A decorative graphic on the left side of the slide consisting of two overlapping parallelograms. The front one is blue and the back one is a light green color. They are positioned diagonally, with the blue one partially covering the green one.

Approximate String matching



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

The image shows two Google search results. The first search is for 'interstellar', showing about 86,200,000 results in 0.38 seconds. The second search is for 'natural lang processing', showing about 627,000,000 results in 0.37 seconds. Below the searches is a diagram illustrating string matching between two DNA sequences, Array t and Array p.

Array t

A	G	C	C	C	A	A	C	A	T	T	T	A	A	G	T	T	T	A	A	A	A	A	T	C	A
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25

Array p

A	G	C	G	T	A	A	A	A	T	A	C	A	G	A	A	G	C	T	G	G	A	A	G	C	A
26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51

Array p

A	A	G	C	G	T	A
0	1	2	3	4	5	6



Considerations

On-line vs off-line

Accuracy

Travel distance

Complexity



Considerations

On-line vs off-line

Accuracy

Travel distance

Complexity

what pwerfol|

 what pwerfol — Search with DuckDuckGo

what **powerful king united ancient egypt**

what **powerful technique is king k known for**

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

Search results

This wiki is using a new search engine. ([Learn more](#))

Search

[Content pages](#) [Multimedia](#) [Translations](#) [Everything](#) [Advanced](#)

Did you mean: *andré emotions*



Algorithms

O , Ω , and Θ notation

Naive - $\Theta(nm)$

Hamming - $\Theta(n)$

Levenshtein - $\Theta(n + d^2)$

N-Gram

BK Tree

Bitap

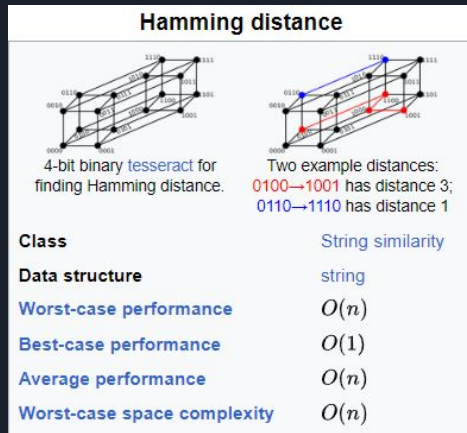
Hamming Distance

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

S = "slide"

T = "pride"

Since $S[0] \neq T[0]$ and $S[1] \neq T[1]$ \Rightarrow The hamming distance is 2



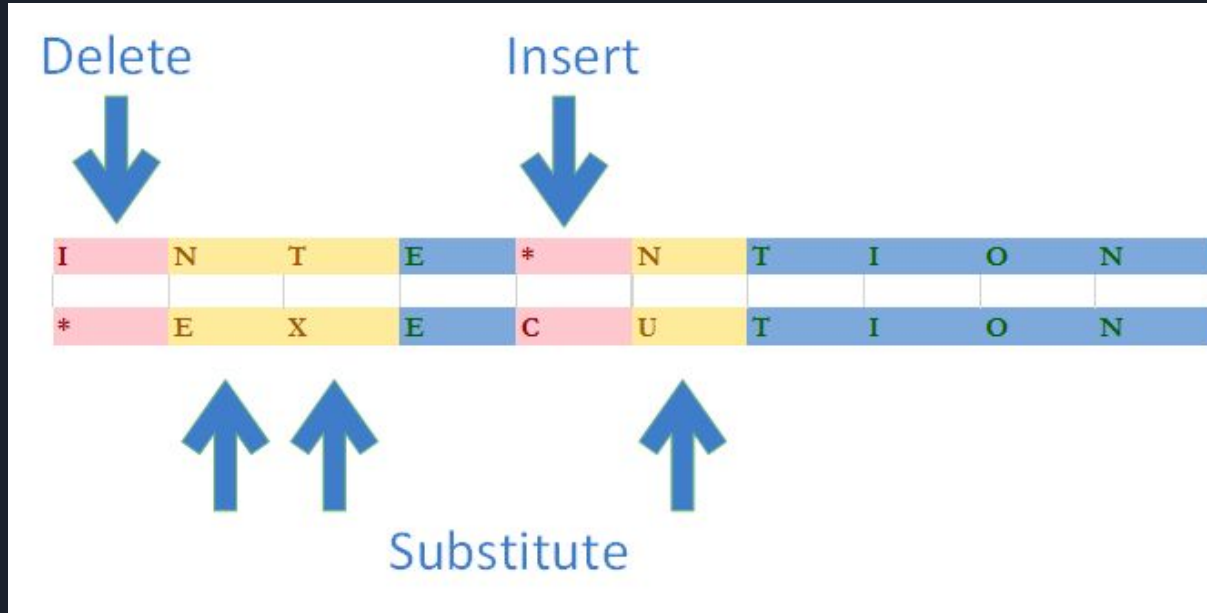
Levenshtein distance

Insertion

Deletion

Substitution

Edit distance



Levenshtein distance

Complexity $O(|a| \times |b|)$

Upper boundary K

$O(\min(|a|, |b|) \times K)$

		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/Approximate_string_matching)

[https://en.wikipedia.org/wiki/Hamming distance](https://en.wikipedia.org/wiki/Hamming_distance)

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



Conclusion

Strings are pretty cool 😎