Biçimsel Diller ve Otomata Teorisi

Hafta 11:
İçerikten Bağımsız Diller (I. Bölüm)

Plan

1. Düzenli İfade Uygulamalari
2. Chomsky Normal Form
3. CKY Algoritmasi

Düzenli İfade Uygulamaları

Düzenli ifadeler, uçsuz bucaksız gen dizilimlerinde spesifik bir bölüm aranırken sıklıkla kullanılan bir yöntemdir.

ör. Kırılgan X Sendromu (Fragile X Syndrome)

Kırılgan X sendromu, genetik bir bozukluktur ve zeka geriliğine yol açar. Erkeklerde her 250 kadınlarda ise her 800 kisiden birinde bu hastaliğa neden olan gen bulunur.

Bu hastalığa neden olan DNA dizilimi şu şekildedir:

'gcg' nukleik asit üçlüsünün ardından 'cgg' veya 'agg' herhangi bir sayıda tekrarlar ve ardından 'ctg' nükleik asit üçlüsü gelir.

Bu D N A dizilimini şu düzenli ifade ile tanımlıyabiliriz:

$$g\ c\ g\ (c\ g\ g \cup a\ g\ g)^*\ c\ t\ g$$

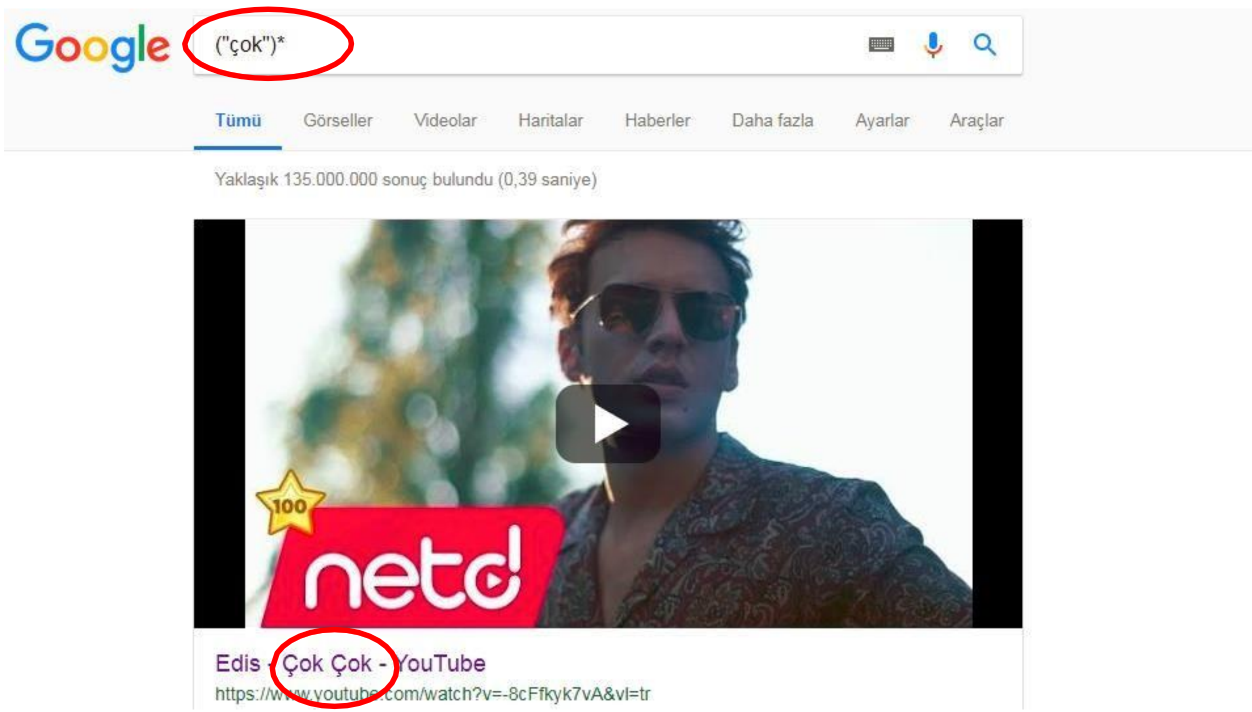
örneğin aşağıdaki D N A dizlimindeki bizim aradığımız bölüm:

gcggcgtgtgtgcgagagagtgggtttaaagctg**gcgcggaggcggctg**gcgcggaggctg

Google'da Duzenli Ifade Kullanimi

Google gibi arama motorlarinda da duzenli ifadeler kullanilir.

| birleşme operatoru, " " bitistirme operatoru, * yildiz operatorudur.



ör. google aramasi: " bir * olsam "

arama sonuclari: " bir marti olsam, " " bir kahve fincan olsam ", " bir bulut olsam ", " bir zengin olsam ", " bir fidan olsam " …

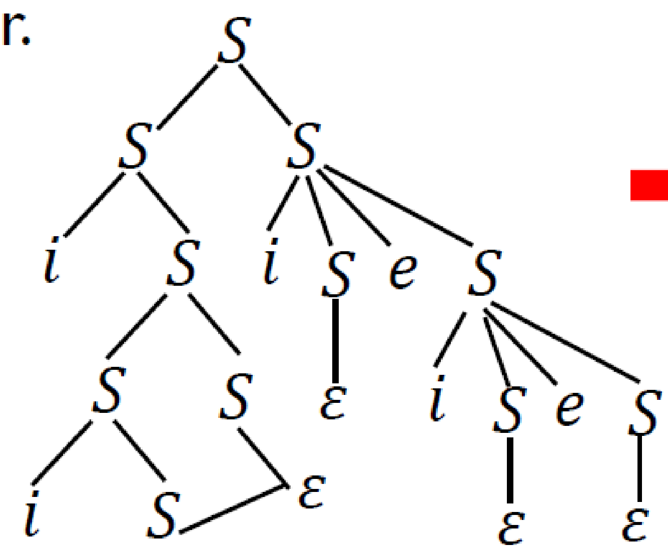
ör. google aramasi: "(nukleer|kirmizi)(baslikli|enerji)"

aramasi " nukleer baslikli ", " nukleer enerji “, " kirmizi baslikli " ve " kirmizi enerji " aramalarini bir defada arar.

ör. $G = (\{S\}, \{i, e\}, R, S)$ ve R kuralları

$$S \longrightarrow \varepsilon|SS|iS|iSeS$$

grammeri bir programlama dilinin sintaksina uygun if-else'ler uretir.



$iiieie$ kelimesini uretir ki bu uygun bir if-else blogudur.

Icerikten Bagimsiz Dillerde Belirsizlik Durumu

Gecen hafta en sol ve en sag turetimi gormustuk. Buna gore bir kelime birden fazla ayni yolla turetilebiliyordu (en sag yada en sol turetimle). Bu durumda ise ayni kelimeye karsilik birden fazla turetim agaci olur. Ve bir kelimenin birden fazla anlami olur.

Belirsizlikle bas etmek icin etkili bir yontem verilen grammeri Chomksky Normal Formu (CNF) ye getirmektir.

CNF ile n uzunlugundaki bir kelime her zaman $2n - 1$ adimda uretilir; boylece bir kelimenin bir dilin elemani olup olmadigi (yada bu dilin grammeri tarafaindan uretilip uretilmedigi) kolayca bulunur.

Ayrica CYK algoritmasiyla verilen bir kelimenin verilen bir grammer tarafindan uretilip uretilmedigini test ederken, grammerin CNF formunda olmasi gerekir.

Chomsky Normal Formu (CNF)

Bir icerikten bagimsiz grammerin her bir kurali asagidaki uc formdan biri seklinde yazilabiliyorsa bu grammere Chomsky normal formuna sahiptir denir:

$$1.\ A \longrightarrow B\ C$$

$$2.\ A \longrightarrow a$$

$$3.\ S \longrightarrow \varepsilon$$

Burada A , B , C birbirinden farkli degiskenler, a bir terminal ve S baslangic degiskenidir.

Teorem: Bir icerikten bagimsiz grammer tarafindan uretilen bir icerikten bagimsiz dil, Chomsky normal formuna sahiptir.

Icerikten Bagimsiz Grammerleri CNF'ye Donusturme

Bir icerkiten bagimsiz grammer $G = (V, \Sigma, R, S)$, CNF'ye donusturulurken asagidaki adimlar sirayla izlenir:

1. Yeni bir baslangic degiskeni S_1 tanimlanir ve bu degisken eski baslangic degiskenine (S'ye) $S_1 \longrightarrow S$ kurali ile baglanir.
2. Sag tarafinda ε terminalini iceren butun kurallar elenir (ortadan kaldirilir).

Diyelimki bir $A \longrightarrow \varepsilon$, kuralimiz olsun ve bunu eleyelim. Bu durumda 3 farkli guncelleme yapmamiz gerekebilir.

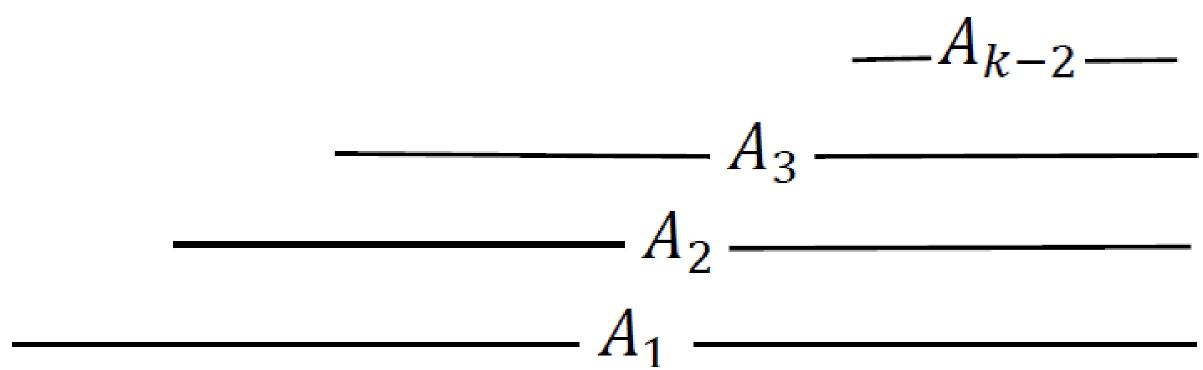
- *i.* $B \longrightarrow A$, kurali $B \longrightarrow \varepsilon$ ye donusur.
- *ii.* $B \longrightarrow uAv$, kurali $B \longrightarrow uv$ ya donusur. (burada u, v birer kelimedir.)
- *iii.* $B \longrightarrow uAvAw$, kurali, $B \longrightarrow uvw, B \longrightarrow uvAw$ ve $B \longrightarrow uAvw$ kurallarina donusur.

Icerikten Bagimsiz Grammerleri CNF'ye Donusturme

3 . $A \longrightarrow B$ formundaki tum unit (birli) kurallar elenir.
Aslinda burda elenen B'dir. Su halde B'nin yerine (varsa) sol tarafinda B olan bir kuralin sag tarafindakiler yazilir. Ornegin $A \longrightarrow B$ ve $B \longrightarrow$ uC$|D$ olsun. Bu durumda yeni kural $A \longrightarrow$ uC$|D$
olur.

4.Sag tarafinda ikiden fazla sembol olan tum kurallar elenir.
Bunun icin

$$A \longrightarrow u_1 \quad u_2 \quad u_3 \quad u_4 \quad \ldots \quad u_{k-2} \; u_{k-1} \; u_k$$



Icerikten Bagimsiz Grammerleri CNF'ye Donusturme

$$A \longrightarrow u_1 A_1$$
$$A_1 \longrightarrow u_2 A_2$$
$$A_2 \longrightarrow u_3 A_3$$
$$\ldots$$
$$A_{k-3} \longrightarrow u_{k-2} A_{k-2}$$
$$A_{k-2} \longrightarrow u_{k-1} u_k$$

ornegin $k = 3$ icin, $A \longrightarrow u_1 u_2 u_3$ icin $A \longrightarrow u_1 A_1$, $A_1 \longrightarrow u_2 u_3$.

5. Son olarak okun $A \longrightarrow u_1 u_2$ formundaki kurallar elenir, ki burada hem u_1 hem de u_2 nin ikisi birden degisken degildir. Yani ornegin $u_1 u_2$; $0B$, $B0$, 00,10 gibi ifadeler olabilir. Burada u_1 ve u_2 turune gore 3 durum dusunulur:

- *i.* u_1 terminal ($u_1 \in \Sigma$), u_2 degisken ($u_2 \in V$): Bu durumda iki yeni kural eklenir: $A \longrightarrow A_1 u_2$, $A_1 \longrightarrow u_1$. A_1 yeni bir degiskendir V'ye eklenir. ($V = V \cup A_1$)

Icerikten Bagimsiz Grammerleri CNF'ye Donusturme

ii. u_2 terminal, u_1 degisken: Bu durumda iki yeni kural eklenir:

$A \longrightarrow u_1 A_1$ ve $A_1 \longrightarrow u_2$. A_1 yeni bir degiskendir V'ye eklenir.

iii. u_1 ve u_2 terminal: Bu durumda uc yeni kural eklenir:

$A \longrightarrow A_1 A_2, A_1 \longrightarrow u_1$ ve $A_2 \longrightarrow u_2$. A_1 ve A_2 yeni degiskenlerdir V'ye eklenir ($V = V \cup A_1 \cup A_2$).

or. $G = (\{A, B\}, \{0,1\}, R, A)$ grammeri icin R kurallari soyle olsun

$$A \longrightarrow BAB|B|\,\varepsilon$$
$$B \longrightarrow 00|\varepsilon$$

bu grammeri CNF'ye donusturelim.

1. Yeni bir baslangic degiskeni olusturup bunu A'ya gonderelim.

$$S \longrightarrow A$$

$$A \longrightarrow BAB|B|\,\varepsilon$$
$$B \longrightarrow 00|\varepsilon$$

2. ε iceren tum kurallari eleyelim.

ε iceren kurallar $A \longrightarrow \varepsilon$ ve $B \longrightarrow \varepsilon$ dir.

$A \longrightarrow \varepsilon$ kuralini eledigimizde sag tarafinda A iceren tum kurallar bundan etkilenir ki bunlar: $S \longrightarrow A$ ve $A \longrightarrow BAB$ dir. $S \longrightarrow A$ ve $A \longrightarrow \varepsilon$ oldugundan kurali $S \longrightarrow \varepsilon$ kuralini elde ederiz. $A \longrightarrow BAB$ ve $A \longrightarrow \varepsilon$ oldugundan $A \longrightarrow BB$ elde edilir. Bu yeni kurallarla:

$$S \longrightarrow A \mid \varepsilon$$
$$A \longrightarrow BAB \mid B \mid BB$$
$$B \longrightarrow 00 \mid \varepsilon$$

$B \longrightarrow \varepsilon$ kuralini eleyelim. Ilk olarak $A \longrightarrow BAB$ kurali icin 3 yeni kural ekleriz (bakiniz 2. adim durum iii.) Bu kurallar:

$A \longrightarrow BA$, $A \longrightarrow AB$ ve $A \longrightarrow A$ kurallaridir.

Ikinci olarak $A \longrightarrow B$ yi dusunelim. $B \longrightarrow \varepsilon$ icin $\mathrm{A} \longrightarrow \varepsilon$ kurali elde edilir (bakiniz 2. adim durum i.) Fakat bu yeni kurali eklemeyiz cunku bu kurali zaten yukarida elemistik.

Ucuncu olarak $A \longrightarrow B$ yi dusunelim. $B \longrightarrow \varepsilon$ icin yine 2. adim iii. durum geregi, $A \longrightarrow B$, $A \longrightarrow B$ ve $A \longrightarrow \varepsilon$ yeni kurallari elde edilir. Bunlarin icinden $A \longrightarrow B$ kuralini bir defa aliriz, $A \longrightarrow \varepsilon$ kuralini almayiz cunku zaten bu kurali daha once elemistik. Sonuc olarak 2. adim sonrasi su kurallari elde ederiz:

$$S \longrightarrow A \mid \varepsilon$$

$$A \longrightarrow B\ A\ B \mid B \mid B\ B \mid A\ B \mid B\ A \mid A$$
$$B \longrightarrow 00$$

3. Unit kurallarin elenmesi. Sahip oldugumuz unit kurallar $A \longrightarrow A$ $S \longrightarrow A$ ve A$\longrightarrow B$ dir. $A \longrightarrow A$ kuralini yeni bir kural eklemeden eleyebiliriz. Kurallarimiz

$$S \longrightarrow A \mid \varepsilon$$

$$A \longrightarrow B\ A\ B \mid B \mid B\ B \mid A\ B \mid B\ A$$
$$B \longrightarrow 00$$

$S \longrightarrow A$ yi elemek icin, bu kuralda A gordugumuz yere A dan turetilebilen $B\ A\ B\ |B|\ B\ B\ |A\ B\ |B\ A$ yapilarini yaziyoruz. Yeni kurallarimiz

$$S \longrightarrow B\ A\ B\ |B|B\ B\ |A\ B\ |B\ A|\ \varepsilon$$

$$A \longrightarrow B\ A\ B\ |B|B\ B\ |A\ B\ |B\ A$$

$$B \longrightarrow 00$$

En son olarak $A \longrightarrow B$ yi eliyoruz. Bunun icin solunda B olan tek kural $B \longrightarrow 00$ yi kullaniyoruz. 3. adim sonunda kurallarimiz:

$$S \longrightarrow B\ A\ B\ |00|B\ B\ |A\ B\ |B\ A|\ \varepsilon$$

$$A \longrightarrow B\ A\ B\ |\ 00\ |\ B\ B\ |A\ B\ |B\ A$$

$$B \longrightarrow 00$$

4. Sag tarafinda ikiden fazla sembol olan tum kurallarin elenmesi.

Bu kurallar $S \longrightarrow B\ A\ B$ ve $A \longrightarrow B\ A\ B$ dir.

$S \longrightarrow B\ A\ B$ i elemek icin $S \longrightarrow B\ A_1$ ve $A_1 \longrightarrow A\ B$ kurallarini ekliyoruz.

$A \longrightarrow B\ A\ B$ i elemek icin $A \longrightarrow B\ A_2$ ve $A_2 \longrightarrow A\ B$ kurallarini ekliyoruz. 4. adim sonunda elde ettigimiz durumlar:

$$S \longrightarrow B\ A_1 \,|00\,|\, B\ B \,|\, A\ B \,|\, B\ A \,|\, \varepsilon$$

$$A \longrightarrow B\ A_2 \,|\, 0\,0 \,|\, B\ B \,|\, A\ B \,|\, B\ A$$
$$B \longrightarrow 00$$
$$A_1 \longrightarrow A\ B$$
$$A_2 \longrightarrow A\ B$$

5. Bu adimda sag tarafinda iki tane sembol olan (fakat ikisi birden degisken olmayan) kurallari eliyecegiz. Bu ornek icin bu kurallar,

$S \longrightarrow 00$, A$\longrightarrow 00$ ve $B \longrightarrow 00$ kurallari. 5. adim iii. durum geregi yeni ekleyecegimiz kurallar sunlar olur:

$S \longrightarrow A_3 A_3,\ \ A_3 \longrightarrow 0$

$A \longrightarrow A_4 A_4, \qquad A_4 \longrightarrow 0$

$B \longrightarrow A_5 A_5, \qquad A_5 \longrightarrow 0$

Sonuc olarak asagidaki CNF grammerine erisilir:

$$S \longrightarrow BA_1 \,|A_3A_3 \,|BB \,|AB \,|BA|\varepsilon$$

$$A \longrightarrow BA_2|A_4A_4 \,|BB \,|AB \,|BA$$

$$B \longrightarrow A_5A_5$$

$$A_1 \longrightarrow AB$$

$$A_2 \longrightarrow AB$$

$$A_3 \longrightarrow 0$$

$$A_4 \longrightarrow 0$$

$$A_5 \longrightarrow 0$$

or: $G = (\{A, B, S\}, \{a, b\}, R, S)$ grammeri icin R kurallari soyle olsun:

$$S \longrightarrow ASA \,|aB$$

$$A \longrightarrow B \,|S$$

$$B \longrightarrow b \,|\varepsilon$$

bu grammeri CNF ye donusturelim.

1.Yeni bir baslangic degiskeni olusturup bunu S ye gonderelim.

$$S_0 \longrightarrow S$$
$$S \longrightarrow A\,S\,A\ |a\,B$$
$$A \longrightarrow B\ |\mathrm{S}$$
$$B \longrightarrow b\ |\varepsilon$$

2. ε iceren kurallarin elenmesi.

$B \longrightarrow \varepsilon$ kuralini eleyelim. Yeni kurallar $A \longrightarrow \varepsilon$ ve $S \longrightarrow a$ olur.

$$S_0 \longrightarrow S$$
$$S \longrightarrow A\,S\,A\ |a\,B\,|\,a$$
$$A \longrightarrow B\ |\ S\ |\ \varepsilon$$
$$B \longrightarrow b$$

$A \longrightarrow \varepsilon$ kuralini eleyelim Yeni kurallar $S \longrightarrow A\ S$ ve $S \longrightarrow S\ A$ ve $S \longrightarrow S$ olur.

$$S_0 \longrightarrow S$$
$$S \longrightarrow A\,S\,A\ |a\,B\,|\,a\ |A\,S\,|S\,A\,|S$$
$$A \longrightarrow B\ |\ S$$
$$B \longrightarrow b$$

3. Unit kurallari eleyelim. $S \longrightarrow S$ i direkt eleyebilriz. $S_0 \longrightarrow S$ kurali $S_0 \longrightarrow ASA \mid aB \mid a \mid AS \mid SA \mid S$ olur.

$$S_0 \longrightarrow ASA \mid aB \mid a \mid AS \mid SA$$
$$S \longrightarrow ASA \mid aB \mid a \mid AS \mid SA$$
$$A \longrightarrow B \mid S$$
$$B \longrightarrow b$$

$A \longrightarrow B$ kurali kurali $A \longrightarrow b$ olur.

$$S_0 \longrightarrow ASA \mid aB \mid a \mid AS \mid SA$$
$$S \longrightarrow ASA \mid aB \mid a \mid AS \mid SA$$
$$A \longrightarrow b \mid S$$
$$B \longrightarrow b$$

$A \longrightarrow S$ kurali $A \longrightarrow ASA \mid aB \mid a \mid AS \mid SA \mid S$ olur.

$$S_0 \longrightarrow ASA \mid aB \mid a \mid AS \mid SA \mid S$$
$$S \longrightarrow ASA \mid aB \mid a \mid AS \mid SA \mid S$$
$$A \longrightarrow b \mid ASA \mid aB \mid a \mid AS \mid SA \mid S$$
$$B \longrightarrow b$$

4. Sag tarafinda ikiden fazla sembol olan kurallari eleyelim.

$S_0 \longrightarrow ASA$ yerine $S_0 \longrightarrow AA_1$ ve $A_1 \longrightarrow SA$

$S \longrightarrow ASA$ yerine $S \longrightarrow AA_1$

$A \longrightarrow ASA$ yerine $A \longrightarrow AA_1$

yazilir.

$$S_0 \longrightarrow AA_1|aB\ \ |a\ |AS\ |SA\ |S$$
$$S \longrightarrow AA_1|aB\ \ |a\ |AS\ |SA\ |S$$
$$A \longrightarrow b\ \big|AA_1|aB\ \ |a\ |AS\ |SA\ |S$$
$$B \longrightarrow b$$
$$A_1 \longrightarrow SA$$

5. $S_0 \longrightarrow aB$ yerine $S_0 \longrightarrow A_2B$ ve $A_2 \longrightarrow a$

$S \longrightarrow aB$ yerine $S \longrightarrow A_2B$; $A \longrightarrow aB$ yerine $A \longrightarrow A_2B$

$$S_0 \longrightarrow AA_1|A_2B\ |a\ |AS\ |SA\ |S$$
$$S \longrightarrow AA_1|A_2B\ |a\ |AS\ |SA\ |S$$
$$A \longrightarrow b\ \big|AA_1|A_2B\ |a\ |AS\ |SA\ |S$$
$$B \longrightarrow b$$
$$A_1 \longrightarrow SA\ , A_2 \longrightarrow a$$

Cocke-Younger-Kasami (CYK) Algoritmasi

CYK algoritmasi Chomsky normal formunda verilen bir grammerin, verilen bir kelimeyi uretip uretmeyecegine karar verilirken kullanilir. Yani giris olarak CNF formunda bir

G = (V, Σ, R, S) grammeri ve bir w kelimesi alir. Cikis olarak eger bu w kelimesi G grammerinden uretilirse "evet", uretilemezse "hayir" yazisini yazdirir.

or. G = $(\{S, A, B, C\}, \{a, b\}, R, S)$ ve CNF formundaki R kurallari

$$S \longrightarrow A\,B \mid B\,C$$

$$A \longrightarrow B\,A \mid a$$

$$B \longrightarrow C\,C \mid b$$

$$C \longrightarrow A\,B \mid a$$

seklinde verilen iceriktan bagimsiz grammer $w = b\;a\;a\;b\;a$ kelimesini uretir mi?

$S \longrightarrow AB|BC$
$A \longrightarrow BA|a$
$B \longrightarrow CC|b$
$C \longrightarrow AB|a$

5	*baaba* S, A, C				
4	*baab* -	*aaba* S, A, C			
3	*baa* -	*aab* B	*aba* B		
2	*ba* S, A	*aa* B	*ab* S, C	*ba* S, A	
1	*b* B	*a* A, C	*a* A, C	*b* B	*a* A, C
	b	a	a	b	a

5. $b \longleftarrow B, a \longleftarrow A, C$

4. $ba \longleftarrow (B)(A, C) = BA, BC \longleftarrow A, S$

$aa \longleftarrow (A, C)(A, C) = AA, AC, CA, CC \longleftarrow B$ (burada yalnizca CC'ye ulasilabilir)

$ab \longleftarrow (A, C)(B) = AB, CB \longleftarrow$ S, C

3. $baa, (b, aa)$ ve (ba, a) seklinde ayristirilir.

$ba, a \longleftarrow (S, A)(A, C) = SA, SC, AA, AC$ (bu ifadelerin hic birine ulasilamaz)

$b, aa \longleftarrow (B)(B) = BB$ (bu ifadeye ulasilamaz, yani buna giden bir ok yoktur)

3. aab, (a, ab) ve (aa, b) seklinde ayristirilir.

$a, ab \longleftarrow (A, C)(S, C) = AS, AC, CS, CC \longleftarrow B$

$aa, b \longleftarrow (B)(B) = BB$ (bu ifadeye ulasilamaz, yani buna giden bir ok yoktur)

aba, (a, ba) ve (ab, a) seklinde ayristirilir.

$a, ba \longleftarrow (A, C)(S, A) = AS, AA, CS, CA$ (hic birine ulasilamaz).

$ab, a \longleftarrow (S, C)(A, C) = SA, SC, CA, CC \longleftarrow B$

4. $baab, (b, aab), (ba, ab), (baa, b)$ seklinde ayristirilir.

$b, aab \longleftarrow (B)(B) = BB$

$ba, ab \longleftarrow (S, A)(S, C) = SS, SC, AS, AC$ (bu ifadelere ulasilamaz)

$baa, b \longleftarrow -, B$

$aaba, (a, aba), (aa, ba), (aab, a)$ seklinde ayristirilir.

$a, aba \longleftarrow (A, C)(B) = AB, CB \longleftarrow S, C$

$aa, ba \longleftarrow (B)(S, A) = BS, BA \longleftarrow A$

$aab, a \longleftarrow (B)(A, C) = BA, BC \longleftarrow A, S$

4. $baaba, (b, aaba), (ba, aba), (baa, ba), (baab, a)$ seklinde ayristirilir.

$b, aaba \longleftarrow (B)(S, A, C) = BS, BA, BC \longleftarrow A, S$

$ba, aba \longleftarrow (S, A)(B) = SB, AB \longleftarrow S, C$

$baa, ba \longleftarrow -, SA$

$baab, a \longleftarrow -(A, C)$

Sonuc olarak piramitin tepesinde S baslangic degiskeni oldugu icin, $w = baaba$ kelimesi verilen grammerden uretilebilir. Baska bir deyisle baslangic degiskeni S den w kelimesine ulasilabilir.

or. G $= (\{A, B, X, T, S\}, \{a, b\}, R, S)$ ve R kurallari

$$S \longrightarrow AB|XB|\varepsilon$$

$$T \longrightarrow AB|XB$$

$$X \longrightarrow AT$$

$$A \longrightarrow a$$

$$B \longrightarrow b$$

seklinde verilen iceriktan bagimsiz grammer $w = aabb$ kelimesini uretir mi?

$S \longrightarrow AB|XB|\varepsilon$

$T \longrightarrow AB|XB$

$X \longrightarrow AT$

$A \longrightarrow a$

$B \longrightarrow b$

4	$aabb$ / S,T			
3	aab / X	abb / $-$		
2	aa / -	ab / S,T	bb / $-$	
1	a / A	a / A	b / B	b / B
	a	a	b	b

4\. $b \longleftarrow B, a \longleftarrow A$

3\. $aa \longleftarrow AA$ (bu ifadeye ulasilamaz)

$aa \longleftarrow AB \longleftarrow ST$, $bb \longleftarrow BB$

2\. aab, (a, ab) ve (aa, b) seklinde ayristirilir.

$a, ab \longleftarrow (A)(S,T) = AS, AT \longleftarrow X$, $aa, b \longleftarrow -B$

abb, (a, bb) ve (ab, b) seklinde ayristirilir.

$a, bb \longleftarrow (A)-,\;\; ab, b \longleftarrow (S,T)(B) = SB, TB$ (bu ifadelere ulasilamaz)

1\. $aabb$, (a, abb), (aa, bb) ve (aab, b) seklinde ayristirilir.

$a, abb \longleftarrow (A), -$

$aa, bb \longleftarrow - -$

$aab, b \longleftarrow (X)(B) = XB \longleftarrow S\ T$