

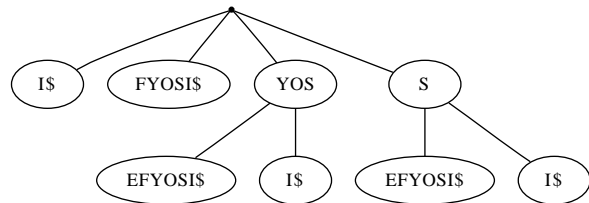
# מבני נתונים ואלגוריתמים - הרצאה 20

26 בינואר 2012

## עצי סיפא

לכל מחזורות  $S$  אפשר להגדיר רישא -  $S(0:k)$  וסיפא  $S(k:n)$ .  
 נרצה לשמור את רשימת כל הסיפות -  $S(k:n)$  עבור  $k = 0 \dots n$ .  
 אפשר להכניס את רשימת כל הסיפות במבנה שנקרא *TRIE*. כל אות היא קדקד ושתי אותיות עוקבות מחוברות.

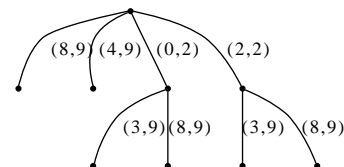
כדי להקל על עצמנו, נוסיף את הסימן \$ לכל סוף סיפא. כל סיפא מסתיימת ב\$.  
 עץ סיפא, Suffix Tree, הוא *Trie* כזה רק שכל ענף שאין בו פיצולים נסמן כקשת אחת.  
 הקדקדים ב*Trie* שאין בהם פיצולים והופכים לחלק מקשת אחת ב*Tree* נקראים Implicit Node.  
 למשל, עץ הסיפא של המחזורות YOSEFYOSI יהיה:



(זה רק חלק מהעץ, חסרים 2 ענפים).  
 יש לכל היותר  $2n$  קדקדים ויש בדיוק  $n$  עלים.  
 אם נסמן אינדקסים על המחזורות:

0	1	2	3	4	5	6	7	8	9
Y	O	S	E	F	Y	O	S	I	\$

אז אפשר לכתוב את העץ כאשר במקום המחזורות עצמן, נחזיק זוג מספרים,  $(a, b)$  שיסמל את המחזורות  $S(a:b)$ .  
 כעת העץ יהיה:



(שוב, זה רק אותו חלק מהעץ).

## אלגוריתם בניה נאיבי

אלגוריתם הבניה הנאיבי לבנות עץ סיפא ייקח  $O(n^2)$  פעולות.

## אלגוריתם Online לבניית עצי סיפא

---

אלגוריתם 1 אלגוריתם לבנית עץ סיפא

עבור על כל אות שנוספת

}

עבור כל אות, עבור על כל סיפא

}

לך לסוף הסיפא והוסף את האות

{

{

---

האלגוריתם הזה לוקח  $O(n^3)$  אך אפשר לשנות אותו קצת כדי שיהיה  $O(n)$  עם טריקים מסויימים. נשים לב שאם משהו הוא עלה, הוא ימשיך להיות עלה עד הסוף - הוא יגיע לסוף המילה. אנחנו מוסיפים רק פיצולים בעץ.

## שימוש לאלגוריתם - בדיקת העתקות

נניח שיש לנו שני טקסטים  $T_1$  ו  $T_2$ . אם אני חושב שהם העתיקו אחד מהשני, אזי "א שאני חושב שקיים Substring של  $T_1$  ושל  $T_2$  זהה וארוך. נוסיף לטקסט הראשון  $S_1$  ולשני  $S_2$ . נוסיף את כל הסיפות של  $T_1 S_1$  ושל  $T_2 S_2$  לאותו עץ סיפא: תחילה נבנה עץ סיפא של  $T_1 S_1$ . נסנה עץ סיפא של  $T_2 S_2$  כאשר העץ סיפא של  $T_1 S_1$  כבר בנוי. העלות של בניית העץ היא  $O(|T_1| + |T_2|)$ . כעת, נעבור על העץ ונבנה לכל קדקד פנימי וקטור  $v$  בגודל 2 שמכיל 1 במקום  $i$  אם  $S_i$  נמצא בתת העץ שמתחתיו. הקדקד הפנימי הכי עמוק שהוקטור שלו הוא  $(1, 1)$  הוא תת המחרוזת הכי ארוכה שמשותפת לשני הטקסטים, ואם הוא מאוד ארוך הטקסט בדר"כ מועתק. שיטה נוספת לבדוק אם העתיקו היא לשאול כמה פעמים מופיע רצף באורך לפחות  $L$  משותף לשני הטקסטים (אם למשל מוסיפים רווחים בין המילים). דרך נוספת היא: אם בוקטור  $v$  נשים במקום  $i$  את כמות הפעמים שמופיע  $S_i$  בתת העץ במתחתיו, אז אפשר לראות כמה פעמים  $(\min(v))$  מופיע רצף באורך לפחות עומק הקדקד בשני הטקסטים. האלגוריתם הוא:

1. חשב לכל קדקד פנימי את מס' המופעים של כל  $S_i$  בתת העץ שמתחתיו וסמן את הסכום ב  $\bar{v}$ .
2. תן לכל קדקד פנימי ערך של  $v_m = \min(\bar{v})$ .
3. עבור על כל הקדקדים הפנימיים שעמוקים או שווים ל  $L$  וסכום את  $v_m$ .