Домашний алгоритм разбиения на слова (с картинками)

Артамонов Павел

В этой статье я расскажу и покажу свой способ сегментации строк на слова. Если вам не интересна жизнь сибиряка в тропиках, можете смело пропускать вступление.

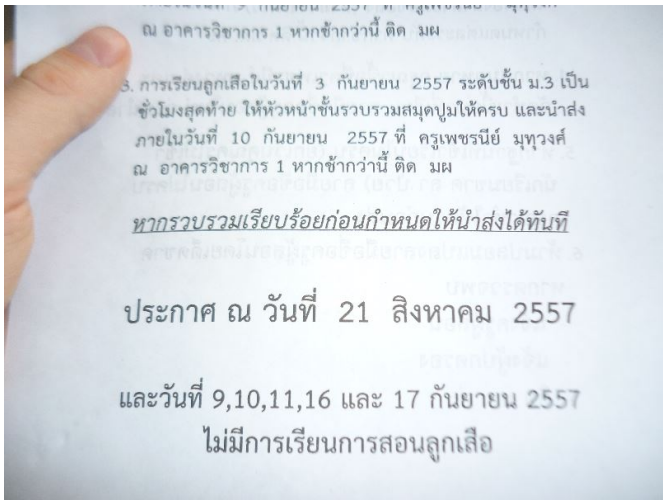
Вступление

Год назад я работал учителем математики в Таиланде. Учил юных джедаев силе абстракции. Учить детей — это очень интересно и приносит много удовольствия. А вот администрация приносит только головную боль. В тайских школах присутствует строгая иерархия: во-первых, по заслуге лет, во-вторых, по доступу к информации. Тот, кто её создает, находится на вершине, за ним идут верные распространители, заканчивая теми, кто информацию получает в последний момент, а уж если у тебя нет информации, то ты никто.

Угадайте кто в этой иерархии находится выше: учителя-иностранцы или пятиклассники? Представьте себе следующее: Пятница, конец рабочего дня, стопка тетрадей для проверки на выходных. Заходит ваш любимый директор и сообщает, что через два часа мы всей школой едем в соседний город на экскурсию на три дня. Занавес.

Это жизнь в полной информационной блокаде. При этом общедоступная информация, та которая развешивается по всей школе, вам тоже не доступна, она ведь на тайском. Бывало придешь на работу, а школа закрыта. На следующий день выясняешь у

директора, как же так, никто ничего не сообщил. А на это директор удивленно разводит руками, все же знали, мол, приказы-то вот же на стене.



Справка: *тайская письменность сильно отличается от письменности других языков, потому что она была независимо создана относительно недавно. Тайский содержит 44 символа согласных, а также 15 символов гласных, которые могут соединяться между собой в 28 разных форм. Гласные могут ставиться слева, справа, сверху и снизу согласной. И еще дополнительно 4 тоновых знака, тон — это что-то вроде интонации, с которой слог должен произноситься. Есть слова, которые отличаются друг от друга только тонами и имеют абсолютно разные значения. Пробелы между словами ставятся редко.*

Знакомых европейских символов в тайском нет и поэтому научиться читать очень трудно. Мне предстояло решить проблему чтения и перевода. Поиграв немножко с распознанием символов и написав небольшую программку, мне удалось распознать символы в строке. Получилась длиннющая строка, которую мне предстояло перевести. И тут я упёрся в стену. Пробелов то нет. Все слова склеены между собой! Для того чтобы их перевести, их нужно было отделить друг от друга.

Поиск по интернетам не принёс результата. Потому как обычные алгоритмы, такие как, например, собраны в этой хабр-статье, опирались на частоту слов. Чтобы собрать эту статистику, нужно было уметь разбивать на слова. А чтобы разбивать на слова, нужна была статистика. Вот такой замкнутый круг.

Пришлось выдумывать с нуля. То, что получилось, я вам продемонстрирую на строке русского языка (для простоты и наглядности).

Алгоритм

Итак, нам понадобится словарь (желательно, чтобы по нему можно было ходить по буквам). И пример-строка: “ядочиталэтодоконцанахабре”.

Первый шаг любого алгоритма — это визуализация.

Составим таблицу, где разместим все возможные варианты слов с учетом их длины. Слова “нах” у нас в словаре не будет, зато будет слово “хабр” и “хабре”.

я	я	яд					
д		до		дочи			дочитал
о	о		очи				
ч		чи	чит	чита	читал		
и	и						
т		та	тал				
а	а						
л							
э			это				
т		то					
о	о						
д		до	док				
о	о		око	окон			
к	к		кон		конца		
о	о	он					
н							
ц							
а	а						
н		на					
а	а	ах					
х		ха	хаб	хабр	хабре		
а	а						
б	б						
р		ре					
е							

Замечательно, теперь мы видим, что вариантов не так много и для программиста перебрать их все не составит труда.

Но как быть в такой ситуации: “яд о читал“ или “я до читал” или “я дочитал”? Какой из вариантов выбрать?

Для этого введем функцию оценки полученного разбиения. Будем выдавать очки за использованные буквы и слова. +3 очка за букву и -1 очко за слово. Чтобы было использовано как можно больше букв с как можно меньшим количеством слов.

И тогда:

	ядочитал	яд о читал	я до читал	я дочитал
Использовано букв из словаря (+3)	0	8	8	8
слов (-1)	0	3	3	2
очки	0	21 = 8*3 - 3	21 = 8*3 - 3	22 = 8*3 - 2

Победила желаемая строка! Осталось только придумать, как оптимальным образом всё перебрать и получить ОДИН единственный ответ. В слове “один” кроется разгадка нашего алгоритма.

Пусть мы нашли это конечное и единственное решение, тогда о каждой букве мы будем знать, начальная ли эта буква и, если да, то сколько букв в этом слове.

Что означает, что перебор должен проходить на уровне одной буквы, а не всех слов со всеми. Например, возьмём первую букву нашей строки “я”. И выберем между

“” 0 очков + x1 “я” 2 очка + x1 “яд” 5 очков + x2.

Где x1, x2 — предыдущие очки (отсчет с нуля).

Как это сделать ведь иксов у нас нет? Предыдущие очки нам известны только для последней буквы и они равны нулю. Значит нам нужно идти с конца!

Достроим два столбца, где будем хранить очки и длины используемых слов.

Будем складывать очки за слово с предыдущими очками (красные коробочки) и искать максимум.

длина	очки		2	5	8	11	14	17	20
		я	я	яд					
		д		до		дочи			дочитал
		о	о		очи				
		ч		чи	чит	чита	читал		
		и	и						
		т		та	тал				
		а	а						
		л							
		э			это				
		т		то					
		о	о						
		д		до	док				
		о	о		око	окон			
		к	к		кон		конца		
		о	о	он					
		н							
		ц							
		а	а						
		н		на					
		а	а	ах					
		х		ха	хаб	хабр	хабре		
		а	а						
		б	б						
		р		ре					
		е							

Ловкость рук и никакой рекурсии. Получено добротное решение без доступа к частоте слов. Алгоритм очень быстр и полностью справляется с поставленной задачей разбиения тайских строк. Его можно легко адаптировать под пропущенные буквы и недописанные слова, особенно если по словарю можно ходить по буквам.

О том как сделать такой словарь я расскажу в следующей статье. Я дочитал эту статью до конца на Хабре