Regular Expressions

Module 1 of our NLP Course

Find all the dates in a Text

(\d{2}[/-]){2}(\d{2}|\d{4})

How to tell if an email is valid

```
([!#-'*+/-9=?A-Z^-~-]+(\.[!#-'*+/-9=?A-Z^-~-]+)*|"([]!#-[^-~
\t]|(\\[\t-~]))+")@([!#-'*+/-9=?A-Z^-~-]+(\.[!#-'*+/-9=?A-Z^
-~-]+)*|\[[\t -Z^-~]*])
```

An Easier Regex that almost always works

```
(^[a-zA-Z0-9_.+-]+@[a-zA-Z0-9-]+\.[a-zA-Z0-9-.]+$)
```

Explained

```
(^[a-zA-Z0-9_.+-]+@[a-zA-Z0-9-]+\.[a-zA-Z0-9-.]+$)
```

```
^ BEGINING OF STRING
[a-zA-Z0-9_.+-]+ ANY OF THESE CHARACHTERS
@ LITERRALY THE @ SYMBOL
[a-zA-Z0-9-]+ ANY OF THESE ON REPEAT
\. LITERRALY THE . SYMBOL
[a-zA-Z0-9-.]+ ANY OF THESE ON REPEAT
$ END OF STRING
```

Try tal@lighttag.io

```
(^[a-zA-Z0-9_.+-]+@[a-zA-Z0-9-]+\.[a-zA-Z0-9-.]+$)
```

```
^ BEGINING OF STRING
[a-zA-Z0-9_.+-]+ ANY OF THESE CHARACHTERS
@ LITERRALY THE @ SYMBOL
[a-zA-Z0-9-]+ ANY OF THESE ON REPEAT
\. LITERRALY THE . SYMBOL
[a-zA-Z0-9-.]+ ANY OF THESE ON REPEAT
$ END OF STRING
```

Light TAG

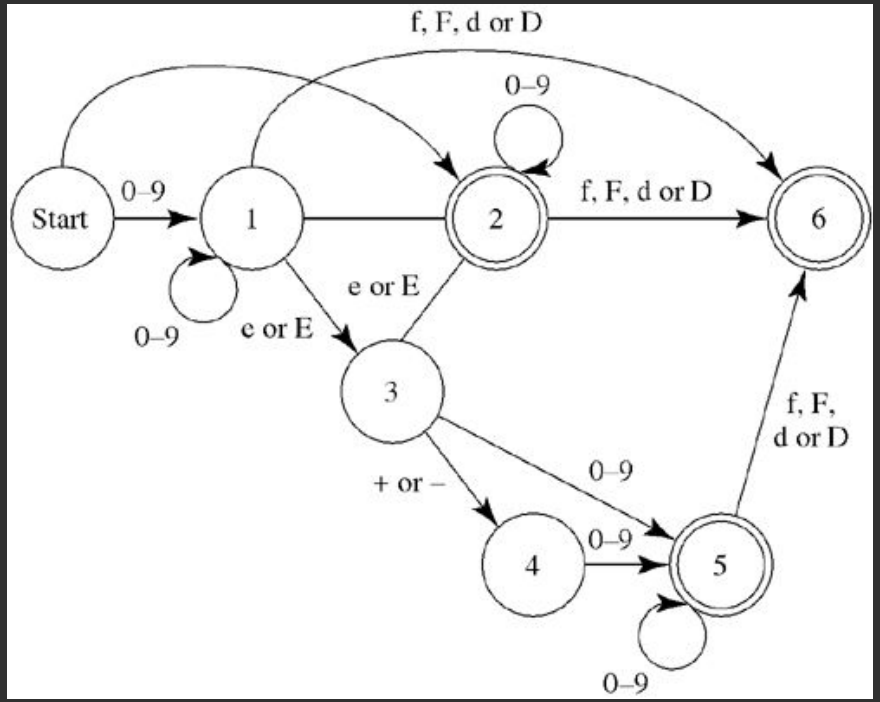
Try “tal is here@lighttag.io”

```
(^[a-zA-Z0-9_.+-]+@[a-zA-Z0-9-]+\.[a-zA-Z0-9-.]+$)
```

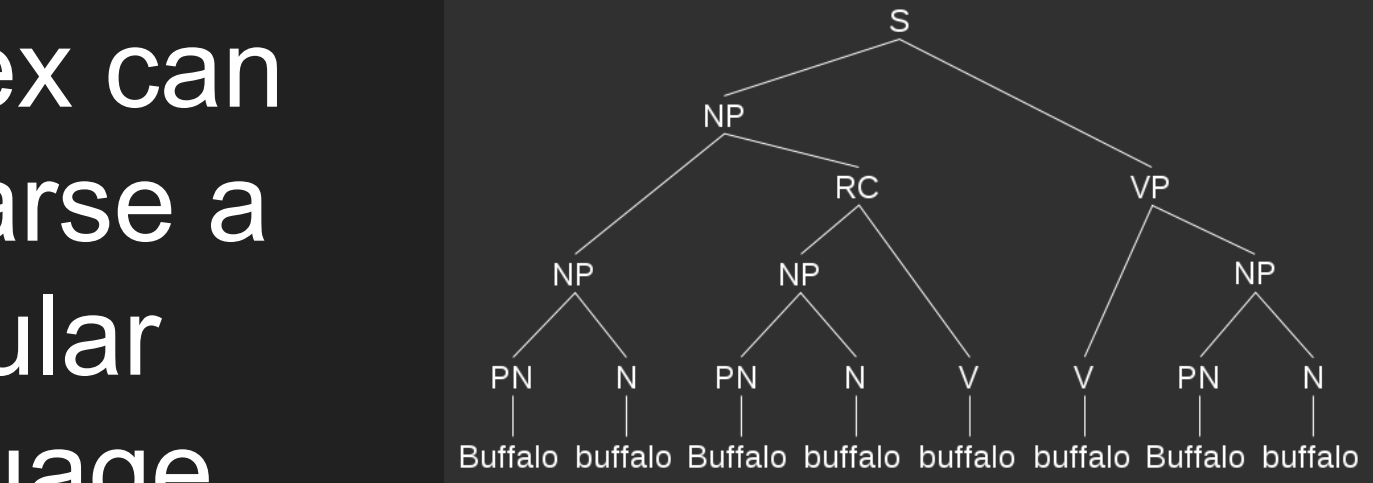
```
^ BEGINING OF STRING
[a-zA-Z0-9_.+-]+ ANY OF THESE CHARACHTERS
@ LITERRALY THE @ SYMBOL
[a-zA-Z0-9-]+ ANY OF THESE ON REPEAT
\. LITERRALY THE . SYMBOL
[a-zA-Z0-9-.]+ ANY OF THESE ON REPEAT
$ END OF STRING
```

Light TAG

A regular language can be recognized by a finite state machine



A Regex can only parse a Regular Language



Light TAG

Many patterns you're looking for are regular

(\d{2}[/-]){2}(\d{2}|\d{4})

Now Practice

- Open Problem set 1
- Do tutorial at RegexOne
- Do Problem set 1

LightTAG

Resources

- RegexOne - Interactive tutorial
- How regex work and why they are fast
- Python docs on regex