CPE 372/641 Natural Language Processing

ASST. PROF. DR. NUTTANART M. FACUNDES

MOST OF THE SLIDES ARE FROM HTTPS://WEB.STANFORD.EDU/~JURAFSKY/SLP3/

Basic Text Processing

Regular Expressions

Regular expressions

- A formal language for specifying text strings, for text normalization
- How can we search for any of these?
  - woodchuck
  - woodchucks
  - Woodchuck
  - Woodchucks

Text normalization

- Converts texts into more convenient, standard forms.
- Includes
  - Tokenization: tokenizes words from running texts
  - Lemmatization: determines if 2 words have the same root, e.g. sings, sang, sung have the same root ‘sing’
  - Stemming: stripes the suffixes from the end of words
  - Sentence segmentation
  - Word comparison: edit distance

Two types of characters in Regular Expressions (REs)

- Literal
  - Every normal text character is an RE, and denotes itself.
- Meta-characters
  - Special characters that allow you to combine REs in various ways
    - Example: a denotes a
    - a* denotes ε or a or aa or aaa or … (zero or more *a*s)

Regular Expressions: Disjunctions

- Letters inside square brackets []

Pattern	Matches
`[wW]oodchuck`	Woodchuck, woodchuck
`[1234567890]`	Any digit

- Ranges `[A-Z]`

Pattern	Matches	
`[A-Z]`	An upper case letter	`Drenched Blossoms`
`[a-z]`	A lower case letter	`my beans were impatient`
`[0-9]`	A single digit	`Chapter 1: Down the Rabbit Hole`

Regular Expressions: Negation in Disjunction

- Negations `[^Ss]`
  - Carat means negation only when first in []

Pattern	Matches	
`[^A-Z]`	Not an upper case letter	`Oyfn pripetchik`
`[^Ss]`	Neither ‘S’ nor ‘s’	`I have no exquisite reason”`
`[^e^]`	Neither e nor ^	`Look here`
`a^b`	The pattern a carat b	`Look up a^b now`

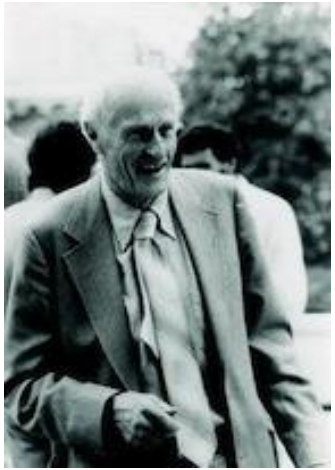
Regular Expressions: More Disjunction

- Woodchucks is another name for groundhog!
- The pipe | for disjunction

Pattern	Matches
`groundhog|woodchuck`	
`yours|mine`	yours mine
`a|b|c`	= [abc]
`[gG]roundhog|[Ww]oodchuck`	

Regular Expressions: ? * + .

Pattern	Matches	
colou?r	Optional previous char	color colour
oo*h!	0 or more of previous char	oh! ooh! oooh! ooooh!
o+h!	1 or more of previous char	oh! ooh! oooh! ooooh!
baa+		baa baaa baaaa baaaaa
beg.n		begin begun begun beg3n



Stephen C Kleene

Kleene *, Kleene +

Regular Expressions: Anchors ^ $

Pattern	Matches
^[A-Z]	Palo Alto
^[^A-Za-z]	1 "Hello"
\.$	The end.
.$	The end? The end!

Example

- Find me all instances of the word “the” in a text.

```
the
```

Misses capitalized examples

```
[tT]he
```

Incorrectly returns `other` or `theology`

```
[^a-zA-Z][tT]he[^a-zA-Z]
```

Errors

- The process we just went through was based on fixing two kinds of errors
  - Matching strings that we should not have matched (there, then, other)
    - False positives (Type I)
  - Not matching things that we should have matched (The)
    - False negatives (Type II)

Errors cont.

- In NLP we are always dealing with these kinds of errors.
- Reducing the error rate for an application often involves two antagonistic efforts:
  - Increasing accuracy or precision (minimizing false positives)
  - Increasing coverage or recall (minimizing false negatives).

More Regular Expressions

–\d	digits	[0-9]
–\D	non-digits	[^0-9]
–\w	alphabetic	[a-zA-Z]
–\W	non-alphabetic	[^a-zA-Z]

In-class Exercise

- To submit at the end of today. Work in Pair. Give two or more examples of strings that each of the following regular expressions matches. And explain how they match.
- \d+ dollars
- \w*oo\w*

Summary

- Regular expressions play a surprisingly large role
  - Sophisticated sequences of regular expressions are often the first model for any text processing text
- For many hard tasks, we use machine learning classifiers
  - But regular expressions are used as features in the classifiers
  - Can be very useful in capturing generalizations

ELIZA: Substitutions Using Memory

User: Men are all alike.

ELIZA: IN WHAT WAY

s/.* all .*/IN WHAT WAY/

They’re always bugging us about something or other.

ELIZA: CAN YOU THINK OF A SPECIFIC EXAMPLE

s/.* always .*/CAN YOU THINK OF A SPECIFIC EXAMPLE/

User: My boyfriend says I’m depressed.

ELIZA: I AM SORRY TO HEAR YOU ARE DEPRESSED

s/.* I’m (depressed|sad) .*/I AM SORRY TO HEAR YOU ARE \1/

Using RE's: Examples

- Predictions from a news corpus:
  - Which candidate for Governor of California is mentioned most often in the news?
- Language use:
  - Which form of comparative is more frequent: 'oftener' or 'more often'?
  - How often do sentences end with infinitival 'to'?
  - What words most often begin and end sentences?

- Personality profiling:
  - Are you more or less polite than the people you correspond with?
  - With labeled data, which words signal friendly messages vs. unfriendly ones?

Regular Expressions

Youtube video by Prof. Jurafsky
https://www.youtube.com/watch?v=EyzTQ0OKeNw

regexpal.com: Regular Expressions Tester

Text to test at regexpal.com

We looked!
Then we saw him step in on the mat.
We looked!
And we saw him!
The cat in the hat!
The other one there, the blithe one.