Laborator nr. 3
Construirea indecşilor direcţi/indirecţi

an univ. 2018 – 2019

1 Aplicaţie propusă

1. Construiţi **index-ul direct cantitativ** pentru un set de documente HTML alese de voi. Indexul va stoca atât cheile (documentele indexate), cât şi valorile asociate (cuvintele din cadrul documentelor) cheilor.

Aplicaţia va primi ca *intrare* calea catre un director ce conţine documentele de indexat (**Atenţie**: directorul poate conţine şi sub-directoare). *Ieşirea* aplicaţiei constă din **2** fişiere, astfel:

1. un prim fişier text de *mapare*: pentru fiecare document HTML (numele fişierului va fi stocat prin cale completă) se va indica numele fişierului care conţine indexul direct asociat;
2. unul sau mai multe fişiere ce vor conţine indexul direct determinat.

2. Pornind de la pseudocodul de mai jos, implementaţi o aplicaţie care va determina indexul invers corespunzător pentru colecţia de documente pentru care a fost determinat indexul direct cantitativ în cadrul modulului anterior.

Algorithm 1 BSBI() (preluare din [1], cap. 4)

```
1: i ← 0
2: while all documents have not been processed do
3:    n ← n + 1
4:    block ← PARSENEXTBLOCK()
5:    BSBI-INVERT(block)
6:    WRITEBLOCKTODISK(block, fn)
7: end while
8: MERGEBLOCKS( f1, . . . , fn; fmerged)
```

Aplicaţia va primi ca intrare **indexul direct cantitativ** determinat în etapa anterioară, şi va furniza, ca ieşire, **2** fişiere, astfel:

1. un prim fişier text de *mapare*: pentru fiecare cuvânt determinat se va indica numele fişierului care conţine indexul invers asociat;
2. un fişier ce conţine **indexul invers cantitativ**.

Bibliografie

[1] Christopher D. Manning et. al. Introduction to Information Retrieval. `http://nlp.stanford.edu/IR-book/information-retrieval-book.html`, 2009.