

*Федеральное государственное автономное учреждение
высшего профессионального образования*

**Московский Физико-Технический Институт
КЛУБ ТЕХА ЛЕКЦИЙ**

**А л г о р и т м ы .
И В Т .**

III СЕМЕСТР

Лекторы: А. И. Гришутин. И. Д. Степанов.



*Автор: А. Ш. Ильдаров.
Проект на [github](#)*

Осень 2020 года

Содержание

1 Паросочетания. Минимальное вершинное покрытие.

Определение 1. Двудольный граф – граф, множество вершин которого можно разбить на две части таким образом, что каждое ребро графа соединяет какую-то вершину из одной части с какой-то вершиной другой части, то есть не существует рёбер между вершинами одной и той же части.

Определение 2. Хроматическое число – минимальное число цветов, в которые можно раскрасить вершины графа так, чтобы концы любого ребра имели разные цвета.

Определение 3. Паросочетание (англ. matching) в двудольном графе — произвольное множество рёбер двудольного графа, такое что никакие два ребра не имеют общей вершины.

Мощность паросочетания – количество ребер в нем.

Максимальное паросочетание – мощность которого *наибольшая* среди всех возможных паросочетаний в данном графе.

Определение 4. Цепь – некоторый простой путь (т.е. не содержащий повторяющихся вершин или рёбер).

Чередующаяся цепь – цепь, в которой рёбра поочередно принадлежат/не принадлежат паросочетанию.

Увеличивающая цепь – чередующаяся цепь, в которой начальное и конечное ребра *НЕ* принадлежат паросочетанию.

Уменьшающая цепь – цепь, в которой начальное и конечное ребра принадлежат паросочетанию.

Определение 5. Вершины двудольного графа, инцидентные рёбрам паросочетания M , называются **насыщенными** или **покрытыми**.

Алгоритм Куна.

Теорема 1.1. Паросочетание M в двудольном графе G – *так* \iff в G нет увеличивающей цепи относительно M .

Доказательство.

\Rightarrow :

От противного: Пусть в G с максимальным паросочетанием M существует увеличивающая цепь.

Тогда заменив в ней все рёбра, входящие в паросочетание, на невходящие и наоборот, мы получим большее паросочетание.

То есть M не являлось максимальным. Противоречие.

\Leftarrow :

Пусть M – не max, покажем что \exists увеличивающая цепь.

Пусть M' – паросочетание: $|M'| > |M|$

Построим подграф $G' = M \oplus M'$, состоящий из ребер \in только одному из паросочетаний.

M и M' – паросочетания \Rightarrow нет вершин, которые смежны с двумя ребрами из паросочетания. То есть у каждой вершины подграфа есть не более одного ребра из M и не более одного из M' . $\Rightarrow \forall v \in G' \hookrightarrow \deg(v) \leq 2$

Как известно, графы с таким свойством степеней вершин представляют из себя наборы цепей и циклов.

При этом длина цикла должна быть четной, ведь иначе мы будем иметь вершину, у которой два ребра, к ней смежных, принадлежат одному паросочетанию.

В циклах поровну ребер из каждого паросочетания, значит их вклад в отрыв M' от M по числу ребер – нулевой.

Значит обогнать M у M' получится только если в графе имеется цепочка нечетной длины, у которой начальное и конечное ребра лежат в M' . Вот эта вот цепочка и есть увеличивающая. \square

1.1 Алгоритм Куна для поиска максимального паросочетания.

1. Фиксируем доли графа: L и R. Изначально считаем паросочетание пустым.
2. В доле L перебираем вершины в порядке увеличения номеров.
3. Если вершина насыщена, то пропускаем ее и идем дальше. В противном случае, пытаемся насытить вершину, запустив поиск увеличивающей цепи из этой вершины следующим образом:
 - 3.1. Стоя в текущей вершине v доли L, посмотрим все ребра из этой вершины.
 - 3.2. Возьмем текущее ребро (v, t_0) : Если t_0 – ненасыщена, то одно ребро (v, t_0) и задает нам увеличивающую цепь, просто увеличим паросочетание с его помощью и прекратим поиск. Если же t_0 насыщена каким-то ребром (t_0, p) , то пойдем вдоль этого ребра, уже в поисках увеличивающейся цепи, проходящей через ребра (v, t_0) и (t_0, p) . Для этого просто перейдем в вершину p и продолжим обход из нее.

4. В конечном итоге, мы либо найдем увеличивающую цепь из вершины v и увеличим паросочетание этой цепью, тем самым насытив вершину, либо же покажем отсутствие увеличивающей цепи и невозможность насыщения вершины.
5. Возьмем следующую по порядку вершину доли L и повторим.
6. После того как мы просмотрим все вершины, попытаемся увеличить паросочетание цепью из них, мы получим максимальное паросочетание.

Минимальное вершинное покрытие.

Определение 6. Вершинное покрытие графа $G=(V,E)$ - такое подмножество S множества вершин графа V , что любое ребро этого графа инцидентно хотя бы одной вершине из множества S .

Определение 7. Минимальное вершинное покрытие графа – вершинное покрытие, состоящие из *наименьшего* числа вершин.

Утверждение 1.2. $|M_{max}| \leq |C_{min}|$. Чтобы покрыть все ребра, нам нужно вершин не меньше, чем мощность наибольшего паросочетания.

Доказательство.

Паросочетание это, как известно, набор непересекающихся рёбер. Мы хотим взять набор вершин, эти ребра покрывающий. Ясно что каждая вершина может покрыть не более одного ребра, ведь они не пересекаются по концам. Оценка снизу доказана. \square

Теорема 1.3. (Кёнига)

В двудольном графе $|M_{max}| = |C_{min}|$.

Доказательство. Явно предъявим покрытие, размер которого равен размеру максимального паросочетания.

1. Разделим граф на доли L и R .
2. Сориентируем ребра: Из $M_{max} \leftarrow$, остальные \rightarrow
3. Обойдем граф из ненасыщенных паросочетанием вершин $\in L$.
4. Получим разбиение графа на четыре множества: L^+, R^+ – вершины, лежащие в L или R соответственно, которые доступны из ненасыщенных вершин, лежащих в L . L^-, R^- – аналогично недоступные из ненасыщенных вершин, лежащих в L .

5. Поймем какие ребра бывают между каждыми из четырех множеств:

5.1. Покажем, что ребер $L^+ \rightarrow R^-$ не бывает, ведь если бы такое ребро имелось, мы из достижимой вершины, лежащей в L^+ добрались бы по этому ребру до вершины, по определению R^- , недостижимой.

5.2. Из аналогичных рассуждений понятно что не бывает ребер $L^- \rightarrow R^+$. Это не отрицает существование ребер $(R^+ \rightarrow L^-)$ тт

5.3. Покажем отсутствие ребер $R^- \rightarrow L^+$: От противного: Пусть (r^-, l^+) – такое ребро. Это ребро вида \leftarrow , то есть принадлежащее паросочетанию.

Вершина l^+ – насыщена, значит чтобы до нее дойти мы должны были начать обход из какой-то другой вершины l' , из которой l^+ достижима.

Так как l' и l^+ лежат в одной доле двудольного графа, маршрут из l' в l^+ в какой-то момент должен пойти справа налево, то есть имеется ребро вида (r', l^+) , которое в силу своего направления также лежит в паросочетании. Значит из смежные ребра (r^-, l^+) и (r', l^+) лежат в одном паросочетании. Противоречие.

6. Получаем что любое ребро графа инцидентно или вершине из L^- или вершине из $R^+ \Rightarrow L^- \cup R^+$ – вершинное покрытие.

7. Покажем, что все вершины из $L^- \cup R^+$ насыщены ребрами из паросочетания:

Это так, ведь если бы в L^- были ненасыщенные вершины, мы бы запускали из них обход и попадали бы в L^+ , чего быть не может в силу отсутствия ребер $L^- \rightarrow R^+$ и если бы в R^+ была бы ненасыщенная вершина, то существовала бы увеличивающаяся цепь, в этой вершине заканчивающаяся, что противоречит максимальности паросочетания.

8. Как известно, не существует ребер $L^- \rightarrow R^+$, которые бы принадлежали паросочетанию, а значит каждому ребру паросочетания инцидентна ровно одна вершина из покрытия $L^- \cup R^+$, а значит $|M| = |L^- \cup R^+|$, и это вершинное покрытие является наименьшим из оценки.

□

2 Потоки

Определение 8. Сеть – ориентированный граф $G = (V, E)$, с двумя выделенными различными вершинами: s – исток и t – сток, а также с функцией $cap : E \rightarrow \mathbb{Z}_{>0}$.

Определение 9. Поток в сети – функция $f : V \times V \rightarrow \mathbb{Z}$:

1. $\forall u, v \in V \hookrightarrow f(u, v) \leq cap(u, v)$
2. $\forall v \in V \setminus \{s, t\} \hookrightarrow \sum_{\substack{u \\ (u,v) \in E}} f(u, v) = \sum_{\substack{w \\ (v,w) \in E}} f(v, w)$
3. $\forall u, v \in V \hookrightarrow f(u, v) = -f(v, u)$

Замечание. Свойства 2 и 3 можно объединить в следующем виде $\forall v \in V \hookrightarrow \sum_{u \in V} f(v, u) = 0$

Определение 10. Остаточная сеть G_f сети G с потоком f – сеть, у которой модифицируются пропускные способности: $cap_f(u, v) = cap(u, v) - f(u, v)$, и удаляются ребра с нулевой cap_f .

Определение 11. Величина потока $|f|$ – то сколько "вытекает" из s – сумма исходящих из s потоков.

Определение 12. Разрез сети G – пара подсетей (S, T) : $s \in S, t \in T, S \cap T = \emptyset, S \cup T = G$

$$cap(S, T) = \sum_{\substack{u \in S \\ v \in T}} cap(u, v)$$

$$f(S, T) = \sum_{\substack{u \in S \\ v \in T}} f(u, v)$$

Лемма 2.1. (S, T) – разрез $\Rightarrow |f| = f(S, T)$

Доказательство. На семинаре. □

Лемма 2.2. (S, T) – разрез $\Rightarrow f(S, T) \leq cap(S, T)$

Доказательство. $\forall u, v \in V \hookrightarrow f(u, v) \leq cap(u, v) \Rightarrow$ чтд. □

Теорема 2.1. (Форда – Фалкерсона) Следующие утверждения эквивалентны:

1. f – макс поток в G
2. в G_f нет пути из s в t
3. $|f| = cap$ некоторого разреза.

Доказательство.

1 \Rightarrow 2: От противного:

Пусть f – max, но в G_f есть путь $s \rightarrow t$.

Рассмотрим $x = \min(\text{cap}_f) > 0$ на этом пути.

Значит вдоль пути можно пустить x единиц потока \Rightarrow в исходной сети можно было пустить на x единиц больше $\Rightarrow f$ – не max.

2 \Rightarrow 3:

Пусть S – множество вершин, доступных из s в G_f , $t \notin S$

$T = V \setminus S$.

По определению (S, T) – разрез.

Покажем что $|f| = \text{cap}(S, T) = \sum_{\substack{u \in S \\ v \in T}} \text{cap}(u, v)$.

Рассмотрим ребро (u, v) : $u \in S, v \in T$

u – доступно в G_f , v – нет \Rightarrow оно удаляется в остаточной сети \Rightarrow

$f(u, v) = \text{cap}(u, v) \Rightarrow \sum_{\substack{u \in S \\ v \in V}} \text{cap}(u, v) = \sum_{\substack{u \in S \\ v \in V}} f(u, v) \Rightarrow$

$f(S, T) = |f| = \text{cap}(S, T)$.

3 \Rightarrow 1:

$\exists(S, T) : |f| = \text{cap}(S, T)$.

При этом $\forall(S, T)$ – разрез $\hookrightarrow f(S, T) \leq \text{cap}(S, T) \Rightarrow f$ – max. □

2.1 Алгоритм Форда – Фалкерсона для поиска максимального потока.

1. Изначально поток равен 0.

2. Пока в G_f есть путь из s в t :

2.1. $x = \min(\text{cap}_f \text{ на этом пути }) > 0$

2.2. Увеличим поток f на найденном пути на x и изменим остаточную сеть соответствующим образом.

$|f|$ увеличивается на целое положительное число и ограничен сверху, значит алгоритм закончится. Асимптотика – $O(\text{ans} \times |E|)$

2.2 Алгоритм Эдмондса – Карпа для поиска максимального потока.

Та же самая идея, только теперь вместо произвольного пути будем выбирать кратчайший путь по ребрам. По сути, DFS меняется на BFS.

1. Изначально поток равен 0.
2. Пока в G_f есть *минимальный* путь из s в t :
 - 2.1. $x = \min(\text{cap}_f \text{ на этом пути }) > 0$.
 - 2.2. увеличим поток f на x и изменим остаточную сеть соответствующим образом.

Алгоритм будет иметь асимптотику $O(|V| \times |E|^2)$

Докажем это.

Определение 13. $\text{dist}(u, v)$ – минимальное число ребер между вершинами.

Лемма 2.3. Пусть поток f' получается из потока f после одной итерации алгоритма Эдмондса – Карпа.

Пусть $\text{dist}'(u, v)$ – кратчайшее расстояние между вершинами в $G_{f'}$, тогда $\forall v \in V \setminus \{s, t\} \hookrightarrow \text{dist}'(s, v) \geq \text{dist}(s, v)$.

То есть с каждой итерацией расстояние не убывает.

Доказательство. От противного:

Пусть $\exists v \in V \setminus \{s, t\} : \text{dist}'(s, v) < \text{dist}(s, v)$ и при этом $\text{dist}(s, v)$ – минимальное возможное

Пусть u – предыдущая вершина перед v на кратчайшем пути из s в v в $G_{f'}$

Тогда $\text{dist}'(s, u) = \text{dist}'(s, v) - 1$ (по определению dist).

Также $\text{dist}'(s, u) \geq \text{dist}(s, u)$ ведь мы выбрали минимально удаленное v .

Рассмотрим 2 случая:

$(u, v) \in E_f$, тогда:

$$\text{dist}(s, v) \leq \text{dist}(s, u) + 1 \leq \text{dist}'(s, u) + 1 \leq \text{dist}'(s, v)$$

Но при этом $\text{dist}'(s, v) < \text{dist}(s, v)$

Противоречие.

$(u, v) \notin E_f$, но известно что $(u, v) \in E_{f'}$.

Появление ребра (u, v) после увеличения потока означает увеличение потока по обратному ребру (v, u) . Увеличение потока производится вдоль кратчайшего пути, а значит вершина

v – предыдущая перед u на пути из s в u

Это значит что $dist(s, v) = dist(s, u) - 1 \leq dist'(s, u) - 1 = dist'(s, v) - 2$, но ведь $dist'(s, v) < dist(s, v)$. Противоречие.

□

Определение 14. Насыщенное – ребро, вдоль которого идет поток, равный его capacity.

Лемма 2.4. Любое ребро сети насыщается не более $O(|V|)$ раз, то есть в алгоритме $O(|V||E|)$ итераций, то есть каждая итерация насыщает хоть 1 ребро.

Доказательство.

Рассмотрим ребро (u, v) в момент его насыщения.

Если оно насыщается, то оно лежит на кратчайшем пути от s до v (так работает наш алгоритм).

Значит $dist(s, u) + 1 = dist(s, v)$.

Теперь посмотрим, когда оно могло насытиться еще один раз:

Оно сначала должно перестать быть насыщенным, для этого нужно отменить поток вдоль (u, v) , то есть пустить поток вдоль (v, u)

Пусть $dist'$ – расстояние в момент, когда пускается поток вдоль (v, u) .

$dist'(s, v) + 1 = dist'(s, u)$ (Так как проталкивание происходит вдоль кратчайшего пути).

Со временем расстояния только увеличиваются, значит $dist'(s, v) \geq dist(s, v) + 1 = dist(s, u) + 2$

Таким образом от момента насыщения, до момента повторного насыщения расстояние $dist(s, u)$ должно вырасти по меньшей мере на 2.

Однако все расстояния ограничены $O(|V|)$, значит и насыщений может быть только $O(|V|)$.

□

3 Продложение потоков. Алгоритм Диница, теоремы Карзанова

3.1 Детали реализации.

```

1  struct Edge {
2      int from = 0,
3          to   = 0;
4
5      long long cap = 0;
6      long long flow = 0;
7  };
8
9  auto edges = vector<Edge>();
10 auto outbound = vector<vector<int>>>();
11 /* v-th element -- numbers of edges from v. */

```

Листинг 1: Удобная реализация ребер

Рядом с каждым ребром всегда есть обратное, изначально с нулевой capacity. Будем присваивать обратному ребру номер исходного, увеличенный на 1. К примеру номера $2n$ и $2n + 1$. Тогда при добавлении нового ребра в граф, мы будем добавлять его и сразу же добавлять обратное, а потом сохранять номера обоих ребер в массив `edges`.

Теперь если мы пускаем поток по ребру с номером i , то можем удобно уменьшить поток по противоположному ребру, как имеющее номер $i \oplus 1$

```

1  edges[i].flow += delta_flow;
2  edges[i ^ 1].flow -= delta_flow;

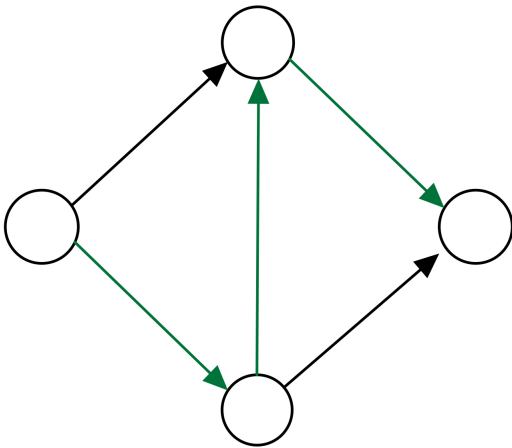
```

При реализации алгоритма Форда-Фалкерсона будем просто выполнять DFS на ребрах, у которых $\text{cap} > \text{flow}$.

3.2 Блокирующий поток.

Определение 15. Блокирующий – поток f в сети G , такой что в сети G (но вовсе не обязательно G_f) больше нет пути из s в t , вдоль которого можно протолкнуть поток.

Пример. Все ребра имеют $cap = 1$, зеленым отмечены насыщенные ребра.



Замечание. Блокирующий поток совершенно не обязательно наибольший.

3.3 Поиск блокирующего потока

За $O(nt)$

Забудем что у нас бывают обратные ребра.

Идея проста, будем пускать поток пока получается, а для оптимизации учтем тот факт, что если мы уже однажды пытались пойти вдоль ребра, и узнали что после прохода по нему дойти до t не получится, то нам не стоит идти по этому ребру снова. Это аналог метки `used` в DFS.

```

1  auto first_worthy_edge_num = vector<int>();
2  /*first_worthy_edge[v] -- relative number of the first edge
3     which is worth considiring from v. */
4
5  long long dfs(int v, long long flow){
6  //If we've followed the path and we can push the flow, we do it.
7
8     if (v == t) return flow;
9
10 //While there is at least one worthy edge.
11 while (first_worthy_edge_num[v] < outbound[v].size()){
12     auto cur_e = g[v][first_worthy_edge_num[v]];
13
14     /*If there is some additional flow we can push and the edge is
15        presented in initial network (is not reverse). */
16
17     //We will code edge presence checker in Dinic's algorithm.
18     if (edges[e].capacity > edges[e].flow && e is not reverse edge){
19         auto next_flow = dfs(edges[e].to,
20                               min(flow, edges[e].cap - edges[e].flow));
21
22
23         if (next_flow > 0){
24             edges[e].flow += next_flow;
25             edges[e ^ 1].flow -= next_flow;
26
27             return next_flow;
28         }
29     }
30
31     ++first_worthy_edge_num[v];
32 }
33
34 return 0;
35 }

```

Тогда сам поиск блокирующего потока будет иметь вид

```

1  while ((auto x = dfs(s, inf)) != 0){
2     blocking_flow += x;
3  }

```

Алгоритм работает за $O(nm)$, ведь если главный вызов dfs (из main) суммарно сдвинул все номера интересных вершин на k , то dfs работал $n + k$, так как каждое ребро либо нас устроило, и мы просто увеличили номер, либо оно нас устроило, и мы пошли в новую

вершину, а сдвигать указатели долго мы не сможем, суммарно сдвигов будет не больше чем m , а значит время работы $= n * \text{кол-во запусков} + m$, а каждый запуск насыщает хоть 1 ребро, при этом из-за отсутствия обратного ребра, ребра не насыщаются. Значит имеем асимптотику $O(nm)$

3.4 Алгоритм Диница для поиска наибольшего потока

За $O(n^2m)$

Пусть G – сеть.

$dist(s, v)$ – кратчайшее расстояние между вершинами.

Определим слоистую сеть:

Вершина s .

Вершины с $dist$ 1 от s .

...

Вершины на расстоянии k от s .

При этом оставим только ребра, идущие из меньшего уровня в следующий по порядку, то есть кратчайшие пути.

Алгоритм

Вспомним что у нас бывают обратные ребра.

Пока не найден наибольший поток:

1. Построим слоистую сеть.
2. Пустим в ней произвольный блокирующий поток.

Утверждение 3.1. *Алгоритм Диница находит наибольший поток.*

Доказательство.

Теорема Форда-Фалкерсона учит нас что поток максимален \iff в остаточной сети нет увеличивающего пути. Если наш алгоритм завершился, то есть не сумел пустить блокирующий поток в слоистой сети, значит в слоистой сети нет пути из s в t , а по построению слоистой сети, это значит что и в исходной сети пути из s в t . \square

Утверждение 3.2. *Алгоритм Диница делает не более n итераций.*

Доказательство.

Покажем что после каждой итерации $dist'(s, t) > dist(s, t)$.

На каждом пути из s в t есть хоть одно насыщенное ребро, ведь мы пустили блокирующий поток. (По определению блокирующего потока)

После того как мы пустили блокирующий поток, у в слоистой сети исчезли некоторые ребра слева направо, а обратные ребра справа налево наоборот появились, значит расстояние в слоистой сети увеличилось, а значит увеличилось и расстояние в исходной сети. \square

3.5 Теоремы Карзанова.

Зачастую асимптотика алгоритма Диница завышена, так как сеть в задаче имеет специфичный вид. Теоремы Карзанова помогают точнее оценить число итераций.

Определение 16. $C_{out}(v) = \sum_{u \in V} cap(v, u)$ $C_{in}(v) = \sum_{u \in V} cap(u, v)$

Замечание. Понятно что тогда через любую вершину течет не более чем $\min(C_{in}(v), C_{out}(v))$

Определение 17. **Общий потенциал сети** $P = \sum_{v \in V, v \neq s, t} \min(C_{out}(v), C_{in}(v))$

Лемма 3.1. В сети G $l = dist(s, t) \leq \frac{P}{F} + 1$, где F – максимальный поток, P – общий потенциал.

Доказательство.

Пусть $V_i = \{v : dist(s, v) = i\}$

Получилась слоистая сеть, в которой ребра есть только из V_i в V_{i+1}

Таким образом, если разрез и пересекает какое-либо ребро, то оно идет между слоями

А сумма capacity ребер между слоями не меньше чем максимальный поток.

$$cap(\cup_0^i V_k, \cup_{i+1}^l V_k) = \sum_{x \in V_i, y \in V_{i+1}} cap(x, y) \geq F$$

Просуммируем это неравенство по всем i :

$$\sum_{i=0}^{l-1} \sum_{x \in V_i, y \in V_{i+1}} cap(x, y) \geq (l-1)F$$

А сверху эту сумму можно оценить общим потенциалом сети, ведь сумма capacity ребер между слоями не может превышать сумму capacity ребер в одном слое, которая, в свою очередь, не превышает сумму потенциалов вершин в V_i .

$$P \geq (l-1)F$$

\square

Лемма 3.2. После проталкивания потока f в сети G , общий потенциал остаточной сети G_f равен общему потенциалу исходной сети.

Доказательство.

Когда мы проталкиваем поток f вдоль ребер (u, v) и (v, w) по обратным ребрам проходит $-f$.

Таким образом уменьшение $C_{in}(v)$ за счет потока по (u, v) будет скомпенсировано увеличением $C_{in}(v)$ за счет увеличения потока по обратному ребру (w, v) . \square

Теорема 3.3. Число итераций алгоритма Диница в сети G с целочисленными $capacity$ – $O(\sqrt{P})$

Доказательство.

Сделаем ровно \sqrt{P} итераций, тогда $l \geq \sqrt{P}$, ведь, как мы уже знаем, каждая итерация увеличивает расстояние между s и t хоть на 1.

Запишем утверждение первой леммы для остаточной сети после \sqrt{P} итераций:

$l \leq \frac{P}{F} + 1$, а $F = F - f$, где f – то сколько потока удалось протолкнуть за прошедшие итерации.

При этом $f \geq \sqrt{P}$, ведь все $capacity$ целочисленные, и мы не могли проталкивать меньше 1 единицы потока за итерацию.

Таким образом получаем $F \leq 2\sqrt{P}$ \square

4 Строки.

4.1 Совпадение подстрок в строке.

Построим **полиномиальную** хеш-функцию, которая каждой строке в соответствие ставит число.

$$h_{p,m}(\overline{c_0c_1\dots c_n}) = (c_0 \cdot p^{n-1} + c_1 \cdot p^{n-2} + \dots + c_{n-2} \cdot p + c_{n-1}) \% m$$

При помощи схемы Горнера, такая функция считается за линейное время, ведь $h_{p,m}(\overline{c_0c_1\dots c_n}) = (\dots(((c_0 \cdot p + c_1))p + c_2)p + c_3)\dots)p + c_{n-1}$

Так как мы имеем дело с подстроками одной большой строки, мы можем сделать предподсчет хеш-функции для всех префиксов большой строки, а дальше воспользоваться свойством хеша:

$$h(S[l, r]) = (h(S[0, r]) - h(S[0, l]) \cdot p^{r-l+1}) \% m.$$

Таким образом, после линейного предподсчета, мы можем сравнивать строки по их хешу за константное время.

4.2 Алгоритм Рабина – Карпа для поиска подстроки в строке.

Для поиска подстроки S длины m в тексте T длины n ,

1. Подсчитаем $h(T[0, m])$ и $h(S[0, m])$, а также p^m .
2. В цикле по всем i от 0 до $n - m$
 - 2.1. Считаем $h(T[i, m + i])$ и сравниваем его с $h(S[0, m])$.
 - 2.2. В случае равенства, опционально, можно провести проверку наивным посимвольным сравнением или выборочным сравнением символов, для исключения малейшей вероятности коллизии.

Алгоритм работает за линейное время $O(n + m)$, однако он не устойчив перед коллизиями хеш-функции, дополнительная обработка которых увеличивает время работы.

4.3 Алгоритм Кнута – Морриса – Пратта для поиска подстроки в строке.

4.3.1 Префикс-функция

Введем для строки S префикс-функцию π

$$\pi(i) = \max_{k \in [0, i]} \{k : S[0, k] == S[i + 1 - k, i + 1]\}$$

– длина наибольшего собственного суффикса подстроки $S[:i]$, который равен префиксу такой же длины. $\pi(0) = 0$ по определению.

Пример. $S = \text{"adam"}$. $\pi(0) = 0$.

$\pi(1) = 0$, ведь у строки "a" есть только пустой собственный суффикс.

$\pi(2) = 0$, ведь у строки "ab" нет собственного суффикса, совпадающего с префиксом той же длины.

$\pi(3) = 1$, ведь у строки "ada" есть собственный суффикс "a" длины 1, который совпадает с префиксом "a".

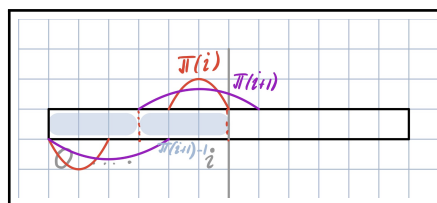
$\pi(4) = 0$.

Нетрудно представить себе наивный алгоритм, подсчитывающий префикс-функцию строки за $O(n^3)$. Оптимизируем его.

Утверждение 4.1. $\pi(i + 1) \leq \pi(i) + 1$

Доказательство. Рассмотрим суффикс, оканчивающийся на позиции $i + 1$ и имеющий длину $\pi(i + 1)$. Удалив из него последний элемент, мы получим суффикс, оканчивающийся на позиции i и имеющий длину $\pi(i + 1) - 1$

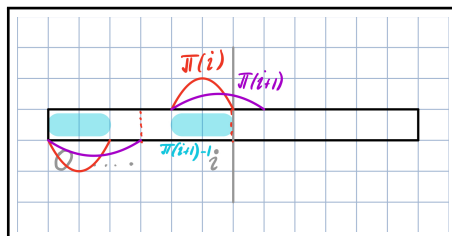
Однако, мы определили $\pi(i)$ максимальную длину суффикса $\Rightarrow \pi(i) \geq \pi(i + 1) - 1$.



□

Утверждение 4.2. Если $S[\pi(i)] == S[i + 1]$, то $\pi(i + 1) == \pi(i) + 1$.

Доказательство. Условие говорит нам, что после префикса длины $\pi(i)$ находится символ, который совпадает с символом, который находится после суффикса, оканчивающегося на позиции i . То есть наибольший суффикс, заканчивающийся на позиции $i + 1$ содержит в себе все символы суффикса, заканчивающегося на позиции i и еще один символ. Это и записано в правой части.



□

Пользуясь вторым свойством, мы сможем частично оптимизировать подсчет. Если же окажется что условие равенства символов не выполняется, мы будем искать суффикс меньшей длины следующим образом: Мы знаем, что $S[1, \pi(\pi(i))]$ – суффикс подстроки $S[1, i]$

Тогда подсчет можно свести к циклу

1. $k = \pi(i)$.
 - 1.1. Если $S[i + 1] == S[k]$, то $\pi(i + 1) = k + 1$.
 - 1.2. Если же $k == 0$, то $\pi(i + 1) = 0$.
2. $k = \pi(k)$. Перейти к пункту 1.

Теперь мы будем проходиться по строке, и для каждой позиции вычислять $\pi(i)$ оптимизированно.

Утверждение 4.3. *Оптимизированный алгоритм подсчета префикс-функции имеет асимптотику $O(n)$.*

Доказательство. Из вышедоказанного утверждения мы знаем что на каждой итерации по индексу строки, k увеличивается не более чем на 1. □

4.3.2 Z-функция

Введем для строки S Z-функцию z

$$z(i) = \max\{k : S[i, i+k) == S[0, k)\}$$

– длина наибольшей подстроки S , начинающейся с позиции i , которая равна префиксу S такой же длины, $z(0) = 0$ по определению.

Пример. $S = \text{"adam"}$. $z(0) = 0$. $z(1) = 0$. $z(2) = 1$. $z(3) = 0$.

Определение 18. Z-блок — подстрока S , начинающаяся с позиции i длины $z(i)$.

Для построения Z-функции строки S будем на каждом шаге хранить самый правый Z-блок – тот, у которого правая граница является самой правой. Назовем позиции начала и конца самого правого Z-блока l и r соответственно. Изначально $l = r = 0$.

Пусть у нас подсчитана Z-функция и Z-блок для всех позиций $< i$, тогда найдем $z(i)$.

Рассмотрим 2 случая.

1. $i \notin [l; r)$:

Тогда нам придется считать функцию наивно, пробегая двумя указателями по строке с начала и по подстроке, начиная с позиции i . Тогда значение $z(i)$ будет определено как первая позиция j , такая что $S[j] \neq S[i+j]$. После этого обновим сохраненное значение самого правого Z-блока.

2. $i \in [l, r)$:

Тут возможны 2 варианта:

2.1. $z(i-l) < r-l$. То есть наибольшая подстрока, начинающаяся с позиции $i-l$, равная префиксу такой же длины, не превышает по длине самый правый Z-блок.

Заметим, что позиция $i-l$ будет лежать в префиксе, равном по длине самый правый Z-блоку.

Тогда $z(i) = z(i-l)$.

2.2. $z(i-l) > r-l$. В этом случае наибольшая подстрока, начинающаяся с позиции $i-l$ выходит за пределы префикса, соответствующего самому правому Z-блоку.

Тогда скажем что $z(i) = r - i + l$ — длина той части наибольшей подстроки, начинающейся с позиции $i-l$, которая лежит внутри префикса, соответствующего самому правому Z-блоку. **А затем будем наивно улучшать оценку, проходясь по строке начиная с позиции i .**

Утверждение 4.4. *Асимптотика вычисления Z-функции — $O(m)$, где m — длина строки.*

Доказательство. Каждая позиция в строке просматривается не более двух раз. При попадании в промежуток $[l; r)$ и наивном улучшении, или же в противном случае при непопадании в промежуток и наивном просчете. При этом, всякий раз когда мы считаем Z-функцию наивно, мы обновляем самый правый Z-блок, то есть сдвигаем его правую границу, а это может происходить не более m раз. Значит мы тратим $O(m)$ на базовые операции и $O(m)$ на суммарные наивные проверки. \square

4.3.3 Алгоритм

Для поиска подстроки S длины m в тексте T :

1. Построим строки вида $S\#T$, где $\#$ — произвольный разделитель, не встречающийся в тексте.
2. По сконструированной строке построим префикс-функцию или Z-функцию.
3. Тогда если в некоторой позиции i $\pi(i) == m$, то эта позиция — конец вхождения подстроки.

Аналогично, если $z(i) == m$, то эта позиция — начало вхождения подстроки.

Утверждение 4.5. *Алгоритм работает за $(|S| + |T|)$.*

Доказательство. Ведь Префикс-функция и Z-функция строятся за линейное время, а затем совершается только один проход всей результирующей строки. \square

5 Продолжение строк. Алгоритм Ахо – Корасик для нахождения всех вхождений набора подстрок.

Мы хотим научиться искать все вхождения строк из набора в некотором тексте.

5.1 Построение бора.

Определение 19. Бор — дерево, в котором каждая вершина обозначает строку, а каждое ребро обозначает букву. Строка, соответствующая вершине (то есть заканчивающаяся в этой вершине), получается конкатенацией всех букв, соответствующих ребрам пути из корня в эту самую вершину. По определению, корню бора соответствует пустая строка.

Определение 20. Терминальная вершина бора для набора слов — вершина, которой соответствует слово из этого набора.

Для добавления строки в бор мы:

Прочитав очередной символ в цикле по строке, переходим по соответствующему ему ребру или создаем такое ребро если потребуется.

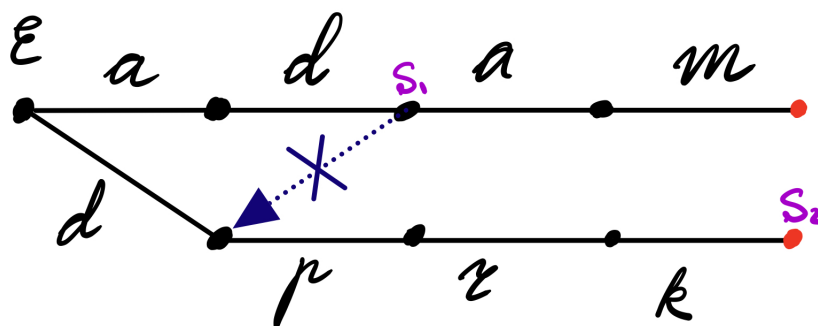
После завершения строки помечаем последнюю вершину как терминальную.

Таким образом, бор строится за линейное по сумме всех строк в наборе время.

5.2 Преобразование бора в автомат.

Будем понимать вершины бора и соответствующие им строки как состояния конечного детерминированного автомата. Однако мы сталкиваемся с проблемой, ребер бора не достаточно для отражения всех возможных переходов между состояниями автомата.

Пример. Ребра бора не отражают тот факт, что перейдя под воздействием символов "ad" в некоторое состояние S_1 мы все еще можем перейти в состояние S_2 .



Определение 21. Суффиксная ссылка вершины v — ссылка (мнимая стрелка в боре) на вершину u , такую, что состояние u — наибольший собственный суффикс состояния v , а если такой вершины u нет, то ссылка на корень. По определению, ссылка из корня ведет в корень.

Пример. Для бора строк "adam" и "dprk" единственной суффиксной ссылкой, не ведущей в корень, будет ссылка из состояния "ad", в состояние "d", зачеркнутая на рисунке выше.

5.2.1 Нахождение суффиксных ссылок.

Чтобы найти суффиксную ссылку для вершины v :

Если v — корень, то и его суффиксная ссылка тоже корень. В противном случае:

- Пусть c — буква, преобразующая из родителя в вершину v .
- Рассмотрим в качестве текущей вершины суффиксную ссылку родителя.
- Пока в рассматриваемой вершине нет ребра, соответствующего букве c , будем прыгать дальше, рассматривая ее суффиксную ссылку.
- Если мы уперлись в корень, то корень и является суффиксной ссылкой вершины v .
- Если мы нашли вершину u , с ребром c , то суффиксной ссылкой будет вершина, в которую ребро c переводит вершину u .

Концептуально, мы берем наибольший собственный суффикс родителя и откусываем от него по одной букве слева, пока не получится подстрока, которая добавлением одного символа становится собственным суффиксом нашей исходной строки.

```
1  auto get_suflink(Node_t* node){
2
3      if (node == root) return root;
4
5      auto last_c = node->c;
6      auto parent = node->parent;
7
8      while(parent != root && parent->go[last_c] == nullptr){
9          parent = get_suflink(parent);
10     }
11
12     if (parent == root) return root;
13     return parent->go[last_c];
14 }
```


Этот алгоритм терминируется, ведь на каждом шаге мы поднимаемся выше по бору, а значит рано или поздно дойдем до корня.

Мы считаем суффиксные ссылки рекурсивно, но этот процесс можно **оптимизировать, записав суффиксную ссылку в вершину**, ведь после построения автомата она меняться не будет. Тогда мы можем ввести универсальную автоматную функцию перехода, которая не делает различия между суффиксными ссылками.

$jump(node, c) =$

son_c , если ребро бора c переводит вершину $node$ в son_c

$null$, если $node$ — корень, у которого нет ребра бора c

$jump(suflink(node), c)$, иначе

С учетом концепции универсальности перехода, переход по суффиксной ссылке вершины — переход из суффиксной ссылки родителя по ребру c , где c — ребро, которое перевело родителя в вершину.

```
1  suflink(node) = jump(suflink(node->parent), node->c);
```

Тогда чтобы посчитать суффиксные ссылки мы пойдем по дереву по слоям, высчитывая `suflink` и `jump` одновременно:

1. Для корня все известно.
2. Для первого слоя `suflink` ведет в корень, а `jump` вниз или по `suflink` в корень и по `go` от него.
3. Для каждого из следующих уровней мы можем найти и `suflink` и `jump` за $O(1)$, ведь для родителя, который находится уровнем выше мы уже все посчитали.

Концептуально это BFS.

5.3 Поиск вхождений при помощи автомата.

Для поиска набора строк в тексте:

1. Построим автомат по набору строк.
2. Каждый раз будем считывать букву текста, и совершать переход по ней в автомате.

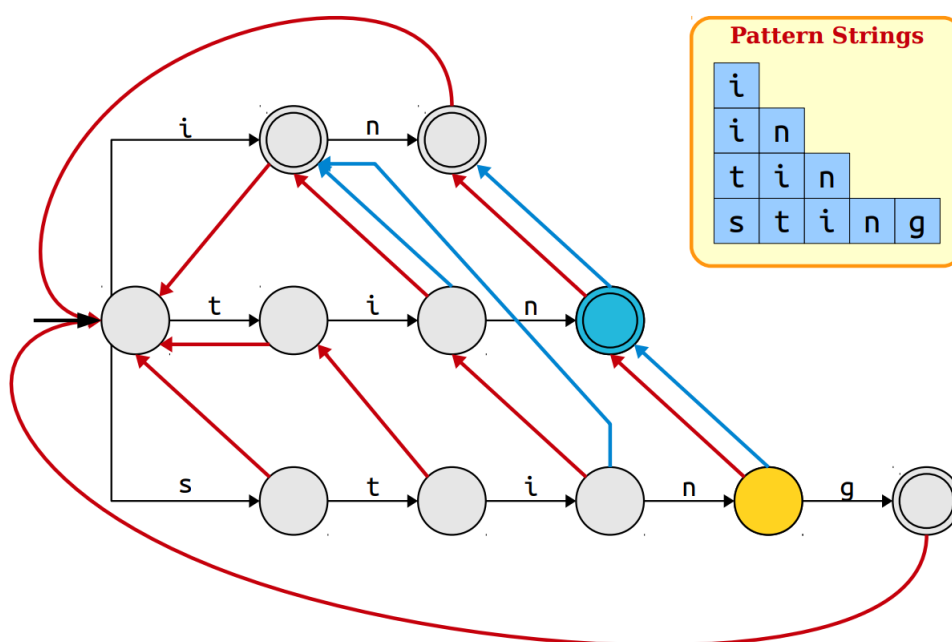
При этом, мы поймем, что нашли совпадение строки из набора, если текущая вершина терминальная, а также если, двигаясь по суффиксным ссылкам, мы можем достигнуть терминальной вершины.

Из-за этогоотягчающего обстоятельства, определение вхождения строки в текст для каждого состояния будет занимать линейное время. Оптимизируем это движение по суффиксным ссылкам.

5.4 Нахождение сжатых терминальных ссылок.

Определение 22. Сжатая терминальная ссылка вершины v — ближайшая достижимая по суффиксным ссылкам из v вершина w , являющаяся терминальной и не совпадающая с v .

Пример. Сжатые терминальные ссылки отмечены синим.



Концептуально, сжатые терминальные ссылки собирают в односвязный список все терминальные вершины, а значит и все подстроки, вхождение которых имело место при достижении текущего состояния. Таким образом, учитывая что сжатая терминальная ссылка ведет только наверх, мы сможем построить эти ссылки по слоям при помощи BFS.

```
1  if (suflink(node)->is_terminal){
2      complete_link(node) = suflink(node);
3  }
4
5  else {
6      complete_link(node) = complete_link(suflink(node));
7  }
```

5.5 Алгоритм с использованием сжатых терминальных ссылок.

1. Построим автомат по набору строк.
2. Построим сжатые терминальные ссылки в автомате.
3. Для каждого считанного символа c :
 - 3.1. $cur_node = jump(root, c)$
 - 3.2. Если $cur_node == null$, вхождение соответствует, идем дальше.
 - 3.3. В противном случае, засвидетельствуем вхождение подстроки, соответствующей текущей вершине, если она терминальная, а также всех ее сжатых терминальных ссылок.

5.6 Сложность алгоритма.

Пусть Σ — суммарная длина подстрок в наборе, T — длина текста, по которому осуществляется поиск, Ans — число найденных вхождений.

1. Построение бора требует $O(\Sigma)$
2. Построение суффиксных ссылок, по сути, BFS по дереву, требует $O(\Sigma)$
3. Построение сжатых терминальных ссылок, также BFS по дереву, требует $O(\Sigma)$
4. Цикл по каждому из символом — $O(T)$
5. Проходы по сжатым терминальным ссылкам суммарно займут $O(Ans)$

Итого, имеем $O(\Sigma + T + Ans)$