

Hledání vzorků v textu

Problém: Text $T[1 \dots n]$, vzorek $P[1 \dots m]$. Existuje i tak, že $T[i \dots i + m - 1] = P[1 \dots m]$?

- ▶ Jaké znáte způsoby řešení problému hledání vzorků v textu?

Hledání vzorků v textu

Uvažujme naivní algoritmus pro hledání vzorků v textu.

Předpokládejme, že všechny znaky v P jsou různé. Jak dokážete naivní algoritmus zrychlit?

Hledání vzorků v textu – varianty

Předpokládejme, že vzorek P může kromě běžných symbolů i speciální symbol $*$, který reprezentuje libovolný počet libovolných znaků (i žádný). Jak můžete upravit známé algoritmy tak, aby hledaly výskyt takovýchto rozšířených vzorků?

K zamyšlení (na doma): Co kdyby vzorek kromě $*$ obsahoval i speciální symbol $?$, který reprezentuje jeden libovolný znak?

Hledání vzorků v textu – konečné automaty

- ▶ Sestrojte konečný automat pro vzorek aabab.

Hledání vzorků v textu – konečné automaty

- ▶ Sestrojte konečný automat pro vzorek aabab.
- ▶ Jak vypadá konečný automat pro vzorek se speciálním symbolem *?

Hledání vzorků v textu – konečné automaty

- ▶ Sestrojte konečný automat pro vzorek aabab.
- ▶ Jak vypadá konečný automat pro vzorek se speciálním symbolem *?
- ▶ Označme $P_k = P[1 \dots k]$, tedy prvních k znaků vzorku P . Vzorek P je bez překryvů, pokud pro každé k, l platí, že je-li P_k sufixem P_l , pak buď $k = 0$ nebo $k = l$. Jak vypadá konečný automat pro vzorek bez překryvů?

Hledání vzorků v textu – konečné automaty

- ▶ Sestrojte konečný automat pro vzorek aabab.
- ▶ Jak vypadá konečný automat pro vzorek se speciálním symbolem *?
- ▶ Označme $P_k = P[1 \dots k]$, tedy prvních k znaků vzorku P . Vzorek P je bez překryvů, pokud pro každé k, l platí, že je-li P_k sufixem P_l , pak buď $k = 0$ nebo $k = l$. Jak vypadá konečný automat pro vzorek bez překryvů?
- ▶ Jak modifikovat algoritmus založený na konečných automatech tak, aby vypsal všechny výskyty vzorku v textu?

Hledání vzorků v textu – varianty

Co kdybychom měli více vzorků P_1, \dots, P_k ? Dokážete upravit některý známý algoritmus tak, aby uměl v textu efektivně hledat libovolný z těchto vzorků?

Poznámka: Pro zjednodušení můžete předpokládat, že všechny vzorky mají stejnou délku.

Související příklady

Cyklická rotace

Cyklickou rotací textu $T[1 \dots n]$ je text $T[i \dots n] \cdot T[1 \dots i - 1]$ pro nějaké i . Mějme na vstupu dva řetězce A, B stejné délky. Chceme rozhodnout, zda je jeden z nich cyklickou rotací druhého.

- ▶ Jak můžeme tento problém vyřešit pomocí hledání vzorku v textu?

Související příklady

Nejdelší palindromický prefix

Máme zadaný text $T[1 \dots n]$ a chceme najít jeho co nejdelší prefix $T[1 \dots i]$, který je zároveň palindromem (čte se stejně zepředu i zezadu).

- ▶ Jak souvisí tento problém s hledáním vzorku v textu?

Související příklady

Hledání podstromů

Mějme dva binární stromy P , T . Chceme najít (všechny) podstromy stromu T , které mají stejnou strukturu jako strom P .

- ▶ Jak souvisí tento problém s hledáním vzorku v textu?

