Working With Text

Pittsburgh useR Group - 04/24/2018

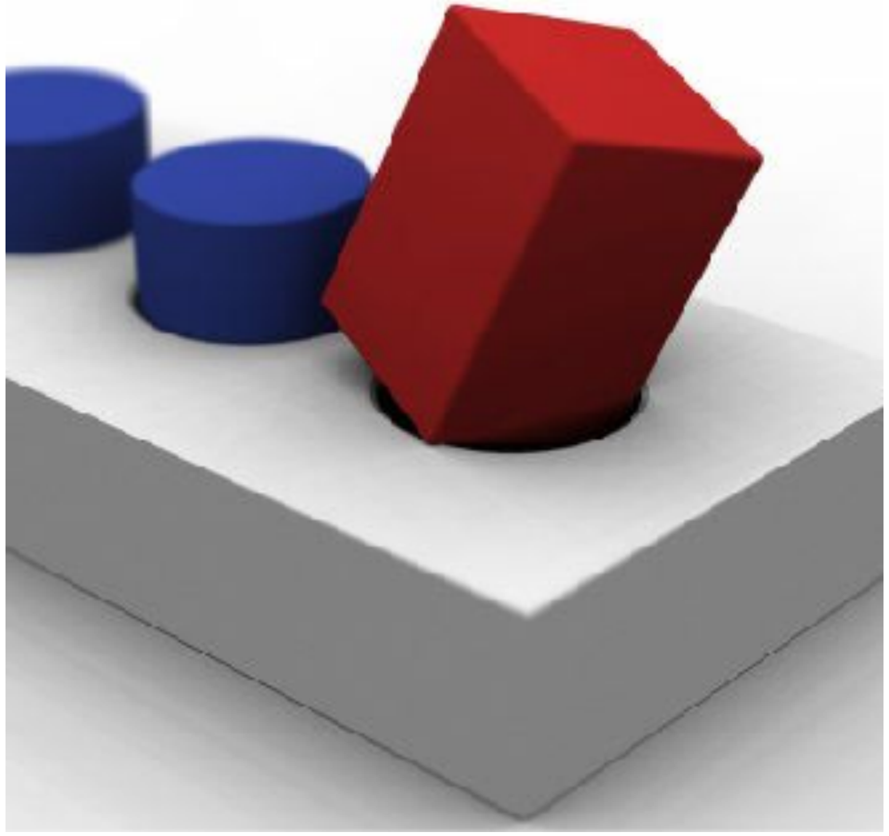
Michael Patnik
AvenueFour Analytics

Avenue 4 Analytics

Challenges

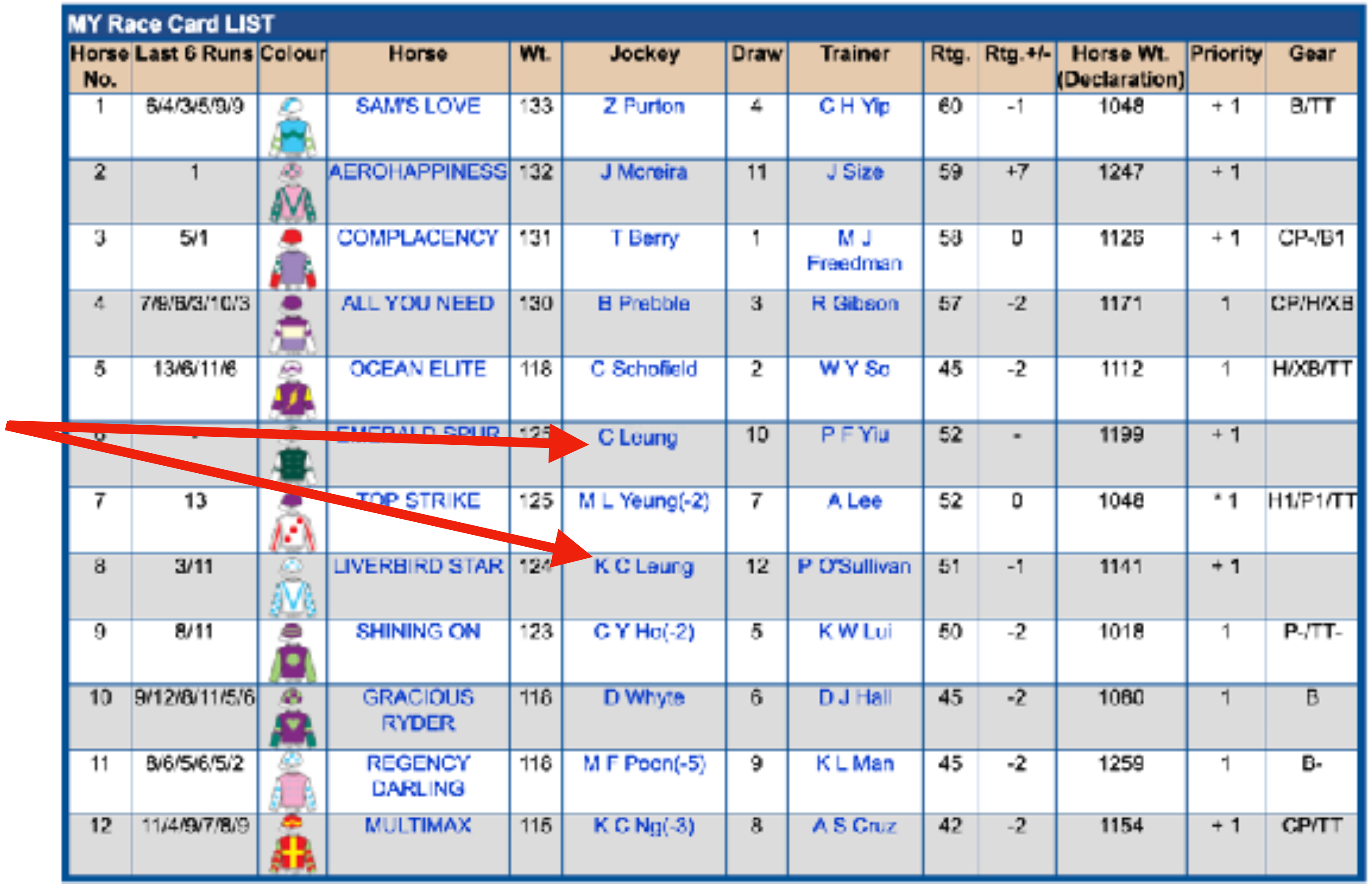
Text Is dirty!

Incompatible with most statistical models.



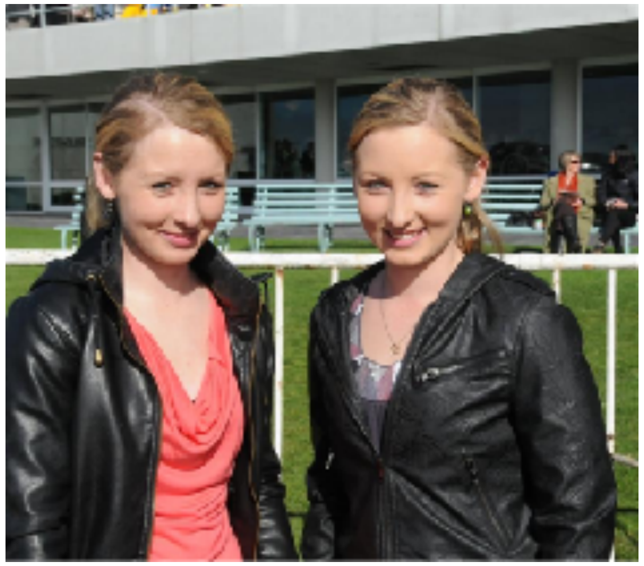
Correctly Identifying Entities

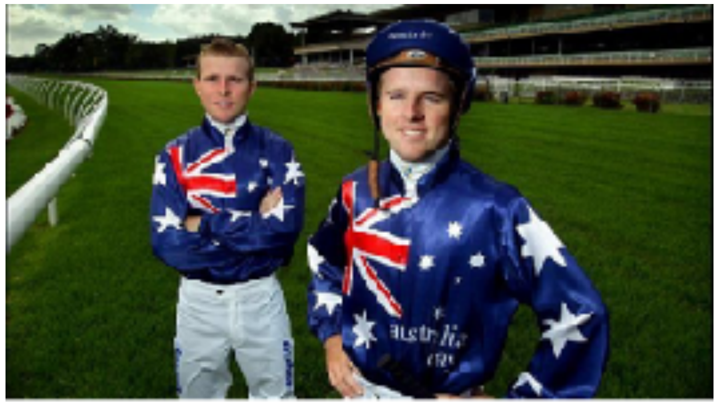
???



MY Race Card LIST

Horse No.	Last 6 Runs	Colour	Horse	Wt.	Jockey	Draw	Trainer	Rtg.	Rtg.+/-	Horse Wt. (Declaration)	Priority	Gear
1	6/4/3/5/9/9		SAM'S LOVE	133	Z Purton	4	C H Yip	60	-1	1048	+ 1	B/TT
2	1		AEROHAPPINESS	132	J Moreira	11	J Size	59	+7	1247	+ 1	
3	5/1		COMPLACENCY	131	T Berry	1	M J Freedman	58	0	1126	+ 1	CP-/B1
4	7/8/8/3/10/3		ALL YOU NEED	130	B Prebble	3	R Gibson	57	-2	1171	1	CP/H/XB
5	13/6/11/6		OCEAN ELITE	118	C Schofield	2	W Y So	45	-2	1112	1	H/XB/TT
6	[illegible]		EMERALD SPUR	125	C Leung	10	P F Yiu	52	-	1199	+ 1	
7	13		TOP STRIKE	125	M L Yeung(-2)	7	A Lee	52	0	1048	* 1	H1/P1/TT
8	3/11		LIVERBIRD STAR	124	K C Leung	12	P O'Sullivan	51	-1	1141	+ 1	
9	8/11		SHINING ON	123	C Y Ho(-2)	5	K W Lui	50	-2	1018	1	P-/TT-
10	9/12/8/11/5/6		GRACIOUS RYDER	116	D Whyte	6	D J Hall	45	-2	1080	1	B
11	8/6/5/6/5/2		REGENCY DARLING	116	M F Poon(-5)	9	K L Man	45	-2	1259	1	B-
12	11/4/9/7/8/9		MULTIMAX	115	K C Ng(-3)	8	A S Cruz	42	-2	1154	+ 1	CP/TT





Similarity Algorithms

Levenshtein Distance

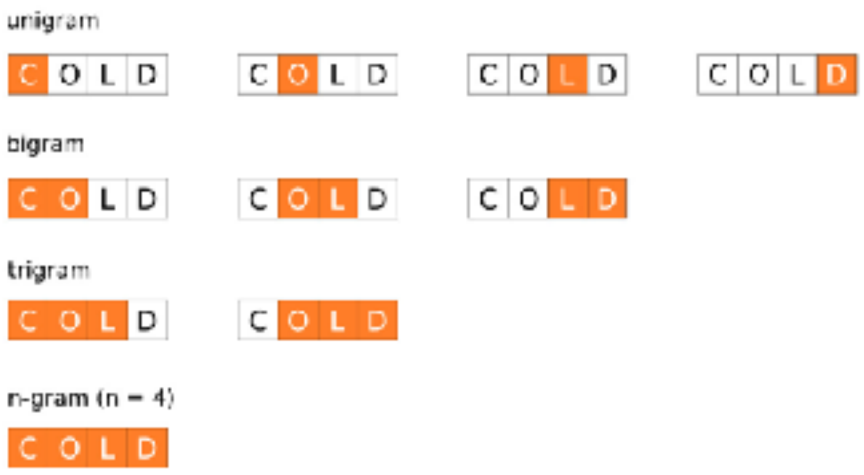
- minimum number of edits necessary to change one string into another

$$\mathrm{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if} \min(i,j)=0, \\ \min \begin{cases} \mathrm{lev}_{a,b}(i-1,j)+1 \\ \mathrm{lev}_{a,b}(i,j-1)+1 \\ \mathrm{lev}_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$$

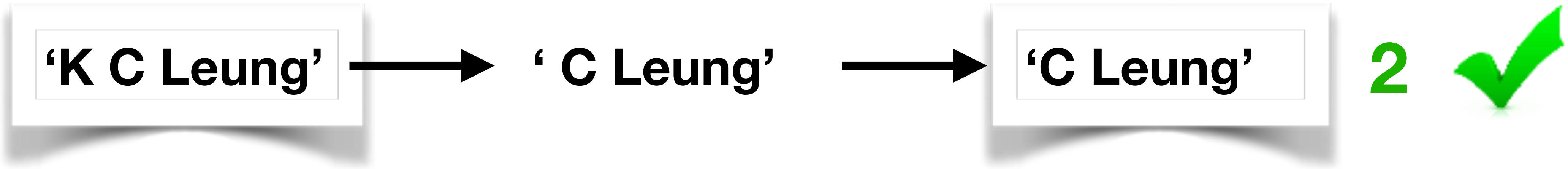
Jaccard Index (using ngrams)

- count of the intersection of ngrams divided by count of the union of ngrams

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

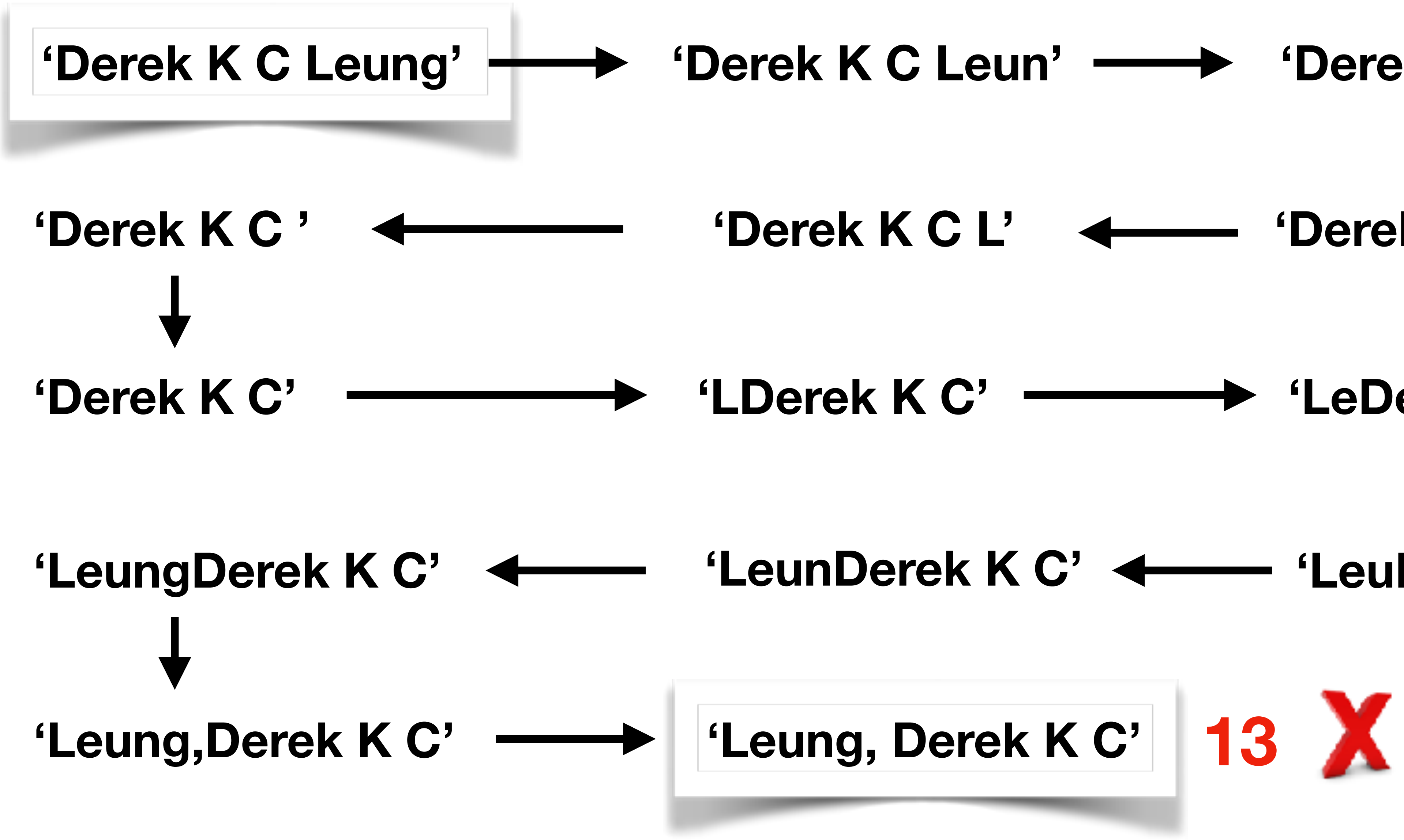


Levenshtein Distance



‘K C Leung’ → ‘ C Leung’ → ‘C Leung’ 2 ✓

Levenshtein Distance



‘Derek K C Leung’

‘Derek K C Leun’

‘Derek K C Leu’

‘Derek K C Le’

‘Derek K C L’

‘Derek K C ’

‘Derek K C’

‘LDerek K C’

‘LeDerek K C’

‘LeuDerek K C’

‘LeunDerek K C’

‘LeungDerek K C’

‘Leung,Derek K C’

‘Leung, Derek K C’

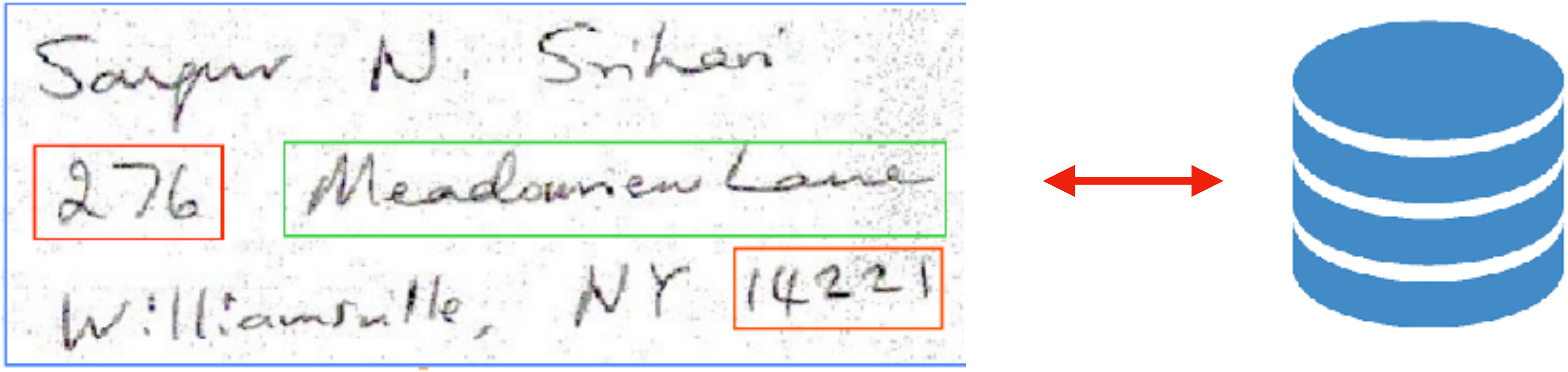
13 ✗

Jaccard Index (using ngrams)

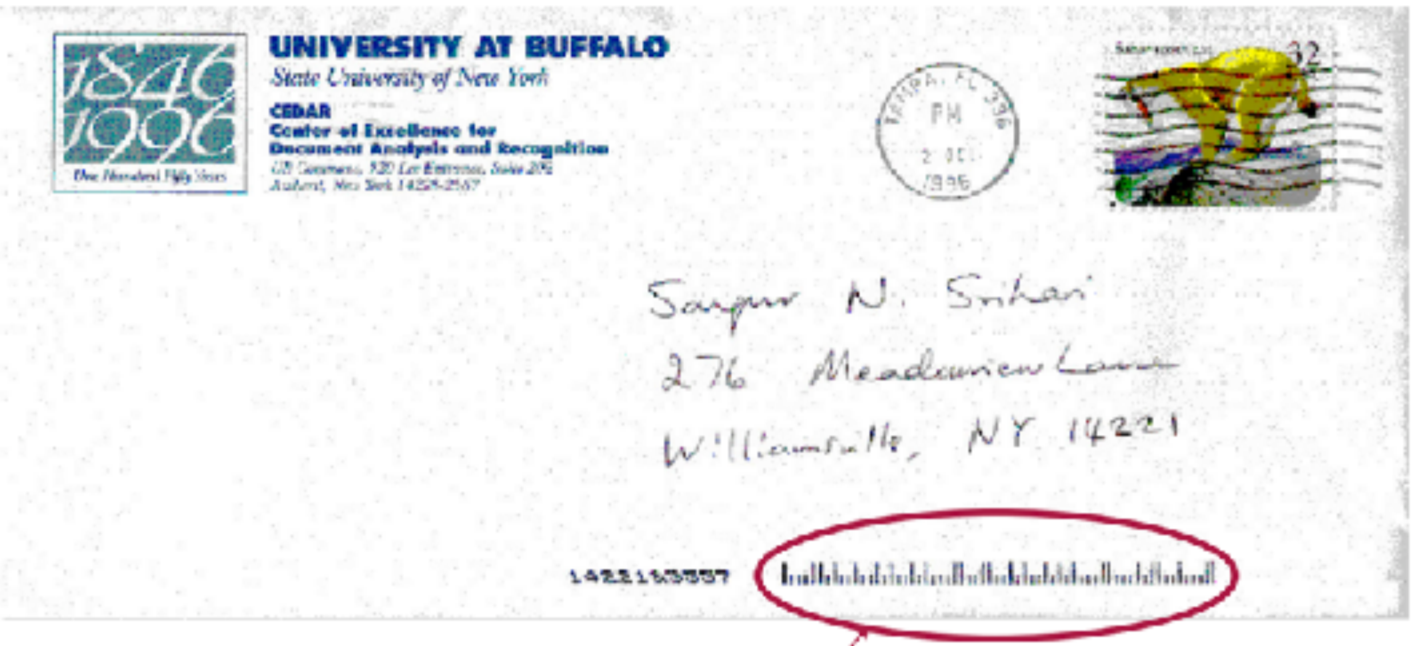
‘K C Leung’	‘K ‘, ‘ C’, ‘C ‘, ‘ L’, ‘Le’, ‘eu’, ‘un’,’ng’	Intersection: ‘C ‘, ‘ L’, ‘Le’, ‘eu’, ‘un’,’ng’ Union: ’K ‘, ‘ C’, ‘C ‘, ‘ L’, ‘Le’, ‘eu’, ‘un’,’ng’ Score: 6 / 8 = 0.75
‘C Leung’	‘C ‘, ‘ L’, ‘Le’, ‘eu’, ‘un’,’ng’	
‘Derek K C Leung’	‘De’, ‘er’, ‘re’, ‘ek’, ‘k ‘, ‘ K’, ’K ‘, ‘ C’, ‘C ‘, ‘ L’, ‘Le’, ‘eu’, ‘un’,’ng’	Intersection: ‘De’, ‘er’, ‘re’, ‘ek’, ‘k ‘, ‘ K’, ’K ‘, ‘ C’, ‘Le’, ‘eu’, ‘un’,’ng’ Union: ‘De’, ‘er’, ‘re’, ‘ek’, ‘k ‘, ‘ K’, ’K ‘, ‘ C’, ‘C ‘, ‘ L’, ‘Le’, ‘eu’, ‘un’,’ng’, ‘g,’, ‘, ’, ‘ D’ Score: 12 / 17 = 0.71
‘Leung, Derek K C’	‘Le’, ‘eu’, ‘un’,’ng’, ‘g,’, ‘, ’, ‘ D’, ‘De’, ‘er’, ‘re’, ‘ek’, ‘k ‘, ‘ K’, ’K ‘, ‘ C’	

Leverage Domain Knowledge and Other Data Elements

Case Study: USPS



Sargur N. Srihari
276 Meadowview Lane
Williamsville, NY 14221



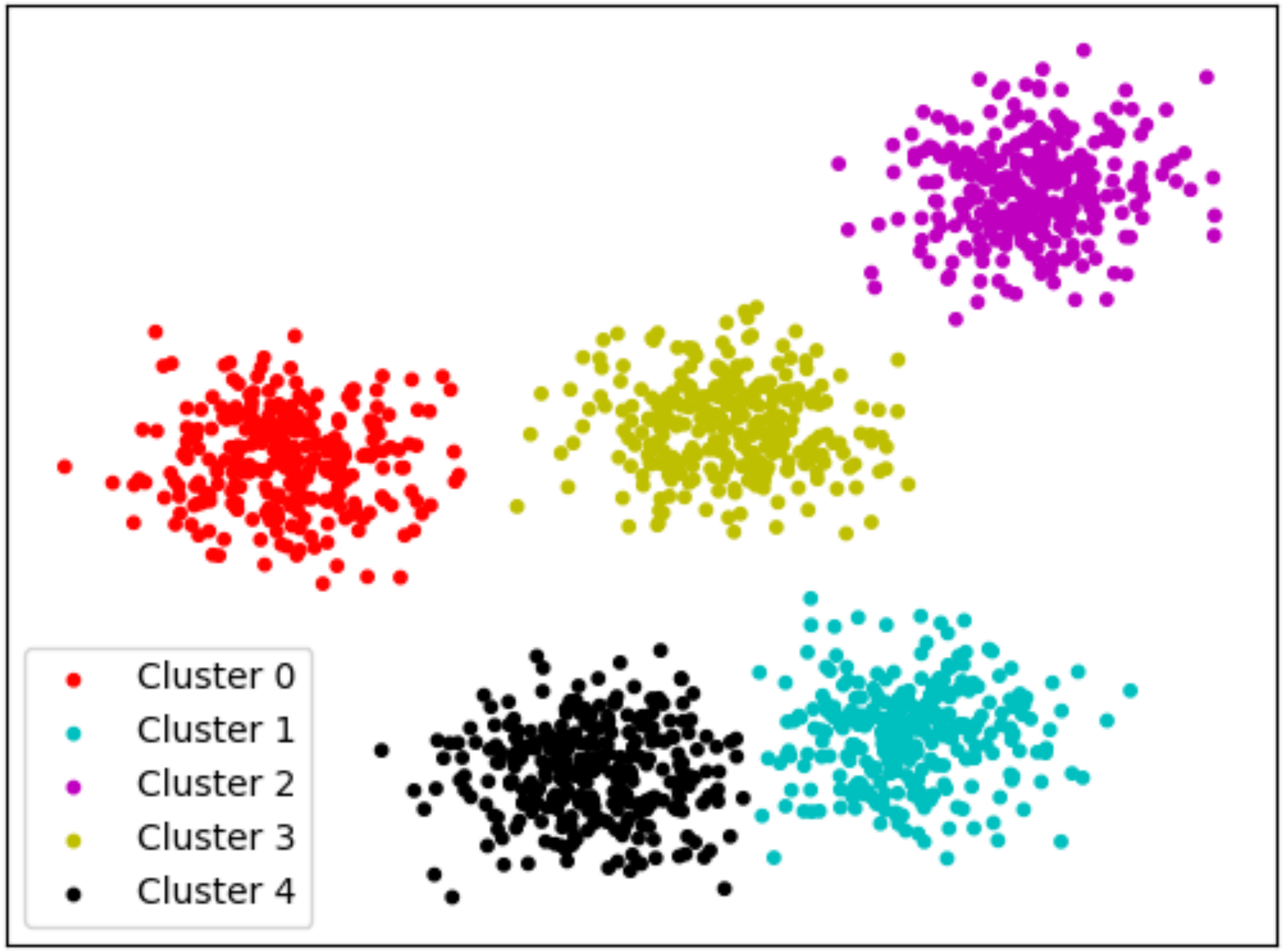
UNIVERSITY AT BUFFALO
State University of New York
CEDAR
Center of Excellence for
Document Analysis and Recognition
UB Commons, 520 Lee Entrance, Suite 202
Amherst, New York 14228-2567

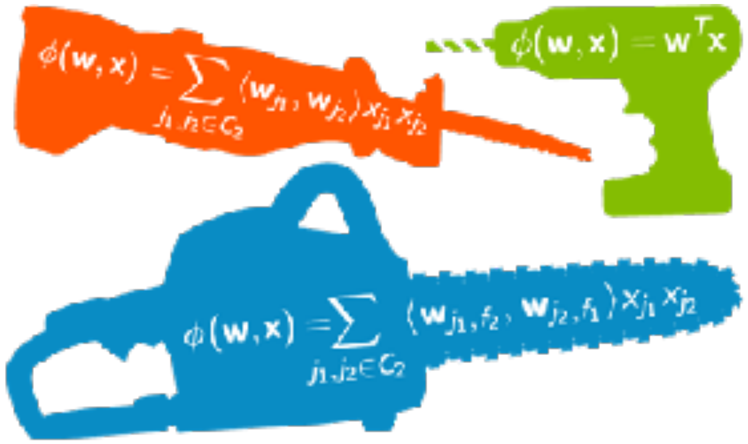
Sargur N. Srihari
276 Meadowview Lane
Williamsville, NY 14221

14221163557

Postnet Bar Code representing Delivery Point

Feature Engineering - Clustering

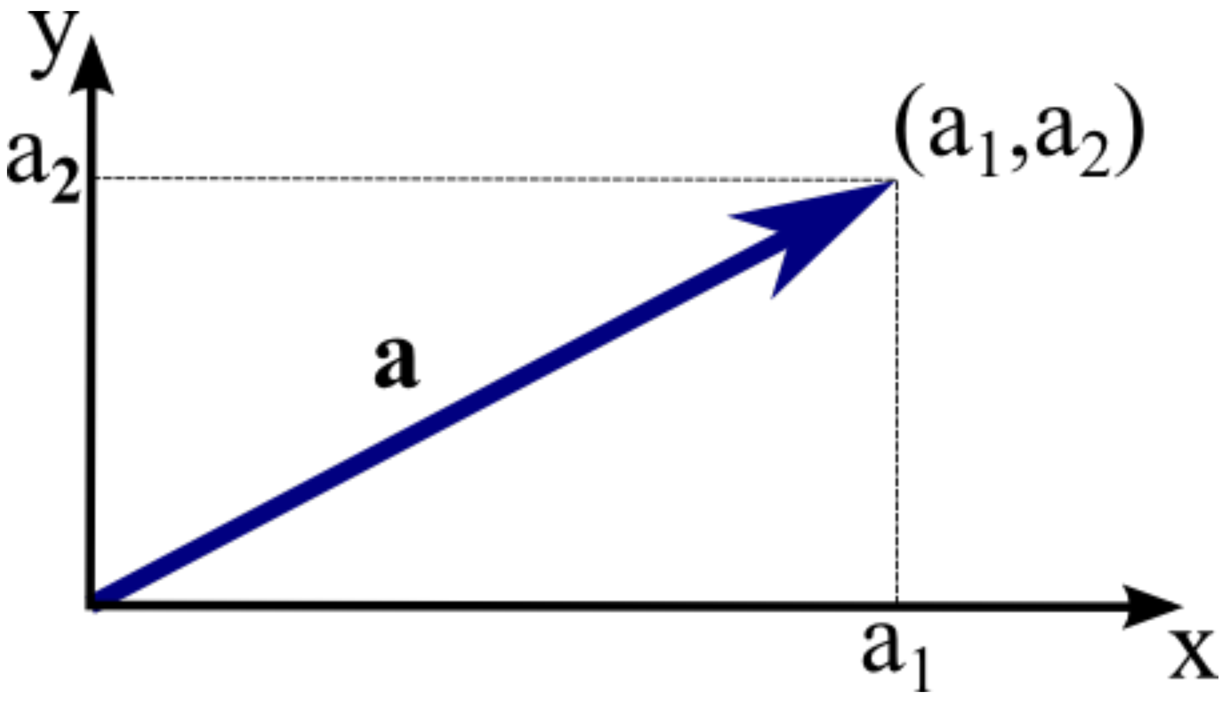



$\phi(\mathbf{w},\mathbf{x}) = \sum_{j_1,j_2 \in C_2} (\mathbf{w}_{j_1}, \mathbf{w}_{j_2}) x_{j_1} x_{j_2}$
$\phi(\mathbf{w},\mathbf{x}) = \mathbf{w}^T\mathbf{x}$
$\phi(\mathbf{w},\mathbf{x}) = \sum_{j_1,j_2 \in C_2} (\mathbf{w}_{j_1,f_2}, \mathbf{w}_{j_2,f_1}) x_{j_1} x_{j_2}$

Text to Vector

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents



Example: Greyhound Comments

Thank You Questions?

Michael Patnik
mike@avenuefour.com

Avenue 4 Analytics