Sheetal Parikh
Problem Set – Module 2
605.744.81 Information Retrieval
Fall 2021

1.) *Character n-gram overlap is used for both automated spelling correction and personal name matching (i.e., deciding whether two names might be the same, a common database problem known as "record linkage"). Using a character 3- gram representation, how many distinct n-grams do "CHEONGSONG" and "CHEONMACHONG" have in common? What is the Dice-coefficient score for these two strings using 3-grams? What is the Dice score using 4-grams instead? Which score is higher?*

3-grams

a.) *Distinct 3-grams of CHEONGSONG:*
CHE
HEO
EON
ONG
NGS
GSO
SON
ONG

b.) *Distinct 3-grams of CHEONMACHONG:*
CHE
HEO
EON
ONM
NMA
MAC
ACH
CHO
HON
ONG

c.) *Distinct 3-grams in common:*
CHE
HEO
EON
ONG

d.) *Dice-coefficient score:*
X = number of 3-grams for CHEONGSONG = 8
Y = number of 3-grams for CHEONMACHONG = 10
$X \cap Y$ = number of 3-grams in common = 4

$$Dice - coefficient\ score = \frac{2|X \cap Y|}{|X| + |Y|} = \frac{2 * 4}{8 + 10} = \frac{8}{18} \approx 0.4444$$

4-grams

a.) *Distinct 4-grams of CHEONGSONG:*
CHEO
HEON
EONG
ONGS
NGSO
GSON
SONG

b.) *Distinct 4-grams of CHEONMACHONG:*
CHEO
HEON
EONM
ONMA
NMAC
MACH
ACHO
CHON
HONG

c.) *Distinct 4-grams in common:*
CHEO
HEON

d.) *Dice-coefficient score:*
X = number of 4-grams for CHEONGSONG = 7
Y = number of 4-grams for CHEONMACHONG = 9
$X \cap Y$ = number of 4-grams in common = 2

$$Dice - coefficient\ score = \frac{2|X \cap Y|}{|X| + |Y|} = \frac{2 * 2}{7 + 9} = \frac{4}{16} = 0.2500$$

The dice-coefficient score using 3-grams is higher than the dice-coefficient score using 4-grams

2.) *Compute the edit distance (or Levenshtein distance) for these two pairs of strings: (a) "CHEBYSHEV" and "TSCHEBYSCHEF"; and (b) "LEVINSTINE" and "LEVENSHTEIN". Then report a sequence of transformations for that cost that converts one string into the other. You should use unit costs for each operation: insertion, deletion, or substitution; that is, each step has a cost of 1.*

a.) *Edit Distance: Chebyshev and Tschebyschef*

Edit Distance Table:

	" "	C	H	E	B	Y	S	H	E	V
" "	0	1	2	3	4	5	6	7	8	9
T	1	1	2	3	4	5	6	7	8	9
S	2	2	2	3	4	5	5	7	8	9

C	*3*	*2*	*3*	*3*	*4*	*5*	*6*	*6*	*7*	*8*
H	*4*	*3*	*2*	*3*	*4*	*5*	*6*	*6*	*7*	*8*
E	*5*	*4*	*4*	*2*	*5*	*5*	*6*	*7*	*6*	*7*
B	*6*	*6*	*5*	*3*	*2*	*3*	*4*	*5*	*6*	*7*
Y	*7*	*7*	*7*	*6*	*4*	*2*	*3*	*4*	*5*	*6*
S	*8*	*8*	*8*	*8*	*7*	*3*	*2*	*3*	*4*	*5*
C	*9*	*8*	*9*	*9*	*9*	*4*	*3*	*3*	*4*	*5*
H	*10*	*9*	*8*	*8*	*9*	*5*	*4*	*3*	*4*	*5*
E	*11*	*10*	*9*	*8*	*9*	*6*	*5*	*4*	*3*	*4*
F	*12*	*11*	*10*	*9*	*9*	*7*	*6*	*5*	*4*	*4*

Based on the edit distance table above, we can see how that the edit distance between CHEBYSHEV and TSCHEBYSCHEF is 4. To convert CHEBYSHEV to TSCHEBYSCHEF we would need to perform the following steps:

1.) Insert T: Cost = 1
2.) Insert S: Cost = 1
3.) The letters "CHEBYS" are the same in both strings so there is no cost in keeping the letters as is.
4.) Insert C: Cost = 1
5.) Replace V to F: Cost = 1

The total cost after the three insert and one replace is 4.

b.) *Edit Distance: Levinstine and Levenshtein*

	" "	**L**	**E**	**V**	**I**	**N**	**S**	**T**	**I**	**N**	**E**
" "	0	1	2	3	4	5	6	7	8	9	10
L	1	0	1	2	3	4	5	6	7	8	9
E	2	2	0	1	2	3	4	5	6	7	7
V	3	3	1	0	1	2	3	4	5	6	7
E	4	4	1	1	1	2	3	4	5	6	6
N	5	4	2	2	2	1	2	3	4	4	5
S	6	5	3	3	3	2	1	2	3	4	5
H	7	6	4	4	4	3	2	2	3	4	5
T	8	7	5	5	5	4	3	2	3	4	5
E	9	8	5	6	6	6	5	3	3	4	4
I	10	9	6	6	6	7	6	4	3	4	5
N	11	10	7	7	7	6	7	5	4	3	4

Based on the edit distance table above, we can see how that the edit distance between LEVINSTINE and LEVENSHTEIN is 4. To convert LEVINSTINE to LEVENSHTEIN we would need to perform the following steps:

1.) The letters "Lev" are the same in both strings so there is no cost in keeping the letters as is.
2.) Replace I to E: Cost = 1
3.) The letters "NS" are the same in both strings so there is no cost in keeping the letters as is.

4.) Insert H: Cost = 1
5.) The letter, “t”, is the same in both strings so there is no cost in keeping the letters as is.
6.) Insert E: Cost = 1
7.) The letter, “IN”, is the same in both strings so there is no cost in keeping the letters as is.
8.) Delete E: Cost = 1

The total cost after one replace step, two inserts, and one delete step is 4.

3.) *Following the method described in the textbook (or lecture materials), what are the Soundex codes for the strings: (a) "Daniels" and (b) “Hrabowski”? Show intermediate steps to produce the final code.*

a.) *DANIELS = D542*

Steps to convert “DANIELS” to a Soundex code:

I. D remains D since we retains the first letter of the word
II. A becomes 0
III. N becomes 5
IV. I becomes 0
V. E becomes 0
VI. L becomes 4
VII. S becomes 2
VIII. We then have D050042. After removing all the zeroes from the resulting string, we are left with the 4-digit Soundex Code: D542

b.) *HRABOWSKI = H612*

Steps to convert “HRABOWSKI” to a Soundex code:

I. H remains H since we retains the first letter of the word
II. R becomes 6
III. A becomes 0
IV. B becomes 1
V. O becomes 0
VI. W becomes 0
VII. S becomes 2
VIII. K becomes 2
IX. I becomes 0
X. We then have H60100220. After removing all the zeroes from the resulting string, we have: H6122
XI. The code has two consecutive 2’s which can be replaced with only the first 2. We then have the 4-digit Soundex code of: H612