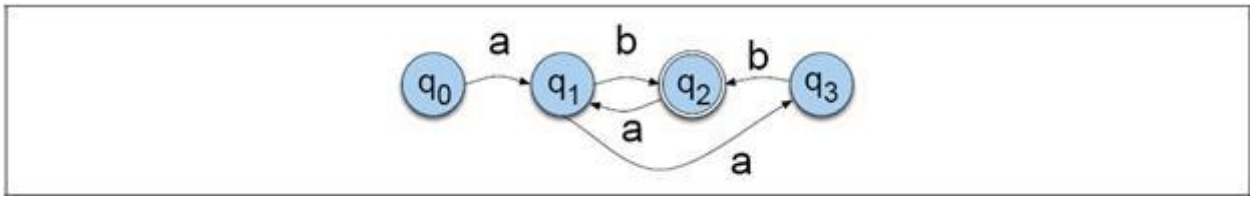
SHIVAM GUPTA{SXG190040}

CS 6320.501 Natural Language Processing: Assignment-2

1. NFSA to Regular Expression (20 points)

Please note that **ONLY** operators presented in the Lectures can be used to answer Regex questions in the homeworks and exams.

a. (**10 points**) Write a regular expression for the language accepted by the FSA:

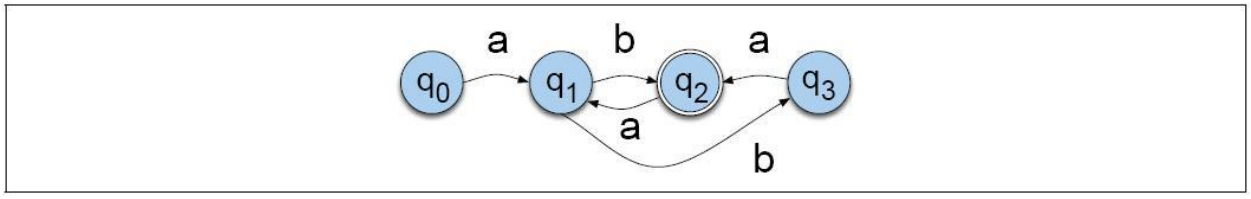


Regular Expression: **\b(ab)+|(aab(ab)*)+\b**

Explanation:

As q0 is the initial state q2 is the final state. Starting the regular Expression with "a" to go from q0 to q1 and then q1 to q2 with "b", So One case would be (ab). There is a possibility it goes from q0->q1->q3>q2 So "aab" is used for the 2nd Case. There is a possibility that (ba) comes with a loop from q1 to q2 then again q2 to q1. So (ab) is put as the loop might occur or might not. There is to possibility of the loop so Or(|) condition is used. (aab(ab)*)+ is used for more than or equal to 1 times.. Or Condition is used because it can go with a loop or directly. Putting the word boundaries(\b) in the end as well.*

b. (**10 points**) Write a regular expression for the language accepted by the NFSA:



Regular Expression: **\b(ab)+|(aba(ab)*)+\b**

Explanation:

As q0 is the initial state q2 is the final state. Starting the regular Expression with "a" to go from q0 to q1 and then q1 to q2 with "b", So One case would be (ab). There is a possibility it goes from q0->q1->q3>q2 So "aba" is used for the 2nd Case. There is a possibility that (ab) comes with a loop from q2 to q1 then again q2 to q1. So (ab) is put as the loop might occur or might not as q2(final)*

is reached. Then again there is a loop so (ba) will be used ending with "b There is to possibility of the loop so Or(|) condition is used. (aab(ab)*)+ is used for more than or equal to 1 times.. Or Condition is used because it can go with a loop or directly. Putting the word boundaries(\b) in the end as well.*

Question 2
(A)

<s> a man a man a man a plan a plan a canal panama panama </s>
count of each word (Unigram):
a : 6
man: 3
plan: 2
canal: 1
panama: 2
Total number of word in the training sentence corpus = 16

The probabilities for Unigrams are:
a: 6/16 = 0.375
man: 3/16=0.1875
plan: 2/16=0.125
canal: 1/16=0.0625
panama: 2/16=0.125
<s> = 1/16 = 0.0625
</s> = 1/16 = 0.0625

Bi-Gram Model

	<s>	a	man	plan	canal	panama	</s>
<s>	0	1	0	0	0	0	0
A	0	0	3	2	1	0	0
Man	0	3	0	0	0	0	0
Plan	0	2	0	0	0	0	0
Canal	0	0	0	0	0	1	0
panama	0	0	0	0	0	1	1
</s>	0	0	0	0	0	0	0

(B)

Bi-Gram Model: Compute the bigram based probability of the following test sentence *<s> plan a panama </s>*

(i) No Smoothning

P(plan|<s>)=0/C(<s>)= 0/1 = 0

P(a|plan)=2/C(plan) = 2/2=1

P(panama|a)=0/C(a) = 0/6=0

P(</s> | panama)= 1/C(panama) = ½ = 0.5

The probability of the Test sentence is (No Smoothing) =

P(plan|<s>) *P(a|plan) *P(panama|a)* P(</s> | panama)

= 0*1*0*0.5 = 0

(ii) Add one Smoothing

Total count of the corpus = 16

V(Total Number of unique Words in Corpus) = 7

	<s>	a	man	plan	canal	panama	</s>
<s>	1	2	1	1	1	1	1
A	1	1	4	3	2	1	1
Man	1	4	1	1	1	1	1
Plan	1	3	1	1	1	1	1
Canal	1	1	1	1	1	2	1
panama	1	1	1	1	1	2	2
</s>	1	1	1	1	1	1	1

P(plan|<s>)=1/C(<s>)+V= 1/(1+7) = 1/8 = 0.125

P(a|plan) = 3/C(plan) +V=3/2+7=0.334

P(panama|a) = 1/C(a) + V = 1/6+7 = 0.0769

P(</s> | panama)= 2/C(panama)+7 = 2/9 = 0.22

The probability of the Test sentence is (Add one Smoothing)=

P(plan|<s>) *P(a|plan) *P(panama|a)* P(</s> | panama)

= 0.125*0.334*0.0769*0.22

= 7.06 X 10^(-4)

(iii) Good Turing Smoothning

Buckets on the basis of the counts:

a : 6

man: 3

plan: 2

canal: 1

panama: 2

Unigram Bucket:

Bucket No.	1	2	3	4	5	6
Bucket Items	Canal	Plan, Panama	Man	-	-	a
Count	1	2	1	0	0	1

<s> a man a man a man a plan a plan a canal panama panama </s>

<u>The 15 Bigrams are as follows:</u>

<s> a
a man
man a
a man
man a
a man
man a
a plan
plan a
a plan
plan a
a canal
canal panama
panama panama
panama </s>

Bucket No.	1	2	3
Bucket Items	*(<s>, a)* *(a canal),* *(canal panama),* *(panama panama)* *(panama </s>)*	*(a plan), (plan, a)*	*(a man), (man a)*
Count	5	2	2

C0 =N1/N

N1= 5, N=15

P(plan|<s>)= as "<s> plan" is having 0 Frequency

So, Prob =N1/N =5/15=1/3 = 0.334

P(a|plan) = as “plan a” is having Frequency of 2
C*2 = (2+1)(N3/N2) = 3*(2/2) = 3
C*2/N = 3/15= 1/5 = 0.2

P(panama|a) = as “a panama” is having 0 Frequency
So, Prob =N1/N =5/15=1/3= 0.334

P(</s> | panama)= as “plan a” is having Frequency of 1
C*1 = (1+1)(N2/N1) = 2*(2/5) = 0.8
C*1/N = 0.8/15 = 0.8/15 = 0.05333

The probability of the Test sentence is (Good Smoothing) =
P(plan|<s>) *P(a|plan) *P(panama|a)* P(</s> | panama)
= 0.334*0.2*0.334*0.05333
= 1.185 X 10^(-3)