Lower bounds for text indexing with mismatches and differences

Laurent Feuilloley (SU)

joint work with
Vincent Cohen-Addad (SU) and
Tatiana Starikovskaya (ENS)

April 2019 · Workshop CoA

Approximate pattern matching

- Pattern matching :
  Is there an occurrence of ‘`seal`’ in the list
  {`plankton, deal, phototrophic, sea, prokaryote`}?

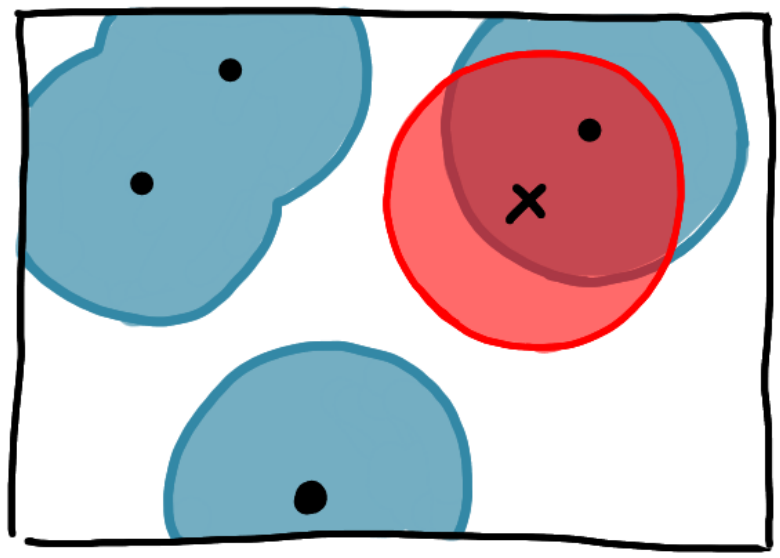
- Approximate pattern matching :
  Is there a word *close to* ‘`seal`’ in the list
  {`plankton, deal, phototrophic, sea, prokaryote`}?

Basic ideas

Approximate pattern matching :
Is there a word close to ‘`seal`’ in the list
{`plankton, deal, phototrophic, sea, prokaryote`}?

Idea 1 : Generate all the words close to ‘`seal`’ and perform exact pattern matching.

Idea 2 : In advance, generate all the words close to a word of the list and perform exact pattern matching.



Let's do the maths

Hamming distance $\leq k$, words of length u, alphabet Σ.
With $k = 2$ and $u = 4$, `seal` $\rightarrow$ `seal` $\rightarrow$ `sfak`.
Then the neighbourhood of a word, has size $\binom{u}{k}|\Sigma|^k \sim u^k|\Sigma|^k$.

$\hookrightarrow$ With the basic ideas :

either the space or the query time, is exponential

Surely, we can do something smarter (?)

Lower bound conjecture

Gonzalo Navarro

Indexed approximate string matching

This is the problem of finding all the approximate occurrences, in a text T of length n, of a pattern Q of length d, both over an alphabet of size $|\Sigma|$.

[...] Although there has been progress on this problem, one still finds that either the index is of exponential size (in k or d or $|\Sigma|$), or the search takes exponential time.

[...] I believe this is a fundamental space/time barrier, but as far as I know this has not been proved.

— *IWOCA open problem session*

Setting

Distances

- Hamming distance : number of replacements
- Edit distance : number of replacements, insertions, deletion

Words at distance 1 from ‘`seal`’ in the list
{`plankton, deal, phototrophic, sea, prokaryote`}

Problems

- Look up : the database is a list of words
- Text indexing : the database is a long text

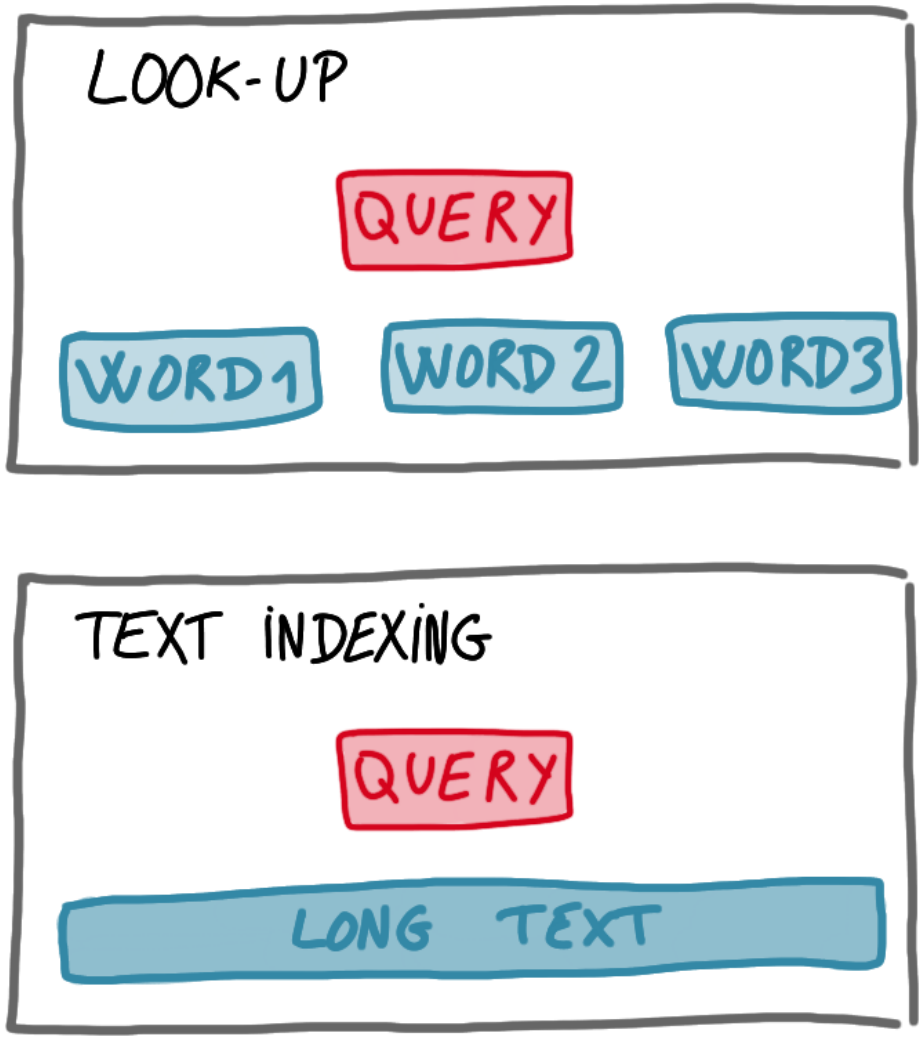
Results

For both distances, and both problems,
two types of exponential space/time lower bounds :

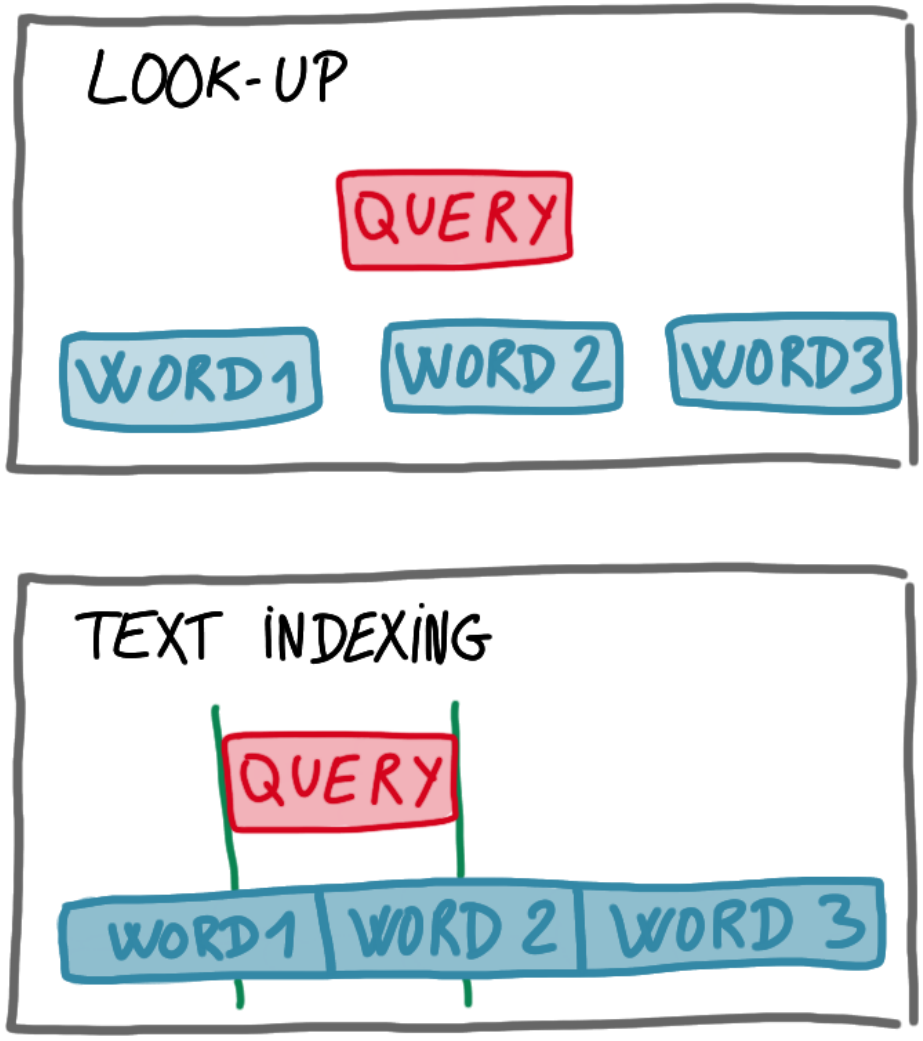
- **Model :** RAM, assuming SETH
  Regime : $k \in \Theta(\log n)$.

- **Model :** pointer-machine
  Regime : $\log(\sqrt{n}) \leq n \ll \log n$.

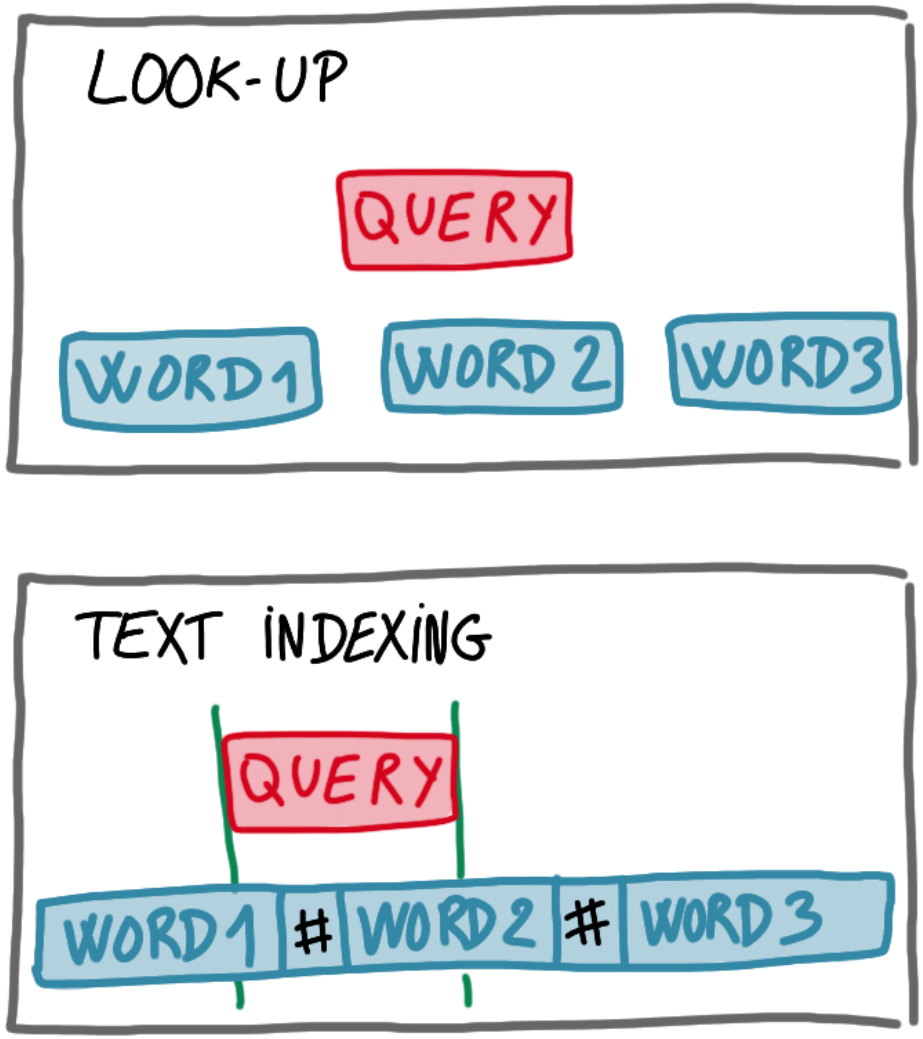
Topic 1 : Look-up $\rightarrow$ Text index



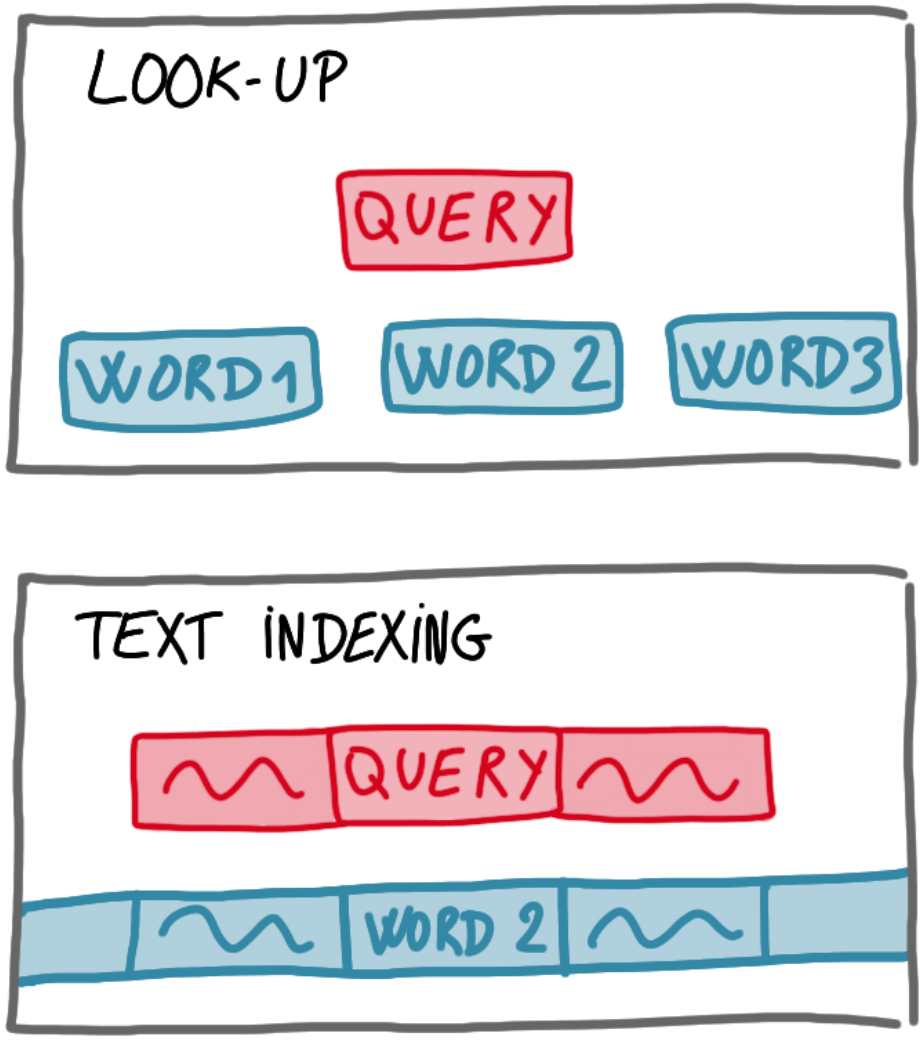
Topic 1 : Look-up $\rightarrow$ Text index



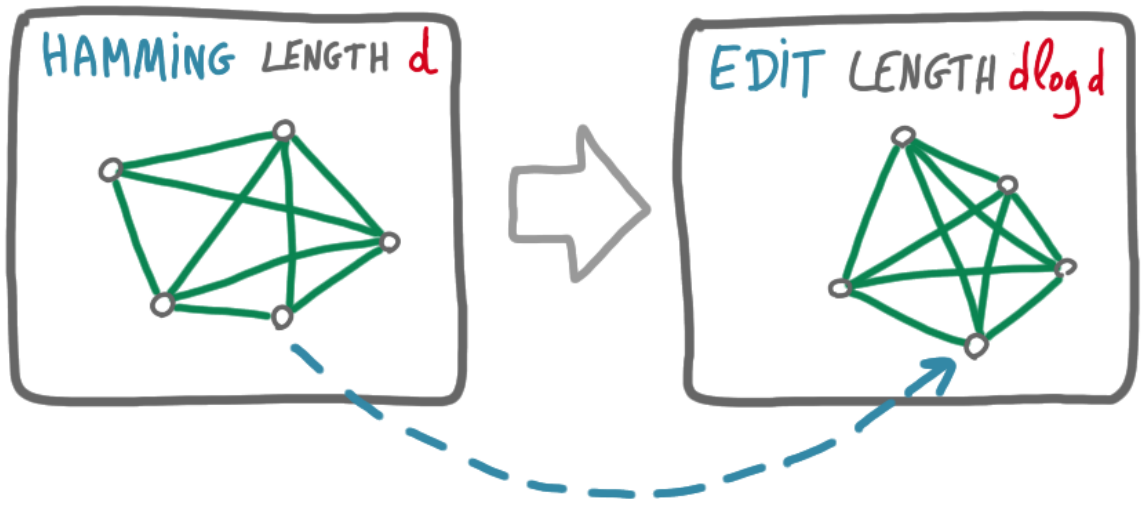
Topic 1 : Look-up $\rightarrow$ Text index



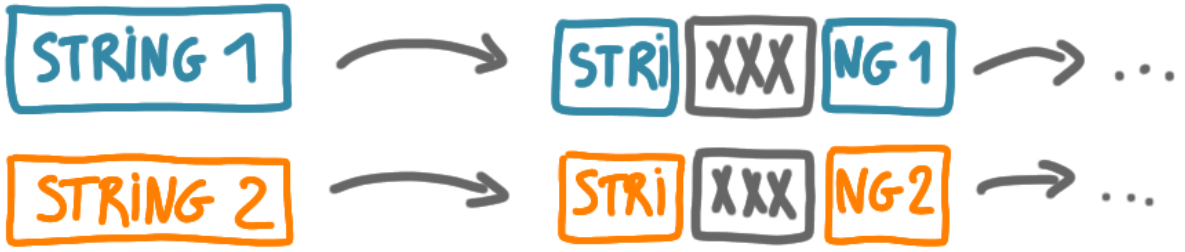
Topic 1 : Look-up $\rightarrow$ Text index



Topic 2 : Hamming to Edit



The idea : inserting "stoppers", to force the alignment.



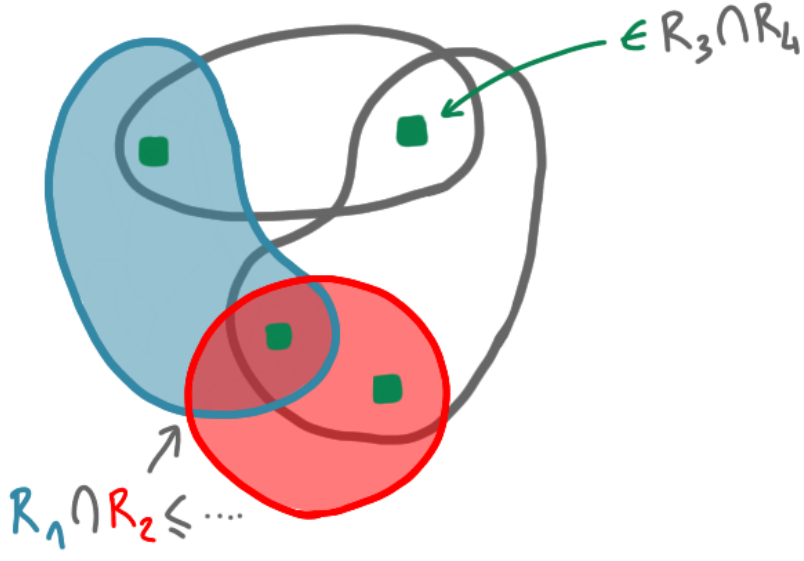
Topic 3 : Transfer to geometry

Theorem (Afshani*) :
For sets of points and regions.
If,

- $\forall p, p \in \geq t$ regions.
- Vol(any $\cap\beta$ regions) $\leq \gamma$

Then,

- stabbing queries has good time-space lower bounds.



The work : find good sets of points and regions, that fit the theorem, for Hamming distance.

Conclusion

- Approximate text-indexing/look-up is an important problem
- We have proved that exponential lower bounds exist.
- But only for some range of parameters.
- Along the way interesting constructions and questions.

Pictures credits : wikicommons ‘File :Seehund.jpg’.