Introduction to Natural Language Processing, Assignment 1

Enrique Mesonero Ronco Sergio Sánchez García Ismael

November 24, 2024

Contents

1 Text Processing

1.1 Tokenization

- First of all, it is necesary to lay down the initial base vocabulary:

$$V_B = \{a, b, c, r, t\}$$

$$V_W(\text{as per} V_P) = \{b, a, t : 10; b, a, r : 5; c, a, t : 8; c, a, r : 4; c, a, r, t : 6\}$$

- Secondly, merge tokens based on frequency:

$$"c" + "a" \text{occur the most, 18 times in total}$$

$$V_B = \{a, b, c, r, t, ca\}$$

$$V_W = b, a, t : 10; b, a, r : 5; ca, t : 8, ca, r : 4; ca, r, t : 6$$

1.2 Levenshtein Distance

	_	H	U	N	D
_	0	1	2	3	4
H	1	0	1	2	3
A	2	1	1	2	3
N	3	2	2	1	2
D	4	3	3	2	1
Y	5	4	4	3	2

hund → handy Total → 1 change operation + 1 add operation → Levensthein Distance = 2

	_	N	A	T	T	Y
_	0	1	2	3	4	5
G	1	1	2	3	4	5
R	2	2	2	3	4	5
I	3	3	3	3	4	5
T	4	4	4	3	3	4
T	5	5	5	4	3	4
Y	6	6	6	5	4	3

natty → gritty Total → 2 change operation + 1 add operation → Levensthein Distance = 3

2 Words and the Company They Keep

2.1 Pearson's Chi-sqaure Test

	B $= b_1$	$B = b_2$	Total
$A = a_1$	9	1770	1779
$A = a_2$	75	219243	219318
Total	84	221013	221097

$$E_i j = \frac{f(w_i) \cdot f(w_j)}{N}$$

	$B = b_1$	$B = b_2$
$A = a_1$	$\frac{84 \cdot 1779}{221097}$	$\frac{1779 \cdot 221013}{221097}$
$A = a_2$	$\frac{84 \cdot 219318}{221097}$	$\frac{221013 \cdot 219318}{221097}$

	$B = b_1$	$B = b_2$
$A = a_1$	0.676	1778.32
$A = a_2$	83.32	219234.6759

$$\chi^2 = \sum_{i,j} \frac{(O_i, j - E_i, j)^2}{E_i, j}$$

$$\chi^2 = \frac{(9 - 0.676)^2}{0.676} + \frac{(1770 - 1778.32)^2}{1778.32} + \frac{(75 - 83.32)^2}{83.32} + \frac{(219243 - 219234.6759)^2}{219234.6759} = 102.5 + 0.04 + 0.83 + 3.16 \cdot 10^{-4} = 103.37$$

2.2 PMI: Pointwise Mutual Information