[a-z] : a...z

[0-9] : 0-9

[^Ss] : neither S nor s I ... etc.

States? : the previous expression is optional

colou?r : color or colour

oo*h! : 0 or more previous char

o+h! : 1 or more previous char

FP : incorrectly catches the words.

FN : incorrectly misses the words

```
/$[0-9]+/

/$[0-9]+\.[0-9][0-9]/

/$[0-9]+(\.[0-9][0-9])?/

/(^|\W)$[0-9]+(\.[0-9]
                 [0-9])?/
```

```
/[0-9]+GB/

/[0-9]+\.[0-9]+GB/

/[0-9]+(\.[0-9]+)?GB/

/(^|\W)[0-9]+(\.[0-9]+)?GB
                              ␣*/
```

/\b[5-9][0-9]{2,}(\.[0-9]+)?

␣* GB \b

500

600

1000 ✗

3 digits

( [5-9][0-9][0-9] |

4 or more

[1-9][0-9]+{3,} )

/(?: Some | a few) (people | cats)

like some \1/

?: as long as the ?: are there,

it won't be registered.

/[A-Za-z]+/

/[a-z]+b/

/b*(ab)*b*/

```
([A-Za-z]+)\b(\1)
```

13 Blue

```
/[0-9]+[a-zA-Z]+/

/\b[0-9]+[a-zA-Z]+\b/

(^|\b)[0-9]+[a-zA-Z]+(\b|$)
/
```

grotto , raven

\b [^a-zA-Z]grotto [^a-zA-Z]\b

[A-Za-z]

Word types = # of vocab.

Word instances = Total # of running words.

$$\underbrace{|V|}_{\text{\# of types}} = k\,\underbrace{N}_{\text{\# of } \underset{\text{instances}}{\underline{\text{running words}}}}^{\beta}$$

# Word Normalisation

U.S.A / USA

am, is, be, are

## Lemmatization

Lemma = shared rule of words

## Morphemes

= Smallest meaningful units that make up words

Morphemes

- stems – core meaning
- affixes – adher to stems

Stemming (chopping off affixes)

simplifies ver. of lemmetization

Porter Stemmer

ATIONAL → ATE ,

ING → $\epsilon$

SSES

Sentence seg:

! ? is very obvious, but:

"." (period) can be used in Dr. Inc.
(etc...)

→ So, tokenise them first:

- either part of the word
- sentence boundary,

then sentence seg.

Minimum Edit Distance :

to measure how similar two strings are.

→ used for spell correction.

$$P(W_3 \mid W_1 , W_2)$$

$$\frac{C(W_1, W_2, W_3)}{C(W_1, W_2)}$$

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| N | 9 | | | | | | | | | |
| O | 8 | 7 | 8 | 9 | 10 | 11 | 10 | 9 | 8 | 9 |
| I | 7 | 6 | 7 | 8 | 9 | 10 | 9 | 8 | 9 | 10 |
| T | 6 | 5 | 6 | 7 | 8 | 9 | 8 | 9 | 10 | 11 |
| N | 5 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 10 |
| E | 4 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 9 |
| T | 3 | 4 | 5 | 6 | 7 | 8 | 7 | 8 | 9 | 8 |
| N | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 7 | 8 | 7 |
| I | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 6 | 7 | 8 |
| # | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| | # | E | X | E | C | U | T | I | O | N |

XECU

~~I~~~~N~~~~T~~E~~N~~TION $4+4=8$

→ EXECUTION .

$$\sqrt[N]{\frac{1}{P(a_1, a_2)}} \quad - \; -$$

$$P(am \mid I)$$

$$= \frac{P(I\ am)}{P(I)} = \frac{\frac{2}{3}}{4}$$

$$P(c) = \frac{c(\overset{w_1,\, w_2}{\qquad}) + 1}{c(w_2) + V}.$$

just plain boring (−)

entirely predictable and lack energy (−)

no surprises and very few laughs (−)

very powerful (+)

the most fun film of the summer (+)

---

predicatable ~~with~~ no fun

$c(-) = 14$ $P(-) =$

$c(+) = 9$ $P(+) =$

and, very, the

$$P(\text{predictable} \mid (-))$$

$$= \frac{C((-), \text{predictable})+1}{C((-))+20} = \frac{1+1}{14+20}$$

$$= \frac{1}{17}$$

$$P(\text{predictable} \mid (+))$$

$$= \frac{C(+, \text{pre})+1}{C((+))+20} = \frac{0+1}{9+20}$$

$$= \frac{1}{29}$$

$$P(\text{no} \mid -) = \frac{1+1}{14+20} = \frac{1}{17}$$

$$P(\text{no} \mid +) = \frac{0+1}{9+20} = \frac{1}{29}$$

$$P(\text{fun} \mid -) = \frac{0+1}{14+20} = \frac{1}{34}$$

$$P(\text{fun} \mid +) = \frac{1+1}{9+20} = \frac{2}{29}$$

$(-)$

$$\left(\frac{1}{17} \times \frac{1}{17} \times \frac{1}{34}\right) \times \frac{3}{5} = 6.1 \times 10^{-5}$$

$$(+) \quad \left(\frac{1}{29} \times \frac{1}{29} \times \frac{2}{29}\right) \times \frac{2}{5} \approx \underline{3.2 \times 10^{-5}}$$

$P(\text{neg}) = \frac{3}{8}$ $c(-) = 9$

$P(\text{mid}) = \frac{2}{8} = \frac{1}{4}$ $c(N) = 9$

$P(\text{pos}) = \frac{3}{8}$ $c(+) = 7$

$V = 24$

Not compelting enough

not interesting enough (-)

failed to impress (-)

dull and uninspiring (-)

it was okay (N)

neither good nor bad, just average (N)

quite compelling (+)

exceptionally ~~good~~ (+/

a thrilling experience (+)

Not

$$P(\text{Not} \mid -) = \frac{1+1}{9+24} = \frac{2}{33}$$

$$P(\text{Not} \mid N) = \frac{0+1}{9+24} = \frac{1}{33}$$

$$P(\text{Not} \mid +) \quad \frac{0+1}{7+24} = \frac{1}{31}$$

---

Compelling

$$P(\text{Com} \mid -) = \frac{1}{9+24} = \frac{1}{33}$$

$$P(\text{com} \mid N) = \frac{1}{9+24} = \frac{1}{33}$$

$$P(\text{com} \mid +) = \frac{1+1}{7+24} = \frac{2}{31}$$

$$P(\text{enough} \mid (-)) = \frac{1+1}{9+24} = \frac{2}{33}$$

$$P(\text{enough} \mid (N)) = \frac{1}{33}$$

$$P(\text{enough} \mid (+)) = \frac{1}{31}$$

$$P(\text{test} \mid -) = \frac{4}{33^3} \times \frac{3}{8}$$

$$= 4.17 \times 10^{-5}$$

$$\mid N) = \frac{1}{33^3} \times \frac{2}{8}$$

$$= 6.95 \times 10^{-6}$$

$$\mid +) = \frac{1}{31^3} \times \frac{3}{8}$$

$$= 1.25 \times 10^{-5}$$

$(3, 2, 1, 3, 0, 4.19)$

$(-5, 2.5, -1.2, 0.5, 2.0, 0.7)$

$(-15, 5, -1.2, 1.5, 0, 2.933)$

$-10 + \cancel{(-1.2 + 1.5)} + 2.933$

$0.3$

$3.333 - 10$

$-6.7$

$-0.67$