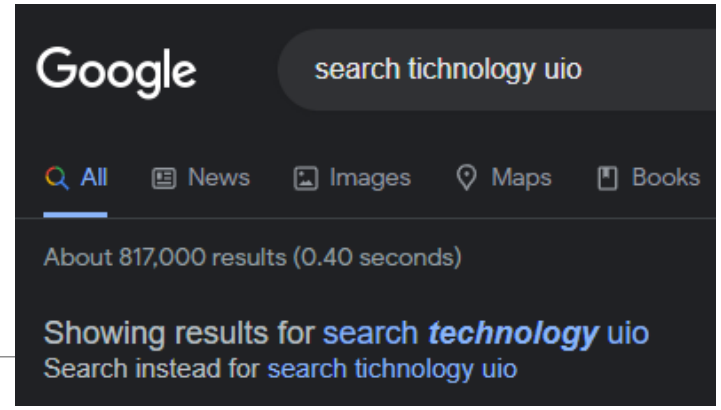
Tries for Approximate String Matching

by H. Shang and T. H. Merrett (1995)

A PRESENTATION BY ANNA AND JONAS

String matching



Applications: spell checking, searching and classification

EXACT

exact_match('abc', 'abc') → True

exact_match('abc', 'abd') → False

APPROXIMATE

apprx_match('abc', 'abc', k=1) → True

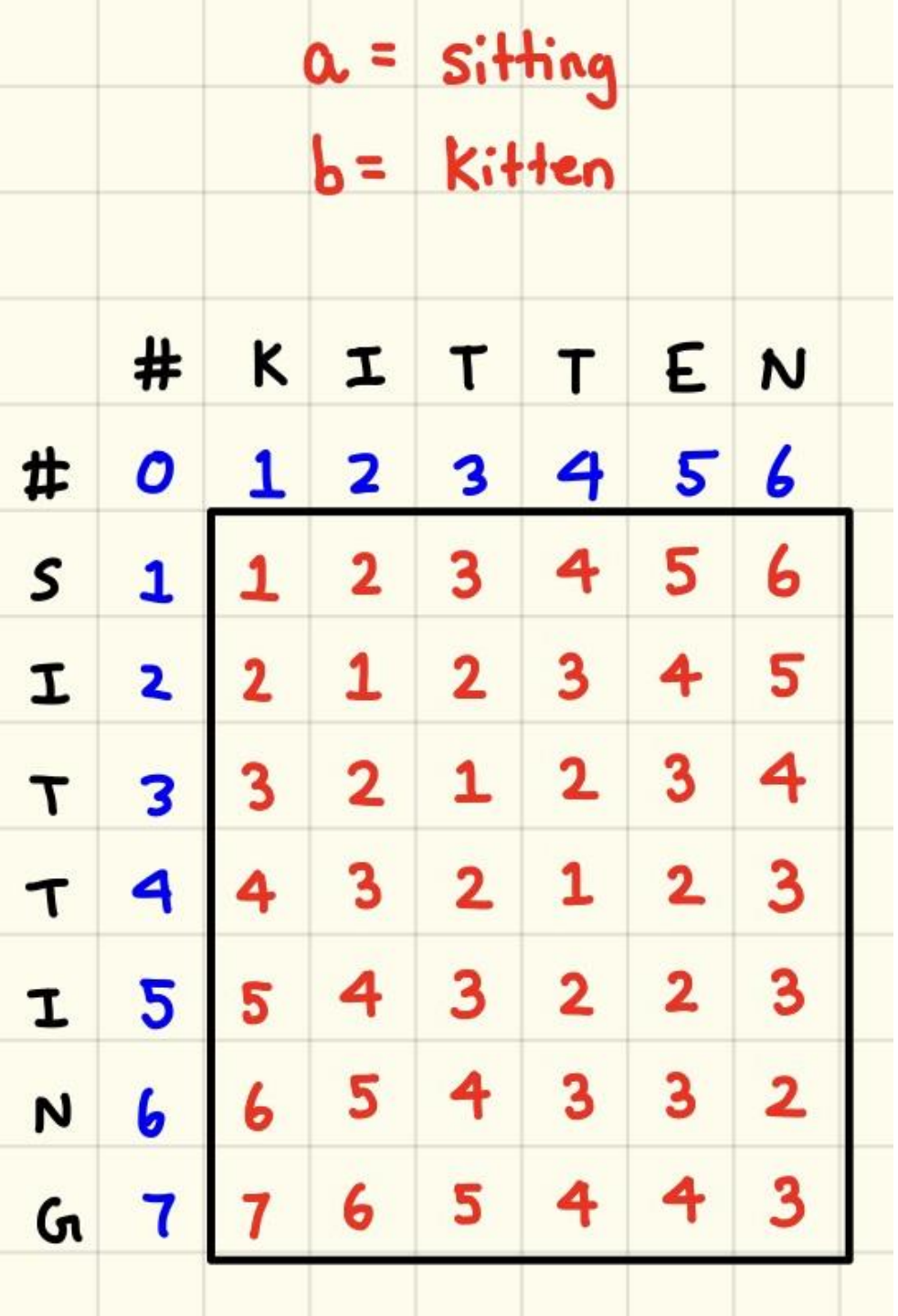
apprx_match('abc', 'abd', k=1) → True

The == operator in Python

```
>>> 'abc' == 'abc'
True
>>> 'abc' == 'abd'
False
```

k determines how approximate the match is

apprx_match('abc', 'abd', **k=0**) → False



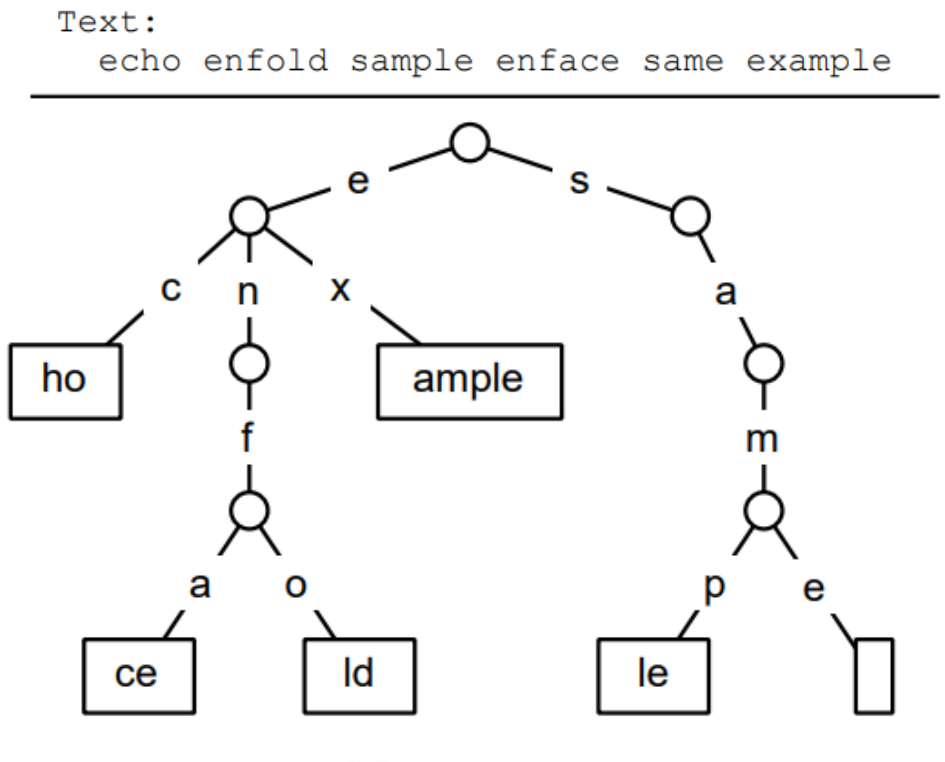
Edit distance

Number of changes/operations required for one word to become another

"How different are words *a* and *b*?"

Operations:

- Insertion: cot → co**a**t
- Deletion: co**a**t → cot
- Substitution: co**a**t → co**s**t
- Transposition: co**st** → co**ts**



Trie

Tree structure

Keys in leaf nodes

Each edge has one character of a key

Compressed because of shared prefixes

Shang and Merrett's algorithm

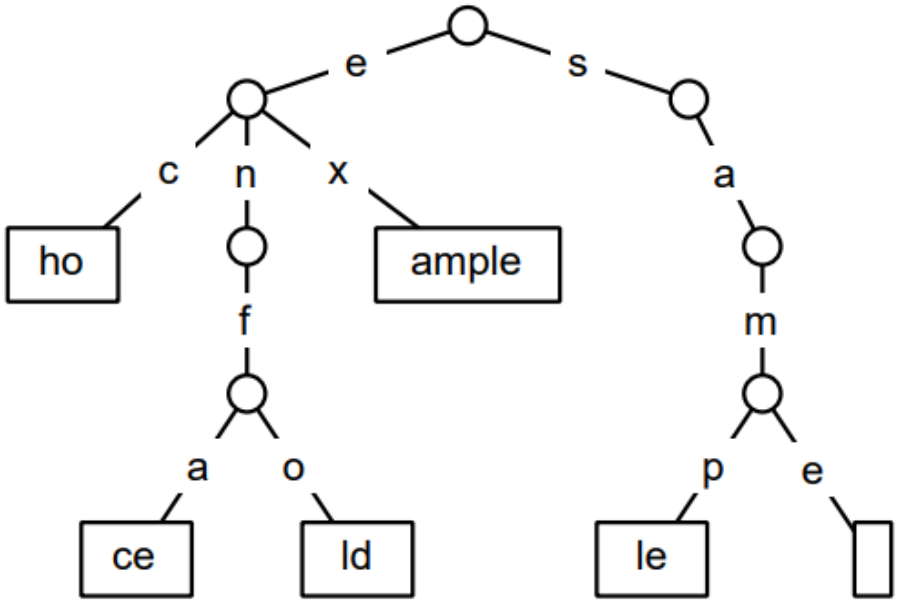
Traverse trie recursively with DFS (Depth First Search)

For each node/character

- Calculate edit distance
- If $editDistance > k$, skip subtrie

Search: 'sane'
$k = 1$

Text:
echo enfold sample enface same example



	ϕ	a	c	d	f	b	d	f
ϕ	0	1	2	3	4	5	6	7
a	1	0	1	2	3	4	5	6
d	2	1	1	1	2	3	4	5
f	3	2	2	2	1	2	3	4
d	4	3	3	2	2	2	2	3

Edit distance

	ϕ	a	c	d	f	b	d	f
ϕ	0	1	2	3	4	5		
a	1	0	1	2	3	4		
d	2	1	1	1	2	3		
f			2	2	1	2		
d						2		

Edit distance with Ukkonen cutoff

Shang and Merrett's algorithm

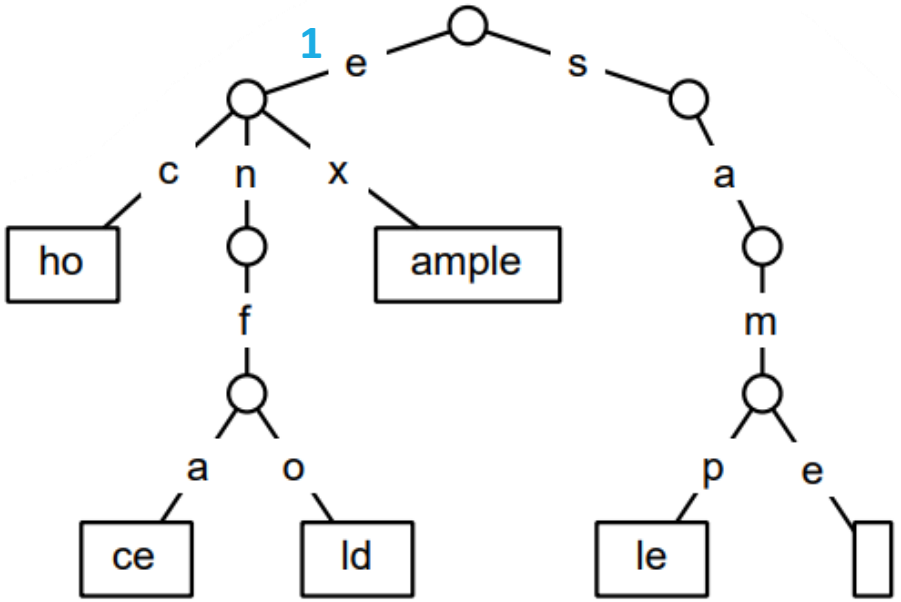
Traverse trie recursively with DFS (Depth First Search)

For each node/character

- Calculate edit distance
- If $editDistance > k$, skip subtrie

Search: 'sane'
$k = 1$

Text:
echo enfold sample enface same example



	ϕ	a	c	d	f	b	d	f
ϕ	0	1	2	3	4	5	6	7
a	1	0	1	2	3	4	5	6
d	2	1	1	1	2	3	4	5
f	3	2	2	2	1	2	3	4
d	4	3	3	2	2	2	2	3

Edit distance

	ϕ	a	c	d	f	b	d	f
ϕ	0	1	2	3	4	5		
a	1	0	1	2	3	4		
d	2	1	1	1	2	3		
f			2	2	1	2		
d						2		

Edit distance with Ukkonen cutoff

Shang and Merrett's algorithm

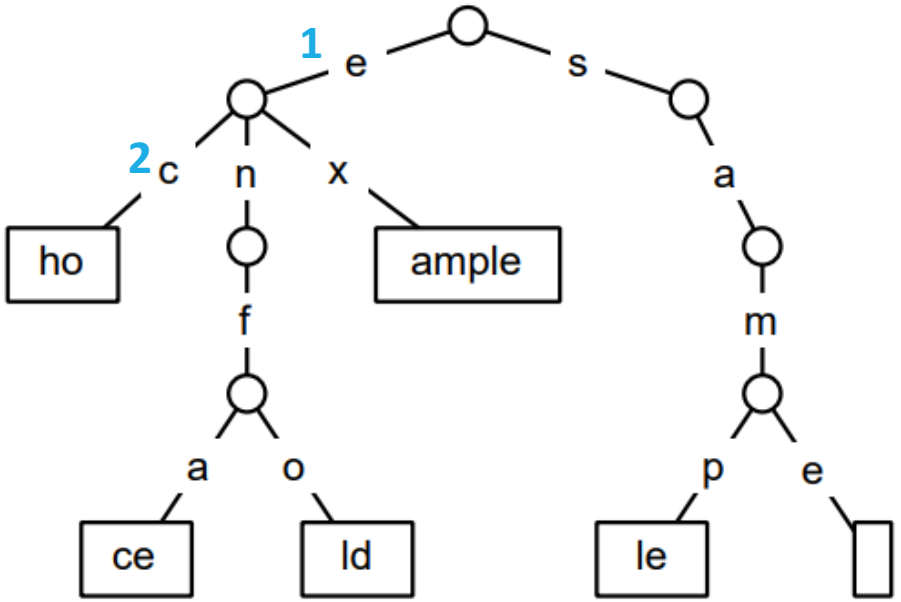
Traverse trie recursively with DFS (Depth First Search)

For each node/character

- Calculate edit distance
- If $editDistance > k$, skip subtrie

Search: 'sane'
$k = 1$

Text:
echo enfold sample enface same example



	ϕ	a	c	d	f	b	d	f
ϕ	0	1	2	3	4	5	6	7
a	1	0	1	2	3	4	5	6
d	2	1	1	1	2	3	4	5
f	3	2	2	2	1	2	3	4
d	4	3	3	2	2	2	2	3

Edit distance

	ϕ	a	c	d	f	b	d	f
ϕ	0	1	2	3	4	5		
a	1	0	1	2	3	4		
d	2	1	1	1	2	3		
f			2	2	1	2		
d						2		

Edit distance with Ukkonen cutoff

Shang and Merrett's algorithm

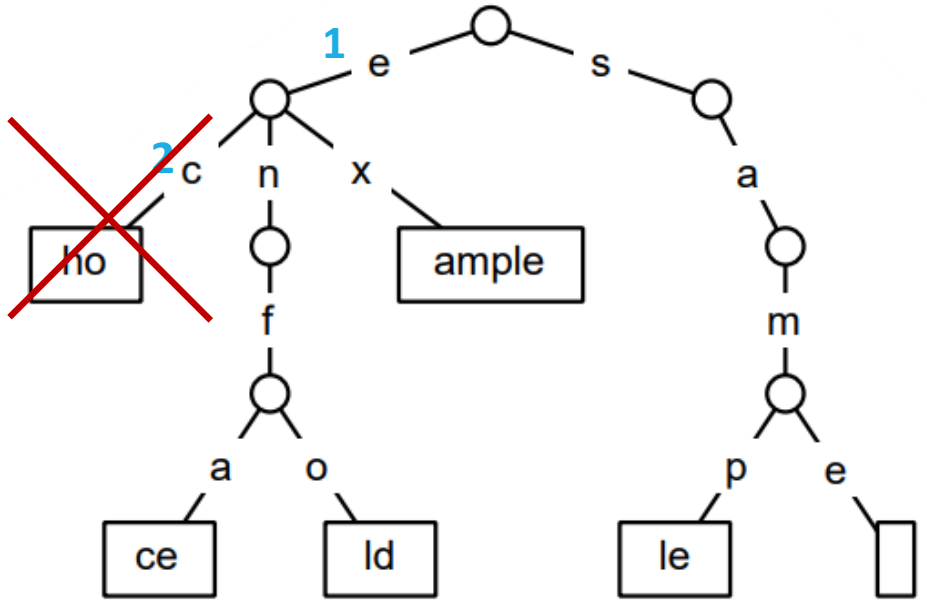
Traverse trie recursively with DFS (Depth First Search)

For each node/character

- Calculate edit distance
- If $editDistance > k$, skip subtrie

Search: 'sane'
$k = 1$

Text:
echo enfold sample enface same example



	ϕ	a	c	d	f	b	d	f
ϕ	0	1	2	3	4	5	6	7
a	1	0	1	2	3	4	5	6
d	2	1	1	1	2	3	4	5
f	3	2	2	2	1	2	3	4
d	4	3	3	2	2	2	2	3

Edit distance

	ϕ	a	c	d	f	b	d	f
ϕ	0	1	2	3	4	5		
a	1	0	1	2	3	4		
d	2	1	1	1	2	3		
f			2	2	1	2		
d						2		

Edit distance with Ukkonen cutoff

Shang and Merrett's algorithm

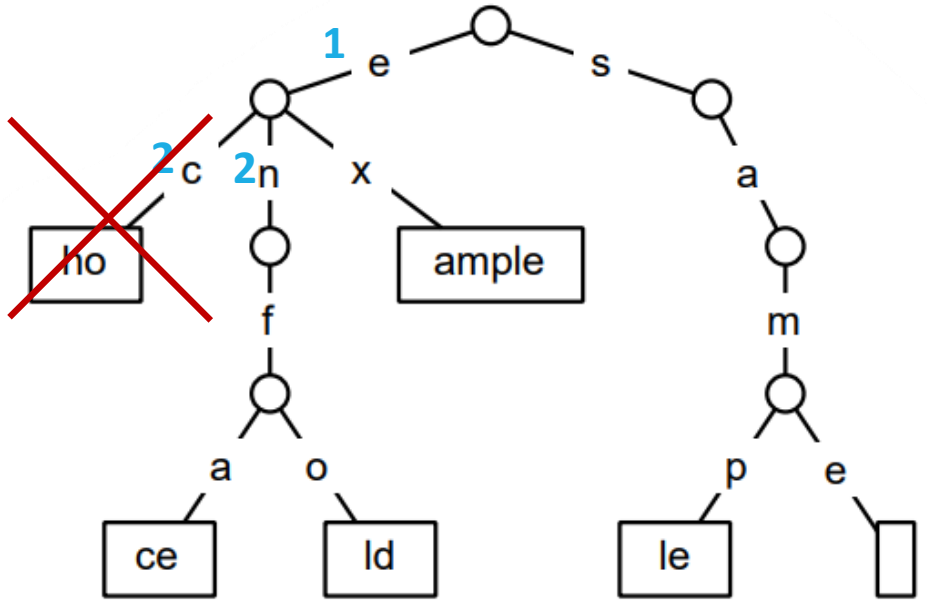
Traverse trie recursively with DFS (Depth First Search)

For each node/character

- Calculate edit distance
- If $editDistance > k$, skip subtrie

Search: 'sane'
$k = 1$

Text:
echo enfold sample enface same example



	ϕ	a	c	d	f	b	d	f
ϕ	0	1	2	3	4	5	6	7
a	1	0	1	2	3	4	5	6
d	2	1	1	1	2	3	4	5
f	3	2	2	2	1	2	3	4
d	4	3	3	2	2	2	2	3

Edit distance

	ϕ	a	c	d	f	b	d	f
ϕ	0	1	2	3	4	5		
a	1	0	1	2	3	4		
d	2	1	1	1	2	3		
f			2	2	1	2		
d						2		

Edit distance with Ukkonen cutoff

Shang and Merrett's algorithm

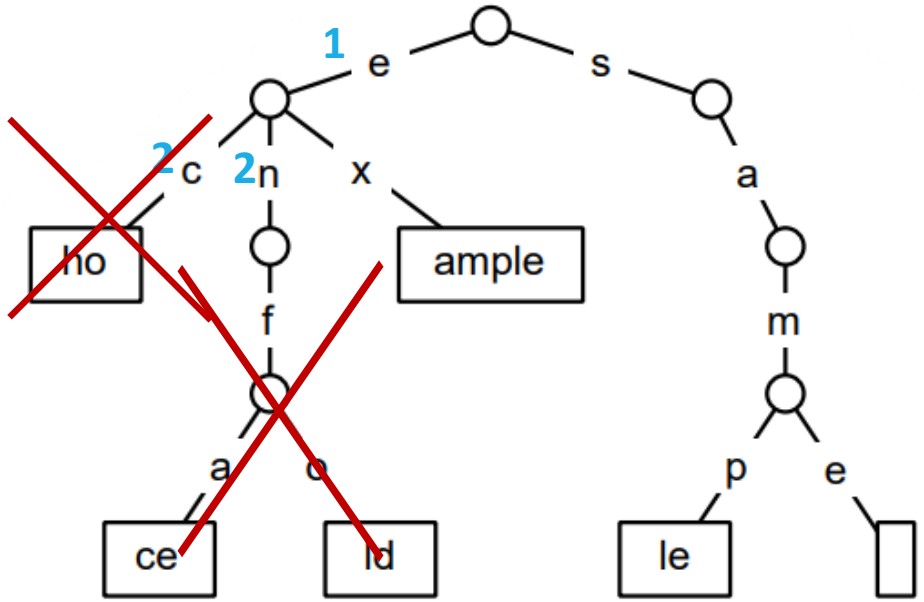
Traverse trie recursively with DFS (Depth First Search)

For each node/character

- Calculate edit distance
- If $editDistance > k$, skip subtrie

Search: 'sane'
$k = 1$

Text:
echo enfold sample enface same example



	ϕ	a	c	d	f	b	d	f
ϕ	0	1	2	3	4	5	6	7
a	1	0	1	2	3	4	5	6
d	2	1	1	1	2	3	4	5
f	3	2	2	2	1	2	3	4
d	4	3	3	2	2	2	2	3

Edit distance

	ϕ	a	c	d	f	b	d	f
ϕ	0	1	2	3	4	5		
a	1	0	1	2	3	4		
d	2	1	1	1	2	3		
f			2	2	1	2		
d						2		

Edit distance with Ukkonen cutoff

Shang and Merrett's algorithm

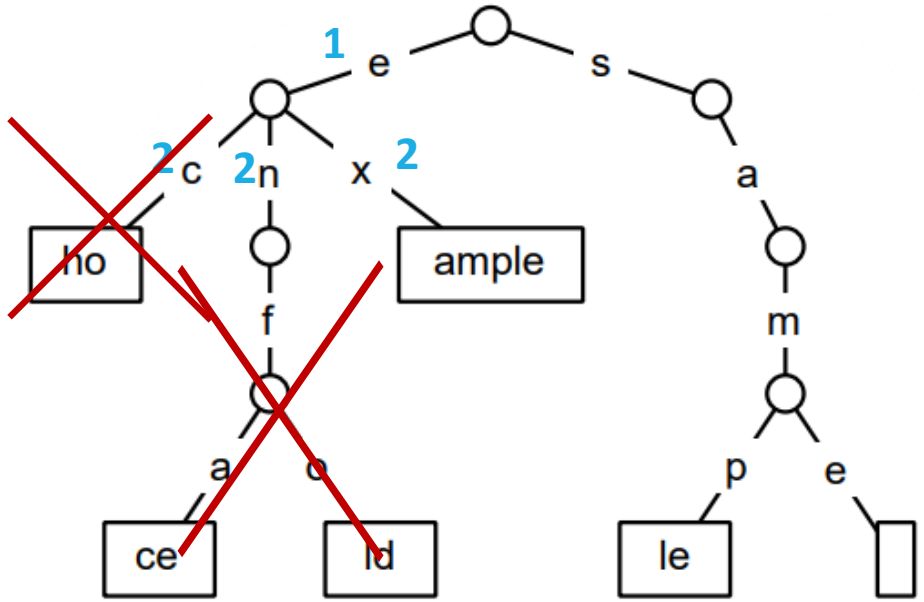
Traverse trie recursively with DFS (Depth First Search)

For each node/character

- Calculate edit distance
- If $editDistance > k$, skip subtrie

Search: 'sane'
$k = 1$

Text:
echo enfold sample enface same example



	ϕ	a	c	d	f	b	d	f
ϕ	0	1	2	3	4	5	6	7
a	1	0	1	2	3	4	5	6
d	2	1	1	1	2	3	4	5
f	3	2	2	2	1	2	3	4
d	4	3	3	2	2	2	2	3

Edit distance

	ϕ	a	c	d	f	b	d	f
ϕ	0	1	2	3	4	5		
a	1	0	1	2	3	4		
d	2	1	1	1	2	3		
f			2	2	1	2		
d						2		

Edit distance with Ukkonen cutoff

Shang and Merrett's algorithm

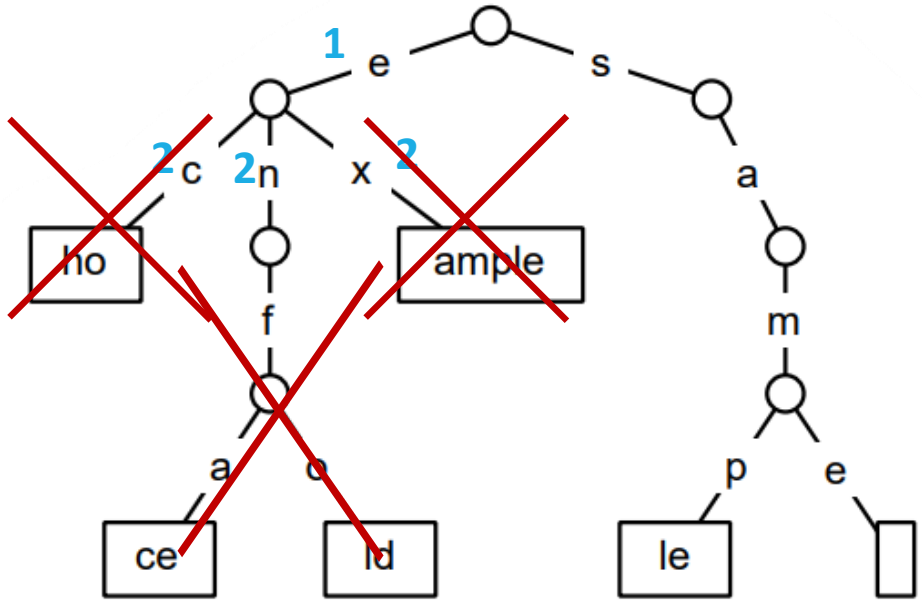
Traverse trie recursively with DFS (Depth First Search)

For each node/character

- Calculate edit distance
- If $editDistance > k$, skip subtrie

Search: 'sane'
$k = 1$

Text:
echo enfold sample enface same example



	ϕ	a	c	d	f	b	d	f
ϕ	0	1	2	3	4	5	6	7
a	1	0	1	2	3	4	5	6
d	2	1	1	1	2	3	4	5
f	3	2	2	2	1	2	3	4
d	4	3	3	2	2	2	2	3

Edit distance

	ϕ	a	c	d	f	b	d	f
ϕ	0	1	2	3	4	5		
a	1	0	1	2	3	4		
d	2	1	1	1	2	3		
f			2	2	1	2		
d						2		

Edit distance with Ukkonen cutoff

Shang and Merrett's algorithm

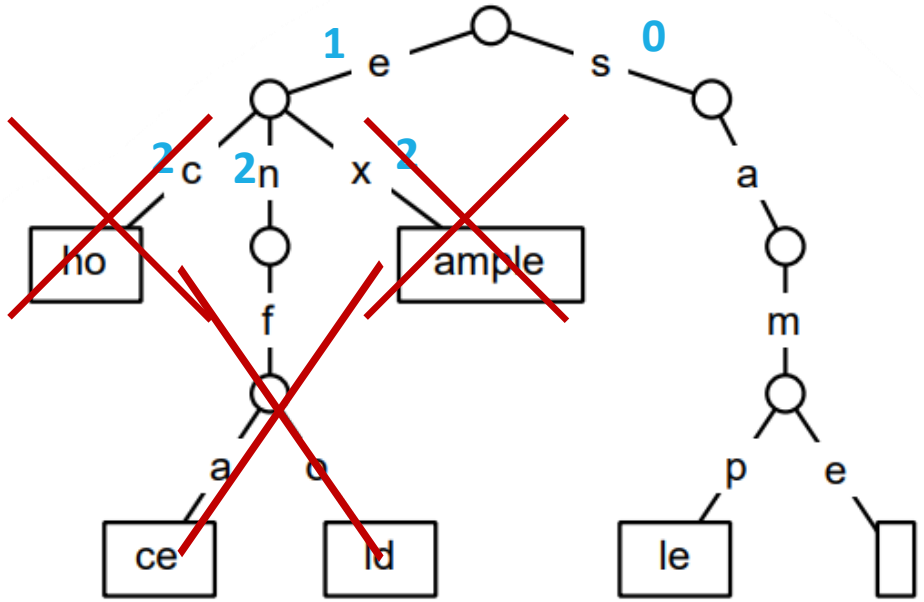
Traverse trie recursively with DFS (Depth First Search)

For each node/character

- Calculate edit distance
- If $editDistance > k$, skip subtrie

Search: 'sane'
$k = 1$

Text:
echo enfold sample enface same example



	ϕ	a	c	d	f	b	d	f
ϕ	0	1	2	3	4	5	6	7
a	1	0	1	2	3	4	5	6
d	2	1	1	1	2	3	4	5
f	3	2	2	2	1	2	3	4
d	4	3	3	2	2	2	2	3

Edit distance

	ϕ	a	c	d	f	b	d	f
ϕ	0	1	2	3	4	5		
a	1	0	1	2	3	4		
d	2	1	1	1	2	3		
f			2	2	1	2		
d						2		

Edit distance with Ukkonen cutoff

Shang and Merrett's algorithm

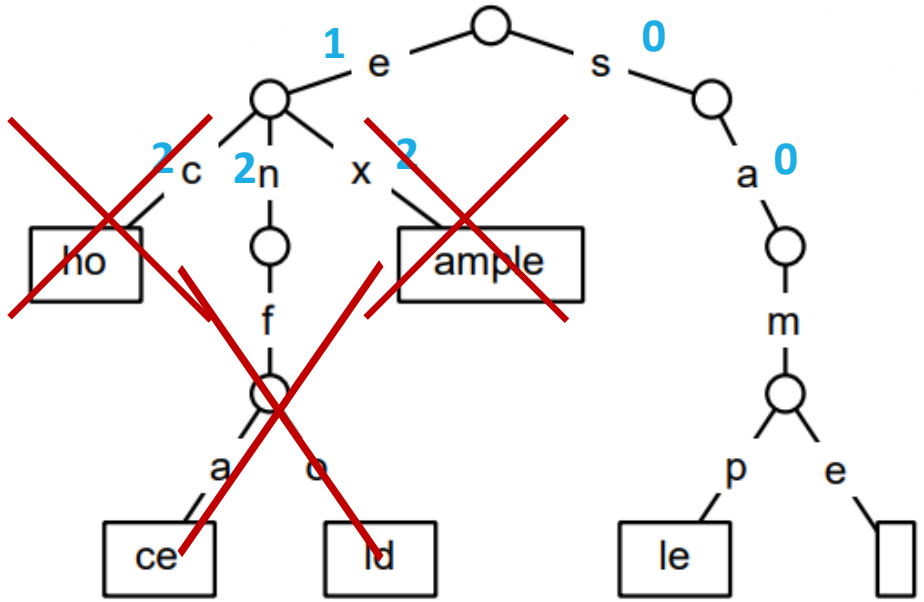
Traverse trie recursively with DFS (Depth First Search)

For each node/character

- Calculate edit distance
- If $editDistance > k$, skip subtrie

Search: 'sane'
$k = 1$

Text:
echo enfold sample enface same example



	ϕ	a	c	d	f	b	d	f
ϕ	0	1	2	3	4	5	6	7
a	1	0	1	2	3	4	5	6
d	2	1	1	1	2	3	4	5
f	3	2	2	2	1	2	3	4
d	4	3	3	2	2	2	2	3

Edit distance

	ϕ	a	c	d	f	b	d	f
ϕ	0	1	2	3	4	5		
a	1	0	1	2	3	4		
d	2	1	1	1	2	3		
f			2	2	1	2		
d						2		

Edit distance with Ukkonen cutoff

Shang and Merrett's algorithm

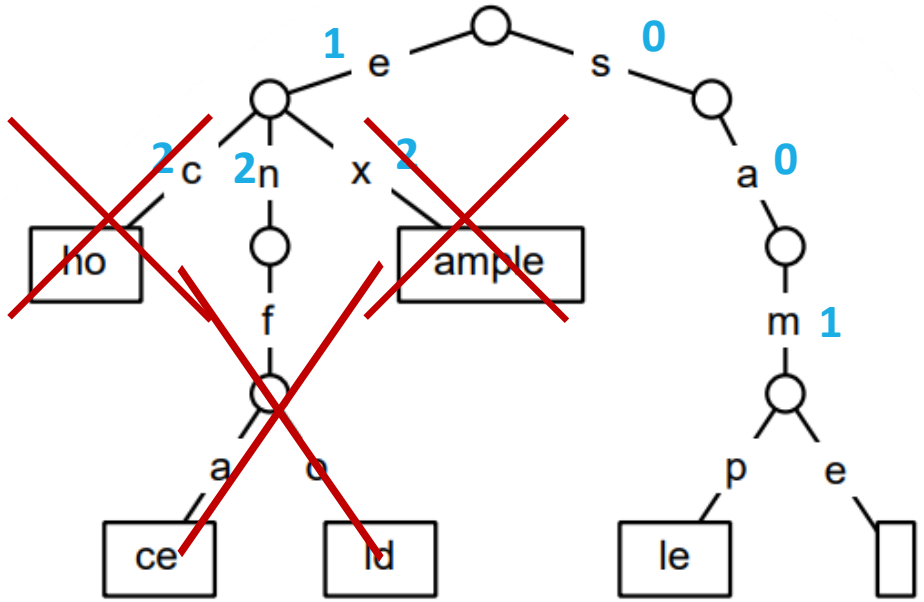
Traverse trie recursively with DFS (Depth First Search)

For each node/character

- Calculate edit distance
- If $editDistance > k$, skip subtrie

Search: 'sane'
$k = 1$

Text:
echo enfold sample enface same example



	ϕ	a	c	d	f	b	d	f
ϕ	0	1	2	3	4	5	6	7
a	1	0	1	2	3	4	5	6
d	2	1	1	1	2	3	4	5
f	3	2	2	2	1	2	3	4
d	4	3	3	2	2	2	2	3

Edit distance

	ϕ	a	c	d	f	b	d	f
ϕ	0	1	2	3	4	5		
a	1	0	1	2	3	4		
d	2	1	1	1	2	3		
f			2	2	1	2		
d						2		

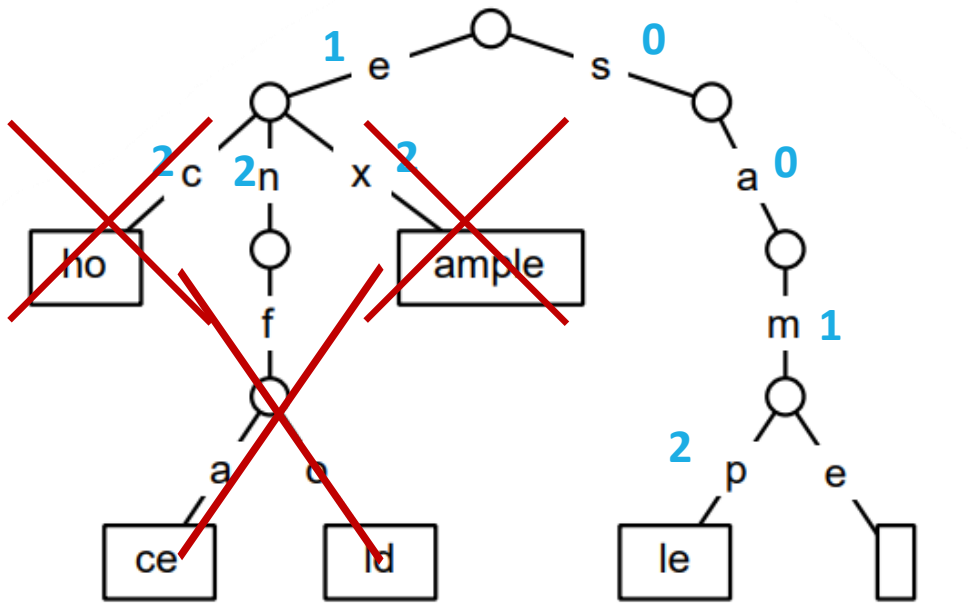
Edit distance with Ukkonen cutoff

Shang and Merrett's algorithm

Traverse trie recursively with DFS (Depth First Search)

For each node/character

- Calculate edit distance
- If $editDistance > k$, skip subtrie



	ϕ	a	c	d	f	b	d	f
ϕ	0	1	2	3	4	5	6	7
a	1	0	1	2	3	4	5	6
d	2	1	1	1	2	3	4	5
f	3	2	2	2	1	2	3	4
d	4	3	3	2	2	2	2	3

Edit distance

	ϕ	a	c	d	f	b	d	f
ϕ	0	1	2	3	4	5		
a	1	0	1	2	3	4		
d	2	1	1	1	2	3		
f			2	2	1	2		
d						2		

Edit distance with Ukkonen cutoff

Shang and Merrett's algorithm

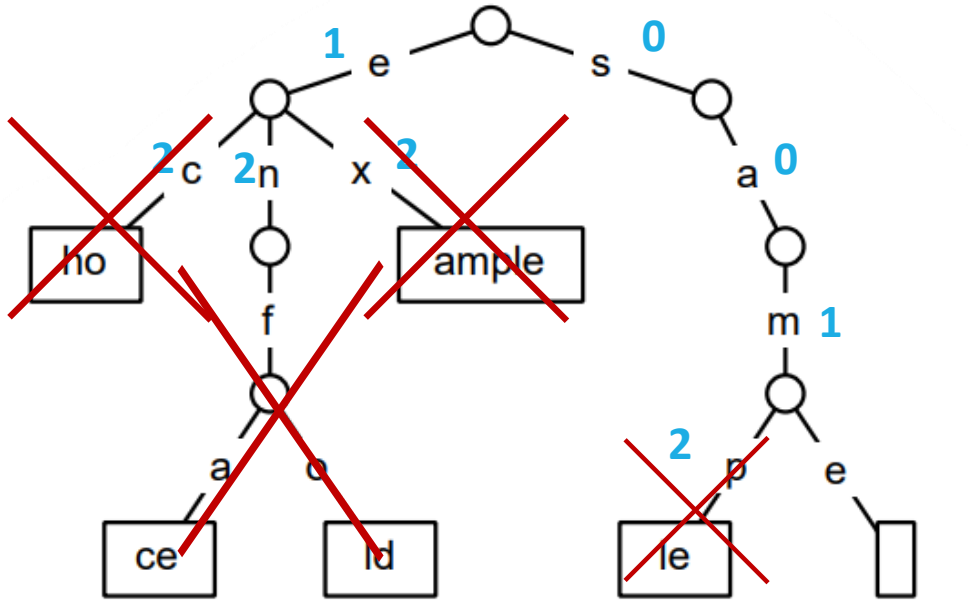
Traverse trie recursively with DFS (Depth First Search)

For each node/character

- Calculate edit distance
- If $editDistance > k$, skip subtrie

Search: 'sane'
$k = 1$

Text:
echo enfold sample enface same example



	ϕ	a	c	d	f	b	d	f
ϕ	0	1	2	3	4	5	6	7
a	1	0	1	2	3	4	5	6
d	2	1	1	1	2	3	4	5
f	3	2	2	2	1	2	3	4
d	4	3	3	2	2	2	2	3

Edit distance

	ϕ	a	c	d	f	b	d	f
ϕ	0	1	2	3	4	5		
a	1	0	1	2	3	4		
d	2	1	1	1	2	3		
f			2	2	1	2		
d						2		

Edit distance with Ukkonen cutoff

Shang and Merrett's algorithm

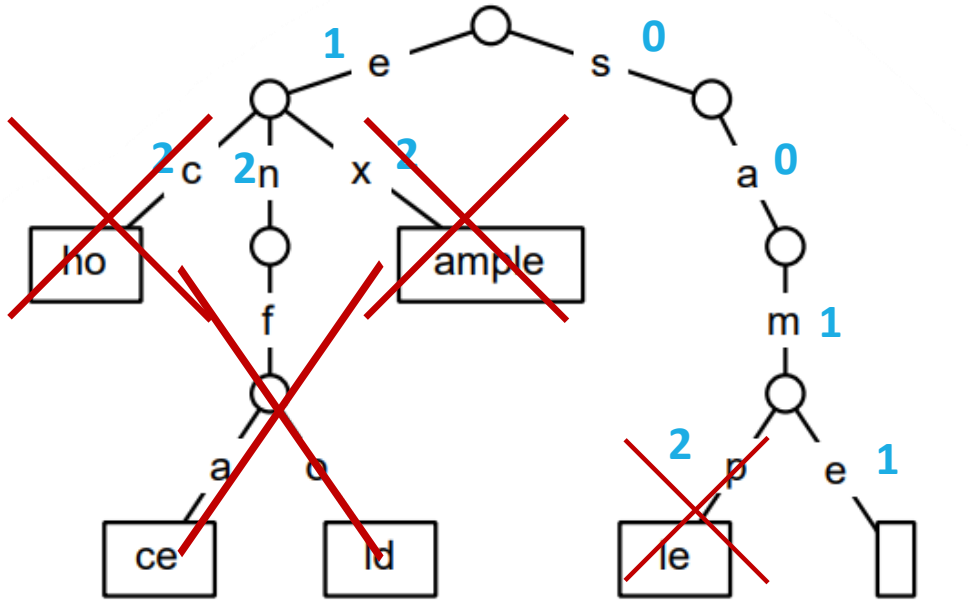
Traverse trie recursively with DFS (Depth First Search)

For each node/character

- Calculate edit distance
- If $editDistance > k$, skip subtrie

Search: 'sane'
$k = 1$

Text:
echo enfold sample enface same example



	ϕ	a	c	d	f	b	d	f
ϕ	0	1	2	3	4	5	6	7
a	1	0	1	2	3	4	5	6
d	2	1	1	1	2	3	4	5
f	3	2	2	2	1	2	3	4
d	4	3	3	2	2	2	2	3

Edit distance

	ϕ	a	c	d	f	b	d	f
ϕ	0	1	2	3	4	5		
a	1	0	1	2	3	4		
d	2	1	1	1	2	3		
f			2	2	1	2		
d						2		

Edit distance with Ukkonen cutoff

Shang and Merrett's algorithm

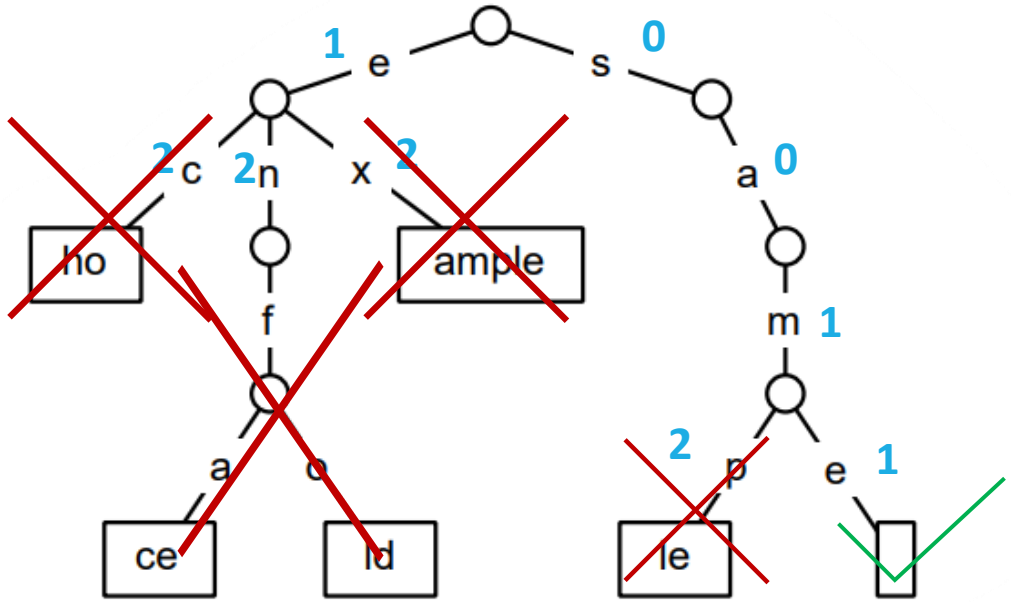
Traverse trie recursively with DFS (Depth First Search)

For each node/character

- Calculate edit distance
- If $editDistance > k$, skip subtrie

Search: 'sane'
$k = 1$

Text:
echo enfold sample enface same example



	ϕ	a	c	d	f	b	d	f
ϕ	0	1	2	3	4	5	6	7
a	1	0	1	2	3	4	5	6
d	2	1	1	1	2	3	4	5
f	3	2	2	2	1	2	3	4
d	4	3	3	2	2	2	2	3

Edit distance

	ϕ	a	c	d	f	b	d	f
ϕ	0	1	2	3	4	5		
a	1	0	1	2	3	4		
d	2	1	1	1	2	3		
f			2	2	1	2		
d						2		

Edit distance with Ukkonen cutoff

Performance

Worst case: $O(k|\Sigma|^k)$

Best case: $k = 0$

Trie

- Grows exponentially
- Better with smaller *k*

agrep

- Linear
- Better with larger *k*

Break point: $k = 2$

Thanks for your attention

T. Shang & H. T. Merrett, *Tries for Approximate String Matching 1995*