Q 1:

I am using perl syntax:

1. grep -P "[a-zA-Z0-9]\*" input.txt

If in case alphabetic string denotes just upper and lowercase characters, below is the command:

grep -P "[a-zA-Z]\*" input.txt

1. As told during discussion that the string should be separated by others using “white space”.

Below is the command used:

grep "\s[a-z]\*b\s" input.txt

If in case the string can also be the last word of the sentence, then the last position can be occupied with any other separator as well.

Below will be the command used in that case:

grep "\b[a-z]\*b\b" input.txt

1. The statement captures the word in \w and \1 refers to the group of similar previous word.

grep -P "\b([a-zA-Z]+)\s+\1\b" input.txt

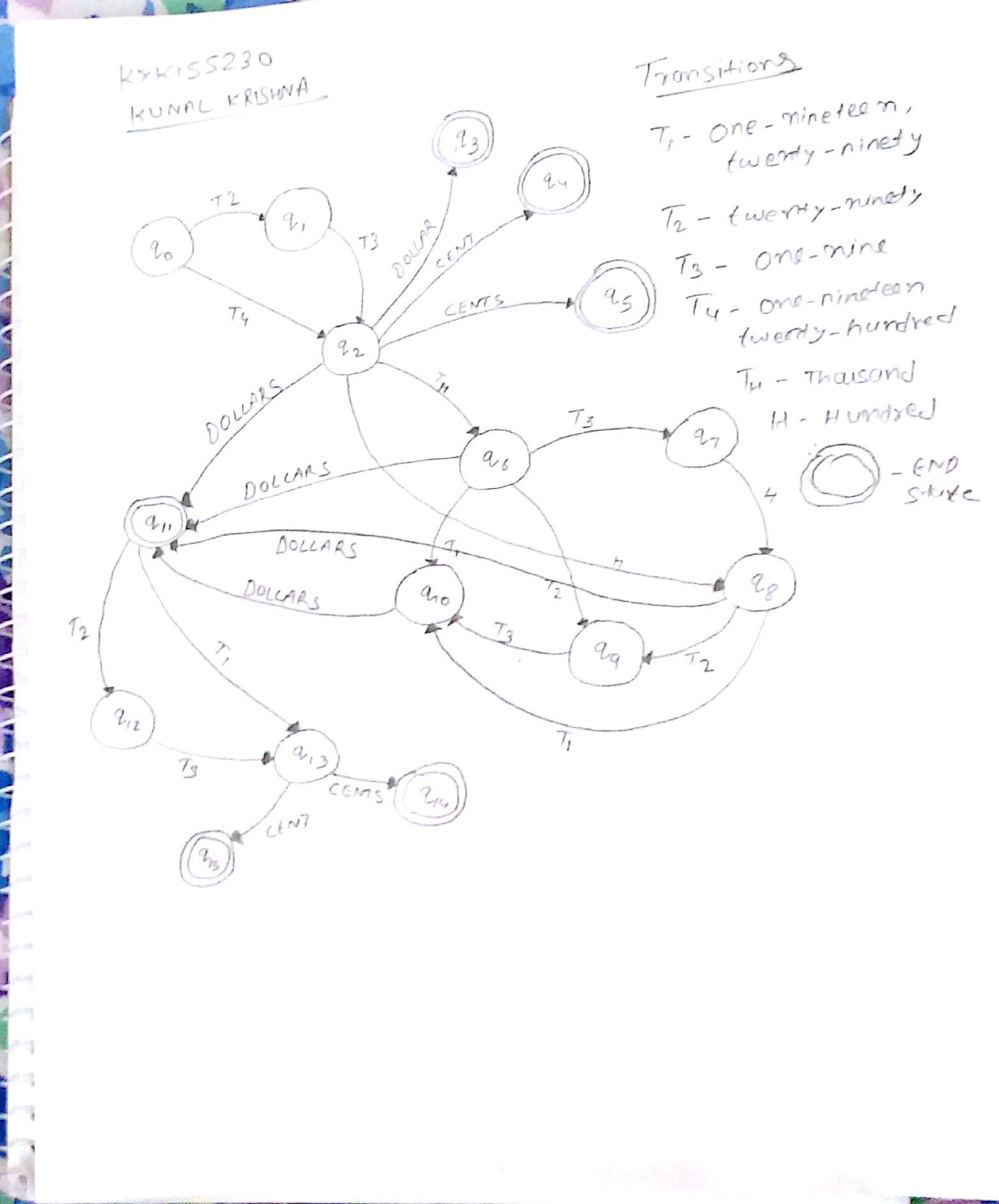
1. grep -P "\b(b+((a)b+)+)\b" input.txt

If in case parts of string can also be considered, then the command would be as below:

grep -P "b+((a)b+)+" input.txt

grep -P "^[\d]+.\*[a-zA-Z]+$" input.txt

Q2: Automata picture



Q3:

I have used java program to implement the Bigram.

Compile Bigram.java file, I have hardcoded the input file, so I have included the input file along with the java program.

I have used unix file name in my program

**NLPCorpusTreebank2Parts-CorpusA-Unix.txt**

**So if you running this program in window machine then please change the file name to below at line number 111.**

**NLPCorpusTreebank2Parts-CorpusA-Windows.txt**

First it will ask for the user input string , then once it accept then it will give the output of given string and then display the required table for S1 and S2.