Natural Language Processing Handout

CSCE 420

November 19, 2019

1 Regular Expression

• Disjunctions [] :

Pattern	Matches
$[Ww]oodchuck$	woodchuck, Woodchuck
[0123456789]	Any single digit

• Disjunctions |:

Pattern	Matches
$abc\|def$	Find 'abc' or 'def'.
$a\|b\|ab$	Find 'a' or 'b' or 'ab'. Example: 'abc'

• Ranges:

Pattern	Matches
$[A-Z]$	An uppercase letter.
$[a-z]$	A lowercase letter.
$[0-9]$	A single digit.

• Negation ˆ. (Note: Carat means negation only when its first in [])

Pattern	Matches
$[\hat{}A-Z]$	Not upper case
$[\hat{}Ss]$	Not 'S' nor 's'
$[\hat{}e\hat{}]$	Not 'e' nor 'ˆ'
$a\hat{}b$	Search for the pattern'$a\hat{}b$'

• Other notations.

?	0 or 1 of previous character
∗	0 or more of previous character
+	1 or more of previous character
.	Any character
ˆ	Start anchor
$	End anchor
\	Escape character

2 Text Classification

Training set:

#	**Text**	**Class**
1	Carla Betty Carla	a
2	Carla Carla Suzanne	a
3	Carla Matt	a
4	Taylor Jessica Carla	b

Document d_5: *Carla Carla Carla Taylor Jessica*