакая часть этого слова не является запретным словом. Сколько в языке L возможно допустимых слов из n букв?

Задача 4. Докажите, что задачи 1 и 2 сводятся к задаче 3.

2 Как записывать ответ?

Зафиксируем какой-нибудь алфавит A из N букв (например, если $A = (a, b, c, \dots, z)$, то $N = 26$). Словом мы будем называть любую конечную последовательность букв алфавита A . Подсловом мы будем называть часть слова, состоящую из идущих подряд в этом слове букв.

Мы будем считать, что в каждом языке L есть ровно одно слово из нуля букв — пустое слово.

Мы будем считать, что запретные слова не содержатся друг в друге, т.е. никакое подслово запретного слова, кроме него самого, не является снова запретным. Кроме того, мы будем считать, что запретные слова состоят как минимум из двух букв, т.е. что пустое слово и отдельные буквы являются допустимыми словами. Напомним, что множество запретных слов конечно.

Задача 5. Свободным языком алфавита A называется язык F_A , в котором вообще нет запретных слов. Докажите, что количество слов из n букв в этом языке равно N^n .

Задача 6. В языке B запретными являются все слова из двух различных букв. Докажите, что для любого натурального n количество допустимых слов из n букв в этом языке равно N .

Пусть M — какое-нибудь множество слов. Обозначим через m_n количество слов в этом множестве, состоящих из n букв. Рядом размеров множества M называется бесконечная сумма

$$M(x) = m_0 + m_1x + m_2x^2 + m_3x^3 + \dots$$

Такого вида бесконечные суммы (с произвольными числами в качестве коэффициентов m_n) мы будем кратко называть просто рядами (их полное название, которое мы не будем использовать — формальные степенные ряды).

Для каждого языка L его рядом размеров $L(x)$ называется ряд размеров множества допустимых слов. Например, для свободного языка F_A ряд размеров — это сумма геометрической прогрессии $F_A(x) = 1 + Nx + N^2x^2 + N^3x^3 + \dots$, а для языка B это $B(x) = 1 + Nx + Nx^2 + Nx^3 + \dots$

Задача 7. Выпишите ряд размеров для языка над алфавитом $\{a, b\}$, в котором запретными являются слова aa и bb .

¹Даже если слова в других словарных формах.

²Подобная история в 1990-е гг произошла с первой версией микропроцессора *Pentium*.

3 Арифметика языков

Если множество M содержит только ограниченное количество слов, то его ряд размеров — это многочлен от переменной x . Для бесконечных множеств и ряды тоже бесконечные, но с ними можно производить разные арифметические операции, похожие на операции с многочленами, то есть складывать, вычитать, умножать друг на друга и на числа и даже иногда делить.

В определениях и задачах этого раздела $S = s_0 + s_1x + s_2x^2 + \dots$ и $R = r_0 + r_1x + r_2x^2 + \dots$ — два ряда, а L_1 и L_2 — какие-то два языка с разными алфавитами A_1 и A_2 . Для определённости, мы будем считать, что алфавит A_1 состоит из заглавных букв, а алфавит A_2 — из строчных. Алфавит A — это объединение двух алфавитов A_1 и A_2 , то есть в него входят и заглавные, и строчные буквы.

Определение 1. а) *Суммой рядов R и S* называется сумма

$$R + S = (s_0 + r_0) + (s_1 + r_1)x + (s_2 + r_2)x^2 + \dots$$

б) *Суммой языков L_1 и L_2* называется язык $L_1 + L_2$ над алфавитом A , у которого множество допустимых слов — объединение множеств допустимых слов языков L_1 и L_2 .

Задача 8. *Задайте язык $L_1 + L_2$ конечным множеством запретных слов.*

Задача 9. *Докажите, что если $L = L_1 + L_2$, то*

$$L(x) = L_1(x) + L_2(x) - 1.$$

Произведение рядов размеров определяется так же, как произведение многочленов.

Определение 2. *Произведением ряда R на одночлен ax^n* называется ряд

$$R \cdot ax^n = ar_0x^n + ar_1x^{n+1} + ar_2x^2x^{n+2} + \dots$$

Произведением рядов R и S называется сумма

$$R \cdot S = R \cdot s_0 + R \cdot s_1x + R \cdot s_2x^2 + \dots$$

Заметим, что эта бесконечная сумма рядов имеет смысл — складывая ряды почленно, мы получаем в качестве коэффициента при каждой степени x конечную сумму чисел.

Задача 10. *Докажите равенство*

$$(1 - x) \cdot (1 + x + x^2 + \dots) = 1.$$

Определение 3. *Произведением двух множеств слов M, N* называется множество MN всех слов вида tn , где t — слово из M , а n — слово из N .

Произведением двух языков L_1 и L_2 называется язык $L_1 \cdot L_2$ над алфавитом A , у которого множество допустимых слов является произведением множеств допустимых слов языков L_1 и L_2 .

Задача 11. *Задайте язык $L_1 \cdot L_2$ конечным множеством запретных слов.*

Задача 12. *Докажите равенство*

$$L(x) = L_1(x) \cdot L_2(x).$$

Деление рядов не имеет аналога для языков, но позволяет сокращённо записывать их ряды размеров. Оно определяется по формуле, похожей на формулу суммы геометрической прогрессии.

Определение 4. Предположим, что ряд S начинается с единицы, то есть $s_0 = 1$, и $S = 1 + \bar{S}$, где $\bar{S} = s_1x + s_2x^2 + \dots$. Тогда *обратным рядом* называется ряд

$$\frac{1}{\bar{S}} = 1 - \bar{S} + \bar{S}^2 - \bar{S}^3 + \dots$$

Частным от деления рядов R и S называется ряд

$$\frac{R}{\bar{S}} = R - R \cdot \bar{S} + R \cdot \bar{S}^2 - R \cdot \bar{S}^3 + \dots$$

Частное двух рядов размеров языков может не соответствовать никакому языку, хотя бы потому, что в получившемся ряде могут появиться отрицательные коэффициенты.

Задача 13. а) *Докажите, что*

$$S \cdot \frac{R}{\bar{S}} = R.$$

б) *Докажите, что если $S \cdot T = R$, где ряд S начинается с единицы, то $T = \frac{R}{\bar{S}}$.*

Полезьа от операции деления рядов состоит в том, что многие бесконечные ряды можно записать с её помощью в виде конечного выражения — частного двух многочленов.

Задача 14. а) Докажите, что

$$F_A(x) = \frac{1}{1 - Nx}.$$

б) Запишите ряды размеров языков из задач 6 и 7 в виде частного двух многочленов.

Задача 15. Докажите, что ряд размеров любого языка может быть записан в виде частного двух многочленов.

Таким образом, ответ к Главной Задаче должен быть представлен в виде частного двух многочленов.

4 Свободное слово

Задача 16. Пусть L — язык над латинским алфавитом, в котором запретным является только слово “mouse”. Найдите $L(x)$.

Определение 5. Пусть a, b — два слова, из которых ни одно не является частью другого. Непустое слово c называется *зацеплением* слов a и b , если оно является окончанием слова a и в то же время началом слова b (например, слово “ко” — зацепление слов “молоко” и “корова”).

Слово называется *свободным*, если у него нет никаких зацеплений с самим собой, кроме самого этого слова.

Задача 17. Пусть в языке L над алфавитом A из N букв имеется только одно запретное слово — некоторое свободное слово из t букв. Докажите, что

$$L(x) = \frac{1}{1 - Nx + x^t}.$$

Задача 18. Решите задачу 2 в предположении, что последовательность, ломающая компьютер, является свободным словом.

5 Преобразования слов

Определение 6. Пусть M и M' — два множества слов. Разобьём множество M на какие-нибудь две части K и L . Функция f из множества L в какое-нибудь подмножество I множества M' называется *преобразованием* множества M в множество M' , если f сохраняет длину слова и является взаимно однозначным отображением из L в I .

В этом случае множество K называется *ядром* отображения f , а множество I — его *образом*.

Преобразование мы будем обозначать стрелкой: $M \implies M'$.

Определение 7. Цепочка преобразований

$$M_1 \implies M_2 \implies \dots \implies M_n$$

называется *точной*, если ядро каждого следующего преобразования совпадает с образом предыдущего.

Задача 19. Пусть L — язык над алфавитом A с множеством допустимых слов G и множеством всех не допустимых слов N . Постройте точную последовательность преобразований

$$\emptyset \implies N \implies F_A \implies G \implies \emptyset,$$

где F_A — множество слов свободного языка, то есть всех слов над алфавитом A , а \emptyset обозначает пустое множество.

Задача 20. В ряд сидят по очереди мальчики и девочки, по 10 тех и других; последней сидит учительница. У детей есть конфеты, поровну в сумме у мальчиков и у девочек. Первый мальчик отдаёт все свои конфеты сидящей за ним девочке. Девочка съедает их, съедает из своих конфет столько же, а остаток отдаёт следующему мальчику. Тот тоже поступает точно так же, за ним — следующая девочка, и так далее. Последняя девочка отдаёт остаток своих конфет учительнице. Сколько ей достанется?

Задача 21. Пусть

$$\emptyset \implies M_1 \implies M_2 \implies \dots \implies M_n \implies \emptyset$$

— точная цепочка преобразований.

а) Докажите, что если в каждом множестве M_i только конечное число m_i слов, то

$$m_1 + m_3 + m_5 + \dots = m_2 + m_4 + \dots$$

б) Докажите формулу для рядов размеров

$$M_1(x) + M_3(x) + M_5(x) + \dots = M_2(x) + M_4(x) + \dots$$

Определение 8. Множество слов M называется *свободным*, если никакое слово из M не является подсловом другого слова из этого множества, все слова в нём свободны и не имеют зацеплений между собой.

Задача 22. Пусть L — язык над алфавитом A , у которого множество запретных слов B свободное. Обозначим через G множество его допустимых слов, через \bar{G} — множество всех допустимых слов, кроме пустого. Постройте точную последовательность преобразований

$$\emptyset \implies B \cdot G \implies A \cdot G \implies \bar{G} \implies \emptyset.$$

Задача 23. Пусть L — язык над алфавитом A из N букв, у которого множество запретных слов B свободное. Докажите формулу

$$L(x) = \frac{1}{1 - Nx + B(x)}.$$

Задача 24. Докажите, что множество заклиний в задаче 1 свободное, и решите её.

Задача 25. Найдите $L(x)$, если алфавит языка L — латинский, а запретные слова — это слова *veni, vidi, vici*.

Определение 9. Пусть L — язык. *Простой сцепкой* называется слово $v = str$, где s, t, r — непустые слова, причём $g = st$ и $f = tr$ — запретные слова, и больше никаких запретных подслов в v нет. Конец r простой сцепки, остающийся после первого запретного слова g , называется её *хвостом*.

Задача 26. Докажите, что множество запретных слов языка является свободным в том и только том случае, если в этом языке нет простых сцепок.

Задача 27. Пусть L — язык над некоторым алфавитом A с множеством запретных слов B и множеством простых сцепок S . Обозначим через G множество его допустимых слов, через \bar{G} — множество всех допустимых слов, кроме пустого. Постройте точную последовательность преобразований

$$S \cdot G \implies B \cdot G \implies A \cdot G \implies \bar{G} \implies \emptyset.$$

Задача 28. Каким условиям должно удовлетворять множество запретных слов языка L , чтобы точную последовательность из задачи 27 можно было продолжить до последовательности

$$\emptyset \implies S \cdot G \implies B \cdot G \implies A \cdot G \implies \bar{G} \implies \emptyset$$

(такие языки назовём *незапутанными*)? Выведите формулу, которая выражала бы ряд размеров $L(z)$ *незапутанного* языка через число N букв в алфавите и ряды размеров множеств B и S .

Задача 29. Вычислите ряд размеров для языка над алфавитом из трёх букв a, b, c с запретными словами abb, bbc, bac .

Задача 30. Вычислите ряд размеров для языка над алфавитом $A = \{x_1, \dots, x_n, y_1, \dots, y_n, z_1, \dots, z_n\}$, в котором запретными являются все слова вида $x_i y_j$ и $y_j z_k$, где $1 \leq i, j, k \leq n$.

Задача 31. Докажите, если множество запретных слов *незапутанного* языка состоит только из одного слова, то это множество свободно.

6 Ещё о свободных множествах

Задача 32. Постройте бесконечное свободное множество в алфавите из двух букв.

Задача 33. Предположим, что множество запретных слов B языка L свободно, а алфавит содержит более одной буквы. Докажите, что множество допустимых слов этого языка бесконечно.

Определение 10. Пусть $S = s_0 + s_1x + s_2x^2 + \dots$ и $R = r_0 + r_1x + r_2x^2 + \dots$ — два ряда. Если для любых коэффициентов s_k и r_k с одинаковыми номерами выполняется неравенство $s_k \geq r_k$, то будем говорить, что между рядами выполняется неравенство

$$S \geq R.$$

Задача 34. Докажите, что если для рядов P, Q и R выполняются неравенства

$$P \geq Q \text{ и } R \geq 0,$$

то

$$PR \geq QR.$$

Задача 35. Предположим, что множества запретных слов B и B' двух языков L и L' над одним алфавитом A содержат одно и то же количество слов каждой длины, так что $B(z) = B'(z)$. Докажите, что если множество B свободно, то выполняется неравенство

$$L'(z) \geq L(z),$$

причём равенство $L'(z) = L(z)$ достигается в том и только том случае, когда множество B' также является свободным.

Задача 36. Известно, что алфавит состоит из двух букв, а множество B содержит не менее двух слов, одно из которых, слово w , имеет длину 2.

- a) Докажите, что множество B не является свободным.
- b) Может ли оно быть свободным, если длина слова w равна 3?

Задача 37. Известно, что алфавит состоит из n букв, а множество B состоит из g слов длины 2. Докажите, что если $g \leq n^2/4$, то множество B может быть выбрано свободным.

Задача 38. Докажите, что если $n = kd$ и $m \leq k^d(d-1)^{d-1}$, где числа d, k, m, n натуральные, то над алфавитом из n букв можно выбрать свободное множество, состоящее из m слов длины d .

Задача 39. a) Докажите, что если B — свободное множество над алфавитом из n букв, то выполняется неравенство

$$\frac{1}{1 - nx + B(x)} \geq 1.$$

b) Верно ли обратное утверждение: если для некоторого натурального n и некоторого многочлена $p(x)$ с неотрицательными целыми коэффициентами и нулевым свободным членом выполняется неравенство

$$\frac{1}{1 - nx + p(x)} \geq 1,$$

то над алфавитом из n букв существует свободное множество B такое, что $B(x) = p(x)$?

Задача 40. Пусть n — натуральное число, $p(x)$ — многочлен с неотрицательными целыми коэффициентами и нулевым свободным членом. Докажите, что свободное множество B с рядом размеров $B(x) = p(x)$ существует в том и только том случае, когда существуют такие многочлены f и g с неотрицательными целыми коэффициентами такие без свободных членов, что

$$(1 - f)(1 - g) \geq 1 - nx + p(x).$$

Задача 41.³ Придумайте условие, описывающее возможные ряды размеров множеств запрещённых слов незапутанных языков (подобно тому, как в задаче 40 охарактеризованы ряды размеров свободных множеств).

7 Слова и цепи

Определение 11. Пусть L — язык. Цепями длины 1 называются все запретные слова, цепями длины 2 — все простые сцепки. Через эти цепи определяются ещё цепи длины 3, 4 и так далее. А именно, слово $v = str$ (где все слова s, t, r — непустые) называется *цепью длины n* , если его начало $g = st$ является цепью длины $n-1$, конец $f = tr$ — запретным словом, причём t является подсловом хвоста p цепи g , и никаких запретных подслов, кроме f , в конечном участке pr нет. *Хвостом* этой цепи называется слово r .

³Жюри не известны ни решение, ни даже ответ к этой задаче

Цепь выглядит примерно так (каждая дуга — это запрещенное слово в ней):



Длина цепи — это количество дуг. Зацепляются только соседние дуги (т. е. их пересечение — зацепление ненулевой длины), и выделенные два последних хвоста не содержат запрещенных слов, кроме последней дуги.

Например, если запрещенное слово — aba , то единственная цепь длины 1 — это aba , длины 2 — $ababa$, длины 3 — $abababa$, и так далее.

Задача 42. Пусть в языке L запрещенными считаются слова “tournament”, “of”, “towns”. Выпишите все цепи длины n .

Задача 43. Антицепь длины n определяется так же, как и цепь длины n , но все слова языка L в определении 11 прочитываются “справа налево”, т. е. хвосты антицепей находятся слева, а начальная цепь длины $n - 1$ — справа. Докажите, что множества цепей длины n и антицепей длины n совпадают.

Задача 44. Докажите, что никакая цепь длины n не содержит в качестве подслова никакую другую цепь длины n .

Задача 45. Докажите, что если слово имеет вид $w = gc$, где g — допустимое слово, а c — цепь, то в случае, если длина цепи c больше 1, слово w представимо в таком виде ровно двумя способами, причём длины цепей в этих представлениях различаются на единицу.

Следующая задача даёт способ решения Главной Задачи.

Задача 46. Пусть L — язык над алфавитом A . Обозначим через G множество его допустимых слов, через \bar{G} — множество всех допустимых слов, кроме пустого. Пусть C_1 — множество цепей длины 1, C_2 — цепей длины 2, и так далее.

Докажите формулу

$$L(x) = \frac{1}{1 - Nx + C_1(x) - C_2(x) + C_3(x) - \dots}$$

Задача 47. Найдите ряд размеров для языка из задачи 42.

Задача 48. Найдите все возможные варианты ответов для задачи 2 в зависимости от того, какие именно команды ломают компьютер.

Задача 49. Назовём подслово s слова w максимальной подцепью, если w представимо в виде $w = gsc$, где g — допустимое слово, а c — цепь, причём для любого другого представления $w = gc'u'$ с другой цепью c' всегда слово c' — подслово слова s . Докажите, что любое недопустимое слово имеет ровно одну максимальную подцепь нечётной длины.

Задача 50. Пусть L — язык над алфавитом A , и A' — новый алфавит, полученный из A добавлением одной буквы. Пусть L' — язык над алфавитом A' , в котором запрещенными являются все запрещенные слова языка L . Докажите формулу

$$L'(x) = \frac{1}{\frac{1}{L(x)} - x}$$

Задача 51. Язык W называется свободным произведением языков L и L' над непересекающимися алфавитами A и A' , если алфавит языка W есть объединение алфавитов A и A' , а множество запрещенных слов — объединение множеств запрещенных слов языков L и L' . Выразите ряд размеров свободного произведения W через ряды размеров языков L и L' .

Задача 52. Предположим, что множество запрещенных слов языка L содержит только слова из двух букв. Рассмотрим другой язык M над тем же алфавитом, в котором запрещенными являются те и только те двухбуквенные слова, которые не являются запрещенными в языке L . Докажите равенство

$$L(x)M(-x) = 1.$$

8 Дополнительные задачи

Задача 53. Докажите, что свободное множество из m слов длины d в алфавите из $n = kd$ букв существует в том и только том случае, когда $m \leq k^d(d - 1)^{d-1}$ (ср. задачу 38), если

а) $d = 2$; б) $d = 3$; в) $d > 3$.

Определение 12. Язык называется d -определённым, если наибольшая из длин его запретных слов равна d . 2-определённый язык называется *квадратичным*.

Задача 54. Квадратичные языки L и M из задачи 52 называются двойственными друг к другу (обозначение: $M = L^!$).

- Докажите, что $(L^!)^! = L$.
- Найдите $(L_1 + L_2)^!$.
- Опишите $(L_1 \cdot L_2)^!$.

Задача 55. Пусть L — d -определённый язык. Определим новый язык $L^{(n)}$, в котором алфавитом являются все допустимые слова языка L длины n , а допустимые слова — все допустимые слова языка L , длина которых делится на n (выраженные через новые буквы).

- Докажите, что язык $L^{(n)}$ задаётся конечным набором запретных слов.
- Всегда ли язык $L^{(n)}$ является d -определённым?
- При каком наименьшем n язык $L^{(n)}$ гарантированно является квадратичным или свободным (вне зависимости от выбора d -определённого языка L)?

Задача 56. Для любого квадратичного языка L над алфавитом x_1, \dots, x_n определим ориентированный граф Γ_L следующим образом: его вершины — n точек, помеченные буквами x_1, \dots, x_n , а ребро (стрелка) $x_i \rightarrow x_j$ проводится в том и только том случае, когда $x_i x_j$ — разрешённое слово. Обозначим через a_k количество допустимых слов из k букв. Докажите, что

- язык L конечный в том и только том случае, когда в графе Γ_L нет циклов;
- язык L имеет полиномиальный рост (т. е. существуют два ненулевых многочлена p, q одной и той же степени d с положительным старшим коэффициентом такие, что $p(k) \geq a_k \geq q(k)$ для всех $k \geq 0$) в том и только том случае, когда в графе Γ_L есть цикл, но нет пересекающихся циклов;
- язык L имеет экспоненциальный рост (т. е. для некоторых $c_1 > c_2 > 1$ и для всех k выполняются неравенства $c_1^k \geq a_k \geq c_2^k$) тогда и только тогда, когда в графе Γ_L есть хотя бы два пересекающихся цикла.

Задача 57. Пусть L и $L^!$ — пара двойственных квадратичных языков. Возможно ли, что оба они имеют экспоненциальный рост?

Задача 58. Для любого d -определённого языка L над алфавитом x_1, \dots, x_n определим ориентированный граф Γ_L следующим образом: его вершины помечены всеми допустимыми словами длины $d - 1$, а ребро (стрелка) $v \rightarrow w$ проводится в том и только том случае, когда при умножении слова v на некоторую букву x_i получается допустимое слово, последние $d - 1$ букв которого составляют слово w . Докажите все три свойства а), б), в) из задачи 56 для построенного графа Γ_L .

Определение 13. Пусть M — некоторое множество слов над алфавитом A . Слова u и v (над тем же алфавитом) называются M -эквивалентными, если для любого слова w слова uw и vw либо оба принадлежат M , либо оба не принадлежат M . Множество M называется *регулярным*, если найдётся такое натуральное число n , что в любом множестве из n слов найдутся два M -эквивалентных друг другу слова.

Задача 59. Докажите, что множество допустимых слов любого языка регулярно.

Определение 14. Конечным автоматом над алфавитом A называется ориентированный граф Γ с конечным множеством вершин V , причём

- стрелки помечены буквами алфавита A , причём для любой буквы $a \in A$ из каждой вершины выходит ровно одна стрелка, помеченная a ;
- выделены начальная вершина $v_0 \in V$ и множество принимающих вершин $W \subseteq V$.

Будем воспринимать каждое слово над алфавитом A как инструкцию для путешествия по стрелкам конечного автомата (Γ, v_0, W) : начинаем с начальной вершины, идём из неё по стрелке, помеченной первой буквой слова, дальше идём по стрелке, помеченной второй буквой, и т.д. Мы говорим, что автомат *принимает* слово, если соответствующий слову путь заканчивается в принимающей вершине.

Задача 60. а) Докажите, что для любого регулярного множества M существует конечный автомат, принимающий слова из M и никаких больше.

- Докажите, что для любого конечного автомата множество принимаемых им слов регулярно.

Задача 61. Докажите, что для любого регулярного множества M его ряд размеров может быть записан в виде частного двух многочленов.

Задача 62. Пусть M_w — множество всех допустимых слов языка L , оканчивающихся на фиксированное подслово w . Докажите, что ряд размеров множества M_w представим в виде частного двух многочленов.

(До промежуточного финиша были предложены части 1–5, после промежуточного финиша добавлены части 6–8.)