Laborator nr. 2
Extragerea cuvintelor din documente text (2)

an univ. 2018 – 2019

1 Aplicaţie propusă

Completaţi aplicaţia dezvoltată în cadrul laboratorului anterior astfel încât să puteţi procesa un set de fişiere TEXT stocate în cadrul unui director [1]. În plus faţă de prelucrările anterioare, pentru fiecare cuvânt determinat se vor realiza următoarele procesări suplimentare:

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

Observaţie 1:
Ar fi util ca aplicaţia dezvoltată de voi să proceseze directoarele *incremental*, pe măsură ce sunt identificate fişierele din cadrul directorului de lucru. Trebuie să ţineţi cont de faptul că un director poate conţine 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`.