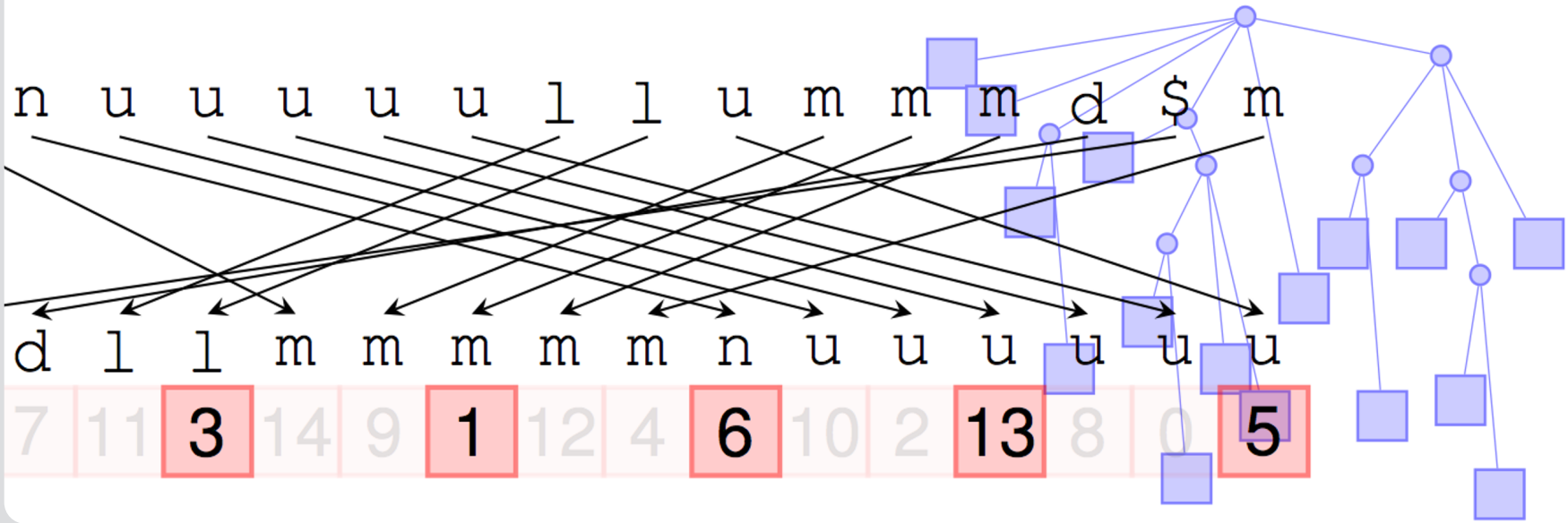
Text Indexing: Lecture 3

Simon Gog – *gog@kit.edu*

Institute of Theoretical Informatics - Algorithmics



Exercise: select(i, c, X)

```
00  select(i, c, X)
01      d ← 0
02      v ← WT.root()
03      while not WT.is_leaf(v) do
04          v ← WT.select_child(v, code(c)[d])
05          d ← d + 1
06      s ← i
07      while not WT.root() = v do
08          s ← WT.select(s, code(v)[d], B_v)
09          d ← d − 1
10          v ← WT.parent(v)
11      return s
```

$code(c)$ is the prefix code of symbol c; B_v is bitvector of node v.
Query time: $\mathcal{O}(\log \sigma)$.

Intermezzo: Top-k document retrieval

Given

- Collection $\mathcal{D}' = \{d_1, \ldots, d_{N-1}\}$
- Each d_i is a string over alphabet $\Sigma' = [2, \sigma]$ sentinel symbol terminated by 1 (also #)
- $\mathcal{D} = \mathcal{D}' \cup d_0$, with $d_0 = 0$.
- „Bag of words" query $Q = \{q_0, q_1, \ldots, q_{m-1}\}$ (unordered set of size m)

Problem

Given a collection $\mathcal{D}$, a query Q of length m, and a similarity measure $\mathcal{S} : \mathcal{D} \times \mathcal{P}_{=m}(\Sigma') \to \mathbb{R}$. Calculate the top-$k$ documents of $\mathcal{D}$ with regard to Q and $\mathcal{S}$. That is a sorted list of document identifiers $\mathrm{T} = \{\tau_0, \ldots, \tau_{k-1}\}$, with $\mathcal{S}(d_{\tau_i}, Q) \geq \mathcal{S}(d_{\tau_{i+1}}, Q)$ for $0 \leq i < k$ and $\mathcal{S}(d_{\tau_{k-1}}, Q) \geq \mathcal{S}(d_j, Q)$ for $j \notin \mathrm{T}$.

Example

Fix a concatenation $\mathcal{C}$ of $\mathcal{D}$.

$i =$	0	1	2	3	4	5	6	7	8	9	10	11	12	13
$\mathcal{C}^{word} =$	LA	O	LA	#	O	LA	LA	LA	#	O	O	LA	#	$
$\mathcal{C} =$	2	3	2	1	3	2	2	2	1	3	3	2	1	0
	d_1				d_3					d_2				d_0

- $\mathcal{S}^{sfreq}(d, q) := f_{d,q}$ (i.e. single term frequency ranking)
- $\mathcal{S}^{sfreq}(d_0, LA) = 0$,
  $\mathcal{S}^{sfreq}(d_1, LA) = 2$,
  $\mathcal{S}^{sfreq}(d_2, LA) = 1$,
  $\mathcal{S}^{sfreq}(d_3, LA) = 3$.
- Top-2: $\mathrm{T} = \{3, 1\}$

Okapi BM25 similarity measure

Successful IR similarity measure:

$$S_{Q,d}^{\text{BM25}} = \sum_{q \in Q} \underbrace{\frac{(k_1+1) f_{d,q}}{k_1 \left(1 - b + b \frac{n_d}{n_{\text{avg}}}\right) + f_{d,q}}}_{=w_{d,q}} \cdot \underbrace{f_{Q,q} \cdot \ln\left(\frac{N - F_{\mathcal{D},q} + 0.5}{F_{\mathcal{D},q} + 0.5}\right)}_{=w_{Q,q}}$$

depends on 3 document-dependent factors:

- $f_{d,q}$ term frequncy
- $F_{\mathcal{D},q}$ document frequency (# of distinct ds which contain q)
- n_d length of document d