Information Retrieval & Text Mining

Marius Popescu

popescunmarius@gmail.com

2024 - 2025

Proiect 1

Sistem de Regasirea Informatiei pentru Limba Romana

Instrumente

- Apache Lucene http://lucene.apache.org/
- Apache Tika https://tika.apache.org/
- Orice alta biblioteca pe care o considerati utila

Doua componente separate

- Indexer: porneste de la un set de documente si creaza un “inverted index” pe care il salveaza.
- Searcher: porneste de la un “inverted index” (creat anterior de indexer) si poate raspunde la (oricate) intrebari

Indexer

Indexeaza continutul tuturor fisierelor de tip txt, doc(x), pdf dintr-o locatie (folder) specificata

Indexer & Searcher

Se vor ocupa / rezolva de problemele specifice limbii romane:

- Eliminarea diacriticelor
- Eliminarea “stop words” pentru limba romana
- Stemming pentru limba romana

Eliminarea Diacriticelor

- Daca un document contine cuvantul “cămaşă” si se cauta “camasa” documentul va fi regasit
- Si invers, daca un document contine cuvantul “camasa” si se cauta “cămaşă” (sau “cămasa”, etc.) documentul va fi regasit

Eliminarea “stop words” pentru limba romana

Daca se cauta “şi” sau “si”, “că” sau “ca”, etc., cautarea nu va intoarce nici un rezultat

Stemming pentru limba romana

- Daca un document contine cuvantul “mamei” si se cauta “mama” (sau “mamele”, etc.) documentul va fi regasit
- Si invers, daca un document contine cuvantul “mama” si se cauta “mamei” (sau “mamelor”, etc.) documentul va fi regasit

Atentie la interactiunea dintre diacritice si celelalte componente

- Lista de “stop words” e cu diacritice sau nu?
- Stemmer-ul pentru romana din Lucene e gandit pentru diacritice. Cand eliminati diacriticele?

  Daca un document contine cuvantul “căruță” si se cauta “carutele” documentul va fi regasit

  Si invers ...

Predarea pana in saptamana 8 la laborator