Laborator 10: Prelucrarea limbajului natural

Construiți un model de limbă pentru poeziile lui Eminescu (http://www.gutenberg.org/ebooks/35323). Modelul va fi construit pe cuvinte întregi. Folosiți modelul pentru a da un scor de similitudine între un document în limba română și poeziile lui Eminescu.

(0.2) Eliminați stop-word-urile din text. Pentru Python puteți folosi https://www.geeksforgeeks.org/removing-stop-words-nltk-python/ . Alternativ, folosiți direct lista de aici: https://raw.githubusercontent.com/stopwords-iso/stopwords-ro/master/stopwords-ro.txt

(0.2) Construiți modelul de limbă pe întreg fișierul. Pentru fiecare cuvânt din text contorizați numărul de apariții. După parcurgerea întregului document, valorile pentru fiecare intrare vor fi ponderate(împărțite) la numărul total de cuvinte unice din fișier. Pentru a nu lucra cu multe zecimale puteți înmulți rezultatul cu 1.000.000. Salvați modelul de limbă într-un fișier txt sau json.

(0.2) Folosiți același proces pe un fișier nou și comparați rezultatul cu modelul de limbă anterior. Calculați un scor de similitudine adunând diferența în modul dintre scorul unui cuvânt în modelul fișierului nou și scorul cuvântului din model (care poate fi considerat 0 în caz ca lipsește). Afișați și media diferenței în modul (suma anterioară împărțită la numărul de cuvinte unice din fișier).