

Laborator nr. 2

Extragerea cuvintelor din documente text (2)

an univ. 2017 – 2018

1 Aplicație propusă

Completați aplicația dezvoltată în cadrul laboratorului anterior astfel încât să puteți procesa un set de fișiere TEXT stocate în cadrul unui director [1]. În plus față de prelucrările anterioare, pentru fiecare cuvânt determinat se vor realiza următoarele procesări suplimentare:

1. cuvântul din text, care a fost determinat în cadrul iterației curente, va fi testat contra unei liste de excepții – un dicționar al unei limbi nu conține, de exemplu, nume proprii; dacă acest cuvânt se regăsește în lista de excepții, atunci se va trece la următoarele etape de procesare (contorizare număr de apariții, etc.);
2. cuvântul din text, care a fost determinat în cadrul iterației curente, va fi testat contra unei liste de *stop-word*-uri – cuvintele de legătură care în mod uzual nu aduc informații noi pentru motoarele de căutare [2]; dacă acest cuvânt curent determinat se regăsește într-o astfel de listă de *stop-word*-uri, atunci acesta va fi eliminat din procesările ulterioare.

Observație 1:

Ar fi util ca aplicația dezvoltată de voi să proceseze directoarele *incremental*, pe măsură ce sunt identificate fișierele din cadrul directorului de lucru. Trebuie să țineți cont de faptul că un director poate conține subdirectoare.

Bibliografie

- [1] Oracle. Creating and Reading Directories. <https://docs.oracle.com/javase/tutorial/essential/io/dirs.html>.
- [2] RANKS NL. Stop Words. <http://www.ranks.nl/stopwords>.