Natural Language Processing (COS4861)

Adriaan Louw (53031377)

March 6, 2018

1 Question 1

1.1 $/[a-zA-Z]?[a-z]*/$

1.2 $/([a-zA-Z]?[a-z]*)\backslash s\backslash 1/$

1.3 $/([a-zA-Z]?[a-z]*)\backslash s[a-zA-Z]?[a-z]*\backslash s\backslash 1/$

1.4.a This string start off at the beginning of a line.It start with 3 asterisks (*), followed by 2 or 3 pluses (+). That is followed by one colon (:) and a space. That is followed by a string containing upper and lower case letters and the underscore. There will be at least one character in this substring. Then we have another colon. Then the same amount of pluses we had before and to finish this string 9 off we have 3 asterisks. This string will also end the line.

1.4.b $***++:CompCSI\_\_N\_L\_P\_\_COS\_4861:++***$
(please note there is a space after the first ":" and a space before the second ":")

1.5 $/0((11)+0)*1*/$

1.6.a $/\backslash w+(\backslash.\backslash w+)*@\backslash w+(\backslash.\backslash w+)*$

1.6.b The initial accuracy was 0.66.
The RE was then changed to $/\backslash w+([\backslash.\backslash w\backslash'\backslash-]+)*@\backslash w+(\backslash.\backslash w+)*$ which has an accuracy of 1 over the training set and it did not match anything but email address's. The data used to determine the correct form of an email address can be found at $https://tools.ietf.org/html/rfc3696$

1.7 Accuracy for this training set was 0.66
The address richard.j.beatty@mvp02.usace.army.mil was not matched completely because the address was over multiple lines.
The address "jomose AT abranch DOT com" was not matched because it uses multiple words and does not have the @ sign.
The RE also incorrectly matched the "address" "trailers@residences", which is not an address at all. This is not taken into account in the accuracy calculation

2 Question 3