

Student Name: _____ Student Number: _____

Question 1: Consider the following small part of a **positional** index with the format:

word: doc#: <position, position,...>; doc#: <position,...>. Example:
good: 2: <6>; 3: <2,12>; 4: <9,17>

Means: word "Good" occurs in document 2 at position 6, in document 3 at locations 2 and 12 and in document 4 at locations 9 and 17.

Black: 1: <1>; 2: <6>; 3: <2,15>; 4: <9>.
Sheep: 3: <4,11>; 4: <3>; 7:<14,89>.
Fish: 1: <2>; 2: <12, 16,21>; 3: <13>; 5: <21,25>.

The **/d** operator, with the format: **word1 /d word2** finds occurrences of **word1** at a distance at most **d**=1 words on either side of **word2**. Thus, **d=1** demands that **word1** is adjacent to **word2**.

- a. Give the set of documents that satisfy the query: **Black /2 Fish**.
{D1,D3}
- b. For which values of **d** the query: **Black /d Sheep** returns the set of documents {D3} as the answer.
d={2,3,4,5} [greater than or equal 2 but less than 6. At d=6 document 4 comes in and the set is no more {D1,D3}.

Question2: Find the Levinshtein distance between words "F1F2**L3**F3F4" and "L1L2**L3**L4" using the table below (or equivalent). Please mark all needed squares. L_i is the i^{th} letter of your last name, F_j is the j^{th} letter of your first name. **Example:** name is "adnan yahya" → $F1='a'$ $F2='d'$ **$L3='h'$** $F3='n'$ $F4='a'$ so word='adhna' and $L1='y'$ $L2='a'$ **$L3='h'$** $L4='y'$ and word2='yahy'

	-	F1=	F2=	L3=	F3=	F4=
-						
L1=						
L2=						
L3=						
L4=						

Test your case at the following link: <http://www.let.rug.nl/~kleiweg/lev/>

Compute the Jackard Similarity between "researcher" and "saercher" using **letter Bi-grams** (2-grams). **Don't worry about spaces before/after.**

"researcher"={re,es,se,ea,ar,rc,che,er}; "saercher"={sa,ae,er,rc,che,er}: size of "researcher"=9, different in "saercher" =2. Total: 11; Common /total=4/11=0.36