Approximate String matching

By Eric and Noor

The different purposes

Common usage

- Search engine
- Spellchecker

Academic

- Plagiarism detection
- DNA sequencing

Industry

- Database matching

The different purposes

Common usage

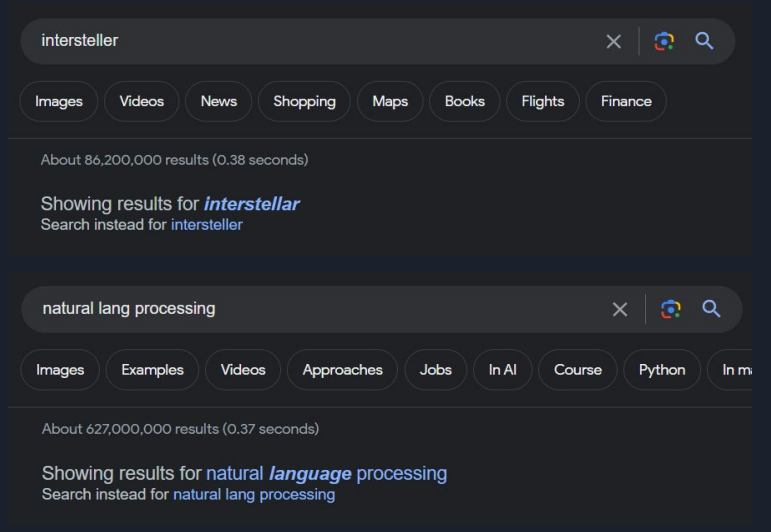
- Search engine
- Spellchecker

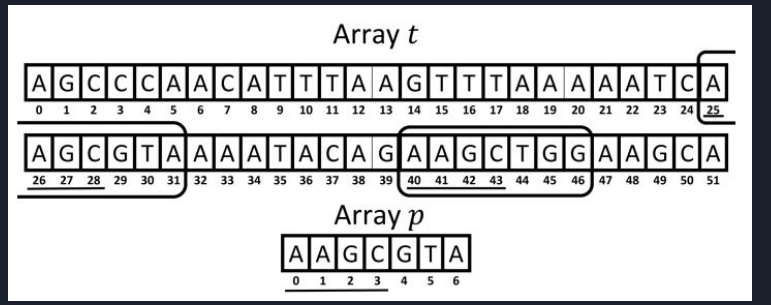
Academic

- Plagiarism detection
- DNA sequencing

Industry

- Database matching





Considerations

On-line vs off-line

Accuracy

Travel distance

Complexity

Considerations

On-line vs off-line

Accuracy

Travel distance

Complexity

Considerations

On-line vs off-line

Accuracy

Travel distance

Complexity

list_A	list_B
Accenture	Addobbe
Adobe	Accent llxure
Akamai Technologies	.
Alexandria Real Estate Equities	duP0nt+
Berkshire Hathaway	
Bio-Techne	Kalcommm
Biogen	cisco-systems
Boeing	cisco-systems

Considerations

On-line vs off-line

Accuracy

Travel distance

Complexity

Algorithms

O, Ω, and Θ notation

Naive - Θ(nm)

Hamming - Θ(n)

Levenshtein - Θ(n + d^2)

N-Gram

BK Tree

Bitap

Hamming Distance

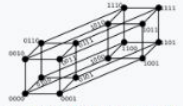
For a string S and T of the same length, the number of positions which are different

`S = “slide”`

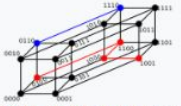
`T = “pride”`

Since `S[0]≠T[0]` and `S[1]≠T[1]` => The hamming distance is 2

Hamming distance



4-bit binary tesseract for finding Hamming distance.



Two example distances: 0100→1001 has distance 3; 0110→1110 has distance 1

Class	String similarity
Data structure	string
Worst-case performance	$O(n)$
Best-case performance	$O(1)$
Average performance	$O(n)$
Worst-case space complexity	$O(n)$

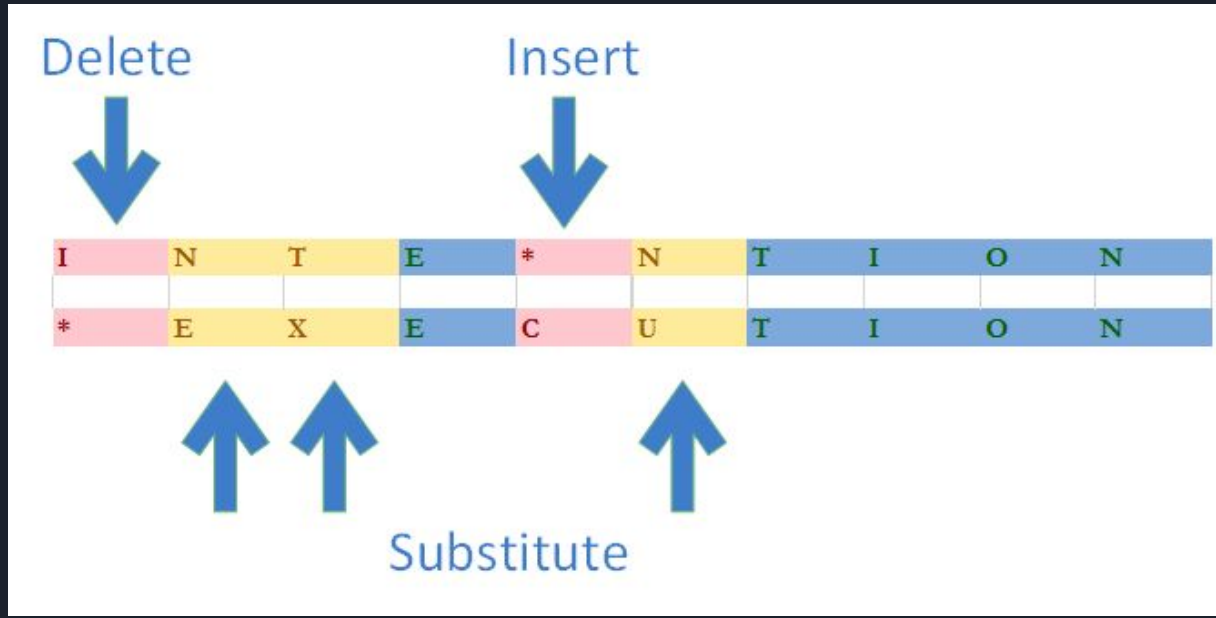
Levenshtein distance

Insertion

Deletion

Substitution

Edit distance



Levenshtein distance

Complexity O(|a| x |b|)

Upper boundary K

O(min(|a|, |b|) x K)

		S	Y	D	N	Y		M	E	Y	E	R
	0	1	2	3	4	5	6	7	8	9	10	11
S	1	0	1	2	3	4	5	6	7	8	9	10
Y	2	1	0	1	2	3	4	5	6	7	8	9
D	3	2	1	0	1	2	3	4	5	6	7	8
N	4	3	2	1	1	2	3	4	4	5	6	7
E	5	4	3	2	1	2	3	4	5	5	6	7
Y	6	5	4	3	2	1	2	3	4	5	6	7
	7	6	5	4	3	2	1	2	3	4	5	6
M	8	7	6	5	4	3	2	1	2	3	4	5
E	9	8	7	6	5	4	3	2	1	2	3	4
I	10	9	8	7	6	5	4	3	2	2	3	4
E	11	10	9	8	7	6	5	4	3	3	2	3
R	12	11	10	9	8	7	6	5	4	4	3	2

The Sauce

https://www.baeldung.com/cs/fuzzy-search-algorithm

https://www.researchgate.net/publication/31594565_Generalized_Hamming_Distance

https://en.wikipedia.org/wiki/Approximate_string_matching

https://en.wikipedia.org/wiki/Hamming_distance

https://www.geeksforgeeks.org/applications-of-string-matching-algorithms/

Conclusion

Strings are pretty cool😎