Lecture - 04 Empherical Laws

Tuesday, 28. January 2025 03:50

Concept — Type Instance of concept — Token eg: Will Will Type: 1 Token: 2

Type - Token Ratio (TTR) : average a new word appears in text

High TTR : tendency to use new words
Low TTR : tendency to use repeated

Problem with length: 30000 tks | 60000 tks repeated
[no. of token ↑ linearly but words do not]
what can be done ? → Running average*X

First: (1 – 1000) + (3 – 1003) average:

Word Frequency, freq of freq :

1	3993	3993 words in corpus with freq 1
>100	102	102 words in corpus with freq more than 1.

Zipf's Law:-

(i) Count the frequency of each word
(ii) List them in decreasing order

eg: (i) the on and Tom
4000 1000 2000 1500

(ii) [the, and, Tom, on] words
[4000, 2000, 1500, 1000] Freq
[1, 2, 3, 4] Rank
) R^n by Zipf's Law

$$\boxed{f \propto \frac{1}{r} \quad \therefore \quad f \cdot r = K \quad (\text{Frequency} * \text{rank} = K)}$$

R	F
50	f_1
150	f_2

$f_1 \cdot 50 = K$
$f_2 \cdot 150 = K$

$f_1 \cdot 50 = f_2 \cdot 150$
$\boxed{f_1 = 3 f_2}$

$$\boxed{P_r = \frac{F_r}{N} \quad \frac{(\text{No. of word})}{\text{Total}}}$$

" Let P_r denote the probability of word of rank r & N denote the total numbers of word occurences. "

Zipf tell that the :
$f \propto \frac{1}{r}$ $\frac{f}{N} = \frac{A}{r}$ $\boxed{K = A \cdot N}$ $A \approx 0.1$ //.

Sage's Law :

The number of meaning m of a word obeys the law :

$$m \propto \sqrt{f} \Rightarrow m \propto \sqrt{\frac{1}{r}} \text{ or } m \propto \frac{1}{\sqrt{r}}$$

⇒ Rank ≈ 1000, avg 2.1 meanings
Rank ≈ 5000, avg 3 meanings

Zip Law :

Word Frequency is inversely proportional to their length.

$f \propto \frac{1}{l}$ short word → more frequency

Heaps Law :

How the size of overall vocab grow with the size of the corpus ?

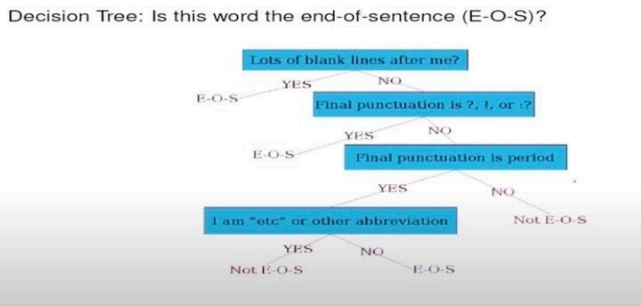
Let $|V|$ be the size of vocab & N be the numb of tokens.

$$|V| = kN^{\beta}$$

Tokenization :

Process of segmenting a string of char → words

Prob 1. Where words start & end.
Approach : End of sent / Not End of sent (Binary classifier)


Decision Tree: Is this word the end-of-sentence (E-O-S)?
Lots of blank lines after me?
YES
NO
E-O-S
Final punctuation is ?, !, or :?
YES
NO
E-O-S
Final punctuation is period
YES
NO
Not E-O-S
I am "etc" or other abbreviation
YES
NO
Not E-O-S
E-O-S

(i) EOL parameters : get unique token
(ii) EOL Hyphen & Lexical Hyphen (co-)
(iii) Problem with compound words

Normalization:

U.S.A and USA should be matched.

Case Folding :

All cases to lower.

Lemmatization :

can't → cant
won't → wont } How ↗

Morphology :

Stem : The core meaning unit
Affix : Bits & pieces adhering to stems

Porter's Algorithm : Step 1 a :

Porter's Algorithm: Step 1a:

- $sses \rightarrow ss$ (caresses → caress)
- $ies \rightarrow i$ (ponies → poni)
- $ss \rightarrow ss$ (caress → caress)
- $s \rightarrow \phi$ (cats → cat)

Size of unique word: $V = K \times N^{\beta}$ $\left.\begin{array}{l} N = \text{total words} \\ K = \text{const.} \end{array}\right\}$ Heap's Law