# Exploratory Text Analysis

# How can you analyse text?

- Matching
  - → Regular Expressions
    - .match()
  - → LIKE
    - SQL
- Extracting
  - "12|12|00" (1, 2, 3)
  - → Regex .findall()
- Replacement
  - Regex .sub()
  - "£15.00" → "€15.00"

... these techniques all use Explicit patterns provided by the analyst

Eg. Regular Expression

FIND "[0-9][0-9]/[0-9][0-9]"

Slash

two digits

two digits

# Aside: Advanced Statistical Techniques

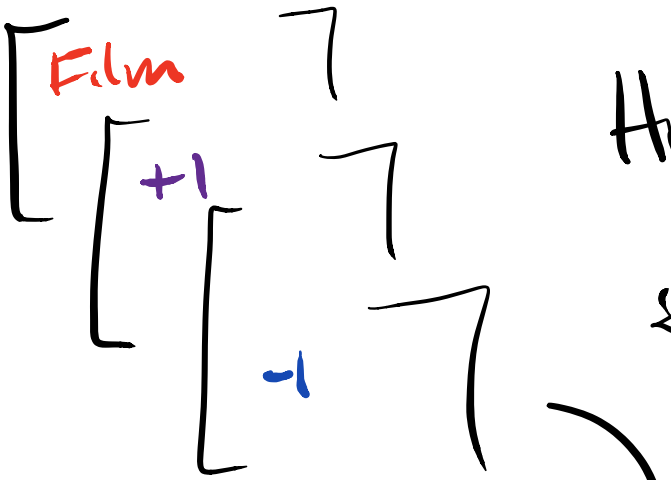
⇒ Patterns not provided by analyst

## "Natural Language Processing"

- Sentiment Analysis
  "I had a great time" → +1
- Topic Analysis
  "I went to see Marvel" → "Cinema"

NB. NLP "Text" —Machine→ Interpretation

# Aside: NLP




Historical Docs with Human Labelled Answers.

find relationship

Eg. WORD FREQUENCIES

$$f(x_{FILM}, x_{SCIFI}, x_{...}) = y$$

# What are Regular Expressions?

a language for describing text patterns

used to...

- Match
- Extract
- Replace.

# Examples: #1 Repeated Nums.

"My BP is 120 mmHg; I'm paying £10.00 for magnesium"

REPEATED

[0-9]+ mm Hg

Number "mmHg"

# Examples: #2

"My food came to: €100.50, I spent €10.00 on the taxi!"

REPEATED REPEATED

:€[0-9]+.[0-9]+

":€" Num Num

# Regular Expression Syntax

| Pattern | Means | Eg |
| --- | --- | --- |
| a | a | Abba |
| € | € | €10.00 |
| . | any single symbol | Hello |
| [012] | any single sym. ∈ 0, 1, 2 | |
| [0-9] | any single sym. ∈ 0, ... 9 | |
| [a-z] | " " " lowercase | |
| [A-Z] | " " " Uppercase. | |

Most symbols are "Literal"

a single Symbol!

[^012] ← "NOT"

any single symbol NOT: 0, 1, 2

€10 for A

ba$ ← END

occuring at the End

Abba Abba

^Ab ← START

occuring at the Start

Abba Abba

(May | June)

↑ Alternative

EITHER: May OR June

I'm considering May, June or July

# Quantifier → Quantity

"How Many?"

Note: prior patterns DO NOT REPEAT!

ANY PATTERN (FROM ABOVE)

"optional"

p* o!* ZERO OR MORE Hello Hello!!!

p+ o!+ ONE OR MORE Hello Hello!!!

p? o!? ZERO OR ONE Hello Hello!!!

"optional"

! means '!'

# Example: #3 Form Fields

"Name: Michael; Age: 31; Location: ;"

:[a-zA-Z0-9]* ; ← Can be Messy

Examples which Match:
- :;
- :a;
- :aaaa;
- :01a;

# Aside: Special & literal

Note that '.' means 'ANY symbol'

and '$' means 'END'

... So how can we say "literally" $ or "Literally" . ?

TWO WAYS

1. \. OR [.]