Natural Language Processing (COS4861)

Adriaan Louw (53031377)

June 2, 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

1.8 We start with $Q' = \{\{1\}\}$
With $q = \{1\}$ we have $t = \bigcup_{p \epsilon q} \delta(p, 0) = \{1\} \cup \{2\} \cup \{3\} \cup \{4\} \cup \{5\} = \{1, 2, 3, 4, 5\}$ add t to Q'. Thus $Q' = \{\{1\}, \{1, 2, 3, 4, 5\}\}$. Then for $t = \bigcup_{p \epsilon q} \delta(p, 1) = \{4, 5\}$. Then we also add t to Q', making $Q' = \{\{1\}, \{1, 2, 3, 4, 5\}, \{4, 5\}\}$. We

then also add $\delta(\{1\}, 0) = \{1, 2, 3, 4, 5\}$ and $\delta(1, 1) = \{4, 5\}$ to δ'. To determine new accepts states we see that $q \cap A = \emptyset$ which means we do not add anything to A'.

For $q = \{1, 2, 3, 4, 5\}$ we get $\delta(\{1, 2, 3, 4, 5\}, 0) = \delta(\{1\}, 0) \cup \delta(\{2\}, 0) \cup (\{3\}, 0) \cup \delta(\{4\}, 0) \cup \delta(\{5\}, 0) = \{1, 2, 3, 4, 5\}$, which we add to Q', but since it is already in Q' there is no effect. Then $\delta(\{1, 2, 3, 4, 5\}, 1) = \{3, 4, 5\}$, which means we add $\{3, 4, 5\}$ to Q'. Thus, $Q' = \{\{1\}, \{1, 2, 3, 4, 5\}, \{4, 5\}, \{3, 4, 5\}\}$. Then we add to mappings $\delta(\{1, 2, 3, 4, 5\}, 0) = \{1, 2, 3, 4, 5\}$ and $\delta(\{1, 2, 3, 4, 5\}, 1) = \{3, 4, 5\}$ to δ'. Calculating whether we have a new accept state, we have $q \cap A = \{4\}$. Thus we add $\{1, 2, 3, 4, 5\}$ to A'.

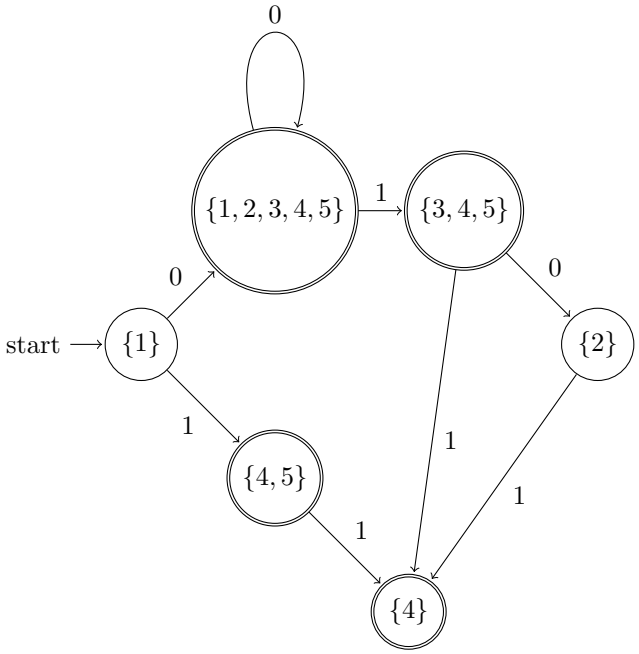
For $q = \{4, 5\}$, $t = \delta(\{4, 5\}, 0) = \emptyset$. We add $\emptyset$ to Q' and the mapping $t = \delta(\{4, 5\}, 0) = \emptyset$ to δ'. Then $t = \delta(\{4, 5\}, 1) = \{4\}$. Which means we add the state $\{4\}$ to Q' and the mapping $\delta(\{4, 5\}, 1) = \{4\}$ to δ'. $q \cap A = \{4\}$, so we add $\{4, 5\}$ to A'. Now $Q' = \{\{1\}, \{1, 2, 3, 4, 5\}, \{4, 5\}, \{3, 4, 5\}, \{4\}, \emptyset\}$.

Next we have $q = \{3, 4, 5\}$. Or 2 new mappings are $\delta(\{3, 4, 5\}, 0) = \{2\}$ and $\delta(\{3, 4, 5\}, 1) = \{4\}$, which we add to δ'. Then we only add $\{2\}$ to Q' because $\{4\}$ is already in Q'. For the accept state check we have $q \cap A = \{4\}$. SO we add $\{3, 4, 5\}$ to A'.

Then for $q = \{2\}$ we have $\delta(\{2\}, 0) = \emptyset$ and $\delta(\{2\}, 1) = \{4\}$. These mappings we add to δ'. We also do not add any new states to Q' since the new states are in Q'. Also $q \cap A = \emptyset$ so no new accept states.

The state $\{4\}$ has no outgoing states so we can ignore this state in terms of generating new states.

Now we have gone through all the states in Q'. The result is
$Q' = \{\{1\}, \{1, 2, 3, 4, 5\}, \{4, 5\}, \{3, 4, 5\}, \{4\}, \emptyset\, \{2\}\}$ and
$A = \{\{1, 2, 3, 4, 5\}, \{4, 5\}, \{3, 4, 5\}, \{4\}\}$.



2 Question 2

2.1 Minimum edit distance from connect to commute is 8

t	7	6	5	6	7	8	7	8
c	6	5	4	5	6	7	8	9
e	5	4	3	4	5	6	7	7
n	4	3	2	3	4	5	6	7
n	3	2	1	2	3	4	5	6
o	2	1	0	1	2	3	4	5
c	1	0	1	2	3	4	5	6
#	0	1	2	3	4	5	6	7
	#	c	o	m	m	u	t	e

Minimum edit distance from connect to contact is 4

t	7	5	5	4	3	4	5	4
c	6	4	4	3	4	5	4	5
e	5	4	3	2	3	4	5	6
n	4	3	2	1	2	3	4	5
n	3	2	1	0	1	2	3	4
o	2	1	0	1	2	3	4	5
c	1	0	1	2	3	4	5	6
#	0	1	2	3	4	5	6	7
	#	c	o	n	t	a	c	t

From this we can see that connect is closer to contact than to commute

2.2 Please run minimumEditDistance.py and please ensure that python 3.7 is installed. It can be found at https://www.python.org/downloads/

2.3 Please run chatbot.py and please ensure that python 3.7 is installed. It can be found at https://www.python.org/downloads/

3 Question 3

3.1 + 3.2 Please run RunCorpora_TaleOfTwoCities.py and RunCorpora_WrightBrothers.py and please ensure that python 3.7 is installed. It can be found at https://www.python.org/downloads/