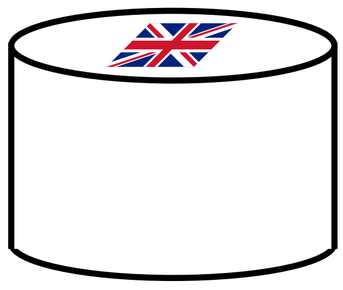
Cross-language High Similarity Search:
Why no Sub-linear Time Bound can be Expected

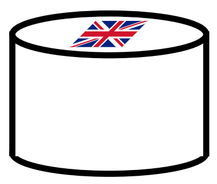
Maik Anderka, Benno Stein, and Martin Potthast

Problem: Cross-language High Similarity Search



Document collection D

Query document q



Subset D_q
$\{d \in D \mid \varphi(q, d) \geq 1 - \varepsilon\}$

Use cases:
- Cross-language plagiarism detection
- Translation search

Naive approach:
- Linear scan using a multilingual IR model
  → compelxity $\mathrm{O}(D)$

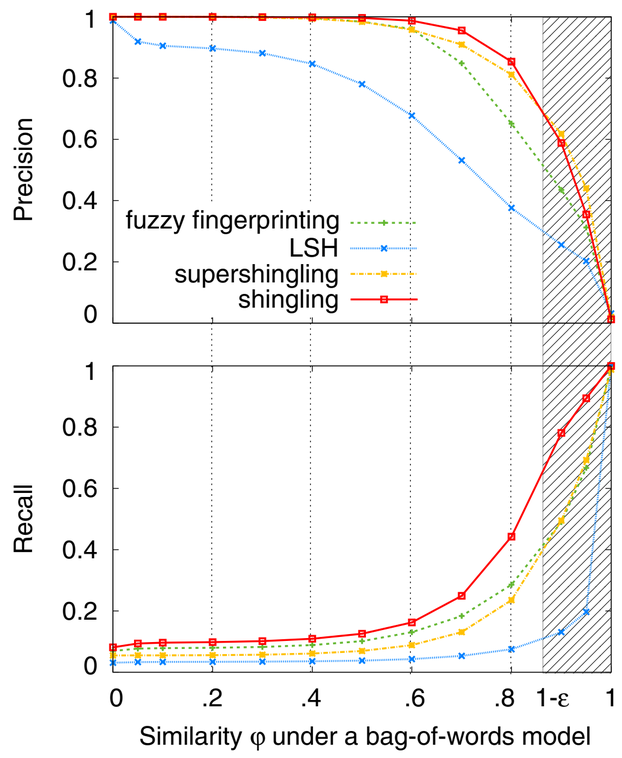
Research question: Can cross-language high similarity search be tackled in sub-linear time?

Background: Monolingual High Similarity Search

Suppose the language of q and D is the same. Then it can be tackled in sub-linear time by fingerprinting or by exhaustive n-gram indexing.

Fingerprinting:
- Compute fingerprints F_q and F_d for q and $d \in D$ using a multi-valued similarity hash-function.
- Consider q and d as similar if their fingerprints intersect:
  $F_q \cap F_d \neq \emptyset \Rightarrow \varphi(q, d) \geq 1 - \varepsilon$, with $0 < \varepsilon << 1$
- Runtime: $\mathrm{O}(D_q)$, whereas $|D_q| << |D|$.

Exhaustive n-gram indexing:
- D is indexed by all n-grams with a reasonable large n, $n \in [5;15]$.
- q is considered as a single n-gram.
- Runtime: $\mathrm{O}(1)$.



Why no Sub-linear Time Bound can be Excpected

Major result: Neither fingerprinting nor exhaustive n-gram indexing can solve *cross-language* high similarity search with an acceptable quality:

- Cross-language similarities are on average 0.5 (cf. plot on the right); hence, with a reasonable ε of ~0.15, D_q nearly contains any document.
- If ε is adjusted to capture more documents (e.g., $\varepsilon = 0.5$) the recall of all fingerprinting approaches drops dramatically as shown above.
- The n-grams of a query and a document written in different languages are not comparable.

Current research is on deriving theoretical performance bounds for cross-language fingerprinting using the locality-sensitive hashing (LSH) framework.