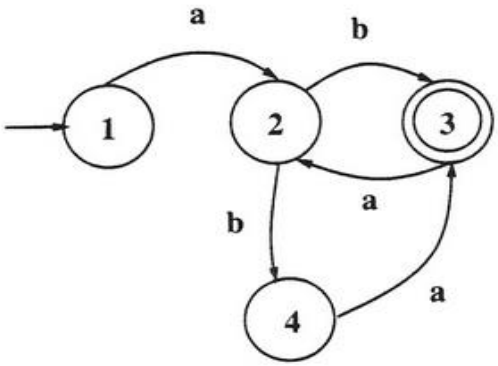
Assignment 1

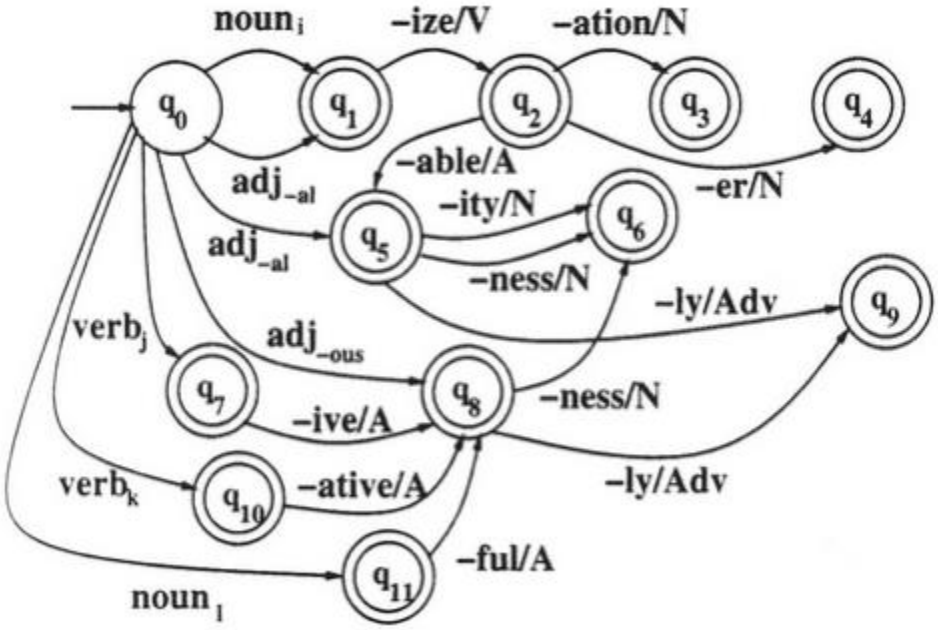
Q1 (1 point)

Write a regular expression for the language accepted by the NFSA in the following figure:



Q2 (1 point)

Give examples of each of the noun and verb classes in the following figure and find some exceptions to the rules.



Q3 (2 points)

To attune your ears to pronounciation reduction, listen for the pronounciation of the word *the, a* or *to* in the spoken language around you. Try to notice when it is reduced, and mark down whatever facts about the speaker or speech situation that you can. What are your observations?

Please post your answer here AND to the discussion board. Note that there is no fixed answer.

© Statistics.com, LLCand Dr. Nitin Indurkhya

Q4 (2 points)

Consider the string *aaaabbbcccdb* as a training corpus. Give a table of counts (over a vocabulary set of letters "a" to "d") for the bigrams that occur given this string. Use a 4x4 table with rows labeled "a", "b", etc and columns likewise. Table(row=a,col=b) should contain the count for bigram "ab". Ignore the start and end tokens.

Q5 (2 points)

Consider the string *aaaabbbcccdb* as a training corpus. Give a table of counts (over a vocabulary set of letters "a" to "d") for the bigrams that occur given this string. Use a 4x4 table with rows labeled "a", "b", etc and columns likewise. Table(row=a,col=b) should contain the count for bigram "ab". Ignore the start and end tokens.

Q6 (2 points)

The following two demos give you an idea of stemming. Try them on your own and post a short discussion of differences (if any) you notice between the two.

1. The following URL has a demo of a stemming algorithm where you can stem large amounts of text http://text-processing.com/demo/stem/.
2. The following demo uses the Porter stemmer which removes derivational affixes: http://snowball.tartarus.org/demo.php.